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

Deliverable 4.1:
Benchmark definitions
Due date: 31/10/2012
Delivery date: 31/10/2012
Delivery type: Report , public

Jean-Marie Beckers Alexander Barth
University of Liège, BELGIUM

Peter Jan Van Leeuwen
University of Reading, UK

Lars Nerger
Alfred-Wegener-Institut, GERMANY

Arnold Heemink Nils van Velzen
Martin Verlaan
Delft University of Technology, NETHERLANDS

Pierre Brasseur Jean-Michel Brankart Guillem Candille
CNRS-LEGI, FRANCE

Pierre de Mey
CNRS-LEGOS, FRANCE

Laurent Bertino
NERSC, NORWAY

Chapter 1

Introduction

The SANGOMA proposal starts with the observation that the majority of current MyOcean products (implementing the GMES marine core service for ocean monitoring and forecasting) are based on suboptimal assimilation methods, providing a limited information about the uncertainties in the model nowcast/forecast. To go beyond the current situation, the main objective of SANGOMA is to advance the status of probablistic assimilation methods and their applicability to operational MyOcean systems. This requires (i) establishing a European network of experts in probabilistic data assimilation, and (ii) providing a harmonized access to state-of-the-art concepts, algorithms and softwares (often developed by individual efforts). To set up such an effective connection between the SANGOMA partners and the MyOcean consortium, a key element of the project is to assess the performance of the methods in a variety of testcases, including realistic assimilation problems. The implementation of these testcases (benchmarks) is the purpose of WP4, and the objective of this first deliverable of WP4 is to document the standard data assimilation problems defining the SANGOMA benchmarks.

Usually, assimilation methods are first developed and tested using quite simple and idealized assimilation problems. This is important to check the mathematical consistency of the method, and to understand how it works without being blinded by real world approximations. On the other hand, the main purpose of assimilation methods is to solve real world problems, and it is also important to evaluate their performance with problems of higher complexity, and their robustness to inescapable approximations. This is why the SANGOMA project includes a hierarchy of benchmarks of increasing complexity: (i) a small case benchmark, based on the Lorenz-96 model with 40 variables, (ii) a medium case benchmark, based on an idealized square ocean model, and (iii) a large case benchmark, based on a realistic North Atlantic model at $1/4^\circ$ resolution. To be complient with most MyOcean systems, the last two benchmarks are based on the NEMO model (see chapter 2 below).

In the definition of each benchmark, what must be specified is (i) the forward model that is used to describe the system (see chapter 2), and (ii) the inverse problem that must be solved (see chapter 3). It is only in a second step (not included in this definition document) that various methods can be compared according to their relative merits in terms of reliability (in the description of the prior

and posterior probability distributions), resolution (information gain or uncertainty reduction), complexity (*e.g.* number of free parameters that must be tuned by the user), numerical cost,... This is why each benchmark requires defining appropriate metrics to measure the assets of every method (see chapter 4). Nevertheless, at this stage of the project, it is also necessary that these definitions and metrics remain flexible enough to be adjusted to the specifities of each inverse problem (which still need to be fully specified in the next deliverable).

Chapter 2

Model configurations

This chapter only refers to the direct problem and the description of the forward models. Model configurations of increasing complexity will be used to elaborate the benchmarks. These model configurations will include:

1. the Lorenz-96 model with 40 variables for the small case benchmark,
2. an idealized square ocean configuration of NEMO for the medium case benchmark, and
3. a realistic North Atlantic configuration of NEMO (at a $1/4^\circ$ resolution) for the large case benchmark.

These model configurations are made available to the SANGOMA partners as explained in Appendix C.

2.1 Small case benchmark

The Lorenz-96 (L96) model is defined by:

$$\frac{dx_i}{dt} = x_{i-1}(x_{i+1} - x_{i-2}) - x_i + F, \quad i = 1, \ldots, n \tag{2.1}$$

with cyclic index: $x_{i-n} = x_i = x_{i+n}$.

Numerics. The differential equation are solved with a fourth-order Runge-Kutta scheme with time step $\Delta t = 0.05$, corresponding to a geophysical time of 6h. Other specifications: $n = 40$, $F = 8$, and the spin-up is initialized by choosing $F = 8.01$ at the 20-th grid point (Van Leeuwen 2010).

2.2 Medium case benchmark

The medium case benchmark is based on an idealized configuration of the NEMO primitive equation ocean model (as described in Cosme et al. 2010): a square and 5000-meter deep flat bottom ocean at mid latitudes (the so called square-box or SQB configuration).

Physics. In this square basin (between 25°N and 45°N), a double gyre circulation is created by a constant zonal wind forcing blowing westward in the northern and southern parts of the basin and eastward in the middle part of the basin. The domain is closed and the lateral boundaries are frictionless. The western intensification of these two gyres produces a western boundary current that feed an eastward jet in the middle of the square basin (see Fig. 2.1 below, showing a snapshot of the model sea surface height). This jet is unstable so that the flow is dominated by chaotic mesoscale dynamics, with largest eddies that are about 100 km wide, and to which correspond velocities of about 1 m/s and dynamic height differences of about 1 meter. All this is very similar in shape and magnitude to what is observed in the Gulf Stream (North Atlantic) or in the Kuroshio (North Pacific).

