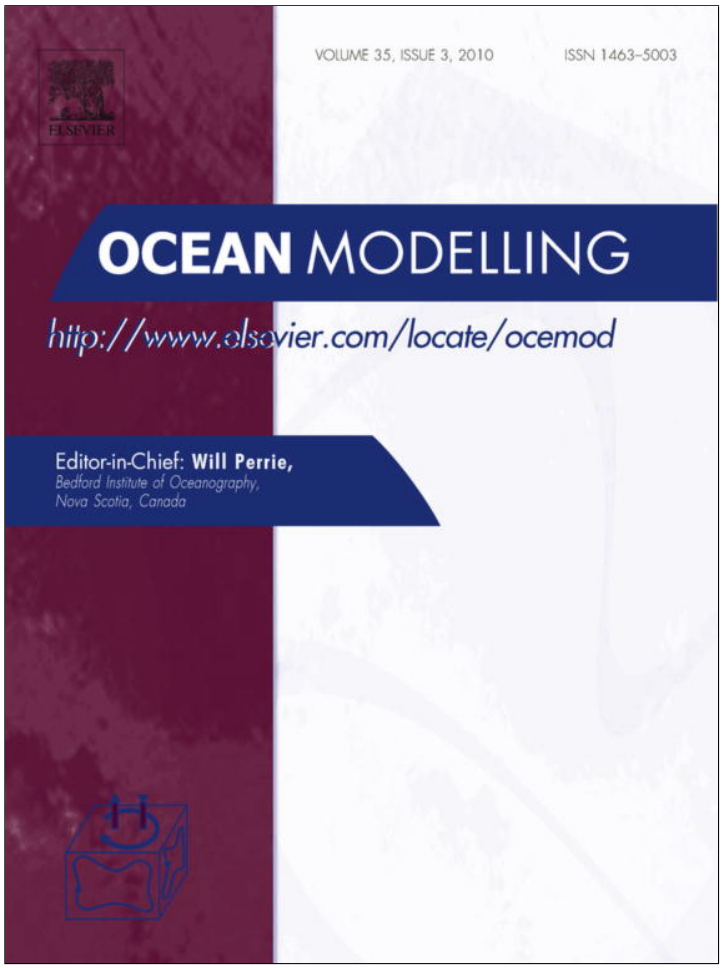

Ocean Modelling 35 (2010) 187–205

Contents lists available at ScienceDirect

Ocean Modelling

journal homepage: www.elsevier.com/locate/ocemod

Assimilation of Argo temperature and salinity profiles using a bias-aware localized EnKF system for the Pacific Ocean

Ziwang Deng [a], Youmin Tang [a,*], Guihua Wang [b]

[a] Environmental Science and Engineering, University of Northern British Columbia, 3333 University Way, Prince George, BC, Canada V2N 4Z9
[b] State Key Laboratory of Satellite Ocean Environment Dynamics, Second Institute of Oceanography, SOA, Hangzhou 310012, China

ARTICLE INFO

Article history:
Received 18 November 2009
Received in revised form 6 July 2010
Accepted 15 July 2010
Available online 17 August 2010

Keywords:
Argo profiles
EnKF data assimilation
Pacific Ocean
Temperature
Salinity

ABSTRACT

In this study, Argo profiles of temperature and salinity for the period from January 2005 to December 2007 are assimilated into a primitive equation model of the Pacific Ocean using a bias-aware localized ensemble Kalman filter (EnKF) with a sequence of 5-day assimilation cycles. Some other in situ observations, including XBT, TAO/TRITON and CTD profiles, used to supplement, are also assimilated into the model. To improve the assimilation performance, several strategies addressing the computational expense and model error statistics are incorporated into the assimilation scheme. Validation is performed by comparing the analyzed ocean states with independent data, including withheld Argo profiles, satellite remote sensing sea level height anomalies (SLA) and the NCEP ocean state re-analysis products. The results show that the assimilation system is capable of significantly reducing the bias and RMSE of ocean temperature and salinity compared with the control run. It can also improve the simulation of zonal currents and SLAs along the equator, especially during strong ENSO events. In addition, a hybrid coupled ENSO prediction model initialized by the assimilation analysis improves the ENSO prediction skill compared against that initialized by the control run without data assimilation.

1. Introduction

A new global oceanic observing system, known as Argo (The Array for Real-time Geostrophic Oceanography), is initiated at the end of 2000. Recently, Argo achieved the target of sampling array of about 3000 (including around 1900 in the Pacific Ocean) floats traveling around the world and supplying observations of the state of the ocean. As a result, Argo has become the main contributor to ocean in situ observations in terms of numbers and offers an opportunity to more accurately estimate ocean states by assimilating Argo profiles into a state-of-the-art ocean general circulation model.

An efficient use of Argo data is to combine it with other data sets, such as the Expendable Bathythermographs (XBT), Tropical Atmosphere and Ocean – TRIangle Trans-Ocean buoy Network (TAO/TRITON, McPhaden, 1995), Conductivity–Temperature–Depth (CTD, Bellucci et al., 2007), satellite altimetry, etc., and models through effective data assimilation techniques. The assimilations of the XBT, TAO/TRITON and CTD have been widely explored; however, the assimilation of Argo profiles has not been sufficiently investigated. Recently, several different attempts were made to assimilate Argo temperature and salinity profiles into ocean models (Kamachi et al., 2004; Oke et al., 2005, 2008; Cummings, 2005; Martin et al., 2007; Huang et al., 2008; Smith and Haines, 2009). The results of these studies show that the inclusion of Argo data into the assimilated observation dataset can significantly improve subsurface ocean state estimation skill. As a new relatively high resolution in situ oceanic observation dataset, Argo data has not been sufficiently explored in estimating the ocean state. Thus, how to effectively use this new information source in improving ocean state estimation is still an active research field. In addition, due to uncertainties in the models and forcing fields, the sparsity of in situ observations and the limitation of model resolution, the analyzed ocean states obtained are still not perfect and contain some errors and uncertainties. Development of new ocean data assimilation systems, improvement of old ocean data assimilation systems and the effective use of new observation products are urgently required. Over the past several decades, many data assimilation methods were developed. One widely used method is the ensemble Kalman filter (EnKF), that was first proposed by Evensen (1994). The EnKF has several appealing properties. For example, the EnKF handles uncertainty in the observations and the prior forecast, approximates the Bayesian update for the forecast state given new observations, provides direct estimates of the forecast covariances from the forecast ensemble and then explicitly updates that ensemble to be consistent with the uncertainty of the analysis (Snyder and Zhang, 2003). However, there are some limitations in the traditional EnKF

* Corresponding author. Tel.: +1 250 9605190; fax: +1 250 9605845.
E-mail address: ytang@unbc.ca (Y. Tang).

1463-5003/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ocemod.2010.07.007

schemes, for example, the high computational cost given the huge model dimension and the sensitivity to the method used to estimate errors. In recent years, some variants of the EnKF were developed to decrease computational cost and/or improve assimilation error characteristics (e.g., Anderson, 2002). For example, one efficient strategy that significantly reduces the computational cost while guaranteeing the analysis quality, is the local ensemble Kalman filter which performs the analysis locally (Fukumori, 2002; Ott et al., 2004). Examples of the techniques used to improve the estimation of the model error characteristics include using an inflator to amplify the background error covariance matrix (Houtekamer and Mitchell, 2001; Desroziers et al., 2005; Li et al., 2009; Zheng, 2009), using innovations to estimate the model error parameters (Mitchell and Houtekamer, 2000), simulating model bias by colored noise (Chui and Chen, 1999) and correcting model bias by a two-stage scheme (Friedland 1969; Dee and da Silva, 1998; Chepurin et al., 2005; Dee, 2005). These techniques have improved the capability and facility of the EnKF and have been applied well mainly in atmospheric data assimilation. However, some of these methods have not been sufficiently explored in oceanic data assimilation, especially in Argo data assimilation.

The purpose of this study is to construct a state-of-the-art ocean data assimilation system that incorporates some advanced strategies. Thus, it can effectively assimilate Argo profiles and other in situ ocean observations into an ocean general circulation model. The ocean model is configured to the Pacific Ocean. Therefore, the data assimilation system can provide initial conditions for ENSO forecast. Some strategies, including the local analysis, bias correction, adaptive stochastic model error estimation, observation error estimation and the incremental analysis update (IAU) are used in this study to address some inherent concerns in EnKF such as computational cost, model bias and ensemble generation. Assimilation experiments from 2005 to 2007 are carried out to validate the assimilation system by comparisons of the analyzed ocean states with independent observations.

This paper is organized as follows. In Section 2, the data and its treatment is given. An overview of the construction of the EnKF data assimilation and the design of the experiments is provided in Section 3. Validation of the data assimilation is discussed in Sections 4 and 5 ends with a summary and discussion.

2. Data

The delayed mode (D-mode) Argo profiles are used as observations in this study. This data has been subjected to detailed scrutiny by oceanographic experts and the final data has been estimated by comparison with high quality ship-based CTD data and climatologies using the process described by Wong et al. (2003) and Böhme and Send (2006). This process is carried out on a 1 year long "data window", so there were only a few D-mode profiles available after 2007 when we began this study in the second half of 2008. Therefore, this study focuses on the assimilation of Argo profiles for the 3-year period from 2005 to 2007. There are about 90,000 paired D-mode Argo T–S profiles available covering most of the Pacific from 60°S to 60°N during the 3-year period. The distribution of Argo profiles in time and space is not uniform. The number of the profiles increases from 2005, peaks in 2006 and decreases in 2007 again. The smaller number of available D-mode Argo profiles in 2007 is due to some of the profiles obtained in 2007 have not been processed yet. If we count Argo profiles within a shifted 11-day time window, there are about 500–800 profiles for 2005, 800–950 for 2006 and 650–700 for 2007 available. The spatial distribution of these profiles is shown in Fig. 1. In this figure the numbers of Argo profiles are counted in a 2°(lon.) × 2°(lat.) bin that is shifted over the whole domain. A dominant feature in Fig. 1 is the significantly heterogeneous distribution of profiles in space; in some regions the profiles are significantly sparser than in other regions. There are some outstanding low density bands, for example, the regions extending approximately along the equator, 37°N, 37°S and between 15°S and 15°N west of the dateline, suggesting the requirement of other supplemental subsurface observations in these regions. The causes of this significant spatial difference are interesting but beyond the scope of this study. The impacts of this heterogeneous distribution of Argo profiles on the performance of the data assimilation will be discussed in Section 4. Other data, i.e., the XBT, CTD and TAO/TRITON profiles, are also used as

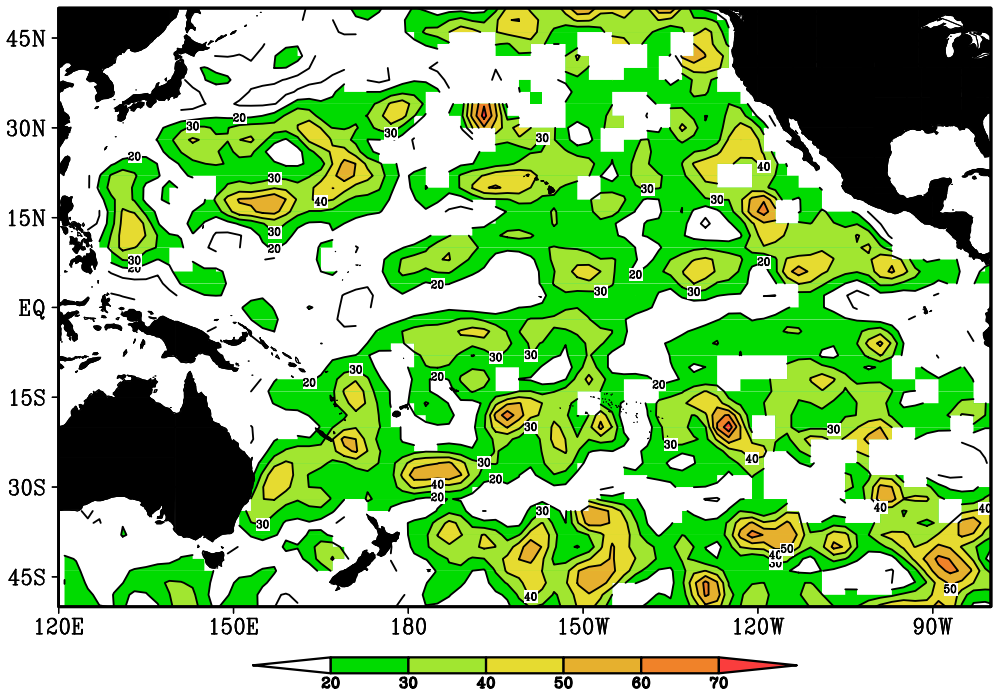

Fig. 1. Spatial distribution of the total number of Argo profiles for 2005–2007. The number is counted at a 2° × 2° cell, shifted over the whole domain.

supplemental observations in this study when Argo profiles are not sufficient within an assimilation window. All these data are available online at http://www.nodc.noaa.gov/GTSPP/). Generally, Argo profiles extend from surface down to 2000 m, XBT data to 800 m but a more typical depth is 500 m, CTD and TAO–TRITON observations to a depth of 500 m.

Comparisons with the independent data that are not used in the assimilation cycles are necessary for the validation of the assimilation system. The independent data used in this study include: withheld Argo T–S profiles obtained using the cross-validation scheme as described in Section 3; the sea level anomaly (SLA) which is a time delayed low resolution ($1^\circ \times 1^\circ$, Mercator grid) gridded product (http://www.aviso.oceanobs.com); and the NCEP re-analysis subsurface temperature and currents. The SLA is the merged map which is produced by combining products of all satellites and suitable for large-scale ocean variation studies (Dibarboure et al. 2009). The time interval of this data is 7 days. The NCEP re-analysis products were gained by some of the data used here but we still consider them as independent data due to the significant differences in the models and the whole observation dataset between the NCEP assimilation system and the one that used here.

