RMetS
Royal Meteorological Society

The US Navy's RELO ensemble prediction system and its performance in the Gulf of Mexico

Mozheng Wei,* Clark Rowley, Paul Martin, Charlie N. Barron and Gregg Jacobs
Naval Research Laboratory, Stennis Space Center, MS, USA

*Correspondence to: Mozheng Wei, Naval Research Laboratory, Stennis Space Center, MS, USA.
E-mail: Mozheng.Wei@nrlssc.navy.mil

The US Navy's relocatable (RELO) ensemble prediction system is fully described and is examined in the Gulf of Mexico for 2010. After briefly describing the ensemble transfer (ET) method for the initial perturbation generation, we introduce a new time-deformation technique to generate the surface forcing perturbations from the atmospheric model fields. The extended forecast time (EFT) is introduced to quantify the advantages of the ensemble mean forecasts over a single deterministic forecast. The ensemble spread and its growth are investigated together with their relations with the ensemble forecast accuracy, reliability and skill.

Similar to many other operational ensemble forecast systems at numerical weather prediction (NWP) centres, the initial analysis error is underestimated by the technique used in the data assimilation (DA) system. Growth of the ocean ensemble spread is also found to lag the growth of the ensemble mean error, a tendency attributed to insufficiently accounting for model-related uncertainties. As an initial step, we randomly perturb the two most important parameters in the ocean model mixing parametrizations, namely the Smagorinsky horizontal and Mellor–Yamada vertical mixing schemes. We examine three different parameter perturbation schemes based on both uniform and Gaussian distributions. It is found that all three schemes improve the ensemble spread to a certain extent, particularly the scheme with Gaussian distribution of perturbations imposed on both the horizontal and vertical mixing parameters.

The findings in this article indicate that the RELO ensemble forecast demonstrates superior accuracy and skill relative to a single deterministic forecast for all the variables and over all the domains considered here. The ensemble spread provides a valuable estimate of forecast uncertainty. However, the RELO uncertainty forecast capability could be further improved by accounting for more model-related uncertainties, for example, by the development of an error parametrization that imposes stochastic forcing at each model grid point.

Key Words: ocean ensemble prediction; ocean data assimilation; ensemble spread and reliability; forecast accuracy and skill; extended forecast time; Smagorinsky horizontal mixing; Mellor and Yamada vertical mixing; Talagrand/rank histogram

Received 29 January 2013; Revised 24 April 2013; Accepted 22 May 2013; Published online in Wiley Online Library 7 August 2013

1. Introduction

An ensemble prediction system is intended to generate a sample of numerical forecasts that represents our knowledge about the possible evolution of a dynamical system. Ensemble forecasts should preferably reflect forecast uncertainties related to both the initial values (analysis) and the representation of the evolving system through a numerical model. During the past 20 years, various perturbation methods have been developed to achieve these goals. It is generally accepted that initial ensemble perturbations must constitute a sample taken from a probability density function (PDF) that represents our best knowledge about the state and uncertainty of the dynamical system (i.e. the 'analysis PDF'). Various initial perturbation methods differ in how they estimate the analysis PDF and how they sample it.

Major meteorological centres have operationally implemented a variety of *first generation* initial perturbation generation methods, including: the perturbed observation (PO) method (Houtekamer *et al.*, 1996), the total energy norm-based singular vector (TE-SV) method (Buizza and Palmer, 1995; Molteni *et al.*, 1996), and the breeding method (BM: Toth and Kalnay,

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.

1993, 1997). These methods were all limited in that, for various reasons, the sample produced was not consistent with the analysis PDF. Summaries and discussions of these methods, including both advantages and disadvantages, are in Wei *et al.* (2008) (hereafter referred to as W08) and are listed in Tab. 1 of that paper. Generally speaking, the initial perturbations used in the first generation of ensemble prediction/forecast systems (EPS or EFS) do not fully represent the uncertainties in the analysis, as is expected from an ideal EPS, since the real initial analysis errors are not being used by these methods. Therefore, they are, in general, not consistent with the data assimilation (DA) systems that generate the analysis fields. Comparisons of the performance of the European Centre for Medium-range Weather Forecasts (ECMWF) and the US National Centers for Environmental Prediction (NCEP) operational EFSs are described in Wei and Toth (2003). A more recent comparison study of these methods in an operational environment, including their performance at ECMWF, the Meteorological Service of Canada (MSC) and NCEP can be found in Buizza *et al.* (2005). With an increased emphasis on the use of the analysis PDF for initial ensemble perturbation generation, a *second generation* of techniques has emerged in recent years. These newer techniques are discussed and summarized in Tab. 2 of W08. Such methods include the ensemble transform Kalman filter (ETKF), ensemble transform (ET), ensemble transform with rescaling (ETR), and the Hessian singular vector (SV) technique (Barkmeijer *et al.*, 1999).

After the ETKF method was proposed for adaptive observation and DA by Bishop *et al.* (2001), it was further applied to ensemble forecasting in an operational environment with the NCEP operational model and real-time observations by Wei *et al.* (2006). A local ETKF has been implemented at the UK Met Office (UKMO) (Bowler *et al.*, 2009). The ET technique was first proposed by Bishop and Toth (1999), also for adaptive observation studies. The work of using ET and ETR for ensemble forecast purposes was carried out by Wei *et al.* (2005), and the important properties of the ET- and ETR-based ensemble perturbations are derived and summarized in W08. The authors also compared results based on the BM, ETKF, ET and ETR methods. All four of these schemes involve the dynamical cycling of ensemble perturbations. In the ET and ETR methods, the initial perturbations are restrained by the best available analysis variance from the operational DA system and centred on the analysis field generated by the same DA system. In this way, the ensemble system remains consistent with the DA. The perturbations are also flow-dependent and orthogonal with respect to the inverse of the analysis error variance. If the analysis variance information is available, then the ET/ETR technique is considerably cheaper than ETKF. The research described in W08 led to the operational implementation of ETR-based EPS at the US National Weather Service on 30 May 2006.

The ET method for EPS has also been developed at the Naval Research Laboratory (NRL) Marine Meteorological Division and implemented in the operational atmospheric forecast model at the Navy's Fleet Numerical Meteorology and Oceanography Center (FNMOC), located in Monterey, CA (McLay *et al.*, 2007, 2008; McLay and Reynolds, 2009). Recently, a more efficient version of ET (banded ET) has been implemented at FNMOC (McLay *et al.*, 2010). A common feature of the second-generation techniques is that the initial perturbations are more consistent with the DA system. A good DA system will provide accurate estimates of the initial analysis error variance for the EPS, while a good, reliable EPS will produce an accurate flow-dependent part of the background covariance for the DA system.

Estimating analysis error covariance is important in building an efficient EPS based on ET and ETR. A simple way to estimate this error is to use multi-centre analysis data, which are routinely available at most major forecast centres. Wei *et al.* (2010) describe a method for estimating the analysis error variance using analysis data from NCEP, ECMWF, UKMO, Canadian Meteorological Centre (CMC) and FNMOC. Estimating analysis errors in a three/four-dimensional variation (3D/4D-Var) DA system is a challenging task. Fisher and Courtier (1995) proposed the Lanczos method for estimating the analysis error variance in the ECMWF DA system. This method produces the analysis error variance estimates by computing the leading eigenvectors of the Hessian matrix, with an obvious drawback that most of the trailing eigenvectors are neglected in the computation. A few calibration schemes for compensating for the loss of the contribution from the less dominant eigenvectors of the Hessian matrix are introduced and tested by Wei *et al.* (2012) in a study carried out for the NCEP 3D-Var DA system.

At the NRL at Stennis Space Center, MS, the Relocatable Circulation Prediction System Version 1.0 (RELO V1.0) is being developed to provide a capability for a rapidly relocatable ocean forecast and data assimilation system for use in operational forecast support for the US Navy's missions (Rowley, 2008, 2010; Rowley *et al.*, 2012). Figure 1 shows a schematic configuration of the RELO system with 32 ensemble members, as used in this article. Basically, the system consists of a forecast model component appropriate for regional to coastal-scale ocean modelling, a data assimilation component, and supporting codes, scripts and databases for domain configuration, data preparation, data assimilation, and post-processing. This system produces real-time forecasts of the ocean state (sea level and 3D temperature, salinity and horizontal currents). Each regular cycle of the system is organized around an analysis that produces an estimate of the ocean state by assimilating newly available observations into the previous best estimate of the ocean state, which is the forecast model output valid at the current analysis time.

The forecast component is the Navy Coastal Ocean Model version 4 (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 NCOM configuration used in the RELO system is fairly flexible, and most of the model configuration parameters are available for the user to define. Default values are assigned to ease model set-up, so most domains can be defined with limited user input. The NCOM in RELO uses a combined sigma−*z* vertical grid with sigma layers near the surface to allow for changes in the surface elevation and a bottom-following vertical coordinate in shallow water, and a switch to *z*-levels below a depth that can be specified by the user. The Arakawa C grid is used in the horizontal direction.

The data assimilation component is the Navy Coupled Ocean Data Assimilation system (NCODA: Cummings, 2005), which was developed at NRL as the ocean data analysis component of the Coupled Ocean Atmosphere Mesoscale Prediction System (COAMPS: Hodur, 1997; Chen *et al.*, 2003). The observational data used for assimilation include satellite sea-surface temperature (SST), satellite altimetry sea-surface height anomaly (SSHA), satellite microwave-derived sea ice concentration, and *in situ* surface and profile data from ships, drifters, fixed buoys, profiling floats, XBTs, CTDs and gliders (see subsection 2.4). The observational data are prepared and processed through the NCODA automated data quality-control system, NCODA QC, which identifies observations with a high probability of error when compared against climatological or model fields with associated variability information. The NCODA analysis employs user-defined thresholds for acceptable error probabilities when accepting data for the analysis and uses the forecasts from NCOM as the background fields in a 3D-Var formulation. Both NCOM and NCODA are used operationally at two of the Navy's operational centres: FNMOC and the US Naval Oceanographic Office (NAVOCEANO), which is located in Stennis Space Center, MS.

The uncertainties from the initial conditions in RELO are represented by the initial perturbations produced by the ET method. Rowley *et al.* (2012) show that perturbing the lateral boundary conditions has only a minor impact on the ensemble

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.

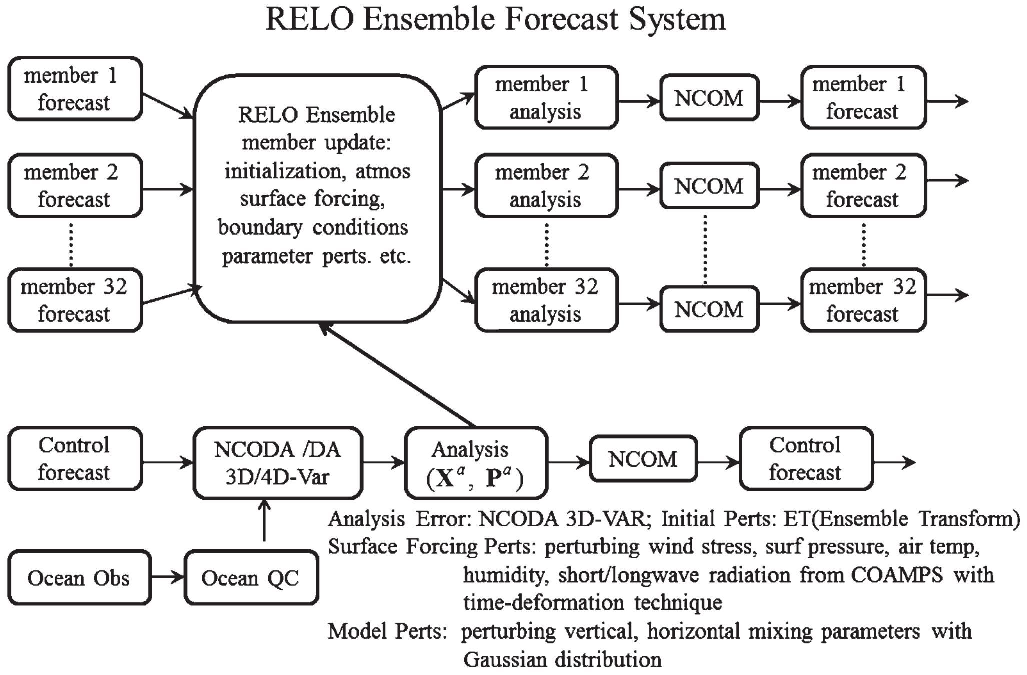
RELO Ensemble Forecast System
member 1 forecast
member 2 forecast
member 32 forecast
RELO Ensemble member update: initialization, atmos surface forcing, boundary conditions parameter perts. etc.
member 1 analysis
member 2 analysis
member 32 analysis
NCOM
NCOM
NCOM
member 1 forecast
member 2 forecast
member 32 forecast
Control forecast
NCODA /DA 3D/4D-Var
Analysis $(\mathbf{X}^a, \mathbf{P}^a)$
NCOM
Control forecast
Ocean Obs
Ocean QC
Analysis Error: NCODA 3D-VAR; Initial Perts: ET(Ensemble Transform)
Surface Forcing Perts: perturbing wind stress, surf pressure, air temp, humidity, short/longwave radiation from COAMPS with time-deformation technique
Model Perts: perturbing vertical, horizontal mixing parameters with Gaussian distribution

Figure 1. Schematic of the US Navy's RELO ensemble prediction system, which contains an ensemble forecast and an ocean DA system. The number of current ensemble members is 32.

spread and is limited to regions close to the boundary. Prior approaches have not accounted for other important uncertainties from the ocean model. The two main sources of model uncertainties include those from the model physical parametrizations and those from dynamics. For a reliable EPS, the ensemble spread should have amplitude and growth rate similar to the ensemble mean error (Buizza *et al.*, 2005; Wei *et al.*, 2006, 2008). Systems that do not account for the model-related uncertainties tend to produce an ensemble spread that grows much slower than the ensemble mean error. As a result, the reliability and the forecast skill and range suffer. Various stochastic parametrization schemes have been developed to account for model-related uncertainties, and these have proven to be effective in atmospheric models at the major world forecasting centres such as ECMWF, NCEP, UKMO and MSC (Buizza *et al.*, 2005; Palmer *et al.*, 2005; Bowler *et al.*, 2009; Charron *et al.*, 2010). For example, four different schemes (addition of isotropic random perturbation fields, multi-parametrization, stochastic physical tendency perturbations and stochastic kinetic energy backscatter (SKEB)) have been used at MSC, three schemes (random parameter, stochastic convective vorticity and SKEB) at the Met Office, one scheme (stochastic parametrization) at NCEP, and two schemes (stochastic perturbed parametrization tendency and SKEB) at ECMWF, are being developed or are in the process of being implemented.

The stochastic forcing, parameter variation, and stochastic kinetic energy backscatter schemes have been studied and developed at the NRL Marine Meteorology Division for atmospheric modelling (Reynolds *et al.*, 2008, 2011a, 2011b). The US Air Force Weather Agency has developed various schemes to account for model errors, including multi-parametrization and perturbation of model parameters and stochastic backscatter stream-function perturbations (Hacker *et al.*, 2011a, 2011b). Bowler *et al.* (2009) described the use of multiple parameters. Each of the parameters evolves in time with a first-order, auto-regressive forcing. Although time-evolving parameters can exploit the impact of an ensemble, Hacker *et al.* (2011b) noted that time-evolving parameters are not necessary to represent parameter uncertainties. Reynolds *et al.* (2011a) used the modified values of different parameters and kept them constant throughout the integrations for all the cycles.

Ensemble prediction methods developed in numerical weather prediction have been more frequently applied to ocean modelling in recent years. Miyazawa *et al.* (2005) used the breeding method with 10 ensemble members to successfully predict a Kuroshio meander position with a lead time of 60 days. The Kuroshio was also studied by Fujii *et al.* (2008) using singular vectors (SVs). Yin and Oey (2007) used 20 members based on the breeding method to study an eddy-shedding event in the Gulf of Mexico. Based on the ensemble, the authors successfully estimated the locations and strengths of the Loop Current and ring for July to September 2005, and found out that bred vectors resemble baroclinic unstable modes in the Gulf of Mexico (GOM). Counillon and Bertino (2009) studied eddy shedding and mesoscale dynamics in the GOM by using a 10-member ensemble based on the HYbrid Coordinate Ocean Model (HYCOM) with 5 km resolution. The initial perturbations are generated by using different values of a parameter in the optimal interpolation DA, while the atmospheric and lateral boundary conditions are perturbed randomly. The Loop Current and eddy fronts from observations were successfully predicted by their ensemble forecast, although the ensemble spread is two to three times smaller than forecast error. O'Kane *et al.* (2011) developed an ocean ensemble prediction system using breeding method to perturb all the model variables, and used all available observations from the operational Ocean Model, Analysis and Prediction System (OceanMAPS) forecasting system at the Australian Bureau of Meteorology to predict the East Australian Current.