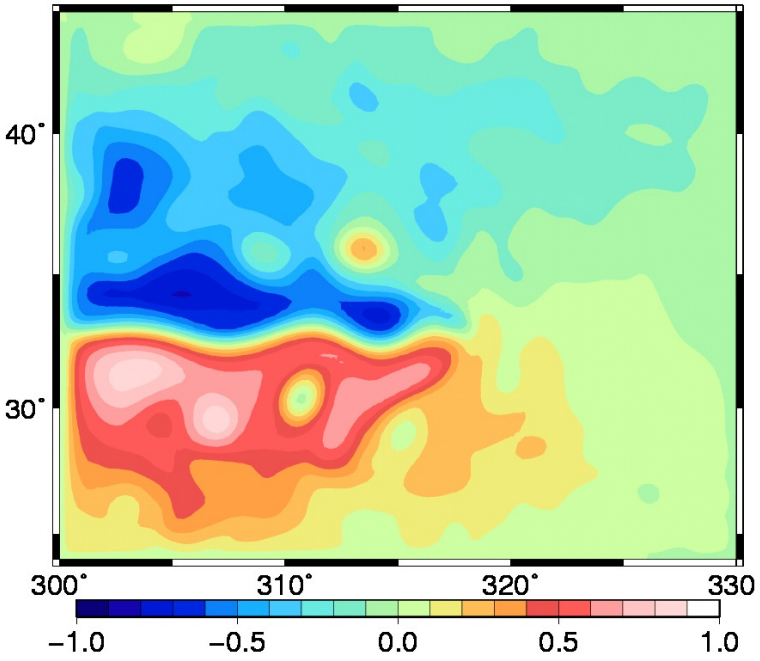

Figure 2.1: Snapshot of sea surface height from the SQB configuration of NEMO.

The model is started from rest with uniform stratification, and the main physical parameters governing the dominant characteristics of the flow are the initial stratification, the wind stress, the bottom friction and the lateral mixing parameterization. The initial stratification is produced using an analytical temperature profile:

$$T(z) = 25 + 24.57[\exp(-z/800) - 1] \quad [\text{in } ^\circ\text{C}] \tag{2.2}$$

and uniform salinity ($S = 35$). The zonal wind stress is prescribed as:

$$\tau_x(\phi) = -10^{-4} \cos\left(2\pi \frac{\phi - \phi_{\min}}{\phi_{\max} - \phi_{\min}}\right) \quad [\text{in N/m}^2] \tag{2.3}$$

where ϕ is the latitude.

Linear bottom friction is applied with drag coefficient $C_D = 4 \times 10^{-4}$. Lateral mixing of momentum and tracer is performed with a bilaplacian operator with coefficient $A_h = -8 \times 10^{10}$ m^4/s.

Numerics. The primitive equations are discretized on an Arakawa C grid, with a horizontal resolution of $1/4^\circ \times 1/4^\circ \cos(\phi)$ and 11 z-coordinate levels along the vertical. The model uses a 'free surface' formulation, with an additional term in the momentum equation to damp the faster external gravity waves. Momentum advection is performed with an energy conserving numerical scheme in vector form, and tracer advection is performed with a 2nd order centered scheme. Time stepping is performed with a leap frog scheme, with a time step $\Delta t = 900$ s.

2.3 Large case benchmark

The large case benchmark is based on a realistic configuration of the NEMO ocean model, for the North Atlantic Ocean, at a $1/4^\circ$ resolution (Barnier et al, 2006), and including an ecosystem model component (Ourmières et al., 2009), figuring the operational MyOcean systems. This model was selected because it has been used in numerous assimilation studies before (Ourmières et al., 2009, Béal et al, 2010, Doron et al. 2011) and because it is based on the NEMO ocean model which is used by most MyOcean Monitoring and Forecasting Centres. This implementation will be used for selected simulations to assess new techniques with real data.

Physics. The model circulation is simulated by the OPA code using the free surface formulation. Prognostic variables are the three-dimensional velocity fields and the thermohaline variables. The model domain covers the North Atlantic basin from 20°S to 80°N and from 98°W to 23°E. The horizontal resolution is 1/4 of a degree, which is considered as eddy-permitting in the mid-latitudes where the Rossby radius of deformation is about 100 km. (see Fig. 2.2 below, showing a snapshot of the model sea surface height). Lateral mixing of momentum and tracers is modelled with a biharmonic operator, vertical mixing is modelled by the TKE turbulence closure scheme, and convection is parameterized with enhanced diffusivity and viscosity. The forcing fluxes are calculated via bulk formulations, using the ERA40 atmospheric forcing fields. Buffer zones are defined at the southern, northern and eastern (Mediterranean) boundaries (which are closed), with restoring to Levitus climatology.

Ecosystem. The biogeochemical model used is LOBSTER (LOCEAN Biogeochemical Simulation Tool for Ecosystem and Resources). It is nitrogen-based and contains six prognostic variables: nitrate, ammonium, phytoplankton, zooplankton, detritus and semilabile dissolved organic matter. In the model, the bottom of the euphotic layer is considered to be at a constant depth of 191 m. In the euphotic layer, the biogeochemical functionalities work as described in Fig. 2.3 (see Lévy et al., 2005 for more detail about the model equations). As LOBSTER features two nutrients, new production and regenerated production can

