Nonlin. Processes Geophys., 20, 621–641, 2013
www.nonlin-processes-geophys.net/20/621/2013/
doi:10.5194/npg-20-621-2013

Nonlinear Processes in Geophysics
Open Access

The impact of initial spread calibration on the RELO ensemble and its application to Lagrangian dynamics

M. Wei[1], G. Jacobs[1], C. Rowley[1], C. N. Barron[1], P. Hogan[1], P. Spence[2], O. M. Smedstad[2], P. Martin[1], P. Muscarella[3], and E. Coelho[4]

[1]Naval Research Laboratory, Stennis Space Center, MS, USA
[2]Qinetiq-North America, Stennis Space Center, MS, USA
[3]ASEE Postdoctoral Program, Naval Research Laboratory, Stennis Space Center, MS, USA
[4]University of New Orleans at NRL, Stennis Space Center, MS, USA

Correspondence to: M. Wei (mozheng.wei@nrlssc.navy.mil)

Received: 15 April 2013 – Revised: 23 July 2013 – Accepted: 24 July 2013 – Published: 11 September 2013

Abstract. A number of real-time ocean model forecasts were carried out successfully at Naval Research Laboratory (NRL) to provide modeling support and numerical guidance to the CARTHE GLAD at-sea experiment during summer 2012. Two RELO ensembles and three single models using NCOM and HYCOM with different resolutions were carried out. A calibrated ensemble system with enhanced spread and reliability was developed to better support this experiment. The calibrated ensemble is found to outperform the un-calibrated ensemble in forecasting accuracy, skill, and reliability for all the variables and observation spaces evaluated. The metrics used in this paper include RMS error, anomaly correlation, PECA, Brier score, spread reliability, and Talagrand rank histogram. It is also found that even the un-calibrated ensemble outperforms the single forecast from the model with the same resolution.

The advantages of the ensembles are further extended to the Lagrangian framework. In contrast to a single model forecast, the RELO ensemble provides not only the most likely Lagrangian trajectory for a particle in the ocean, but also an uncertainty estimate that directly reflects the complicated ocean dynamics, which is valuable for decision makers. The examples show that the calibrated ensemble with more reliability can capture trajectories in different, even opposite, directions, which would be missed by the un-calibrated ensemble. The ensembles are applied to compute the repelling and attracting Lagrangian coherent structures (LCSs), and the uncertainties of the LCSs, which are hard to obtain from a single model forecast, are estimated. It is found that the spatial scales of the LCSs depend on the model resolution. The model with the highest resolution produces the finest, small-scale, LCS structures, while the model with lowest resolution generates only large-scale LCSs. The repelling and attracting LCSs are found to intersect at many locations and create complex mesoscale eddies. The fluid particles and drifters in the middle of these tangles are subject to attraction and repulsion simultaneously from these two kinds of LCSs. As a result, the movements of particles near the Deepwater Horizon (DWH) location are severely limited. This is also confirmed by the Lagrangian trajectories predicted by the ensembles.

1 Introduction

The Grand Lagrangian Deployment (GLAD) was an at-sea experiment that was conducted in the northern Gulf of Mexico (GOM) from 17 July to 3 August 2012 by the Consortium for Advanced Research on Transport of Hydrocarbon in the Environment (CARTHE, http://www.carthe.org/). CARTHE is one of the consortia supported by the Gulf of Mexico Research Initiative (GoMRI, http://gulfresearchinitiative.org/) and it comprises 26 principal investigators from 12 universities and research institutions, including the Naval Research Laboratory (NRL) at Stennis Space Center, MS. These universities and research institutions are distributed across four Gulf of Mexico states and four other states.

As the modeling team for CARTHE at NRL, our focus was on the numerical modeling, data assimilation (DA), and forecasting to support and provide numerical guidance to the GLAD experiment. To support this mission, we have successfully run two RELO (Relocatable Circulation Prediction System) ensembles, each with 32 members, and three single-model deterministic forecasts using the Navy

Coastal Ocean Model (NCOM; Martin, 2000; Barron et al., 2006) and the Hybrid Coordinate Ocean Model (HYCOM, http://www.hycom.org, Bleck, 2002; Chassignet et al., 2003; Halliwell, 2004) with different resolutions. All of these five ocean forecast systems were run in real-time, assimilating routine in situ and satellite observations processed at the US Naval Oceanographic Office (NAVOCEANO), located at Stennis Space Center, MS. To prepare for these important numerical and in situ experiments, all the forecast experiments started on 16 May 2012 to initialize the ocean models and ensembles and test the support software needed to distribute the real-time forecasts. The implementation and operation of these forecast systems were conducted smoothly without delays in delivery. These forecast products provided real-time guidance to the GLAD drifter deployment.

It is known that the first generation of ensemble prediction/forecast systems (EPS or EFS) was implemented at the major meteorological centers about 20 yr ago, and the details have been described in numerous publications, e.g., Toth and Kalnay (1993), Houtekamer et al. (1996), Molteni et al. (1996), Descamps and Talagrand (2007), and Leutbecher and Palmer (2008). A number of improvements have been made regularly over the past few years in areas such as initial perturbation generation techniques, methods for representing model-related uncertainties, and computing efficiency. The performance of different ensemble methods or systems has been studied and compared in the literature, e.g., Hamill et al. (2000), Wei and Toth (2003), Buizza et al. (2005), Bowler (2006), Wei et al. (2006), Descamps and Talagrand (2007), and Magnusson et al. (2009). The basic properties, advantages, and disadvantages of the different ensemble methods are summarized in Tables 1 and 2 of Wei et al. (2008). Some state of the art statistical post-processing techniques (calibration) such as Bayesian model averaging (BMA) can be found in Raftery et al. (2005).

At NRL, the RELO ensemble forecast system has been developed to provide a capability for a rapidly relocatable ocean ensemble forecast and data assimilation system for use in operational forecast support for the U.S. Navy's missions (Rowley, 2008, 2010; Rowley et al., 2012; Wei et al., 2013, hereafter referred to as W13). A schematic showing the configuration of the RELO system with 32 ensemble members as used in this paper is presented in Fig. 1 of W13. The forecast component of the RELO ensemble system is NCOM, (Martin, 2000; Barron et al., 2006). NCOM is a primitive-equation ocean model developed at NRL for local, regional, and global forecasting of temperature, salinity, sound speed, and currents. The data-assimilation component is the Navy Coupled Ocean Data Assimilation System (NCODA; Cummings, 2005), which is based on a 3D-Var formulation. Both NCOM and NCODA are used operationally at two US Navy operational centers, namely the Fleet Numerical Meteorology and Oceanography Center (FNMOC), located in Monterey, CA, and NAVOCEANO.

In the RELO ensemble, the analysis error estimated from NCODA is used to generate the initial perturbations by using the Ensemble Transform (ET) method. However, it was found that the current method used in NCODA underestimates the analysis error (W13). As a result, the RELO ensemble is under dispersive and the spread is smaller than the ensemble mean error. In theory, an ideal EPS should have a spread that has amplitude comparable to the ensemble mean error and grows at a similar rate (Hamill et al., 2000; Buizza et al., 2005; Wei et al., 2006, 2008; Descamps and Talagrand, 2007; Leutbecher and Palmer, 2008). Although estimating analysis error in a 3D-Var-based DA system such as NCODA is challenging, the Lanczos method, with proper calibration, can be used to produce reasonably good analysis error variance with extra computational cost (Wei et al., 2012). Another simpler, poor-man's method is to use multi-analysis data from different DA systems or operational centers as demonstrated in Wei et al. (2010). Work on improving the analysis error estimate in NCODA is continuing at NRL.

Accounting for model-related uncertainties in the RELO ensemble is another direction that can be taken to enhance the ensemble spread. As an initial step to achieve this, W13 proposed and examined three different schemes for perturbing the horizontal and vertical mixing parameters. The results show that a scheme perturbing both the horizontal and vertical mixing parameters based on a Gaussian distribution produces the largest spread increment. The ensemble based on this scheme will be used in this paper. However, the RELO ensemble is still under dispersive, even with this parameter perturbation scheme. Similar cases have been found in atmospheric ensemble systems, e.g., Reynolds et al. (2011). To further improve the RELO ensemble in such a short period of time for the CARTHE GLAD experiment, we proposed a calibration to enhance the initial spread inside the ET based on previous estimates of the ensemble spread and forecast error. The superior performance of this calibrated ensemble will be explored and demonstrated. Some other efficient methods, such as stochastic forcing (Lermusiaux, 2006), which can potentially enhance the spread, will be pursued in the next step.

In this study, we will compare the two RELO ensembles with three, single-model, deterministic forecasts with different resolutions, which will showcase the advantages of ensembles over single model forecasts. The impacts and benefits of the proposed initial spread calibration will be examined and demonstrated in terms of the most commonly used verification metrics. These include RMS error, anomaly correlation, PECA, Brier score, spread-reliability, and Talagrand rank histogram. In addition, we will extend the RELO ensembles to Lagrangian dynamics, including particle trajectory prediction and the Lagrangian coherent structure (LCS). The advantages of using the ensemble, especially the calibrated, more reliable ensemble, will be demonstrated in all of these cases. The details of the

complicated tangles formed by the repelling and attracting LCSs over the region near the location of DWH and their impacts on the movements of particles are revealed and compared with the trajectories predicted by the ensembles. The DWH (which is located about 60 km off the Louisiana coast at 88.39° W, 26.74° N) is the location of the largest oil spill incident in US history, in which over 4.9 million barrels of oil were released into the GOM between 20 April and 15 July 2010.

Another goal of this paper is to describe these numerical forecast systems and their corresponding products, including the RELO ensembles and, particularly, the calibrated ensemble that was developed for this mission. We believe that the material presented in this study will be valuable for scientists both inside and outside the CARTHE project, especially after the CARTHE GLAD data are released to the public in the future.

Section 2 provides very brief descriptions of the ET formulation for initial perturbations, the time-deformation technique to generate surface forcing perturbations from an atmospheric model for the RELO ensembles, the methodology for perturbing the mixing parameters, the configurations for the RELO ensembles, NCOM and HYCOM, and the experimental set up. The major results are presented in Sect. 3. Also shown in Sect. 3 are the results from applications of the ensembles to Lagrangian dynamics and particle trajectory prediction. A discussion and conclusions are presented in Sect. 4.

2 Methodology, ocean ensembles and models, and experimental setup

2.1 Initial and surface forcing perturbations

The NRL RELO ensemble prediction system uses the ET method that transfers forecast perturbations from the previous cycle into new perturbations using the estimated initial analysis error variance. The same initial analysis error variance will be used in the following rescaling process. The method and its properties in general are described in Wei et al. (2005, 2008) and McLay et al. (2007).

In the RELO ensemble, the analysis fields are generated by the NCODA DA system and the estimate of the analysis error variance is also derived from NCODA. The ET method has the advantage that the ensemble perturbations span a subspace that has a maximum number of degrees of freedom. The orthogonality of the initial perturbations will increase as the number of ensemble members increases. If the number of ensemble members approaches infinity, the transformed perturbations will be orthogonal under the inverse of the analysis error variance norm. In addition to the flow-dependent spatial structure, the covariance constructed from the initial perturbations is approximately consistent with the analysis covariance from the DA if the number of ensemble members is large. More details about the initial perturbations based on ET in the US Navy's RELO ensemble can be found in W13.

The surface forcing perturbations are produced from real-time, meteorological, forecast fields obtained from FNMOC, including wind stress, surface pressure, shortwave and long-wave radiation, air temperature, and specific humidity. FNMOC produces operational forecasts using the Navy Operational Global Atmospheric Prediction System (NOGAPS) for global and the Coupled Ocean Atmosphere Mesoscale Prediction System (COAMPS) for regional forecasts. Throughout our experiments with both the RELO ensemble and single forecasts, we use COAMPS atmospheric data fields, which are available at 3 h intervals and are updated using a 12 h analysis-forecast cycle.

For the RELO ensemble, perturbed surface forcing fields for different ensemble members are generated with a random shifting technique from the single-model prepared forcing. At every cycle, 32 completely independent random fields are generated every 24 h with a specified de-correlation length. For each ensemble member, forcing is prepared at the same 3 h interval by linear interpolation of the forcing, but with the values computed at randomly shifted times. The time shifts are defined using a set of independent random fields generated every 24 h with a defined spatial de-correlation, so that any interpolated field is not correlated with any other interpolated field 24 h away, and the atmospheric forcing for each ensemble member will be independent. More detailed mathematical formulae are given in W13.

2.2 Parameter perturbations

Our previous studies showed that the technique used in the NCODA DA system underestimates the initial analysis error. Consequently, the initial perturbations generated by the ET method in the RELO ensemble are relatively small. At the same time, the initial perturbations do not grow fast enough to describe the forecast errors due to the lack of model uncertainty representations in the ensemble. As an initial step to address this issue, the parameter perturbations were introduced to the RELO ensemble in W13. The impacts on the ensemble spread, reliability, accuracy, and forecasting skill were investigated in that study. The two key parameters that play critical roles in describing the horizontal and vertical mixing in NCOM (Martin, 2000; Barron et al., 2006) are perturbed, namely the scaling parameter (smag) for the Smagorinsky horizontal mixing formulation (Smagorinsky, 1963), and the turbulent kinetic energy dissipation coefficient (b1_myl2) for the Mellor–Yamada Level 2 (MYL2, Mellor and Yamada, 1974; Mellor and Durbin, 1975) vertical mixing scheme. Note that there are options for other horizontal and vertical mixing schemes in NCOM. The advantages and disadvantages of using the Smagorinsky and MYL2 schemes in comparison with the other choices are discussed in Martin (2000). In both the RELO ensemble and NCOM single-model runs, the default values are $\mathrm{smag} = 0.1$ and $\mathrm{b1\_myl2} = 15.0$.

