Multivariate error covariance estimates by Monte-Carlo simulation for assimilation studies in the Pacific Ocean

Anna Borovikov, Michele M. Rienecker, Christian L. Keppenne, and Gregory Johnson

(Paper submitted to Monthly Weather Review"

One of the most difficult aspects of ocean state estimation is the prescription of the model forecast error statistics. The paucity of ocean observations limits our ability to estimate the characteristics of the error from model-observation differences. In most practical applications, simple functional forms of the error distributions for the individual variables are usually prescribed. Rarely are cross-covariances between different model variables used. Here a comparison is made between a univariate Optimal Interpolation (UOI) scheme and a multivariate OI algorithm (MvOI) in the assimilation of ocean temperature. In the UOI case only temperature is updated using a Gaussian covariance function. In the MvOI salinity, zonal and meridional velocities are updated in addition to temperature using empirically estimated multivariate statistical relationships.

Earlier studies have shown that a univariate OI has a detrimental effect on the salinity and velocity fields of the model. Apparently, in a sequential framework it is important to analyze temperature and salinity together. For the MvOI estimation of the model error statistics is made from an ensemble of model integrations. An important advantage of using an ensemble of ocean states is that it provides a natural way to estimate cross-covariances between the fields of different physical variables constituting the model state vector, at the same time incorporating the model's dynamical and thermodynamical constraints as well as the effects of physical boundaries.

Only temperature observations have been assimilated in this study. In order to investigate the efficacy of the multivariate scheme two data assimilation experiments are validated with a large independent set of recently published subsurface observations of salinity, zonal velocity and temperature. For reference, a third control run with no data assimilation is used to check how the data assimilation affects systematic model errors. While the performance of the UOI and MvOI is similar with respect to the temperature field, the salinity and velocity fields are greatly improved when multivariate correction is used, as evident from the comparison with independent observations. The MvOI assimilation is found to improve upon the control run in generating the water masses with properties close to the observed, while the UOI failed to maintain the temperature and salinity structure.

Multivariate error covariance estimates by Monte-Carlo simulation for assimilation studies in the Pacific Ocean.

Anna Borovikov*
SAIC, Beltsville, Maryland

Michele M. Rienecker
Global Modeling and Assimilation Office,
NASA/Goddard Space Flight Center,
Greenbelt, Maryland

Christian L. Keppenne
SAIC, Beltsville, Maryland

Gregory C. Johnson
NOAA/Pacific Marine Environmental Laboratory
Seattle, Washington

December 2, 2003

[*] *Corresponding author address:* Anna Borovikov, Code 900.3 NASA/Goddard, Greenbelt, MD 20771, *ayb@mohawk.gsfc.nasa.gov*

Abstract

One of the most difficult aspects of ocean state estimation is the prescription of the model forecast error covariances. The paucity of ocean observations limits our ability to estimate the covariance structures from model-observation differences. In most practical applications, simple covariances are usually prescribed. Rarely are cross-covariances between different model variables used. Here a comparison is made between a univariate Optimal Interpolation (UOI) scheme and a multivariate OI algorithm (MvOI) in the assimilation of ocean temperature. In the UOI case only temperature is updated using a Gaussian covariance function and in the MvOI salinity, zonal and meridional velocities as well as temperature, are updated using an empirically estimated multivariate covariance matrix.

Earlier studies have shown that a univariate OI has a detrimental effect on the salinity and velocity fields of the model. Apparently, in a sequential framework it is important to analyze temperature and salinity together. For the MvOI an estimation of the model error statistics is made by Monte-Carlo techniques from an ensemble of model integrations. An important advantage of using an ensemble of ocean states is that it provides a natural way to estimate cross-covariances between the fields of different physical variables constituting the model state vector, at the same time incorporating the model's dynamical and thermodynamical constraints as well as the effects of physical boundaries.