In this study, we use the ET method to generate initial perturbations. A time-deformation technique is introduced to generate the surface forcing perturbations from real-time atmospheric fields. All the ocean model variables at all the levels are perturbed in an operational environment with all available observations from NCODA. We focus on accounting for model-related uncertainties by perturbing the key mixing turbulence parameters. In Reynolds *et al.* (2011a), a number of parameters are held constant within each ensemble member for the entire experiment, while several parameters continuously change throughout the integration in the method implemented by Bowler *et al.* (2009). In contrast to these methods used in atmospheric ensembles, the two most important parameters in NCOM responsible for the horizontal and vertical mixing are perturbed with the prescribed statistics imposed at every cycle and held constant during each cycle's integration. Thus, each ensemble member has a different value of these parameters at every cycle, with the changing parameter values reflecting the temporal

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.

variation of the parameters and model uncertainties. Once the integration starts, the parameters are held constant, simulating a complete parametrization package for this member. Of course, we do not expect that perturbing these model parameters will account for all of the model-related uncertainties; additional packages accounting for different sources of model uncertainties will be developed in the future.

Section 2 provides brief descriptions of the ET formulation for initial perturbations, the time-deformation technique to generate surface forcing perturbations from an atmospheric model for RELO, the methodology for perturbing the mixing parameters, the configuration for the RELO ensemble, and the experimental set-up. The major results from using different ensembles with different perturbing schemes are presented in section 3 in separate subsections. A discussion and conclusions are presented in section 4.

2. Methodologies for various perturbations in the RELO system and experimental set-up

Before we provide the experimental results in section 3, we briefly describe the methodologies used for generating the initial perturbations, the surface forcing perturbations for the atmospheric fields, and the schemes accounting for model-related uncertainties. The RELO configuration over the domain of interest and the experiment design will also be described.

2.1. *Initial perturbations for RELO*

In the ensemble transform (ET) method, the forecast perturbations from the previous cycle are first transferred into new perturbations using the ET with the estimated initial analysis error variance, followed by a rescaling using the same initial analysis error variance information. The details and properties of the method are described in Wei *et al.* (2005, 2008), McLay *et al.* (2007, 2008) and McLay and Reynolds (2009). Only a very brief description is provided here. Let

$$\mathbf{Z}^{\mathrm{f}} = \frac{1}{\sqrt{k-1}}[\mathbf{z}_1^{\mathrm{f}}, \mathbf{z}_2^{\mathrm{f}}, \ldots \ldots, \mathbf{z}_k^{\mathrm{f}}],\ \mathbf{Z}^{\mathrm{a}} = \frac{1}{\sqrt{k-1}}[\mathbf{z}_1^{\mathrm{a}}, \mathbf{z}_2^{\mathrm{a}}, \ldots \ldots, \mathbf{z}_k^{\mathrm{a}}], \tag{1}$$

where the n-dimensional state vectors $\mathbf{z}_i^{\mathrm{f}} = \mathbf{x}_i^{\mathrm{f}} - \mathbf{x}^{\mathrm{f}}$ and $\mathbf{z}_i^{\mathrm{a}} = \mathbf{x}_i^{\mathrm{a}} - \mathbf{x}^{\mathrm{a}}(i = 1, 2, \ldots \ldots k)$ are k ensemble forecast and analysis perturbations for all the model variables, respectively. Here $\mathbf{x}^{\mathrm{f}}$is the mean of k ensemble forecasts from NCOM, and $\mathbf{x}^{\mathrm{a}}$ is the analysis from the Navy's independent DA system NCODA. Unless stated otherwise, the lower and upper-case bold letters will indicate vectors and matrices, respectively. In the ensemble representation, the $n \times n$ forecast and analysis covariance matrices are approximated, respectively, as

$$\mathbf{P}^{\mathrm{f}} = \mathbf{Z}^{\mathrm{f}}\mathbf{Z}^{\mathrm{fT}} \text{ and } \mathbf{P}^{\mathrm{a}} = \mathbf{Z}^{\mathrm{a}}\mathbf{Z}^{\mathrm{aT}}, \tag{2}$$

where the superscript T indicates the matrix transpose. For a given set of forecast perturbations $\mathbf{Z}^{\mathrm{f}}$ at time t, the analysis perturbations $\mathbf{Z}^{\mathrm{a}}$are obtained through an ensemble transformation $\mathbf{T}$ such that

$$\mathbf{Z}^{\mathrm{a}} = \mathbf{Z}^{\mathrm{f}}\mathrm{T}. \tag{3}$$

In RELO, the best analysis error variances are derived from NCODA, which is based on 3D-Var and uses all the operational real-time observations. Suppose $\mathbf{P}_{\mathrm{op}}^{\mathrm{a}}$ is a diagonal matrix with the diagonal values being the analysis error variances obtained from the operational NCODA system. The ET transformation matrix $\mathbf{T}$ can be constructed as follows. For an ensemble forecast system, the forecast perturbations $\mathbf{Z}^{\mathrm{f}}$ can be generated by Eq. (1). One can solve the following eigenvalue problem:

$$\mathbf{Z}^{\mathrm{fT}}\mathbf{P}_{\mathrm{op}}^{\mathrm{a}-1}\mathbf{Z}^{\mathrm{f}} = \mathbf{C}\Gamma\mathbf{C}^{-1}, \tag{4}$$

where $\mathbf{C}$ contains the column orthonormal eigenvectors ($\mathbf{c}_i$) of $\mathbf{Z}^{\mathrm{fT}}\mathbf{P}_{\mathrm{op}}^{\mathrm{a}-1}\mathbf{Z}^{\mathrm{f}}$ (also the singular vectors of $\mathbf{P}_{\mathrm{op}}^{\mathrm{a}-1/2}\mathbf{Z}^{\mathrm{f}}$), and Γ is a diagonal matrix containing the associated eigenvalues (λ_i) with magnitude in decreasing order, that is, $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots \ldots, \mathbf{c}_k]$, $\mathbf{C}^{\mathrm{T}}\mathbf{C} = \mathbf{I}$ and $\Gamma = \mathrm{diag}(\lambda_1, \lambda_2, \ldots \ldots, \lambda_k)$.

Let us suppose $\mathbf{F} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots \ldots, \lambda_{k-1})$ and $\mathbf{G} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots \ldots, \lambda_{k-1}, \alpha)$, where α is a non-zero constant, i.e. $\mathbf{G} = \mathrm{diag}(g_1, g_2, \ldots \ldots, g_k) = \begin{pmatrix} \mathbf{F} & 0 \\ 0 & \alpha \end{pmatrix}$ and $\Gamma = \begin{pmatrix} \mathbf{F} & 0 \\ 0 & 0 \end{pmatrix}$.

The new analysis perturbations can be constructed through transformation:

$$\mathbf{Z}^{\mathrm{a}} = \mathbf{Z}_{\mathrm{p}}^{\mathrm{a}}\mathbf{C}^{\mathrm{T}} = \mathbf{Z}^{\mathrm{f}}\mathbf{C}\mathbf{G}^{-1/2}\mathbf{C}^{\mathrm{T}}. \tag{5}$$

It can be shown that the new analysis perturbations in Eq. (5) are centred (sum of all perturbations is zero). In addition, this has the advantage that the ensemble perturbations span a subspace that has a maximum number of degrees of freedom. W08 also showed that the orthogonality of the initial perturbations will increase as the number of ensemble members increases. If the number of ensemble members approaches infinity, then 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.

2.2. *Surface forcing perturbations for RELO using time-deformation technique*

To generate the surface forcing perturbations for RELO, the real-time meteorological fields are obtained from the Navy's meteorological operational centre FNMOC, which produces operational data fields using the Navy Operational Global Atmospheric Prediction System (NOGAPS, for global) and COAMPS (for regional) forecast systems. These operational data are used to produce surface forcing fields for RELO and NCOM by the Navy's ocean operational centre NAVOCEANO. In general, these fields include atmospheric wind stress, surface pressure, short-wave and long-wave radiation, air temperature and specific humidity. Throughout our experiments, the COAMPS atmospheric data fields, which are available every 24 hours, have been used to produce surface forcing for single and ensemble forecasts of the ocean. For a single forecast, the atmospheric forcing at each hour is computed from the linear temporal interpolation of forcing terms from the neighbouring forcing fields produced from COAMPS.

For the EPS, the perturbed surface forcing fields for different ensemble members are also drawn from the same dataset, but by using time-deformation with the random shifting technique in which a number of completely independent random fields are generated every 24 hours with a desired de-correlation length. However, the linear interpolation is computed with randomly shifted time, and the two neighbouring dates are selected randomly among a set of fixed dates with available atmospheric fields. The random shifts are designed so that any interpolated field is not correlated with any other interpolated field 24 hours away. Therefore, the atmospheric forcing for each ensemble member is independent from forcing for the others.

For example, we generate $nr = 4$ or 5 completely random fields in spectral space:

$$\overline{R}(i,j,k) = r(i,j,k)\overline{c}(i,j,k),$$

where i, j are longitude and latitude indices in the NCOM horizontal domain, $k = 1, 2, \ldots nr$, $r(i,j,k)$ is a randomly generated number, and overbar indicates the variables are in spectral space. $\overline{c}(i,j,k)$ is the square root of the eigenvalues of the correlation matrix in spectral space, and it depends on the size

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.

of horizontal domain. The eigenvalues of the correlation matrix in spectral space will determine the correlation length-scale in physical space. These randomly generated fields with a specified correlation length-scale are transferred back to physical space using a Fast Fourier Transform (FFT), i.e.

$$R(i,j,k) = \mathrm{FFT}(\overline{R}(i,j,k)).$$

The amplitudes of these randomly generated fields are adjusted by multiplying predefined coefficients $p(k)$, e.g.

$$R_2(i,j,k) = p(k) * R(i,j,k).$$

The random time shift for any time t is generated by linear interpolation of the adjusted random field,

$$s(i,j,t) = w_1(t) * R_2(i,j,k_1) + w_2(t) * R_2(i,j,k_2).$$

Here, $w_1(t), w_2(t)$ are the weights computed at time t and $k_1,k_2$are the two neighbouring times that are identified by the interpolation subroutine, and $k_1 < t < k_2$.

The surface forcing $F(i,j,t,m)$ for any ensemble member m at time t is produced by linearly interpolating the atmospheric forcing fields from COAMPS, e.g. $a(i,j,t)$, at randomly shifted time $T = t + s(i,j,t)$. Thus,

$$F(i,j,t,m) = W_1(T) * a(i,j,T_1) + W_2(T) * a(i,j,T_2),$$

where $W_1(t), W_2(t)$ are the weights computed at the randomly shifted time T, and $T_1,T_2$are the two neighbouring times that are identified by the interpolation subroutine such that $T_1 < T < T_2$. The whole process is repeated for different ensemble members. Since a new random time shift is generated independently each time, the final surface forcing fields for different ensemble members are independent.

2.3. *Perturbing the horizontal and vertical mixing parameters in NCOM*

In this study, we investigate the impact of varying model parameters on the RELO ensemble spread, reliability, accuracy and forecast skill. We choose two parameters that play critical roles in describing the horizontal and vertical mixing in NCOM (Martin, 2000; Barron *et al.*, 2006). Like other parametrization schemes in atmospheric and ocean models, the schemes described below are attempts to describe phenomena at scales smaller than those resolved by the model. The parametrized formulae are approximate representations of unresolved ocean mixing processes in terms of model variables at the resolved scales.

NCOM has two options for horizontal mixing parametrizations, one based on maintaining a maximum horizontal grid-cell Reynolds number and the other following the Smagorinsky scheme (Smagorinsky, 1963). The relatively simple grid-cell Reynolds number scheme is designed to suppress noise generated by numerical advection and scales the mixing coefficients according to the velocity magnitude. The Smagorinsky scheme scales the rate of mixing according to the horizontal velocity shear and is considered more physically based, and the eddy coefficients are isotropic and independent of coordinate rotation. All simulations in this article utilize the Smagorinsky formulation; the control run scaling has the default value $Smag = 0.1$.

NCOM also has multiple options for vertical mixing parametrization following the Mellor–Yamada Level 2 (MYL2: Mellor and Yamada, 1974; Mellor and Durbin, 1975) and the Mellor–Yamada Level 2.5 (MYL2.5: Mellor and Yamada, 1982) turbulence closure schemes. There is also an option to adjust the mixing of the MYL2 scheme using the Large *et al.* (1994) mixing enhancement in an attempt to account for unresolved mixing processes by extending the mixing of typical oceanic turbulence models above the normal critical Richardson number value. The MYL2.5 scheme provides a prognostic equation to compute the turbulent kinetic energy (TKE), which includes advective and diffusive transport, and uses a second prognostic equation to provide an estimate for the vertical turbulence length scale (Martin, 2000). In contrast, the simplified MYL2 scheme assumes that there is an approximate local balance between shear production, buoyancy production, and dissipation in the TKE equation, which allows the TKE to be calculated algebraically from the mean vertical density and velocity gradients. The turbulence length-scale is estimated from an empirical formula. While the MYL2.5 scheme can be more accurate in high-resolution simulations, where the transport of TKE is significant, the MYL2 scheme is more efficient than the MYL2.5 scheme due to the extra computational cost of the latter's prognostic treatment of the ice and turbulent length-scale. Ocean forecasts produced by the Navy's operational centre typically use the MYL2 formulation due to the overriding importance of efficiently using operational resources. The RELO experiments reported here use the default MYL2 scheme for vertical mixing parametrization, with the ensemble experiments examining variations in *b1_myl2*, the MYL2 parameter that scales the TKE dissipation, which affects the predicted depth of mixing. The default operational value $b1\_myl2 = 15.0$ is used in the control run.

The experiments perturb these two critical parameters in the horizontal and vertical mixing turbulence parametrization, *smag* and *b1_myl2*. Since the unresolved mixing processes are not known, we test their representation by treating these key parameters as stochastic variables. In the absence of *a priori* knowledge regarding the stochastic distribution of the mixing processes, two common distributions, Gaussian and uniform, are evaluated. The first experiment perturbs only the vertical mixing parameter with a uniform distribution, while a second experiment perturbs both of the mixing parameters with uniform distributions. The third experiment, expected to be the preferred case, imposes a Gaussian distribution on both parameters. All the ensemble results from the experiments perturbing these parameters are compared with the control ensemble, a default RELO ensemble with default, unperturbed mixing parameters. To demonstrate the advantage of the ensemble over the single deterministic forecast, a single NCOM forecast with the same resolution is evaluated as well. A summary of these experiments is listed and described in Table 1.

The ranges selected for the random generator in the uniform distributions and the mean and standard deviation in the Gaussian distributions are selected such that the mixing parameters fall within the limits appropriate for NCOM. If any parameter values are too extreme, unphysical values may occur for some variables and cause NCOM to crash. With the values chosen in Table 1, 99.99% of the random values generated by Gaussian distributions for *smag* and *b1_myl2* will be in the range of

$$smag\_range = mean \pm 4 * std = [0.05, 0.2],$$
$$b1\_myl2\_range = mean \pm 4 * std = [15.0, 20.0].$$

Under these distributions, values of these two randomly generated parameters are expected to fall within reasonable ranges and allow NCOM to run smoothly.

2.4. *RELO configuration and experiments*

Our RELO ensemble experiments are carried out for a period of 102 days from 0000 UTC 15 April to 0000 UTC 25 July 2010 with 32 perturbed members plus an unperturbed single run. Additional ensembles with different parameter perturbation schemes as described in Table 1 are run over the same period. The forecast length during the experiments is 72 hours with output every 6 hours. The NCOM horizontal domain covers the whole Gulf of Mexico (GOM) from 98 to 79°W and 18 to 31°N with model grid spacing 3 by 3 km. The number of vertical levels is 49, with 34 sigma levels in the upper ocean and z-levels starting from level 35 to the bottom of the sea. The advantages of this

Published 2013. This article is a U.S. Government work and is in the public domain in the USA.

Table 1. A summary of experiments with various distributions for perturbations of the horizontal and vertical mixing parameters.