Three different parameter-perturbation schemes based on different statistical distributions were tested and compared in W13. In this study, we use the scheme that produced the largest spread increment for the RELO ensemble, i.e., the scheme in which these two parameters in the horizontal and vertical mixing turbulence parameterization, smag and b1_myl2, are perturbed using a Gaussian distribution. The mean and standard deviation of smag are chosen as 0.125 and 0.01875, while for b1_myl2, the values are 17.5 and 0.625, respectively. Under these distributions, values of these two randomly generated parameters are expected to fall within reasonable ranges and allow NCOM to run smoothly. The RELO ensemble based on this choice of parameter perturbation is denoted by gom32r (or r in the figures). Again, more details can be found in W13.

2.3 NCOM, HYCOM, RELO ensemble with calibration and experimental design

The introduction of parameter perturbations in gom32r to account for model uncertainties from the mixing parameterizations improved the ensemble spread to a certain extent. However, the ensemble spread in gom32r is still smaller than the ensemble mean error as shown in W13. To prepare for the GLAD at-sea experiment and provide the best possible real-time ensemble forecasts and uncertainty estimates for the scientists in this experiment, we needed to address this issue efficiently and quickly. Any efforts that needed a long period of development could not be considered. Hence, we took an ad hoc approach to calibrate the initial spread magnitude based on the difference between the RMS error of the ensemble mean and the spread from the data accumulated during past experiments. This kind of ad hoc approach has proven to be effective in operational ensemble systems at major NWP centers (Houtekamer et al., 1996; Buizza et al., 2005; Bowler et al., 2009; Wei et al., 2008). It is necessary to note that this is a simple calibration for initial spread only, which is different from more sophisticated calibration methods for postprocessing, such as the BMA method (Raftery et al., 2005). Our initial spread calibration is done before ensemble forecasts start, while the BMA is applied to the ensemble products with clever statistical techniques after the forecasts. Hence, in this study, we ran another, real-time, RELO ensemble with a calibrated ensemble spread, in which the magnitudes of the initial perturbations generated in the ET in gom32r were increased by 50 %. One of the advantages of this ad hoc calibration is that the spatial structure of the initial perturbations is not altered. This system is denoted as gom32q in this study or q in the figures. In Sect. 3, we will show the results from both gom32r and gom32q using various verification metrics. The improved performance in terms of forecast accuracy, skill, and reliability due to this calibration process will be demonstrated in comparison with gom32r.

W13 is mainly concerned about the impacts from various parameter perturbation schemes with one model and one resolution, no initial perturbation calibration is involved and the experiments were carried for 15 April to 25 July 2010. However, some common measures and techniques have been used both in this study and W13, such as spread-reliability diagram, Talagrand histogram, RMS error and anomaly correlation. In addition to the two RELO ensembles (i.e., gom32r and gom32q), each with 32 ensemble members, three single-model forecasts were also carried out using NCOM at both 3-km and 1-km resolution, and HYCOM at 4-km resolution. For validation, we chose the forecast series from 1 June to 17 September 2012, a total of 109 days, although the real-time experiments started on 16 May 2012. The forecast length for all the model runs was 72 h, with output every 6 h.

Both ensembles r and q and the 3-km NCOM single forecast have a horizontal domain covering the GOM from 98 to 79° W and 18 to 31° N with a grid spacing of 3 km × 3 km. The grid dimensions are 640 and 481 in the longitude and latitude directions, respectively. This single NCOM forecast is denoted as ncom3km (or 3k in the figures). The number of vertical levels is 49, with 34 bottom-following sigma layers in the upper ocean and z levels from the bottom of level 34 to the bottom of level 49 at a depth of 5500 m. The advantages of this kind of hybrid σ–z coordinate were discussed in Martin (2000) and Barron et al. (2006).

Another single forecast with NCOM has a horizontal resolution of 1 km × 1 km (denoted by ncom1km, or 1k in the figures) covering the GOM from 97.95 to 80.25° W and 18.05 to 30.79° N, with grid dimensions of 1800 and 1420 in the longitude and latitude directions, respectively. The vertical coordinate and resolution are the same as ncom3km. The only difference between ncom1km and ncom3km is in the horizontal resolution. Tidal forcing (the barotropic tidal height and transports at the open boundaries and the tidal potential in the interior) is turned on for gom32r, gom32q, ncom3km, and ncom1km.

The last single forecast is produced with the Gulf of Mexico HYbrid Coordinate Ocean Model (HYCOM). This model is on a Mercator projection covering the region from 18 to 32° N, and from 98 to 76.4° W. The horizontal grid resolution is 1/25°, ∼ 4 km resolution (indicated by hycom4km, or 4k in the figures). The model employs 20 hybrid vertical coordinate surfaces. Vertical coordinates can be isopycnals (density tracking), often best in the deep stratified ocean, levels of equal pressure (nearly fixed depths), best used in the mixed layer and unstratified ocean, and sigma levels (terrain following), often the best choice in shallow water. HYCOM combines all three approaches by choosing the optimal distribution at every time step. The model makes a dynamically smooth transition between coordinate types by using the layered continuity equation. The model is nested in a climatology generated from a multi-year, climatologically forced, 0.08° HYCOM Atlantic Ocean

simulation. There is no tidal forcing turned on during this run. All the ensembles and single forecast models together with their configurations are summarized in Table 1.

3 Results from the RELO ensembles, NCOM and HYCOM

3.1 Impact of the calibration on ensemble spread

The ensemble mean, which provides forecast, and the ensemble spread, which provides the forecast uncertainties, are the very basic attributes of an ensemble prediction system. For a carefully designed, reliable ensemble, the ensemble mean generally outperforms a single deterministic forecast in terms of the root mean square (RMS) error and the absolute error. The ensemble spread is closely related to the range, reliability, and sharpness or resolution of the EPS (Wei and Toth, 2003; Wei et al., 2006, 2008; W13). One of the contributions of this project is to introduce a calibrated ensemble q that, we hope, outperforms the original ensemble r. Before the RELO ensemble is compared with the single forecasts, we concentrate on the comparisons between the ensemble r and the calibrated ensemble q in this section.

The GLAD at-sea experiment started on 17 July 2012 and lasted until 3 August 2012. We choose to show a snapshot of the horizontal spread distributions at 00:00 UTC on 20 July 2012 to depict the immediate enhancement of the initial ensemble spread from the calibration in Fig. 1. The spreads of the main model variables T (temperature), S (salinity), u (velocity component along longitudinal direction), and v (velocity component along latitudinal direction) on the surface are shown (from top to bottom) for ensembles r (left panel) and q (right panel). As expected, the calibrated ensemble spreads q (right panel) are larger than the spreads from ensemble r (left panel) for all the variables. The largest temperature spreads are located near the Yucatan Current, the Florida Current, and the Loop Current (LC) eddy, reflecting a larger ocean state variability near these regions. Relatively high uncertainty is also found at the surface south of the Mississippi River delta, which is near the DWH site. As expected, there is a large uncertainty in the surface salinity near the Mississippi and Atchafalaya river outflows. The low salinity values near the coasts of Louisiana and Mississippi are due to the large fresh water inputs. The largest surface salinity variations are located in this area of mixing. For the surface velocity, large variations occur within 200 km of the Louisiana coast and in regions near the LC and Yucatan Current.

The spread comparisons between ensembles r and q shown in Fig. 1 are just snapshots of the two ensembles at 00:00 UTC 20 July 2012, from one particular vertical level. In order to obtain more statistically meaningful comparisons, we need to compute various verification metrics over a much larger number of samples. To compute the values of metrics related to accuracy and reliability, we have interpolated all the ensemble forecasts to the observation locations. The routine in situ and remote sensing observations from the NAVOCEANO operational NCODA DA system are used as truth to compare against. In this study, our evaluations are carried out for different observation spaces, including the full observation space, near the surface (upper 1 m), and in the ocean interior over a range from 0–100 m. The numbers and locations of observations vary each day, and the locations are most likely not on model grids. On a particular day during the experimental period, the numbers of temperature observations are about 12750 (full space), 8098 (0–100 m) and 6505 (surface). While for salinity, the observation numbers are about 7200 (full space) and 2500 (0–100 m). The reasons for evaluating the forecasts over different observation spaces include the fact that these are dynamically distinct domains. The surface is normally dominated by air–sea interactions and highly variable wind-driven currents. The interior is generally controlled by mesoscale dynamics and internal mixing processes. Secondly, the density of observations is different for the different domains. For instance, the number of observations near the surface is much larger than that in the interior. Therefore an average over the entire observation space may be skewed toward the near-surface.

Figure 2 shows the ensemble spreads at 00:00 UTC on each day during our experimental period from 1 June to 17 September 2012. To have the best statistical meaning, all the spread values are averaged over the layer from 0–100 m where most of the observations are located. The average values over the whole period are indicated in the figures. It is clear that the spreads for the calibrated ensemble q are consistently larger than those of ensemble r for both temperature and salinity at 24, 48, and 72 h forecast lead times. The results in this figure also indicate large spread variability on different dates during this period for both variables, particularly salinity. Spread values averaged over the full observation space (not shown), are also consistent with these results.

One may notice higher values of temperature spreads for forecast lead times of 24, 48 and 72 h at 00 UTC of 2 September 2012, 1 September 2012, and 31 August 2012. There are two reasons. Hurricane Isaac entered the GOM from around 29 August 2012 with changing intensities, and lasted until 3 September 2012 before moving northeastward away from the Gulf coast. It had caused a lot of flooding around the coast states including Louisiana and Mississippi where NRL is located near the coast. The larger spread reflects the larger variability of temperature in the water during the period of Hurricane Isaac. The long-lasting flooding also caused power outage and loss of communication. There were very few observation data from operational center around 3 September 2012 due to the lost transmissions. Since all these verifications are

Table 1. Experimental setup and model description.

	gom32r (r)	gom32q (q)	ncom3km (3k)	ncom1km (1km)	hycom4km (4km)
Model	NCOM	NCOM	NCOM	NCOM	HYCOM
Resolution	3km	3km	3km	1km	4km
	49 hybrid $\sigma - z$ levels	49 hybrid $\sigma - z$ levels	49 hybrid $\sigma - z$ levels	49 hybrid $\sigma - z$ levels	20 hybrid levels
Tidal forcing	on	on	on	on	off
Number of members	32	32	1	1	1
Perturbations	Analysis: NCODA 3D-Var. Initial Perts: ET, Surface forcing perturbs from COAMPS atmospheric fields based on time-deformation technique. Model error: perturbing vertical and horizontal turbulence mixing parameters with Gaussian distribution.	gom32r + Initial pert calibration	N/A	N/A	N/A

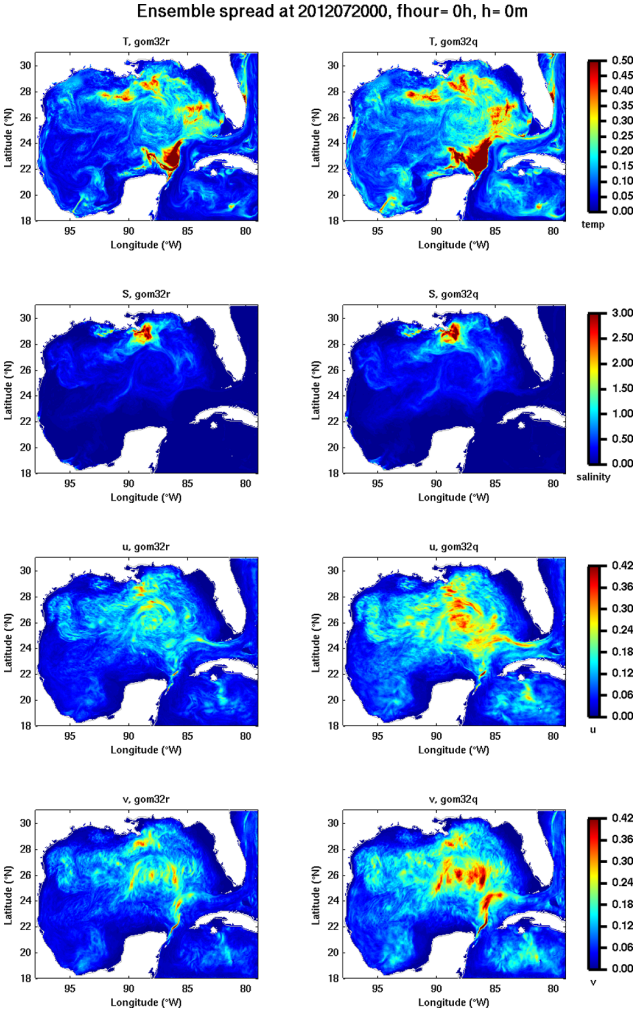

Fig. 1. Initial ensemble spread for r (left panel) and q (right panel) at 00:00 UTC 20 July 2012 on the surface for temperature (T), salinity (S), u, and v (from top to bottom).

carried out against observations in observation space, too few observations can also lead to larger spread.

To compare the spread from another perspective, the temperature spreads for ensembles are plotted as a function of forecast lead time in Fig. 3. They are averaged over the full observation space (a), and for the layer from 0–100 m (b). It can be seen that the ensemble spreads from both ensemble systems grow slightly over 72 h in both observation spaces. In addition, the enhancement of the spread from the calibrated ensemble q is evident.

