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

Y. Yan, A. Barth, J.-M. Beckers, Guillem Candille, J. M. Brankart, Pierre Brasseur

AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Oceans

RESEARCH ARTICLE

10.1002/2014JC010349

Key Points:
- Ensemble assimilation into a realistic North Atlantic Ocean model
- Both deterministic and probabilistic validations of the results
- Consistency and complementarity of both validations highlighted

Correspondence to:
Y. Yan,
yajing.yan@univ-smb.fr

Citation:
Yan, Y., A. Barth, J. M. Beckers, G. Candille, J. M. Brankart, and P. Brasseur (2015), 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, *J. Geophys. Res. Oceans*, *120*, 5134–5157, doi:10.1002/2014JC010349.

Received 29 JUL 2014
Accepted 26 JUN 2015
Accepted article online 1 JUL 2015
Published online 23 JUL 2015

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

Y. Yan[1,2], A. Barth[1], J. M. Beckers[1], G. Candille[3], J. M. Brankart[3], and P. Brasseur[3]

[1]GeoHydrodynamics and Environment Research, MARE, AGO, University of Liège, Liège, Belgium, [2]LISTIC, Université Savoie Mont Blanc, Annecy-le-Vieux, France, [3]CNRS, Laboratoire de Glaciologie et Géophysique de l'Environnement, Université de Grenoble, Grenoble, France

Abstract Sea surface height, sea surface temperature, and temperature profiles at depth collected between January and December 2005 are assimilated into a realistic eddy permitting primitive equation model of the North Atlantic Ocean using the Ensemble Kalman Filter. Sixty ensemble members are generated by adding realistic noise to the forcing parameters related to the temperature. The ensemble is diagnosed and validated by comparison between the ensemble spread and the model/observation difference, as well as by rank histogram before the assimilation experiments. An incremental analysis update scheme is applied in order to reduce spurious oscillations due to the model state correction. The results of the assimilation are assessed according to both deterministic and probabilistic metrics with independent/semiindependent observations. For deterministic validation, the ensemble means, together with the ensemble spreads are compared to the observations, in order to diagnose the ensemble distribution properties in a deterministic way. For probabilistic validation, the continuous ranked probability score (CRPS) is used to evaluate the ensemble forecast system according to reliability and resolution. The reliability is further decomposed into bias and dispersion by the reduced centered random variable (RCRV) score in order to investigate the reliability properties of the ensemble forecast system. The improvement of the assimilation is demonstrated using these validation metrics. Finally, the deterministic validation and the probabilistic validation are analyzed jointly. The consistency and complementarity between both validations are highlighted.

1. Introduction

Nowadays, advanced numerical ocean circulation models, the increase in supercomputing facilities, as well as the development of ocean observing systems make operational prediction systems possible. Data assimilation plays an important role in the operational prediction system. Numerous data assimilation methods are now well-established. The Ensemble Kalman Filter (EnKF) [*Evensen*, 2003], one of the most used stochastic methods, has been extensively used and it works well for very large problems like operational prediction problems. Compared to other assimilation methods, one of the major advantages of EnKF is the benefit of flow-dependent background error which is of fundamental importance for both analysis and information measures [*Buehner*, 2004].

Verification of assimilation results helps operational forecasters to interpret forecasts, to understand model biases, and to select models for use in different conditions. Therefore, meaningful verification approaches are of particular importance and have received considerable attention. The conventional evaluation of the assimilation results (deterministic validation) is essentially based on the root-mean square (RMS) error, which provides a score to evaluate the deterministic system. When it is applied to an ensemble forecast system, the ensemble mean is usually taken into consideration. In this way, the difference between the forecast and the observations are considered as the forecast error. However, the uncertainty associated with the forecast is not taken into account. Thus, the richness of the probabilistic forecast given by the ensemble is not investigated exhaustively. Sophisticated methods that take into account the distribution properties of the ensemble seems necessary to better exploit the probabilistic characteristics of the ensemble. For this

reason, new methods for the evaluation of forecast probability distributions, and further investigation into the properties of the conventional verification measures for probability forecasts have been developed. Proper verification practice and correct interpretation of verification statistics have been extensively promoted in recent publications [*Hamill*, 2000; *Hersbach*, 2000; *Candille et al*., 2006; *Casati et al*., 2008; *Candille et al*., 2014]. Probabilistic verification relies on two criteria: reliability and resolution, which correspond to the main attributes of an ensemble forecast system. The reliability corresponds to the statistical consistency between a priori predicted probabilities and a posteriori observed frequencies of the occurrence of the event under consideration. The resolution indicates the ability of a forecast system to separate a priori cases when the event under consideration occurs more or less frequently than the climatological frequency [*Toth et al*., 2003]. The continuous ranked probability score (CRPS) [*Stanski et al*., 1989] and the Brier Score (BS) [*Brier*, 1950; *Murphy*, 1973; *Wilks*, 1995] allow for an evaluation of an ensemble forecast system according to these two criteria. The reliability can also be evaluated by reliability diagram [*Hartmann et al*., 2002], rank histogram [*Anderson*, 1996; *Talagrand et al*., 1999], the reduced centered random variable (RCRV) score [*Talagrand et al*., 1999], and so on. Until now, both deterministic and probabilistic scores have been used by meteorological centers [*Talagrand et al*., 1999; *Candille et al*., 2006]. However, there have been few issues published that have focused on the similarity and/or difference of the behaviors of deterministic and probabilistic scores in both perfectly and imperfectly reliable ensemble forecast systems. It seems interesting to use both scores jointly to validate an ensemble forecast system in order to highlight the redundant and complementary information provided by both scores.

In this paper, we implemented assimilation experiments with an ocean circulation model for an operational ocean circulation prediction system over the North Atlantic Ocean, the NATL025 configuration of the NEMO model [*Barnier et al*., 2006], with the EnKF. The observations include Jason-1 sea surface height (SSH) data [*Ménarda et al*., 2003], AVHRR sea surface temperature (SST) data [*Casey et al*., 2010], and ARGO temperature profiles [*Davis*, 1991]. The ensemble, with 60 members, is created by generating perturbations in the forcing variables related to the temperature. The assimilation is performed every 10 days for 1 year, 2005, with the first 180 days as ensemble spin up time. In order to reduce the spurious oscillations induced by intermittent model state correction, the Incremental Analysis Update (IAU) [*Yan et al*., 2014] is applied instead of the conventional intermittent assimilation scheme. The assimilation results are validated according to both deterministic and probabilistic metrics, which goes further than most previous studies and constitutes one of the original points of this paper. For deterministic validation, the assimilation results, more precisely the ensemble means, are compared to the observations used in the assimilation experiments and independent/semi-independent observations. Thermohaline variables (SSH, SST, temperature, salinity) and horizontal velocities are considered. For probabilistic validation, scores such as rank histogram, CRPS, and RCRV are computed for thermohaline variables in order to diagnose the ensemble distribution properties. Through comparisons between the free run and the assimilation experiments, the positive impact of the data assimilation is demonstrated. Furthermore, the deterministic validation and the probabilistic validation are investigated jointly in order to highlight the consistency of both validations, as well as their complementarity. The connections between the ensemble mean/spread versus observation plot and the CRPS and RCRV scores are highlighted. Highly reliable situations, in which the RMS error and the CRPS give similar information, are identified for the first time in this paper.

This paper is organized as follows: the model configuration is described in section 2. In section 3, observations, ensemble generation and validation, assimilation methods, and experimental setups are introduced in detail. Section 4 is dedicated to metrics for both deterministic and probabilistic validations. Discussions of results are given in section 5. Finally, the conclusion is derived in section 6.

2. Model Configuration

The circulation of the North Atlantic is simulated by the OPA code (NATL025 configuration of the NEMO model [*Barnier et al*., 2006]) using 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 primitive equations are discretized on an Arakawa C grid, with a horizontal resolution of $1/4^\circ \times 1/4^\circ \cos(\phi)$ (where ϕ is the latitude), which is considered as eddy-permitting in the midlatitudes where the Rossby radius of deformation is about 100 km. The effective resolution, which

becomes finer with increasing latitude, is ~27.75 km at the equator and ~13.8 km at 60°S or 60°N. Vertical discretization takes place on 46 geopotential levels, with a grid spacing increasing from 6 m at the surface to 250 m at the bottom. The maximum depth in the model is 5844 m.

Partial step (PS) topography [*Adcroft et al*., 1997], making the depth of the bottom cell variable and adjustable to the real depth of the ocean, is used to represent flow-topography interactions. Momentum advection is performed with an energy and enstrophy conserving (EEN) numerical scheme [*Arakawa and Lamb*, 1981; *Barnier et al*., 2006] in vector form, with an additional term in the momentum equation to damp the faster external gravity waves. Tracer advection is performed with a total variance diminishing advection scheme to avoid the generation of overshoots in the case of sharp gradients. Lateral mixing of tracers is modeled with a Laplacian lateral isopycnal diffusion operator, 300 m^2s^{-1} at the equator and decreasing poleward, proportional to the grid size. Lateral mixing of momentum is modeled with a horizontal biharmonic viscosity operator, -1.5×10^{11} m^4s^{-1} at the equator and decreasing poleward by the cube of the grid size. Surface boundary layer mixing and interior vertical mixing are parametrized according to a turbulence kinetic energy (TKE) turbulence closure model [*Blanke and Delecluse*, 1993]. In case of static instability, a viscosity/diffusivity enhancement of up to 10 m^2s^{-1} is used. The forcing fluxes are calculated via bulk formulations, using the ERAinterim atmospheric forcing fields [*Dee et al*., 2011]. The temperature and salinity fields are initialized using the Levitus climatology [*Levitus et al*., 1998]. The horizontal and vertical velocity fields are initially set to zero, as is the SSH.

The model spin up time is 16 years, starting from January 1989. According to *Testut et al*., [2003], on one hand, numerical integration that is too long tends to deteriorate the three-dimensional distribution of temperature and salinity; yet on the other hand, a too short integration leaves some dynamical features unadjusted. Therefore, a model spin-up of 16 years is chosen [*Kantha and Clayson*, 2000]. Time stepping is performed with a leap frog scheme, with a time step of $\Delta t=2400$ s.

3. Assimilation Experiments

The assimilation was performed in 2005 for a period of 1 year with an assimilation window of 10 days. The assimilation cycle must be long enough to accumulate a 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 [*Skachko et al*., 2009]. The first 180 days are ensemble spin up time. It is important to integrate the ensemble over a time interval covering a few characteristic time scales of the dynamical system [*Evensen*, 2003] to ensure dynamic stability and correct multivariate correlations before commencing the assimilation. Afterward, on one hand, the free ensemble run is performed over the last 180 days in order to compare the forecasts for this period to the hindcast experiments. On the other hand, the EnKF is activated during the last 180 days in order to assimilate the observations into the model integration over time.

3.1. Assimilation Method

The assimilation tool used for the experiments is the Ocean Assimilation Kit (OAK) [*Vandenbulcke et al*., 2006; *Barth et al*., 2007a, 2008; *Vandenbulcke et al*., 2010]. The assimilation method provided in OAK is the square root analysis scheme of EnKF [*Evensen*, 2004]. The IAU scheme is applied instead of intermittent assimilation in order to reduce, by keeping the mass and momentum fields in balance, the spurious oscillations produced by intermittent assimilation. Specifically, the IAU 0 scheme [*Yan et al*., 2014] is used because of its advantages in reducing high-frequency analysis-induced oscillations but not increasing the computation time compared to other IAU assimilation schemes. In this scheme, at the end of each assimilation window, the analysis is performed using observations around each analysis step. An increment is calculated from the difference between the analyzed and the forecasted model states. This increment is then added to the model integration for the subsequent 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].

3.2. Observations

Three types of observations, Jason-1 SSH data [*Ménarda et al*., 2003], AVHRR SST data [*Casey et al*., 2010], and ARGO temperature profiles [*Davis*, 1991], are available at all assimilation steps. Examples of observation

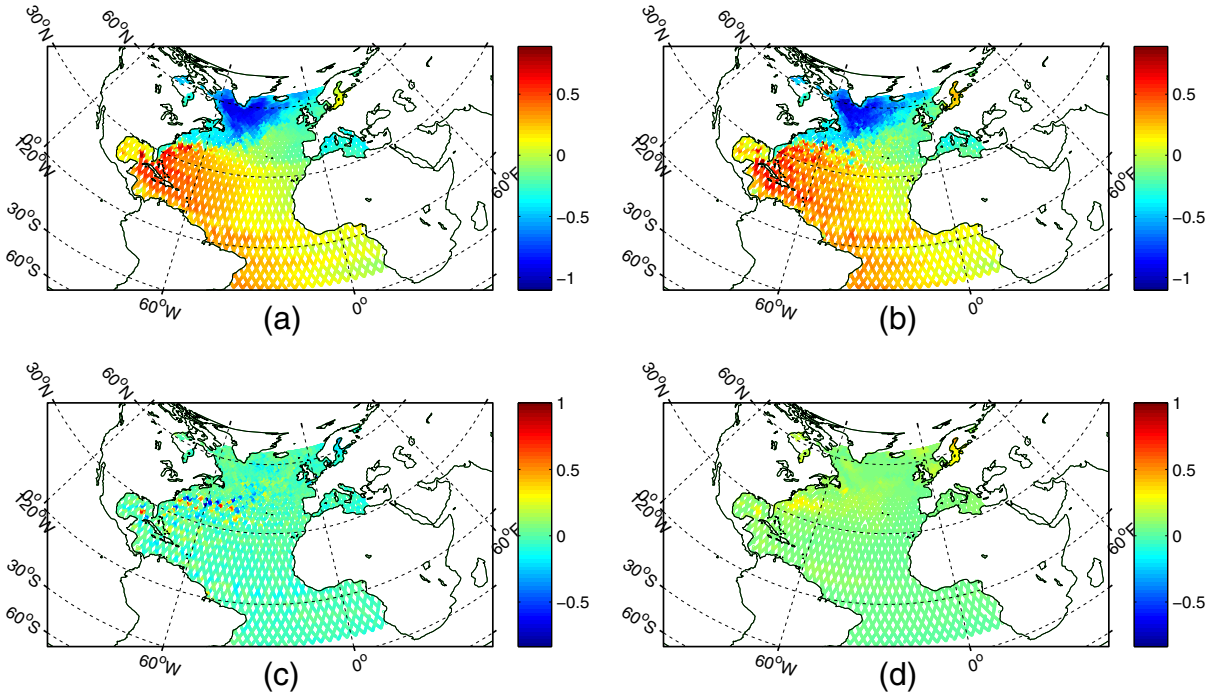

Figure 1. Snapshot of (a) model (without perturbation in forcing) (b) observation, Jason-1 (c) model/observation difference (model - observation), and (d) ensemble spread for SSH at the end of the ensemble spin up.

distributions are shown in Figures 1b, 2b, and 4b. In general, SSH observation grids correspond to a typical along/across track resolution (11.2 km × 5.1 km), with SSH observations globally covering the North Atlantic basin except for the subpolar area. For SST, the observation grids are very dense with a resolution of $1/4^\circ \times 1/4^\circ$, covering the whole North Atlantic basin. However, the observation grids of temperature profile are sparse, with observation points located between the surface and 2000 m depth only in part of the North Atlantic basin (about 2500 points at each analysis step). The observational errors for SSH and ARGO temperature data are 5 cm and 0.3°, respectively. These values are based on nominal error of these data and representative error found in the literature [*Testut et al*., 2003]. For SST data, the observational error is the standard deviation map associated with the temperature value, with a mean value of 0.20° (maximum of 1.28° and minimum of 0.11°). Moreover, ENVISAT SSH data [*Resti et al*., 1999], Mercator reanalysis SST data [*Ferry et al*., 2012], and ARGO salinity profiles are used as validation observations to evaluate the assimilation results. Note that, the Mercator reanalysis SST data are not completely independent of the AVHRR SST data assimilated in the experiments (hereafter, semiindependent observation). However, since the accuracy of the Mercator reanalysis is improved with respect to that of the AVHRR data, they can be used to judge the assimilation results.

