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

Deliverable 4.5: Metrics obtained with the large case benchmark

Due date: 31/10/2015
Delivery date: 31/10/2015
Delivery type: Report , public

Jean-Marie Beckers Alexander Barth
Yajing Yan
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 purpose of this deliverable is to present the results that have been obtained with the large case SANGOMA benchmark, focusing on the probabilistic metrics that have been used to evaluate the results (described in deliverable 4.3) . The large case benchmark is based on a realistic North Atlantic model at $1/4^\circ$ resolution, assimilating real world observation (described in deliverable 4.1); it has been designed to be complient with MyOcean systems (as close as possible to a realistic system).

As planned in the SANGOMA proposal, assimilation experiments with the large case benchmark have been carried out by GHER and CNRS/LGGE. Each of these works has already been published in scientific journal: in one paper by Candille et al. (2015) in Ocean Science, and in one paper by Yan et al. (2015) in the Journal of Geophysical Research. This deliverable is based on the work described in these two papers, but presents the two assimilation experiments in the same framework.

As explained below, the two assimilation experiments use the same model configuration, the same assimilation method (ensemble Kalman filter, with square root analysis scheme), and applies the same probabilistic metrics to evaluate the results. The main difference is in the way modelling errors are taken into account in the assimilation system. In the first experiment (by GHER), modelling error is simulated by introducing stochastic air-sea forcing perturbations; and in the second experiment (by CNRS/LGGE), modelling error is simulated by introducing stochastic perturbations in the equation of state (see chapter 2 below). An appropriate simulation of modelling error is indeed a key element in ensemble assimilation systems, which it is often difficult to do correctly.

The plan of the deliverable is as follows: chapter 2 describes the ensemble model system (without assimilation), and a first comparison to observations (using rank histograms); chapter 3 describes the assimilation experiments; chapter 4 describes the probabilistic metrics that have been obtained; and chapter 5 summarizes the conclusions that have been obtained from this work.

Chapter 2

Ensemble model system

2.1 Model configuration

The definition of the large case SANGOMA benchmark has been provided in deliverable 4.1. Here is a short summary of the model configuration.

The large case benchmark is based on a realistic configuration of the NEMO ocean model, for the North Atlantic Ocean, at a $1/4°$ resolution (Barnier et al, 2006), 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) and because it is based on the NEMO ocean model which is used by most MyOcean Monitoring and Forecasting Centres.

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.

Numerics. The primitive equations are discretized on an Arakawa C grid, with a horizontal resolution of $1/4° \times 1/4° \cos(\phi)$. 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. Time stepping is performed with a leap frog scheme, with a time step $\Delta t = 2400$ s

2.2 Ensemble simulations

Two different methods are used to produce the ensemble simulations. In the first one (GHER), perturbations are introduced in the model using a stochastic air-sea forcing function; and in the second one (CNRS/LGGE), perturbations are introduced in the equation of state.

Stochastic perturbation of the forcing (GHER) Uncertainties in the system occur for many different reasons: model dynamics, parameters, forcing, initial and boundary conditions. Model predictions incorporate error even when forcing data are perfect due to the choices or limitations in regard to the model physics and parameters. This includes error associated with assumptions, theory and/or conceptualisations within the underlying equations, errors due to the computational grid and its discretisation, numerical errors associated with the time step or numerical methods used to solve the mathematical equations, and uncertainty associated with any model parameters adopted. Errors in forcing data are associated with the measurement (or prediction) of forcing data and its spatial representation.

It is an important task of the assimilation system to make correct assumptions about the uncertainties. In these experiments, the ensemble is generated by adding realistic noise in the forcing parameters. For this, the air temperature at 2 m (t2), wind velocities at 10 m (u10, v10), the long wave radiation (radlw) and short wave radiation (radsw) are considered. The perturbation is obtained by using the Fourier decomposition of the forcing variable vectors. The details of the principle is explained in Barth et al. (2011). For a realistic ocean circulation model, the ensemble should be representative of the impact of forcing errors on monthly time scale ocean dynamics. The monthly variability is thus taken into account during the perturbation computation.

The principle is as follows. Let $\mathbf{p}$ be the vector of forcing variables of the year 2005:

$$\mathbf{p}(x, y, t) = \sum_k a_k(x, y) exp(i\omega_k t) \tag{2.1}$$

where ω_k is the kth angular frequency and $a_k(x, y)$ are complex fields corresponding to the Fourier coefficients of the angular frequency ω_k:

$$\omega_k = \frac{2\pi k}{\Delta t} k = -\frac{k_{max}}{2}, \cdots, \frac{k_{max}}{2} - 1 \tag{2.2}$$

Δt is the time interval of the forcing variables, for example, every 3 hours, and k_{max} is the total number of 3-hourly field during the time span of the forcing variables, here one year.

The perturbation vector $\mathbf{p}_p(x, y, t)$ is obtained as follows:

$$\mathbf{p}_p = \alpha Re(\sum_k a_k(x, y) z_k(t)) \tag{2.3}$$

where z_k is a complex random time series with a temporal correlation scale of $T_k = 2\pi/|\omega_k|$, zero mean and unit variance. The factor α takes into account that the expected perturbation is in general smaller than the temporal variability.

The construction of the random time series is explained in Evensen (1994). Usually, we do not take into account all the frequencies, since the variability can be dominated by only some frequencies or we take interest in only short time variation, therefore, we combine only the frequencies that we are interested in. For a realistic ocean circulation model, the ensemble should be representative of the impact of forcing errors on monthly time scale ocean dynamics. The variability of one month is thus taken into account for the air temperature and wind velocities.

The variability of three months is taken into account for the long wave and short wave radiation.

Moreover, since the air temperature, the wind velocities are correlated. The same time series are used for these three variables. Regarding the choice of the factor α, it is determined in order that the perturbation is realistic. In order to validate the ensemble, the ensemble spread at the end of the ensemble spin up, is compared to the difference between the model prediction without perturbation and the observations. Here, $\alpha = 1.5$ is chosen for the air temperature and the wind velocities. For the long wave and short wave radiation, 3.0 is taken as α value.

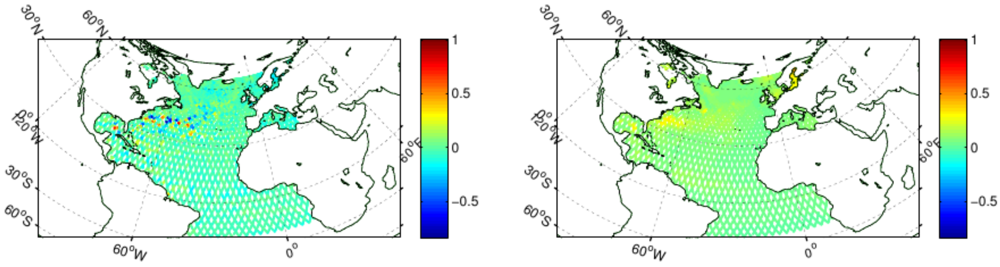


Figure 2.1: Model/observation difference (left panel) and ensemble spread (right panel) for SSH at the end of the ensemble spin up.