Only temperature observations from the Tropical Atmosphere-Ocean array have been assimilated in this study. In order to investigate the efficacy of the multivariate scheme two data assimilation experiments are validated with a large independent set of recently published subsurface observations of salinity, zonal velocity and temperature. For reference, a third control run with no data assimilation is used to check how the data assimilation affects systematic model errors.While the performance of the UOI and MvOI is similar with respect to the temperature field, the salinity and velocity fields are greatly improved when multivariate correction is used, as evident from the analyses of the rms differences of these fields and independent obsevations. The MvOI assimilation is found to improve upon the control run in generating the water masses with properties close to the observed, while the UOI failed to maintain the temperature and salinity structure.

1. Introduction

Data assimilation provides a framework for the combination of the information about the state of the ocean contained in an incomplete data stream with our knowledge of the ocean dynamics included in a model. The problem of data assimilation may be formulated in statistical terms, where because of uncertainty in both observations and models, an estimate of the state of the ocean at any given time is considered to be a realization of a random variable. An estimate of the state of the ocean is produced as a blend of observation and model estimates based on prior knowledge of the error statistics of each, with some measure of the uncertainty in the estimate. The differences between assimilation methods lie primarily in the approaches taken to estimate the error statistics associated with the forward (dynamical) model, the so-called background or forecast error statistics. Since an accurate representation of the data and model error statistics is crucial to a successful data assimilation, a lot of effort has been expended in this direction.

One simplifying assumption that is often made is that these error statistics do not change significantly with time and thus can be approximated by a constant probability distribution. This is the basis of the Optimal Interpolation (OI) data assimilation scheme. An alternative to this assumption is to allow for time evolution of the probability distribution. An example of such a data assimilation scheme is the Kalman Filter (Kalman 1960), in which the error is assumed to be normally distributed and the forecast error covariance matrix is evolved prognostically. The Kalman Filter can be shown to give an optimal estimate in the case of linear dynamics. To account for nonlinear processes a generalization of the Kalman Filter, the Extended Kalman Filter uses instantaneous linearization of the model equations during the data assimilation analysis and the full equations to update the model (e.g., Daley 1991; Ghil & Malanotte-Rizzoli 1991). However the cost of time stepping the model error covariance matrix is computationally expensive, rendering this method impractical when used with high-resolution general circulation models. Under certain conditions it is possible to use an asymptotic Kalman Filter, developed by Fukumori et al. (1993), where a steady-state covariance matrix replaces the time-evolving one. An Ensemble Kalman Filter (EnKF) was introduced by Evensen (1994) based on a Monte Carlo technique, in which the model error statistics are computed from an ensemble of model states evolving simultaneously. An application of this method with the Poseidon ocean model used in this study has been developed by Keppenne and Rienecker (2002, 2003).

This study focuses on the importance of the multivariate aspect of the forecast error covariance in the context of OI data assimilation. Provided a fairly good observing network, the background error structure can be estimated using analysis of spatial and temporal decorrelation scales, as done in numerous meteorological applications (Ghil & Malanotte-Rizzoli, 1994). However, even for atmospheric data assimilation, the observing system is not adequate to support a full calculation of background error covariance statistics and so the model itself has been used to estimate these statistics. The vastness and complexity of the domain and relative scarcity of oceanographic observations require additional simplifying assumptions in similar calculations (e.g., homogeneity of statistics). This paper explores an estimation of the model error from an ensemble of model integrations using Monte-Carlo techniques in a manner similar to the EnKF. An important advantage of using an ensemble of ocean states is that it provides a natural way to estimate cross-covariances between the fields of different physical variables constituting the model state vector.

There are many questions that arise with this approach. For example, how large should the ensemble be, and more generally, how should it be generated. Other questions are related to the underlying assumption of the stationarity and the unbiased nature of error statistics in the OI algorithm. Will a one-time estimate of the model error, derived from a Monte Carlo ensemble, be a good representation of this error at another time, at any time during assimilation? Or, in other words, what is the variability of the model error covariance structure? What are the dominant time scales? Can this information be acquired and, if so, used to improve the assimilation scheme?