Exp. ID	Perturbation Scheme
s:	Single deterministic forecast (NCOM + NCODA 3D-Var, same resolution as ensemble).
c:	Control RELO with 32 members without perturbing any parameters ($smag = 0.1 b1\_myl2 = 15.0$)
v:	perturbing $b1\_myl2 = [15.0, 20.0]$ with a uniform distribution.
h:	perturbing $smag = [0.05, 0.2]$ and $b1\_myl2 = [15.0, 20.0]$ with uniform distribution.
g:	perturbing $smag$ ($mean = 0.125, std = 0.01875$) and $b1\_myl2$ ($mean = 17.5, std = 0.625$) with Gaussian (normal) distribution.

kind of hybrid sigma−*z* coordinate have been discussed in Martin (2000) and Barron *et al.* (2006). The vertical grid extends down to 5500 m. The configuration of the RELO experiments, consisting of 32-member ensembles using the NCOM and NCODA DA system, is depicted in Figure 1.

In all the experiments we have carried out for this article, real observations for this same period from the US Navy's operational centre (i.e. NAVOCEANO) are used as the verifying truth. Before the observations are assimilated into NCOM via NCODA, the data go through a real-time ocean data quality control (QC) process. These data include remotely sensed sea-surface temperature (SST), sea-surface height anomaly (SSHA), and sea ice concentration as well as *in situ* observations. The *in situ* surface temperature observations are collected from ships and buoys while subsurface profiles of temperature and salinity are gathered from eXpendable BathyThermographs (XBT), Conductivity, Temperature, Depth (CTD) instruments, Argo floats and gliders. Observations are also supplemented with synthetic subsurface profiles of temperature and salinity using SST and SSHA via the Modular Ocean Data Assimilation System (Cummings, 2005).

3. Results from RELO ensembles

3.1. RELO ensemble spread and mean distribution

The very basic attributes of any ensemble prediction system are the ensemble mean and ensemble spread. It is expected that the ensemble mean normally outperforms a single deterministic forecast in terms of the root mean square (RMS) error and the absolute error. The ensemble spread strongly influences the range, reliability, and sharpness or resolution of the EPS. Our results from different ensemble configurations show that the spread differences among different perturbation schemes (v, h and g) are visually small; quantitative differences are shown in later sections. Thus, in this section we concentrate on the comparisons between the control ensemble (c) and one of the perturbed parameter ensembles (g).

The DeSoto Canyon area spans a range of dynamical processes including deep-water mesoscale structures that vary on long timescales, shelf processes that are more rapidly varying due to wind forcing and strong interactions with the coast due to freshwater outflow providing buoyancy forcing. This region is impacted by diverse influences of wind-driven currents and circulation associated with some eddy-like features that are related to the Loop Current. The wind stress and eddies are combined, and they can produce a complicated pattern of currents within the DeSoto Canyon. This region is also of great interest for the many active oil and gas explorations. The initial examination of the ensemble spread indicates the range of uncertainty captured under these varying dynamical regimes. For these reasons, we choose a location P = (88.39°W, 28.74°N) at 0000 UTC on 20 April 2010 to show a few snapshots of the ensemble structure which is always helpful to introduce our ensemble before we move on to more detailed studies of performance in the later sections.

In order to get a glimpse of the vertical distribution at P, the ensemble mean (left) and spread (right) at 72 hours forecast from 0000 UTC 20 April 2010 for ensemble c are shown in Figure 2. The ensemble mean temperature shows that the mixed layer continues down to about 50 m, then the temperature gets cooler indicating the beginning of the thermocline layer until below 500 m. Salinity increases with depth due to the surface mixing process with fresh river water on the surface and tropical precipitation. Both components of the water current are larger at the surface due to the mixing process induced from wind and solar radiation. The distributions shown in the right panels indicate that the spread is smaller for deeper water for every variable we have computed, and does not change much after 200 m. One of the main reasons is the lack of observations in deeper water as explained further in later sections. The distributions for the other parameter-perturbed ensembles show similar structures, and the differences are small (not shown).

Figure 3 shows ensemble plumes originating at the surface from P, including temperature, salinity, and the velocity components *u* and *v*, respectively. Comparisons between the left and right columns show that the impact of perturbing the mixing turbulence parameters is small for all variables for ensembles originating from this location. The ensemble mean and median are very similar in all cases. It is a common practice that the ensemble mean is normally used to predict forecast events, as it has been shown that the mean from a reliable ensemble performs better than a single deterministic forecast (Toth and Kalnay, 1993, 1997).

The equivalent ensemble plumes originating from the same location at 1500 m are computed, but not shown. The difference is that the spreads at this depth are even smaller compared with those at the surface. This is mostly a consequence of the small analysis error variance produced by the NCODA DA system in deeper water. When the analysis error variance used in the ET ensemble generation is small, the initial ensemble amplitude will be small. Little if any growth as a function of forecast lead time is evident in the ensemble spreads in all cases with all variables. This evidence can be observed more clearly in Figure 4, showing the ensemble spread as a function of the forecast lead time and depth for ensembles c and g. Since there is so little variation below 200 m, we restrict the plots to the upper 200 m where spread variations are better resolved by the colour range. For temperature and salinity, the largest spread is not at the surface but in a 50 m thick range centred at 50 m depth.

As expected, the spreads of *u* and *v* are largest near the surface due to the atmospheric forcing perturbations introduced through the time-deformation technique, which is consistent with the results in Figure 2. However, the impact from this atmospheric forcing propagates downward very slowly as the forecast lead time increases, and only a very small impact can be seen within 72-hour forecasts. Again, similar to the other variables, the spreads for *u* and *v* near and below 200 m are small and show little variation. Because the assimilation data stream has very few observations at these depths, NCODA has little new information to produce anything but very small analysis increments. Since the analysis error variance estimated from NCODA is small, the initial ensemble spread at these depths is small as well.

The ocean dynamics of the GOM are strongly influenced by the Yucatan Current inflow which forms the powerful Loop Current which is connected to the Florida Current. From the Florida Straits, it travels to the Atlantic to form the Gulf Stream. Ensemble spreads shown in Figures 2 and 4 indicate the largest variability is around 50 m depth for both temperature and salinity. It is interesting to compare the dynamics between the surface and 50 m level. The horizontal contrasts over the Gulf between distributions at 0 and 50 m are shown in Figures 5 and 6 with 0000 UTC 20 April 2010 snapshots of the control ensemble mean and spread for temperature, salinity, *u* and *v*. The ensemble mean shows warmer surface temperature from the Caribbean Sea through the Yucatan Channel and along the western edge of the Loop Current, while regions of high temperature at 50 m are found in the midst of the Loop Current extension and approaching the

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.

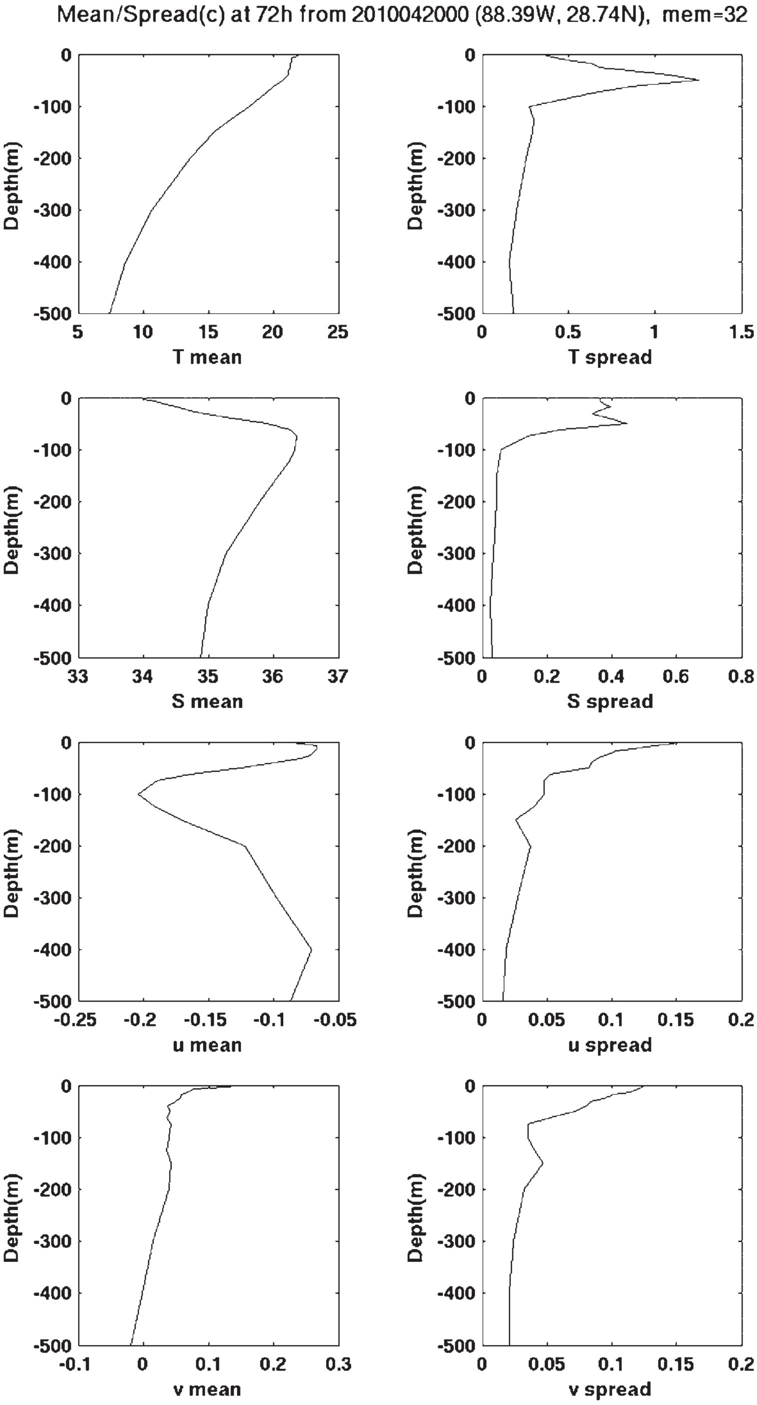

Figure 2. Ensemble mean (left) and spread (right) at 72 hours forecast from 0000 UTC 20 April 2010 at (88.39°W, 28.74°N) for ensemble c. Panels from top to bottom are for temperature (°C), salinity (PSU), u and v (m s^{-1}).

shelf break in the central western Gulf. Maxima in the ensemble spread shows relatively high uncertainty in surface temperature north of the Loop current, about 200 km south-southeast of the Mississippi River mouth. Temperature variability at 50 m is similarly high along the northern east side of GOM and near the entrance to the Florida Straits.

The surface salinity is distributed relatively evenly except near the Mississippi and Atchafalaya river outlets, where large freshwater river input mixes with the sea water. This leads to smaller salinity values near the coasts of Louisiana and Mississippi. It is also in this mixing area where the largest surface salinity variations are located. Relatively high salinity variations are more widespread at 50 m, with the largest variations near the north of GOM. The magnitudes and variations of the velocity components are larger on the surface than at 50 m. The ensemble mean shows clear signals of the Yucatan Current and Florida Current, which lead to the strong Gulf Stream system in the North Atlantic. The ensemble spread indicates large variations in the surface velocity within 200 km of the Louisiana coast. At 50 m, variations in u and v are smaller than at the surface, thanks to the atmospheric forcing perturbations implemented in the RELO ensemble.

In order to see the impact on the ensemble forecasts from the different mixing parameter perturbation schemes as described in subsection 2.3, we look at the differences between the parameter-perturbed and control (unperturbed) ensembles at the surface (Figure 7) and at 50 m (Figure 8) at a forecast lead time of 72 hours starting from 0000 UTC 20 April 2010. These depths are selected because, as demonstrated in Figure 4, the largest spread in velocity is found at the surface, while the largest spreads in temperature and salinity are more often found near 50 m depth. Spreads are small below 200 m for all the ensembles.

The impacts on the mean and spread of the surface temperature are mostly located in the northeastern GOM from Louisiana to Florida. The impact patterns from v and h, in which the perturbed

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.

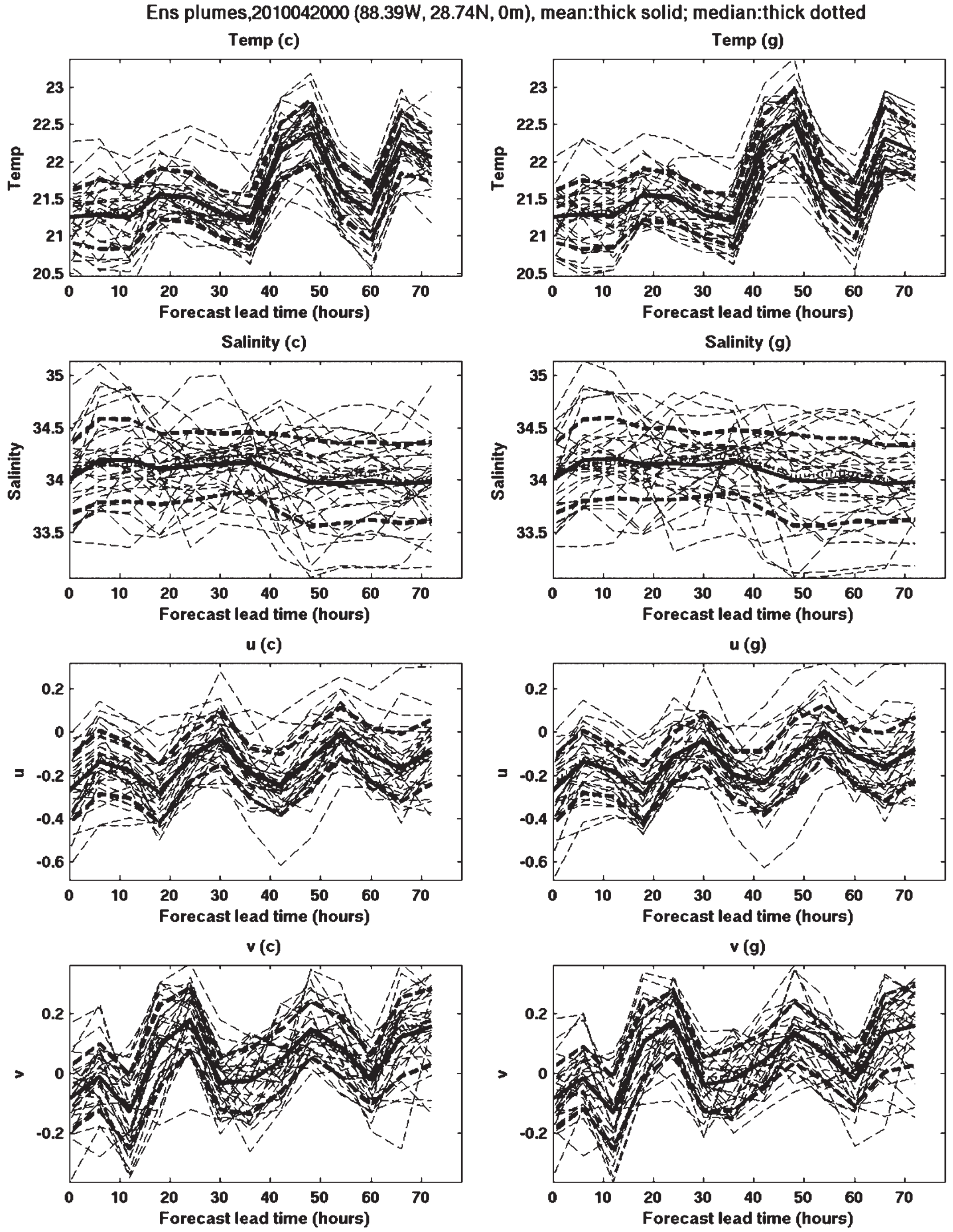

Figure 3. Ensemble plumes from 0000 UTC 20 April 2010 at (88.39°W, 28.74°N, 0 m) for ensembles c (left) and g (right). Panels from top to bottom are for temperature, salinity, *u* and *v*. Dashed lines are for the 32 individual ensemble members; the ensemble mean and median are indicated by thick solid and dotted lines. The thick dashed lines indicate ensemble mean +/− one standard deviation.

mixing parameters are distributed uniformly and affect either the vertical or the horizontal and vertical mixing, respectively, are different from the impact of g with the normally (Gaussian) distributed horizontal and vertical mixing parameters (row 1). Cases v and h exhibit similar spatial characteristics with increased mean SST, higher in case v, off the continental shelf in the northeastern Gulf and reduced SST, cooler in case h, from the open ocean to the coast in the central Gulf. Case g exhibits a different pattern of changes, with little change in the northeastern Gulf and a dipole of warming and cooling in the central northern Gulf that is in reverse phase from the changes in cases v and h. While again concentrated in the northeast quarter of the GOM, the anomalies in surface temperature spread are sprinkled indistinctively and do not show clear differences among the perturbation schemes. At 50 m (Figure 8), the mean temperature is lower than at the surface, as expected, but the spread at 50 m is larger. Coherent patterns in the temperature spread differences at 50 m are difficult to identify among runs using these three perturbation schemes. These differences will be evaluated more quantitatively in later sections.