3.2 Impact of the calibration on ensemble reliability

As demonstrated in the previous section, the ensemble spread is clearly enhanced by the calibration we introduced. But, does this enhancement increase the ensemble reliability, forecast accuracy, or skill? We will show the reliability comparison in this section. The comparison of forecast accuracy and skill will be provided in the next section. The spread of a reliable ensemble system should capture the forecast errors as a function of the forecast lead time. An ensemble with too small a spread will miss important dynamic events, especially extreme ones, while an ensemble with too large a spread will make the ensemble less sharp and less reliable with lower resolution.

In this section, we compare the two ensemble spreads with metrics that are especially designed for ensemble systems. The first one is the perturbation versus error correlation analysis (PECA) introduced by Wei and Toth (2003). PECA is designed to reduce the influence of initial analysis error, instead it evaluates the ensemble perturbations by measuring their ability to explain the forecast error variance. Therefore, PECA is a more appropriate, independent metric for the comparison of ensembles generated using different analysis schemes. A brief description and summary of PECA is provided in Appendix A, and more details can be found in Wei and Toth (2003).

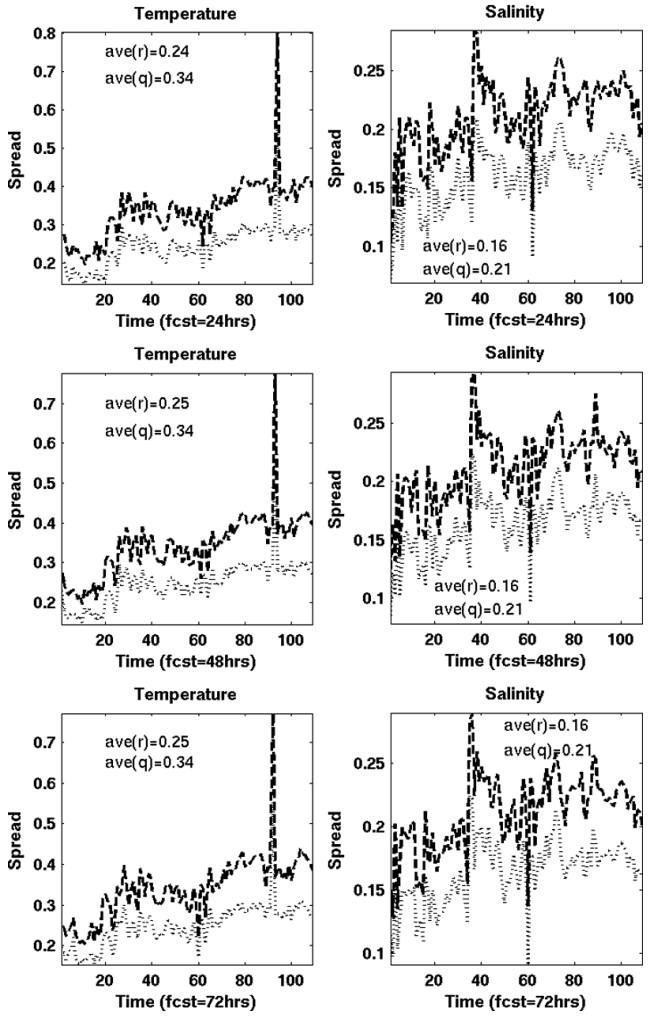

Fig. 2. Ensemble spreads for r (dotted) and q (dashed) as functions of day during the experimental period from 1 June to 17 September 2012. The spreads of T (left panel) and S (right panel) are shown for forecast lead times of 24, 48 and 72 h from top to bottom panels. The spread values are averaged over the observation space between 0 and 100 m.

Figure 4 shows PECA values from the optimally combined perturbations as defined in Eq. (A2) for both ensembles for temperature and salinity over full observation space, surface and space of the layer 0–100 m. The results clearly show that the calibration increases the optimal PECA. This means that as forecast lead time increases, the dimension of the subspace spanned by the ensemble perturbations is increased compared with the original ensemble r. The optimally combined perturbation in q can explain more forecast errors. The increment of PECA value is larger in the observation space with higher dimensions such as the full observation space and the space between 0 and 100 m in Fig. 4a, c, and d. The difference is smaller at the surface (Fig. 4b).

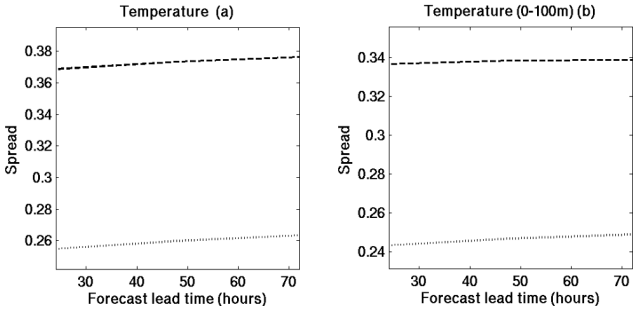

Fig. 3. Ensemble spreads of temperature for r (dotted) and q (dashed) as functions of forecast lead time. The spread values are averaged over the 109 days from 1 June to 17 September 2012, and over the full observation space **(a)**, the layer between 0 and 100 m **(b)**.

In order to test whether the differences between the two ensembles are statistically significant, we plot error bars (EBs) for both ensembles at forecast lead times of 24, 48, and 72 h. It is known that there are different ways of defining statistical significance which is also dependent on the user's needs. In this paper, we use the standard error (SE). Each vertical EB covers the range of mean value on the curve minus and plus SE, thus, showing the confidence interval (CI) around the mean. The values of lower and upper limits for mean value $\overline{x}$ on the curve with CI are $\overline{x} - \mathrm{SE}$ and $\overline{x} + \mathrm{SE}$ respectively. With the CIs specified by EBs, one can easily estimate whether the differences between the values of ensembles r and q are statistically significant. Since the sample sizes for both ensembles are the same, we will consider the differences between ensembles q and r as (not) statistically significant if the two EBs do not (do) overlap. Here SE is estimated as the sample standard deviation divided by the square root of the sample size as in any standard text books. The mean and SE describe bounds on a sample mean. The SE is an estimate of how close the sample mean is likely to be in comparison to the population mean, whereas the standard deviation is the degree to which individuals within the sample differ from the sample mean. More details about SE, CI and statistical significance can be found in statistical books such as Wilks (2006). The same definition and method will be used in some other figures with EBs later. The EBs in Fig. 4 show that the advantages of ensemble q over r are statistically significant in all forecast lead times and spaces, except for the sea surface temperature (SST). The differences for SST are not statistically significant, perhaps due to the relatively small verifying observation space.

The second metric is the spread-reliability diagram (Talagrand et al., 1997; Leutbecher and Palmer, 2008) . It is computed with 20 bins based on our 32-member ensembles, using observations as the truth. The exact steps for our RELO ensembles are outlined in Appendix A of W13. As an example of a comparison using this metric, Fig. 5

shows ensemble spread-reliability diagrams for temperature using observations as the truth. To have maximum statistical significance, all the values are averaged over a large number of samples within the full observation spaces from 1 June to 19 September 2012. Since the ensemble spread is expected to represent the forecast uncertainty, the spread-reliability curve over such a large sample should be close to diagonal line, which indicates perfect reliability.

The results in Fig. 5 show that the ensemble spread of r is small or under dispersive for all the ranges for all the forecast lead times. The calibrated ensemble q is also under dispersive, especially for smaller variance, but it is closer to the diagonal line for larger variances. For larger variance at 24 h lead time, it is slightly over dispersive. It is evident that the spread-reliability curve for the calibrated ensemble q is closer to the diagonal line for all the forecast lead times in the full observation space. This means that the reliability of the ensemble is enhanced by the calibration. The spread-reliability diagrams over the other observations spaces, such as the surface and the space between 0 and 100 m, are also computed (not shown) and yield similar conclusions. The CI based on sample uncertainty can be calculated using the bootstrapping technique as shown in Hamill (1999) and Hamill et al. (2008).

The ensemble spread reliability and its consistency can also be diagnosed by another popular metric called the rank histogram or the Talagrand histogram, which is described in Talagrand et al. (1997), Candille and Talagrand (2005), and Wilks (2006). The procedures for computing rank histograms and consistency indices for our RELO ensembles using observations as truth are described in Appendix B of W13. Our consistency index defined in Appendix B of W13 is just a modified version from Talagrand et al. (1997), Candille and Talagrand (2005). The advantage is that the modified index is exactly the ratio of RMS distances. The rank histograms for both temperature and salinity are computed at three forecast lead times, namely 24, 48, and 72 h, and in three domains, including the full observation space, the space between 0 and 100 m, and the surface (for temperature only). Shown in Fig. 6 are salinity rank histograms for both ensembles.

A quantitative measure of flatness is given by the consistency index of rank histogram. The value of consistency index indicated in each of the histograms shows that the index value of ensemble q is about half that of ensemble r in each of the cases. Based on this index, ensemble q is flatter than ensemble r at all 3 forecast lead times over these two observation spaces. In addition, the index values for ensemble q in the observation space of 0–100 m (right panel) are much closer to 1, which is the value for an ideal ensemble system. However, one notices the larger values near the middle ran for q. This indicates that there are more observations falling near the center of the ensemble q in comparison with r, and ensemble q is a little over dispersive for salinity. This is also confirmed by the spread-reliability diagram of salinity (not shown). For

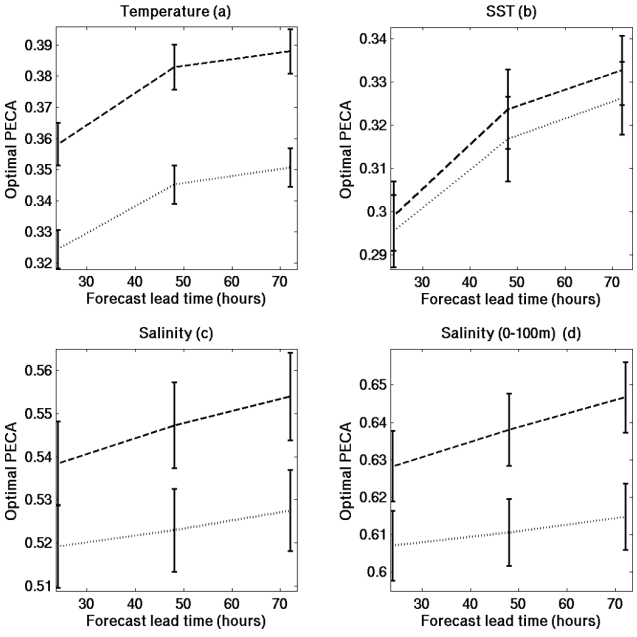

Fig. 4. Optimal PECA as a function of forecast lead time for ensembles r (dotted) and q (dashed) for temperature in full observation space **(a)**, surface temperature **(b)**, salinity in full observation space **(c)**, and salinity in layer between 0 and 100 m **(d)**. Error bars based on the standard error indicate the confidence interval.

temperature rank histogram, both ensembles q and r both are U-shaped, indicating both ensembles are under dispersive for temperature, which is confirmed by Fig. 5. However, the consistency indices for q are much smaller than those for r over all observation spaces (not shown). In general, the rank histograms for ensemble q are much flatter than for ensemble r in all cases. This is a clear indication that the calibrated ensemble q is much more consistent than ensemble r.

3.3 Impact of the calibration on forecast accuracy and skill

The RELO ensemble is expected to provide forecast uncertainty with its spread and superior forecasts with its ensemble mean, which is expected to be more accurate and skillful than an individual forecast. In this study, we use the RMS difference between the ensemble mean and subsequent observations corresponding to the forecast time to measure the forecast error. The RMS error is one of most commonly used metrics to quantify forecast accuracy. It is the difference between the model forecast and the truth represented by unassimilated observations valid during the forecast interval; thus it is a direct measure of forecast accuracy. Forecast accuracy generally decreases as the forecast lead time increases, and this change in accuracy is represented as a growth in the RMS forecast error. An ideal ensemble is

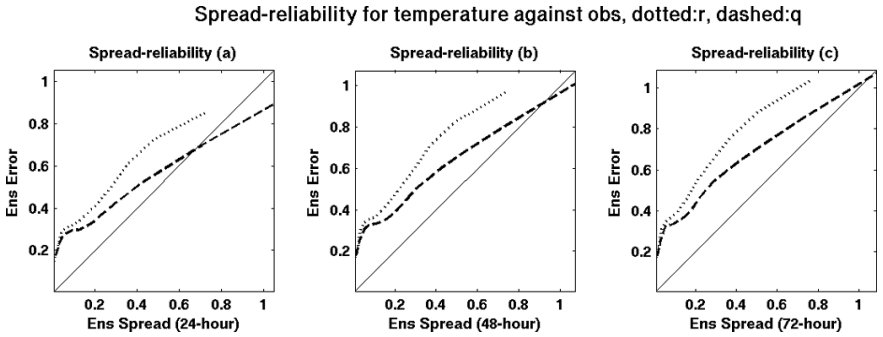

Fig. 5. Ensemble forecast spread-reliability diagrams for temperature (r: dotted, q: dashed) using observation as truth for lead times of 24, 48, and 72 h (from left to right). Values are averaged over the 109 days period from 1 June to 17 September 2012, and over the full observation space.

expected to have an ensemble spread that has a similar magnitude and growth rate to the ensemble RMS error.