3. The ocean data assimilation system

3.1. The ocean general circulation model (OGCM)

The OGCM is the latest version of Ocean Parallelise (OPA9.2), a primitive equation OGCM widely used in oceanic studies (e.g., Delecluse and Madec, 1999; Tang et al., 2004; Moore et al., 2006; Bellucci et al., 2007; Deng et al., 2009). The ocean model is described in Madec (2008). The domain of the model used here is the Pacific Ocean between 60°N and 60°S and between 116°E and 66°W, for a total of 90×95 horizontal grid points. The horizontal resolution in the zonal direction is 2°, while the resolution in the meridional direction is 0.5° within 5° of the equator, smoothly changing up to 2.0° at 30°N and 30°S and then changing down to 1.0° at 60°N and 60°S. There are 31 unevenly spaced levels in the vertical, with 24 levels concentrated in the upper 2000 m. The thickness of the levels varies from 10 m at the surface (within the first 100 m) to 500 m below the 3000 m level. The maximum depth is set to 5000 m and a realistic topography based on the ETOPO5′ global atlas is used (Ferry et al., 2007). The spatial resolution of this version is somewhat coarse for performing high resolution data assimilations. However, due to the limitation of computation resource, we are forced to use this moderate resolution ocean model. In addition, the goal of this study is to construct a data assimilation system, through incorporating some advanced strategies into the EnKF to provide initial conditions for oceanic and seasonal climate prediction (ENSO forecast).

The model is forced for 200 years with the NCEP monthly climatological mean wind stress, derived from the 50-year NCEP re-analysis wind stress, and the heat flux Q_s to get an initial state. From this initial condition, the model is forced by the actual monthly NCEP wind stress to simulate conditions for the period 1981–2004. The ocean state at the end of 2004 provided the initial conditions for both the control and the assimilation runs, starting on January 1, 2005. The control run for the period 2005–2007 without data assimilation provides a basis for comparison. The heat flux Q_s is given by

$$Q_s = Q_0 + \lambda(T - T_0) \quad (1)$$

where Q_0 is the climatological heat flux, obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF) re-analysis project for the base period 1971–2000. T is the model SST, T_0 is Levitus' observed climatological SST (Levitus and Boyer, 1998), and λ is the relaxation rate, set to $-40\ \mathrm{W\,m^{-2}\,K^{-1}}$ (Tang et al., 2004; Moore et al., 2006). For a 50 m mixed-layer depth, this value corresponds to a relaxation time scale of two months (Madec, 2008).

3.2. EnKF data assimilation system

The data assimilation system constructed here is based on the EnKF, which has gained popularity because of its simple conceptual formulation and relative ease of implementation (Evensen, 2003). However, there are still some issues demanded to be addressed when developing a realistic assimilation system by the EnKF, e.g., the estimation of observation error covariance for a multi-source dataset, the efficiency of computation, the correction of model bias, etc. In this section we will discuss some strategies used in this study for these concerns.

3.2.1. Observation error

Argo profiles and other observations are used in this study. For some regions, we observed that some profiles are very close to each other at specific data assimilation steps, causing costly expense in the matrix inversion with little improvement in assimilation performance. To solve this problem, a sub data set is constructed by the "data thinning" approach prior to each assimilation step, ensuring that there is only one observation within a model grid cell and assimilation window (of 11 days, see below). Priority is given to Argo profiles in the construction of the sub data set due to its importance in this study. Such a data thinning approach is not optimal in the data assimilation context because it may loss some information in some analyses. However it can effectively reduce the cost of computation.

Argo floats have a 10-day re-sampling period thus Argo profiles available are sparse on a specific day. To solve the temporal sparsity of Argo data, we treat all Argo data within a time window of 11-days as the observations used for the assimilation. Oke et al. (2008) and Balmaseda et al. (2007) used this method when they assimilated Argo data. The 11-day window is much longer than the assimilation interval of 5 days here, which was chosen in order to include more observations at a single assimilation step as in Huang et al. (2008), Oke et al. (2008) and Smith and Haines (2009). This 11-day time window is centered at the analysis day. Thus, the observations for the day of the analysis and the observations for 5 days before and after the analysis day are used at each assimilation step. In this time window, all Argo profiles within a model grid cell are sorted according to the differences between the time of analysis and that of the observations. The profile that has the minimum time difference is finally used in the assimilation. If there is no Argo observation available in this process, other data sets are explored using the same scheme as for Argo profiles. With this strategy, the density of the constructed observations in the sub data set is no more than one observation per model grid cell. This approach can greatly simplify the process and save the computational cost although it rejects some useful profiles. However it can be expected that number of rejected Argo profiles is not large since Argo observation is sparser in space relative to the model horizontal resolution. We will find in the following discussion that these rejected profiles generally have large observation errors.

Since the model resolution varies with latitude, the constructed sub data sets used for the assimilation have relatively sparse samples at middle latitudes compared with dense samples at higher latitudes and in the equatorial regions. In the vertical direction, the nominal resolutions of Argo and the other data sets are typically much finer than the model grid. To ensure that there is no more than one observation falling within each model level, an observation is obtained for each model level by a linear interpola-

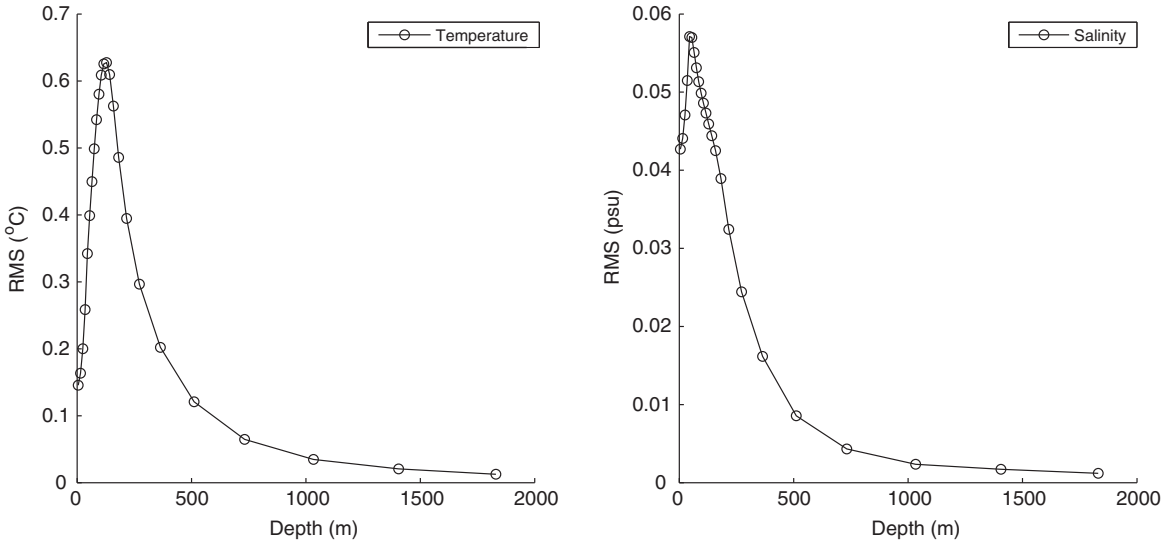

Fig. 2. Mean RMS of temperature (left panel) and salinity (right panel) as a function of depths. The value on each level is averaged over model domain based on the control run for the period 1985–2004 relative to the climatology of the model.

tion in the vertical using observations closest to the model level. This is a significant saving for profiles that are well-resolved over depth (Oke et al., 2008).

Observation errors, denoted by R_k, include measurement errors and representation errors. In order to take into account the time differences between the observations and the analysis, different weights are given to the observations made before or after the analysis (e.g., Oke et al., 2008; Bellucci et al., 2007). The weights, in principle, represent the inverse of the data error variances (Maes et al., 2000). Therefore, we use a simple linear function to adjust observation error variances according to the absolute differences between the analysis time and the observation times. In addition, we further consider the impacts of depth on the adjusted observation errors under the assumption that at deeper levels the impacts of time difference are less than that at the surface. Therefore, according to the method proposed by Oke et al. (2008), the error for an observation is estimated by

$$\varepsilon_o^2 = \varepsilon_{instr}^2 + \varepsilon_{RE}^2 + \varepsilon_{\Delta t}^2 \tag{2}$$

where ε_o^2 is the observation error, ε_{instr}^2 the instrument error, ε_{RE}^2 the representation error and $\varepsilon_{\Delta t}^2$ the error associated with the time difference between the analysis and observation (called age error). In this study, the D-mode data sets that have passed strict quality control are assumed to have very small instrument errors. Thus, we set $\varepsilon_{instr} = 0.1$ °C for the temperature observations and $\varepsilon_{instr} = 0.1$ psu for the salinity observations at all depths. The representation error ε_{RE}^2, which could be attributed to any physical process appearing in the observation but not in the model and referred to as the forward interpolation error, is subject to the effects of discretization error and limited resolution (Desroziers et al., 2005; Cummings, 2005). Indeed, Janic and Cohn (2006) demonstrated that ε_{RE}^2 is also state dependent and correlated in time. For simplicity, we only consider the impacts of model resolution on the representation error. It is reasonable to assume that at a given model resolution the representation error of a variable should be larger in a region where the variation of the model variable is large than in a region where the variation of the variable is small. Thus, we approximate the representation error in terms of model standard deviation (S_{mod}) using the following equation:

$$\varepsilon_{RE} = \kappa S_{\mathrm{mod}} \tag{3}$$

where κ is an adjustable coefficient with $\kappa = 1.0$ for temperature and $\kappa = 1.5$ for salinity. The larger κ for salinity is due to the small S_{mod} values (shown in Fig. 2) is not large enough to stand for the amplitude of the representation error. The model standard deviation S_{mod} is calculated using the simulated temperature (salinity) anomalies based on the control run for the period from 1985 to 2004 for the Pacific Ocean forced by NCEP forcing. For the error $\varepsilon_{\Delta t}$, we assume that it depends on both the time difference and depth. The impacts of depth can be represented by the variation of S_{mod} with depth. Thus, we express $\varepsilon_{\Delta t}$ as a function of S_{mod} and the time difference $|t^a - t^o|$ using the following equation:

$$\varepsilon_{\Delta t} = \kappa S_{\mathrm{mod}} |t^a - t^o| / 10.0 \tag{4}$$

where the coefficient κ has same meaning and value as in Eq. (3), and both the time difference $|t^a - t^o|$ and the number 10.0 are in time unit of days. We use Smod in Eq. (4) is due to the fact that we do not have enough observations to calculate observed standard deviation. Therefore, we assume that the ocean model can capture realistic standard deviation. However, we found that there are significant variations in Smod in both the horizontal and vertical directions. For example, there are very large Smod values in the thermocline at the equator and very small values at higher latitudes. Using these values to tune the amplitudes of the representation error and age error is difficult, and easily causes the overestimate/underestimate of the errors for some regions. To simplify this process, we average the Smod in the horizontal direction and get its vertical profile, namely that assuming the Smod is only the function of depth.[1] Fig. 2 shows that the Smod of temperature increases and peaks around 140 m and then decreases with the increasing depth. The Smod of salinity increases and peaks at 50 m and then decreases with the increasing depth. The vertical variation of the Smod is strong.

After the instrument, representation and age errors are estimated, the total observation error can be obtained by Eq. (2). The observation error is usually assumed to be uncorrelated between locations, thus its covariance matrix R_k is diagonal, obtained by the observation error of all data within the 11-day time window. In this study, the observation perturbation is drawn from the Gaussian distribution with the mean equal to zero and the variance equal to the error square. If we denote by $\overline{y_k^o}$ the observation vector used at the kth data assimilation cycle, the ith member of perturbed observation, $y_{k,i}^o$, can be expressed by

[1] It should be noted that this methods used to estimate observation errors is a little artificial and possibly affects the performance of the data assimilation.

$$y^o_{k,i} = \overline{y^o_k} + y'_{k,i} \tag{5}$$

where $y'_{k,i}$ is the perturbation.

3.2.2. Random model error

In estimating the random model error, we should consider both its amplitude and spatial structure. After the observation error covariance matrix R_k is given, the spread of the forecast error determines the Kalman gain. If the spread is too large, the background error covariance matrix will be overestimated and the analysis will tend to overfit the observations. Conversely, if the spread of the error covariance matrix is too small, the covariance matrix will be underestimated and the state analysis will tend to under utilize the observations (Turner et al., 2008). One newly developed method to solve this problem is using an inflator to adjust the amplitude of the background error covariance matrix (Houtekamer and Mitchell, 2001; Desroziers et al., 2005; Li et al., 2009; Zheng, 2009). This method assumes that the stochastic model error covariance matrix has a similar spatial structure but different amplitude to that of the forecast error caused by the errors in initial condition. Thus, the inflator is used to amplify the size of the background covariance matrix obtained from the ensemble members. Another method is to parameterize the error covariance matrix and then estimate the parameters by minimizing the −2 log-likelihood of observed-minus-forecast residuals (Dee and da Silva, 1998; Mitchell and Houtekamer, 2000). However, this method is very complex and impractical for a high dimensional model. A simple way of considering the random error is through the following three methods: (i) using stochastic equations, (ii) adding noise to the forecast ensemble at the analysis time (without integrating noise in the model), and (iii) using multimodel ensemble (Hamill, 2002). In this study, we used the method ii), i.e., adding noise to each member of the forecast ensemble. Thus, the forecast ensemble member $X^f_{k,i}$ is updated by

$$X^f_{k,i} = M(X^a_{k-1,i}) + q_{k,i} = X^p_{k,i} + q_{k,i} \tag{6}$$

where M is the nonlinear model, $X^a_{k-1,i}$ is the analysis at step $k-1$, $q_{k,i}$ is the random model error and $X^p_{k,i}$ is the model prediction at step k. The subscript denotes the ensemble member. We construct $q_{k,i}$, which is white in time but red in space, using pseudorandom fields produced following the scheme proposed by Evensen (2003). This procedure produces smooth pseudo random fields with mean equal to zero, variance equal to one and a specified covariance, which determines the smoothness of the fields. Considering the vertical coherence of the pseudorandom fields between adjacent model levels, we use a method to construct three-dimensional pseudorandom fields (Deng et al., 2009). In this study, we only perform data assimilation for the upper 24 model levels ($L = 24$) and leave other deeper levels to be adjusted by the model dynamics, due to (i) saving computational cost and (ii) simplifying assimilation process.