For the surface salinity, the ensemble mean is the smallest, and the spread is the largest near the Mississippi River and more broadly near the major freshwater inputs from Louisiana and Mississippi along the northern coast of the GOM. The differences among the different ensemble schemes are also largest near the northern coast. A clear difference at 50 m is that a relatively low salinity is broadly distributed across most of the northern half of the GOM with the maximum salinity spread concentrated in the northern Gulf from the Mississippi River outlet to the west Florida Shelf. The ensemble schemes h and g, which have both the horizontal and vertical mixing parameters perturbed, have slightly larger impacts on the salinity ensemble mean and spread than does scheme v, which perturbs only one parameter. The ensemble mean of *u* shows clearly the Loop Current and Florida

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.

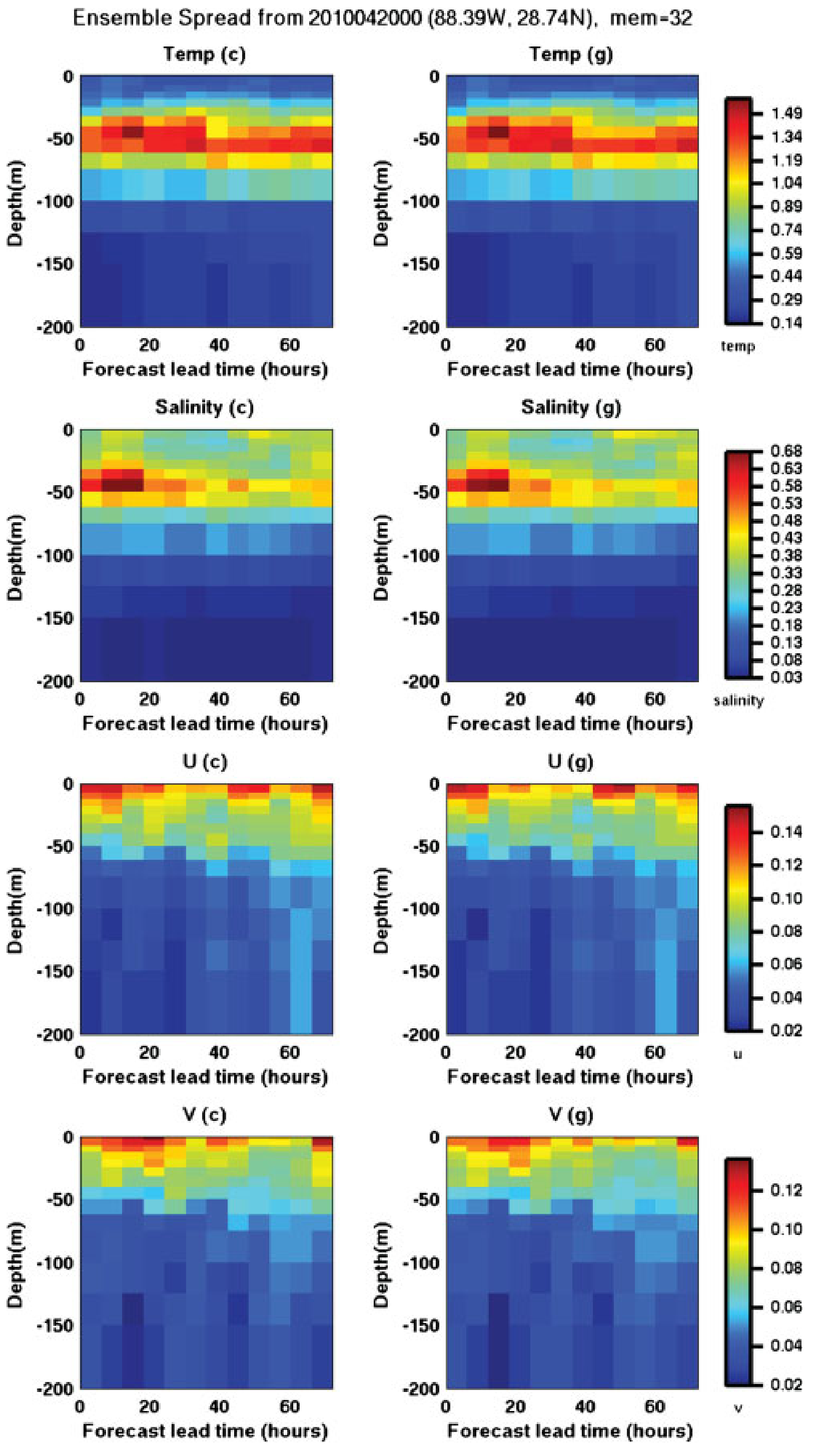

Figure 4. Ensemble spread as a function of forecast lead time and water depth from 0000 UTC 20 April 2010 at (88.39°W, 28.74°N) for ensembles c (left) and g (right). Panels from top to bottom are for temperature, salinity, *u* and *v*.

Current at both the surface and at 50 m. However, regions with a large ensemble velocity spread are more broadly distributed at the surface than at 50 m. The largest variations in the ensemble mean and spread are concentrated near the Florida Straits. While the various ensemble schemes do not produce large differences in the ensemble means and spreads, schemes h and g, with two mixing parameters perturbed, have slightly larger impact than alternatives.

3.2. *Forecast accuracy and ensemble spread*

The goals of ensemble forecasts are to predict the ocean state with the ensemble mean and to predict the forecast uncertainty with the ensemble spread. We quantify the forecast error as the RMS difference between the forecast ensemble mean and subsequent observations corresponding to the forecast time. Forecast accuracy generally decreases as the lead time of the forecast increases; this change in accuracy is represented as a growth rate in the RMS forecast error. The estimated uncertainty of a forecast is proportional to the ensemble spread, and it can be shown that an ideal ensemble should have an ensemble spread that has a similar magnitude and growth rate to the ensemble RMS error (Wei and Toth, 2003; Buizza *et al.*, 2005; Wei *et al.*, 2006, 2008). One of the main reasons for perturbing the mixing parameters in RELO is to account for model-related uncertainties and their contributions to ensemble spread. Without representations of model-related uncertainty, the ensemble will be under-dispersive and underestimate the true forecast uncertainty. If important sources of uncertainty are neglected, the reliability of forecast uncertainties will be reduced.

Ocean dynamics vary horizontally and vertically in the GOM as we can see from Figures 2–6. The temperature does not

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.

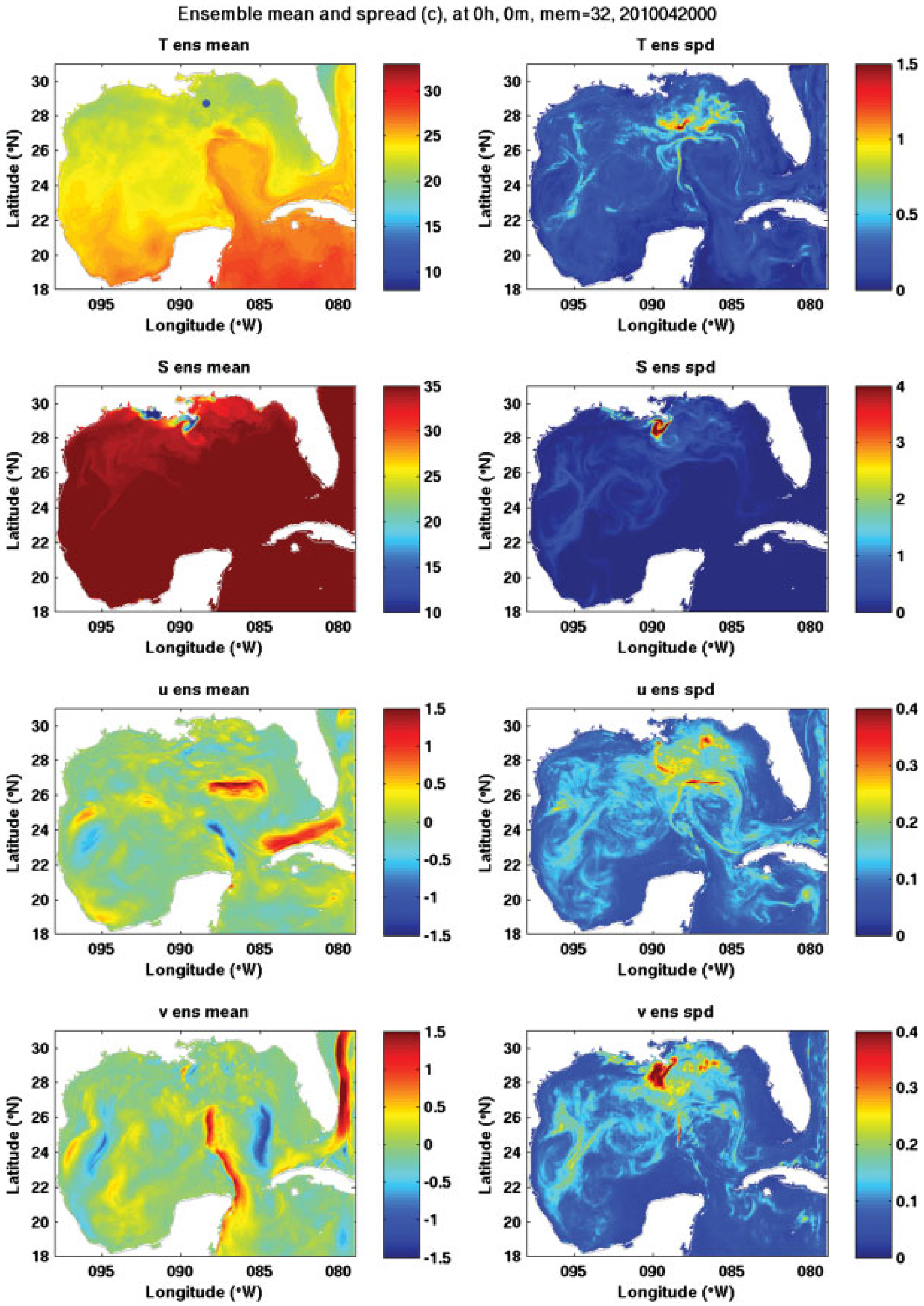

Figure 5. Control ensemble mean (left) and spread (right) at 0000 UTC 20 April 2010 at the surface for temperature, salinity, *u* and *v* (from top to bottom). (88.39°W, 28.74°N) is marked with a black dot in top left panel.

change much in the mixed layer until about 50 m as shown in Figure 2. Water becomes cooler from 50 m, indicating the start of the thermocline layer. Due to the complicated mixing processes with fresh river water, air–sea interactions such as highly variable wind-driven currents, the tropical precipitation and solar radiation on the surface, both components of water current have largest spread, while salinity has smallest spread at the surface. But salinity increases with depth until about 100 m. The interior layer is mostly controlled by internal mixing processes and shear between geostrophically balanced flows. The ensemble spreads are generally larger above 200 m for most variables. In addition, the density of observations for the evaluations is much larger near the surface than in the interior. To better identify different dynamics in different domains, evaluations in the following sections are carried out in three different domains, namely surface (about the upper 1 m), ocean interior over a range from 50 to 200 m, and the whole domain.

To evaluate RELO NCOM forecast accuracy, we plot in Figure 9(a)–(c) the RMS error of the ensemble means for temperature as a function of lead time for the control and parameter-perturbed ensembles. The RMS error is the difference between the forecast and the truth embodied in unassimilated observations valid during the forecast interval; thus it is a direct measure of forecast accuracy. To increase the statistical significance, all the RMS values are averaged over the 102 days from 0000 UTC 15 April to 0000 UTC 25 July 2010 and in various observation spaces. The RMS errors of the ensemble means for salinity as a function of lead time for the control and parameter-perturbed ensembles are plotted in Figure 10(a) and (b) for averages over the whole observation space and for the layer between 50 and 200 m, respectively.

It is very clear that all the ensemble means of both temperature and salinity have lower RMS errors than the single deterministic forecast at all lead times. This applies to the control and different parameter-perturbed schemes, although the differences between

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.

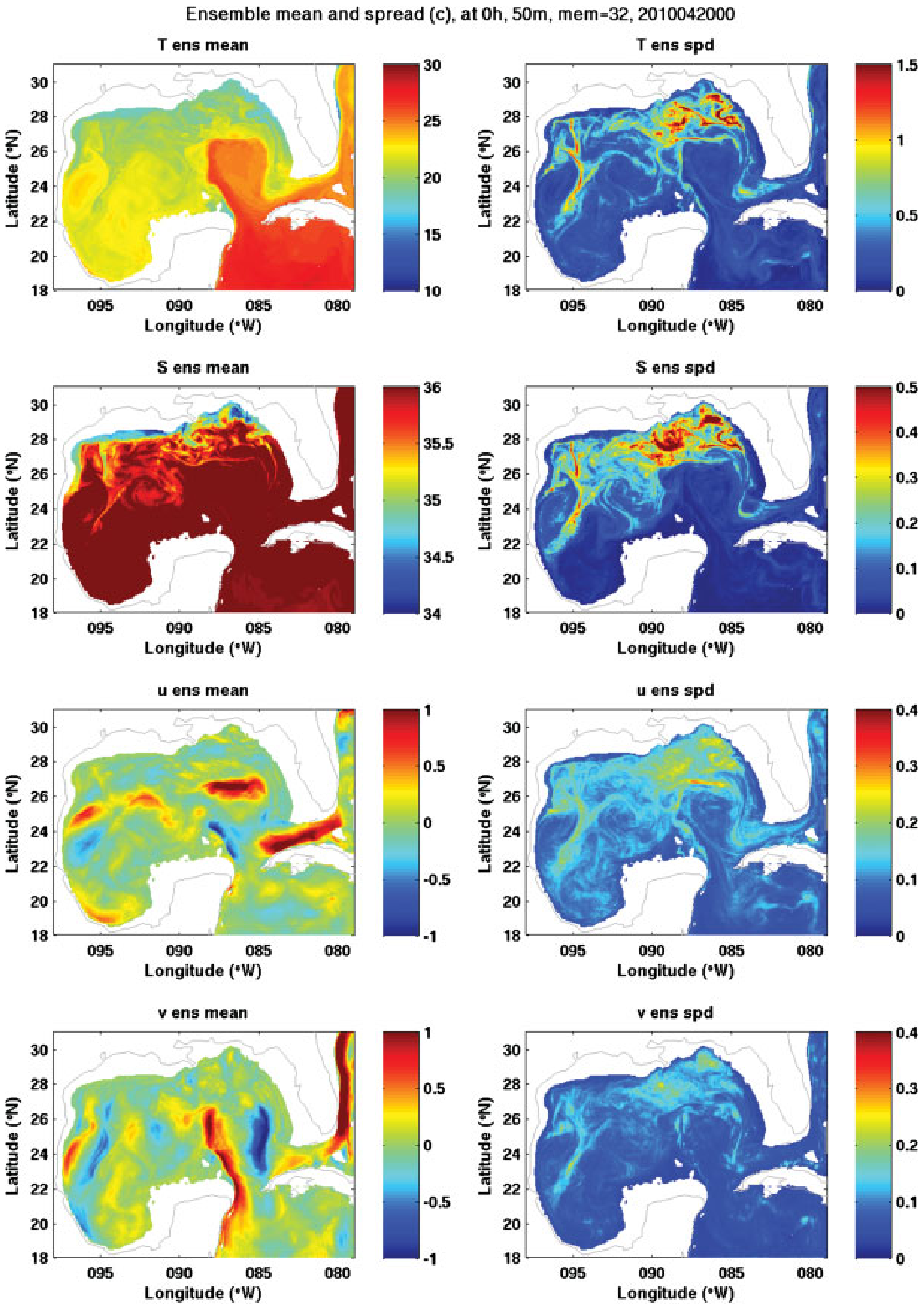

Figure 6. Same as Figure 5, but at 50 m depth.

the different schemes are small. There is no doubt that any of the ensemble means of temperature and salinity offer more accurate forecasts than the single deterministic forecast.