The ensemble spin up time is 180 days to ensure dynamic stability and to obtain correct multivariate correlations before starting assimilation. At the end of the ensemble spin up time, the ensemble is diagnosed and validated by comparison between the ensemble spread and the difference between the model prediction without perturbation and the observations. for instance, the SSH model/observation difference (model - observation) at the end of the spinup, as well as the ensemble spread are shown in Figure 2.1. There is a good spatial consistency between the model/observation difference and the ensemble spread is observed: large errors are located in the Gulf Stream region. The RMS error of the model compared to the obser- vations is 0.097 m, while the ensemble spread is 0.086 m.

Stochastic perturbation of the equation of state (CNRS/LGGE) Before building a NATL025-based ensemble required for data assimilation, we have to wonder how to represent the uncertainties in the model, and how to simulate the impacts of the unresolved small scales on the larger scales circulation. Brankart (2013) has shown that the unresolved scales in the nonlinear seawater equation of state represent a major source of uncertainties in the computation of the large scale horizontal density gradient (from T/S large scale fields), and the impacts of these uncertainties can be simulated by random processes representing unresolved T/S fluctuations.

Following these conclusions, the NATL025-based ensemble is built by introducing stochastic perturbations in the equation of state. In practice, a single non-perturbed integration is first performed from Levitus (1989) to January 1st 2005 to reach the regime state of the model. Then a 96-member ensemble of perturbed simulations is run for 6 months with the following stochastic formulation of the equation of state:

$$\rho = \frac{1}{2}\{\rho[T + \Delta T, S + \Delta S, p_o(z)] + \rho[T - \Delta T, S - \Delta S, p_o(z)]\} \tag{2.4}$$

where $p_o(z)$ is the reference pressure depending on the depth and ΔT and ΔS are a set of T/S perturbations defined as the scalar product of the respective local T/S gradients with random walks $\boldsymbol{\xi}$:

$$\Delta T = \boldsymbol{\xi} \cdot \nabla T \qquad \text{and} \qquad \Delta S = \boldsymbol{\xi} \cdot \nabla S \tag{2.5}$$

$\boldsymbol{\xi}$ are produced by a first-order autoregressive process (AR-1) with a 10-day decorrelation time scale, and horizontal and vertical standard deviations σ_s equal to 1.4 and 0.7 grid points respectively. $\boldsymbol{\xi}$ are uncorrelated over the horizontal and fully correlated along the vertical.

These stochastic parameters are chosen to produce an ensemble spread that is large enough for our purpose while keeping the model numerically stable. Nevertheless, in order to avoid numerical instabilities, limiting factors are introduced (1.5 σ_s) on the perturbations and the time step of the stochastic model is divided by 4 compared to the time step of the classical model (600s instead of 2400s). Such an ensemble is thus built to spread mostly over areas with strong gradients and where the equation of state is strongly nonlinear, for instance in the Gulf Stream area.

The size of the ensemble (96 members) is chosen in order to satisfy several factors. Without considering the numerical cost, the larger the size of the ensemble, the more accurate the descriptions of the probability density functions (pdfs) and the covariance matrices. The ensemble size is then a compromise between the numerical constraints and the accuracy of the pdfs and covariances matrices associated with the ensemble. Moreover, we also have to take into account the saturation —depending on the ensemble size— of the probabilistic measures with which the ensemble is evaluated (Candille et al 2014).

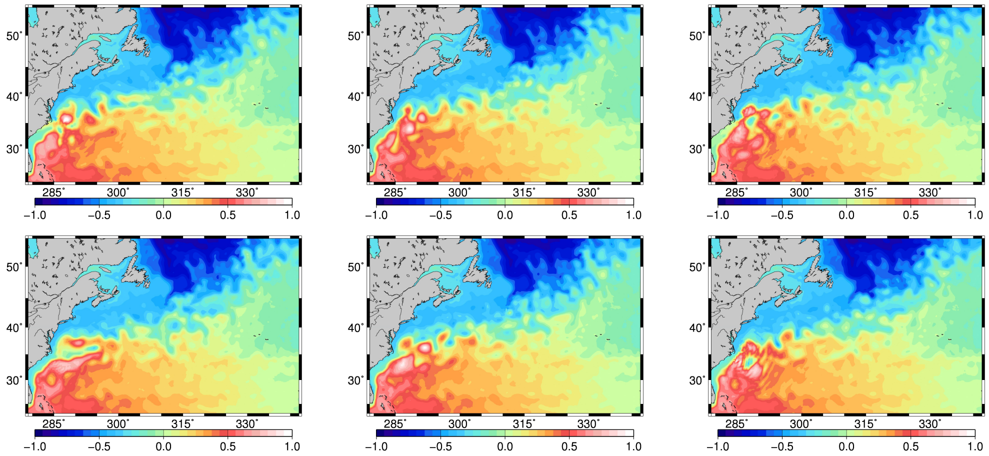


Figure 2.2: 6 members from the free run 96-member ensemble SSH; snapshots for July 9th 2005.

These perturbations are designed to represent the major part of the uncertainties on the large scale horizontal density gradient. We thus expect an impact on the mesoscale circulation that is observable by altimetric data. But we cannot reasonably expect that these perturbations can simulate all kinds of uncertainties that significantly influence the thermohaline circulation of the North Atlantic. This ensemble is designed to allow an effective control of the mesoscale circulation by altimetric data, not to compensate for any model deficiencies in the description of the thermohaline strcuture of the ocean.

We now present a qualitative description of the 96-member ensemble running for 6 months (without any assimilation process). Figure 2.2 shows the variability of the dynamical fields (SSH) depending on the stochastic perturbations (only 6 members are shown). We can notice that the large scale patterns are similar in all members, but every member presents different eddy-pattern: as expected, the largest variability of the ensemble is mostly located around the Gulf Stream front.