After a series of two-dimensional pseudorandom fields W_j ($j = 1, \ldots, L$) is constructed, the component of the three-dimensional pseudorandom field at jth level ε_j ($j = 1, 2, 3, \ldots, L$) can be constructed by the following equation:

$$\begin{aligned} \varepsilon_1 &= W_1 \\ \varepsilon_j &= \alpha_j \varepsilon_{j-1} + \sqrt{1-\alpha_j^2} W_j \end{aligned} \tag{7}$$

where $a_j \in [0 \quad 1]$ is an adjustable coefficient determining the correlation or coherent structure between level $j-1$ and level j. We assume that the vertical correlations between the model error fields of two adjacent levels can be approximated by the correlations between the variations of the variable. Thus, we use the output of the control run for the period from 1985 to 2004 to calculate the correlation coefficients of temperature anomalies and of salinity anomalies and used these correlation coefficients to estimate the parameter α_j. Following this procedure, a random model error is generated that is coherent in both the horizontal and vertical directions. The perturbation on the initial states is also constructed using the same way. In this study, the ensemble size is 61, i.e., randomly producing 61 perturbations superimposed onto model states to integrate the dynamical model forward to producing 61 prediction members.

After determining the spatial coherent structure of a random model error field, a major difficult and critical issue is how to determine its amplitude. Here, we determine the amplitude using a method similar in principle to the inflation scheme (Houtekamer and Mitchell, 2001; Desroziers et al., 2005; Li et al., 2009; Zheng, 2009). In the following, we present a brief description of the method. According to Mitchell and Houtekamer (2000):

$$\langle vv^T \rangle = H_k P^f_k H^T_k + R_k = H_k P^p_k H^T_k + H_k Q_k H^T_k + R_k \tag{8}$$

where $v = \overline{y^o_k} - H_k(\overline{X^f_k} - \overline{\beta^f_k})$ is the innovation vector, i.e., the difference between the observations and the forecast ensemble mean minus bias prediction interpolated to the observations. $\langle vv^T \rangle$ is the observation-minus-forecast covariance. $\overline{\beta^f_k}$ is the model bias, which is the deterministic part of the model error. In next subsection, we will discuss it in details. P^p_k is the prediction error covariance matrix calculated by:

$$P^p_k \approx P^p_e = \overline{(X^p_k - \overline{X^p_k})(X^p_k - \overline{X^p_k})^T} \tag{9}$$

where the overline denotes an average over the ensemble, X^p_k is the prediction ensemble containing the members $X^p_{k,i}$ ($i = 1, 2, \ldots, 61$). The model error covariance matrix Q_k is defined as

$$Q_k \approx Q_e = \overline{(q_k q^T_k)} \tag{10}$$

where $q_k = \sigma_k q^0_k$, and σ_k is an adjustable coefficient. From Eq. (8) we obtain:

$$tr(\langle vv^T \rangle) = tr(H_k P^p_k H^T_k) + tr(H_k Q_k H^T_k) + tr(R_k) \tag{11}$$

where tr is an operator that is used to obtain the trace of a matrix. After obtaining P^p_k, $\langle vv^T \rangle$ and R_k, the term $tr(H_k Q_k H^T_k)$ can be determined by Eq. (11). The coefficient σ_k, which will be used to determine the amplitude of the pseudo random field, can be calculated using the following equation:

$$\sigma_k = \sqrt{\frac{tr(HQ_k H^T)}{N_k - 1}} \tag{12}$$

where N_k is the total number of the observations. Thus, the final stochastic model error field for a variable is calculated by

$$q_{k,i} = \sigma_k q^0_{k,i} \tag{13}$$

Using this method, an adaptive spatial coherent model error ensemble is constructed for each data assimilation cycle. A final forecast ensemble X^f_k is constructed by summing up the random model error ensemble q_k and the prediction ensemble X^p_k using Eq. (6). Thus, the background error covariance matrix P^f_k can be computed using this final forecast ensemble:

$$P^f_k = P^p_k + Q_k \approx P^f_e = \overline{(X^f_k - \overline{X^f_k})(X^f_k - \overline{X^f_k})^T} \tag{14}$$

The Kalman gain is then calculated by

$$K^X_k = (\rho \circ P^f_k) H^T_k (H_k(\rho \circ P^f_k) H^T_k + R_k)^{-1} \tag{15}$$

where ρ is a covariance localization function defined by

$$\rho(x, y) = e^{\left(-\frac{d^2_x}{L^2_x} - \frac{d^2_y}{L^2_y}\right)} \Big/ \cos^2(y) \tag{16}$$

where d_x and d_y are longitude and latitude differences between neighboring grids and the grid assimilated, L_x and L_y are associated decorrelation scales at the equator, L_x is taken as 15° and L_y is taken as 7.5° in this study; x and y are indices of the model grid representing longitudes and latitudes; and the open circles between ρ and P_k^f denote a Schur product (an element by element matrix multiplication). Considering the effects of latitude on the decorrelation scale, a factor $\cos^2(y)$ is used to adjust the decorrelation scales in both the latitude and longitude directions. The "covariance localization" in Eq. (16) reduces spurious correlations between distant locations in the background covariance matrix P_k^f, which is caused by the limited ensemble size (Gaspari and Cohn, 1999; Houtekamer and Mitchell, 2001; Oke et al., 2005).

3.2.3. Model bias

The generalized two-step bias-correction algorithm for the kth analysis cycle in the EnKF can be rewritten as below (Dee, 2005):

$$\beta_{k,i}^a = \beta_{k,i}^f - K_k^\beta\left[y_{k,i}^o - H_k\left(X_{k,i}^f - \beta_{k,i}^f\right)\right] \tag{17}$$

$$X_{k,i}^a = (X_{k,i}^f - \beta_{k,i}^a) + K_k^X\left[y_{k,i}^o - H_k\left(X_{k,i}^f - \beta_{k,i}^a\right)\right] \tag{18}$$

where the superscripts f and a refer respectively to the forecast and analysis of a given variable, subscript i refers to the ith ensemble member, k refers to the kth analysis cycle. K_k^β and K_k^X are gain matrices for the bias and state estimation, respectively. Since the bias is slowly varying in time and errors of the bias estimation are much smaller than the background errors, the gain of the bias can be simplified as follows (Dee 2005):

$$K_k^\beta = \gamma K_k^X \tag{19}$$

where $\gamma(\ll 1)$ is a small constant which controls the adaptivity of the bias estimate. In this study we set $\gamma = 0.01$. Following, we will describe the method of estimating the model bias.

The model bias varies slowly and has a mean different from zero, it can be estimated using the difference between the model state and the observations under the assumption that the observations are unbiased (Drecourt et al., 2006). The bias value of a state variable on a model grid could be defined as the mean difference averaged over a neighboring region (Chu et al., 2004; Frank and Colby, 1997). In this study, we use the difference between the analysis and the observations to define the bias after kth analysis step finished:

$$\beta_k = \frac{1}{N_k}\sum_{|dx|<l_x,|dy|<l_y}(H_k\overline{X_k^a} - \overline{y_k^o}) \tag{20}$$

where $|dx| < l_x$, $|dy| < l_y$ defines a rectangle neighboring region around the model grid point, $|dx|$ and $|dy|$ are spatial differences between the model grid point and the locations of the observations in the zonal and meridional directions, and we set $l_x = 1500$ km, $l_y = 500$ km in this study. N_k is the number of the observations within the region at time k, H_k the measurement operator, $\overline{X_k^a}$ the mean of the analyzed state ensemble X_k^a, $\overline{y_k^o}$ the observation vector. In Eqs. (8) and (17) the bias forecast is needed. However, before the kth data assimilation cycle finished, the bias cannot be calculated because the $\overline{X_k^a}$ is not available. To solve this problem, we approximate the bias using its value at the previous assimilation step according to the fact that the bias varies slowly. Since the bias is mean error, we could further assume that the bias difference between ensemble members is much smaller comparing the amplitude of the bias mean. Therefore, we may approximate the bias at the kth assimilation step by:

$$\begin{aligned}\beta_{1,i}^f &= 0\\ \beta_{k,i}^f \approx \overline{\beta_k^f} &\approx \beta_{k-1} = \frac{1}{N_{k-1}}\sum_{|dx|<l_x,|dy|<l_y}(H_{k-1}\overline{X_{k-1}^a} - \overline{y_{k-1}^o})\end{aligned} \tag{21}$$

Thus, combining Eqs. (17)–(21), the bias analysis of each ensemble member can be approximated by:

$$\beta_{k,i}^a \approx \overline{\beta_k^a} = \beta_{k-1} - \gamma K_k^X\left[y_k^o - H_k\left(\overline{X_k^f} - \beta_{k-1}\right)\right] \tag{22}$$

After correcting the bias in the state forecast, a new ensemble is produced with the ith member:

$$X_{k,i}^e = X_{k,i}^f - \overline{\beta_k^a} \tag{23}$$

Then Eq. (18) becomes

$$X_{k,i}^a = X_{k,i}^e + K_k^X\left[y_{k,i}^o - H_k\left(X_{k,i}^e\right)\right] \tag{24}$$

The final state analysis for the kth analysis cycle is obtained using the mean of the analysis ensemble:

$$\overline{X_k^a} = \frac{1}{n}\sum_{i=1}^{n} X_{k,i}^a \tag{25}$$

In summary, we use the mean difference between the model analysis and observation of previous step as the bias forecast at the present step. The bias forecast is analyzed by (17) for the current step, producing the bias analysis that is used to correct model states by (18) and serves as the bias prediction of the next step.

3.2.4. Local analysis

The local analysis scheme proposed by Bishop et al. (2001) is used to relieve the burden of computation. It assimilates all observations that may affect the analysis at a given grid point simultaneously and obtains the analysis independently for each model grid point (Houtekamer and Mitchell, 2001; Ott et al., 2004; Szunyogh et al., 2008; Hunt et al., 2007). In this study, we perform local analysis grid by grid for the top 24 levels. To obtain the analysis of a model grid point, we define a local domain around the grid point, and all observations within this domain are assimilated. The domain is 3000 km in longitude, 1500 km in latitude and 3 model levels in the vertical. The grid point is at the center of the rectangle horizontal region and in the middle level. Having carried out the assimilation on all grids, all analyses combine a final analysis ensemble. It is clear that the horizontal region is the same area of the bias forecast estimation in Eq. (20).

As a sequential data assimilation method, the EnKF has the significant drawback of the time discontinuity of the solution resulting from intermittent corrections of the model state. This discontinuity can lead to spurious high frequency oscillations. To avoid exciting these waves and allow the model dynamics to adjust gradually to the changes in the density field, an algorithm called incremental analysis update (IAU) proposed by Bloom et al. (1996) is used in this study. The principle of this method is to incorporate the sequential analysis increment directly in prognostic equations of the model as an additional forcing term (Balmaseda, 2007; Castruccio et al., 2008). In practice, the increment is added slowly over the subsequent 5 days, after which a new background field is produced and the assimilation cycle is repeated. Details about this method were described by Castruccio et al. (2008).

3.3. Experimental setup

Since the in situ subsurface observations are sparse in the Pacific, to effectively use the available observations for the assessment of the performance of the system, we design four experiments:

(i) No-cross-validation assimilation experiment. In this experiment, all Argo profiles and other observations are assimilated. The assimilation performance is evaluated against the NCEP re-analysis data and satellite SLA data. The analyzed model states of this experiment are also used to initialize a hybrid coupled ENSO prediction model for hindcast experiments.

(ii) Cross-validation assimilation experiment. In this experiment, the Argo dataset is randomly split into four independent groups, each containing 25% of Argo profiles. Four assimilation runs are performed respectively, each using three of the four groups Argo data (75%) and the other observations and withholding a group (25%) for validation that are never used in the assimilation. As such, four runs generate a complete withheld dataset (100%) that is used for the validation.

(iii) A simple data assimilation experiment. In this experiment, the data used are same as (ii), but without considering the bias correction ($\beta^a_{k,i} = 0$) and the additive random model error ($q_{k,i} = 0$). A multiplicative inflator λ_k is used to avoid the underestimation/overestimation of background error covariance and determined by

$$\lambda_k = \frac{tr(\langle vv^T \rangle) - tr(H_k P^p_k H^T_k) - tr(R_k)}{tr(H_k P^p_k H^T_k)} \tag{26}$$

Thus, the amplified background error covariance matrix is

$$P^f_k = (1 + \lambda_k) P^p_k \tag{27}$$

This simple experiment severs as a justification that some complicated strategies used in (ii) are necessary in developing a realistic oceanic data assimilation system.

(iv) Using the analyses from (i) as initial conditions, ENSO hindcast experiments are conducted until the leading time of 12 months for the period from 2005 to 2007. For reference, hindcast experiments initialized from the control run for the same time period are also performed.

A run with no data assimilation is also performed (CTL), as the reference for comparison. All the experiments are performed for the period of 3-year from January 1, 2005 to December 31, 2007.

4. Methods for validation

To quantify the impacts of the data assimilation on the model ocean states, some statistics, including mean difference between model and observation (MD), root mean square error (RMSE) and ENSO hindcast correlation skill are calculated. As the observations are not on model grids, these statistics are estimated based on sub-regions or/and grid cells. The MD and RMSE are calculated using the following equations:

$$MD = \frac{1}{N} \sum_{i=1}^{N} (Y^m_i - Y^o_i) \tag{28}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (Y^m_i - Y^o_i)^2} \tag{29}$$

where Y is temperature or salinity; the superscript "o" denotes the independent observations not used in the assimilation and "m" denotes the model equivalents at the observation positions and time; the subscript i indicates the individual measurements within a specific sub-region or grid cell and N is the total number of the observations within it.

Providing high quality initial conditions for ENSO prediction is an important goal of ocean data assimilation in the tropical Pacific. Thus, a hybrid coupled ENSO prediction model is used to test whether the assimilation can significantly improve ENSO hindcast skill compared with the control run. The hybrid coupled model is identical to that used in Deng and Tang (2009) and Deng et al. (2009) which is composed of the same OGCM used in the data assimilation system coupled to a statistical atmospheric model. The hybrid coupled model is initialized with the analyzed states from the assimilation experiment (i) and from the control run respectively. Since the adjustment of the ocean model to the assimilation needs some time, the analyzed states of the first 3 months are discarded when we performed hindcast experiments. As a result, the hindcast experiments start from the first day of each month and last 12 months for the period from April 2005 to December 2007. Finally, 33 hindcast results are available and are used to estimate hindcast skill, the correlation and root mean square error (RMSE), by comparing predicted sea surface temperature anomalies (SSTA) against observed counterparts from the ERSST.V2 (Smith and Reynolds, 2004).