To investigate whether the calibration introduced will enhance the forecast accuracy, we compute the RMS errors of the ensemble means from both ensembles r and q for temperature and salinity. In addition, the RMS errors of the single deterministic forecasts based on ncom3km, which has the same resolution as the ensembles, are also calculated for the same period of time and against the same operational observations as the ensembles. The results will provide a direct, fair comparison between ensemble (r) and a single model (ncom3km) that uses the same model and resolution, and comparison between the original ensemble (r) and the calibrated ensemble (q).

Figure 7 shows the RMS errors of the two ensemble means and one single forecast from ncom3km for salinity as a function of lead time. The RMS values are averaged over the full observation space (a), and the layer between 0 and 100 m (b). In both observation spaces, the ensemble with calibration (q) has lower RMS values than ensemble r, which has lower RMS values than ncom3km for all the forecast lead times. In other words, the calibrated ensemble (q) is more accurate than the original ensemble (r), which is more accurate than the single deterministic forecast with the same model and configuration (ncom3km). One also notices that the differences between ensemble r and ncom3km are smaller for shorter lead times, and not statistically significant, but the differences grow as the lead time increases, and become statistically significant after lead time of 48 h over both observation spaces. In contrast, the calibration makes larger impacts on the RMS values of the ensemble mean, even for shorter lead times although these differences fall just short of statistical significance and do not change much as the lead time increases.

We next turn to assess the forecast skill of these two ensembles and compare them with the single forecast generated by ncom3k. One of the most common metrics to quantify forecast skill is the anomaly correlation (AC). As for other metrics, we use observations as the truth. A simple correlation coefficient (CC), which is defined as the correlation between forecast and the observed values, is also sometimes used in the literature. But AC is preferred, since CC does not take forecast bias into account and it is quite possible for a forecast with large error to have a high CC value. It is a common practice to use climatology as the reference to account for seasonal variations when AC is computed (Wilks, 2006). For a forecast variable f at a particular forecast lead time, with c as the climate data and y as the observation field at the same verifying locations as the forecast, AC is defined as the correlation between the forecast and observation anomalies with respect to climatology, i.e.,

$$\mathrm{AC} = \frac{\overline{(f-c)(y-c)}}{\sqrt{\overline{(f-c)^2}}\sqrt{\overline{(y-c)^2}}}, \quad (1)$$

where the over-bar indicates the geographical mean over the verifying space. Therefore, the AC quantifies similarities in the pattern of departure (or anomalies) from the climatology field; it is a pattern correlation and regarded as a skill score relative to climatology. It is arguably the most commonly used metric in NWP centers (Buizza et al., 2005). We have used the climatological data obtained from NAVOCEANO.

In Fig. 8 the AC values of salinity averaged over the full observation space, the space between 0 and 100 m for both ensembles and the single forecasts from ncom3km are shown. In both observation spaces, the calibrated ensemble (q) has the highest skill score, while the AC values for ensemble r are higher than those from ncom3km for all the forecast lead times. The increment generated by the calibration for the ensemble is even larger than the advantage of ensemble r over the single forecast from ncom3km for short lead times. However, similar to the RMS errors in Fig. 7, the advantages of ensemble r over the single model forecast from ncom3km become larger as the forecast lead time increases, and the differences are becoming statistically significant at about 72 h forecast lead time. The enhancement generated by the calibration does not change much as the forecast lead time increases, and these differences are not

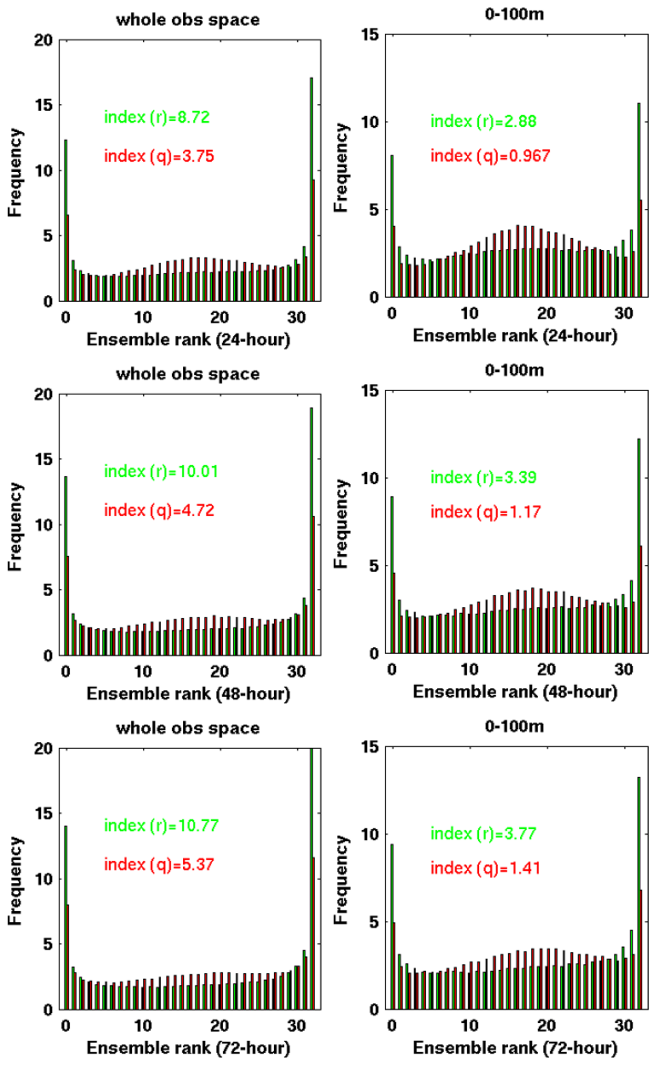

Fig. 6. Talagrand rank histograms for salinity (r: green, q: red) using observation as truth for lead times of 24, 48, and 72 h (from top to bottom). All the values are averaged over the 109 day period from 1 June to 17 September 2012, and over the full observation space (left panel), the layer between 0 and 100 m (right panel). Consistency index is also indicated in each case for both ensembles.

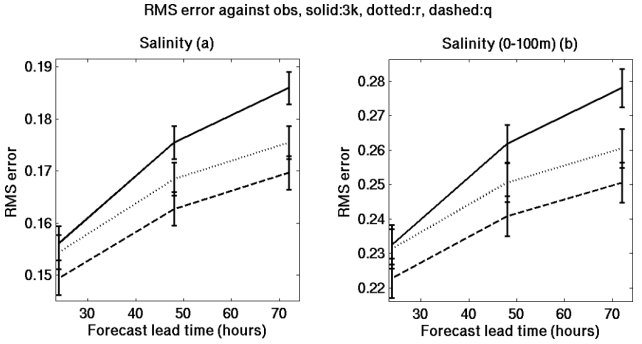

Fig. 7. The RMS errors of salinity for ensembles r (dotted), q (dashed), and ncom3km (solid) as functions of lead time. All the RMS values are averaged over the 109 days from 1 June to 17 September 2012, and averaged over the full observation space **(a)**, and the layer between 0 and 100 m **(b)**. Error bars based on the standard error indicate the confidence interval.

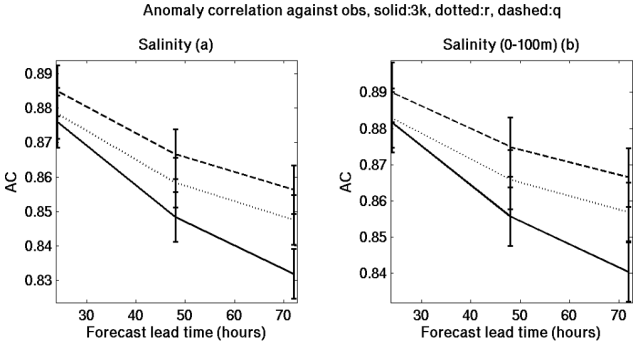

Fig. 8. The anomaly correlation of salinity for ensembles r (dotted), q (dashed) and ncom3km (solid) as functions of lead time. All AC values are averaged over the 109 days from 1 June to 17 September 2012, and averaged over the full observation space **(a)**, and the layer between 0 and 100 m **(b)**. Error bars based on the standard error indicate the confidence interval.

statistically significant based on the SE over these two spaces.

The most common measure of accuracy for a probabilistic forecast is the Brier score (BS) (Candille and Talagrand, 2005; Wilks, 2006). The BS is essentially the mean square error of the probabilistic forecasts considering if the event occurs. It is analogous to the mean square error of a deterministic forecast. It is the average of squared differences between pairs of forecast probabilities (p_i) and the subsequent binary observations (o_i), i.e.,

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2, \tag{2}$$

where the index i denotes a numbering of the n forecast-event pairs, $o_i = 1$ if the event occurs and $o_i = 0$ if the event does not occur. In this article, the dichotomous events are defined according to the climatology. To increase the statistical significance, we choose 10 climatologically equally likely categories based on the climatology data for temperature and salinity at each observation location. The BSs will be averaged over these 10 event categories. This is better than using a single dichotomous event (Buizza et al., 2005). It is clear that the BS is negatively oriented, with a perfect forecast being BS = 0. Less accurate forecasts receive higher values of BS, with the worst score being BS = 1.0. The BS can be decomposed into three terms called reliability, resolution, and uncertainty by using binned probabilities, i.e.,

$$\mathrm{BS} = \mathrm{Reliability} - \mathrm{Resolution} + \mathrm{Uncertainty}. \tag{3}$$

The detailed algebraic derivation and discussion of these three terms can be found in Wilks (2006).

Figure 9 shows the BS values of temperature and salinity as a function of forecast lead time for ensembles averaged over the full observation space and the space of 0–100 m. All the BS values for ensemble q are smaller than those for ensemble r, indicating q performs better as a probabilistic forecast system over these two spaces. The enhancement of the probabilistic accuracy generated by the calibration is clearly evident for both temperature and salinity in both observation spaces. The gap between the two ensembles tends to grow as the lead time increases for salinity in both observation spaces (bottom panel). The EBs in Fig. 9 show that the differences between the two ensembles are statistically significant for all the forecast lead times and over all spaces. In contrast, the differences of two ensembles based on deterministic metrics such as RMS error and AC are not statistically significant (Figs. 7 and 8). Thus, the calibration of ensemble initial spread has larger impact on the probabilistic forecasts than deterministic forecasts provided with ensemble mean.

Both reliability and resolution are computed based on Eq. (3) with 33 probability categories based on our 32-member ensemble for the decomposition. The conclusion that ensemble q performs better still holds. For example, the resolution values of temperature over the two observation spaces for ensembles (r, q) at forecast lead times of 24, 48, and 72 h are shown in Table 2. It is clear that the gaps between the two ensembles are almost constant as forecast lead time increases. However, for salinity over the two observation spaces, the performance gap increases slightly as a function of forecast lead time, which is shown in Table 3. The Brier skill score (BSS) based on BS using climatology as reference has also been computed, the results confirm the about conclusions (not shown).

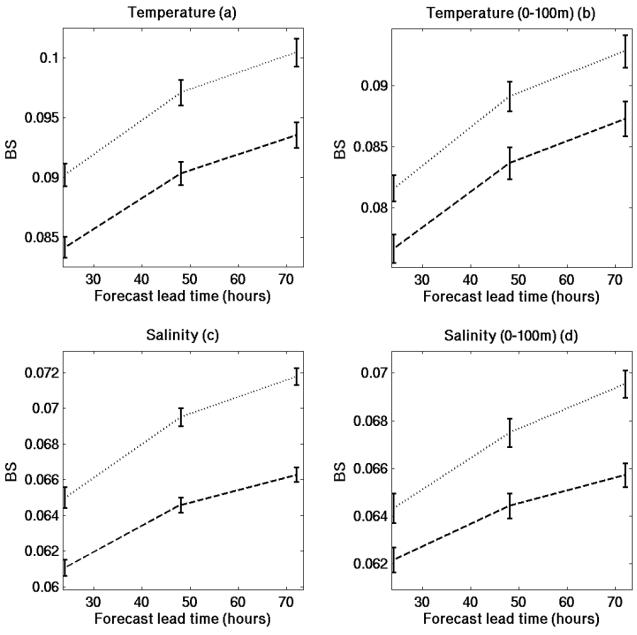

Fig. 9. Brier scores of temperature (top panel) and salinity (bottom panel) for ensembles r (dotted) and q (dashed) as functions of lead time. All BS values are averaged over the 109 days from 1 June to 17 September 2012, and averaged over the full observation space (left panel), and the layer between 0 and 100 m (right panel). Error bars based on the standard error indicate the confidence interval.

3.4 Impact of the calibration on Lagrangian trajectory prediction

The results and findings discussed in all the previous sections are based on the Eulerian formulation, i.e., the ocean states are described at fixed grid points. In this section, we turn our attention to the application of ensembles to Lagrangian dynamics and prediction, including particle trajectory prediction using ensembles.