2.3 Ensemble validation

We check the reliability of the ensemble simulations using rank histograms (Anderson, 1996). Let us consider the pdf is described by an ensemble of size N. For each realisation of the system, the N ensemble members are ranked in increasing order, thereby defining $N+1$ intervals (or bins). Then we check the rank of the verification within these bins: the rank histogram is built by accumulation over all realisations —assumed independent— of the ranks. The ensemble system is reliable, *i.e.* the verification is statistically indistinguishable from the ensemble when it falls with equal probability in each of the $N+1$ intervals and then shows a flat rank histogram. A shape different from the flatness of the rank histogram characterizes a lack of reliability of the system. For instance, a system with many outliers, meaning the verification values fall outside the ensembles, presents a U-shape and is called underdispersive system.

Stochastic perturbation of the forcing (GHER) The rank histogram of SSH over the Gulf Stream region (Fig. 2.3) confirms the good representativity of the uncertainty by the ensemble spread. Here, the rank histogram is only computed in the Gulf Stream region, because the variability among ensemble members is mainly located in this area. There, the model error represented by the ensemble spread is considered to be realistic.

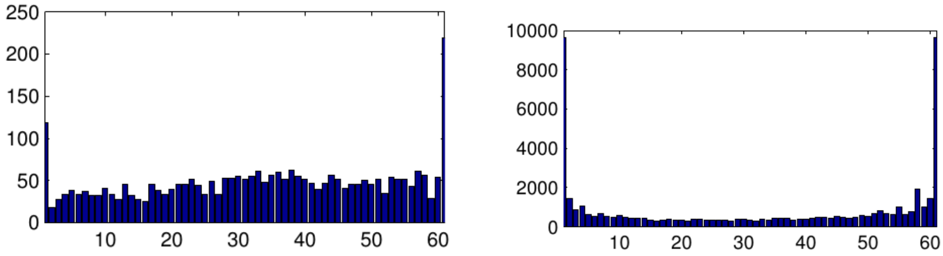


Figure 2.3: Rank histogram for SSH in the Gulf Stream region (left panel) and temperature over the whole basin (right panel) at the end of the ensemble spin-up.

Stochastic perturbation of the equation of state (CNRS/LGGE) In Figure 2.4, the reliability verification of the SSH is performed against the along-track altimetric observations from JASON during the 10-day cycle from June 29th to July 9th 2005. The left panel shows the observations rank. Many outliers are observed, especially in the subtropical gyre. This is mainly due to the very small spread of the ensemble in this area combined to the model error (ensemble mean minus observation). Actually, the ensemble spread is so small in this area that the outliers only reflects the local model error. For the rank histogram construction (right panel), the statistics are only accumulated over the Gulf Stream verification area to avoid aggregating heterogeneous data from the frontal region and the gyre. Graphically, we observe a weak positive bias (asymmetric rank histogram to the left) and a slight underdispersion. Note that observational error is taken into account in the rank histogram construction by adding it to the ensemble members through a white noise (with standard deviation consistent with the observation error used in the assimilation process, $\sigma_o = 10$cm, see next chapter).

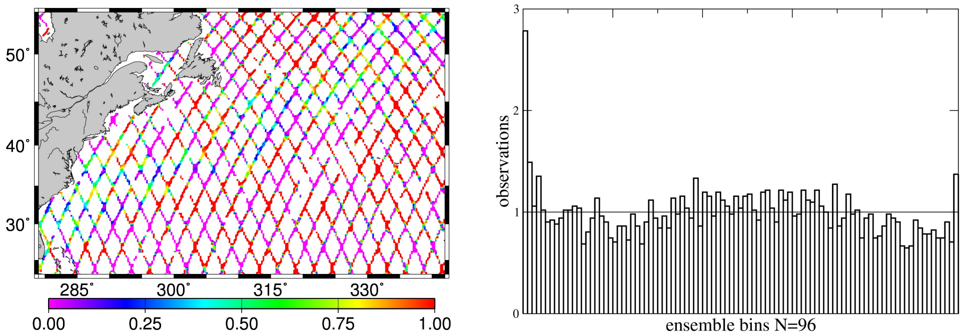


Figure 2.4: SSH JASON tracks, local rank of the observations over 10-day JASON-cycle from June 29th to July 9th 2005 (left panel) and rank histogram over the Gulf Stream verification area (right panel).

Chapter 3

Assimilation experiments

3.1 Setup of the experiments

The assimilation is performed in 2005, during one year. The first 180 days are ensemble spin up period (see previous chapter). It is indeed important to integrate the ensemble over a time interval covering a few characteristic time scales of the dynamical system to ensure dynamic stability and to trigger multivariate correlations before starting the assimilation of observations.

Assimilated observations are not the same in the GHER experiment and in the CNRS/LGGE experiment.

- On the one hand, in the GHER experiment, 3 types of observations are assimilated alltogether: JASON1 SSH data, AVHRR SST data and ARGO temperature profiles.

  The assimilation is performed with an assimilation window of 10 days. The assimilation cycle must be long enough to accumulate sufficient amount of observations to correct the model state accordingly. The 10-day interval corresponds to the characteristic time scale of the ARGO data collection.

  The IAU scheme is applied instead of the intermittent assimilation in order to reduce the spurious oscillations produced in the intermittent assimilation by keeping the mass and momentum fields in balance. More precisely, the IAU scheme is used because of its advantages in reducing the high frequency analysis-induced oscillations but not increasing the computation time compared to other IAU assimilation schemes (Yan et al, 2014). In this scheme, at the end of each assimilation window, the analysis is done using observations around the analysis time. An increment is calculated from the difference between the analysed and the forecasted model states. This increment is then added to the model integration for the subsequent assimilation window. Therefore, the model integration is always forward, there is no model integration repeat for each assimilation window. Moreover, a time scale in accordance with the observation decorrelation is applied to the weighting function of the increment update. A linearly decreasing function is applied, because the increment is more correlated with the observations near the current analysis step. The advantage of this linearly decreasing weighting function compared to the constant weighting function is shown in Yan et al., 2014.

In addition, the observations are localised in the analysis to avoid the effect of spurious long range correlations. The localisation length scale is determined according to the auto-correlation length of SST and SSH, here 300 km. Then the observations are weighted depending on their distance from the water column being analysed. The observation weight function is the Gaussian function with an e-folding scale of the localisation length.

- On the other hand, the CNRS/LGGE experiment focuses on the assimilation of altimetric observations, with the aim of controlling the eddy dynamics in the Gulf Stream region. The observations are along-track data coming from two different satellites: 10-day cycle JASON (long inter-tracks) and 35-day cycle ENVISAT (short inter-tracks).

  The assimilation cycles are defined to fit with the 10-day cycle of JASON and are then performed within 10-day assimilation windows $[t_k, t_{k+10}]$. The update (see details below) is performed at the middle of the assimilation window t_{k+5} with all observations and model equivalents —model outputs projected to the exact observations times and locations— contained in the 10-day assimilation window. Increments are then computed for each ensemble member as explained below (Ourmières et al 2006).