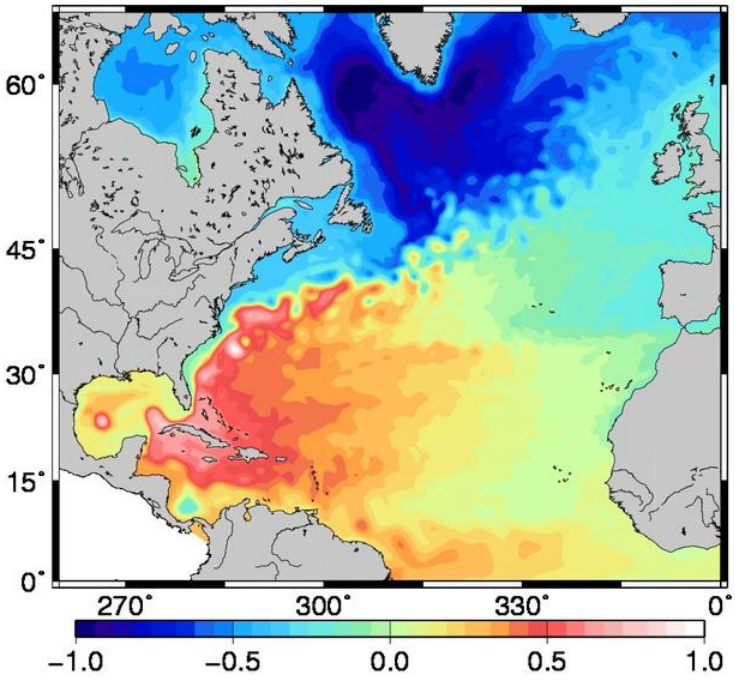

Figure 2.2: Snapshot of sea surface height from the NATL025 configuration of NEMO.

be distinguished. Below the euphotic layer, the model considers very simple parameterizations of decay to nitrate, detritus sedimentation and remineralization of zooplankton mortality.

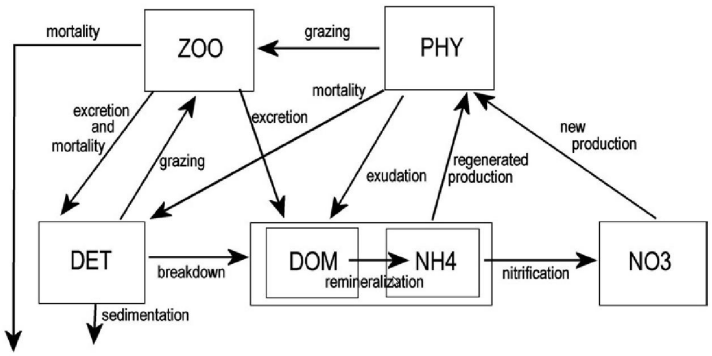

Figure 2.3: Functionalities of the LOBSTER model in the euphotic layer.

Numerics. The primitive equations are discretized on an Arakawa C grid, with a horizontal resolution of $1/4^\circ \times 1/4^\circ \cos(\phi)$. Vertical discretization is done on 45 geopotential levels, with a grid spacing increasing from 6 m at the surface to 250 m at the bottom (with partial step to better discretize the bottom topography). The model uses a 'free surface' formulation, with an additional term in the momentum equation to damp the faster external gravity waves. Momentum advection is performed with an energy and enstrophy conserving numerical scheme in vector

form, and tracer advection is performed with the TVD scheme. Time stepping is performed with a leap frog scheme, with a time step $\Delta t = 2400$ s. The bio-geochemical model is coupled on-line to the circulation model, with a coupling frequency equal to the circulation model time step.

Chapter 3

Specification of benchmarks

This chapter is dedicated to the inversion problem and the inherent specifications for the assimilation process. The SANGOMA benchmark system is made of a hierarchy of assimilation problems of increasing complexity, specified as follows. The aim of this chapter is to describe a framework rigid enough to enable relevant comparisons between stochastic methods, but flexible enough to enable large investigations on each method. For instance, the ensemble size is considered as a free parameter.

3.1 Small case benchmark

The small case benchmark is based on identical twin experiments performed with the Lorenz-96 model with 40 state variables.

Time settings. The model is spun up from $t = 0$ to $t = 150$ (remind $\Delta t = 0.05$). Assimilation experiments start at $t_0 = 150$ and end at $t_1 \in [1000, 2000]$. These values are in agreement with 'standards' used in several assimilation studies with the Lorenz-96 model (*e.g.* Nakano et al. 2007, Sakov et al. 2012) and could be adjusted.

Uncertainties in the system. Uncertainty is introduced in the initial condition of the assimilation experiments at t_0. This uncertainty is assumed Gaussian with zero mean and covariance $\mathbf{P}_0$ which is set to the covariance of the variability of the model spin-up in $t \in [100, 150]$.

Observations. Observations are extracted from the reference simulation (without perturbation of the initial condition). Observations are available for all state variables at every time step. Observation error is assumed Gaussian with zero mean and covariance $\mathbf{R} = \sigma^2\mathbf{I}$, with $\sigma \in [1, 2]$. However, to increase the complexity of this small case benchmark and make non-Gaussian behaviours more apparent, it must also be possible to reduce the spatio-temporal density of the observations (*cf.* references above).

3.2 Medium case benchmark

The medium case benchmark is based on identical twin experiments performed with the square-box NEMO configuration.

Time settings. The model is spun up from rest for 40 years (remind $\Delta t = 15$ min). Assimilation experiments start at $t_0 = 40$ years and end at $t_1 = 42$ years (flexible values). The model restart file after a 40-years spin-up will be delivered to all participants.

Uncertainties in the system. Uncertainty is introduced in the initial condition of the assimilation experiments at t_0. This uncertainty is assumed Gaussian with zero mean and covariance $\mathbf{P}_0$. $\mathbf{P}_0$ is set to the covariance of the variability of the model spin-up between years 20 and 40 (sampled every 5 days).

Observations. Observations are extracted from the reference simulation (without perturbation of the initial condition). Observations are assumed available for sea surface height and for some vertical profiles for temperature, every 5 days. As in small case benchmark, the observations coverage could be degraded by simulating satellites traces (see Fig. 3.1 as examples; the code will be delivered to all participants or at least the 'observations' files corresponding to the restart file). Observation error is assumed Gaussian with zero mean and covariance $\mathbf{R} = \sigma^2\mathbf{I}$, with $\sigma = 5$ cm.