The amplitude of uncertainties contained in an analysis is a major concern when using the analysis. It can be measured by the spread of the analysis ensemble in the EnKF. The mean spread at the model grid point for the whole assimilation period (3-year) can be calculated using the following equation:

$$s = \sqrt{\frac{1}{mn-1} \sum_{j=1}^{m} \sum_{i=1}^{n} \left(X^a_{j,i} - \overline{X^a_j} \right)^2} \tag{30}$$

where $X^a_{j,i}$ is the value of an analysis member, $\overline{X^a_j}$ is the ensemble mean, the subscript j indicates m assimilation cycles, and i, the n individual members. The spreads for temperature and salinity at each model grid point in all levels are calculated.

5. Results

In this section, we validate the Argo data assimilation system by the T and S fields from the withheld Argo observations, satellite remote sensing SLA and the heat content and zonal currents along the equator from the NCEP re-analysis data. A comparison of the hindcast skills at 6-month and 12-month leads of the tropical sea surface temperature anomalies (SSTA) initialized from the assimilation experiment (i) and from the CTL is also performed.

5.1. Comparison of mean difference (MD)

A comparison of MDs of temperature and salinity for all model levels is performed. For simplicity, only the MDs of three levels, at 10 m, 140 m and 500 m, are displayed here. The three levels are chosen to represent the near surface, thermocline and deeper ocean layers respectively. Note that (i) there is not sufficient Argo observations at surface (0 m) to compare; (ii) the depth of 140 m is the place where there is maximum RMSE of temperature along the equator; (iii) at depths deeper than 500 m only Argo observations are available and other observations are sparse; and (iv) at deeper levels (>500 m) the MD is relatively small.

Fig. 3 shows spatial distribution of temperature MD for the CTL, simple experiment and cross-validation assimilation run at the three levels, against the withheld Argo profiles. In the CTL, there are considerable discrepancies between model and observation, as shown in the left panels in Fig. 3. In the central and eastern equatorial Pacific, the model produces colder near surface temperatures but warmer subsurface temperatures (140 m) than those observed in the real ocean. In the western equatorial Pacific the model presents colder temperatures for all of the three levels,

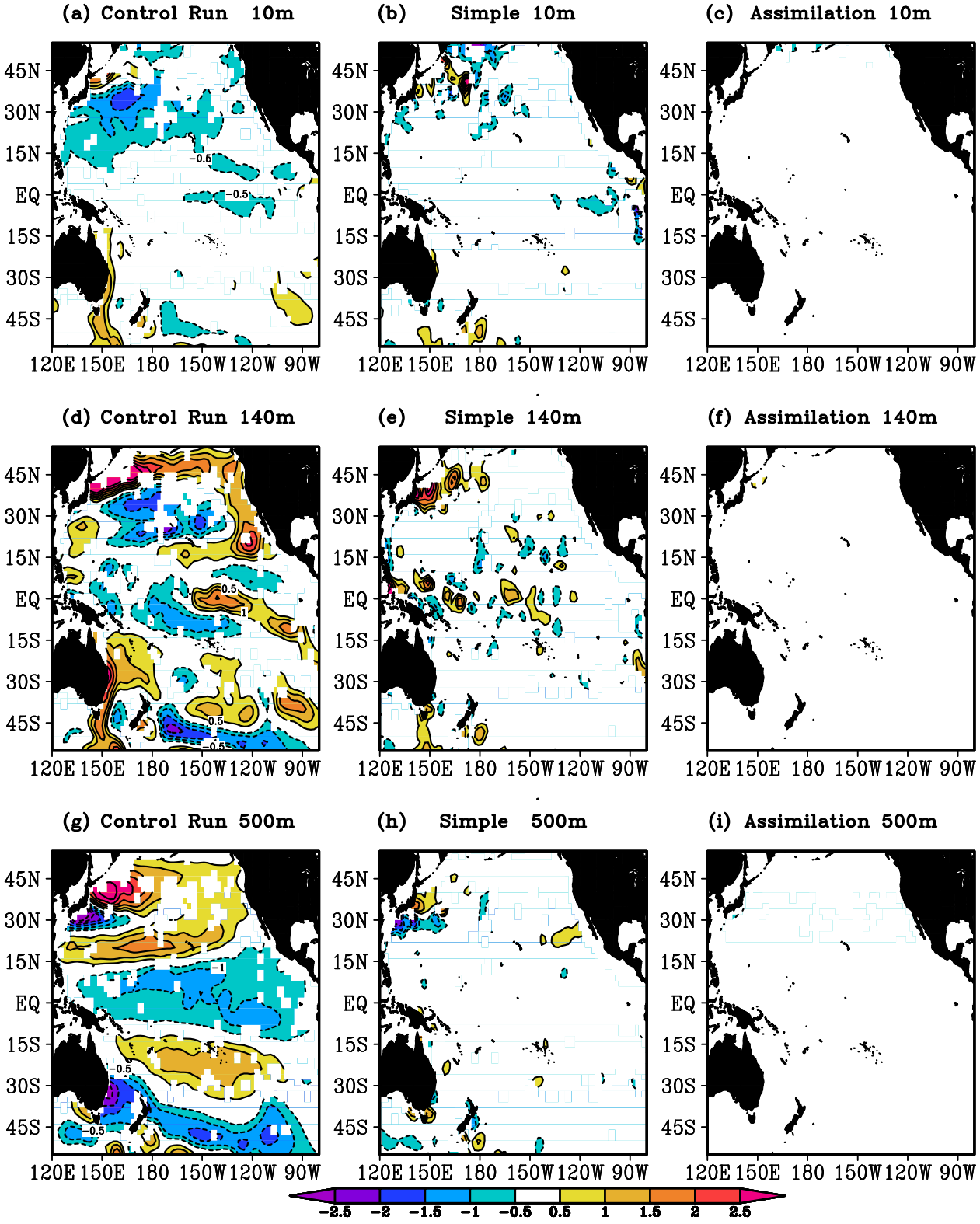

Fig. 3. Differences between model and observation temperature averaged over the 3-year period at 10 m (top), 140 m (middle) and 500 m (bottom) for CTL (left), the simple assimilation (middle) and cross-validation assimilation run (right). Contour interval is 0.5 °C and the zero lines are ignored. Areas with absolute MD over 0.5 °C are shaded.

making the thermocline depth shallower than in the real ocean, causing the west–east slope of the thermocline to be reduced. Outside the equatorial region, the near surface water is obviously cooler than observation over most areas, and the minimum negative temperature MD values appear in the Kuroshio area. Some warmer areas mainly appear in the Southeastern Pacific, Northwestern Pacific and around Australia. At 140 m and 500 m, there are also significant large-scale MDs indicating the existence of model biases in temperature field. In contrast to the CTL, both the simple and cross-validation assimilation run significantly reduce these

biases over almost the whole region at all levels. However, the simple assimilation scheme cannot remove the biases in the surface and subsurface near the Kuroshio and in the equatorial eastern Pacific.

Similarly, the salinity MDs between the model and the observations are shown in Fig. 4. In the CTL, considerable errors appear at all levels. The dominant feature is that the model underestimates the salinity in the tropical Pacific whereas in mid latitude salinity is overestimated as compared to the observed salinity. These considerable large-scale biases in the salinity field are also significantly reduced in both the simple and the cross-validation assimilation runs. However, Fig. 4b, e and h shows that the simple data assimilation scheme cannot thoroughly remove the large

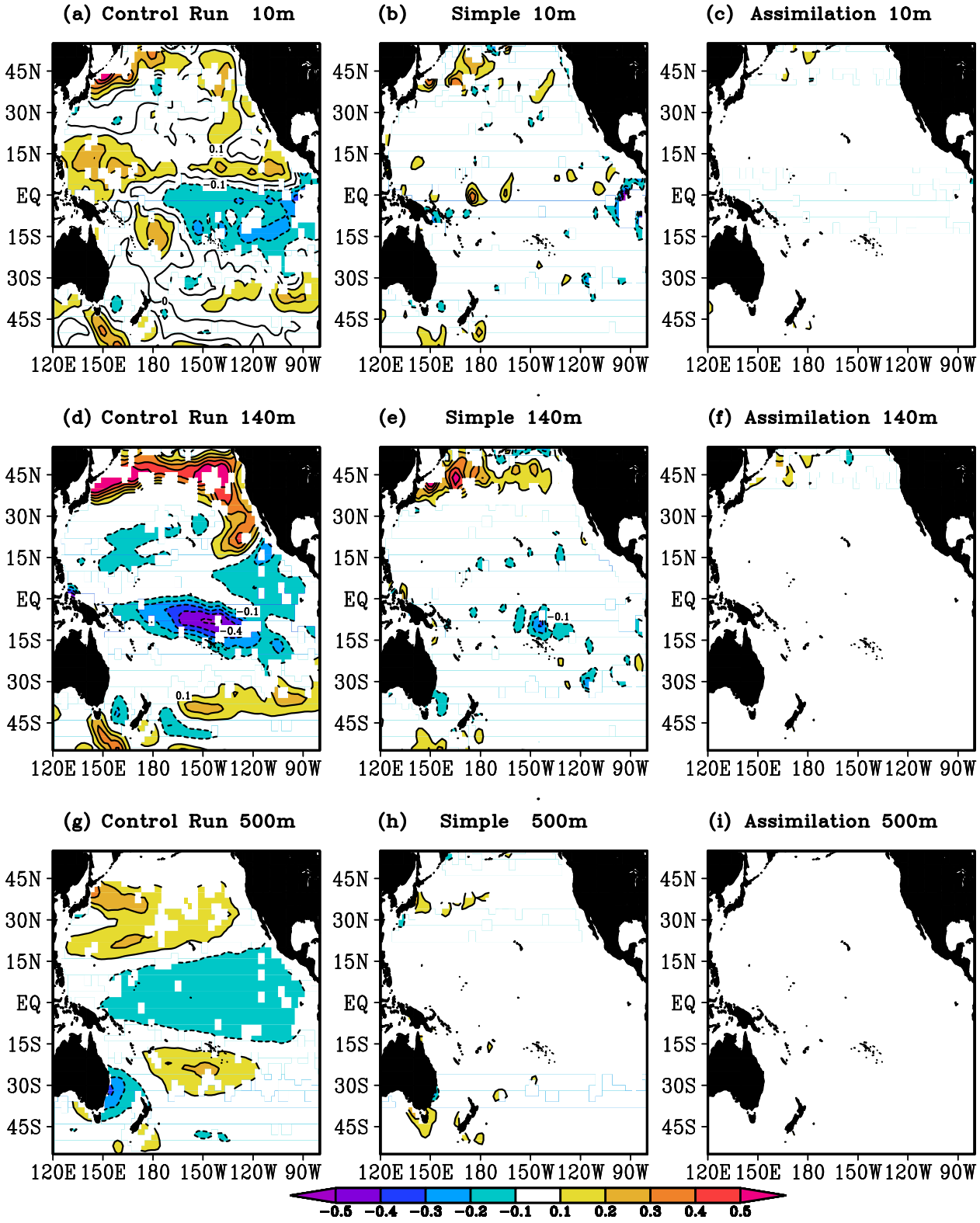

Fig. 4. Differences between model and observation salinity averaged over the 3-year period at 10 m (top), 140 m (middle) and 500 m (bottom) for CTL (left), the simple assimilation and cross-validation assimilation run (right). Contour interval is 0.1 psu.

biases in some regions, for example the biases in the Kuroshio area at all levels. The significant contrasts between the CTL and cross-validation run shown in Figs. 3 and 4 demonstrate that the cross-validation assimilation scheme is capable of correcting the model biases appearing in the CTL. It should be noted that these comparisons are based on the period 2005–2007, during which there are two El Niño events (2004–2005, 2006–2007) and two La Nina events (early 2006 and 2007–2008). These events are responsible for the large interannual variations of ocean temperature. The cooler upper ocean temperature in the CTL might be due to the El Niño events that cause the real ocean warmer than normal years. However, the ocean model may be incapable of simulating adequately the large positive anomalies corresponding to these El Niño events. The mean errors of the temperature and salinity by the simple assimilation scheme show that this method is incapable of removing biases in some regions. Therefore, it is necessary to develop a bias-correction assimilation scheme for this model as discussed above.

5.2. Comparison of root mean square error (RMSE)

The RMSE is attributed to both the mean and anomaly differences between the model and observations. In this study, we have analyzed the combined contributions of these two differences according to Eq. (29). Fig. 5 shows the averaged RMSE, as a function of depth, for temperature (Fig. 5a) and salinity (Fig. 5b) over the entire Pacific Ocean. For reference, the RMSE profiles of the no-cross-validation are plotted in the same figures as well. Fig. 5a shows that the temperature RMSE ranges from 0.2 °C to 1.7 °C and peaks at the depth of the thermocline for the CTL. Fig. 5b shows that the salinity RMSE ranges from 0.02 psu to 0.25 psu and peaks at a depth around 100 m for the CTL. The small temporal variation causes the decreased RMSEs at deeper levels. With the data assimilation, both the RMSEs of temperature and salinity significantly reduce at all levels to about less than half of the values in the CTL. The simple assimilation run produces smaller RMSEs of temperature than the CTL but still larger than the cross-validation run for all levels. The difference of RMSEs between the simple assimilation run and cross-validation assimilation run is not significant for salinity. Comparing the temperature RMSEs of the cross-validation experiment and no-cross-validation experiment reveals that their differences are small. This suggests that Argo temperature profiles have provided sufficient information for the ocean analysis at the given model resolution. Thus, the reduction in number of Argo profiles in the cross-validation assimilation does not significantly impact the performance of the temperature assimilation. Another possible reason is that there are plenty of other temperature profiles used in both the cross-validation and no-cross-validation experiments.