  The increment is then introduced into the model by the Incremental Analysis Update (IAU) algorithm: a 10-day integration runs from t_k by injecting fractions of the increment step by step all along the assimilation window. The full increment is thus introduced in the system at the final time t_{k+10} of the assimilation cycle. The IAU ensemble at the final time of the cycle provides the initial conditions for the forecast ensemble of the next cycle. The forecast ensemble trajectories are thus discontinuous from a cycle to another, while the IAU ensemble trajectories remains continuous (and avoid possible numerical shocks which would occur if the increments were fully injected at one single time).

  In addition, to avoid the spurious effect of inaccurate long-range correlations the update is also performed with a localisation process: the local assimilation areas are limited by a radius of 4.5° (≈ 450km at 30N) and the observation influence is defined by Gaussian functions with standard deviation of 1.5° (≈ 150km at 30N).

3.2 Assimilation method

The same assimilation method is used in the two assimilation experiments (by GHER and CNRS/LGGE), but the implementation is different: GHER uses OAK (Ocean Assimilation Kit), and CNRS/LGGE uses SESAM (System of Sequential Assimilation Modules).

The assimilation scheme is a ensemble-based Kalman filter method, using a square root algortihm to update the ensemble (like in LETKF, Bishop et al 2001).

Each member i of the forecast ensemble of vector state $\mathbf{x}$ is written

$$\mathbf{x}_i^f = \overline{\mathbf{x}^f} + \delta \mathbf{x}_i^f \tag{3.1}$$

where $\overline{\mathbf{x}^f}$ is the ensemble mean and $\delta \mathbf{x}_i^f$ the associated anomalies which define the columns (with factor $1/\sqrt{N-1}$) of the forecast square root covariance matrix $\mathbf{S}^f$ (covariance matrix $\mathbf{P}^f = \mathbf{S}^f \mathbf{S}^{fT}$). Each ensemble member is projected into the observation space by $\mathcal{H}$ so that $\overline{\mathcal{H}\mathbf{x}^f}$ is the forecast ensemble mean in observation space and $\delta\left(\mathcal{H}\mathbf{x}_i^f\right)$ are the associated anomalies defining the columns (with factor $1/\sqrt{N-1}$) of the forecast square root covariance matrix in the observation space $\mathcal{H}\mathbf{S}^f$.

The ensemble mean is then updated with a square root algorithm, meaning that no observation perturbation is needed but requiring to work in the eigenspace of

$$\mathbf{\Gamma} = (\mathcal{H}\mathbf{S}^f)^T \mathbf{R}^{-1} (\mathcal{H}\mathbf{S}^f) = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1} \tag{3.2}$$

where $\mathbf{U}$ and $\mathbf{\Lambda}$ are respectively the unitary matrix and the diagonal matrix of the singular value decomposition of $\mathbf{\Gamma}$. Note that the observation error covariance matrix $\mathbf{R}$ is diagonal: $\mathbf{R} = \sigma_o^2 \mathbf{I}$. The analysis ensemble mean $\overline{\mathbf{x}^a}$ is then updated as follows

$$\overline{\mathbf{x}^a} = \overline{\mathbf{x}^f} + \mathbf{S}^f \mathbf{U} \left(\mathbf{I} + \mathbf{\Lambda}\right)^{-1} (\mathcal{H}\mathbf{S}^f)^T \mathbf{R}^{-1} \left(\mathbf{y}_o - \overline{\mathcal{H}\mathbf{x}^f}\right) \tag{3.3}$$

$\overline{\mathbf{x}^a}$ thus depends on the innovation $\mathbf{y}_o - \overline{\mathcal{H}\mathbf{x}^f}$, the observation error covariance $\mathbf{R}$, and the anomalies expressed both in model and observation spaces: $\delta \mathbf{x}_i^f$ and $\delta\left(\mathcal{H}\mathbf{x}_i^f\right)$.

After that, each ensemble anomaly i can be updated as follows

$$\delta \mathbf{x}_i^a = \sqrt{N-1} \left(\mathbf{S}^f \mathbf{U} \left(\mathbf{I} + \mathbf{\Lambda}\right)^{-\frac{1}{2}} \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{U}^T\right)_i \tag{3.4}$$

And then, each updated ensemble member $\mathbf{x}_i^a$ is rebuilt from the updated ensemble mean and its associated updated ensemble anomaly:

$$\mathbf{x}_i^a = \overline{\mathbf{x}^a} + \delta \mathbf{x}_i^a \tag{3.5}$$

The increments $\delta \mathbf{x}_i = \mathbf{x}_i^a - \mathbf{x}_i^f$ are then computed and are introduced into the model by an IAU method in order to continuously update the ensemble and the covariance matrix. Note that the stochastic version of the model is used for these assimilation cycles (same as for the free run integration).

Chapter 4

Validation metrics

4.1 Definition of the probabilistic metrics

The definition of the probabilistic metrics used in the SANGOMA benchmarks has been provided in deliverable 4.3. Here is a short summary of the metrics that are used in this chapter.

In probabilistic validation, the performance of the ensemble system is diagnosed according to *reliability* and *resolution*.

To numerically assess the lack of reliability graphically observed on the rank histogram, we use the Reduced Centered Random Variable (RCRV) as follows. For each realisation of the system, the RCRV is defined as

$$y = \frac{o - m}{\sigma} \tag{4.1}$$

where m and σ are the mean and the standard deviation of the produced pdf respectively and o the observed value. Note that the observation error σ_o can be simply introduced in y by considering $\sigma = \sqrt{\sigma_{ens}^2 + \sigma_o^2}$. The system is reliable, if the mean of y over all the realisation of the probabilistic system is null and its standard deviation is equal to 1. Thus, the reliability is decomposed into (normalized) bias $b = E[y]$ and dispersion $d^2 = E[y^2] - b^2$. For the example in Fig. 2.4, the normalized bias and the dispersion from the RCRV are equal to $b = -0.1$ and $d = 1.15$ respectively, *i.e.* the system has a weak positive bias (10%) and a slight underdispersion (15%). These two diagnoses —the RCRV and the rank histogram— only measure the reliability of the system. We need to use other scores evaluating the resolution to get a full probabilistic skill assessment of the ensemble system.

The Continuous Rank Probability Score (CRPS) measures the global skill of a probabilistic system by evaluating both reliability and resolution. It is based on the square difference between the produced Cumulative Distribution Functions (cdf) of a univariate variable x and the corresponding cdf of the observation:

$$\text{CRPS} = E\left[\int_{\mathbb{R}} (F_p(x) - F_o(x))^2 \, dx\right] \tag{4.2}$$

where F_p is the cdf associated with the produced pdf, F_o the cdf associated with the observation (a simple Heaviside distribution when no observation error is considered), and $E\left[\cdot\right]$ is the average of the integrals over all the verification dataset.