The primary interest of this study is ocean phenomena taking place on seasonal-to-interannual time scales. One example of such phenomena is the quasi-regular occurrence of El Niño - a large scale warming of near-surface temperature in the eastern equatorial Pacific Ocean accompanied by a basin wide perturbation in the tilt of the thermocline across the equatorial ocean. The estimate of error statistics derived below attempts to capture errors associated with such variability. The logical organization of the paper is as follows. Next OI assimilation algorithm, model and data are described (Section 2). Then the forecast error covariance model, a traditional Gaussian model of the forecast error covariance and the empirical multivariate model of interest are detailed (Section 3). Then the multivariate error covariance model properties are explored (Section 4). After the experimental setup is decribed, the results of multivariate assimilation are compared with univariate assimilation (Section 5). The paper concludes with discussion of the results and further directions of research (Section 6).

2. OI assimilation

a. OI framework

A detailed discussion of the sequential data assimilation algorithms can be found in earlier literature (see for example, Lorenc (1988)). Here, only a brief outline is given.

A dynamic (prediction) model can be represented in terms of a nonlinear operator $\Psi(\mathbf{x})$, where $\mathbf{x}$ is a state vector of length n_x. Let $\mathbf{d}$ denote a vector of observations which has dimension $n_d \ll n_x$ (typically) and an element of $\mathbf{d}$ is not necessarily an element of the state vector $\mathbf{x}$. The aim of a data assimilation algorithm is to determine the best estimate of the state vector based on the estimates available from both model and observations. Formally, an optimal estimate of the state would minimize a "cost" functional, which can be defined, for example, to represent the total variance of the system - a measure of the misfit between the estimate and observations and other desired constraints, each with their own "cost" or "risk". For example, written as

$$\mathcal{J}(\mathbf{x}) = (\mathbf{x}-\mathbf{x}^f)^T\mathbf{P}(\mathbf{x}-\mathbf{x}^f) + (\mathbf{d}-\mathcal{H}(\mathbf{x}))^T\mathbf{W}(\mathbf{d}-\mathcal{H}(\mathbf{x})), \tag{1}$$

the cost functional $\mathcal{J}(\mathbf{x})$ contains a model error term and a data misfit term. Here $\mathbf{x}^f$ denotes the model simulated state, and $\mathcal{H}$ denotes the observation transformation operator, which relates the observed quantities and the model variables. Other terms, such as boundary condition error, may be explicitly included in $\mathcal{J}(\mathbf{x})$. $\mathbf{P}$ and $\mathbf{W}$ are weights representing our confidence in the model and the data respectively. Specification of these weight matrices requires some prior knowledge of the model and data error statistics.

A discrete form of the model can be written as $\mathbf{x}_k = \Psi_{k-1}(\mathbf{x}_{k-1})$, where $\mathbf{x}_k$ is the forecast state vector at time level k and Ψ_{k-1} is the numerical approximation to the set of model equations describing the evolution of the state forward from time $k-1$ to k. Similarly, observations available at time k can be denoted as $\mathbf{d}_k$ and the observation transformation operator as $\mathcal{H}_k(\mathbf{x}_k)$.

A sequential unbiased assimilation scheme for the time-varying $\mathbf{x}_k$ is given by:

$$\mathbf{x}_k^f = \Psi_{k-1}(\mathbf{x}_{k-1}^a) \tag{2}$$

$$\mathbf{x}_k^a = \mathbf{x}_k^f + \mathbf{K}_k\left(\mathbf{d}_k - \mathcal{H}_k(\mathbf{x}_k^f)\right) \tag{3}$$

Here superscript f stands for the forecast and a for the analysis. All sequential data assimilation schemes have the form of equation (3) and differ from each other by the weight matrix $\mathbf{K}_k$ often called the *gain matrix.*

The optimality of $\mathbf{K}_k$ can be defined under certain assumptions about the error statistics. Most sequential data assimilation algorithms are based on assumptions that the observational and model errors are unbiased, white in time, spatially uncorrelated with each other and that their spatial covariances are known (usually it is assumed that at least initially the errors are Gaussian).