Forecast uncertainty is a function of uncertainty in the initial state, uncertainty in quantifying the external forces that modify the ocean state, and uncertainty in the model representation of the physical processes of the ocean state. The purpose of introducing the perturbed mixing parameters is to account for the model-related uncertainties in dealing with unresolved horizontal and vertical mixing processes. While the unresolved subgrid processes are parametrized in these examples with simple schemes, as described in subsection 2.3, a more realistic approach would recognize that the mixing is stochastic by nature. Thus, introducing random perturbations in these parameters should improve the representations of model uncertainty and thereby increase the accuracy of the ensemble forecasts and spread.

The ensemble spread between 50 and 200 m for temperature as a function of forecast lead time is shown in Figure 9(d). It is also clear that the ensemble with two randomly perturbed parameters using the Gaussian distribution has the largest spread increment (g in solid), with ensemble h (perturbing both the horizontal and vertical mixing parameters with a uniform distribution) second. This is also the case for salinity (not shown).

In most operational ensemble systems in both meteorology and oceanography, the ensemble spreads are generally under-dispersive and grow more slowly than the ensemble mean error, particularly in ensembles that do not account for model-related errors (Wei and Toth, 2003; Buizza *et al.*, 2005; Wei *et al.*, 2006, 2008; McLay *et al.*, 2007, 2008, 2010; Bowler *et al.*, 2009; McLay and Reynolds, 2009; Reynolds *et al.*, 2011a, 2011b). This is also the case in our RELO ensemble system, where the ensemble spread is much smaller than the ensemble RMS error. The small initial ensemble spread is a consequence of the underestimation of the analysis error variance computed from the 3D-Var NCODA system. This results in a smaller initial ensemble spread during

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.

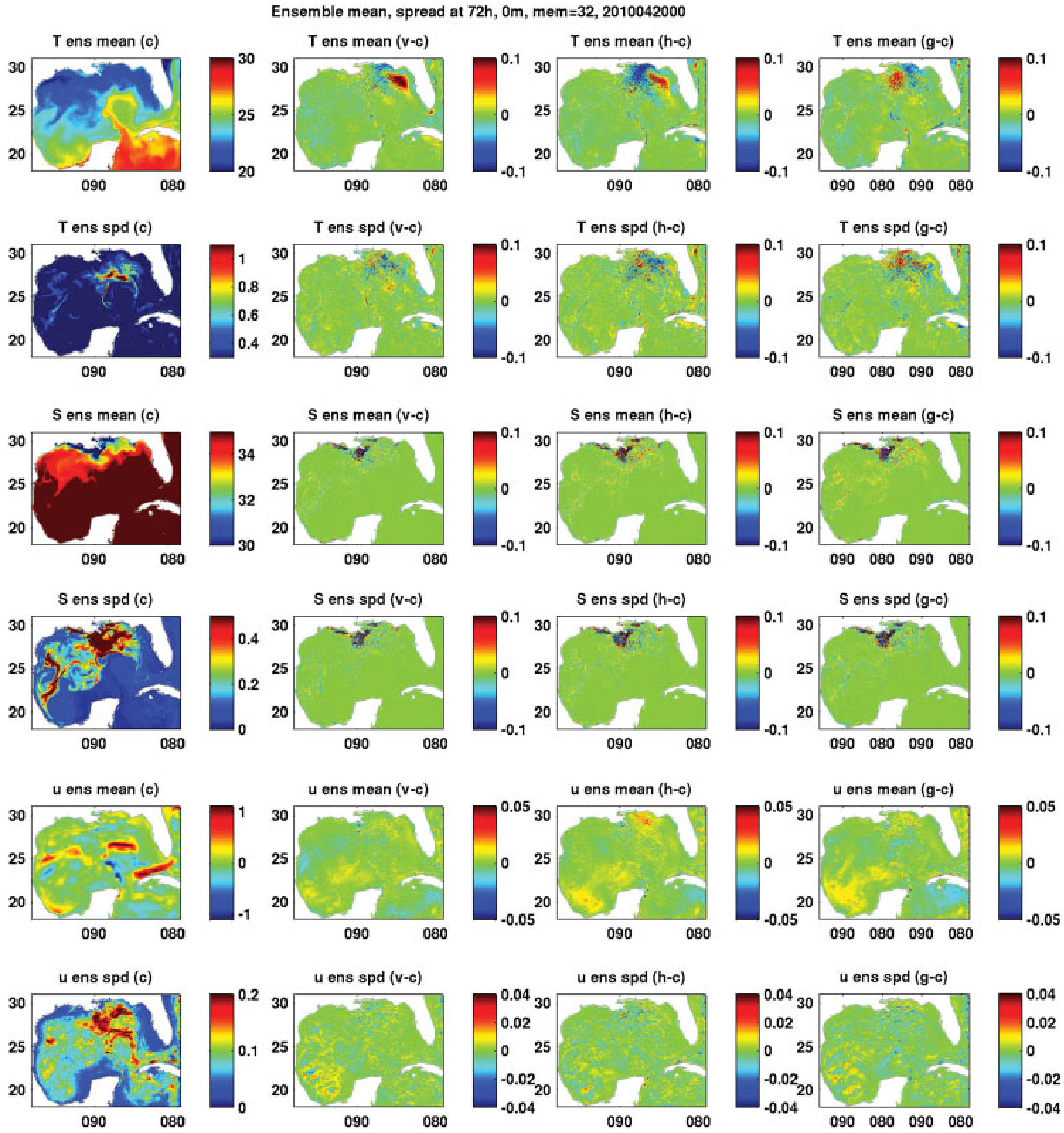

Figure 7. Left column from top: the control ensemble surface mean and spread at 72 hours forecast lead time from 0000 UTC 20 April 2010 for temperature, salinity and *u*. From the 2nd to 4th column: the differences between the control and parameter-perturbed ensembles v, h and g, respectively.

the initial perturbation generation process. The future plans for improving this are discussed in the Discussion and conclusions section.

3.3. *Ensemble spread and reliability*

As discussed in the previous subsection, the ensemble spread is important for the whole forecast system. The spread of an ensemble forecast varies in space and time, and should capture the forecast errors as a function of the forecast lead time. It is expected that a reliable ensemble spread should have a magnitude similar to the ensemble mean error and a growth rate similar to the forecast error. An ensemble spread that is too small will miss some important dynamic events, especially extreme ones, while an ensemble spread that is too large will make the ensemble less sharp and less reliable, with lower resolution. In this section, we quantify the quality of the spread of our 32-member ensembles by computing a spread–reliability diagram with 20 bins. Similar methods have been used in Majumdar *et al.* (2002), Wei *et al.* (2006), and Leutbecher and Palmer (2008). The steps for doing this are outlined in appendix A.

Figure 11 shows ensemble forecast spread–reliability diagrams for temperature using observations as the truth for various lead times and observation spaces. The equivalent spread diagrams for salinity is shown in Figure 12. Again, to increase their statistical significance, all the values are averaged over a large number of samples within the respective observation spaces from 15 April to 25 July 2010. As the ensemble spread is supposed to represent the forecast uncertainty, the spread–reliability curve over such a large sample should coincide with the diagonal line denoting equality between the ensemble spreads and ensemble errors. The spread–reliability for temperature in Figure 11 shows that all the ensemble spreads are too small or under-dispersive for all the ranges. All the ensembles are overconfident and under-predict the forecast error variance, which is consistent with what is seen in Figure 9.

In comparison with the temperature, the ensemble salinity forecasts are more reliable, especially for 24-hour lead time evaluations over the whole observation space (top left in Figure 12). When the verification is restricted to the observation space between 50 and 200 m, the spread–reliability reveals a transition near the 0.08 ensemble spread as the ensembles under-dispersive, under-predicting forecast error

Published 2013. This article is a U.S. Government work and is in the public domain in the USA.

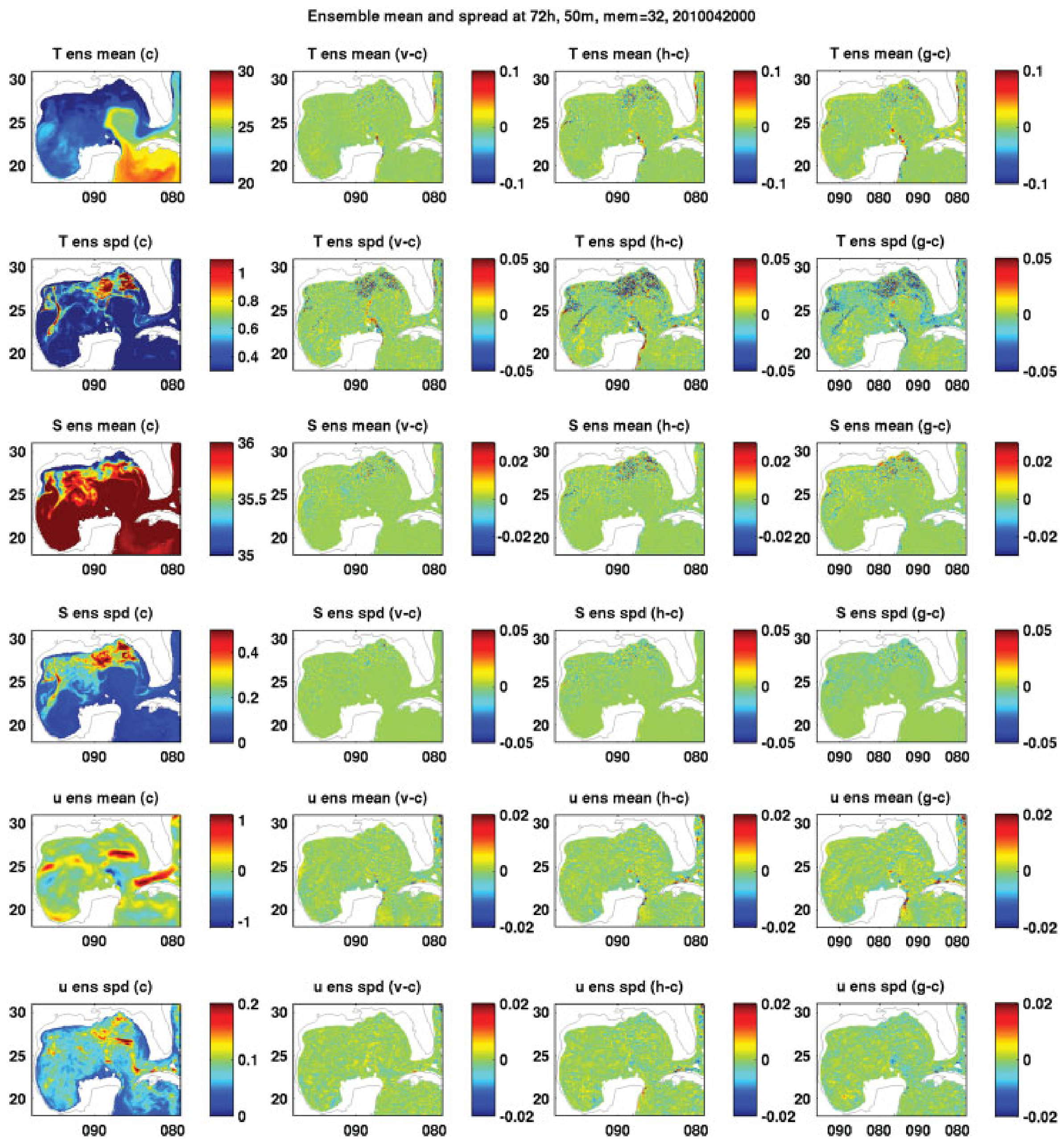

Figure 8. Same as Figure 7, but for 50 m depth.

variance (overconfident) at the smaller ensemble spreads to over-dispersive, over-predicting (underconfident) the forecast error variance at the larger ensemble spreads. The reliability differences are small for different parameter-perturbed ensembles for both temperature and salinity. This is partly due to the small number of bins (20) we have chosen. It is expected that larger differences would be seen if a larger number of bins were used.

The spread reliability and consistency can also be assessed using the rank histogram or Talagrand histogram (Talagrand *et al.*, 1997; Wilks, 2006). The idea, interpretation, and computing steps are described in detail in appendix B. In this article, we compute the rank histograms for both temperature and salinity at three forecast lead times, namely 24, 48 and 72 hours and in three vertical domains, including the whole observation space, the space between 50 and 200 m, and, for temperature only, the surface. Since previous versions of figures with three forecast lead times in one row were too hard to read, we choose to show only the histograms with 72-hour forecasts. Figure 13 shows the rank histograms for temperature at a lead time of 72 hours in various observation spaces. The same is shown for salinity in Figure 14.

First of all, one notices that there are only small differences in terms of the consistency indices among the different parameter perturbation schemes for both temperature and salinity over all three domains, and these small differences are not significant. This indicates that the consistency index of the rank histogram is not very sensitive to small variations in the ensemble spread. The spread consistency decreases (index value increases) as a function of the forecast lead time. This is particularly true for temperature in the whole water column and at the surface (not shown). For temperature, the spread in the observation space between 50 and 200 m (middle panel in Figure 13) is much more consistent than in the whole observation space and at the surface for all three forecast lead times. Overall, the ensemble spread of salinity (Figure 14) shows more consistency than the temperature spread (Figure 13) in these two observation spaces. This is consistent with the spread–reliability diagrams shown in Figures 11 and 12. In addition, the salinity spread shows more consistency in the space between 50 and 200 m than in the whole observation space at each of the three forecast lead times.

3.4. Ensemble forecast skill

Having studied the ensemble forecast accuracy using observations as truth in comparison with a single deterministic forecast and the ensemble spread reliability in previous sections, we look into

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.

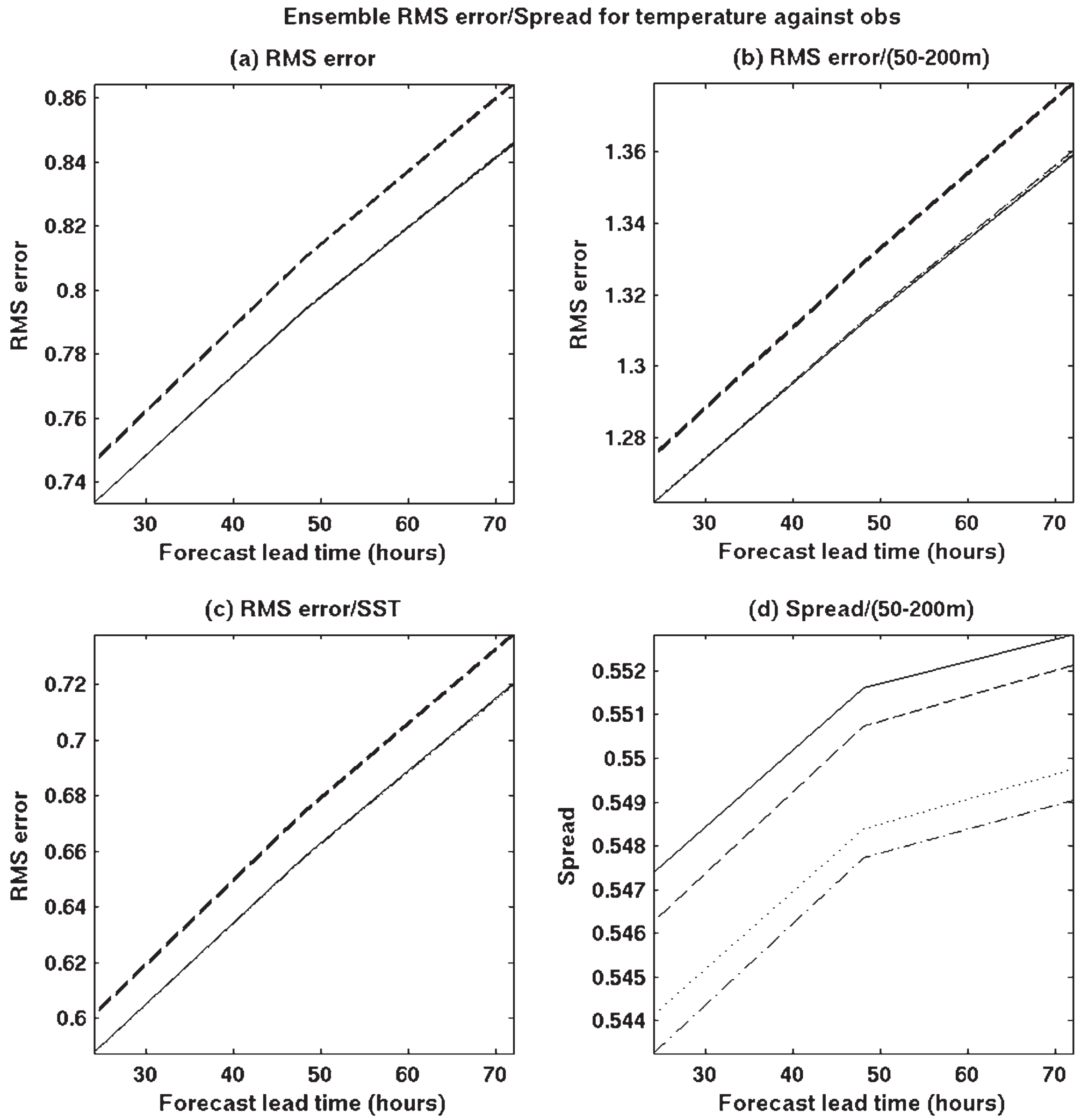

Figure 9. RMS errors of ensemble means and single forecast (thick dashed) for temperature averaged over (a) whole observation space, (b) 50–200 m and (c) surface. All RMS values are computed against observations. (d) Ensemble spread for temperature from different ensembles as a function of forecast lead time averaged over 50–200 m. All RMS and spread values are averaged from 0000 UTC 15 April to 0000 UTC 25 July 2010. Ensembles schemes are indicated in dotted (c), dash-dot (v), dashed (h), and solid (g).