Contrary to the reliability scores presented above, the CRPS has the dimension of the evaluated variables (for instance expressed in meters for the SSH, in Kelvin for the temperature, etc...). The CRPS can be decomposed into the reliability/resolution parts in many different ways. But for practical and numerical considerations (see Candille and Talagrand 2005), the decomposition described by H. Hersbach (see details in Hersbach 2000) is chosen:

$$CRPS = Reli + Resol \tag{4.3}$$

This decomposition is based on the same principle as the rank histogram construction. These scores are negatively oriented, *i.e.* the reliability part ($Reli$) is null for a reliable system and the resolution part ($Resol$) s equal to 0 for a perfect deterministic system.

The goal of the assimilation is to improve the information contained in the system, *i.e.* reduction of the probabilistic resolution, by keeping the system as reliable as possible, *i.e.* without deteriorating the reliability.

4.2 Results of the experiments

As a first step, we look at the experiments using the classic metrics used in operational oceanography systems. Probabilistic metrics will be described afterwards in section 4.3.

Stochastic perturbation of the forcing (GHER) The time evolution of the spatially averaged RMS errors of the forecast/analyses and of the free run with respect to observations of SSH, SST and temperature profiles are shown in Figure 4.1. For the temperature profiles, the RMS values are calculated by taking into account the volume represented by the model grid points, since the vertical grid is not regular.

For SSH, the RMS errors of the analyses are about 4 cm smaller than those of the forecast at the beginning of the experiment. Towards the end of the experiment, the RMS difference between the analyses and forecasts reach 7 cm. Note that the RMS errors of the forecasts are very close to those of the free run, this can be explained by the fact that no SSH increment is updated in the assimilation experiment, the correction of the SSH in the model state depends on the interactions between the temperature, the salinity and the SSH during the model integration. Detailed inspection shows that punctual large residual (0.6 m) is mainly located in the Gulf Stream region. Off the African coast, residual on the order of 0.2 m is observed.

For SST, the RMS error of the model prediction is more than 2°C at the beginning of the experiment. In the free run, the RMS errors increase slightly, then decrease to 1°C from the 5th step until the 10th step. Then the RMS errors remain stable at about 1°C until the end of the year. The RMS errors of the forecast are reduced quickly at the first two analysis steps, then stay stable until the 6th step. Between the 6th and 10th steps, they decrease to 0.5°C. After that, the RMS errors remain stable at about 0.5°C.

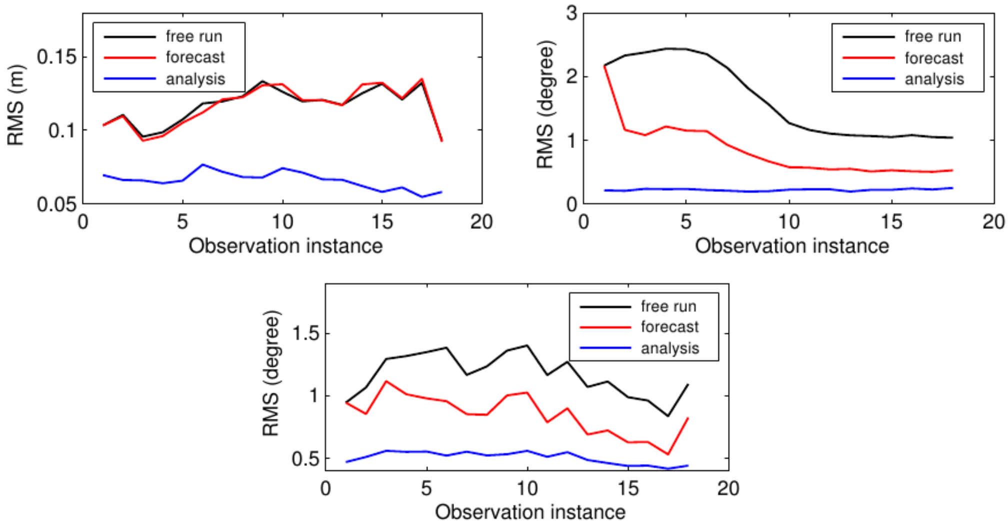


Figure 4.1: Time evolution of the spatially averaged RMS errors of the free run, the forecasts and the analyses compared to the observations used in the assimilation experiments for SSH (top left), SST (top right) and temperature profiles (bottom).

For the temperature profiles, at the beginning of the experiment, the RMS differences between the analyses and the forecasts (also the free run) are 0.5°C. The RMS errors of the free run and the forecast then increase, which results of a RMS difference between the analyses and the free run on the order of 0.9°C and a RMS difference between the analyses and the forecasts on the order of 0.5°C. Detailed insight shows that large residual exists near the surface in the subpolar area and in depth in the Gulf Stream region. Further investigation confirms the presence of the instability (density inversion) induced by the model state correction in the Gulf Stream region, which can explain partly the large residual in depth in this area.

Stochastic perturbation of the equation of state (CNRS/LGGE) As an illustration of the assimilation process described in the previous chapter, local time evolutions of the 96-member ensemble is shown in Figure 4.2 for 18 months (6 free run months and 12 assimilation cycle months). As already mentioned, we observe the saturation of the spread of the free run ensemble (cyan curves) with different amplitudes and time-saturations depending on the locations. This saturation shows that the local climatological variabilities are reached. The updated ensembles (blue dots for the forecast ensemble and green curves for the IAU ensemble) present a noticeable spread reduction and a temporal variability globally included in the climatological envelop defined by the free run ensemble saturation. This kind of ensemble behaviors could foreshadow an improvement of the probabilistic resolution without degrading the reliability. We also note the spread reduction and a slight bias correction with the IAU ensemble compared to the forecast one.