An analysis localization method is used to rule out corrections due to distant observations. Such corrections exist when error correlations occur between distant grid points. These corrections are often unreliable. For the EnKF, the localization approach was discussed in *Houtekamer and Mitchell* [2001]; *Hamill and Whitaker* [2001]. Here, we apply an approach similar to *Testut et al*. [2003] adapted to the square root analysis scheme of EnKF provided by OAK. To compute the correction at each water column, the observations are weighted by a factor of $\exp\left(-r^2/d^2\right)$ with d the localization length scale. The localization length scale is determined according to the autocorrelation length of SST and SSH, here 300 km.

3.3. Ensemble Generation and Validation

Uncertainties in an assimilation system occur for many different reasons: model dynamics, parameters, forcing, and initial and boundary conditions. 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),

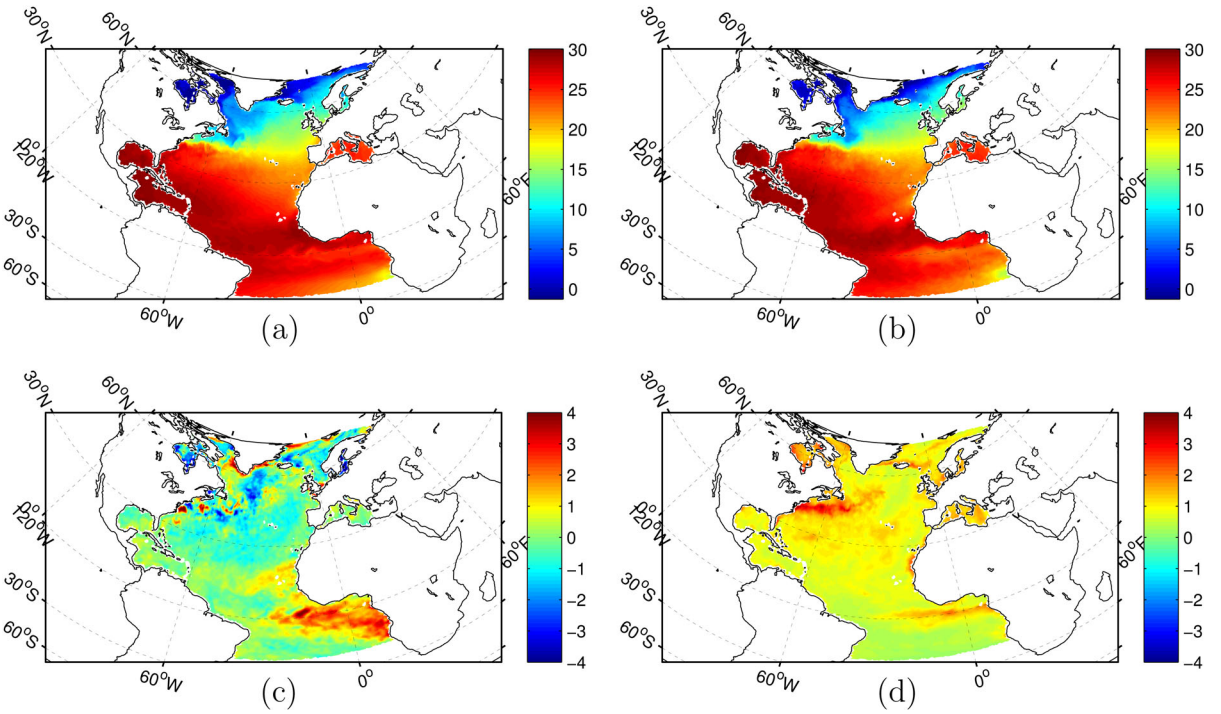

Figure 2. Snapshot of (a) model (without perturbation in forcing) (b) observation (c) model/observation difference (model - observation), and (d) ensemble spread for SST at the end of the ensemble spin up.

the long-wave radiation (radlw) and short-wave radiation (radsw) are considered. The temporal variability of the forcing variable is obtained by the Fourier decomposition (a series of angular frequencies and the associated Fourier coefficients) of the forcing variable vector. The perturbation is generated by combining the angular frequencies that we are interested in (where different frequency corresponds to different variability) and a random time series with a temporal decorrelation scale determined by the corresponding angular frequency. More details of the principle are explained in *Barth et al.* [2011] and *Marmain et al.* [2014]. 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 ensemble spin up time is 180 days to ensure dynamic stability and correct multivariate correlations before commencing the 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 in forcing variables and the observations. For this, SSH, SST, and temperature profiles are considered. Moreover, rank histograms for these three variables are computed in order to validate the ensemble distribution properties in a probabilistic way. A flat rank histogram indicates an on average good ensemble spread. An asymmetrical distribution is usually an indication of a bias in the mean of the ensemble, while a U or inverted U-shape distribution may be an indication of an underdispersive or overdispersive system [*Hamill*, 2000]. Note that, however, a perfectly reliable ensemble system might display a U-shape distribution due to observation uncertainties in some cases.

The model prediction, the observation, the model/observation difference (model - observation), and the ensemble spread at the end of the ensemble spin up time for SSH and SST are shown in Figures 1 and 2. There is a good global agreement between the model and the observation for these two variables. For SSH, general 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 observations is 0.097 m, while the ensemble spread is 0.086 m. Furthermore, the rank histogram of SSH over the Gulf Stream region (Figure 3a) confirms the good representation of the model error by the ensemble (Here, the rank histogram is only computed in the Gulf Stream region, because the variability among ensemble members is mainly located in this area.).

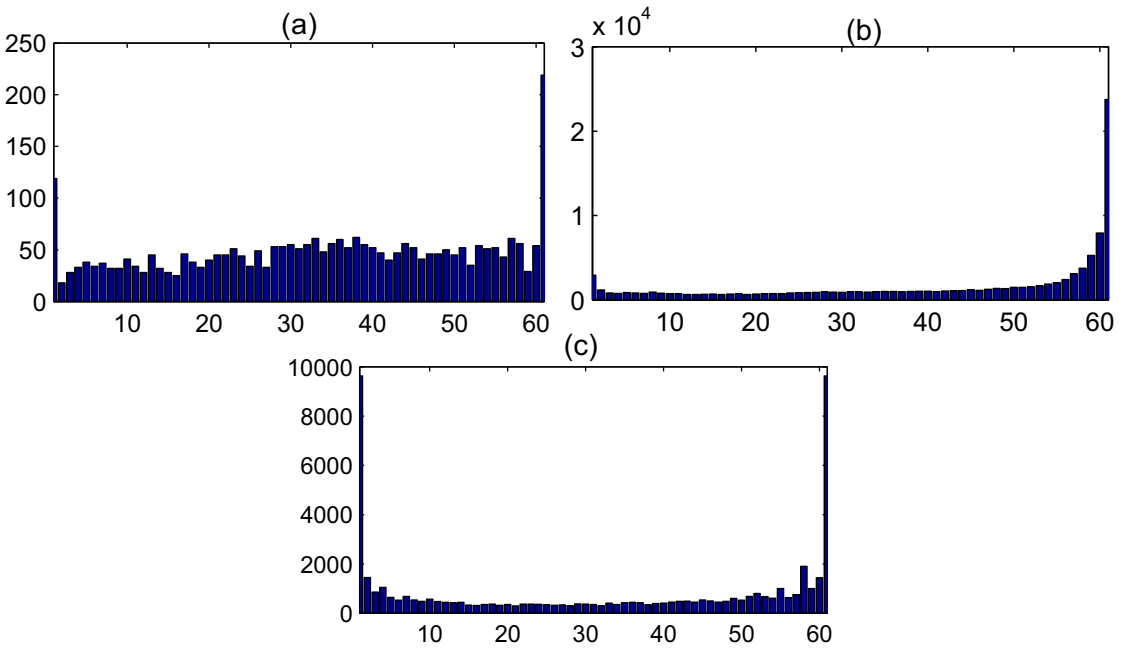

Figure 3. Rank histogram for (a) SSH in the Gulf Stream region, (b) SST, and (c) temperature over the whole basin at the end of the ensemble spin-up.

For SST, large errors are located in the subpolar area, in the Gulf Stream region, and off the African coast. The RMS error of the model compared to the observations is 1.252°, while the ensemble spread is 0.9206°. These errors are consistent and consistent with the general realistic ocean circulation model error (∼1°) [*Kaplan et al*., 1998]. Moreover, the rank histogram of SST over the whole North Atlantic basin (Figure 3b) confirms the sufficiency of the ensemble spread. Consequently, the model error represented by the ensemble spread is considered to be realistic for SSH and SST.

Note that the rank histogram of SST is not symmetrical (Figure 3b), which indicates the presence of bias in the ensemble mean. In order to highlight the model bias, the averaged model/observation difference of SST over 6 months (the first 6 months of 2005) is analyzed. The SST bias distribution is consistent with that of the NATL3 configuration of the NEMO model (1/3° resolution) used in *Testut et al*. [2003] (not shown). Strong anomalies are observed in three areas: the subpolar area (probably related to forcing errors), the Gulf Stream region (related to systematic errors in the Gulf Stream pathway in the model and insufficient resolution of the forcing), and off the African coast (reflecting a weakness in the representation of the African upwelling off Senegal).

For the temperature profiles, the global variation predicted by the model is consistent with the observations (Figure 4). However, according to the model/observation difference, larger differences exist under the surface and above 1500 m depth. Compared to the model/observation difference, the ensemble spread is smaller in these areas. The global RMS error of the model compared to the observation is 1.511°, while the ensemble spread is 0.5°, three times smaller than the RMS error. The U-shaped rank histogram (Figure 3c) also indicates an underdispersion of the present ensemble. However, note that the model/observation difference for SST reaches 5° in many areas: at the surface off the African coast, in the subpolar area and at depth in the Gulf Stream region. The spatial distribution of this large difference corresponds to that of the SST model bias mentioned previously. Therefore, the large model/observation difference is mainly due to large model bias in these areas that we cannot represent by stochastic perturbation. Regarding the spatial distribution of the ensemble spread, large values are observed near the surface over the whole basin and at depth in the Gulf Stream region. The ensemble is generated from forcing perturbations, thus ocean model state variables at the surface are more involved. The variation between ensemble members at depth depends only on the model dynamic variation during the ensemble spin up time. In the Gulf Stream region, the model dynamic is strong, large variation is thus observed at depth in this area. While in other areas