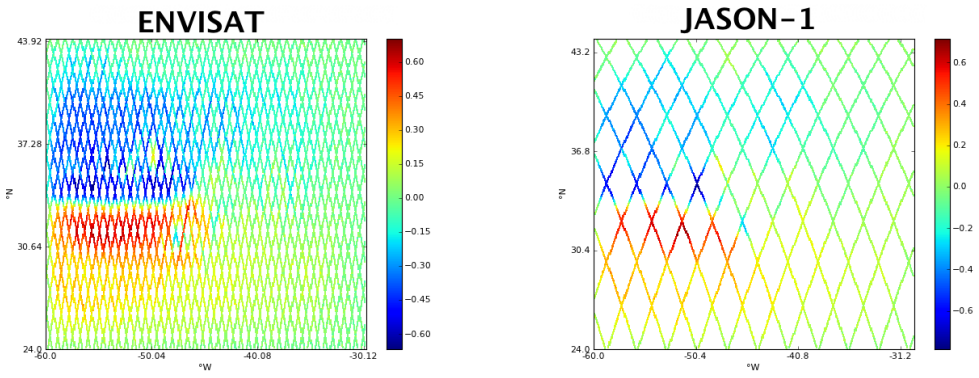

Figure 3.1: Examples of simulated altimetric SSH measurements (in meters) in SQB domain for Envisat and Jason-1, after a whole cycle for both.

3.3 Large case benchmark

The large case benchmark is based on realistic assimilation experiments performed with the North Atlantic NEMO configuration at $1/4°$ resolution.

Time settings. A reference interannual simulation is available between January 1989 and December 2010 (with restart files every year). Assimilation experiments start on January 1st, 1998 (t_0) and last for at least 4 years.

Uncertainties in the system. Uncertainties in the system occur for many different reasons (model dynamics, parameters, forcing, initial and boundary conditions). It is an important duty of the assimilation system to make correct assumptions about the uncertainties. The purpose of a realistic benchmarks is thus also to evaluate the quality of these assumptions.

Observations. The three following observations datasets will be assimilated:

- Altimetric data (together with appropriate geoid data);
- ARGO float data (temperature and salinity);
- Ocean colour data.

An independent observations dataset should be defined for validation purpose (to be determined between LEGI-CNRS and ULg partners who are committed to work on this benchmark).

Chapter 4

Definition of metrics

Metrics are defined to *objectively* compare the different stochastic assimilation methods tested on each benchmark. Here, the mathematical measure (or score) **and** the evaluated objects form what we call 'metric'. Basically the scores are the same for each experiment, but the hierarchy of SANGOMA benchmarks sets a hierarchy of designed specifications to correspond to the purpose of each one.

Scores. The quality of the probability distributions produced by the stochastic assimilation systems are evaluated and compared through 2 probabilistic attributes: the statistical consistency, or reliability, and the statistical variability, or resolution/information/entropy (*e.g.* Toth et al 2003). Details on these scores are presented in appendix A.

The reliability measures the agreement between an estimated and a verified distribution. Many scores can be used to assess this attribute:

- Rank histogram and its multivariate extension (Gneiting et al. 2008). A perfectly reliable system shows a flat rank histrogram. The non-reliability can be compared to the flatness by a χ^2 test. For the multi-variate verification, a rank histogram can also be built from a minimum spanning tree process (Gombos et al. 2007).
- Random centered reduced variable and its matricial extension (Candille et al. 2007). This diagnosis enables a partition of the reliability into bias and dispersion (at least for univariate case): for a perfectly reliable system, one has to get a null bias and a unit dispersion.
- Reliability component of the Brier score and of the continuous ranked probability score (CRPS, Candille and Talagrand 2005). Scores negatively oriented and equal to zero for a perfect reliable system.

The resolution measures the system ability in separating relevant situations. For instance a system always producing the climatological distribution is perfectly reliable but provides no information (except the climatology of course). Many scores can be used to assess the resolution attribute:

- Resolution component of the Brier score (Murphy 1973). Score negatively oriented and equal to zero for a perfect deterministic system (best value).

The resolution is usually compared to the *uncertainty* (part of the score which only depends on the verification data set). If the resolution is greater than the *uncertainty*, the system is considered useless.

- The resolution component of the CRPS and its mutivariate extension, the energy score (Gneiting et al. 2008). Score negatively oriented and equal to zero for a perfect deterministic system. Also, if the resolution is greater than the uncertainty, the system is considered useless.
- Entropy (Gneiting et al. 2008). Score negatively oriented and equal to zero for a perfect deterministic system. Also, if the resolution is greater than the *uncertainty*, the system is considered useless. This measure is equivalent to the resolution part of the CRPS, but does not provide any diagnosis on the reliability of the system.

Since the ensemble verification is statistically performed, we can add confidence interval (for instance by resampling method, bootstrap) on the scores in order to get *objective* comparisons between the assimilation systems.

Remark: the RMSE, which is a deterministic diagnosis, can also be used as a first assessment of the ensemble quality.

Now, for each benchmark, we specify (i) the questions that we want to answer, and (ii) the metrics that we are going to use to answer these questions.

4.1 Small case benchmark

The first benchmark involves a very small size dynamical system and a very idealized assimilation problem (twin experiment, all variables could be observed, no model error). The questions can thus be stated with full mathematical generality, and the metrics can be defined without any kind of approximation. Conversely, no answers on the numerical cost or on the robustness to uncontrolled approximations can be expected at this stage.
Nevertheless, even with this small state vector size of 40, it would be ambitious (but still possible) to evaluate the multivariate probability distribution as a whole. It would take very large ensemble size ($\approx 100 - 1000$) in order to well define the whole probability distribution. Even at this stage, the verification could be limited to the marginal and the N-variate distributions ($N < 40$).