As a first comparison to the observations, Figure 4.3 shows the time series

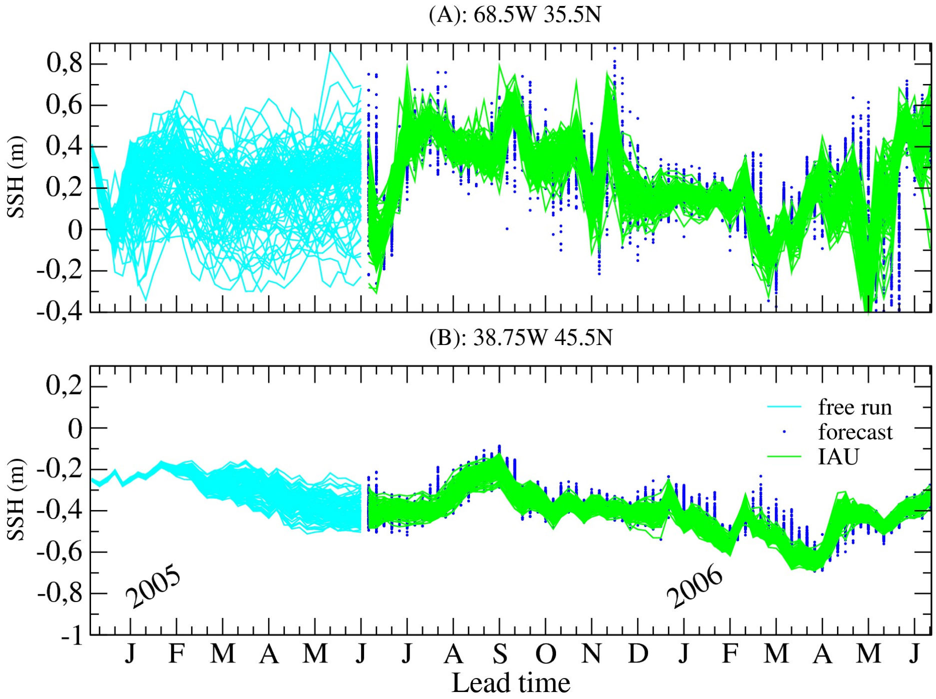


Figure 4.2: SSH time series of the ensembles (free run, forecast and IAU) at locations (A) and (B).

of the ensembles standard deviation averaged over Gulf Stream (left panel) and Z-focus areas (right panel). We first observe in the figure that the saturation of the standard deviation is not reached over the global Gulf Stream area contrary to the Z-focus one. This is mainly due to the fact that the Gulf Stream verification area contains locations where the perturbations grow very slowly (*e.g.* near the subtropical gyre). The main point here is the standard deviations reduction observed with the introduction of the altimetric corrections. This reduction is effective from the first assimilation cycle. After that, the averaged standard deviations are globally stabilized by the subsequent assimilation cycles. We also notice the clear reduction of the standard deviations of the IAU ensembles as compared to the forecast, meaning the stochastic perturbations are strong enough to produce significant spreads within 10 days and thus to avoid the ensemble collapse without introducing any inflating factor. The 10-day oscillations of the standard deviation come from the discontinuity of the forecast ensembles.

As mentioned above, the altimetric updates tend to reduce and stabilize the standard deviation, except the clear increase observed around September 2005 (especially over the large verification area). This results from a lack of observed data caused by lots of JASON-tracks missings. This circumstantial JASON failure shows the impact of the satellite coverage on the updated ensembles: this increases the ensembles standard deviation. That may sound obvious, but this shows that the accuracy of the correction is very sensitive to the number of available altimetric data.

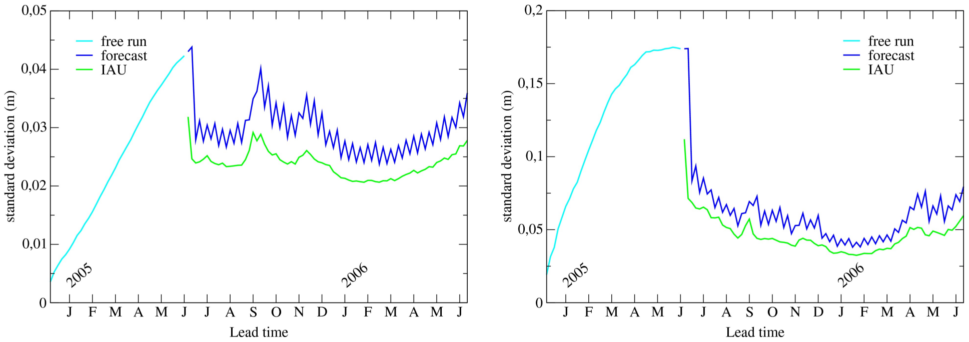


Figure 4.3: SSH time series of standard deviations for free run, forecast and IAU ensembles over the Gulf Stream verification area (left panel) and the Z-focus area (right panel).

4.3 Probabilistic validation

In this section, the ensemble distributions of the free run, the forecast and the analysis are diagnosed in a probabilistic way according to two criteria: reliability and resolution. For each assimilation experiment, the RCRV and CRPS scores are successively presented.

Stochastic perturbation of the forcing (GHER). The reliability is first investigated by RCRV score in terms of bias and dispersion. The RCRV scores for SSH is shown in Figure 4.4. For SSH, the bias of the analysis is closer to 0 compared to that of the forecast and of the free run, but the dispersion of the analysis is larger than that of forecast (slightly larger than 1), which indicates slight underestimation of the ensemble error. Towards the end of the assimilation experiment, the RCRV dispersion for SSH becomes close to 1, which indicates the improvement of the ensemble performance. For SST (not shown), the bias is greatly reduced by the analysis, which results of a value very close to 0 during the whole period of assimilation. The dispersion of the analysis is very close to 1. Therefore, the ensemble system is very reliable for SST with assimilation.

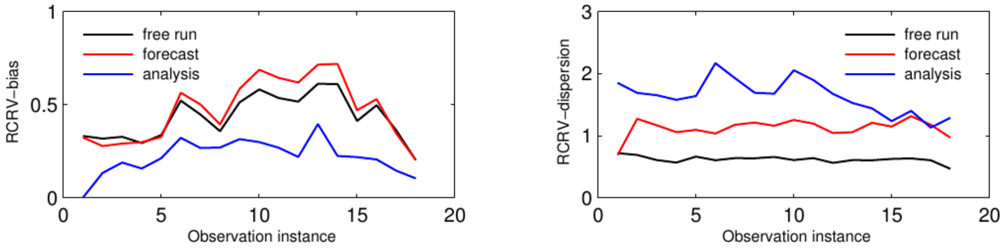


Figure 4.4: SSH bias (left panel) and dispersion (right panel) from RCRV.

For the temperature profiles (fig. 4.5), the bias is reduced by the analysis, with values close to 0. However, the relatively significant under-dispersion of ensemble exists from the beginning of the experiment until the 13th analysis step.

From these analyses, we can see that the slight degradation of reliability with assimilation for SSH and temperature profile is mainly due to the more or less significant underestimation of the ensemble spread.