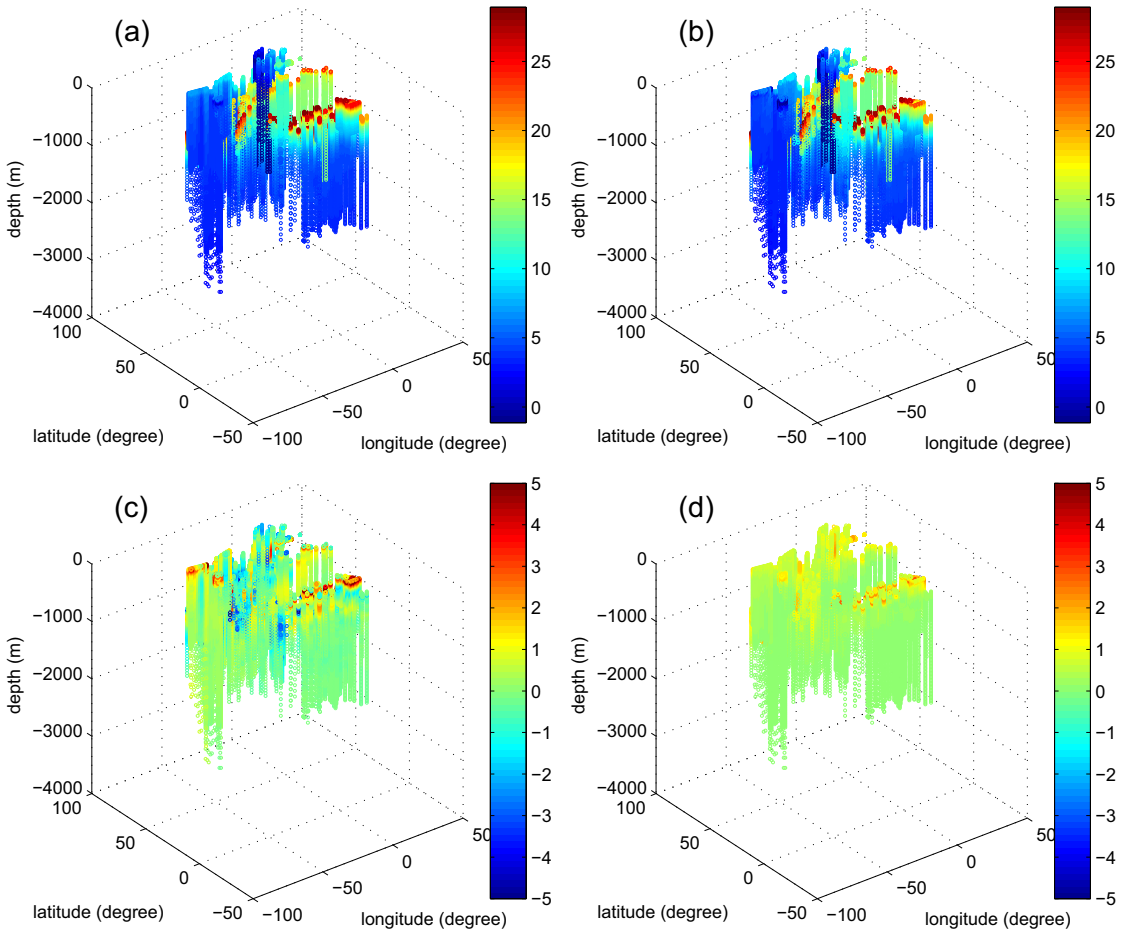

Figure 4. Snapshot of (a) model (without perturbation in forcing), (b) observation, (c) model/observation difference (model - observation), and (d) ensemble spread for the temperature profile at the end of the ensemble spin up.

where the model dynamic is weak, even with longer ensemble spin up time (1 year), no larger variation is observed at depth. On the contrary, longer ensemble spin up results in a larger spread of SSH. Although the magnitude of the ensemble spread is smaller than the model/observation difference, it is comparable with the general realistic ocean circulation model error. Taking into account the model bias, the ensemble spread for temperature is also considered realistic.

3.4. Model State Vector

The model state vector for the assimilation consists of three variables: SSH, temperature, and salinity. The model state here is referred to the assimilation tool, and is different from the prognostic variables of the ocean circulation model mentioned in section 2. The velocities are not directly corrected with the other prognostic variables in the assimilation experiments in order to avoid instability of the model dynamics. Their corrections depend on the geostrophic balance relationship between the temperature, the salinity, and the velocities. Note also that SSH constitutes one of the model state variables, it is analyzed in the assimilation experiments, but its increment is not included during the increment update. Only the temperature and salinity increments are incorporated in the model integration. The main reason for this lies in the fact that the variation of SSH is related to variations of the stratification and the current. Based on the temperature and salinity correction, the model will adjust the SSH following the stratification and velocity changes to retrieve the geostrophic state. Even if the SSH increment was included during the increment update, similar balanced model state would be obtained. However, inappropriate SSH increment with respect to the temperature and salinity increments may cause instability of the model dynamics.

4. Validation Metrics

In order to objectively evaluate the assimilation experiments, we rely on both deterministic and probabilistic metrics. For deterministic validation, first, the ensembles (mean and spread) of thermohaline variables (SSH, SST, temperature, salinity) are compared to independent/semiindependent observations. Second, the horizontal (zonal and meridional) velocity is assessed by comparison to the velocity field generated from four satellites (ENVISAT [*Resti et al*., 1999], Jason-1 [*Ménarda et al*., 2003], TOPEX/Poseidon [*Fu et al*., 1994], and GFO [*Hancock et al*., 2001] data (DEOS) (http://rads.tudelft.nl/gulfstream) in order to diagnose if the correction of the state vector is sufficiently robust to permit the adjustment of the model dynamics.

In probabilistic validation, the performance of the ensemble forecast system is diagnosed according to reliability and resolution. The CRPS measures the distance of how far the ensemble system is found from the verifying observations according to these two criteria. It has tended to be the first choice method for the verification of operational ensemble forecasts [*Casati et al*., 2008]. According to *Hersbach* [2000]; *Candille et al*. [2006], the CRPS can be decomposed into CRPS-Reli and $CRPS_{pot}$: where CRPS=CRPS-Reli+$CRPS_{pot}$. CRPS-Reli measures the reliability of an ensemble system. $CRPS_{pot}$ can be further decomposed into CRPS-Uncert and CRPS-Reso: where $CRPS_{pot}$=CRPS-Uncert−CRPS-Reso. CRPS-Reso expresses the resolution of an ensemble system. CRPS-Uncert is the lowest possible CRPS value based on climatology. It is solely determined by the climatology and does not depend on the performance of the forecast model. Here, the climatology is defined by the verifying observation. Thus, for a given variable and verifying observations, any changes in $CRPS_{pot}$ are due to changes in CRPS-Reso. Therefore, $CRPS_{pot}$ is often used to provide information on the resolution of an ensemble forecast system, due to the ease of its computation. A detailed illustration of the computation of CRPS and its decompositions is given in Appendix A. CRPS, CRPS-Reli, and $CRPS_{pot}$ are negatively oriented. The smaller they are, the better an ensemble is. An ensemble system with CRPS value of 0 always exactly reproduces the verifying observation without any ensemble spread. CRPS-Reli is equal to 0 if the system is perfectly reliable. A significant positive value of CRPS-Reli indicates the lack of reliability of the system. $CRPS_{pot}$ reaches its minimum for a perfect deterministic system and positive values quantify a lack of resolution [*Candille et al*., 2006]. In general, a broad distribution of the verification sample corresponds to a large CRPS-Uncert, while a sharp distribution of the verification sample corresponds to a small CRPS-Uncert. Since the ensemble system has positive resolution if it outperforms the climatological probabilistic forecast or the verification observations, $CRPS_{pot}$ is smaller than CRPS-Uncert. The smaller $CRPS_{pot}$ is than CRPS-Uncert, the more informative the ensemble system is.

The reliability can be further decomposed into bias and dispersion. For this, the RCRV score allows for the investigation of the reliability property of an ensemble forecast system in terms of bias and dispersion. The definition of the RCRV score is given in equation (1). The average of RCRV, referred to as RCRV-bias, is computed over all realizations of the system and represents the weighted bias between the ensemble and the observation. The standard deviation of RCRV, referred to as RCRV-dispersion, constitutes an indicator of systematic over and under dispersion of the ensemble. It measures the agreement of the ensemble spread and the specified observational error with the observed amplitude of the forecast error. A perfectly reliable system has no bias and a dispersion equal to 1. A significant negative (positive) value of bias indicates a positive (negative) bias. A value of dispersion significantly larger (smaller) than 1 characterizes the underdispersion (overdispersion) of the system.

$$\mathrm{RCRV}=\frac{y_o-\bar{x}}{\sqrt{\sigma_o^2+\sigma^2}} \tag{1}$$

where y_o is the observation, σ_o represents the observation error, $\bar{x}$ corresponds to the ensemble mean, and σ denotes the ensemble spread.

5. Results Analyses and Discussions

In this section, the assimilation results are analyzed and discussed. Comparisons are made between the free run (model prediction without data assimilation), the forecast (model prediction based on data assimilation at previous steps), and the analysis (combination of the model prediction and the observation). Evaluation is performed according to the metrics defined in section 4. First, deterministic validation is performed on both thermohaline variables (assimilated variables: SSH, SST, temperature profile and unassimilated variable:

salinity profile) and horizontal velocity. For thermohaline variables, the ensemble mean is compared to the observations and the RMS error is analyzed. The coupled ensemble mean/spread of thermohaline variables is then compared with the observations in order to evaluate the ensemble distribution properties. Second, probabilistic validation is performed on thermohaline variables. The ensemble system is evaluated by CRPS in terms of reliability and resolution. The reliability of the ensemble system is also evaluated by RCRV in terms of bias and dispersion. Finally, both deterministic validation and probabilistic validation are investigated jointly in order to highlight the consistency and complementarity of both validations. Note that for SSH and salinity profile, independent observations, ENVISAT altimetric data and ARGO profiles, are used for validation. For SST, semiindependent observations, Mercator reanalysis, are used for validation. While for temperature profile, assimilated observations are used since no independent observations are available for validation. In this analysis, the scores for temperature and salinity are calculated without taking into account the volume represented by the model grid point, because the degradation of salinity after the analysis is mainly located at the surface and its contribution will be very small if the volume represented by the point is taken into account. The objective of this paper is to show under which circumstances the deterministic validation and the probabilistic validation are consistent/inconsistent rather than to demonstrate the good performance of the assimilation system. The emphasis of the degradation of the salinity at the surface by assimilation provides an unreliable situation where a significant difference between both validations is observed.

5.1. Deterministic Validation

5.1.1. Thermohaline Variables

The temporal evolutions of spatially averaged RMS errors of the ensemble means of SSH, SST, temperature, and salinity profiles in the free run/forecast/analysis are shown in Figure 5.

For SSH (Figure 5a), the reduction of the RMS error in the analysis is not so significant at the beginning of the experiments. At the sixth and eighth steps, the RMS errors of the analyses are slightly larger than those of the free run/forecast. Detailed inspections showed that besides in the Gulf Stream region and off the African coast, a large residual is observed at each step in the subpolar area where no Jason-1 observations were available in the assimilation experiments (not shown). At the sixth and eighth steps, the smaller number of ENVISAT observations available for validation causes larger global RMS errors compared to other steps. Note that the RMS errors of the forecast are very close to those of the free run, this can be explained by the fact that SSH increment is not updated in the assimilation experiments, the correction of SSH in the model state depends on the interactions between temperature, salinity, and SSH during the model integration. For SST (Figure 5b), the RMS error of the free run is more than 2° at the beginning of the experiments, increases slightly and then decreases to 1° from the tenth step. The RMS errors of the analyses remain stable at 0.2° from the beginning until the end of the experiments. Detailed analyses show that large residuals (more than 1°) are located in the Gulf Stream region and subpolar area where a large model bias has been identified (not shown). Both the model bias and random error are reduced, especially off the African coast. This significant improvement results from the good observation distribution and small observation error that allow the assimilation system to be well constrained by SST observations. For temperature profile (Figure 5c), the RMS errors of the analyses are much smaller than those of the free run since the beginning of the experiments. According to detailed insight, large residuals exist near the surface in the subpolar area and at depth in the Gulf Stream region (not shown). Further investigation confirms the presence of the instability (density inversion) induced by the model state correction in the Gulf Stream region (not shown), which can partly explain the large residual at depth in this area. Regarding salinity profile (Figure 5d), RMS reduction of about 0.02 psu is obtained by assimilation compared to the free run. Large residuals are observed near the surface over the whole basin and at depth in the Gulf Stream region (not shown). The former may be related to errors present in the covariance matrix mainly due to the use of a relatively small ensemble size. Detailed analyses of the covariance matrix show that the salinity at the surface is over correlated with SST at the beginning of the experiments (not shown). Since the assimilation system is strongly constrained by SST, the salinity at the surface is thus corrected too much at the beginning of the experiments. The latter can be explained by instability (density inversion) induced by the model state correction in the Gulf Stream region.

Besides the ensemble mean, the coupled ensemble mean/spread of SSH, SST, temperature, and salinity profiles are compared to the observations and the associated observation errors (Figures 6–9). The behaviors of

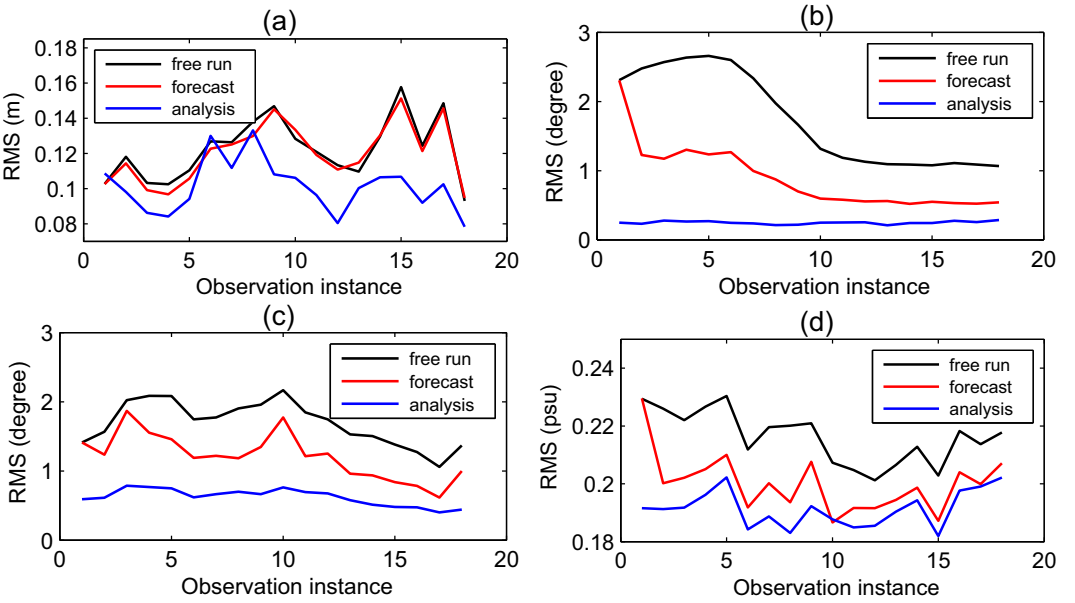

Figure 5. Temporal evolution of spatially averaged RMS errors of the ensemble means of (a) SSH (b) SST (c) temperature profile, and (d) salinity profile in the free run/forecast/analysis. For SSH, SST, and salinity profile, independent/semiindependent observations are used. For temperature profile, assimilated observations are used since no independent observations are available.