In contrast to temperature, the differences of the salinity RMSEs between the two experiments are relatively large, implying that the reduction in profile number significantly impacts the salinity assimilation. This is reasonable because only Argo salinity profiles are available and there are no other salinity profiles to supplement the reduction in number of Argo salinity profile in the cross-validation assimilation.

Above comparisons show that the bias-correction assimilation run can produce better results than the simple assimilation scheme. For simplicity and clarity, in the following parts, we only focus on oceanic analysis from bias correction with cross-validation scheme, against the CTL.

In Fig. 6, the spatial distributions of temperature RMSEs at the three levels for the CTL and cross-validation assimilation experiment are shown. The RMSEs are calculated for the 3-year period using Eq. (29). There are large temperature RMSE values (>1.0 °C) in some regions, for example the North Equatorial Countercurrent

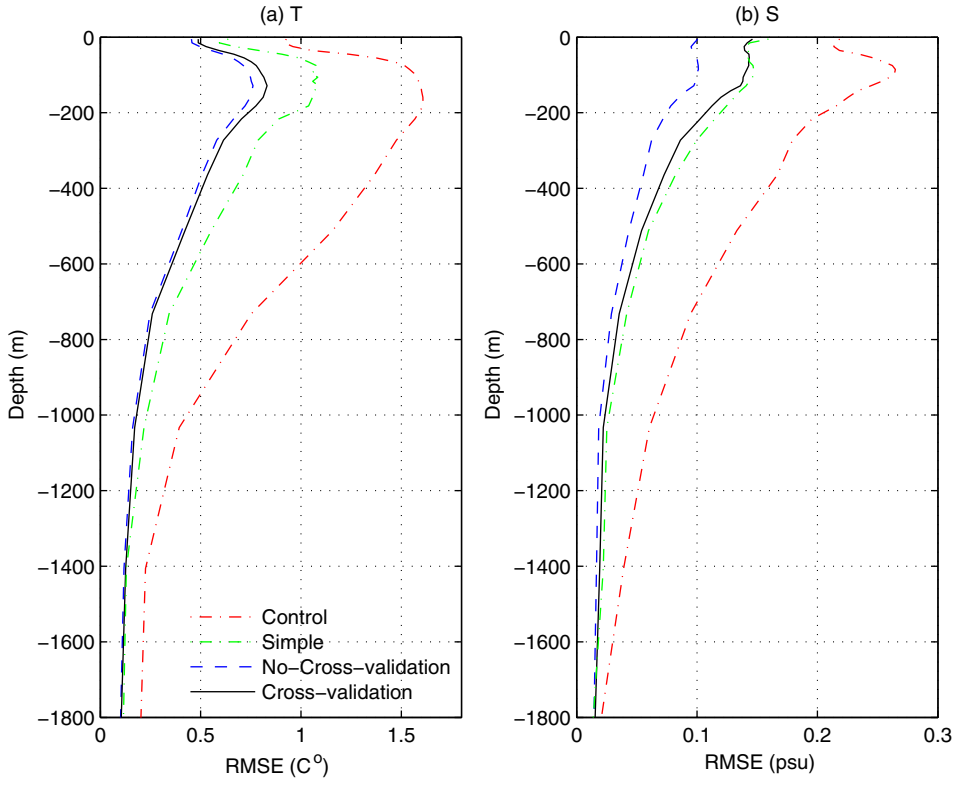

Fig. 5. RMSEs of temperature (a) and salinity (b), as a function of depth, from analyses of the control experiment (red dash-dotted line), simple assimilation (green), no-cross-validation assimilation (blue) and cross-validation assimilation run (black) for the Pacific Ocean, where the observations are Argo profiles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(NECC), the South Equatorial Current (SEC), the Kuroshio and West Wind Drift. This probably associated with strong currents there. Comparing the spatial distributions of RMSEs (Fig. 6a, c and e) and that of MDs (Fig. 3a, d and g) shows similar locations of some centers, i.e., large RMSE centers generally correspond to positive (or negative) MD centers, suggesting the dominant impacts of the model biases on the RMSEs in some regions in the CTL. The similarity in the distributions between the MD and RMSE centers implies that locations where the model has large biases may have large anomaly error as well. With the data assimilation, the RMSEs of temperature on the three levels (Fig. 6b, d and f) are significantly reduced. However, over some sub-regions, such as the Kuroshio regions, around Australia, etc., there are still relatively large RMSE values, which are not thoroughly removed by the data assimilation. The possible causes for these large RMSE values are the insufficiency of observation, coarse model spatial resolution, improper expression of the forecast error and complicated dynamics that the ocean model does not resolved.

Similar to the temperature RMSE, the spatial structures of the salinity RMSEs for the CTL at the three levels (shown in Fig. 7a, c, and e) have some local maximum centers as well. The positions of some centers generally correspond to the large MD values shown in Fig. 4a, d and g, and they appear approximately at the same locations as the maximum temperature RMSE values, suggesting the existence of some obvious model errors. These errors are mainly caused by coarse model resolution, unrealistic physics presentation and unreasonable external forcing. With the data assimilation, the amplitudes of the salinity RMSE at the three

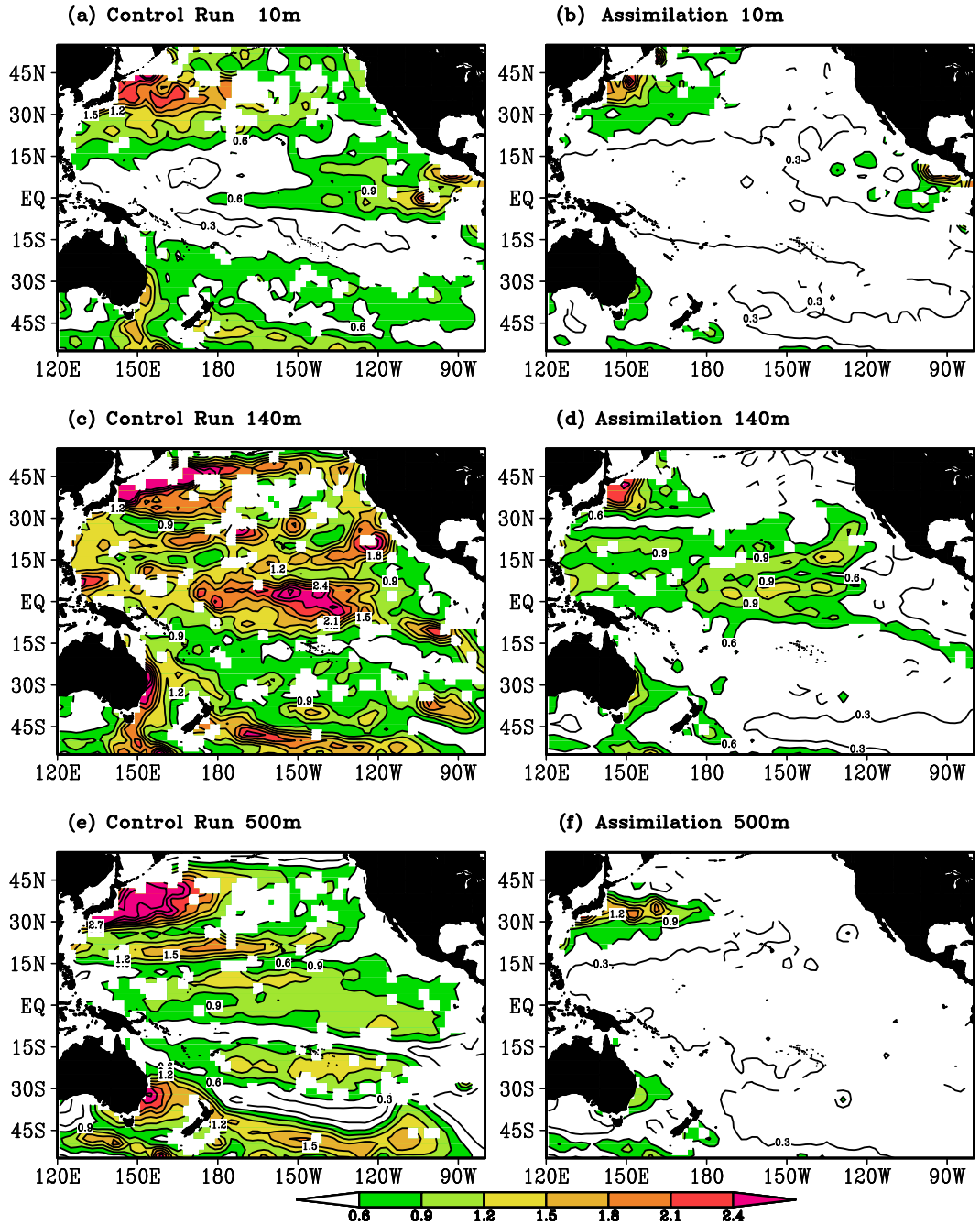

Fig. 6. RMSE of modeled subsurface temperature from (left) the control experiment and (right) the cross-validation run at depths (a, b) 10 m, (c, d) 140 m and (e, f) 500 m relative to Argo profiles during January 2005–December 2007. Contour interval is 0.3 °C. The areas with RMSE over 0.6 °C are shaded.

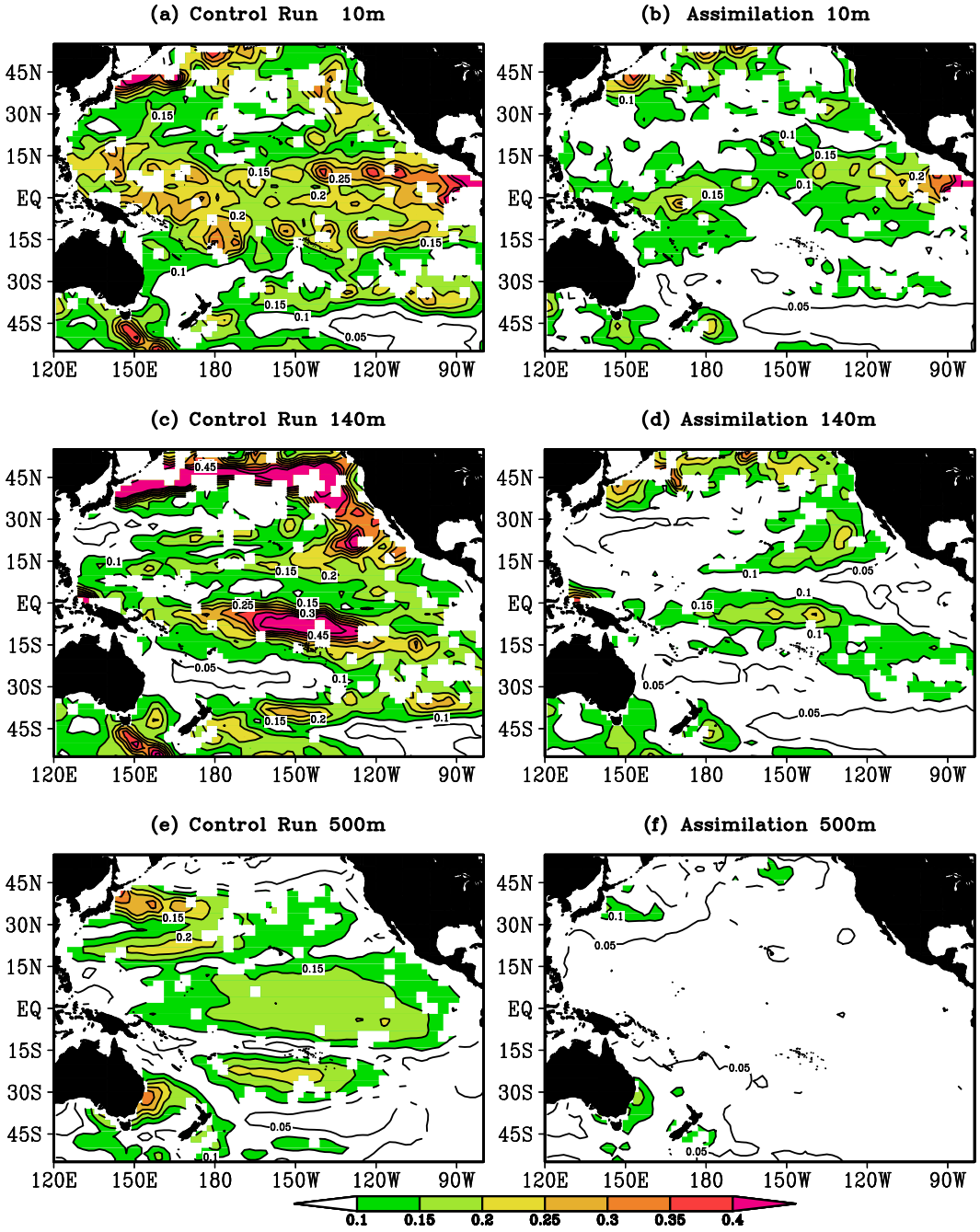

Fig. 7. RMSE of modeled subsurface salinity from (left) the control experiment (right) and Exp3 at depths (a, b) 10 m, (c, d) 140 m and (e, f) 500 m. Contour interval is 0.05 psu. The areas with RMSE over 0.1 psu are shaded.

levels, shown in Fig. 7b, d and f, are significantly reduced. The above results demonstrate that the assimilation scheme can improve the oceanic temperature and salinity simulation over the entire Pacific Ocean at all levels.