Suppose the true evolution of the system is governed by

$$\mathbf{x}_k^t = \Psi_{k-1}(\mathbf{x}_{k-1}^t) + \epsilon_{k-1}^t, \tag{4}$$

where ϵ_k^t, called *system noise* or model error, is a (Gaussian) white-noise sequence:

$$E\epsilon_k^t = 0, \ E\epsilon_k^t \left(\epsilon_l^t\right)^T = \mathbf{S}_k \delta_{kl}.$$

The observations may be described by

$$\mathbf{d}_k = \mathcal{H}_k(\mathbf{x}_k^t) + \epsilon_k^o, \tag{5}$$

where ϵ_k^o, the *observational noise* or measurement error, is also a (Gaussian) white-noise sequence,

$$E\epsilon_k^o = 0, \ E\epsilon_k^o \left(\epsilon_l^o\right)^T = \mathbf{W}_k \delta_{kl}.$$

This ϵ_k^o may also include any error of representations of the processes of interest, although such errors will not in general satisfy the assumption of a white, Gaussian sequence. Without any loss of generality, it is also assumed that the system noise and the observational noise are uncorrelated with each other,

$$E\epsilon_k^t \left(\epsilon_k^o\right)^T = 0. \tag{6}$$

Under these assumptions, for a linear model and a linear observation transformation operator $\mathcal{H}_k \equiv \mathbf{H}_k$ in the equations (4) and (5), the cost functional (1) is exactly minimized in a least-squares sense when

$$\mathbf{K}_k = \mathbf{P}_k^f \mathbf{H}_k^T (\mathbf{H}_k \mathbf{P}_k^f \mathbf{H}_k^T + \mathbf{W}_k)^{-1}. \tag{7}$$

Here $\mathbf{P}_k^f$ is the forecast error covariance matrix, which, in general, is time-dependent and the accuracy of its estimation relies on our knowledge of $\mathbf{S}_k$ and $\mathbf{W}_k$. For a high resolution ocean model with the number of state variables on the order of 10^6, $\mathbf{P}_k^f$ is extremely expensive to store and evaluate in full. Thus, numerous approaches have been suggested to simplify the computation of $\mathbf{P}_k^f$. The traditional OI method assumes that $\mathbf{P}_k^f \equiv \mathbf{P}$ is approximately constant. In the case of observational errors, the matrix $\mathbf{W}$ is often assumed to be diagonal and to contain only information about the level of variance in the measurement error due to instrumental imperfection and unresolved small-scale signal. There are means of allowing for simple time evolution of the forecast error variance (see, for example, Ghil 1991, Rienecker and Miller 1991), but they are not considered here. A full evolution of $\mathbf{P}_k^f$ would be a Kalman filter.

The effects of non-linear dynamics and inhomogeneities associated with ocean boundaries are implicitly taken into account when the empirical forecast error covariance matrix $\mathbf{P}$ is constructed from model integrations as presented in the next section.

b. Model and forcing

The model used for this study is the Poseidon reduced-gravity, quasi-isopycnal ocean model introduced by Schopf and Loughe (1995) and used by Keppenne and Rienecker (2002, 2003) for test of the Ensemble Kalman Filter. The model described by Schopf and Loughe (1995) has been updated to include the effects of salinity (e.g., Yang et al,. 1999). The model was shown to provide realistic simulations of tropical Pacific climatology and variability (Borovikov et al., 2001).

The model equations are:

$$\begin{aligned}
\frac{\partial h}{\partial t} + \nabla\cdot(\mathbf{v}h) + \frac{\partial w_e}{\partial \zeta} &= 0,\\
\frac{\partial hT}{\partial t} + \nabla\cdot(\mathbf{v}hT) + \frac{\partial w_e T}{\partial \zeta} &= \frac{\partial}{\partial \zeta}\left(\frac{\kappa}{h}\frac{\partial T}{\partial \zeta}\right) + \frac{\partial Q}{\partial \zeta} + h\mathcal{F}_H(T),\\
\frac{\partial hS}{\partial t} + \nabla\cdot(\mathbf{v}hS) + \frac{\partial w_e S}{\partial \zeta} &= \frac{\partial}{\partial \zeta}\left(\frac{\kappa}{h}\frac{\partial S}{\partial \zeta}\right) + h\mathcal{F}_H(S),\\
\frac{\partial P}{\partial \zeta} &= -g\rho h,\\
P'(0) &= g\rho_0\eta,\\
\frac{\partial P'}{\partial \zeta} &= \rho_0 bh,\\
\eta &= \frac{1}{g}\int bh\,d\zeta,\\
\frac{\partial(\mathbf{v}h)}{\partial t} + \nabla\cdot(\mathbf{v}h\mathbf{v}) + \frac{\partial w_e\mathbf{v}}{\partial \zeta} &= -\frac{h}{\rho_0}\nabla P' - bh\nabla z\\
-fh\mathbf{k}\times\mathbf{v} + \frac{\partial}{\partial \zeta}\left(\frac{\nu}{h}\frac{\partial \mathbf{v}}{\partial \zeta}\right) &+ \frac{1}{\rho_0}\frac{\partial \tau}{\partial \zeta} + h\mathcal{F}_v'(\mathbf{v}).
\end{aligned}$$