SSH are similar to those of the temperature profile. In both the free run and the forecast, the observations are always included within the ensemble spread intervals, although the difference between the ensemble mean and the observation is large, since the ensemble spread is large. In the analysis, the difference between the ensemble mean and the observation is reduced. Meanwhile, the ensemble spread is reduced and the ensemble slightly underestimates the analysis error at some steps. Therefore, the observations are not always included within the ensemble spread intervals. However, taking also the observation errors into account, according to Figures 6d and 7d, the SSH and temperature anomalies are always included within the observation error intervals. Consequently, the ensemble can be considered good enough to represent

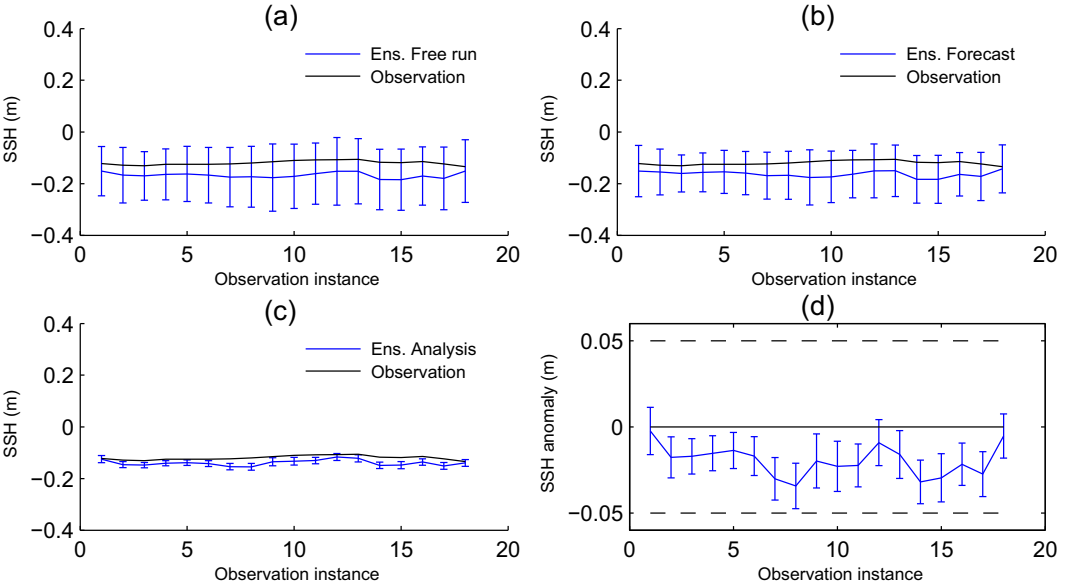

Figure 6. Ensemble mean/spread versus independent observations (Envisat altimetric data) of the (a) free run (b) forecast (c) analysis for SSH, averaged over the whole domain of the model grid. (d) SSH anomaly (ensemble mean of the analysis - observation) versus ensemble spread (error bar) and observation error (dashed line). The black line corresponds to the observation.

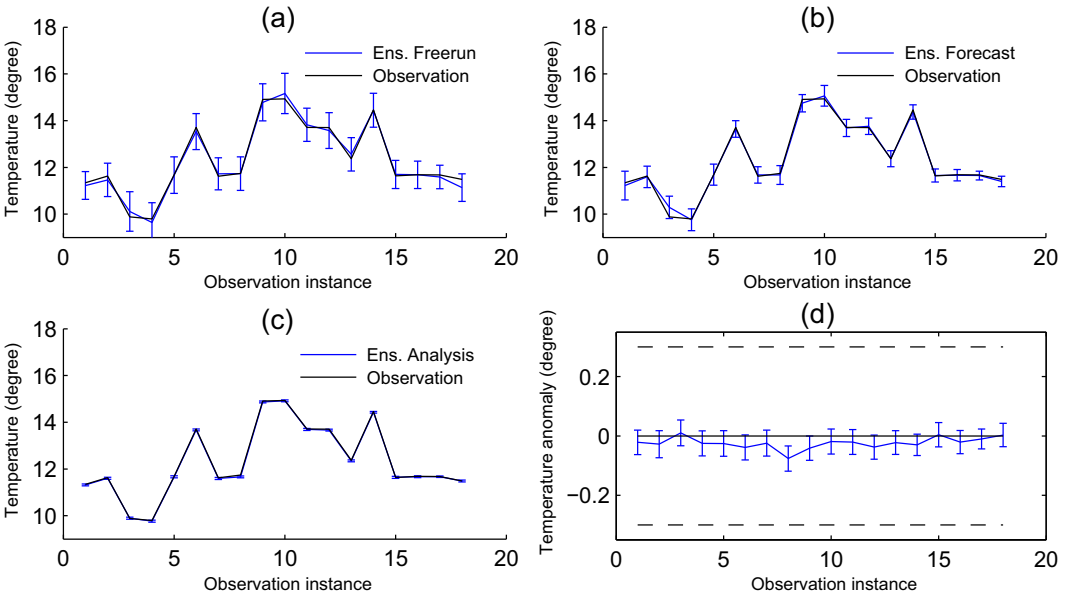

Figure 7. Ensemble mean/spread versus assimilated observations of the (a) free run (b) forecast (c) analysis for temperature profile, averaged over the whole domain of the model grid. (d) Temperature anomaly (ensemble mean of the analysis - observation) versus ensemble spread (error bar) and observation error (dashed line). The black line corresponds to the observation.

the ensemble error for these two variables. Note that the ensemble members are always smaller than the observations, therefore negative bias exists in the ensemble for these two variables. Potential improvement for slight underestimation of ensemble error can be obtained by slightly increasing the observation error in the assimilation experiments. In this way, the correction of the model state by the observation will be slightly smaller, the difference between the ensemble mean and the observation will thus be slightly larger, but the ensemble spread will not be reduced significantly by the analysis and it can represent the ensemble error appropriately. Moreover, potential improvement can also be obtained by using a nondiagonal matrix of observation error. Because of the use of a diagonal matrix for the observation error, the underestimation of the ensemble error after the analysis can also come from a misrepresentation of a horizontal and/or vertical correlation in the observation error.

For SST (Figure 8), at the beginning of the experiment, the difference between the ensemble mean and the observation is large in both the free run and the forecast, with the presence of large negative model bias. The ensemble spread seems insufficient and the observations are thus not included in the ensemble spread intervals. Toward the end of the experiments, this situation is improved. In the analysis, the difference between the ensemble mean and the observations is very small since the beginning of the experiments, large model bias is efficiently reduced by the analysis. The ensemble spread is also very small, in the order of 0.1°, but corresponds to good representation of the ensemble error in the analysis. The SST anomaly in Figure 8d confirms these conclusions.

For salinity profile, the ensemble spread seems large in the free run, but the observations always lie at the upper limit of the ensemble spread intervals, which indicates a negative bias in the forecast model. After the analysis, the difference between the ensemble mean and the observation is always reduced compared to the free run, but the ensemble spread is reduced significantly. Due to this, the difference between the observations and the ensemble distribution becomes larger. Taking 0.02 psu as observation error [*Oka and Ando*, 2004], the salinity anomaly is situated below the observation error interval and there is no intersection between the lower observation error line and the upper ensemble spread line, which implies that the distance between the observation and the ensemble is larger than the system uncertainty (a combination of observation error and ensemble error, $\sqrt{\sigma_o^2+\sigma^2}$, see equation (1)). Compared to the other assimilated variables (SSH, SST, temperature profile), the correction for the unassimilated salinity is not satisfactory. However, note that these plots are made without taking into account the volume represented by the model grid point and that salinity degradation is mainly situated at the surface. Taking into account the volume

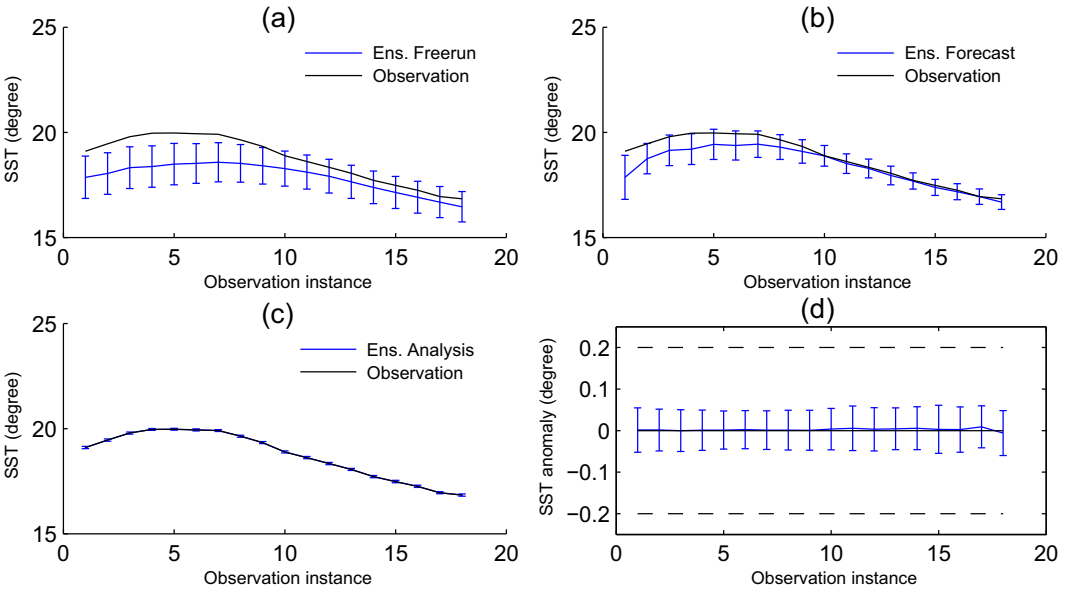

Figure 8. Ensemble mean/spread versus semiindependent observations (Mercator reanalysis) of the (a) free run (b) forecast (c) analysis for SST, averaged over the whole domain of the model grid. (d) SST anomaly (ensemble mean of the analysis - observation) versus ensemble spread (error bar) and observation error (dashed line). The black line corresponds to the observation.

effect, the weighted averaged difference between the observation and the ensemble is much smaller and the ensemble is always included within the observation error intervals. Potential improvement of salinity can be obtained by increasing the ensemble size and/or assimilating the ARGO salinity profiles.

5.1.2. Horizontal Velocity

Figure 10 shows an example of the difference of the horizontal velocity (at 3 m depth) in the free run and in the assimilation experiments in the Gulf Stream region on 16 November 2005. A semiindependent Gulf Stream velocity field generated from four altimetric satellite data (ENVISAT, Jason-1, TOPEX/Poseidon, and GFO) of the same date is available for validation. Compared to the free run, the assimilation intensifies the

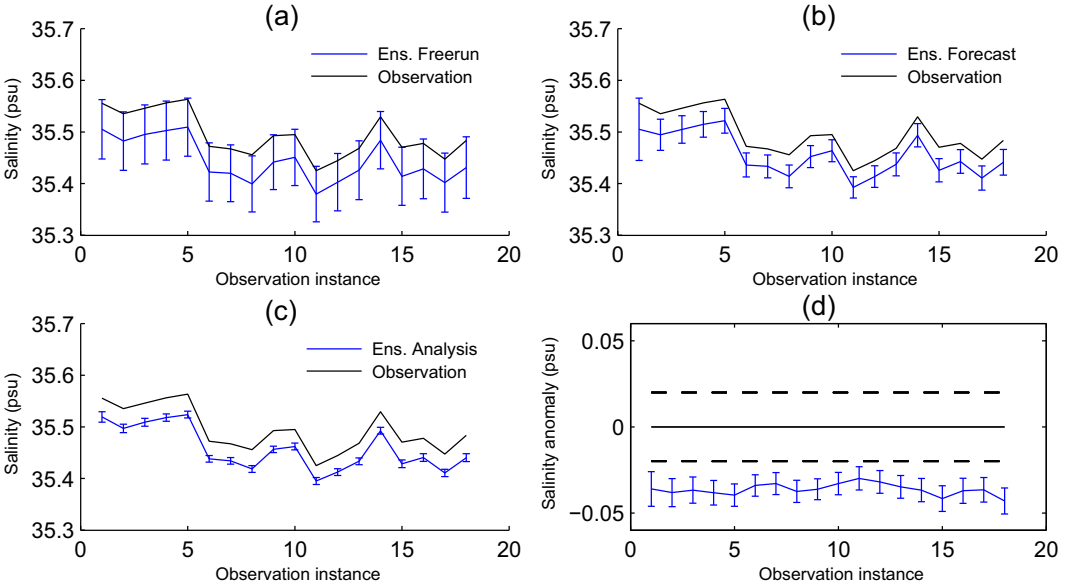

Figure 9. Ensemble mean/spread versus independent observations (ARGO profile) of the (a) free run (b) forecast (c) analysis for salinity, averaged over the whole domain of the model grid. (d) salinity anomaly (ensemble mean of the analysis - observation) versus ensemble spread (error bar). The black line corresponds to the observation.