Forecast velocity fields from numerical models have been used to predict particle or drifter trajectories on the water surface in the application of ocean models to Lagrangian dynamics. Some of the earlier work on the prediction of drifter trajectories in ocean simulations includes Özgökmen et al. (2000, 2001). The significance of the contribution ocean models have made to disaster response, search, rescue, and contaminant monitoring and mitigation has been manifested in the aftermath of the DWH oil spill accident. The value of trajectory predictions by ocean models has been demonstrated particularly well after the 2010 DWH incident by Maltrud et al. (2010), Huntley et al. (2011b), and Mariano et al. (2011). A special collection of papers about using ocean models to predict particle trajectories entirely devoted to the 2010 DWH oil spill can be found in Liu et al. (2011). Although most of these studies are based on single model forecasts, a few of the studies used ensembles. However, those ensemble results are actually composed of different forecasts from different models or organizations. In this case, it is not easy to estimate the uncertainties of the predicted trajectories. In the following, we will apply RELO ensemble forecasts to Lagrangian prediction. The advantage of using an ensemble is that it provides not only the more accurate ensemble mean to describe the ocean state, but also valuable uncertainty information, which is not directly available from traditional, single, deterministic forecasts. For the sake of computational efficiency, we advect particles with the 2-D, interpolated surface velocity using a fourth-order Runge–Kutta integration. In this study, we do not attempt to account for diffusive processes or subgrid-scale uncertainties.

In addition to demonstrating the extra information provided only by the RELO ensemble, more attention will be paid to the impact and value that is offered by the calibration introduced in this study. We choose 7 particles to represent drifters or surface oil patches from the DWH disaster, which are indicated by A, B, C, D, E, F, and G on the surface of the GOM. The trajectories of these particles can be predicted

Table 2. Resolution for temperature.

	24 h (r, q)	48 h (r, q)	72 h (r, q)
Full Observation Space	(0.017, 0.018)	(0.014, 0.015)	(0.013, 0.014)
0–100 m	(0.022, 0.023)	(0.018, 0.019)	(0.017, 0.018)

Table 3. Resolution for salinity.

	24 h (r, q)	48 h (r, q)	72 h (r, q)
Full Observation Space	(0.0300, 0.0305)	(0.0264, 0.0276)	(0.0248, 0.0263)
0–100 m	(0.0291, 0.0294)	(0.0266, 0.0274)	(0.0252, 0.0263)

from the water currents generated from any of the ocean prediction models and ensembles. To show the importance of uncertainty information provided by the calibrated ensemble, particles F and G are placed near the sensitive hyperbolic locations based on our water-current forecasts. For each of these particles, we integrate the velocity fields provided by each ensemble member to obtain a particle trajectory. Thus, there are 32 different possible trajectories from each location from one ensemble forecast. The trajectories of these particles from ensemble r for the period of 00:00 UTC 20 July to 00:00 UTC 23 July 2012 are shown in the top panel of Fig. 10. The corresponding trajectories generated by ensemble q are displayed in the bottom panel of Fig. 10. Also plotted in both panels of Fig. 10 are the sea surface height (SSH) in colored contours, and the surface current velocity vectors. The correlation between the SSH and velocity is clear. In addition, the particle trajectories tend to follow the directions indicated by the velocity vectors, as expected.

The predicted trajectory for each particle based on the ensemble mean is plotted in a thick red curve. When ensemble spread distribution is close to Gaussian, ensemble mean and mode will be close. This is shown to be the case in our RELO from the plume distribution in W13. Thus, the trajectory based on ensemble mean is almost the same as the one based on mode with the highest probability, i.e., the most likely scenario for the particle to follow. If a single model such as ncom3km, ncom1km, or hycom4km is used, it will generate just one trajectory (one possible outcome) for each particle. Due to the highly nonlinear, chaotic nature of the GOM, there are uncertainties in the initial conditions. Hence, there should be a range of possibilities for the trajectory that each particle might follow. Instead of a single trajectory produced by a single model for a particle, ensemble r or ensemble q provides 32 possible trajectories representing different possibilities, together with the trajectory based on ensemble mean. This extra information should help users and decision makers make better and more scientifically sound decisions.

The impact of the calibration on the particle trajectories is evident by comparing the possible trajectories of each particle in these two panels. It is clear that the spread of the ensemble trajectories for each particle predicted by ensemble q (bottom panel) is larger. The spread of the ensemble trajectories can be defined and computed as the RMS distance between the trajectories predicted by the

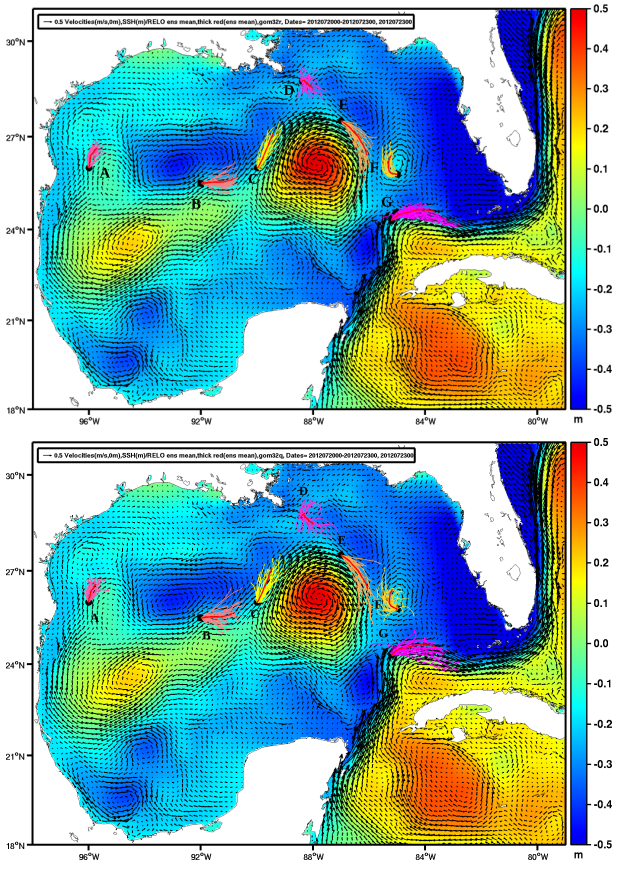

Fig. 10. The trajectories of 7 particles (A, B, C, D, E, F, G) predicted by ensemble members from r (top panel) and q (bottom panel). The predicted trajectories by ensemble means are denoted by thick red curves. Particle D is chosen to be at the location of DWH accident. Superimposed are the SSH in color contour, and surface water velocity indicated by arrows.

different ensemble members. Therefore, ensemble q can capture a wider range of possibilities of future movement of the particle than ensemble r (top panel). This can be seen from the predicted trajectories for particles B, C, D, and E. Particle D is located at the position of the DWH oil spill. The trajectory spread from both ensembles shows that the ocean dynamics near this location has large uncertainties. The oil particles spilled in this area could meander in many different directions; hence, it will make the trajectories more unpredictable around this area.

In some other cases, when both ensembles are used to predict a particle's possible trajectories, ensemble q could capture advection in some possible different directions that might be missed by ensemble r. Prediction for particle F is one example. All 32 members of ensemble r predict the particle moving northward. But, among the 32 members of ensemble q, there are two members predicting that particle F will move southward, i.e., ensemble q predicts a $2/32$ probability for particle F to move southward; these probabilities are missed by the less reliable ensemble r. For particle G, most members of both ensembles predict that the particle will follow the LC toward the Florida Strait. But, ensemble q predicts a $2/32$ probability for particle G to move westward, while ensemble r predicts only a $1/32$ probability for advection in this direction. Therefore, there are some uncertainties that are missed by the less reliable ensemble r. We have also noticed other examples on other dates when the calibrated ensemble q captures greater uncertainty than ensemble r near those sensitive, hyperbolic locations during our real-time ensembles runs for the GLAD experiment.

3.5 Ensembles in Lagrangian coherent structure

Another particularly interesting application of ensembles to the Lagrangian framework is the Lagrangian coherent structure (LCS). Not only can an ensemble provide the most likely LCS, but also its associated uncertainties. The LCS has been used in computational fluid dynamics (Haller and Yuan, 2000; Shadden et al., 2005; Haller and Sapsis, 2011) and it has been adopted in the ocean modeling community to study tracer distribution and prediction (Lekien et al., 2005; Coulliette et al., 2007; Olascoaga et al., 2008; Beron-Vera et al., 2008; Shadden et al., 2009; Olascoaga, 2010; Huntley et al., 2011a, b; Olascoaga and Haller, 2012). The LCSs are the locally most strongly attracting or repelling material surfaces in the flow. They move with the flow and provide coherent surfaces organizing the advection of tracers. Perhaps the most commonly used method to identify the LCS is to use the finite-time Lyapunov exponent (FTLE); some slightly varying definitions can be found in the literature. FTLE is a measure of the finite-time averaged maximum separation rate of two initially close fluid particles.

Shadden et al. (2005) provides a robust mathematical definition, and defines LCSs as ridges of maxima of FTLE fields. Equating LCSs with FTLE ridges of maxima offers an attractive easy tool to locate the LCS structures in real ocean models with a lot of data. The exact mathematical definition involving the derivatives of FTLEs to compute the ridges is also provided in Shadden et al. (2005). However, a simple, efficient way to estimate the LCS without the complicated computation of derivatives is to plot the FTLE field on a two dimensional surface and then the maximum values will be identified. With some caveats, the structure of these maximum values of FTLE can be used as the detector of LCS. This approach has been widely used in ocean predictions. In this paper, we also use the high values of FTLE to locate LCS. These maximum FTLE values, which are typically well-defined curves indicating high stretching between the fluid particles, appear as ridges in the graph of the FTLE field and serve as the definition of LCS (Lekien et al., 2005; Coulliette et al., 2007; Beron-Vera et al., 2008, 2010; Olascoaga et al., 2008; Beron-Vera and Olascoaga, 2009; Shadden et al., 2009; Olascoaga, 2010; Andrade-Canto et al., 2013). The maximum FTLEs identified this way have proven to be effective to identify LCSs. Following these studies, we concentrate on the 2-D surface flow of the ocean. The mathematical definition and computation of the FTLE are provided in Appendix B.

The attracting LCS, which is a material surface that attracts neighboring fluid particles at the locally highest rate over a time interval, has been used in tracer prediction and pollutant dispersal modeling in Olascoaga et al. (2008, 2012) and Olascoaga (2010). Beron-Vera et al. (2008) computed attracting LCSs to identify mesoscale eddies. Repelling LCSs are material lines that act as moving barriers to transport. They have important impacts on the movements of particles and pollutants on the ocean surface. Lekien et al. (2005) and Coulliette et al. (2007) have used repelling LCSs for optimizing pollution management and release in the coastal ocean in California and Florida. The repelling LCS was also used by Shadden et al. (2009) to help optimize drifter release in Monterey Bay. All of the above studies are based on deterministic forecasts from single ocean models. With the application of an ensemble, the uncertainties of these LCSs can be identified more easily.

In the following, we use our real-time forecast data generated for the GLAD drifter deployment to identify the LCSs and their uncertainties. Since several models or resolutions have been used in our experiments, the sensitivity of the LCS to the model and its resolution can be studied. The LCS from the ensemble mean velocity can be compared with those from single ocean models. The time interval $t - t_0$ is 3 days, which is the forecast length of all our model and ensemble forecasts for this experiment. In general, both repelling and attracting LCSs depend on the time interval chosen. However, the LCSs identified by using the FTLE method are reasonably robust. More discussion of this can be found in Shadden et al. (2005).

Figure 11 shows the repelling LCS on the ocean surface at 00:00 UTC 23 July 2012, which was three days after

deployment of the GLAD Large Scale Spiral (LSS) drifters. The repelling LCSs are computed based on the 3-day forecasts of the ensemble mean velocity of gom32q (a), ncom3km (b), ncom1km (c), and hycom4km (d). Since there is little difference between the LCSs from the ensemble means of gom32r and gom32q in terms of magnitude and scale, only the LCS from the ensemble mean of gom32q is shown.

The LCS of each ensemble member is computed for ensemble q. Each LCS from a different ensemble member represents a possible realization of the structure within this ensemble. The standard deviation (STD) of 32 LCSs based on the 32 individual members is a good estimate of the uncertainty of the LCSs described by the ensemble. The procedure and formula can be expressed as the following. For an ensemble with K members, the FTLE field σ_i for each member i can be computed based on Eqs. (B3) and (B4). The STD of FTLE fields from all ensemble members can be computed as

$$\mathrm{STD}(\boldsymbol{x}) = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(\sigma_i - \sigma_0)^2} \tag{4}$$

where σ_0 is the mean of the individual σ_i. We notice that σ_i itself is not the LCS which can be estimated as the ridge (maximum) of σ_i. In fact, $\mathrm{STD}(\boldsymbol{x})$ provides the uncertainty distribution of all the FTLE fields σ_i including σ_i with maximum values. Thus, $\mathrm{STD}(\boldsymbol{x})$ should provide a good estimate of the uncertainties of LCSs. This may not be the best approach to estimating uncertainties of LCSs, but it is a simple, efficient way of using an ensemble which is not easily available from single model forecasts.

The results in the previous sections show that ensemble q has a larger, more realistic spread and is more reliable than ensemble r for all the variables over the various domains. Thus, we choose the uncertainties of the LCSs identified by ensemble q, which are plotted as shaded colors in each of the four panels in Figs. 11–13.