Here ζ is the generalized vertical coordinate, h is layer thickness, $\mathbf{v}$ is the 2D horizontal velocity vector, w_e is mass flux across ζ surfaces, T is potential temperature, S is salinity, Q is external heat flux, P is pressure, ρ is density, η is dynamic height, b is buoyancy, τ is wind stress, κ and ν are vertical diffusivities and friction, and $\mathcal{F}$ is a horizontal smoothing operator. The generalized vertical coordinate of the model includes a turbulent well-mixed surface layer with entrainment parameterized according to a Kraus-Turner (1967) bulk mixed layer model.

For this study, the domain is restricted to the Pacific Ocean (45°S to 65°N) with realistic land boundaries. At the southern boundary the model temperature and salinity are relaxed to the Levitus (1994) climatology. The horizontal resolution of the model is 1° in longitude; and in the meridional direction a stretched grid is used, varying from 1/3° at the equator to 1° poleward of 10°S and 10°N. The calculation of the effects of vertical diffusion, implemented at three hour intervals through an implicit scheme, are parameterized using a Richardson number-dependent vertical mixing following Pacanowski-Philander (1981). The diffusion coefficients are enhanced when needed to simulate convective overturning in cases of gravitationally unstable density profiles. Horizontal diffusion is also computed daily using an 8th-order Shapiro (1970) filter. The net surface heat flux is estimated using the atmospheric mixed layer model by Seager et al. (1994) with the monthly averaged time-varying air temperature and specific humidity from the NCEP-NCAR reanalysis (e.g., Kalnay et al. 1996) and climatological shortwave radiation from the Earth Radiation Budget Experiment (ERBE) (e.g., Harrison et al. 1993) and climatological cloudiness from the International Satellite Cloud Climatology Project (ISCCP) (e.g., Rossow and Schiffer 1991).

Surface wind stress forcing is obtained from the Special Sensor Microwave Imager (SSM/I) surfaces wind analysis produced by Atlas et al. (1991) based on the combination of the Defense Meteorological Satellite Program (DMSP) SSM/I data with other conventional data and with the ECMWF 10m surface wind analysis. The surface stress was produced from this analysis using the drag coefficient of Large and Pond (1982). Daily averaged wind stress forcing for the time period of 1996-1997 was applied to the model. The precipitation is given by monthly averaged analyses of Xie and Arkin (1997).

For example, model mean (1988-1997) temperature, salinity and zonal velocity sections along the equator compare very well with estimates made from observations (Johnson et al., 2002) taken during an overlapping period (figure 1).

c. *Data*

The TAO/Triton Array, consisting of more than 70 moored buoys spanning the equatorial Pacific (http://www.pmel.noaa.gov/toga-tao/home.html and McPhaden et al., 1998) as shown in figure 2, measures oceanographic and surface meteorological variables: air temperature, relative humidity, surface winds, sea surface temperatures and subsurface temperatures down to a depth of 500 meters. By 1994 these measurements became available daily across the equatorial Pacific Ocean approximately uniformly spaced at 15 degrees.