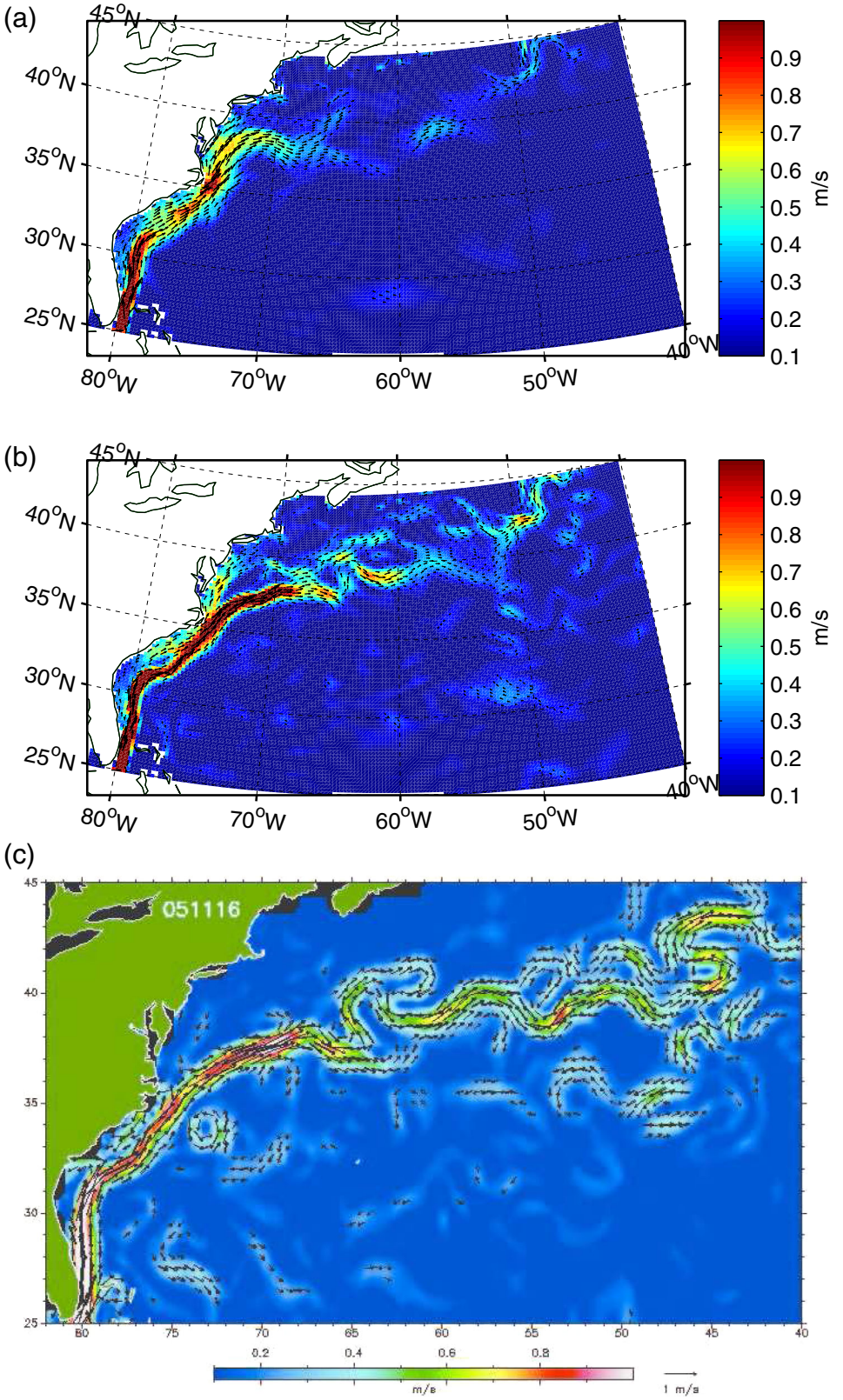

Figure 10. Horizontal velocity in the Gulf Stream region (at 3 m depth) on 16 November 2005. (a) Free run (b) assimilation, and (c) semiin-dependent observation from four altimetric satellite data (DEOS).

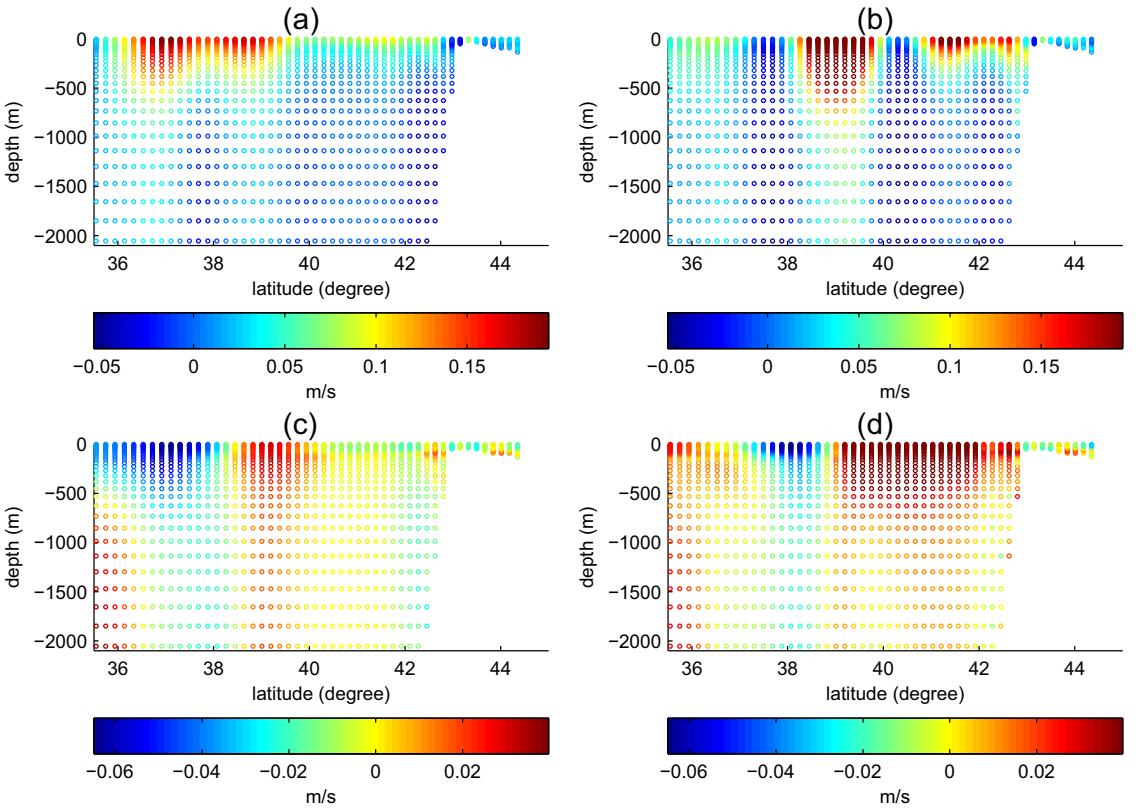

Figure 11. Meridional section in the Gulf Stream region (at 61.5° W) for the zonal (a) (b) and meridional (c) (d) velocities until 2000 m depth. Figures 11a and 11c in the free run and Figures 11b and 11d in the assimilation experiments averaged over 6 months.

main current in this area and changes the direction of the current efficiently at about 70° W. The velocity field obtained in the assimilation experiments is more similar to that of semiindependent observation. More eddy activities and meanders are generated by assimilation between 60° W and 40° W around 40° N (the front area). The positive impact of the assimilation experiments on horizontal current and the associated transport is thus highlighted. According to detailed inspections, the variances and covariances of the model state variables are significant in the front area, the correction of the temperature and the salinity by the assimilation gradually modifies the density of the water and then the current direction following the geostrophic adjustment during the model integration. The benefit of flow-dependent background error of EnKF is thus highlighted.

In order to analyze the impact of the assimilation on zonal and meridional velocities at depth, a vertical section in the Gulf Stream is performed. The averaged zonal and meridional velocities over 6 months in the free run and in the assimilation experiments are shown in Figure 11. Near the surface, the velocities are consistent with the DEOS horizontal velocity field. More eddy activities and meanders are generated by assimilation at the front area (about 40° N) where a sharp temperature change is located (not shown). At depth, the zonal and meridional velocities are intensified as expected. This vertical section demonstrates that the assimilation consistently modifies the three-dimensional flow.

5.2. 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. First, the CRPS and its decompositions (CRPS-Reli, $CRPS_{pot}$, and CRPS-Uncert defined in section 4 and Appendix A) of SSH, SST, temperature, and salinity profiles are analyzed. Second, the RCRV scores are computed for these four variables in order to investigate further the reliability of the ensemble distribution.

The CRPS and its decompositions for SSH are shown in Figure 12. The behaviors of the CRPS are similar to those of the RMS error (Figure 5a). They have similar temporal variation. The CRPS of the analysis is smaller

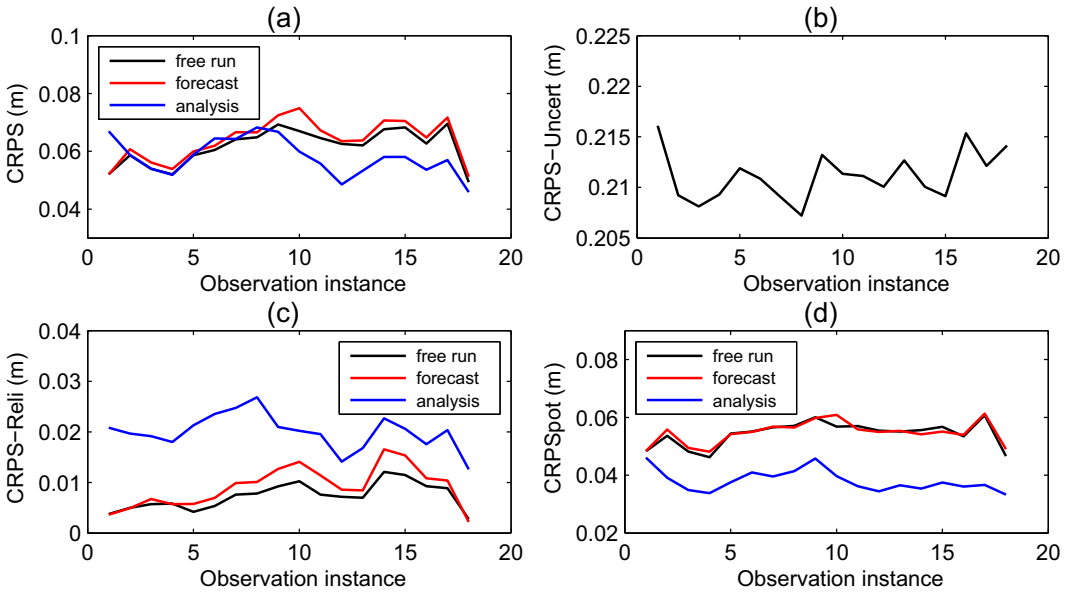

Figure 12. (a) CRPS, (b) CRPS-Uncert, (c) CRPS-Reli, and (d) $CRPS_{pot}$ for SSH. Independent observations are used for verification.

than those of the free run and the forecast only from the ninth step. The improvement by assimilation is thus not so significant, because of the presence of a large residual in the subpolar area without observations in the assimilation experiments. Note that the CRPS of the forecast is slightly larger than that of the free run, this can be explained by the parasitic correction present during a certain period because of the presence of gravity waves [*Barth et al.*, 2007b]. The decompositions of CRPS show that the assimilation improves the resolution, but degrades the reliability of the ensemble system. The degradation of the reliability is in the order of 0.01 m and compared to the system uncertainty, this degradation is small. Moreover, the peaks of reliability degradation around the eighth and fourteenth steps (Figure 12c) correspond to situations where observations lie outside the ensemble spread intervals on the ensemble mean/spread versus observation plots (Figure 6c). For the resolution, stable improvement of $CRPS_{pot}$ values by assimilation is observed, which is consistent with the ensemble spread reduction and the closeness between the ensemble mean and the observation, as well as the fact that the ensemble always lies entirely within the observation error interval. CRPS-Uncert is in the order of 0.21 m, which corresponds to the CRPS based on the verification observations only and without the performance of the forecast model. $CRPS_{pot}$ are much smaller than CRPS-Uncert, which implies an informative system, as well as the important role of the forecast model in the assimilation experiments. Also, CRPS and its decompositions are computed eliminating observation points in the subpolar area where a large residual is present. Smaller CRPS and CRPS-Reli values are obtained, but no obvious change of $CRPS_{pot}$ value is observed (not shown).

The CRPS and its decompositions for SST are shown in Figure 13. The behaviors of CRPS, CRPS-Reli, and $CRPS_{pot}$ are very similar to each other, and also to the RMS error (Figure 5b). The values of these scores for the analysis are close to 0 from the beginning and stay stable until the end of the experiments, which corresponds to an almost perfectly reliable system. The differences of $CRPS_{pot}$ between the free run/forecast and the analysis are large throughout the experiments, which indicates a good performance of the assimilation in terms of resolution improvement. $CRPS_{pot}$ are much smaller than CRPS-Uncert, which implies an informative ensemble system for SST. Consequently, the assimilation experiments improve both the reliability and resolution of SST. The quality of the analysis is ensured over the whole period of assimilation. These conclusions are very close to what we observe from the RMS error (Figure 5b) and the coupled ensemble mean/spread versus observation plot (Figure 8).

Figure 14 shows the CRPS and its decompositions for temperature profile. Compared to SSH, the behavior of the CRPS is even more similar to that of the RMS error and they have exactly the same temporal variation (Figure 5c). The assimilation improves the resolution, but slightly degrades the reliability (degradation in the

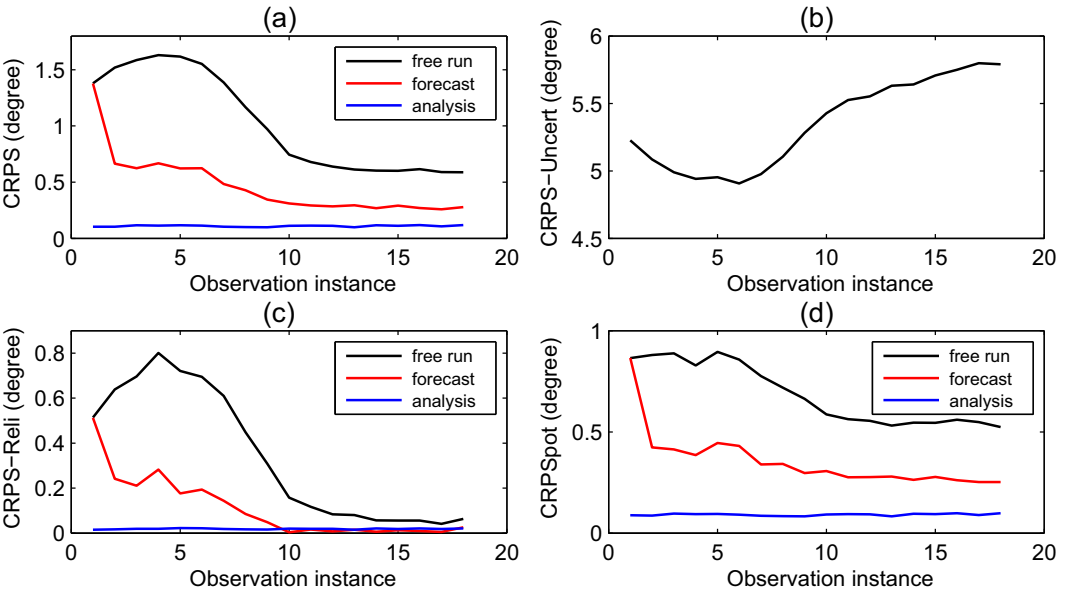

Figure 13. (a) CRPS, (b) CRPS-Uncert, (c) CRPS-Reli, and (d) $CRPS_{pot}$ for SST. Semiindependent observations are used for verification.

order of 0.07°, which is very small compared to the system uncertainty). CRPS-Uncert is in the order of 4°, indicating the poor quality of the verification observations in terms of CRPS without the performance of the forecast model. Compared to CRPS-Uncert, $CRPS_{pot}$ is much smaller, which indicates an informative ensemble system for the temperature.