5.3. Comparison with NCEP re-analysis results

In the last section, we have shown that the assimilation scheme improves the simulation of the temperature and salinity fields. Despite the comparisons are performed using independent observations, these improvements are still not surprising because though the comparisons are based on the independent data, the two variables compared are the observational variables assimilated into the model. In this section, we will perform some further comparisons using the NCEP re-analysis data set. The NCEP re-analysis is much easier and more convenient to use compared with sparse and sporadic subsurface in situ observations, and it is often used as a proxy for grid observations in many studies. We will compare the upper ocean heat content anomaly (HCA) and zonal current between the simulations and the counterparts of the NCEP re-analysis. We calculated the HCA for the upper ocean from the surface down to 250 m along the equator for the 3 years from 2005 to 2007. Because the NCEP data are not used in the assimilation system, the results of the no-cross-validation assimilation experiment are used for the comparisons. The time evolutions of the HCA for the CTL, assimilation run and NCEP are shown in Fig. 8. In this figure, the HCA of the CTL and assimilation run are calculated relative to the seasonal cycle climatology based

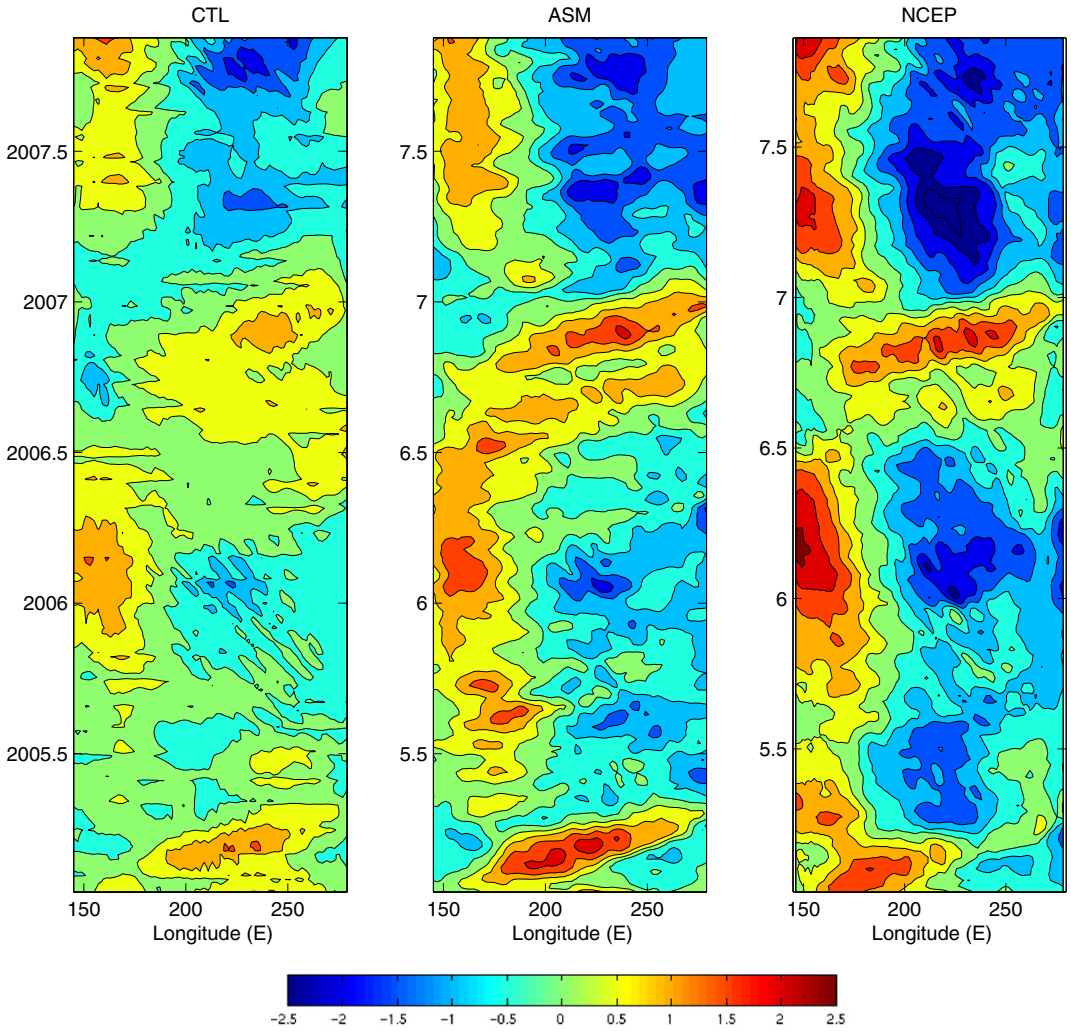

Fig. 8. Time–longitude plots of HCA along the equator during 2005–2007 from (left) the control experiment, (middle) the Exp1, and (left) the observations. Contour interval is 0.5 °C.

on the 20-year (1985–2004) control run. The anomaly for the NCEP is relative to its seasonal cycle climatology for the same period. One can see that the assimilation result is more similar to the NCEP result than the CTL. Comparing the result of the CTL with that of the NCEP reveals that the CTL can capture the general features of the HCA evolution with time, especially some large amplitude centers can be simulated by the model. However, the amplitude of the HCA of the CTL is much weaker than that of the NCEP product. In contrast to the CTL, the data assimilation results are much closer to the NCEP re-analysis product in both phase and amplitude, such as the increased amplitude and similar locations of maxima and minima.

Fig. 9 shows the zonal current anomaly at a depth of 15 m along the equator for the 3-year period for the CTL, assimilation run and NCEP re-analysis. Fig. 9a demonstrates negative zonal current anomalies propagating from the eastern equatorial Pacific to the western equatorial Pacific from spring to fall in each year. Along with that there are some small scale easterly zonal current anomaly propagating westward starting from different longitude during periods when the zonal current anomalies are negative. These small scale fast propagating waves are unrealistic and may be associated with the unstable waves excited by the free surface scheme used in the model. In contrast to the CTL, the assimilation run presents a more realistic zonal current time evolution and spatial pattern (Fig. 9b) compared with that of the NCEP re-analysis (Fig. 9c). It should be noted that the current velocity is an independent variable since it was not assimilated into the model. Fig. 9 suggests that the effects of the data assimilation on the zonal current velocity are produced by the dynamical adjustment of the model.

5.4. Comparison with satellite SLA observation

Sea level anomaly is a comprehensive indicator of upper ocean temperature, salinity and mass distribution, etc. A comparison between the model SLA and observation SLA is useful to show the impacts of the data assimilation on the ocean state estimation. In this study, the SLA observations are satellite data for the same period of 3-year. The resolution of the SLA is higher than that of the ocean model. The purpose of this comparison is to explore the ocean analysis in term of capturing large-scale characteristics of real SLA variation, thus we do not have to interpolate the SLA onto the model grid. Fig. 10 shows the time evolutions of the SLA for the CTL, assimilation run and satellite observations along the equator. Comparing the SLA time evolutions shows that the assimilation run presents a more realistic simulation than the CTL, especially during the 2006–2007 El Niño period. The amplitude of SLA in the CTL is much weaker relative to the assimilation run and observations. In the observations, there are many small-scale anomalies, which are not captured by both the CTL and assimilation run due to the coarse spatial resolution of the OGCM.

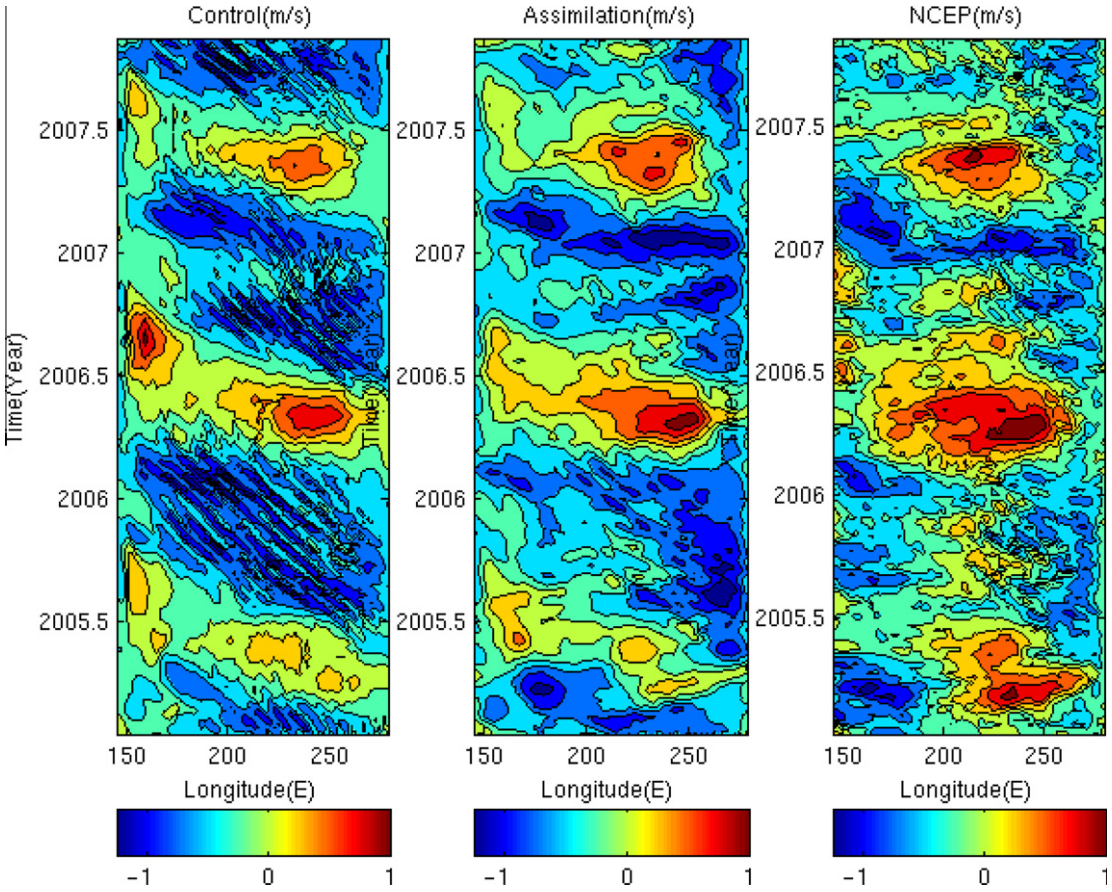

Fig. 9. Zonal velocity in the tropical Pacific at the equator at 15 m depth during the period 2005–2007: (left) control run (middle) assimilation run and (right) the NCEP re-analysis.

In addition, both the CTL and assimilation run cannot accurately capture a realistic sea–saw structure between the east and west, especially during the first half of 2007. This result suggests that there are limitations in the data assimilation system when it is used to improve SLA simulation.

5.5. The comparison of ENSO hindcast skill

In this subsection, we will explore the role of Argo assimilation in the seasonal climate prediction. The comparison of ENSO prediction skills from two initialization schemes, i.e., CTL and assimilation analysis, also serves as a further verification of the assimilation performance. It has been well recognized that ENSO is basically an initial value problem, and initial condition from which a prediction start to run greatly determines prediction skill.

A hybrid coupled model is applied to perform ENSO hindcast experiments. It is composed of the OGCM and a linear atmospheric model, as reported in Deng and Tang (2009). Starting from the initial conditions provided by the assimilation run and CTL, 33 hindcast experiments are conducted respectively for the period from April 2005 to December 2007. Based on the results of these experiments the sea surface temperature anomalies (SSTA) over the tropical Pacific are calculated. The observed SSTA for the same period are computed with respect to the seasonal cycle from 1971 to 2000. The correlation and RMSE skills at 6-month lead and 12-month lead are shown in Figs. 11 and 12, respectively. If we define the correlation coefficient greater than 0.35 (0.05 significant level) as a useful skill, both the predictions initialized from the CTL and assimilation present useful skills over the Niño 3 region (5°S–5°N, 150°W–90°W) at the 6-month lead as shown in Fig. 11. However, the predictions initialized from the assimilation presents useful and much higher skills than those initialized from the CTL over the Niño 4 region (5°S–5°N, 160°E–150°W), where the control predictions cannot provide useful skill at all. The correlation differences shown in Fig. 11e verify this conclusion. Comparing Fig. 11b against Fig. 11d shows that both of the prediction experiments have large RMSE values along the equator and the predictions initialized from the assimilation present smaller RMSE values over most areas (shaded in Fig. 11f). Fig. 12 shows that at the 12-month lead the predictions initialized from the assimilation present useful correlation skills over the Niño 3 region, whereas the one initialized from the CTL have no useful skill over the same region. Fig. 12e shows that the improvements mainly appear over areas along the equator. Fig. 12b, d and f shows similar improvements as shown in Fig. 11b, d and f. It should be noted that the assimilation cannot improve the correlation skill over some areas and the reduction in RMSE is always smaller than 0.2 °C. Although the improvements are limited over some areas, the Argo data assimilation system is capable of improving ENSO hindcast skill, especially the correlation skill.