Questions:

1. What is the consistency between the exact prior probability distribution and the one that is simulated or assumed by the assimilation method ?
2. Is the posterior probability distribution statistically consistent with the real error ?
3. To what extent is the uncertainty about the system reduced by the assimilation method ?

Metrics:

1. Compute every score evaluating both reliability and resolution on the prior distribution. Since all the errors are under control, the emphasis should be put on the reliability attribute in order to estimate the impact of the ensembles size on the assimilation stochastic methods.
2. Compute the scores evaluating both reliability and resolution on the posterior distribution.
3. Compute the scores evaluating the resolution/entropy. Compare the gain between the prior and posterior distributions. A better resolution/entropy for the posterior distributions is expected.

4.2 Medium case benchmark

The intermediate benchmark is meant to be a direct transposition of the first benchmark metrics to a mesoscale ocean flow. For this reason, the ocean system is kept as simple as possible (square ocean, simple physics) and the assimilation problem is still an idealized problem (twin experiments, no model error). The additional difficulties come from the much larger size of the system, and from the fact that not all state variables are observed. With this benchmark, the question of the numerical efficiency of the assimilation method starts to become an issue.

Questions:

1. To what extent can the prior probability distribution be described by a moderate size ensemble ? What is the best way to combine the ensemble description with additional assumptions about the prior distribution (like adaptive procedures) ?
2. Are the marginal posterior probability distributions consistent with the real error ? Is there a difference between observed and non-observed variables or as a function of depth ?
3. What is the posterior uncertainty for every single model variable ? How does it change in space and time ?

Metrics:

1. Produce an estimate of exact marginal probability distributions using a very large ensemble, and explore the variations of the scores with respect to this exact distribution as a function of the ensemble size and/or additional assumptions (which are related to the numerical cost). This can be done for univariate marginal distributions and for several bi- or tri-variate marginal distributions to see if the dependence between variables is correctly reproduced as a function of the distance and/or time. If the distributions are close to Gaussian, this can be reduced to exploring the modifications in the ensemble variance and in the linear correlation structure.

2. Compute the scores of every model variable given the corresponding posterior marginal distribution.

3. Compute and compare the scores resolution/entropy corresponding to the marginal prior and posterior probability distributions.

4.3 Large case benchmark

The purpose of the last benchmark is to provide an intercomparison of the assimilation methods using a real-world assimilation problem, which is close to the current MyOcean systems. With respect to the intermediate benchmark, the additional difficulties are: (i) the much larger complexity of the system (larger variety of dynamical processes, about $n = 6 \times 10^7$ state variables), (ii) the use of real-world observations (so that the true state of the system is no longer known), and (iii) the presence of various sources of model errors. For these reasons, the questions must be reformulated and the metrics adapted to provide a similar kind of intercomparison in this more complex situation.

Questions:

1. To what extent is it possible to provide a consistent description of the prior probability distribution ? Is it consistent with the available observations ? What is the best compromise between exploring the probability distribution using a large size ensemble (e.g. to identify non-Gaussian behaviours) and making prior assumptions about the shape of the distribution ?

2. Are the marginal posterior probability distributions consistent with the *independent* observations ?

3. What is the posterior uncertainty ? Is the estimation compatible with the available prior knowledge of the dynamics ?

Metrics:

1. Compute and compare the scores for the marginal (and maybe bivariate) prior distributions, with emphasis on the reliability scores.

2. Compute the scores evaluating both reliability and resolution for the marginal (and bivariate if possible) posterior distributions.

3. Define the list of key diagnostics (variables function) to be evaluated. Estimate and compare the scores resolution/entropy corresponding to the marginal prior and posterior probability distributions for each diagnosis.

Remark :

- a stochastic assimilation system should not only be evaluated on the quality of the posterior distribution it produces but also on the quality of the multi-range ensemble integrations - or *predictions*- it generates. The probabilistic evaluation of ensemble predictions starting from the considered ensemble analysis is then recommanded.

Chapter 5

Conclusion

In WP4, the benchmarks and metrics described in this document will be applied to compare and assess the impact of data assimilation methods of various kinds. This includes: (i) an evaluation of the common approach as defined in WP2, (ii) intercomparison of the existing assimilation tools, (iii) assessment of the stochastic assimilation methods developed in WP3. The small and medium case becnhmarks will be especially used to assess the statistical and numerical effectiveness of the methods and tools gathered in the project. On the other hand, the large case benchmark will be used to evaluate the potential impact of advanced stochastic approaches (including non-Gaussian assumptions) in the real MyOcean systems. The final step of this assessment will include the assimilation of real ocean observations (satellite altimetry, ARGO float data, ocean colour) in the coupled circulation/eocystem model.

Our expectation is that this can provide a reliable information about the current status of stochastic data assimilation methods in the perspective of improving the quality of the MyOcean products. With the metrics that have been defined, our focus is clearly the improvement of the information that is provided by MyOcean users about the expected accuracy of the data that they receive. We indeed believe that associating reliable information about uncertainties (in terms of probability distributon, histograms, error bars,...) is crucial to any practical application of the data, whether it is for technical or political decision making (navigation, fisheries management,...) or for scientific research (model validation, forcing,...).

Appendix A

Scores

A.1 Rank histogram

The univariate rank histogram (RH) checks the reliability of an ensemble system. For each of M realisations of the system, the N ensemble members are ranked in increasing order, thereby defining $N+1$ intervals. If the verification is statistically indistinguishable from the ensemble, it must fall with equal probability in each of those $N+1$ intervals and then shows a flat RH. The shape of the RH caracterizes the lack of reliability of the system. For instance, a strong U-shape with overpopulated outliers shows an underdispersive system. The deviation from the flatness, considering the finitness of M, is measured by the following quantity

$$\delta = \frac{MN}{N+1} \sum_{i=1}^{N} \left(s_i - \frac{M}{N+1} \right)^2 \tag{A.1}$$

where s_i is the observed population of the i-th interval. For a reliable system, δ must be equal to 1.
The program sangoma_ComputeHistogram.F90, available in the SANGOMA assimilation data tools, provides the RH. A subroutine computing the score δ should be added.