In Figure 15, the CRPS and its decompositions for salinity profile are shown. Remind that the difference between salinity and the other variables mentioned previously is that the salinity observations are not assimilated in the experiments. According to Figure 15, no improvement is observed by the assimilation in terms of CRPS for salinity at the beginning of the experiments. Toward the end of the assimilation experiments, the CRPS of the analysis is even degraded compared to those of the free run and of the forecast.

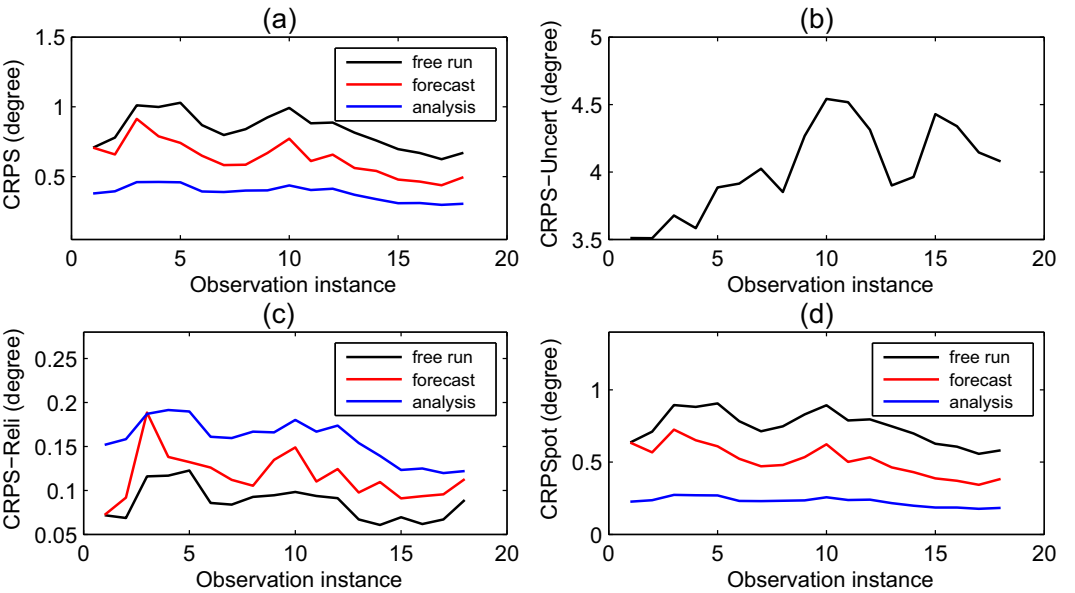

Figure 14. (a) CRPS, (b) CRPS-Uncert, (c) CRPS-Reli, and (d) $CRPS_{pot}$ for the temperature profile. Assimilated observations are used for verification.

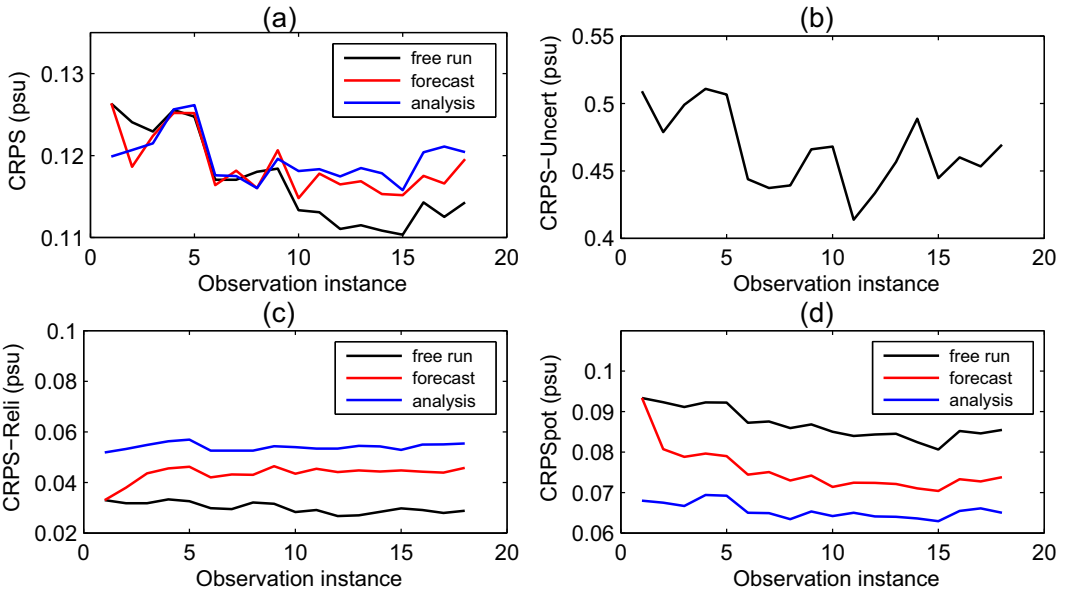

Figure 15. (a) CRPS, (b) CRPS-Uncert, (c) CRPS-Reli, and (d) $CRPS_{pot}$ for the salinity profile. Independent observations are used for verification.

The behavior of the CRPS here is not similar to that of the RMS error (Figure 5d), which is different from the other assimilated variables. CRPS-Reli and $CRPS_{pot}$ show that the assimilation significantly degrades the reliability with degradation in the order of 0.02 psu (large value compared to the system uncertainty), but still improves the resolution of the ensemble system. Indeed, the remark on the coupled ensemble mean/spread versus observation plot (Figure 9) (the difference between the ensemble means and the observations are reduced by assimilation at each step, but the observations lie farther away from the ensemble members after the analysis) explains the different behaviors of the CRPS and the RMS, since the CRPS measures the squared distance between the observation and the ensemble, while the RMS measures the squared distance between the observation and the ensemble mean. Furthermore, detailed analyses on the distribution of the ensemble members with respect to the observation at point scale (Figure 16) clearly show the difference between the temperature and the salinity. For the temperature, the observation value is 6.21°, the ensemble means for the free run and the analysis are 5.98° and 6.12°, respectively. The ensemble mean of the analysis is closer to the observation, which results in smaller RMS error of the analysis. At the same time, the ensemble spread is much reduced by the analysis, but it can still be representative enough for the ensemble error. The distribution of the ensemble members is closer to the observation distribution after the analysis. For the salinity, the observation value is 35.35 psu, the ensemble means of the free run and the analysis are 35.63 and 35.60 psu, respectively. The ensemble mean of the analysis is closer to the observation, thus we observe that the RMS error is reduced by the analysis. However, the dispersion of the ensemble becomes very small after the analysis, the ensemble spread is no longer representative of the ensemble error. The distribution of the ensemble members is farther from the observation distribution after the analysis. Therefore, even though the RMS error of the ensemble mean is reduced, the CRPS degrades, because the latter takes into account the ensemble distribution, not only the ensemble mean. It therefore follows that the reliability of the salinity is degraded, but the resolution of the salinity is improved as the ensemble mean is improved and the corresponding histogram becomes sharper. The lack of dispersion after analysis is due to the overestimation of the covariance (in absolute terms) between the salinity and the other observed variables (mainly SST). Probable reasons for this overestimation are that some model errors have not been taken into account in the stochastic model simulations, as well as the limited ensemble size.

Since the reliability is degraded for SSH, temperature, and salinity profiles, the reliability is further investigated by RCRV score in terms of bias and dispersion. The RCRV scores for SSH, SST, temperature, and salinity profiles are shown in Figure 17. For SSH, positive RCRV-bias value indicates negative bias present in the forecast model, which is consistent with the coupled ensemble mean/spread versus observation plot (Figure 6). Globally, the bias is not reduced by assimilation because of a large residual that exists in the subpolar area.

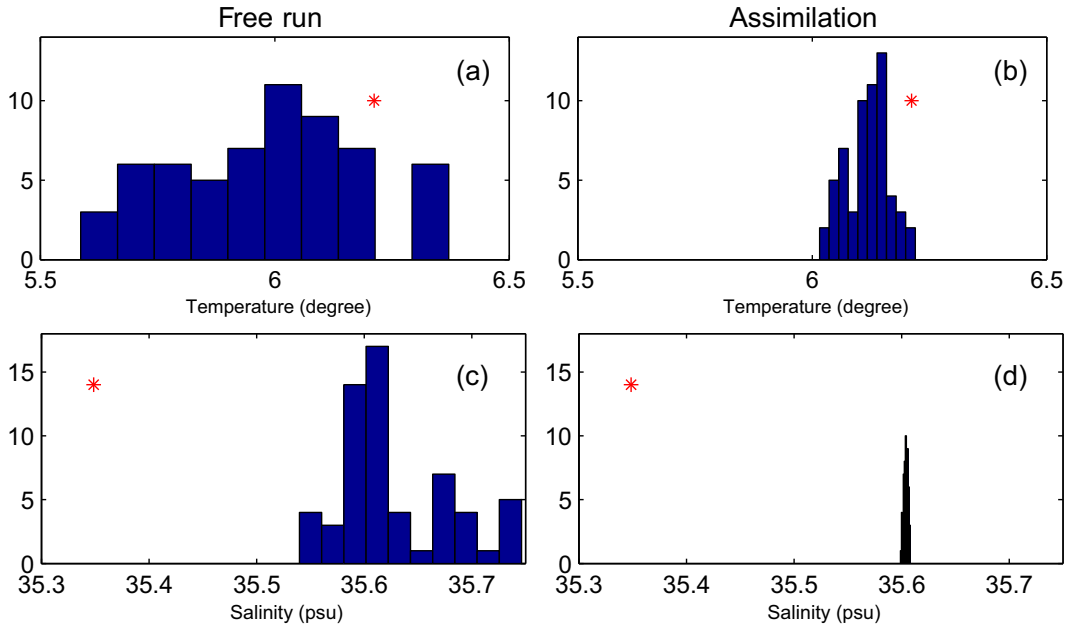

Figure 16. Example of ensemble members distribution with respect to the observation on a single point for the (a) (b) temperature (c) (d) salinity in the free run and in the assimilation experiments at the sixteenth step. The abscissa of the red star corresponds to the observation value.

The average RCRV-bias value in the order of 0.4 indicates the percentage, 40%, of the bias with respect to the system uncertainty. According to RCRV-dispersion values, the ensemble is slightly overdispersive in the free run and slightly underdispersive in the forecast. After the analysis, RCRV-dispersion values are larger than 2, which indicates an ensemble twice as underdispersive in comparison to a perfectly reliable case. When the points in the subpolar area are eliminated, RCRV bias is greatly reduced. However, RCRV dispersion is always larger than 2, which indicates that an ensemble underdispersion problem still exists (not shown). For SST, the behaviors of RCRV are very similar to those of CRPS and its decompositions (Figure 13). The large negative bias (1.5 times the system uncertainty) is greatly reduced by the analysis, which results in a RCRV-bias value very close to 0 during the whole period of assimilation. The RCRV-dispersion of the analysis is very close to 1, indicating a good agreement between the ensemble spread and the analysis error. Therefore, the ensemble system is very reliable for SST after the assimilation. For temperature profile, negative bias (on average 25% of the system uncertainty) is present in the forecast model, which is consistent with the coupled ensemble mean/spread versus observation plot (Figure 7d) and this bias is reduced by the analysis, with the RCRV-bias value very slightly larger than 0. The ensemble underdispersion problem exists in the free run (3 times underdispersive with respect to a perfectly reliable case) and it becomes more pronounced by the analysis, on average 5 times underdispersive with respect to a perfectly reliable case. From these analyses, we can see that the slight degradation of CRPS-Reli by assimilation for SSH and temperature profile is mainly due to the underestimation of the ensemble spread after the analysis. For salinity profile, negative bias (25% of the system uncertainty) is present in the forecast model, which is consistent with the coupled ensemble mean/spread versus observation plot (Figure 9) and the bias is increased by the assimilation to 1.5 times the system uncertainty. Moreover, a serious ensemble underdispersion problem exists, which results in very large RCRV-dispersion value. This significant ensemble underdispersion can also be seen in Figure 16. For this variable, the significant degradation of the CRPS-Reli is due to significant increase of the bias and serious underdispersion of the ensemble at the same time.