The immediate difference one can notice among these repelling LCSs from the different models and the ensemble mean is the spatial scale due to the different resolutions of the models being used. The LCS generated from ncom3km shows smaller-scale structures outside the LC than hycom4km, which has lower resolution. Surprisingly, hycom4km displays smaller repelling LCS around the LC than ncom1km, ncom3km, and gom32q. Although gom32q uses NCOM with the same resolution as ncom3km, the LCS from the ensemble mean shows larger spatial structures than that from ncom3km. This is probably due to the filtering effect of the ensemble mean, which removes some of the smaller-scale features of the individual member forecasts. It is an advantage to use the ensemble mean if one is more interested in larger-scale features. As expected, the smallest scale features of the repelling LCS are revealed by ncom1km in Fig. 11c. Thus, ncom1km is an ideal model to study the sub-mesoscale eddy structures in the GOM. The large-scale repelling LCS around the LC is clearly identified by gom32q, ncom3km and ncom1km, but not by hycom4km, which instead shows a smaller, circular, repelling LCS near the north of the LC. All the models demonstrate a long, robust, large-scale repelling LCS starting from the Yucatan Current, connecting the LC, Florida Current, and Gulf Stream Current. To see the smaller-scale differences among these LCSs from the different models in the region of the GLAD experiment, an enlargement of the boxed region will be shown in Fig. 12.

In general both repelling and attracting LCSs act as transport barriers, if a drifter is on the ridge of the repelling LCS, it could fall to either side of the LCS. Once the drifter is on one side of the repelling LCS, it will be trapped on that side, as it is almost impossible for a drifter to cross the LCS barriers. Thus, the movement of a drifter depends greatly on how repelling LCSs change with time, and the drifter movement is generally constrained by the repelling LCSs. Accurately identified repelling LCSs will provide helpful guidance to drifter deployment and to forecasting the trajectories of drifters (Lekien et al., 2005; Coulliette et al., 2007; Shadden et al., 2009).

The uncertainty estimate associated with these LCSs is provided by gom32q and is plotted as shaded color. It can be seen that there are relatively larger uncertainties of the LCS along the Yucatan Current, connecting the LC, Florida and the Gulf Stream Currents where the large scale of LCS structures exist. Other regions with larger uncertainties are located around the DWH location, which is also the focused region (boxed in Fig. 11) for the GLAD LSS drifter deployment experiment. The repelling LCS at 00:00 UTC 23 July 2012 over the GLAD region for each model is shown in Fig. 12. Again, the uncertainty estimates are displayed as shaded color for each model. The DWH location is indicated by a white square. Now much more detailed LCS structures and their uncertainties can be seen clearly from all four forecasts, particularly the one ncom1km (Fig. 12c), which displays well-organized, small-scale repelling LCSs, which cannot be seen from the other low-resolution models. The ensemble mean shows similar LCS to ncom3km, but with some of the smaller scales filtered out. The LCS from hycom4km seems to have slightly different structures due to the differences in resolution and model coordinate and physics schemes. The uncertainty distribution is more pronounced over the LCS ridges, for gom32q, ncom3km, and hycom4km. But for the LCS from ncom1km, which has much higher resolution than the other three models, there is little correlation between the LCS and its uncertainty. This is expected as the uncertainty is estimated by the ensemble with $3\,\mathrm{km} \times 3\,\mathrm{km}$ resolution. To better estimate the uncertainty of the LCS from ncom1km, we need to run an ensemble based on NCOM with $1\,\mathrm{km} \times 1\,\mathrm{km}$ resolution. But this is too expensive to run with our current computing resources.

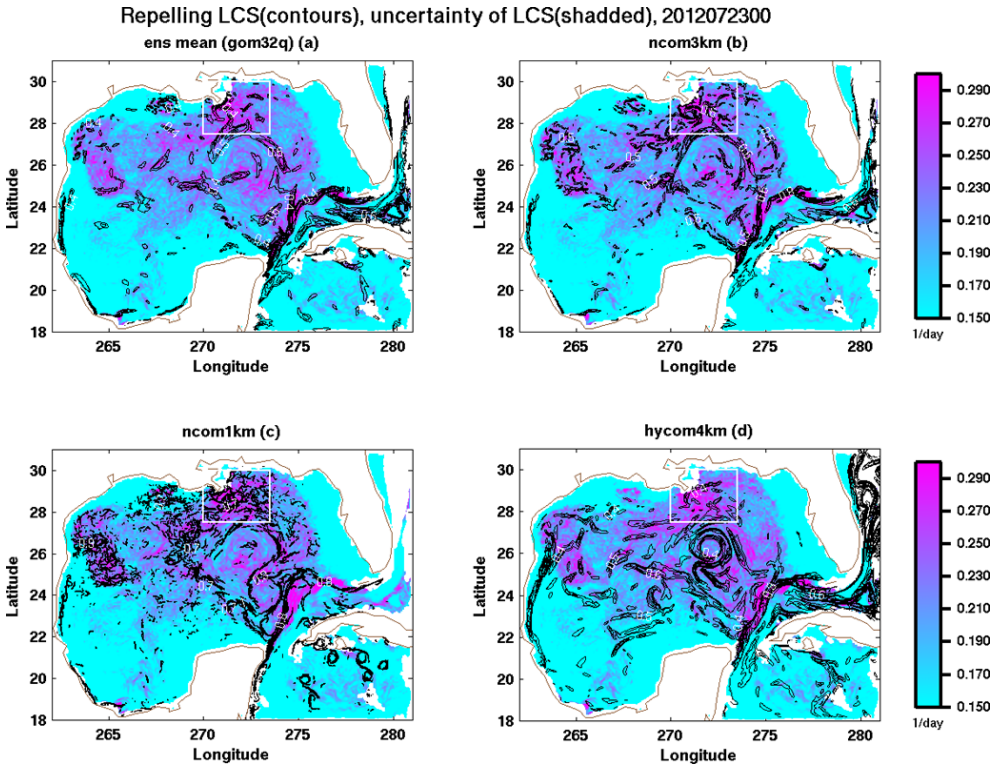

Fig. 11. The repelling LCSs (black contours) and their associated uncertainties of LCS (from gom32q, color shaded) on the ocean surface over the GOM at 00:00 UTC, 23 July 2012, generated by ensemble mean of gom32q **(a)**, ncom3km **(b)**, ncom1km **(c)**, and hycom4km **(d)**.

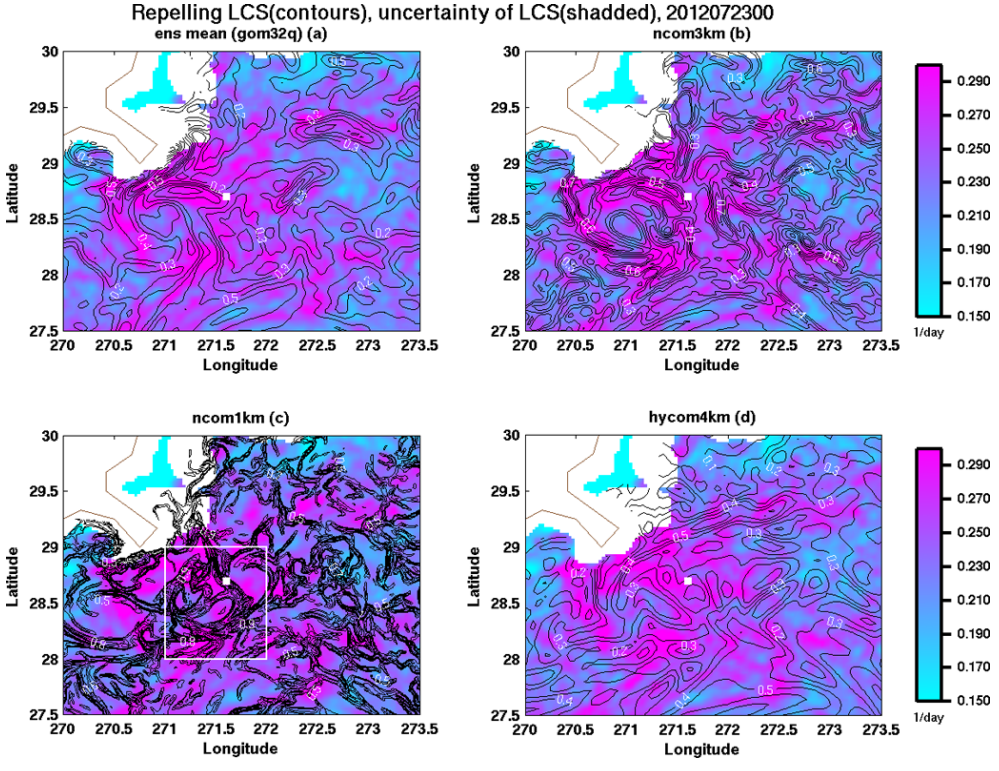

Fig. 12. The same as Fig. 11, but for a smaller domain around the GLAD drifters area indicated by the white rectangle in Fig. 11. The DWH location is indicated by the white rectangle.

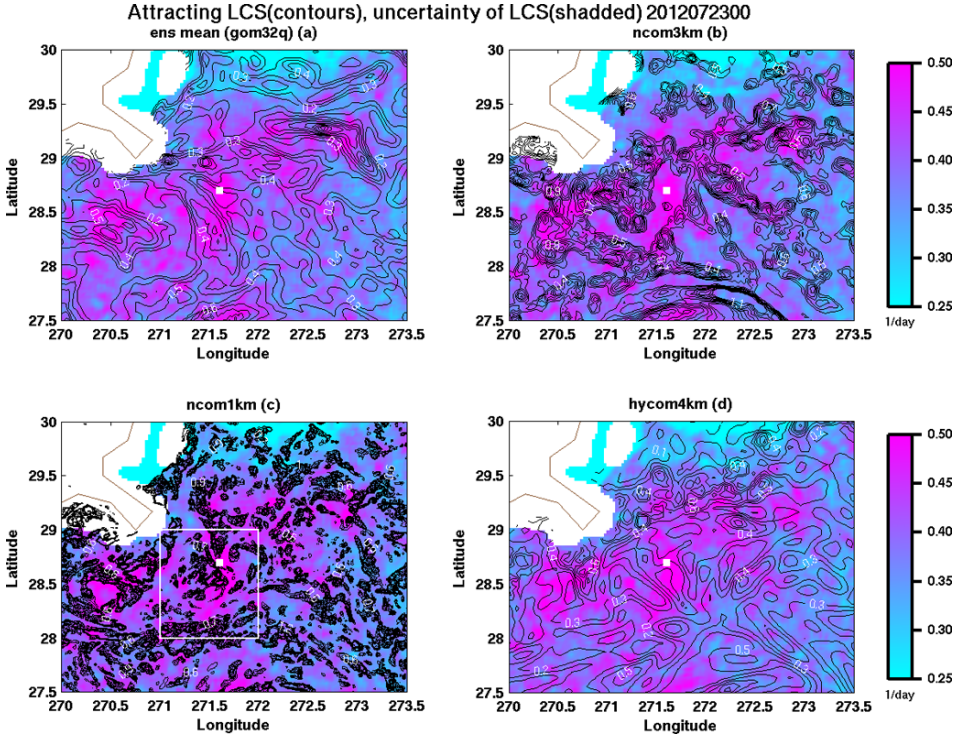

Fig. 13. The same as Fig. 12, but for attracting LCS.

Figure 13 is similar to Fig. 12, but shows the corresponding attracting LCS for each model. Both repelling and attracting LCSs are considered as material lines. If fluid particles straddle the attracting (repelling) LCS, they will converge (diverge) in forward time. The attracting LCS from ensemble has similar spatial-scale structure to the repelling LCS, while the spatial scales of attracting LCS from hycom4km are similar to those from the ensemble. Interestingly, the attracting LCS from ncom1km (Fig. 13c) appears to have smaller spatial scales than the repelling LCS from the same model (Fig. 12c). This might be an artifact from the way the attracting LCS is computed. A potential drawback with his method is that the grid of the smallest exponent is plotted at the final trajectory locations, which typically becomes deformed. To confirm this, future work is needed to compare these with the attracting LCS based on the conventional backward time FTLE method. To see more details of both the repelling and attracting LCS from ncom1km, we zoom to an even smaller region around the DWH location which is shown in Fig. 14. It is also very interesting to compare the repelling and attracting LCSs, their relative locations, and how they are interwoven in the region around the DWH location.