5.6. The estimation of uncertainty of the ocean state analysis

As discussed above, the assimilation system can improve the ocean state estimation and ENSO prediction skill. However, because of the imperfections in both the OGCM and data assimilation scheme, as well as the errors in the observations, the assimilation analyses still contain some uncertainties. According to the principles of the EnKF, the forecast errors are assumed to be proportional to the spread of the ensemble before the assimilation and the

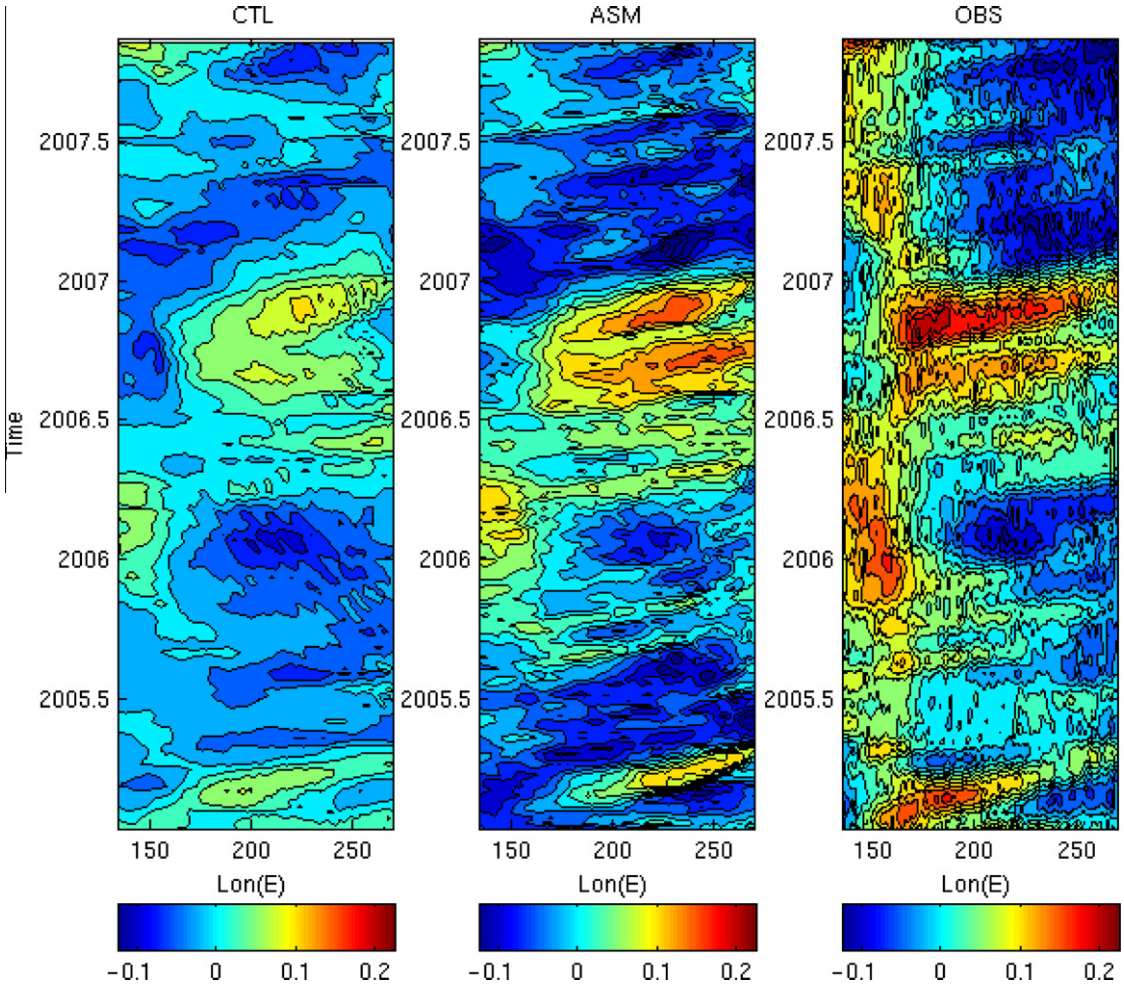

Fig. 10. Time-longitude plots of SLA along the equator during 2005–07 from the control experiment (left panel), the Exp1 (middle panel), and the observations (right panel). Contour interval is 2.5 cm.

analysis error is assumed to be proportional to the spread of the ensemble after assimilation. Thus, we can estimate the analysis errors using Eq. (30). Here, we focus on the spatial distributions of the analysis errors. For consistency with the presented results, we only present the spreads of the selected three levels in Fig. 13. This figure shows that the time-averaged ensemble spreads of the analyses vary in space. The uncertainties of the temperature and salinity analyses share some common characteristics, for example the relatively large uncertainties in the Northeastern Pacific and the Southeastern Pacific, the maximum values that appear along the latitudes of about 40°N and 30°S, the coasts of North America, Australia, New Zealand and Indonesia. In general, the maximum spread centers at the three levels appear at similar locations. What factors determine the spatial pattern of the spread is an interesting topic. Comparing Fig. 13 with Fig. 1 reveals that the small spread areas usually correspond to the areas of dense Argo observations and vice versa, for example the area between 5°N and 30°N where the spreads are small and Argo observations are dense, and the area around 30°S where the spreads are large and Argo observations are sparse. This is not surprising, because the analysis step is equivalent to a forcing added to the model equations and this observational forcing in principle reduces the error growth rate and the number of unstable directions with respect to those of the original system (Carrassi et al., 2007). The observations constrain the model not to diverge too far away from real state. So, in regions where there are more observations, the members of the ensemble converge and in turn the spreads of the analysis ensembles decrease and vice versa.

6. Summary and discussion

The objective of this study is to construct an assimilation system for the Pacific Ocean for effectively assimilating Argo profiles and other in situ ocean observations by the state-of-the-art EnKF techniques. The OGCM used is the primitive equation ocean model OPA9.2. To deal with model biases, a method similar to the simplified two-stage bias-correction scheme (Dee, 2005) is used in the system. The stochastic model errors are simulated using pseudo random fields, which are white in time but have coherent structure in space and constructed by the method proposed by Evensen (2003). The amplitudes of these errors are determined according to an adaptive error estimation scheme similar to the method proposed by Mitchell and Houtekamer (2000). All analyses are conducted locally. To avoid the impacts of significant instable waves excited by large increments during the assimilation cycles, the IAU strategy is used as well. Argo T–S profiles, XBT, CTD and TAO/TRITON profiles for a 3-year period (2005–2007) are assimilated into the OGCM. Evaluation of this system is performed by comparing model output with independent observations, such as the withheld Argo profiles, NCEP re-analysis products and satellite remote sensing sea level anomaly. ENSO

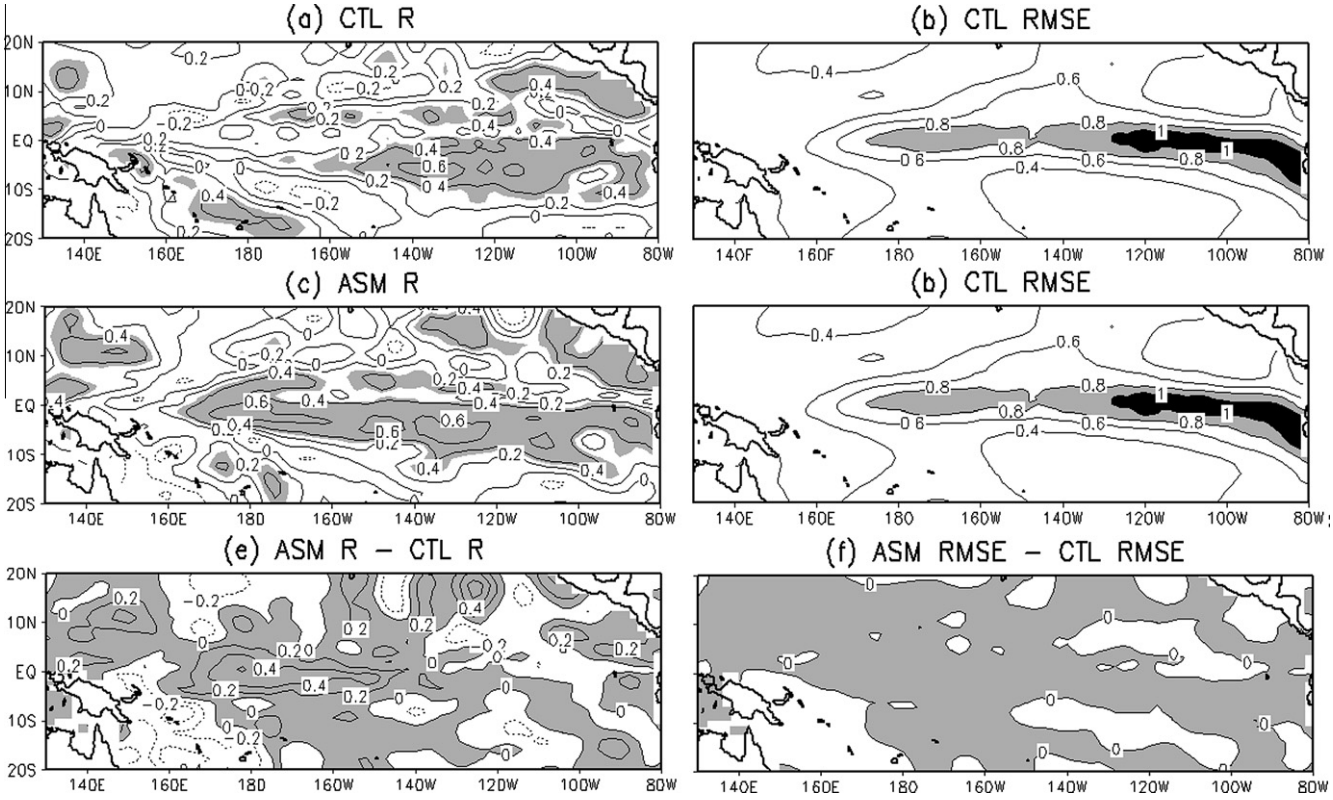

Fig. 11. The correlation (left panel) and RMSE (right panel) skills between predicted SSTA at 6-month lead against the observation for period from 2005 to 2008. The skills of predictions initialized from control run (upper panel) and from the assimilation (middle panel) are shown here. Their differences are shown in the bottom panel. The contour interval is 0.2 in (a), (c) and (e) and 0.2 °C in (b), (d) and (f). Shaded are the values over 0.35 (0.05 significant level) in (a) and (c), over 0.8 °C in (b) and (d), over 0.0 in (e) and under 0.0 °C in (f).

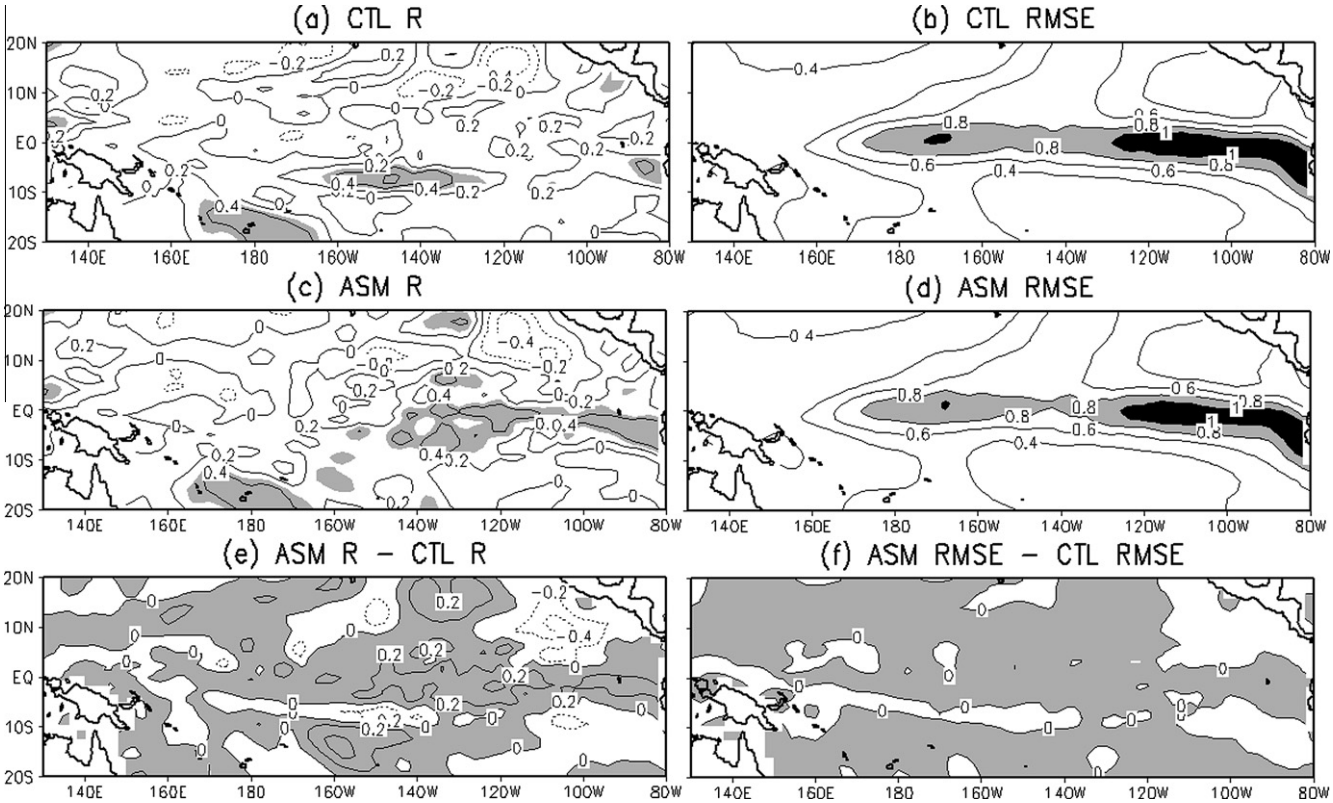

Fig. 12. As in Fig. 11, but for the 12-month lead.

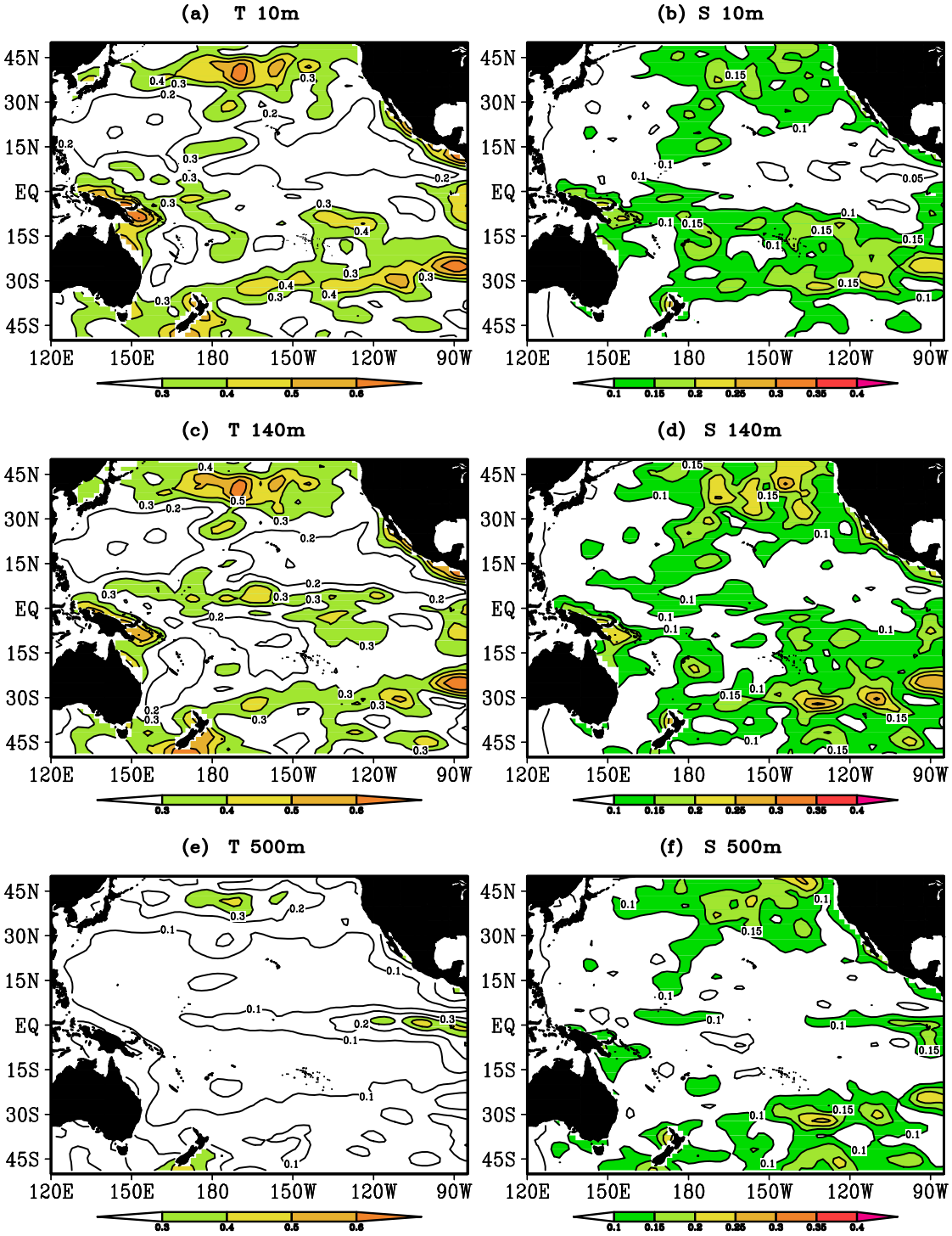

Fig. 13. Ocean temperature (a, c, e) and salinity (b, d, f) analysis ensemble spreads averaged over the period 2005–2007 at the depths of 10 m (a, b), 140 m (c, d) and 500 m (e, f).

hindcast experiments initialized by the assimilation and CTL run are also performed to test the data assimilation system. Finally, the uncertainties in the analyzed fields are estimated using analysis ensembles.