5.3. Joint Analysis of Deterministic Validation and Probabilistic Validation

SSH, SST, temperature, and salinity profiles, these four variables represent four different situations according to the improvement by assimilation. They are summarized in Table 1. For SST, an almost perfectly reliable system has been obtained with assimilation. For SSH and temperature profile, although the reliability is

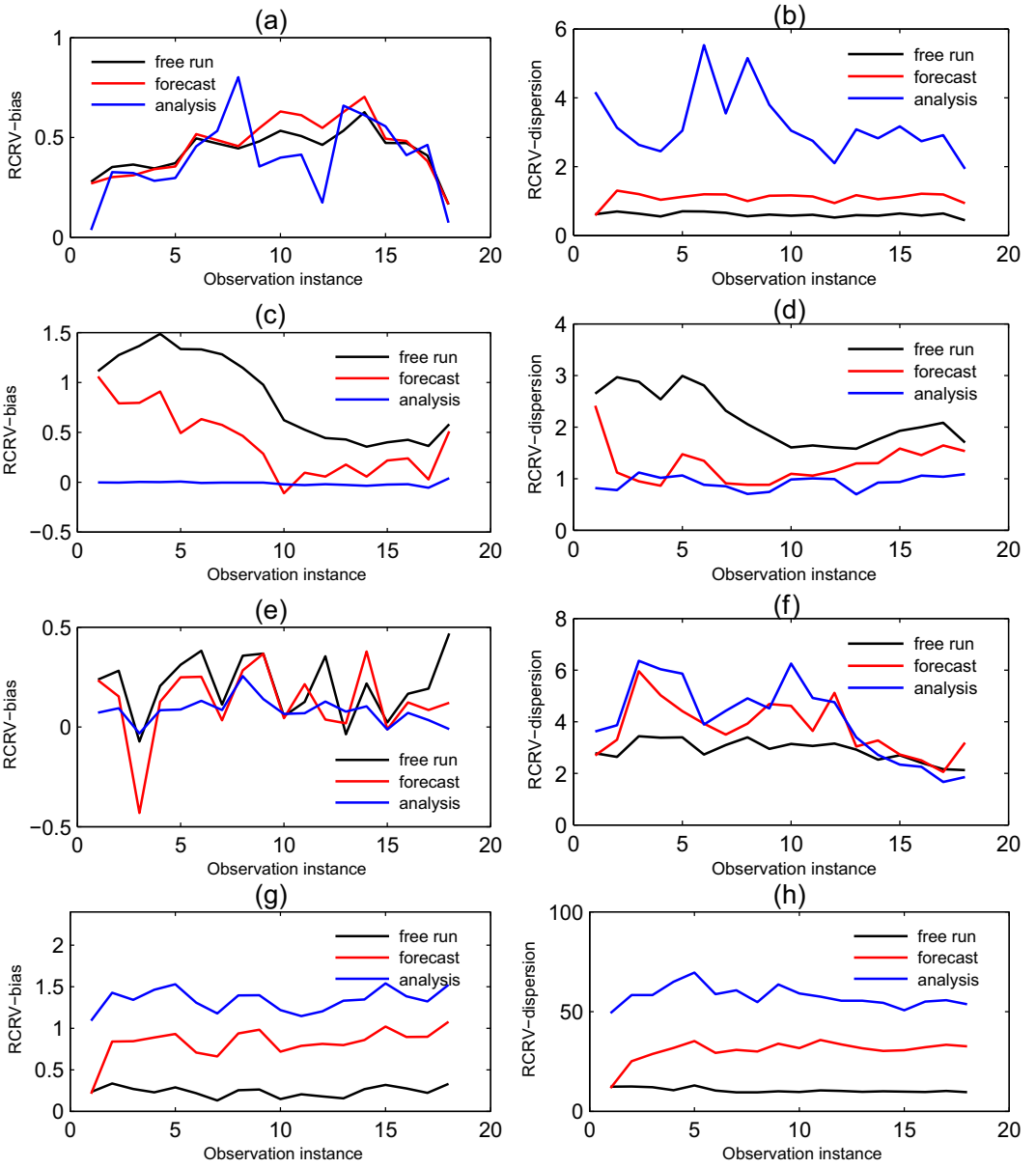

Figure 17. RCRV score of the free run, the forecast and the analysis for (a) (b) SSH (c) (d) SST (e) (f) temperature (g) (h) salinity profile at each analysis step. Independent/semiindependent observations are used for verification for SSH, SST, and salinity profile. For temperature profile, the scores are computed with observations used in the assimilation experiments.

Table 1. Summary of the Improvement by Assimilation of SSH, SST, Temperature, and Salinity Profiles[a]

	Deterministic Metrics		Probabilistic Metrics				
Variables	RMS	Ensemble/Observation Plot	CRPS	CRPS-Reli	CRPS-Reso	RCRV-Bias	RCRV-Disp
SST	~ 0	$y_o \in [\bar{x} \pm \sigma]$, $x \in [y_o \pm \sigma_o]$	~ 0	~ 0	~ 0	~ 0	~ 0
T	$\downarrow$	$y_o \in [\bar{x} \pm \sigma]$, $x \in [y_o \pm \sigma_o]$	$\downarrow$	$\uparrow$	$\downarrow$	$\downarrow$	$\uparrow$
SSH	$\downarrow$	$y_o \notin [\bar{x} \pm \sigma]$, $x \in [y_o \pm \sigma_o]$	$\downarrow$	$\uparrow$	$\downarrow$	$\rightarrow$	$\uparrow$
S	$\downarrow$	$y_o \notin [\bar{x} \pm \sigma]$, $x \notin [y_o \pm \sigma_o]$	$\uparrow$	$\uparrow$	$\downarrow$	$\uparrow$	$\uparrow$

[a]$\sim$ denotes the closeness, $\uparrow$ denotes the increase, $\downarrow$ denotes the decrease, and $\rightarrow$ denotes no change. T and S correspond to temperature and salinity, respectively. x represents ensemble members, $\bar{x}$ represents the ensemble mean, and σ represents the ensemble spread. y_o denotes the observation and σ_o denotes the observation error.

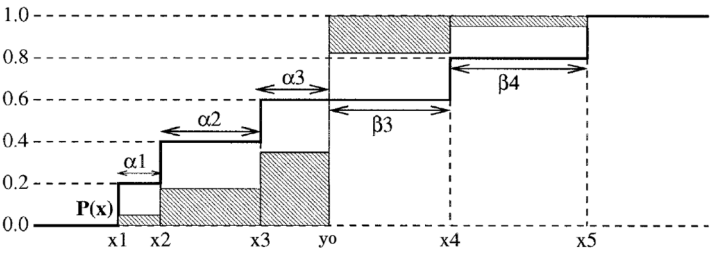

Figure 18. Illustration of the CRPS computation (according to *Hersbach* [2000]). The cumulative distribution for an ensemble of five members ($x_1, \ldots, x_5$) and the verifying observation y_o is shown. The CRPS is represented by the shaded area.

degraded slightly by assimilation, the ensemble system can still be considered reliable. While for salinity profile, the degradation of reliability by assimilation is significant, the ensemble system is no longer reliable.

Different behaviors of both deterministic and probabilistic scores are observed for these four variables. First, for a reliable ensemble system (SST, temperature profile, and SSH), the behaviors of the RMS and the CRPS are similar, both RMS and CRPS provide similar information on the squared distance between the predicted state and the observation. On the other hand, for an unreliable ensemble system (salinity profile), the behavior of the RMS is not similar to that of the CRPS. Moreover, the similarity between CRPS and RMS error for SST and temperature profile is more significant than that of SSH, which indicates that the more reliable the ensemble system is, the more similar the RMS error is to the CRPS. Note also that the difference of the decrease in RCRV-bias with assimilation between SSH and temperature profile causes the difference in the RMS/CRPS similarity between these two variables, which implies that an appropriate RCRV-bias value (smaller than 50% of the system uncertainty) is of particular importance to ensure a reliable system. Further, for SSH, SST, and temperature profile, the resolution component of the CRPS dominates the reliability component, while it is not the case for the salinity profile. The resolution is also relative to the sharpness of the ensemble distribution for reliable ensemble system [*Candille et al*., 2014]. Therefore, in reliable system, the RMS has significance on the resolution of the ensemble system. This is consistent with the interpretation that for a deterministic forecast system, the CRPS is equal to the mean absolute error [*Hersbach*, 2000].

Second, the plots of ensemble mean/spread versus observation are directly connected to CRPS-Reli, $CRPS_{pot}$, and RCRV-bias. The remark that the observation lies outside the ensemble intervals is a sign of lack of reliability, which is consistent with the degradation of reliability according to CRPS-Reli. The decrease of the ensemble spread and the closeness between the ensemble mean and the observation after the analysis can be related to the improvement of the resolution, which is consistent with $CRPS_{pot}$. In a reliable system, the remark that the ensemble always lies entirely within the observation error interval can be considered as a sign of resolution. Moreover, the relative positions of the observations with respect to the ensembles on the coupled ensemble mean/spread versus observation plots gives qualitative information on the sign (positive or negative) and the magnitude of the bias of the ensemble which is consistent with RCRV-bias. RCRV-bias provides more quantitative information on the magnitude of the bias compared to the system uncertainty. Regarding the ensemble spread evaluation, the plots of ensemble mean/spread versus observations alone seem insufficient. RCRV-dispersion provides further precise information on the significance of the over/underdispersion.

Table 2. Values of α_i and β_i Depending on the Position of the Verifying Observation y_o With Respect to the Ensemble Members Ordered From Small to Large[a]

		α_i	β_i
$0 < i < N$	$y_o > x_{i+1}$	$x_{i+1} - x_i$	0
	$x_{i+1} > y_o > x_i$	$y_o - x_i$	$x_{i+1} - y_o$
	$y_o < x_i$	0	$x_{i+1} - x_i$
outlier	$y_o < x_1$	0	$x_1 - y_o$
	$x_N < y_o$	$y_o - x_N$	0

[a]N is the ensemble size.

6. Conclusions

In this paper, assimilation of Jason-1 altimetric data, AVHRR SST data, and ARGO temperature profiles into an eddy permitting primitive equation model of the North Atlantic ocean is performed with the EnKF. To represent the uncertainty present in the model, 60 ensemble members are generated

by adding realistic noise to the atmospheric forcing variables related to the temperature. An IAU scheme is applied instead of intermittent assimilation in order to reduce high-frequency oscillations due to instantaneous model state correction.

The assimilation results are evaluated through both deterministic and probabilistic validations and against independent/semiindependent observations. For deterministic validation, the RMS error of the ensemble mean compared to the observation, as well as the comparison between the coupled ensemble mean/spread and the observation, is analyzed. For probabilistic validation, the ensemble distribution is mainly diagnosed by CRPS according to reliability and resolution. The reliability is further investigated by a RCRV score that decomposes the reliability into bias and dispersion. The deterministic validation and the probabilistic validation are analyzed jointly. The consistency and complementarity between both validations are highlighted.

According to the results, great improvement is obtained for SST. With assimilation, both the random error and the model bias are corrected and an almost perfectly reliable ensemble system has thus been obtained. This benefit is essentially due to a good representation of the model error by the ensemble generated from forcing perturbation and a large number of observations of sufficient quality (small observation error). For SSH, the benefit of the assimilation mainly lies in the improvement of the resolution of the ensemble system and the reduction of the RMS error of the ensemble mean. The decrease of the difference between the observation and the ensemble is not obvious, since a large residual exists in the subpolar area where no Jason-1 observations are available for the assimilation experiments. The ensemble spread seems slightly insufficient to correctly represent the ensemble error, slight degradation of the reliability is thus present. Note that in the assimilation experiments, the SSH increment is not used. After the analysis, the SSH correction mainly comes from the model adjustment with temperature and salinity corrections. Therefore, potential improvement of SSH forecast would be possible if the SSH increment was used in the model state correction. For the temperature profile, the benefit of the assimilation lies in the improvement of the resolution of the ensemble system, the decrease of the RMS error of the ensemble mean and the decrease of the distance between the observation and the ensemble. Note that the ensemble is underdispersive in the free run, with the limited ensemble size and a small quantity of observations, the ensemble underdispersion problem is worse after the analysis. However, for both SSH and temperature profile, even if the ensemble spread seems insufficient to represent the ensemble error after the analysis, if the observation error is taken into account, the anomaly of the ensemble compared to the observation is included within the observation error interval. Therefore, the ensemble distributions of the analysis are considered good enough for these two variables. For salinity profile, the RMS error of the ensemble mean is reduced and the resolution of the ensemble system is improved by the assimilation. However, the reliability of the ensemble system is significantly degraded due to an increase in the distance between the observation and the ensemble and a serious ensemble underdispersion problem. Salinity observations are not directly used in the assimilation experiments, the correction of the salinity depends on the covariance matrix. Because of the relatively small ensemble size, larger error in the covariance matrix can exist. Therefore, the improvement of salinity by assimilation is not obtained as for the other variables.

Regarding the joint analysis of the deterministic validation and the probabilistic validation, the behaviors of the RMS error are very similar to those of the CRPS in a reliable system (given appropriate RCRV-bias). Both can provide useful global information about how far the predicted state is from the observation. In reliable systems, the RMS of the ensemble mean has significance on the resolution and the resolution component of the CRPS dominates the reliability component. For the plots of ensemble mean/spread versus the observation, giving information about the position of the observation with respect to the ensemble mean/spread interval can be further connected to the CRPS-Reli, $CRPS_{pot}$, and RCRV-bias scores. The fact that the observations lie outside the ensemble interval is a sign of lack of reliability. The decrease of the ensemble spread which results in sharper ensemble distribution and the closeness of the ensemble mean and the observation can be considered as a sign of the resolution. In reliable systems, the fact that the ensemble always lies within the observation error interval is a sign of resolution. The position of the ensemble with respect to the observation also provides qualitative information on the sign and the magnitude of the difference between the observation and the ensemble, which is consistent with RCRV-bias. RCRV-bias provides further quantitative information on the bias compared to the system uncertainty and RCRV-dispersion provides precise information on the dispersion of the ensemble.

From previous analyses, it is of particular importance to use probabilistic scores to validate the ensemble distribution properties, especially for unassimilated variables which are probably subject to low reliability and insufficient ensemble dispersion, the deterministic metrics alone seem not sufficient to objectively assess these variables.

Appendix A: Computation of CRPS

For an ensemble system x, including N members $(x_1, x_2, \ldots, x_N)$ ordered from small to large and with equal weight to each member, depending on the position of the verifying observation y_o, $H(x-y_o)$ will be either 0, or 1, or partly 0, partly 1 in the interval $[x_i, x_{i+1}]$, with H the well-known Heaviside function defined in equation (A1).

$$H(x)=\begin{cases}0 & \text{for} \quad x<0\\ 1 & \text{for} \quad x\geq 0\end{cases} \tag{A1}$$

For each of these three possible situations, the CRPS can be written as

$$\text{CRPS}=\sum_{i=0}^{N}\alpha_i p_i^2+\beta_i(1-p_i)^2 \tag{A2}$$

p_i is the fraction i/N. α_i and β_i are illustrated in Figure 18 and their values are shown in Table 2.