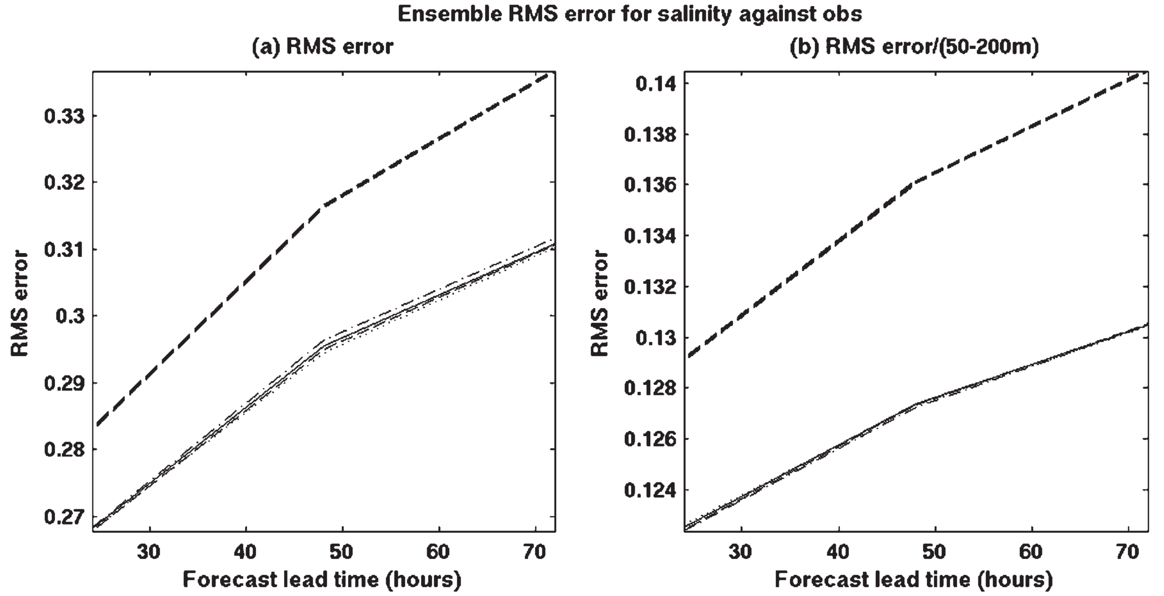

Figure 10. RMS errors of ensemble means and single forecast (thick dashed) for salinity averaged over (a) whole observation space and (b) 50–200 m. All RMS values are computed against observations, and averaged from 0000 UTC 15 April to 0000 UTC 25 July 2010. Ensembles schemes are indicated in dotted (c), dash-dot (v), dashed (h), and solid (g).

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.

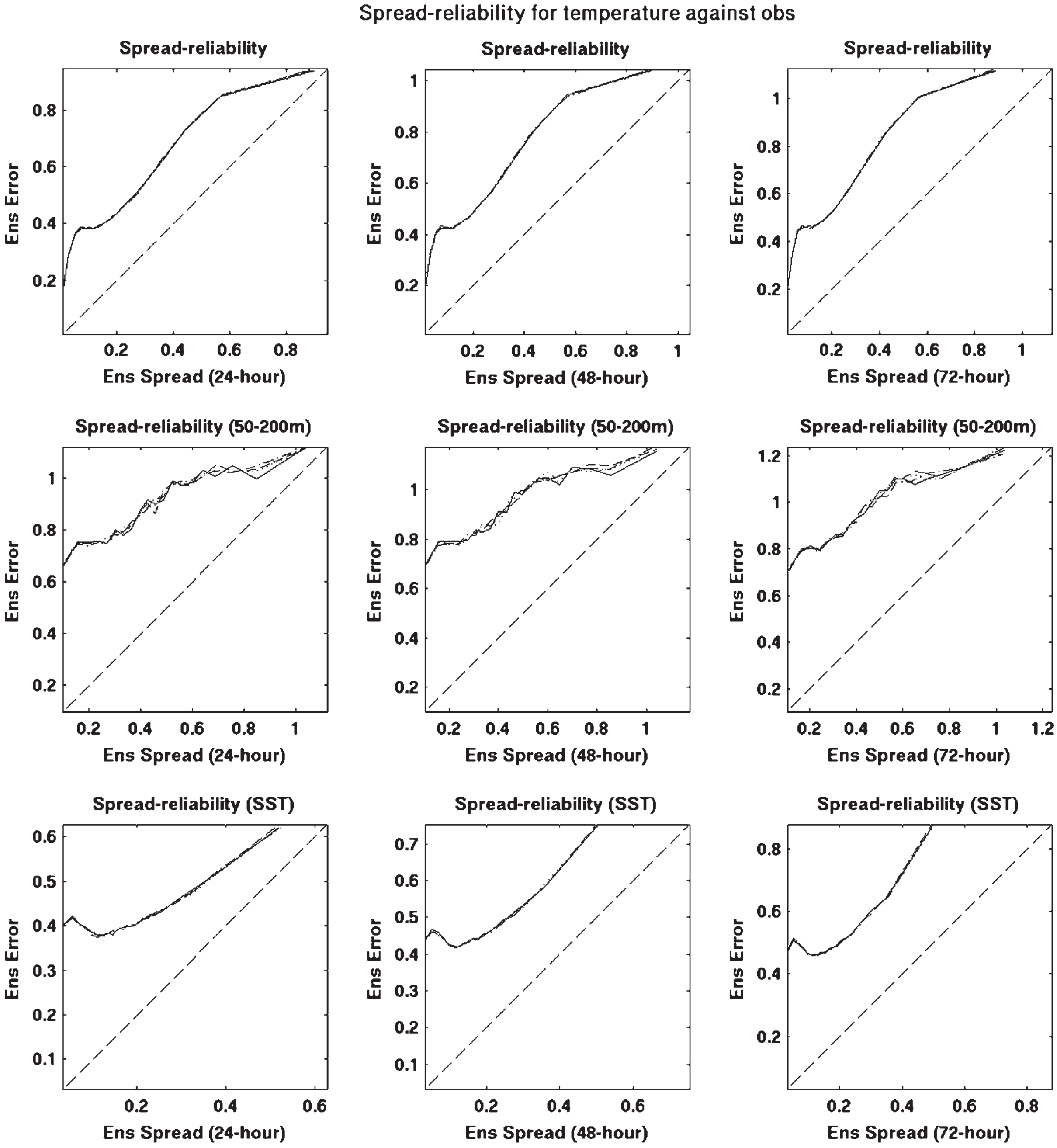

Figure 11. Ensemble forecast spread–reliability diagrams for temperature using observation as truth for lead times of 24, 48 and 72 hours (from left to right). All values are averaged from 0000 UTC 15 April to 0000 UTC 25 July 2010, and over the whole observation space (1st row), 50–200 m (2nd row) and surface (3rd row). Ensembles schemes are indicated in dotted (c), dash-dot (v), dashed (h) and solid (g).

the forecast skill of various ensembles and compare them with the single forecast. In order to quantify the ensemble forecast skill, we compute the anomaly correlation (AC) of the ensemble mean and the single forecasts. Again, the observations are used as the truth. The AC is preferred to a simple correlation coefficient (CC), which is defined as the correlation between the forecast and the observed values. The CC does not take forecast bias into account. It is possible for a forecast with large error to have a high CC value. It is well established practice to use the AC with climatology as a reference to account for seasonal variation (Wilks, 2006). The AC for any forecast variable *f* at a particular forecast lead time is defined as the correlation between the forecast and observation anomalies with respect to climatology, i.e.

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

where *c*, *y* are the climate data and observation fields respectively at the same verifying locations as the forecast; the overbar indicates the geographical mean over the verifying space. Therefore, the AC measures similarities in the pattern of departure (or anomalies) from climatology, thus it is a pattern correlation and is regarded as a skill score relative to climatology. It is arguably the most commonly used metric in numerical weather prediction centres (Buizza *et al.*, 2005). The climatological data we used were obtained from the Navy's ocean operational centre NAVOCEANO. More details can be found in Carnes *et al.* (2010).

Shown in Figure 15 are the AC values of the temperature averaged over various observation spaces and the same 102-day period. It is noticeable that the differences among the different ensemble schemes are very small. However, it is clear that all the ensemble means from the different parameter perturbation schemes are more skilful than the single forecast for all the lead times in all three spaces. To quantify the advantages of the ensemble mean over the single forecast, we introduce the extended forecast time (EFT) by the ensemble mean for different forecast lead times, e.g. the AC values for the ensemble mean AC_e and single forecast AC_s as functions of time can be described as $AC_e = f_e(t_e)$ and $AC_s = f_s(t_s)$. For two forecast systems having

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.

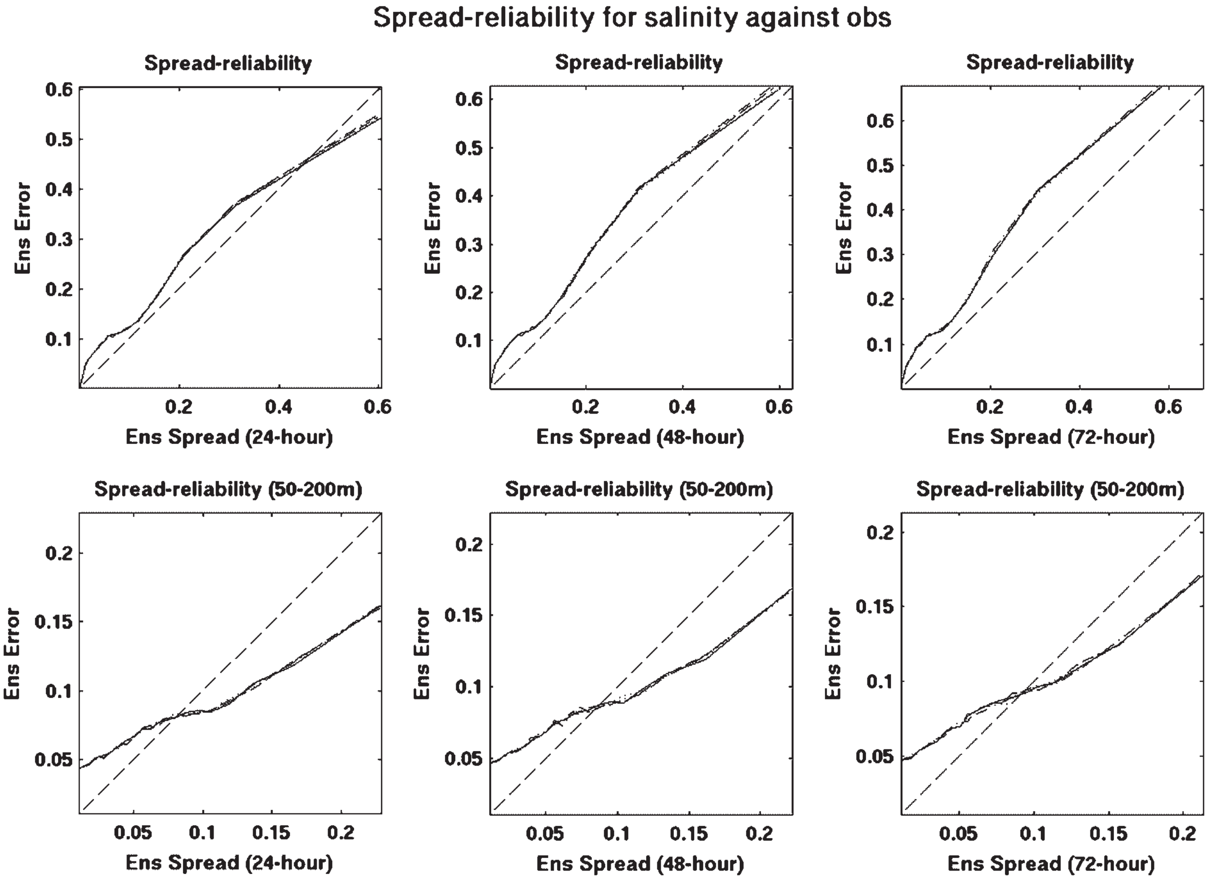

Figure 12. Same as Figure 11, but for salinity, and the values are averaged over the whole observation space (1st row), 50–200 m (2nd row).

an equal AC score, the forecast lead times are different since the ensemble mean is more skilful than the single forecast. We define the difference between these two lead times with the same AC score as the EFT by the ensemble mean. The lead time for the ensemble mean is

$$t_{\mathrm{e}} = f_{\mathrm{e}}^{-1}(AC_{\mathrm{e}}) = f_{\mathrm{e}}^{-1}(AC_{\mathrm{s}}) = f_{\mathrm{e}}^{-1}(f_{\mathrm{s}}(t_{\mathrm{s}})).$$

The EFT can be computed as

$$EFT = t_{\mathrm{e}} - t_{\mathrm{s}} = f_{\mathrm{e}}^{-1}(f_{\mathrm{s}}(t_{\mathrm{s}})) - t_{\mathrm{s}}.$$

For forecast lead time up to 72 hours, both AC functions $(f_{\mathrm{e}}(t_{\mathrm{e}}), f_{\mathrm{s}}(t_{\mathrm{s}}))$ are approximately linear. With this approximation, the EFT is easily estimated for the ensemble (g) and the single forecast using the above formula. The corresponding EFT is depicted as a function of the forecast lead time on the right panel. The advantage of the ensemble mean over the single forecast is clear. The EFT for temperature is about 4–6.5 hours in the whole observation space. This can be understood equivalently as the ensemble having the same skill score as the single forecast for a forecast that is 4–6.5 hours longer. The EFT hours for temperature for the space between 50 and 200 m and for the surface are 5–7.5 hours and 3–5 hours, respectively.

Figure 16 is similar to Figure 15, but is for salinity. Unlike the temperature scores, the differences between the different ensemble schemes are larger, especially over the whole observation space where the ensemble (g) with the Gaussian distribution for both the horizontal and vertical mixing parameters is most skilful. There is no significant difference in the space between 50 and 200 m. The advantages of ensembles over a single forecast of salinity are larger than those for temperature. The EFT hours are about 7.5–11.5 hours and 25–36 hours in the whole observation space and within the space of 50–200 m. This may not be surprising if we look at the ensemble spread–reliability difference between temperature and salinity, such as Figures 11, 12, and Figures 13, 14. Both the reliability and consistency of the ensembles are much higher for salinity in these two observation spaces.

4. Discussion and conclusions

In this article, the US Navy's RELO ensemble prediction system is fully described, and its performance is examined in the Gulf of Mexico for the period from 0000 UTC 15 April to 0000 UTC 25 July 2010. After briefly describing the initial perturbation generation method using the ensemble transfer (ET), a new time-deformation technique is introduced to produce the surface forcing perturbations from the operational atmospheric model fields. The results presented in this article demonstrate that the RELO ensemble mean forecasts are clearly superior to a single deterministic forecast in terms of accuracy and skill for all the variables and over all the domains considered in this article. In order to quantify the advantages of the ensemble forecasts in comparison with the single deterministic forecast, the extended forecast time (EFT) is introduced and computed. At the same time, the ensemble spread and its growth are investigated, and the impacts on the ensemble forecast accuracy, reliability and skill are also examined in detail.