Mathur et al. (2007) used the LCS to uncover the Lagrangian building blocks of turbulence. They used the LCS to quantify the hyperbolicity of material lines in the Lagrangian skeleton. The authors argued that the complex tangle formed by the repelling and attracting LCSs is the underlying cause of turbulent particle motion. The LCS was also used by Beron-Vera et al. (2008) to unambiguously identify mesoscale oceanic eddies using the surface ocean currents. The authors noticed that the intersection of repelling and attracting LCSs define “lobes” that enclose and restrain fluid over time due to the material nature of the LCSs. In Fig. 14, both repelling and attracting LCSs from ncom1km are plotted together on 20–23 July 2012, which is the same period as Fig. 10 showing the Lagrangian trajectories predicted by the ensembles. The DWH location is indicated by a little blue square. In unsteady flows like the GOM, the repelling and attracting LCSs do not coincide, they transversely intersect each other many times. It can be seen that the repelling and attracting LCSs exist separately over the majority of the areas in this domain around the DWH location during this 4-day period. However, they do intersect at many locations where mesoscale eddies may be created. The movement of the particles at these locations, such as the particles from the DWH oil spill and the GLAD drifters, will be severely restricted by the complicated LCS structures. Most of the GLAD drifters released were most likely to be deployed in the troughs of LCSs; very few were on the ridges of LCSs. Fluid particles or drifters in the middle of these tangles are subject to attraction of the attracting LCS and simultaneous repulsion of the repelling LCS. One example is the trajectories of particle D from the DWH location predicted by both ensembles in Fig. 10. This particle moves a much shorter distance than the particles at the other

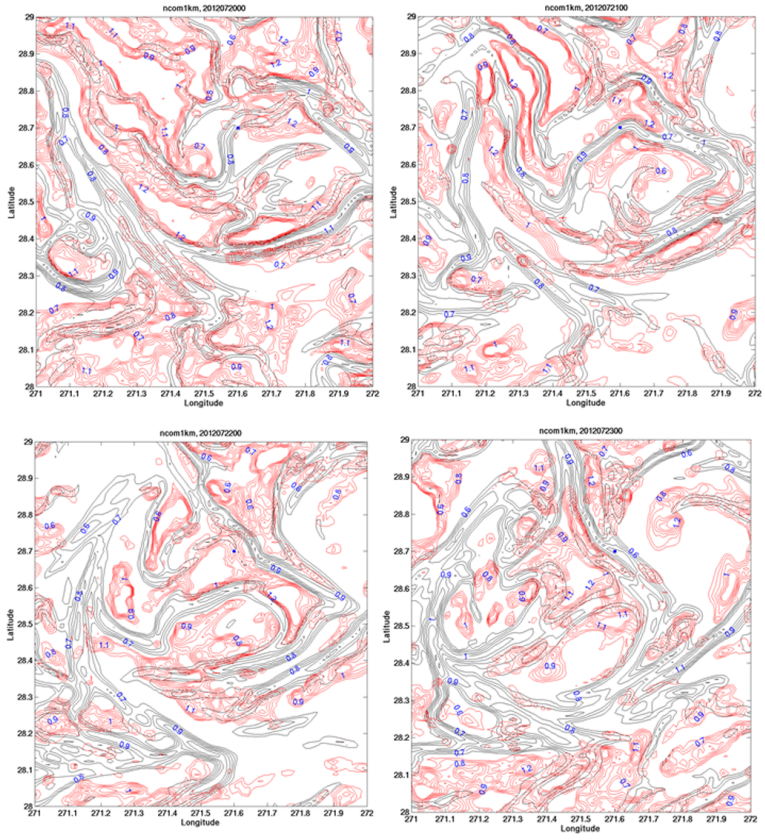

Fig. 14. The repelling (black) and attracting (red) LCSs generated by ncom1km over a small domain near DWH location that is boxed in Figs. 12c and 13c from 20 to 23 July 2012. The DWH location is indicated by the blue square.

locations over the three-day period we have predicted, such as C, E, F, and G. The ensemble predicted trajectories from D are restricted to a small area with many different directions by the complicated repelling and attracting LCSs.

Our next step is to compare the predicted drifter trajectories and the identified LCSs by the ensembles, and their associated uncertainties against the observed drifter trajectories from the GLAD data set. The full advantages of the ensembles are expected to be exploited and demonstrated with this valuable source of drifter data. We will report those results in the future.

4 Discussion and conclusions

As the designated modeling team within CARTHE to support and provide numerical guidance to the GLAD at-sea experiment in the summer of 2012, we carried out several real-time ocean model forecasts starting on 16 May 2012, well before the GLAD drifter deployment. Two ensembles (gom32r and gom32q), three single-model forecasts using ncom3km, ncom1km, and hycom4km were run. The output from all of these forecasts was archived and made available on web servers for all the CARTHE scientists and students involved in this project. The forecasts with different models and resolutions provide various choices for the different needs for CARTHE scientists during the GLAD drifter experiment. In this paper, we offer brief descriptions of these numerical forecast systems, with particular attention paid to the RELO ensemble with calibration, which was proposed to improve the ensemble performance. The advantages and disadvantages of the different systems and models are studied and summarized. Another goal of this study was to use the ensembles for the prediction of Lagrangian trajectories and Lagrangian coherent structures.

All our forecasts from both the ensembles and single models are evaluated and verified against the Navy's operational observations used in NCODA from 00:00 UTC 1 June to 00:00 UTC 19 September 2012. The verifications are based on the most commonly used verification metrics. Since the calculations used in NCODA underestimates the analysis error, the initial ensemble perturbations generated through the ET cannot match the real analysis error variance. Consequently, the ensemble spread is smaller than the ensemble mean error, and the reliability of the ensemble is compromised (W13). To overcome this difficulty efficiently in a short period of time for the targeted GLAD experiment, we tested the use of a calibrated ensemble (gom32q) with an enhanced initial spread. Another separate effort has been underway to improve the analysis error in NCODA, but this

will take a longer time to develop and evaluate. In fact, the mixing parameter perturbation scheme introduced in W13 is also part of these efforts to improve the RELO ensemble spread and overall reliability. The proposed calibration has been proven to be an efficient and effective method to further improve the ensemble spread.

To understand how much the spread has been enhanced by the calibration introduced, these two ensemble spreads are compared directly from different perspectives. These include direct comparisons of horizontal distributions, long-time averages over the whole period of the experiment, and averages over different observation spaces with different dynamics. All the results show that the ensemble spread is clearly enhanced by the calibration scheme. In addition, it is found that this calibrated ensemble (gom32q) is superior to the un-calibrated ensemble (gom32r) for all the variables in all the observation spaces we have tested in terms of quantitative forecasting accuracy, skill, and reliability. The metrics we have evaluated include RMS error, anomaly correlation, Brier score, PECA, spread-reliability, and Talagrand rank histogram using observations as truth and climatology as a reference to account for seasonal variation. It is also demonstrated that even the un-calibrated ensemble (gom32r) is more accurate and skillful than the single model forecast with the same resolution (ncom3km) based on the RMS error and anomaly correlation for all the variables and all the different observation spaces. Tests on statistical significance show that the differences between the two ensembles based on probabilistic metric BS are statistically significant, while the differences based on deterministic metrics (RMS error and AC) using ensemble means are not statistically significant. Thus, the initial spread enhancement has larger impact on the probabilistic forecasts than deterministic forecasts provided with ensemble mean.

The extra value of ensemble system in application to Lagrangian trajectory prediction is also demonstrated in this study. In contrast to a single ocean model forecast, ensemble can generate important uncertainty estimates in addition to predicting the most likely particle trajectory. In addition, the trajectory spread generated by the ensemble system directly reflects the complicated ocean dynamical properties near the area of interest, which cannot be revealed by single trajectory produced by a single model. All of this information is important for decision makers during drifter deployments and disaster relief efforts, such as the aftermath of the DWH oil spill incident. Moreover, the importance of ensemble reliability in predicting particle trajectories is demonstrated by the direct comparison of the two ensembles. The calibrated ensemble q, with more reliability, can pick up completely different trajectory directions, which are missed by the less reliable, uncalibrated ensemble r.

To reveal more details about the complex ocean dynamics in the GOM and the regions around the DWH location, both repelling and attracting LCSs are computed from ensembles q and r, and compared with those generated from the single models with different resolutions (ncom3km, ncom1km, hycom4km). The LCSs based on the ensemble means of both ensembles q and r are similar as expected. It is interesting to note that the LCSs identified by the ensemble means have larger spatial scales than those produced by ncom3km due to the filtering effect of the ensemble mean, which removes some small-scale features. This can be an advantage in situations where only the larger scales of the transport barriers are needed, such as for tracer prediction on longer timescales.

Our results also show that both repelling and attracting LCSs are sensitive to model resolution. The LCSs produced by hycom4km have the largest scales, while ncom1km, with the highest resolution in our experiments, is able to produce the finest small-scale LCS structures that cannot be generated by using lower-resolution models such as ncom3km, hycom4km, or the ensemble means. One advantage of the ensemble in this application is the capability for estimating the uncertainties of these LCSs.

To take advantage of our highest resolution model (ncom1km), we compared the repelling and attracting LCSs directly over the same domain and followed their time evolution. It was found that these two opposite LCSs do not exist in the same locations most of the time, but they indeed transversely intersect many times in the small region around the DWH location. These complicated tangles formed by the repelling and attracting LCSs are the underlying cause of the turbulent particle motion, and they define various "lobes" that restrain the movement of fluid particles, such as those from oil spill. This is found to be consistent with the Lagrangian trajectories predicted by the ensembles over the region.

The application of ensemble approach in the Lagrangian framework of ocean prediction is still largely unexplored. It is planned that the application of ensembles to Lagrangian trajectory and LCS prediction will be exploited further in the near future. The benefits of the ensemble over a single forecast have been widely recognized and accepted by the public, not just by researchers. The work presented in this study is just a first step in this direction. Some particularly interesting areas include the impact of Lagrangian spread on the trajectory prediction and LCS. How the repelling and attracting LCSs interact with each other over a region with complex turbulent particle motions. How the repelling and attracting LCSs control particle movements. One immediate task is to use the large amount of drifter data collected during the GLAD experiment to verify the Lagrangian trajectories predicted by the ensembles. With the observed drifter data, we will be able to exam the controls imposed by the repelling and attracting LCSs over the region around the DWH location.

Appendix A

PECA (Perturbation vs. Error Correlation Analysis)

It is known that ensemble performance depends on the quality of the model and the DA system that generates the analysis for the ensemble forecast (Buizza et al., 2005). Thus, with conventional verification metrics, it is difficult to distinguish the contributions from the improvements to the model, the DA system or the ensemble system design. In other words, it is difficult to assess the real performance of the ensemble initial perturbations, which are supposed to capture the initial analysis error variance. PECA was designed to supplement other conventional verification metrics as an ensemble verification tool that measures the performance of an ensemble system. More discussion of the PECA verification metric can be found in Wei and Toth (2003). The main properties of this metric can be summarized as (a) it is less sensitive to the performance of and errors in the model and DA system; (b) it evaluates the degree of independence of the ensemble members; (c) it measures how much of the forecast error can be explained by individual or optimally combined perturbations; (d) it reflects more on the quality of the ensemble method; and (e) a higher PECA score indicates a more skillful ensemble. Briefly, ensemble perturbations z_i^f are defined as the differences between individual perturbed forecasts and the ensemble mean. The forecast errors are normally defined as the differences between the ensemble mean forecast and the best available analysis or observations at verification time, i.e. $\boldsymbol{e} = \boldsymbol{x}^f - \boldsymbol{x}^a$.

The correlation between each perturbation and the forecast error can be computed. The mean correlation from all the perturbations should measure how much forecast error can be represented by the ensemble perturbations. Furthermore, we can also optimally combine all the perturbations to have one combined perturbation such that the final combined perturbation will be as close to the forecast error as is mathematically possible. This is achieved by solving a least square problem:

$$\mathrm{Min}||\boldsymbol{e} - \sum_{i=1}^{K} \alpha_i z_i^f||_{L2}. \tag{A1}$$

Having obtained α_i, the final optimally combined perturbation is defined as

$$\mathbf{p}_{\mathrm{optimal}} = \sum_{i=1}^{K} \alpha_i z_i^f. \tag{A2}$$

PECA values contain the correlations between the forecast errors and the optimally combined perturbations as well as the individual perturbations.

Appendix B

LCS (Lagrangian coherent structure)

Suppose the ocean velocity field generated by NCOM or HYCOM is $\mathbf{v}(x, y, t) = (u(x, y, t), v(x, y, t))$. The dynamical equation is given by

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{v}(x, y, t). \tag{B1}$$

If we follow a particle at time t_0 to a later time t, the integration of the above equation will provide a flow map $F(t_0, t)$ that maps the particle at the initial position to the current position at time t, i.e., $\boldsymbol{x}(t) = F(t_0, t)\boldsymbol{x}(t_0)$. A matrix can be formed using the gradients of the flow map as

$$\mathbf{C} = \left\{\frac{\mathrm{d}\mathbf{F}}{\mathrm{d}\boldsymbol{x}}\right\}^T \left\{\frac{\mathrm{d}\mathbf{F}}{\mathrm{d}\boldsymbol{x}}\right\}, \tag{B2}$$

with the superscript T indicating matrix transformation. This symmetric matrix is called the right Cauchy–Green deformation tensor, and is a function of t_0, $\boldsymbol{x}_0$, t, $\boldsymbol{x}$. The largest FTLE associated with this trajectory over the time interval $[t_0, t]$ is defined as

$$\sigma_{\mathrm{rep}}(\boldsymbol{x}_0, t_0, \boldsymbol{x}, t) = \frac{1}{|t - t_0|} \log\sqrt{\lambda_{\max}(C)}, \tag{B3}$$

where $\lambda_{\max}(C)$ denotes the largest eigenvalue of C. Therefore, the FTLE is the time-averaged, maximum, exponential stretching about the trajectory from time t_0 to t. There are two types of LCSs. The first kind is the repelling LCS, which is the material surface formed by the trajectories of the dynamical system that repel other trajectories at locally highest rate for the time interval $t - t_0$. The second one is the attracting LCS, which is the material surface that attracts nearby trajectories at locally highest rate for the time interval $t - t_0$.

A common way of computing the repelling LCS at time t_0 is to integrate a set of trajectories forward in time starting from an array of initial conditions up to a time t. Equation (B3) gives the largest FTLE, which can be used to identify the repelling LCS at time t_0. It is associated with the stable manifold. The ridges of the largest FTLE indicate the repelling LCSs. Another, separate, backward integration from time t to t_0 is needed to locate the attracting LCS at time t, which is associated with the unstable manifold. More details can be found in Shadden et al. (2005). However, in this study we use the new development by Haller and Sapsis (2011) to compute the attracting LCS. Instead of carrying out a separate backward integration, the authors proved that the attracting LCS at time t identified by the backward time integration described above can be computed using the minimal eigenvalue generated from the same forward time integration from time t_0 to t, i.e.

$$\sigma_{\mathrm{att}}(\boldsymbol{x}, t) = -\frac{1}{|t - t_0|} \log\sqrt{\lambda_{\min}(C)}, \tag{B4}$$

where $\lambda_{\min}(C)$ denotes the smallest eigenvalue of C. Equation (B4) is used to identify the attracting LCS. As a result, the computing cost is reduced by half.