The temperature observations from the TAO/Triton array were the only data type used during these assimilation experiments since the focus is on well-known deleterious effects of temperature assimilation in the equatorial waveguide. (However, in the global assimilation conducted by the NASA Seasonal-to-Interannual Prediction Project to initialize seasonal forecasts, the global XBT data base is included.) The standard deviation of the observational error, σ_{TAO}, is set to 0.5°C and the errors are assumed to be uncorrelated in space and time. This value is high compared to the instrumental error of 0.1°C (Freitag et al , 1994) since it also has to reflect the representativeness error, i.e. the data contains a mixture of signals of various scales including frequencies much higher than the target scales of assimilation. By tuning σ_{TAO} we effectively control the ratio of the data error variance to the model error variance.

3. Forecast error covariance modeling

In error covariance structure modeling, one is striving for an accurate representation of the error statistics and usually for simple and efficient implementation for computational viability. With little knowledge of the true nature of the model error covariances, one often has to make assumptions and settle for simple methods which usually have the advantage of being easy to implement. This section describes two different models for the forecast error covariance structure, a simpler and less computationally intense and a more elaborate and hopefully more accurate model. For both, an OI framework is used wherein the forecast error covariance matrix, $\mathbf{P}^f$, is assumed to be time-invariant.

a. Univariate functional model

A commonly used method of analytical error covariance function (see, for example, Carton, 1990 and Ji, 1995) has been employed here in the tropical Pacific Ocean region. In this study, the spatial structure of the model temperature (T) forecast error is assumed to be Gaussian in all three dimensions with scales 15°, 4° and 50 m in zonal, meridional and vertical directions, respectively. These scales were estimated from the ensemble of model integrations described below in the next subsection. Those spatial scales are also resolved by the observing system of equatorial moorings which are generally separated by 10° to 15° in the zonal direction and by 2° to 3° in the meridional direction. Horizontal scales are comparable to scales used in the similar assimilation schemes, for example, by Ji et al. (1995) and Rosati et al. (1996). There are several advantages of this error covariance model. For the Gaussian form of the covariance function, the minimum variance estimate for the least squares minimizing functional is the maximum likelihood estimate, and the analysis error covariance function is also Gaussian. It is easy to implement and adapt to the parallel computing architecture. In this implementation the temperature observations have been processed and the correction was only made to the model temperature field during each assimilation cycle, while other variables adjusted according to the model's dynamic response to the temperature correction.

b. Monte Carlo method for estimating the multivariate forecast error covariance

A more realistic covariance structure that is consistent with model dynamics and the presence of ocean boundaries was sought through an application of the Monte Carlo method. The variability across an ensemble of ocean state estimates was used for a one-time estimate of the model forecast error statistics. In spirit, this approach is similar to the Ensemble Kalman Filter except that the error covariance does not evolve with time and does not feel the impact of prior data assimilation, although it could.

The design of this forecast error covariance model was influenced by the need to assimilate TAO mooring observations for seasonal forecasts. While the Poseidon model has layered configuration, the TAO observations are taken at approximately constant depth levels. In the implementation for this study, the covariances are calculated on pre-defined depth levels. At each assimilation cycle the model fields are interpolated to these depths, the assimilation increment is computed on these pre-specified levels, and are then interpolated back to the temperature grid points at the center of the model layers. The discussion below deals with the three-dimensional model error covariance matrix whose horizontal structure coincides with the model grid, and in the vertical is arranged at constant depths coincident with the nominal TAO instrument depths.

Consider the non-dimensionalized model state vector

$$\mathbf{x} = \begin{bmatrix} T/\sigma_T \\ S/\sigma_S \\ U/\sigma_U \\ V/\sigma_V \\ ssh/\sigma_{ssh} \end{bmatrix}, \tag{8}$$

here T, S, U, V and ssh are model variables: temperature, salinity, zonal and meridional velocities and dynamic height respectively, and $\sigma_{[T,S,U,V,ssh]}$ are non-dimensionalizing factors. For the latter we took the global variance within each of the model fields at a depth of 100 m: σ_T=0.65, σ_S=0.08, σ_U=0.09, σ_V=0.08 and σ_{ssh}=0.08 in the corresponding units. The multivariate covariance matrix is