The RELO ensemble spread provides valuable uncertainty forecast information; however, it is also clear that the uncertainty forecast capability should be improved further by accounting for more model-related uncertainties. As discussed in the introduction, it is challenging to compute the initial analysis error from a 3D-Var-based DA system such as NCODA. In the current version of the RELO ensemble, it is found that the analysis error that is used in the ET initialization is underestimated. As a result of the underestimation, the initial spread is found to be too small to have a magnitude comparable to the ensemble mean forecast error. In spite of this, the superiority of ensemble mean to the single forecast is clearly demonstrated in this study. Ensembles based on HYCOM generated with other methods such as that in Counillon and Bertino (2009) were also found to be under-dispersive. The authors found out that the ensemble spread is two to three times smaller than the forecast error with perturbations on DA parameter, atmospheric and lateral boundary conditions. At NRL, efforts are being made to make a better estimate of the

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.

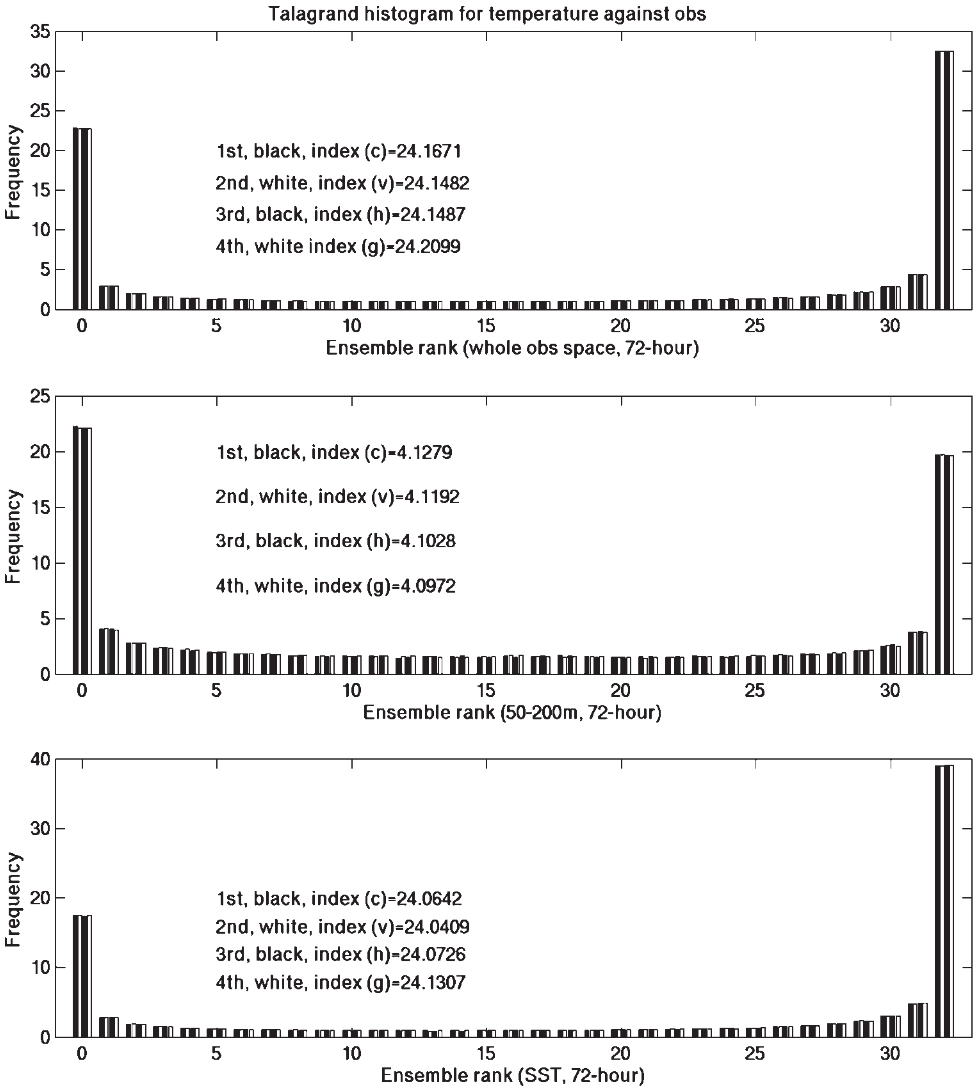

Figure 13. Talagrand rank histograms for temperature using observation as truth for lead times of 72 hours. All the values are averaged from 0000 UTC 15 April to 0000 UTC 25 July 2010, and over the whole observation space (1st row), 50–200 m (2nd row) and surface (3rd row). Ensembles schemes are indicated in black and white bars from left to right.

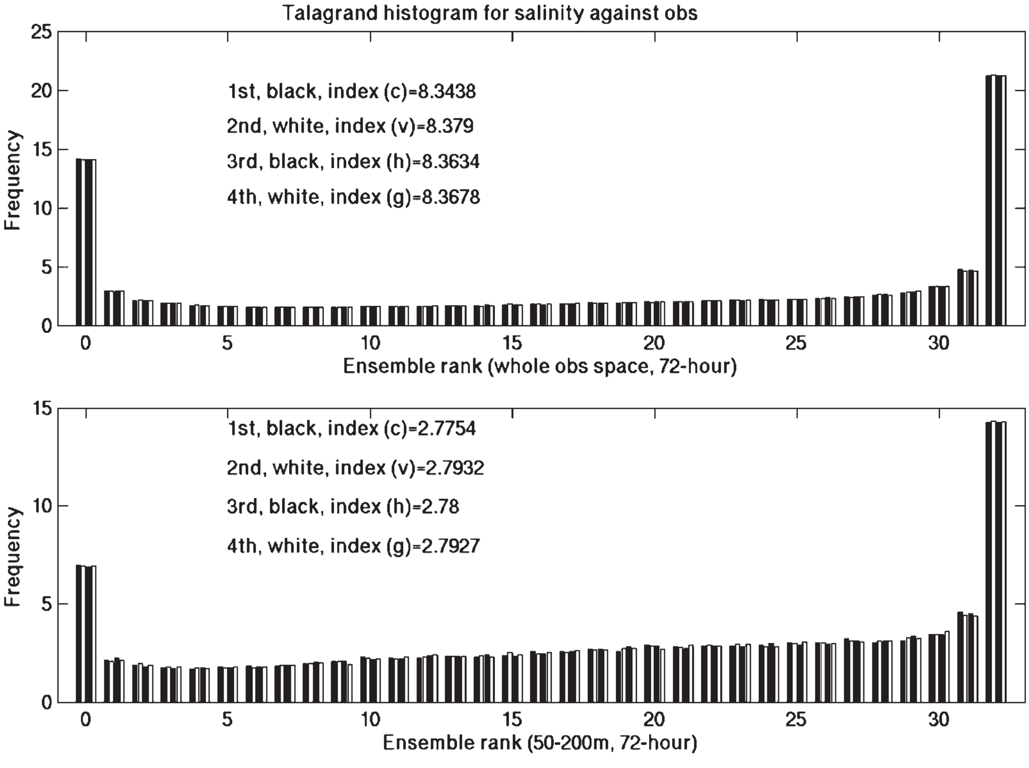

Figure 14. Same as Figure 13, but for salinity, and the values are averaged over the whole observation space (1st row), 50–200 m (2nd row).

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.

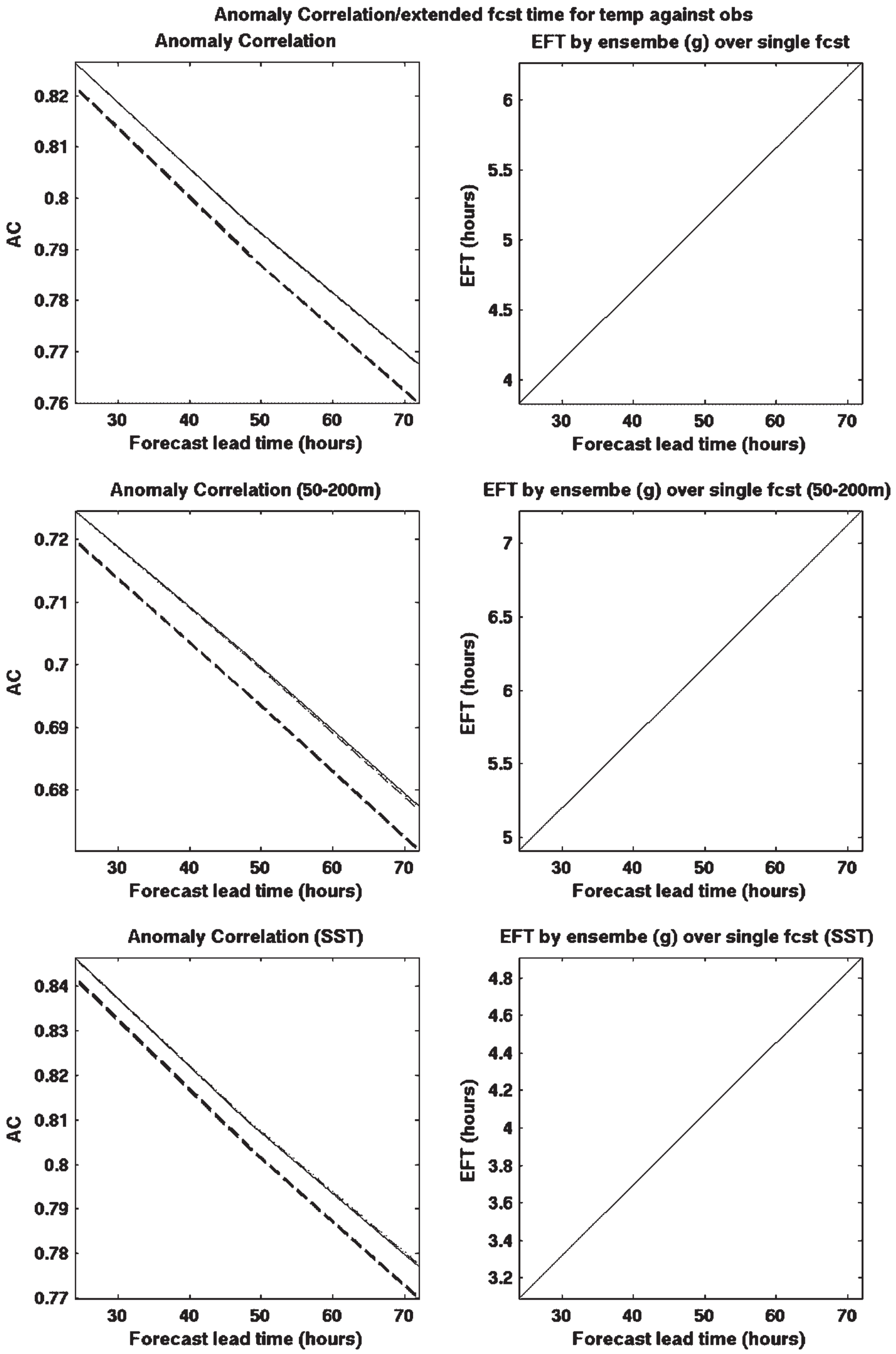

Figure 15. The anomaly correlation (left column) of ensemble mean and single forecast (thick dashed) for temperature for different ensembles as a function of forecast lead time. Right column: the extended forecast time by ensemble mean over single forecast as a function of time. All the values are averaged from 0000 UTC 15 April to 0000 UTC 25 July 2010, and over the whole observation space (1st row), 50–200 m (2nd row) and surface (3rd row). AC is computed using observation as truth and real-time climatology as reference. Ensembles schemes are indicated in dotted (c), dash-dot (v), dashed (h) and solid (g).

analysis error from NCODA; this will improve our future RELO ensemble. More advanced methods, such as the Lanczos method with calibration in 3D-Var, could be explored to improve the analysis error estimate (Wei *et al.,* 2012).

The results in this study also show that, without accounting for model-related uncertainties, the spread growth of the RELO ensemble cannot match that of the ensemble mean error. As an initial step in our long-term research plan towards accounting for more ocean-model-related uncertainties in the future, we introduce randomly generated perturbations to the two most important parameters in the ocean model mixing parametrizations, namely the Smagorinsky horizontal and Mellor–Yamada vertical mixing schemes. In order to study the impact of these parameter perturbations on the ensemble, we carry out experiments with a single deterministic forecast and the default control ensemble, in addition to the three different parameter perturbation schemes based on uniform and Gaussian distributions.

It is found that all three schemes improve the ensemble spread to a certain extent, particularly the scheme with a Gaussian distribution imposed on both the horizontal and vertical mixing parameters simultaneously. The improvement on the ensemble forecast accuracy, reliability and skill is found to be small. This is consistent with the findings for atmospheric ensembles as described in Bowler *et al.* (2009), Charron *et al.* (2010), Hacker *et al.* (2011a) and Reynolds *et al.* (2011b). It is also clear that the ensemble spreads still do not grow as fast as the ensemble mean RMS error with addition of parameter perturbations. These findings indicate that just perturbing these two mixing parameters is not sufficient to account for most of the model-related uncertainties. The results will provide some insights with respect to perturbing parameters in ocean models for other researchers and developers in the ocean community. Further

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.

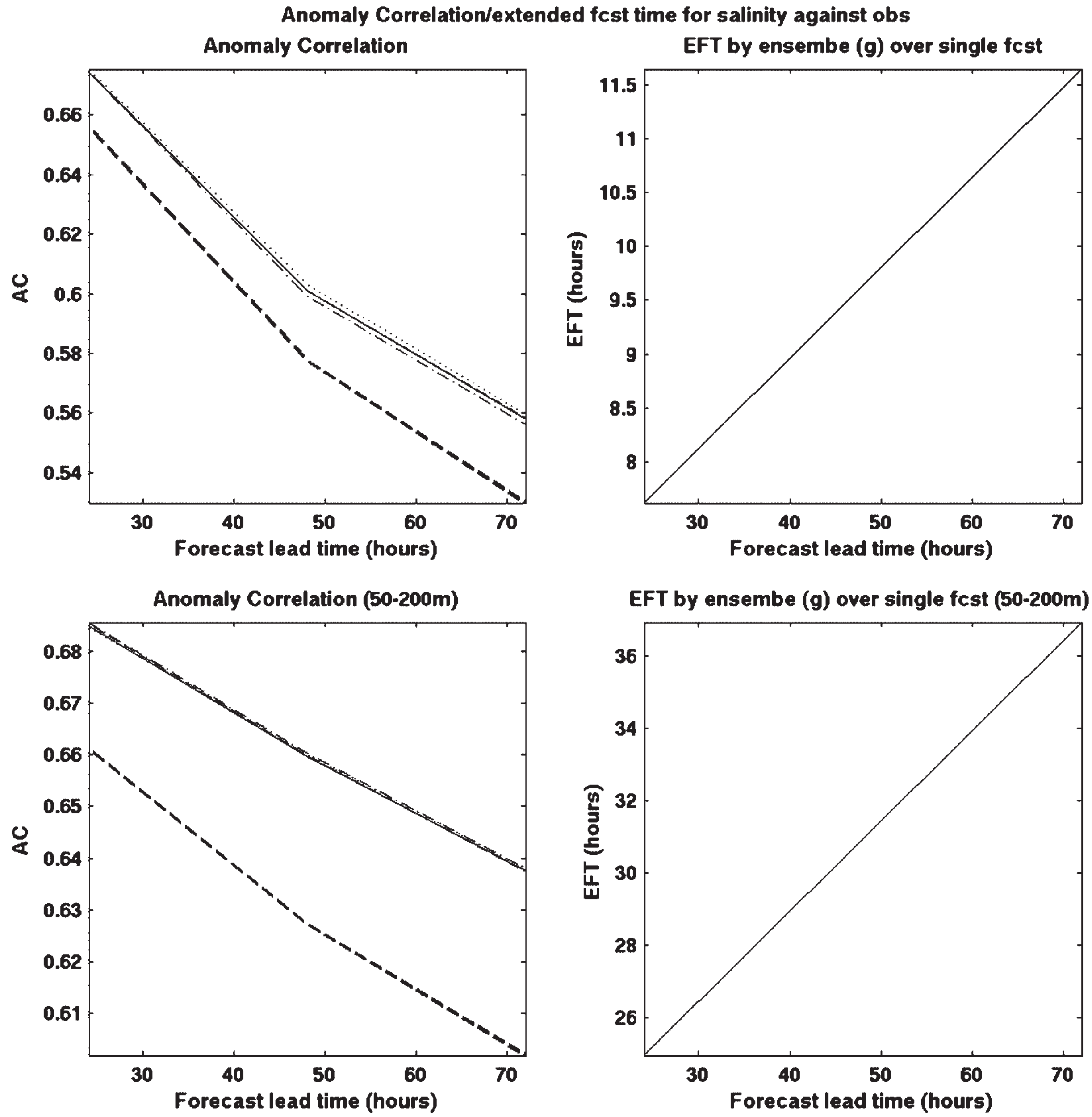

Figure 16. Same as Figure 15, but for salinity, and the values are averaged over the whole observation space (1st row), 50–200 m (2nd row).

improvements for RELO ensemble could be made from two possible aspects. The first one is to perturb more parameters in addition to these two in NCOM simultaneously. The second option is to increase the ranges of perturbations for these two parameters. Including both of these options at the same time will probably produce largest impacts on the ensemble spread and performance. However, care must be taken when the ranges are increased, as too large or too small a value of either parameter could force the model to produce unphysical variable values. These options will be explored in future studies.