The results show that the data assimilation can significantly improve the ocean thermal state estimation compared against the CTL. Over some regions, the large model biases in the mean temperature fields and salinity fields in the CTL are removed or significantly reduced by the assimilation at all levels. Also the assimilation system significantly reduces the RMSE in both temperature and salinity fields in the whole Pacific at all levels. Comparing the simulated ocean state with the NCEP re-analysis products shows that the assimilation run significantly improves the thermal structure and current structure in the tropical Pacific. The assimilation also improves the sea level anomaly simulation along the equator, and the prediction skill of the tropical Pacific SSTA, especially the SSTA over the western equatorial Pacific.

The results presented in this work suggest that the bias-aware localized EnKF data assimilation system, which makes use of some newly developed strategies, is capable of improving the ocean thermal state estimation. Comparing this data assimilation system with a simple data assimilation system reveals that the incorporation of those advanced methods can considerably improve oceanic analysis. However, in some patchy areas, such as the Kuroshio region, eastern equatorial coast region and the southeast of Australia, the RMSEs of the analyzed fields are still large after assimilating T–S profiles, suggesting that the model resolution is probably too coarse to capture the fine scale variation in these regions. The improvement of SLA by the assimilation system is limited, suggesting the necessity of including the assimilation of satellite SST and SLA in future researches. Currently, the development of a similar assimilation system but with SST and SLA assimilation and for a higher resolution OGCM is in progress.

One significant advantage of Argo data over some in situ observation data sets (e.g., XBT, CTD and TAO/TRITON) is its better spatial and temporal coverage, enabling Argo data to play a stronger role in the ocean state estimation, especially for regions where the other in situ observations are sparse such as the regions beyond the equatorial Pacific. Thus, this study focuses on the assimilation of Argo profiles with other in situ observations used only as supplemental data. Thus, only temperature and salinity observations are used in this assimilation system. The neglect of other observations such as satellite SST and SLA, as well as oceanic currents, etc. may have significant impacts on the oceanic state estimation. In addition, the costly computational expense forced us to simplify some assimilation processes. For example, the bias-correction method used in this study is similar to that by Dee (2005) but a simpler bias prediction model is used. Also, the representation error was not considered variable in the horizontal; and moreover the parameter κ in Eqs. (3) and (4)) is decided somewhat subjectively. All these issues may impact the assimilation performance and need to be addressed in future studies.

Another important issue in EnKF is the ensemble size. In this study, the ensemble size is not very large, probably affecting the assimilation performance. However, the local analysis strategy used here can reduce the impact of the small ensemble size on the efficiency of the EnKF, as suggested by several sensitive experiments (not shown).

Acknowledgements

This work is supported by BC-China Innovation and Commercialization Strategic Development Program (ICSD-2007-Tang-Y). We would like to thank two anonymous reviewers for their constructive comments. G. Wang is supported by the International Corporation Program of China (2008DFA22230).

References

Anderson, J.L., 2002. A local least squares framework for ensemble filtering. Mon. Weather. Rev. 131, 634–642.
Balmaseda, M., Anderson, D., Vidard, A., 2007. Impact of Argo on analyses of the global ocean. Geophys. Res. Lett. 34, L16605. doi:10.1029/2007GL030452.
Bishop, C., Etherton, B., Majundar, S., 2001. Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. Mon. Weather. Rev. 129, 420–436.
Bellucci, A.S., Masina Dipietro, P., Navarra, A., 2007. Using temperature-salinity relations in a global ocean implementation of a multivariate data assimilation scheme. Mon. Weather. Rev. 135, 3785–3807.
Bloom, S.C., Takas, L.L., da Silva, A.M., Ledvina, D., 1996. Data assimilation using incremental analysis updates. Mon. Weather. Rev. 124, 1256–1271.
Böhme, L., Send, U., 2006. Objective analyses of hydrographic data for referencing profiling float salinities in highly variable environments. Deep Sea Res. Part II 53 (1–2), 246.
Carrassi, A., Trevisan, A., Uboldi, F., 2007. Adaptive observations and assimilation in the unstable subspace by breeding on the data-assimilation system. Tellus 59A, 101–113.
Castruccio, F., Verron, J., Gourdeau, L., Brankart, J.M., Brasseur, P., 2008. Joint altimetric and in-situ data assimilation using the GRACE mean dynamic topography: a 1993–1998 hindcast experiment in the Tropical Pacific Ocean. Ocean Dyn. 58, 43–63.
Chepurin, G., Carton, J., Dee, D., 2005. Forecast model bias correction in ocean data assimilation. Mon. Weather. Rev. 133, 1328–1342.
Chu, P.C., Wang, G., Fan, C., 2004. Evaluation of the U.S. navy's modular ocean data assimilation system (MODAS) using South China Sea monsoon experiment (SCSMEX) data. J. Oceanogr. 60, 1007–1021.
Chui, C.K., Chen, G., 1999. Kalman Filtering with Real Time Applications, third ed. Springer-Verlag.
Cummings, J.A., 2005. Operational multivariate ocean data assimilation. Quart. J. Roy. Meteorol. Soc. 131, 3583–3604.
Dee, D., da Silva, A., 1998. Data assimilation in the presence of forecast bias. Quart. J. Roy. Meteorol. Soc. 117, 269–295.
Dee, D., 2005. Bias and data assimilation. Quart. J. Roy. Meteorol. Soc. 131, 3323–3343.
Delecluse, P., Madec, G., 1999. Ocean modelling and the role of the ocean in the climate system. In: Holland, W.R., Joussaume, S., David, F. (Eds.), Modeling the Earth's Climate and its Variability, Les Houches, Session LXVII 1997. Elsevier Science, pp. 237–313.
Deng, Z., Tang, Y., 2009. The retrospective prediction of ENSO from 1881–2000 by a hybrid coupled model – (II): Interdecadal and decadal variations in predictability. Clim. Dyn. 32, 415–428.
Deng, Z., Tang, Y., Zhou, X., 2009. The retrospective prediction of ENSO from 1881–2000 by a hybrid coupled model – (I): SST assimilation with ensemble Kalman filter. Clim. Dyn. 32, 397–413.
Desroziers, G., Berre, L., Chapnik, B., Poli, P., 2005. Diagnosis of observation, background and analysis error statistics in observation space. Quart. J. Roy. Meteorol. Soc. 131, 3385–3396.
Dibarboure, G., Lauret, O., Mertz, F., Rosmorduc, V., Maheu, C., 2009. SSALTO/DUACS user handbook: (M)SLA and (M)ADT near-real time and delayed time products. CLS-DOS-NT-06.034.
Drecourt, J-P., Madson, H., Rosbjerg, D., 2006. Bias aware Kalman filters: comparison and improvements. Adv. Water Res. 29, 707–718.
Evensen, G., 1994. Sequential data assimilation with a nonlinear quasigeostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. 99, 10143–10162.
Evensen, G., 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. Ocean Dyn. 53, 343–367.
Frank, P., Colby, J.R., 1997. A preliminary investigation of temperature errors in operational forecasting models. Weather Forecast. 13, 187–205.
Ferry, N., Rémy, E., Brasseur, P., Maes, C., 2007. The Mercator global ocean operational analysis system: assessment and validation of an 11-year reanalysis. J. Marine Syst. 65, 540–560.
Fukumori, I., 2002. A partitioned Kalman filter and smoother. Mon. Weather. Rev. 130, 1370–1383.
Friedland, B., 1969. Treatment of bias in recursive filtering. IEEE Trans. Automat. Contr. AC-14 (4), 359–367.
Gaspari, G., Cohn, S.E., 1999. Construction of correlation functions in two and three dimensions. Quart. J. Roy. Meteorol. Soc. 125, 723–757.
Hamill, T.H., 2002. Ensemble-based Data Assimilation: A Review. Unpublished manuscript, University of Colorado and NOAA-CIRES Climate Diagnostics Centre.
Houtekamer, P.L., Mitchell, H.L., 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. Mon. Weather. Rev. 129, 123–137.
Huang, B., Xue, Y., Behringer, D., 2008. Impacts of Argo salinity in NCEP Global Ocean Data Assimilation System: The tropical Indian Ocean. J. Geophys. Res., 113.
Hunt, B.R., Kostelich, E.J., Szunyogh, I., 2007. Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. Physica D 230, 112–126.
Janic, T., Cohn, S.E., 2006. Treatment of observation error due to unresolved scales in atmospheric data assimilation. Mon. Weather. Rev. 134, 2900–2915.
Kamachi, M., Kurangano, T., Ichikawa, H., Nakamura, H., Nishina, A., Isobe, A., Ambe, D., Arai, M., Gohda, N., 2004. Operational data assimilation system for the Kuroshio South of Japan: reanalysis and validation. J. Oceanogr. 60, 303–312.
Levitus, S., Boyer, T., 1998. NOAA/OAR/ERSRL PSD. Boulder, Colorado, USA. <http://www.cdc.noaa.gov>.
Li, H., Kalnay, E., Miyoshi, T., 2009. Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. Q.J.R. Meteorol. Soc. 135, 523–533.
Madec, G., 2008. "NEMO ocean engine". Note du Pole de modélisation, Institut Pierre-Simon Laplace (IPSL), France, No. 27. ISSN 1288-1619.
Maes, C., Behringer, D., Reynolds, R.W., Ji, N., 2000. Retrospective analysis of the salinity variability in the western tropical Pacific Ocean using an indirect minimization approach. J. Atmos. Ocean. Technol. 17, 512–524.
Martin, M.J., Hines, A., Bell, M.J., 2007. Data assimilation in the FOAM operational short-range ocean forecasting system: a description of the scheme and its impact. Quart. J. Roy. Meteorol. Soc. 133, 981–995.
McPhaden, M.J., 1995. The tropical atmosphere–ocean array is completed. Bull. Am. Meteorol. Soc. 76, 739–741.
Mitchell, H.L., Houtekamer, P.L., 2000. An adaptive ensemble Kalman filter. Mon. Weather. Rev. 128, 416–433.
Moore, A., Zavala, J., Tang, Y., Kleeman, R., Weaver, A., Vialard, J., Sahami, K., Anderson, D., Fisher, M., 2006. Optimal forcing patterns for coupled models of ENSO. J. Clim. 19, 4683–4699.

Oke, P.R., Schiller, A., Griffin, D.A., Brassington, G.B., 2005. Ensemble data assimilation for an eddy-resolving ocean model of the Australian Region. Quart. J. Roy. Meteorol. Soc. 131, 3301–3311.
Oke, P.P., Brassington, G.B., Griffin, D.A., Schiller, A., 2008. The bluelink ocean data assimilation system (BODAS). Ocean Modell. 21, 46–70.
Ott, E., Hunt, B.H., Szunyogh, I., Zimin, Z.V., Kostelich, E.J., et al., 2004. A local ensemble Kalman filter for atmospheric data assimilation. Tellus 56A, 415–428.
Smith, G.C., Haines, K., 2009. Evaluation of the S(T) assimilation method with Argo dataset. Quart. J. Roy. Meteorol. Soc. 135, 739–756.
Smith, T.M., Reynolds, R.W., 2004. Improved extended reconstruction of SST (1854–1997). J. Clim. 17, 2466–2477.
Snyder, C., Zhang, F., 2003. Assimilation of simulated Doppler Radar observations with an Ensemble Kalman Filter. Mon. Weather. Rev. 131, 1663–1677.
Szunyogh, I., Kostelich, E.J., Gyarmati, G., Kalnay, E., Hunt, B.R., Ott, E., Satterfield, E., Yorke, J.A., 2008. A local ensemble transform Kalman filter data assimilation system for the NCEP global model. Tellus 60A, 113–130.
Tang, Y., Kleeman, R., Moore, A., 2004. SST assimilation experiments in a tropical Pacific Ocean model. J. Phys. Oceangr. 34, 623–642.
Turner, M.R.J., Walker, J.P., Oke, P.R., 2008. Ensemble member generation for sequential data assimilation. Remote Sens. Environ. 112, 1421–1433.
Wong, A.P.S., Johnson, G.C., Owens, W.B., 2003. Delayed-mode calibration of autonomous CTD profiling float salinity data by theta-S climatology. J. Atmos. Ocean. Technol. 20 (2), 308–318.
Zheng, X., 2009. An adaptive estimation of forecast error covariance parameters for Kalman filtering data assimilation. Adv. Atmos. Sci. 26, 154–160.