Remarks The extension to the multi-variate and the Minimum Spanning Tree RH should be investigated.

A.2 Reduced Centered Random Variable

Like the RH, the reduced centered random variable (RCRV) measures the reliability of an ensemble system.
For each realisation of the system, the following variable is built

$$y = \frac{o - m}{\sigma} \tag{A.2}$$

where m is the ensemble mean, σ the ensemble standart deviation and o the verification value. Note that the verification error σ_o can be simply introduced in y

by considering $\sigma = \sqrt{\sigma_{ens}^2 + \sigma_o^2}$.
The system is reliable, *i.e.* the verification is indistinguisable from the ensemble, if the mean of y is null and its standard deviation is equal to 1. Thus, the reliability is decomposed into (normalized) bias $b = E[y]$ and dispersion $d^2 = E[y^2] - b^2$.

RMSE Considering the notations above, the RMSE can be written as

$$\text{RMSE} = E[(o-m)^2] = E[(\sigma y)^2] \tag{A.3}$$

The RMSE is linked to the RCRV at the first order approximation:

$$\text{RMSE}^2 \approx E[\sigma^2](b^2 + d^2) \tag{A.4}$$

Multi-variate A matricial extension of the RCRV is defined for L variables

$$M = DD^T S^{-1} \tag{A.5}$$

where D is the difference vector between the L ensemble means and associated verifications, and S the covariance matrix of the ensembles. Note that if the L variables are uncorrelated, S is invertible if $N \geq L+1$ (N is the ensemble size). The system is reliable when $E[M] = \mathbf{I}_L$.

Remark: Codes for RCRV and its matricial extension will be provided.

A.3 Brier score and Entropy

Brier score The Brier score is a RMSE in probability space.
For simplifications, we denote p the probability $p(x)$, where x is the state vector. The Brier score is then written

$$\mathcal{B} = E[(p-o)^2] \tag{A.6}$$

where o is the probability associated with the verification value.
By definning $p' = E_p[o]$, the probability of the verification when p is produced, and $p_c = E[o] = E[p']$ the probability associated with the verification data set, the Brier score can be decomposed into reliability and resolution parts:

$$\mathcal{B} = \underbrace{E[(p-p')^2]}_{reliability} + \underbrace{p_c(1-p_c) - E[p'-p_c)^2]}_{resolution} \tag{A.7}$$

The term $p_c(1-p_c)$ is also called *uncertainty* and only depends on the the verification data set. A skill score can be thus defined by

$$\mathcal{B}_s = 1 - \frac{\mathcal{B}}{p_c(1-p_c)} \tag{A.8}$$

The reliability part is null for a perfectly reliable system. The resolution part goes from 0 for a perfect deterministic system to 1 for a useless system, *i.e.* a system providing no more information than the *uncertainty*.

Entropy The Entropy only measures the information content of the system (related to the resolution). Considering the notations above, the entropy is written

$$\mathcal{S} = -E[p' \ln p'] \tag{A.9}$$

This score goes to 0 for a perfect deterministic system to $p_c \ln p_c$ for a useless system, *i.e.* a system providing no more information than the *uncertainty*.

Remark: codes for Brier score and entropy will be provided, at least for the univariate case. The multi-variate extension of these measures should be investigated.

A.4 Continuous Ranked Probability Score

The CRPS is the extension of the Brier score to the Cumulative Distribution Functions (cdf) of a variable x (in the univariate case):

$$\text{CRPS} = E\left[\int_{\mathbb{R}} \left(F_p(\xi) - F_o(\xi)\right)^2 d\xi\right] \tag{A.10}$$

where F_p is the cdf associated with the produced ensembles and F_o the cdf associated with the verification values. The CRPS can be decomposed into the reliability/resolution parts in the same way as the Brier score. But for practical and numerical considerations (see Candille and Talagrand 2005), the decomposition described by H. Hersbach (Hersbach 2000) is chosen. This decomposition is based on the same principle as the rank histogram construction. The reliability part is null for a reliable system and the resolution part goes from 0 for a perfect deterministic system to $\int_{\mathbb{R}} F_c(\xi)\left(1 - F_c(\xi)\right) d\xi$ for a useless system (F_c is the cdf associated with the verification data set).
A routine computing the CRPS and its Hersbach's decomposition will be delivered to all the participants. An extension of the CRPS to the multi-variate case should be investigated.

Appendix B

References

Barnier B., G. Madec, T. Penduff, J.-M. Molines, A.-M. Treguier, J. Le Sommer, A. Beckmann, A. Biastoch, C. Böning, J. Dengg, C. Derval, E. Durand, S. Gulev, E. Remy, C. Talandier, S. Theetten, M. Maltrud, J. McClean, and B. DeCuevas. 2006. Impact of partial steps and momentum advection schemes in a global ocean circulation model at eddy permitting resolution. *Ocean Dynamics*, **56**, pp 543–567.

Béal D., P. Brasseur, J.-M. Brankart, Y. Ourmières, and J. Verron. 2010. Characterization of mixing errors in a coupled physical biogeochemical model of the North Atlantic: implications for nonlinear estimation using Gaussian anamorphosis. *Ocean Science*, **6**, pp 247–262.