Another area to which we are going to pay special attention in the near future, to account for more model-related errors, is to develop a stochastic parametrization scheme that imposes stochastic forcing at all the model grid points. Lermusiaux (2006) studied the use of stochastic forcing in ocean prediction and DA systems and found a positive impact. The ensemble experience at the major NWP centres has shown that a number of stochastic parametrization schemes are normally needed to account for various sources of model uncertainties in a mature, reliable ensemble. In general, the ensemble performance is best when a number of different schemes are used simultaneously for different sources of model error. The parameter-perturbation scheme is most effective when it is combined with other schemes in an atmospheric ensemble (Palmer *et al.,* 2005; Bowler *et al.,* 2009; Charron *et al.,* 2010; Hacker *et al.*, 2011a, 2011b; Reynolds *et al.,* 2011a, 2011b). Theoretically, model errors related to subgrid-scale parametrizations can also be dealt with by more general statistical dynamics and stochastic models such as those implemented in O'Kane and Frederiksen (2008). A recent review of these broader theoretical methods can be found in Frederiksen *et al.* (2012). The discussion of these methods is beyond the scope of this study.

The parameter perturbation scheme developed in this article will be one of the components of the future RELO ensemble system. It is our plan at NRL that a few different stochastic schemes will be developed and implemented in the RELO to account for various sources of model uncertainty in order to achieve the best probabilistic forecast performance. Hopefully, when more model uncertainties are accounted for, the ensemble spread will grow at a similar rate to the ensemble mean RMS error and raise the RELO ensemble reliability to a new level. We will report our progress when it becomes available.

Acknowledgement

This work was funded through the SEMESTER 6.2 project at NRL and supported by the Office of Naval Research (Program Element 0602435N). We are grateful for the assistance of many

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.

of our colleagues at NRL at Stennis Space Center, particularly Pat Hogan, Emanuel Coelho, Peter Spence, Jan Dastugue and Philip Muscarella.

References

Barkmeijer J, Buizza R, Palmer TN. 1999. 3D-Var Hessian singular vectors and their potential use in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **125**: 2333–2351.

Barron CN, Kara AB, Martin PJ, Rhodes RC, Smedstad LF. 2006. Formulation, implementation and examination of vertical coordinate choices in the Global Navy Coastal Ocean Model (NCOM). *Ocean Modelling* **11**: 347–375.

Bishop CH, Toth Z. 1999. Ensemble transformation and adaptive observations. *J. Atmos. Sci.* **56**: 1748–1765.

Bishop CH, Etherton BJ, Majumdar SJ. 2001. Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Weather Rev.* **129**: 420–436.

Bowler NE, Arribas A, Beare SE, Mylne KR, Shutts GJ. 2009. The local ETKF and SKEB: Upgrades to the MOGREPS short-range ensemble prediction system. *Q. J. R. Meteorol. Soc.* **135**: 767–776.

Buizza R, Palmer TN. 1995. The singular-vector structure of the atmospheric global circulation. *J. Atmos. Sci.* **52**: 1434–1456.

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

Carnes MR, Helber RW, Barron CN, Dastugue JM. 2010. '*Validation test report for GDEM4.*' NRL Technical Report, NRL/MR/7330-10-9271, Naval Research Laboratory, Stennis Space Center, MS, USA, 42 pp.

Charron M, Pellerin G, Spacek L, Houtekamer PL, Gagnon N, Mitchell HL, Michelin L. 2010. Toward random sampling of model error in the Canadian ensemble prediction system. *Mon. Weather Rev.* **138**: 1877–1901.

Chen S, Cummings J, Doyle J, Hodur RH, Holt T, Liou C, Liu M, Mirin A, Ridout J, Schmidt JM, Sugiyama G, Thompson WT. 2003. '*COAMPS version 3 model description: General theory and equations.*' NRL/PU/7500–03-448, 145 pp.

Counillon F, Bertino L. 2009. High-resolution ensemble forecasting for the Gulf of Mexico eddies and fronts. *Ocean Dyn.* **59**: 83–95.

Cummings JA. 2005. Operational multivariate ocean data assimilation. *Q. J. R. Meteorol. Soc.* **131**: 3583–3604.

Fisher M, Courtier P. 1995. '*Estimating the covariance matrices of analysis and forecast error in variational data assimilation.*' ECMWF Technical Memorandum 220.

Frederiksen JS, O'Kane TJ, Zidikheri MJ. 2012. Stochastic subgrid parameterizations for atmospheric and oceanic flows. *Phys. Scr.* **85**: 1–29.

Fujii Y, Tsujino H, Usui N, Nakano H, Kamachi M. 2008. Application of singular vector analysis to the Kuroshio large meander. *J. Geophys. Res.* **113**: C07026, DOI: 10.1029/2007JC004476.

Hacker JP, Ha S-Y, Snyder C, Berner J, Eckel FA, Kuchera E, Pocernich M, Rugg S, Schramm J, Wang X. 2011a. The U.S. Air Force Weather Agency's mesoscale ensemble: Scientific description and performance results. *Tellus* **63A**(:): 625–641.

Hacker JP, Snyder C, Ha S-Y, Pocernich M. 2011b. Linear and non-linear response to parameter variations in a mesoscale model. *Tellus* **63A**: 429–444.

Hodur RM. 1997. The Naval Research Laboratory's Coupled Ocean/ Atmosphere Mesoscale Prediction System (COAMPS). *Mon. Weather Rev.* **125**: 1414–1430.

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

Large WG, McWilliams JC, Doney SC. 1994. Oceanic vertical mixing: A review and a model with a nonlocal boundary layer parameterization. *Rev. Geophys.* **32**: 363–403.

Lermusiaux PFJ. 2006. Uncertainty estimation and prediction for interdisciplinary ocean dynamics. *J. Comput. Phys.* **217**: 176–199.

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

McLay JG, Reynolds CA. 2009. Two alternative implementations of the ensemble-transform (ET) analysis-perturbation scheme: The ET with extended cycling intervals, and the ET without cycling. *Q. J. R. Meteorol. Soc.* **135**: 1200–1213.

McLay JG, Bishop CH, Reynolds CA. 2007. The ensemble-transform scheme adapted for the generation of stochastic forecast perturbations. *Q. J. R. Meteorol. Soc.* **133**: 1257–1266.

McLay JG, Bishop CH, Reynolds CA. 2008. Evaluation of the ensemble transform analysis perturbation scheme at NRL. *Mon. Weather Rev.* **136**: 1093–1108.

McLay JG, Bishop CH, Reynolds CA. 2010. A local formulation of the ensemble transform (ET) analysis perturbation scheme. *Weather and Forecasting* **25**: 985–993.

Majumdar SJ, Bishop CH, Etherton BJ, Toth Z. 2002. Adaptive sampling with ensemble transform Kalman filter. Part II: Field program implementation. *Mon. Weather Rev.* **130**: 1356–1369.

Martin PJ. 2000. '*A description of the Navy Coastal Ocean Model Version 1.0.*' Technical Report NRL/FR/7322–00-9962, Naval Research Laboratory, Stennis Space Center, *MS*, **42**(p):p.

Mellor GL, Durbin PA. 1975. The structure and dynamics of the ocean surface mixed layer. *J.Phys. Oceanogr.* **5**: 718–728.

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

Mellor GL, Yamada T. 1982. Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys.* **20**: 851–875.

Miyazawa Y, Yamane S, Guo XY, Yamagata T. 2005. Ensemble forecast of the Kuroshio meandering. *J. Geophys. Res.* **110**: C10026, DOI: 10.1029/2004JC002426.

Molteni F, Buizza R, Palmer TN, Petroliagis T. 1996. The ECMWF Ensemble Prediction System: Methodology and validation. *Q. J. R. Meteorol. Soc.* **122**: 73–119.

O'Kane TJ, Frederiksen JS. 2008. A comparison of statistical dynamical and ensemble prediction methods during blocking. *J. Atmos. Sci.* **65**: 426–447.

O'Kane TJ, Oke PR, Sandery PA. 2011. Predicting the East Australian Current. *Ocean Modelling* **38**: 251–266.

Palmer TN, Shutts GJ, Hagedorn R, Doblas-Reyes FJ, Jung T, Leutbecher M. 2005. Representing model uncertainty in weather and climate prediction. *Ann. Rev. Earth Planet. Sci.* **33**: 163–193.

Reynolds CA, Teixeira J, McLay JG. 2008. Impact of stochastic convection on the ensemble transform. *Mon. Weather Rev.* **136**: 4517–4526.

Reynolds CA, McLay JG, Goerss JS, Serra EA, Hodyss D, Sampson CR. 2011a. Impact of resolution and design on the U.S. Navy global ensemble performance in the Tropics. *Mon. Weather Rev.* **139**: 2145–2165.

Reynolds CA, Ridout JA, McLay JG. 2011b. Examination of parameter variations in the U.S. Navy global ensemble. *Tellus* **63A**: 841–857.

Rowley C. 2008. '*RELO system user guide.*' Oceanography Division, Naval Research Laboratory. Stennis Space Center, MS, USA, 59 pp.

Rowley C. 2010. '*Validation test report for the RELO system.*' Oceanography Division, NRL Memorandum Report NRL/MR/7320–10-9216. Naval Research Laboratory, Stennis Space Center, *MS*, **69**: pp.

Rowley C, Richman J, Ferreira-Coelho E. 2012. 'Boundary condition uncertainty in the NRL relocatable ocean ensemble forecast system.' *AGU Ocean Science Meeting*, Salt Lake City, UT, 20–25 February 2012.

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

Talagrand O, Vautard R, Strauss B. 1997. 'Evaluation of probabilistic prediction systems.' Pp 1–26 in *Proc. Workshop on Predictability, ECMWF, Reading,* UK.

Toth Z, Kalnay E. 1993. Ensemble forecasting at NMC: The generation of perturbations. *Bull. Am. Meteorol. Soc.* **74**: 2317–2330.

Toth Z, Kalnay E. 1997. Ensemble forecasting at NCEP and the breeding method. *Mon. Weather Rev.* **125**: 3297–3319.

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

Wei M, Toth Z, Wobus R, Zhu Y, Bishop CH, Wang X. 2005. 'Initial perturbations for NCEP ensemble forecast system.' *Proceedings of the First THORPEX International Science Symposium, 6–10 December 2004,* Montreal, Canada. Pp 227–230 in WMO TD No. 1237, WWRP THORPEX No. 6, 2005.

Wei M, Toth Z, Wobus R, Zhu YJ, Bishop CH, Wang XG. 2006. Ensemble transform Kalman filter-based ensemble perturbations in an operational global prediction system at NCEP. *Tellus* **58A**: 28–44.

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

Wei M, Toth Z, Zhu YJ. 2010. Analysis differences and error variance estimates from multi- center analysis data. *Austr. Meteorol. Oceanogr. J.* **59**: 25–34.

Wei M, De Pondeca MSFV, Toth Z, Parrish D. 2012. Estimation and calibration of observation impact signals using the Lanczos method in NOAA/NCEP data assimilation system. *NonlinearProcesses in Geophys.* **19**: 541–557.

Wilks DS. 2006. *Statistical Methods in the Atmospheric Sciences*, 2nd edition. Cambridge University Press.

Yin X-Q, Oey L-Y. 2007. Bred-ensemble ocean forecast of loop current and rings. *Ocean Modelling* **17**: 300–326.

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.

Appendix A. Steps for computing spread reliability

To quantify the ensemble spread reliability, we compute the forecast error variance explained by the ensemble variance. Similar methods have been used in Majumdar *et al.* (2002), Wei *et al.* (2006), and Leutbecher and Palmer (2008). The following steps can be carried out to achieve this. (1) Select a variable at a specific forecast lead time over a targeted domain. (2) Choose a truth, such as an observation in our case, and compute the ensemble mean RMS error at each grid point. (3) Compute the corresponding ensemble spread. (4) Stratify the ensemble spread and mean RMS error, followed by partitioning into equally populated bins of increasing spread. (5) Compute the mean values of the ensemble spread and RMS error for each bin. (6) Draw a curve connecting the average value from each bin. A good ensemble with a reliable spread should have a curve that closely follows the diagonal line.

Appendix B. Talagrand histogram and consistency index

The Talagrand histogram (Talagrand *et al.,* 1997; Wilks, 2006) or rank histogram checks whether the ensemble spread is consistent with the assumption that the observation (which is used as truth) is statistically just another member of the forecast distribution, and all the observations are equally distributed amongst the predicted ensemble. It measures how well the ensemble spread represents the true uncertainty of the truth/observations. The general steps to construct a rank histogram are: (1) Select an ensemble variable at a specific forecast lead time and a domain you want to verify. (2) Choose a truth that you want to verify against. This is normally an observation or analysis. We use an observation as truth in most of this article. (3) Sort the ensemble members into increasing order at every observation point as depicted in Figure 17 for the RELO ensemble. This creates 33 possible bins (for 32 members) that the observation falls into at each point. (4) Count where the observation falls with respect to the ensemble forecast. (5) Tally over all the observations to create a histogram of rank.

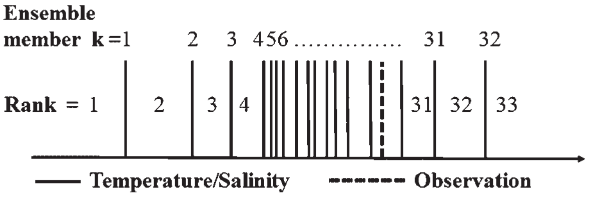

Figure 17. Schematic for sorting RELO ensemble forecasts to compute Talagrand rank histogram.

For an ensemble with a perfect spread, each member represents an equally likely scenario, so the observation is equally likely to fall among any member. A uniform, flat histogram means that the truth is indistinguishable from any ensemble member, and the ensemble spread is about right to represent the forecast uncertainty. A U-shaped histogram indicates that the ensemble spread is too small, and many of the observations fall outside the extremes of the ensemble. In this case, the ensemble is under-dispersive and overconfident. A dome-shaped histogram indicates that the ensemble spread is too large, and most of the observations fall near the centre of the ensemble. In this case the ensemble is over-dispersive and underconfident. An asymmetric shape means that the ensemble contains some biases.

To have a quantitative measure of the consistency of the rank histogram, a consistency index is defined in this article. The basic idea is motivated by Talagrand *et al.* (1997), but the formulation is modified. Let k be the number of ensemble members, and n be the number of samples. The total rank is $k+1$. For an ideal uniform distribution, the number of elements in each bin is $n/(k+1)$. Thus, if the number in bin i is n_i, then the expected value of n_i is

$$\langle n_i \rangle = \frac{n}{k+1}. \tag{A1}$$

The RMS distance between the real and the expected values can be written as

$$rmsd = \sqrt{\frac{1}{k+1}\sum_{i=1}^{k+1}\left(n_i - \frac{n}{k+1}\right)^2}. \tag{A2}$$

It can be shown that $\langle rmsd \rangle = \sqrt{\frac{nk}{k+1}}$. The consistency index is defined as

$$c = \frac{rmsd}{\langle rmsd \rangle}. \tag{A3}$$

For an ideal, reliable ensemble system, $c \approx 1.0$.

Published 2013. This article is a U.S. Government work
and is in the public domain in the USA.