Acknowledgements. This research was made possible in part by a grant from BP/The Gulf of Mexico Research Initiative (GoMRI) through the Consortium for Advanced Research on the Transport of Hydrocarbon in the Environment (CARTHE). It was also partly funded through the SEMESTER 6.2 project at NRL and supported by the Office of Naval Research. We are grateful to the assistance of many of our colleagues at NRL at Stennis Space Center, particularly Germana Peggion, Jan Dastugue, Michael Phelps, and the scientists from the other organizations that participated in CARTHE. We are grateful to the editor and two anonymous reviewers for the constructive comments and suggestions which have improved the manuscript. The explanation to spatial scale difference between the repelling and attracting LCSs produced by the high resolution model is kindly suggested by one of the reviewers.

Edited by: T. Gneiting
Reviewed by: two anonymous referees

References

Andrade-Canto, F., Sheinbaum, J., and Zavala Sansón, L.: A Lagrangian approach to the Loop Current eddy separation, Nonlin. Processes Geophys., 20, 85–96, doi:10.5194/npg-20-85-2013, 2013.

Barron, C. N., Kara, A., Martin, P., Rhodes, R., and Smedstad, L.: Formulation, implementation and examination of vertical coordinate choices in the Global Navy Coastal Ocean Model (NCOM), Ocean Model., 11, 347–375, 2006.

Beron-Vera, F. J. and Olascoaga, M. J.: An assessment of the importance of chaotic stirring and turbulent mixing in the West Florida Shelf, J. Phys. Oceanogr., 9, 1743–1755, 2009.

Beron-Vera, F. J., Olascoaga, M. J., and Goni, G. J.: Oceanic mesoscale eddies as revealed by Lagrangian coherent structures, Geophys. Res. Lett. 35, L12603, doi:10.1029/2008GL033957, 2008.

Beron-Vera, F. J., Olascoaga, M. J., and Goni, G. J.: Surface ocean mixing inferred from multisatellite altimetry measurements, J. Phys. Oceanogr., 40, 2466-2480, doi:10.1175/2010JPO4458.1, 2010.

Bleck, R.: An oceanic general circulation model framed in hybrid isopycnic-Cartesian coordinates, Ocean Model., 4, 55–88, 2002.

Bowler, N. E.: Comparison of error breeding, singular vectors, random perturbations and ensemble Kalman filter perturbation strategies on a simple model, Tellus, 58A, 538–548, 2006.

Bowler, N. E., Arribas, A., Beare, S., Mylne, K., and Shutts, G.: The local ETKF and SKEB: Upgrade to the MOGREPS short-range ensemble prediction system, Q. J. Roy. Meteor. Soc., 135, 767–776, 2009.

Buizza, R., Houtekamer, P. L., Toth, Z., Pellerin, P., Wei, M., and Zhu, Y.: A comparison of the ECMWF, MSC and NCEP global ensemble prediction systems, Mon. Weather Rev., 133, 1076–1097, 2005.

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

Chassignet, E. P., Smith, L. T., Halliwell, G. R., and Bleck, R.: North Atlantic simulations with the HYbrid Coordinate Ocean Model (HYCOM): Impact of the vertical coordinate choice, reference pressure, and thermobaricity, J. Phys. Oceanogr., 33, 2504–2526, 2003.

Coulliette, C., Lekien, F., Paduano, J., Haller, G., and Marsden, J.: Optimal pollution mitigation in Monterey Bay based on coastal radar data and nonlinear dynamics, Environ. Sci. Technol., 41, 6562–6572, 2007.

Cummings, J.: Operational multivariate ocean data assimilation, Q. J. Roy. Meteor. Soc., 131, 3583–3604, 2005.

Descamps, L. and Talagrand, O.: On Some Aspects of the Definition of Initial Conditions for Ensemble Prediction, Mon. Weather Rev., 135, 3260–3272, 2007.

Haller, G. and Sapsis, T.: Lagrangian Coherent Structures and the Smallest Finite-Time Lyapunov Exponent, Chaos, 21, 1–5, 2011.

Haller, G. and Yuan, G.: Lagrangian coherent structures and mixing in two-dimensional Turbulence, Physica D, 147, 352–370, 2000.

Halliwell, G. R.: Evaluation of vertical coordinate and vertical mixing algorithms in the HYbrid Coordinate Ocean Model (HYCOM), Ocean Model., 7, 285–322, 2004.

Hamill, T. M.: Hypothesis Tests for Evaluating Numerical Precipitation Forecasts, Weather Forecast., 14, 155–167, 1999.

Hamill, T. M., Snyder, C., and Morss, R. E.: A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles, Mon. Weather Rev., 128, 1835–1851, 2000.

Hamill, T. M., Hagedorn, R., and Whitaker, J. S.: Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part II: Precipitation, Mon. Weather Rev., 136, 2620–2632, 2008.

Houtekamer, P. L., Lefaivre, L., Derome, J., Ritchie, J., and Mitchell, H. L.: A system simulation approach to ensemble prediction, Mon. Weather Rev., 124, 1225–1242, 1996.

Huntley, H. S., Lipphardt, B. L., and Kirwan, A. D.: Lagrangian predictability assessed in the East China Sea, Ocean Model., 36, 163–178, 2011a.

Huntley, H. S., Lipphardt, B. L., and Kirwan, A. D.: Surface Drift Predictions of the Deepwater Horizon Spill: The Lagrangian Perspective, in: Monitoring and Modeling the Deepwater Horizon Oil Spill: A Record-Breaking Enterprise, edited by: Liu, Y., MacFadyen, A., Ji, Z.-G., and Weisberg, R. H., AGU, Washington, D. C., Geoph. Monog. Series, 195, 179–195, 2011b.

Lekien, F., Coulliette, C., Mariano, A. J., Ryan, E., Shay, L. K., Haller, G., and Marsden, J.: Pollution release tied to invariant manifolds: A case study for the coast of Florida, Physica D, 210, 1–20, 2005.

Lermusiaux, P. F. J.: Uncertainty Estimation and Prediction for Interdisciplinary Ocean Dynamics, J. Comput. Phys., 217, 176–199, doi:10.1016/j.jcp.2006.02.010, 2006.

Leutbecher, M. and Palmer, T. : Ensemble forecasting, J. Comput, Phys., 227, 3515–3539, 2008.

Liu, Y., MacFadyen, A., Ji, Z.-G., and Weisberg, R. H. (Eds.): Monitoring and Modeling the Deepwater Horizon Oil Spill: A Record-Breaking Enterprise, AGU, Washington D.C., Geoph. Monog. Series, 195, 271 pp., 2011.

Magnusson, L., Nycander, J., and Kallen, E. : Flow-dependent versus flow-independent initial perturbations for ensemble prediction, Tellus, 61A, 194–209, 2009.

Maltrud, M., Peacock, S., and Visbeck, M.: On the possible long-term fate of oil released in the Deepwater Horizon incident, estimated using ensembles of dye release simulations, Environ. Res. Lett., 5, 1–7, 2010.

Mariano, A. J., Kourafalou, V. H., Srinivasan, A., Kang, H., Halliwell, G. R., Ryan, E., and Roffer, M.: On the modeling of the 2010 Gulf of Mexico Oil Spill, Dynam. Atmos. Oceans, 52, 322–340, 2011.

Martin, P. J.: A description of the Navy Coastal Ocean Model Version 1, , Naval Research Laboratory, Stennis Space Center, MS, Technical Report, NRL/FR/7322-00-9962, 42 pp., 2000.

Mathur, M., Haller, G., Peacock, T., Ruppert-Felsot, T., and Swinney, H.: Uncovering the Lagrangian Skeleton of Turbulence, Phys. Rev. Lett., 98, 144502, doi:10.1103/PhysRevLett.98.144502, 2007.

McLay, J., Bishop, C. H., and Reynolds, C. A.: The ensemble-transform scheme adapted for the generation of stochastic forecast perturbations, Q. J. Roy. Meteor. Soc., 133, 1257–1266, 2007.

Mellor, G. L. and Durbin, P.: The structure and dynamics of the ocean surface mixed layer, J. Phys. Oceanogr., 5, 718–728, 1975.

Mellor, G. L. and Yamada, T.: A hierarchy of turbulence closure models for planetary boundary layers, J. Atmos. Sci., 31, 1791–1806, 1974.

Molteni, F., Buizza, R., Palmer, T., and Petroliagis, T.: The ECMWF ensemble prediction system: Methodology and validation, Q. J. Roy. Meteor. Soc., 122, 73–119, 1996.

Olascoaga, M. J.: Isolation on the West Florida Shelf with implications for red tides and pollutant dispersal in the Gulf of Mexico, Nonlin. Processes Geophys., 17, 685–696, doi:10.5194/npg-17-685-2010, 2010.

Olascoaga, M. J. and Haller, G.: Forecasting sudden changes in environmental contamination Patterns, P. Natl. Acad. Sci. USA, 109, 4738–4743, 2012.

Olascoaga, M. J., Beron-Vera, F. J., Brand, L. E., and Kocak, H.: Tracing the early development of harmful algal blooms with the aid of lagrangian coherent structure, J. Geophys. Res., 113, C12014, doi:10.1029/2007JC004533, 2008.

Özgökmen, T. M., Griffa, A., Mariano, A. J., and Piterbarg, L. I.: On the predictability of Lagrangian trajectories in the ocean, J. Atmos. Ocean. Tech., 17, 366–383, 2000.

Özgökmen, T. M., Piterbarg, L. I., Mariano, A. J., and Ryan, E.: Predictability of drifter trajectories in the tropical Pacific Ocean, J. Phys. Oceanogr., 31, 2691–2720, 2001.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian model averaging to calibrate forecast ensembles, Mon. Weather Rev., 133, 1155–1174, 2005.

Reynolds, C. A., Ridout, J., and McLay, J.: Examination of parameter variations in the US Navy global ensemble, Tellus, 63A, 841–857, 2011.

Rowley, C.: RELO SYSTEM USER GUIDE. Oceanography Division Naval Research Laboratory, Stennis Space Center, MS, USA, 59 pp., 2008.

Rowley, C.: Validation Test Report for the RELO System, Oceanography Division, Naval Research Laboratory, Stennis Space Center, MS, NRL Report, NRL/MR/7320–10-9216, 69 pp., 2010.

Rowley, C., Richman, J., and Emanuel, C.: Boundary Condition Uncertainty in the NRL Relocatable Ocean Ensemble Forecast System, AGU Ocean Science Meeting, Salt Lake City, UT, 20–25 February 2012.

Shadden, S. C., Lekien, F., and Marsden, J. E.: Definition and properties of Lagrangian coherent structures from finite-time Lyapunov exponents in two-dimensional aperiodic flows, Physica D, 212, 271–304, 2005.

Shadden, S. C., Lekien, F., Paduan, J. D., Chavez, F., and Marsden, J. E.: The correlation between surface drifters and coherent structures based on HF radar in Monterey Bay, Deep-Sea Res. Pt II, 56, 161–172, 2009.

Smagorinsky, J.: General circulation experiments with the primitive equations. I: The basic experiment, Mon. Weather Rev., 91, 99–164, 1963.

Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of probabilistic system. In Proc. Of
Workshop on Predictability, ECMWF, Reading, UK, 1-25, 1997.

Toth, Z. and Kalnay, E.: Ensemble forecasting at NMC: the generation of perturbations, B. Am. Meteorol. Soc., 174, 2317–2330, 1993.

Wei, M. and Toth, Z.: A new measure of ensemble performance: Perturbations versus Error Correlation Analysis (PECA), Mon. Weather Rev., 131, 1549–1565, 2003.

Wei, M., Toth, Z., Wobus, R., Zhu, Y., and Bishop, C.: Initial Perturbations for NCEP Ensemble Forecast System, in: Thorpex Symposium Proceedings for the First THORPEX Internal Science Symposium, 6–10 December 2004, Montreal, Canada, The Symposium Proceedings in a WMO Publication 2005, WMO TD No. 1237, WWRP THORPEX No. 6, 227–230, 2005.

Wei, M., Toth, Z., Wobus, R., Zhu, Y., Bishop, C., and Wang, X.: Ensemble Transform Kalman Filter-based ensemble perturbations in an operational global prediction system at NCEP, Tellus, 58A, 28–44, 2006.

Wei, M., Toth, Z., Wobus, R., and Zhu, Y.: Initial perturbations based on the Ensemble Transform (ET) technique in the NCEP global operational forecast system, Tellus, 60A, 62–79, 2008.

Wei, M., Toth, Z., and Zhu, Y. : Analysis differences and error variance estimates from multi-center analysis data, The Australian Meteorological & Oceanographic Journal, 59, 25–34, 2010.

Wei, M., De Pondeca, M. S. F. V., Toth, Z., and Parrish, D.: Estimation and calibration of observation impact signals using the Lanczos method in NOAA/NCEP data assimilation system, Nonlin. Processes Geophys., 19, 541–557, doi:10.5194/npg-19-541-2012, 2012.

Wei, M., Rowley, C., Martin, P., Barron, C., and Jacobs, G.: The U.S. Navy's RELO Ensemble Prediction System and its performance in the Gulf of Mexico, Q. J. Roy. Meteor. Soc., online first, 139, doi:10.1002/qj.2199, 2013.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Cambridge Press, 627 pp. 2006.