Candille G., C. Côté, P. L. Houtekamer, and G. Pellerin. 2007. Verification of an ensemble prediction system against observations. *Mon. Wea. Rev.*, **135**, pp 2688–2699.

Candille G. and O. Talagrand. 2005. Evaluation of probabilistic prediction systems for a scalar variable. *Quart. J. Roy. Meteor. Soc.*, **131**, pp 2131–2150.

Cosme E., J.-M. Brankart, J. Verron, P. Brasseur, and M. Krysta. 2010. Implementation of a reduced rank square-root smoother for high resolution ocean data assimilation. *Ocean Modeling*, **33**, pp 87–100.

Doron M., P. Brasseur, and J.-M. Brankart. 2011. Stochastic estimation of biogeochemical parameters of a 3D ocean coupled physical-biogeochemical model: twin experiments. *J. Mar. Sys.*, **87**, pp 194–207.

Gneiting T., L.I. Stanberry, E.P. Grimit, L. Held, and N.A. Johnson. 2008. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. Tech. Report **537**, *Dpt Statistics*, Univ. Washington, pp 1–26.

Gombos D., J.A. Hansen, J. Du, and J. McQueen. 2007. Theory and applications of the minimum spanning tree rank histogram. *Mon. Wea. Rev.*, **135**, pp 1490–1505.

Hersbach H. 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weatther and Forecasting*, **15**, pp 559–570.

Lévy M., M. Gavart, L. Mémery, G. Caniaux, and A. Paci. 2005. A four-dimensional mesoscale map of the spring bloom in the northeast Atlantic (POMME experiment): Results of a prognostic model. *J. Geophys. Res.*, **110**. pp 1–23.

Murphy A. H. 1973. A new vector partition of the probability score. *J. Appl. Meteor.*, *12*, pp 595–600.

Nakano S., G. Ueno, and T. Higuchi. 2007. Merging particle filter for sequential data assimilation. *Nonlin. Processes Geophys.*, **14**, pp 395–408.

Ourmières Y., P. Brasseur, M. Lévy, J.-M. Brankart, and J. Verron. 2009. On the key role of nutrient data to constrain a coupled physicalâbiogeochemical assimilative model of the North Atlantic Ocean. *J. Mar. Syst*, **75**, pp 100–115.

Sakov P., D.S. Oliver, and L. Bertino. 2012. An iterative EnKF for strongly nonlinear systems. *Mon. Wea. Rev.*, **140**, pp 1988–2004.

Toth Z., O. Talagrand, G. Candille, and Y. Zhu. 2003. 'Probability and ensemble forecasts' in Forecast Verification: A Practitionerâs Guide in Atmospheric Science, Jolliffe I., Stephenson D.B. (eds), Wiley: UK. pp 137–163.

Van Leeuwen P. J. 2010. Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Quart. J. Roy. Meteor. Soc.*, **136**, pp 1991–1999.

Appendix C

Distribution of the model configurations

C.1 Small case benchmark

The small case benchmark is based on the portable Lorenz-96 model with 40 variables. The model is available:

- in Fortran, in the PDAF software (http://pdaf.awi.de/),
- in Java, in the openDA software (http://www.openda.org), or
- in Matlab, in the EnKF Matlab code (http://enkf.nersc.no/Code/EnKF-Matlab/).

C.2 Medium case benchmark

The medium case benchmark is based on the an idealized square ocean configuration of NEMO. This configuration can be installed and used as follows:

1. Download the NEMO model
   - Register to the NEMO website http://www.nemo-ocean.eu/: get login and password. After login, you can access to the full NEMO documentation: reference manuals, user's guides, description of the reference configurations,...
   - Download NEMO (as also explained in the NEMO Quick Start Guide):
     svn co http://forge.ipsl.jussieu.fr/nemo/svn/tags/nemo_v3_3/NEMOGCM
     This creates a directory structure starting from NEMOGCM, with the source code, the reference NEMO configurations and associated tools.
2. Prepare the SQB configuration
   - Go to the directory with the NEMO reference configurations:
     cd NEMOGCM/CONFIG

- See the list of available compilers and platforms:
  ./makenemo -m help
  Define the compiler and platform that you want to use (with option '-j0' to avoid compiling anything at this stage), for instance:
  ./makenemo -j0 -m ifort_linux
- Copy the GYRE reference configuration to create a new SQB configuration (without compiling, with option '-j0'):
  ./makenemo -j0 -r GYRE -n SQB
- Go to the directory with the newly created SQB configuration:
  cd NEMOGCM/CONFIG/SQB
- Modify the precompilation options file "cpp_SQB.fcm", using the following CPP options: cpp_SQB.fcm (appropriate for the SQB configuration). Please refer to the NEMO CPP options guide (this page requires to be logged in) for the meaning of the CPP options. Additional keys are needed to parallelize the computation using domain decomposition (key_mpp_mpi, key_nproci>1 and key_nprocj>1, for instance key_nproci=4 and key_nprocj=4 to use 16 processors).
- Copy the following source directory (specific to SQB): SQB/MY_SRC in the directory:
  NEMOGCM/CONFIG/SQB/MY_SRC
- Edit the file "par_SQB.h90" to change the model resolution: set jp_cfg=4 for the 1/4° resolution SQB configuration.