For M verifying observation points, each with a weight ω_k ($\omega_k=1/M$ in case of equal weight for all points), the averaged CRPS can be expressed as

$$\overline{\text{CRPS}}=\sum_{i=0}^{N}[\bar{\alpha}_i p_i^2+\bar{\beta}_i(1-p_i)^2] \tag{A3}$$

where

$$\bar{\alpha}_i=\sum_{k=0}^{N}\omega_k\alpha_i^k \tag{A4}$$

$$\bar{\beta}_i=\sum_{k=0}^{N}\omega_k\beta_i^k \tag{A5}$$

The quantities $\bar{\alpha}_i$ and $\bar{\beta}_i$ can be expressed into two quantities $\bar{g}_i$ and $\bar{o}_i$ which have a physical interpretation. $\bar{g}_i$ is the averaged Euclidean distance between consecutive ensemble members for $0<i<N$ and Euclidean distance between the smallest/largest ensemble members and the outliers (when the verifying observation y_o is outside the range of the ensemble) for $i=0$ and $i=N$. $\bar{o}_i$ corresponds to the average frequency that the verifying observation y_o is less than the middle of the bin i (range delineated by consecutive ensemble members x_i and x_{i+1}).

For $0<i<N$,

$$\bar{g}_i=\bar{\alpha}_i+\bar{\beta}_i \tag{A6}$$

$$\bar{o}_i=\frac{\bar{\beta}_i}{\bar{\alpha}_i+\bar{\beta}_i} \tag{A7}$$

For outliers,

$$\bar{o}_0=\sum_{k=0}^{N}\omega_k H(x_1^k-y_o^k) \tag{A8}$$

$$\bar{g}_0=\frac{\bar{\beta}_0}{\bar{o}_0} \tag{A9}$$

and

$$\bar{o}_N = \sum_{k=0}^{N} \omega_k H(x_N^k - y_o^k) \tag{A10}$$

$$\bar{g}_N = \bar{\alpha}_N (1 - \bar{o}_N) \tag{A11}$$

It can be verified that for all $i = 0, 1, \ldots, N$,

$$\bar{\alpha}_i p_i^2 = \bar{g}_i (1 - \bar{o}_i) p_i^2 \tag{A12}$$

$$\bar{\beta}_i (1 - p_i)^2 = \bar{g}_i \bar{o}_i (1 - p_i)^2 \tag{A13}$$

From equation (A3), the averaged CRPS can now be expressed as:

$$\overline{\text{CRPS}} = \sum_{i=0}^{N} \overline{g_i} [(1 - \overline{o_i}) p_i^2 + \overline{o_i} (1 - p_i)^2] \tag{A14}$$

Its decompositions can be expressed as:

$$\text{CRPS-Reli} = \sum_{i=0}^{N} \overline{g_i} (\overline{o_i} - p_i)^2 \tag{A15}$$

$$\text{CRPS}_{\text{pot}} = \sum_{i=0}^{N} \overline{g_i o_i} (1 - \overline{o_i}) \tag{A16}$$

$$\text{CRPS-Uncert} = \sum_{k=1}^{M-1} q_k (1 - q_k)(x^{k+1} - x^k) \tag{A17}$$

where q denotes the cumulative probability based on the verification observations.

$$q_k = q_{k-1} + \omega_k, \; q_0 = 0 \tag{A18}$$

$\omega_k = 1/M$ assuming equal weight for all verification observations.

Acknowledgments

This work is funded by the European Sangoma project (FP7-SPACE-2011, grant 283580). The ENVISAT and Jason-1 altimetric data are downloaded from the AVISO site (http://www.aviso.altimetry.fr/en/data/data-access/ftp.html) (data set name: dt_ref_global_en_sla_2005.tar, dt_ref_global_j1_sla_2005.tar, need to fill a registration form). The ARGO temperature and salinity profiles are downloaded from Met Office EN3 (http://www.metoffice.gov.uk/hadobs/en3/data/EN3_v2a/download_EN3_v2a.html) (Data set name: EN3_v2a_Profiles_2005.tar) and Coriolis sites (http://www.coriolis.eu.org/Data-Services-Products/View-Download/Data-selection) (select the ARGO profiles covering the model domain (20°S-80°N, 98°W-23°E) between the 1 January and the 31 December 2005, then download them). The AVHRR SST data and the Mercator reanalysis are provided by Mercator Ocean (contact: laurent.parent@mercator-ocean.fr). The horizontal velocity field data in the Gulf Stream region are obtained from the DEOS site (http://rads.tudelft.nl/gulfstream/gif) (Data set name: gulf_051116_vel.gif). We want to thank the reviewers for their valuable comments and suggestions.

References

Adcroft, A., C. Hill, and J. Marshall (1997), Representation of topography by shaved cells in a height coordinate ocean model, *Mon. Weather Rev.*, *125*(9), 2293–2315.

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

Arakawa, A., and V. Lamb (1981), A potential enstrophy and energy conserving scheme for the shallow water equations, *Mon. Weather Rev.*, *109*, 18–36.

Barnier, B., et al. (2006), Impact of partial steps and momentum advection schemes in a global ocean circulation model at eddy-permitting resolution, *Ocean Dyn.*, *56*, 543–567.

Barth, A., A. Alvera-Azcárate, J. Beckers, M. Rixen, and L. Vandenbulcke (2007a), Multigrid state vector for data assimilation in a two-way nested model of the Ligurian sea, *J. Mar. Syst.*, *65*(1–4), 41–59.

Barth, A., J.-M. Beckers, A. Alvera-Azcárate, and R. H. Weisberg (2007b), Filtering inertia-gravity waves from the initial conditions of the linear shallow water equations, *Ocean Modell.*, *19*, 204–218.

Barth, A., A. Alvera-Azcárate, and R. Weisberg (2008), Assimilation of high-frequency radar currents in a nested model of the west Florida shelf, *J. Geophys. Res.*, *113*, C08033, doi:10.1029/2007JC004585.

Barth, A., A. Alvera-Azcárate, J. Beckers, J. Staneva E. V. Stanev, and J. Schulz-Stellenfleth (2011), Correcting surface winds by assimilating high-frequency radar surface currents in the German Bight, *Ocean Dyn.*, *61*, 599–610.

Blanke, B., and P. Delecluse (1993), Variability of the tropical Atlantic ocean simulated by a general circulation model with two different mixed layer physics, *J. Phys. Oceanogr.*, *23*, 1363–1388.

Brier, G. (1950), Verification of forecasts expressed in terms of probabilities, *Mon. Weather Rev.*, *78*, 1–3.

Buehner, M. (2004), Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting, *Mon. Weather Rev.*, *131*(607), 1013–1043.

Candille, G., C. Côté, P. Houtekamer, and G. Pellerin (2006), Verification of an ensemble prediction system against observations, *Mon. Weather Rev.*, *135*, 2688–2699.

Candille, G., J. Brankart, and P. Brasseur (2014), Assessment of an ensemble system that assimilates Jason-1/Envisat altimeter data in a probabilistic model of the North Atlantic ocean circulation, *Ocean Sci. Discuss.*, *11*(6), 2647–2690.

Casati, B., L. J. Wilson, D. B. Stephenson, P. Nurmi, A. Ghelli, M. Pocernich, U. Damrath, E. E. Ebert, B. G. Brown, and S. Masonh (2008), Forecast verification: Current status and future directions, *Meteorol. Appl.*, *15*(1), 3–18.

Casey, K. S., T. B. Brandon, P. Cornillon, and R. Evans (2010), The past, present, and future of the AVHRR pathfinder SST program, in *Oceanography from Space*, edited by V. Barale, J. Gower, and L. Alberotanza, pp. 273–287, Springer, N. Y.

Davis, R. (1991), Observing the general circulation with floats. *Deep Sea Res., Part A*, *38*, s531–s571.

Dee, D. P., et al. (2011), The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, *Q. J. R. Meteorol. Soc.*, *137*(656), 553–597.
Evensen, G. (2003), The ensemble Kalman filter: Theoretical formulation and practical implementation, *Ocean Dyn.*, *53*, 343–367.
Evensen, G. (2004), Sampling strategies and square root analysis schemes for the EnKF, *Ocean Dyn.*, *54*, 539–560.
Ferry, N., L. Parent, G. Garric, M. Drevillon, C. Desportes, C. Bricaud, and F. Hernandez (2012), Scientific Validation Report (ScVR) for Reprocessed Analysis and Reanalysis, technical report MYO-WP04-ScCV-rea-MERCATOR-v1.0, Mecartor Océan, Toulouse, France.
Fu, L., E. J. Christensen, C. A. Yamarone Jr., M. Lefebvre, Y. Ménard, M. Dorrer, and P. Escudier (1994), Topex/poseidon mission overview, *J. Geophys. Res.*, *99*(C12), 24,369–24,381.
Hamill, T. (2000), Interpretation of rank histograms for verifying ensemble forecasts, *Mon. Weather Rev.*, *129*, 550–560.
Hamill, T., and J. Whitaker (2001), Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter, *Mon. Weather Rev.*, *129*, 2776–2790.
Hartmann, H., T. Pagano, S. Sorooshiam, and R. Bales (2002), Confidence builder: Evaluating seasonal climate forecasts from user perspectives, *Bull. Am. Meteorol. Soc.*, *84*, 683–698.
Hersbach, H. (2000), Decomposition of the continuous ranked probability score for ensemble prediction system, *Weather Forecast.*, *15*(5), 559–570.
Houtekamer, P., and H. Mitchell (2001), A sequential ensemble Kalman filter for atmospheric data assimilation, *Mon. Weather Rev.*, *129*, 123–137.
Hancock, D. W., III, G. S. Hayne, R. L. Brooks, and D. W. Lockwood (2001), Geosat Follow-On (GFO) Altimeter Document Series, volume 1, GFO altimeter engineering assessment report from launch to acceptance 10 February 1998 to 29 November 2000, technical report NASA/TM-2001-209984/VER1/vol.1, NASA, USA.
Kantha, L. H., and C. A. Clayson (2000), *Numerical Models of Oceans and Oceanic Processes*, 1st ed., Academic, Waltham, Mass.
Kaplan, A., M. Cane, Y. Kushnir, A. Clement, M. Blumenthal, and B. Rajagopalan (1998), Analyses of global sea surface temperature 18561991, *J. Geophys. Res.*, *103*(C9), 18,567–18,589.
Levitus, S., T. Boyer, M. Conkright, T. O. Brien, J. Antonov, C. Stephens, L. Stathoplos, D. Johnson, and R. Gelfeld (1998), Volume 1: Introduction, *NOAA Atlas NESDIS 18*, *World Ocean Database 1998*, U.S. Gov. Print. Off., Washington, D. C.
Marmain, J., A. Molcard, P. Forget, A. Barth, and Y. Ourmières (2014), Assimilation of HF radar surface currents to optimize forcing in the northwestern Mediterranean sea, *Nonlinear Process. Geophys.*, *21*, 659–675.
Ménarda, Y., L. Fub, P. Escudiera, F. Parisota, J. Perbosa, P. Vincenta, S. Desaib, B. Hainesb, and G. Kunstmannb (2003), The Jason-1 Mission special issue: Jason-1 calibration/validation, *Mar. Geod.*, *26*(3–4), 131–146.
Murphy, A. (1973), A new vector partition of the probability score, *J. Appl. Meteorol. Climatol.*, *12*, 595–600.
Oka, E., and K. Ando (2004), Stability of temperature and conductivity sensors of Argo profiling floats, *J. Oceanogr.*, *60*, 253–258.
Resti, A., J. Benveniste, M. Roca, G. Levrini, and J. Johannessen (1999), The Envisat radar altimeter system (RA-2), *ESA Bull.*, *98*, 1–8.
Skachko, S., J. Brankart, F. Castruccio, P. Brasseur, and J. Verron (2009), Improved turbulent air-sea flux bulk parameters for controlling the response of the ocean mixed layer: A sequential data assimilation approach, *J. Atmos. Oceanic Technol.*, *26*, 538–555.
Stanski, H., L. Wilson, and W. Burrows (1989), Survey of common verification methods in meteorology, in *Atmospheric Environment Service*, Forecast Res. Div., WMO World Weather Watch Technical Report No. 8 TD No. 358, World Meteorological Organisation, Geneva.
Talagrand, O., R. Vautard, and B. Strauss (1999), Evaluation of probabilistic prediction systems, in Proceedings of *Workshop on Predictability*, pp. 1–25, European Centre for Medium-Range Weather Forecasts, U. K.
Testut, C., P. Brasseur, J. Brankart, and J. Verron (2003), Assimilation of sea-surface temperature and altimetric observations during 1992–1993 into an eddy permitting primitive equation model of the North Atlantic Ocean, *J. Mar. Syst.*, *40–41*, 291–316.
Toth, Z., O. Talagrand, G. Candille, and Y. Zhu (2003), Probability and ensemble forecasts, in *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, edited by I. Jolliffe and D. Stephenson, John Wiley & Sons, Hoboken, N. J.
Vandenbulcke, L., A. Barth, M. Rixen, A. Alvera-Azcárate, Z. B. Bouallegu, and J. Beckers (2006), Study of the combined effects of data assimilation and grid nesting in ocean models: Application to the Gulf of Lions, *Ocean Sci.*, *2*(2), 213–222.
Vandenbulcke, L., A. Capet, and J. Beckers (2010), Onboard implementation of the GHER model for the black sea, with SST and CTD data assimilation, *J. Oper. Oceanogr.*, *3*(2), 47–54.
Wilks, D. (1995), *Statistical Methods in the Atmospheric Sciences: An Introduction*, 3rd ed., Academic, Waltham, Mass.
Yan, Y., A. Barth, and J. Beckers (2014), Comparison of different assimilation schemes in a sequential Kalman filter assimilation system, *Ocean Modell.*, *73*, 123–137.