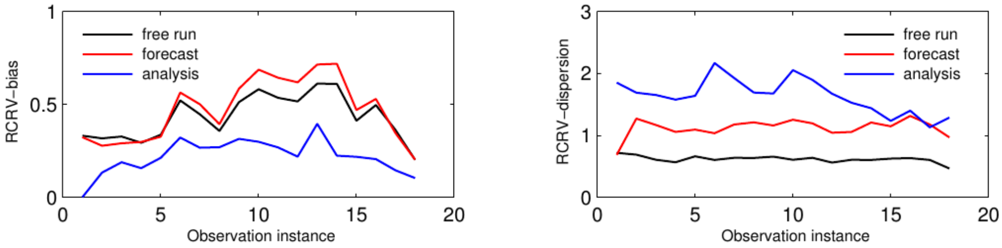


Figure 4.5: Temperature profile bias (left panel) and dispersion (right panel) from RCRV.

The CRPS and the associated decomposition for SSH are shown in Fig. 4.6. The CRPS of the analysis is smaller than for the free run and the forecast. In particular, between the 5th step and the 17th step, the CRPS of the free run and the forecast increase, but the CRPS of the analysis decreases. However, the CRPS of the forecast is slightly larger than that of the free run, this can be explained by the fact that the SSH analysis is not used in the increment update. The correction of SSH in the model depends only on the model adjustment following the correction of the temperature and the salinity. Parasitic correction of SSH can be present during a certain period because of the presence of the gravity waves. The decomposition of CRPS shows that the assimilation improves the resolution, but degrades slightly the reliability of the ensemble forecast system. For the reliability, before the 16th step, the CRPS-Reli of the analysis remains relatively stable with a small decreasing tendency. While large fluctuation is observed for the CRPS-Reli of the free run and the forecast because of the model state oscillation, especially at the 10th and 14th steps. At these two steps, the CRPS-Reli of the forecast is larger than that of the analysis. For the resolution, the CRPS-Reso of the free run, the forecast and the analysis remain stable with the assimilation experiment going on. Very small difference exists between the free run and the forecast, while large difference exists between the analysis and the free run.

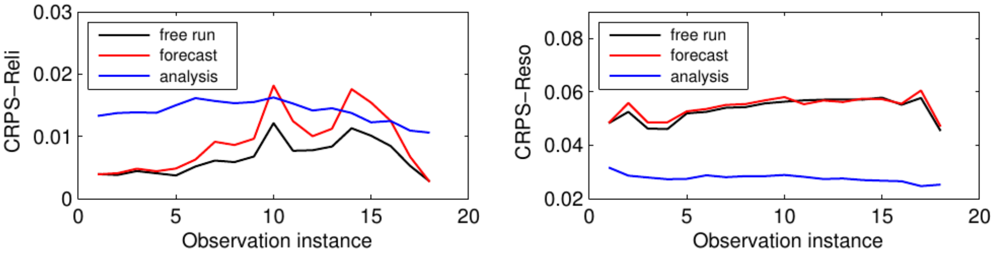


Figure 4.6: SSH: reliability (left panel) and resolution component of the CRPS compared to the uncertainty (right panel).

Figure 4.7 shows the CRPS for the temperature profile. Again, the assimilation improves the resolution, but degrades slightly the reliability. Compared to the uncertainty, the resolution is much smaller, which indicates an informative ensemble system for the temperature.

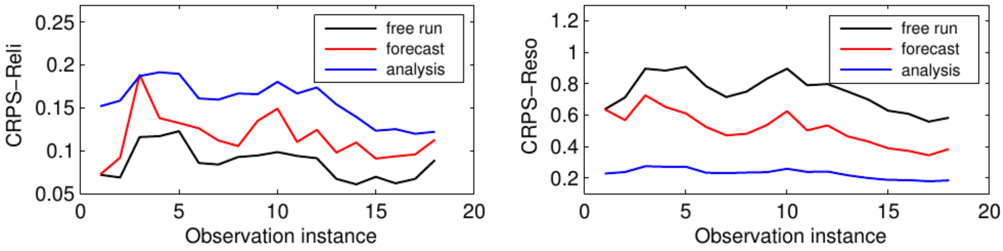


Figure 4.7: Temperature profile: reliability (left panel) and resolution component of the CRPS compared to the uncertainty (right panel).

Stochastic perturbation of the equation of state (CNRS/LGGE). As a first probability diagnostic, we again investigate the reliability property through the bias and the dispersion related to the RCRV for SSH. In Figure 4.8, 3 kinds of curve are shown: dashed curves for verification against JASON data, dotted curves for verification against ENVISAT data and full curves for verification against both satellites data. Also note that —for all the presented probabilistic scores in this section— the statistics are accumulated over one month and over the Gulf Stream area. Remark also, that unlike the forecast ensemble, the IAU ensemble is checked against observations that have been used to compute the increments. The data from the ensemble system compared to the observations are thus no longer independent. In this figure, the forecast ensemble and observations are the real independent data.

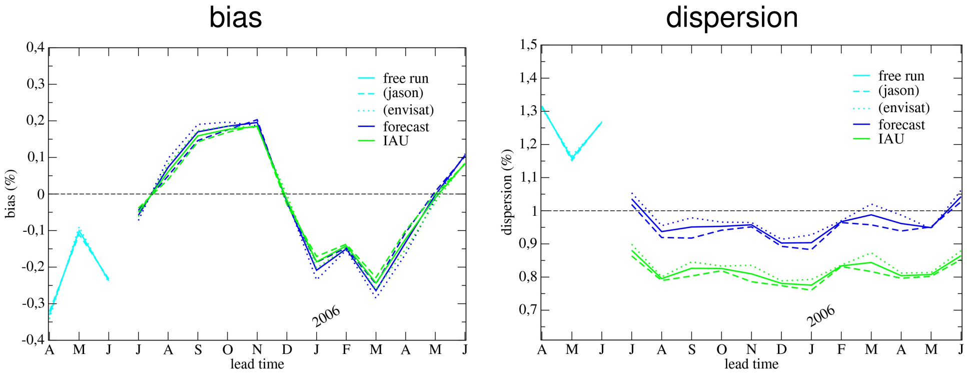
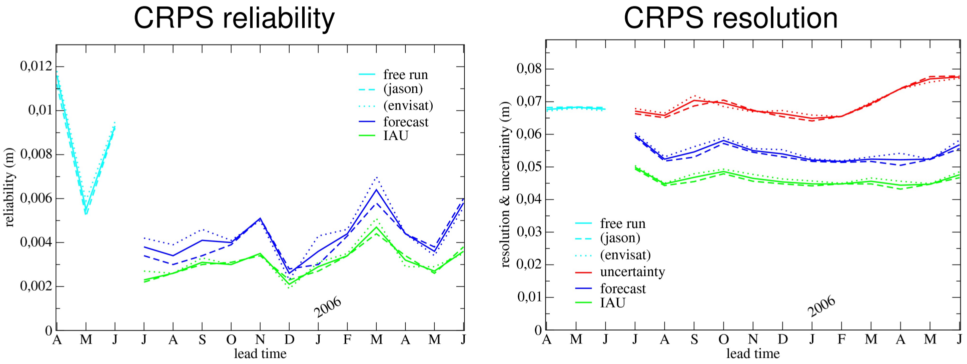


Figure 4.8: SSH bias (left panel) and dispersion (right panel) from RCRV.