3. Compile the SQB configuration

- Go to the directory with the compiling options:
  cd NEMOGCM/ARCH
- Edit the file corresponding to your compiler and platform, for instance:
  vi arch-ifort_linux.fcm
  to change the directory with the NetCDF library.
  IMPORTANT: The same Fortran compiler must have been used to produce the NETCDF library (so that the format of the 'netcdf.mod' file is appropriate).
- Go to the directory with the NEMO configurations:
  cd NEMOGCM/CONFIG
- Compile NEMO (current configuration and compiler have already been set above):
  ./makenemo

4. Run the SQB configuration

- Go to the directory where to run your first test experiment:
  cd NEMOGCM/CONFIG/SQB/EXP00

- Replace the namelist file "namelist" (copied above from the GYRE configuration), by the SQB namelist: namelist_04_biharm for the 1/4Â° resolution SQB configuration, with biharmonic horizontal mixing. Please refer to the NEMO manual for the meaning of the model parameters.
- Run your first one year simulation (opa is the name of the executable):
  ./opa
  The model output includes:
  - ocean.output: model control print (text file);
  - time.step: current timestep (text file);
  - SQB*restart.nc: restart files (NetCDF files);
  - SQB*grid*.nc: output files (NetCDF files).
- To continue the model simulation from the final restart file, the namelist must be edited to modify the initial and finl timesteps. Here is an example ksh script iterating several years: run.ksh, together with the skeleton namelist edited by the script: namelist_04_harm.skel for the 1/4° resolution SQB configuration, with harmonic horizontal mixing; namelist_04_biharm.skel for the 1/4° resolution SQB configuration, with biharmonic horizontal mixing.
- This is just a starting point. Please set up your own directory structure to organize your assimilation experiments. The only two things that you need to run the SQB configuration are the executable and the namelist with the parameters.

C.3 Large case benchmark

The large case benchmark is based on a realistic North Atlantic configuration of NEMO (at a $1/4°$ resolution). This configuration can be installed and used as follows:

1. Download the NEMO model

   - Register to the NEMO website http://www.nemo-ocean.eu/: get login and password. After login, you can access to the full NEMO documentation: reference manuals, user's guides, description of the reference configurations,...
   - However, NATL025 is a NEMO configuration developed by the DRAKKAR project. It is thus easier to donwload the NEMO code with all required specific developement together with the DRAKKAR configuration manager (DCM). More information about DCM on the webpage https://servforge.legi.grenoble-inp.fr/projects/DCM (login and password required).
   - Download NEMO and associated tools from the DCM distribution (with the same login and password):
     svn - -username *your_name* co https://servforge.legi.grenoble-inp.fr/svn/DCM/DCM/trunk NEMODRAK_3.4

svn - -username *your_name* co https://servforge.legi.grenoble-inp.fr/svn/DCM/BUILDNC/trunk/TIME TIME

svn - -username *your_name* co https://servforge.legi.grenoble-inp.fr/svn/DCM/BUILDNC/trunk/DIMGPROC DIMGPROC

svn - -username *your_name* co https://servforge.legi.grenoble-inp.fr/svn/DCM/BUILDNC/trunk/Macrolib Macrolib

- Please note that DCM requires to set up several environment variables (check the documentation on the webpage https://servforge.legi.grenoble-inp.fr/projects/DCM/wiki/DcmInstall).

2. Prepare the NATL025 configuration

   - Create the directory structure for the new configuration/case:

     mkconfdir NATL025 "your_case_name"

     A new case should be created each time you want to perform a new simulation with different settings (modified code, parameters, input data). A standard naming convention for the case is proposed in the DCM documentation.

   - Edit the configuration "makefile":

     $UDIR/CONFIG_NATL025/NATL025-"your_case_name"/makefile

     Modify especially the following variables:

     CASEREF = 'none' PREV_CONFIG = '$HOMEDCM/CONFIGS/NATL025-GREF3.4' MACHINE = PW6_VARGAS NCOMPIL_PROC = 64

   - Copy all relevant files from the reference configuration:

     make copyconfig

   - Mofify the precompilation options separately (in the file CPP.keys), for instance:

     P_P = key_natl025 key_dynspg_flt key_zdftke key_traldf_c2d key_dynldf_c2d key_ldfslp key_dimgout key_lim2 key_lim2_vp key_mpp_mpi

3. Compile the NATL025 configuration

   - Prepare the code before compilation:

     make install

   - Compile NEMO (current configuration and case):

     make

   - Compile the peripheric tools in the directories TIME and DIMGPROC (the second one is to combine the binary output files produced by each process into a single NetCDF file), e.g.:

     cd DIMGPROC ln -s ../Macrolib/macro.vargas make.macro make ; make install

4. Run the NATL025 configuration

- Download the DRAKKAR scripts to run the model:
  svn co https://servforge.legi.grenoble-inp.fr/svn/DCM/RUNTOOLS/trunk RUN_TOOLS
  These must be considered as example scripts, which need to be adjusted to your computer.
- Prepare the control script for the current configuration/case
  cd RUN_TOOLS/TEMPLATE mkctl NATL025-"your_case_name"
  The following assumes that you install the scripts for the machine "vargas" (IBM SP Power6).
- Go to the directory with the control script for the current configuration/case
  cd $HOME/RUN_NATL025/NATL025-"your_case_name"/CTL
- Prepare the model parameter files:
  - namelist: template in NEMODRAK_3.4/CONFIGS/NATL025-GREF3.4
  - namelist_ice: template in NEMODRAK_3.4/CONFIGS/NATL025-GREF3.4
  - namelistio: template in DIMGPROC
- Edit the file "includefile.ksh", and adjust the parameters and directories. In particular, set the number of jobs to be resubmitted (MAXSUB).
- Check class limits in the script:
  NATL025-"your_case_name"_vargas.ksh
- Set the first and last time step of the first job in:
  NATL025-"your_case_name".db
  All subsequent jobs are automatically resubmitted with the same number of timesteps.
- Prepare the input data files in the data directory ($SDIR/NATL025/NATL025-I).
- Run the model:
  ./run_nemo_vargas.ksh
- The results of the simulations are produced in the directory:
  $SDIR/NATL025/NATL025-"your_case_name"-S