The monthly time series (left panel) show no clear bias reduction compared to the free run ensemble with altimetric corrections. Nevertheless, the IAU ensemble reduces the bias as compared to the forecast. This is due to the inbreeding between the IAU ensembles and the observations in this case. About the dispersion (right panel), the underdispersion of the free run ensembles is removed by the altimetric corrections. The forecast ensemble becomes close to perfectly dispersive while a slight overdispersion ($\approx 85\%$) remains in the IAU ensemble (mostly due to the inbreeding with the observations). On the other hand, regarding the spread, two results look contradictory at first sight: we observe the reduction of the spread in the IAU ensembles (Fig. 4.3) and at the same time, the IAU ensembles overdispersion (Fig. 4.8 right panel). Actually, the spread of the ensemble is reduced, but also the bias against the observations so that the dispersion is finally degraded down to an overdispersive system.


CRPS reliability

CRPS resolution

Figure 4.9: SSH: reliability (left panel) and resolution component of the CRPS compared to the uncertainty (right panel).

We now investigate the global CRPS score for the SSH variable. Figure 4.9 shows the reliability (left panel) and the resolution (right panel) components of the CRPS. The resolution part is compared to the uncertainty associated with the verification dataset (this represents the climatological variability over the verification area and period). The reliability of the ensembles is improved by the altimetric corrections compared to the free run ensemble. Also, the system becomes more informative after assimilation processes (better resolution), *i.e.* the ensemble system is more accurate in space and time for instance by correctly translating the eddies along the Gulf Stream front. For both components of the CRPS, the IAU ensemble performs better than the forecast (these scores are negatively oriented), partially due to the inbreeding between the IAU ensembles and the observations.

The resolution curves in Fig. 4.9 (right panel) show that the free run ensemble has no resolution, *i.e.* $\mathcal{G} \approx 0\%$. If we only look at the independent verification data, *i.e.* the forecast ensemble, the potential gain we get with the assimilation process is up to 30%. For the SSH variable, the assimilation process leads to an information gain with a reliability improvement: the uncertainty on the state of the flow is reduced by 30% and this information is reliable.

Chapter 5

Conclusion

The main originalities of this study are:

1. to explicitly represent uncertainties inherent to the ocean circulation, using two different approaches: one with perturbations of the forcing (GHER), and one with perturbations of the equation of state (CNRS/LGGE);
2. to use this as a basis to the implementation of a full 4D-ensemble assimilation system in a realistic configuration, using the same ensemble assimilation method;
3. to validate the results using probabilistic metrics (rank histogram, RCRV and CRPS scores).

The study shows that both ensembles correctly represent the climatological variability of the eddy-dynamics over the Gulf Stream area, especially in the frontal regions. The ensemble systems (without assimilation) tend to be reliable —even if globally slightly underdispersive— but provide no useful information (null probabilistic resolution associated with climatological system) as a result of the chaotic nature of the eddy flow.

These ensembles are then updated by assimilating observations using the same full 4D-ensemble assimilation method: covariance matrix propagated by the ensemble and ensemble anomalies to observations computed at the exact observation time and location. The updated ensemble systems then become more reliable —the underdispersion is reduced— and more informative compared to the climatology. The updated ensemble systems are thus probabilistically more skillful.

The experiments presented in this study show promising results in uncertainty reduction, especially in the way it is diagnosed using probabilistic metrics, but we are conscious that the two systems are still far from optimal. First, the methods used to simulate uncertainty in the model are still very specific. Other sources of uncertainty should be introduced in order to make assimilation more effective everywhere. Moreover, many parameters could be optimized all along the process. For instance, the stochastic perturbations amplitudes were only roughly tuned to produce a reasonable dispersion of the ensemble. The same has been done to choose the localisation parameters, in order to keep a sufficient dispersion after the observational update. Despite of our efforts to go to more objective tuning and validation of the method, many choices —even if reasonable— are thus still subjective. This is probably the point where more effort should be put in future work.

Recommandation to COPERNICUS

As a final conclusion to SANGOMA WP4, we thus strongly recommand to follow the same general route in the development of the Copernicus operational systems, i. e.

- *by making progress in the explicit simulation of model uncertainties using a stochastic approach;*
- *by progressively moving to a probabilistic description of the operational products (using ensemble simulations); and*
- *by generalizing the use of probabilistic metrics to evaluate the quality of the products, and their impact on end-user applications.*

We believe that this is presently the easiest and most effective way to increase the quality of the operational products, and to improve the accountability of the operational system.

Chapter 6

References

Anderson J. 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 1518–1530.

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.

Barth, A., Alvera-Azcarate, A., Beckers, J., Staneva, J., Stanev, E. V., and Schulz-Stellenfleth, J., 2011. Correcting surface winds by assimilating High-Frequency Radar surface currents in the German Bight. *Ocean Dynamics*, **61**, 599–610.

Bishop H.C., B.J. Etherton and S.J. Majumdar, 2001. Adaptive sampling with the Ensemble Transform Kalman Filter. Part I: theoretical aspects. *Mon. Wea. Rev.*, **129**, 420–436

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

Candille G., Brankart J.-M., and Brasseur P., 2015. Assessment of an ensemble system that assimilates Jason-1/Envisat altimeter data in a probabilistic model of the North Atlantic ocean circulation. *Ocean Science*, **11**, 425-438.

Evensen G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, **99**, 10143–10162

Ourmières Y., J.-M. Brankart, L. Berline, P. Brasseur and J. Verron, 2006. Incremental analysis update implementation into a sequential ocean data assimilation system. *J. Atmos. Oceanic Technol.*, **23**, 1729–1744.

Yan, Y., Barth, A., and Beckers, J., 2014. Comparison of different assimilation schemes in a sequential Kalman filter assimilation system. *Ocean Modelling*, **73**, 123–137.

Yan Y., A. Barth, J.-M. Beckers, G. Candille, J.-M. Brankart and P. Brasseur, 2014. Ensemble assimilation of ARGO temperature profile, sea surface temperature and altimetric satellite data into an eddy permitting primitive equation model of the North Atlantic Ocean. *JGR - Oceans*. **120**, 5134–5157.