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}^{T,T} & \mathbf{P}^{T,S} & \mathbf{P}^{T,U} & \mathbf{P}^{T,V} & \mathbf{P}^{T,ssh} \\ \mathbf{P}^{T,S} & \mathbf{P}^{S,S} & \mathbf{P}^{S,U} & \mathbf{P}^{S,V} & \mathbf{P}^{S,ssh} \\ \mathbf{P}^{U,T} & \mathbf{P}^{U,S} & \mathbf{P}^{U,U} & \mathbf{P}^{U,V} & \mathbf{P}^{U,ssh} \\ \mathbf{P}^{V,T} & \mathbf{P}^{V,S} & \mathbf{P}^{V,U} & \mathbf{P}^{V,V} & \mathbf{P}^{V,ssh} \\ \mathbf{P}^{ssh,T} & \mathbf{P}^{ssh,S} & \mathbf{P}^{ssh,U} & \mathbf{P}^{ssh,V} & \mathbf{P}^{ssh,ssh} \end{bmatrix}. \qquad (9)$$

If the matrix $\mathbf{A}^{m \times n_x}$ contains the m-member ensemble of (anomalous) ocean states as columns, then $\mathbf{P}$ can be computed as

$$\mathbf{P}^{n_x \times n_x} = \frac{\mathbf{A}\mathbf{A}^T}{m-1}, \text{ with } rank(\mathbf{P}) \leq \min\{m, n_x\}. \qquad (10)$$

The size of $\mathbf{P}$ is on the order of $n_x \approx 10^6$ (the dimension of the state vector), while its rank is no greater than the size of the ensemble, m (on the order of 10^2 in the case of this study). The estimate of the error covariance matrix was stored on file and read in during every assimilation cycle of the OI algorithm. Since the rank of the error covariance matrix $\mathbf{P}$ estimated using this method is no greater than the Monte Carlo ensemble size, it can be conveniently represented using a basis of empirical-orthogonal functions (eofs), $\mathbf{E}$. To compute the eof representation of $\mathbf{P}$, observe that $\mathbf{A}\mathbf{A}^T$ has the same eigenvalues as $\mathbf{A}^T\mathbf{A}$, which is only $m \times m$ and the eigenvectors of $\mathbf{A}\mathbf{A}^T$ are related to those of $\mathbf{A}^T\mathbf{A}$ as

$$\mathbf{E} = \mathbf{A}\mathbf{U}(\Lambda)^{-1/2}, \qquad (11)$$

where $\mathbf{E}^{n_x \times m}$ contains the eigenvectors of $\mathbf{A}\mathbf{A}^T$, $\mathbf{U}^{m \times m}$ contains the eigenvectors of $\mathbf{A}^T\mathbf{A}$ and $\Lambda^{m \times m} = diag(\lambda_1^2, ..., \lambda_m^2)$ has the eigenvalues of $\mathbf{A}^T\mathbf{A}$. Then

$$\mathbf{P} = \frac{\mathbf{A}\mathbf{A}^T}{m-1} = \frac{\mathbf{E}\Lambda\mathbf{E}^T}{m-1} = \mathbf{Q}\mathbf{Q}^T. \qquad (12)$$

The columns of $\mathbf{E}$ are orthonormal and the eigenvalues λ_i^2, $i = 1, .., m$, are the variances. Equation (7) can thus be rewritten as

$$\mathbf{K} = \mathbf{Q}\mathbf{Q}^T\mathbf{H}^T(\mathbf{H}\mathbf{Q}\mathbf{Q}^T\mathbf{H}^T + \mathbf{W})^{-1}, \text{ with } \mathbf{Q} = \mathbf{E}\Lambda^{1/2}(m-1)^{-1/2}. \qquad (13)$$

1) Ensemble generation

As the first test of this methodology, the ensemble of states was generated by forcing the ocean model with an ensemble of air-sea fluxes:

$$\mathbf{F}_n = \mathbf{F} + \delta\mathbf{F}_n. \tag{14}$$

$\mathbf{F}$ is the climatology of forcing used for control run, $\delta\mathbf{F}_n$ are interannual anomalies - in phase with interannual SST anomalies but with different internal atmospheric chaotic variations. Surface forcing is used for the ensemble generation because this is probably the dominant source of error in the upper ocean in the equatorial Pacific. Although errors in the synoptic forcing will be large, the focus here is on the longer time scales of interest for seasonal prediction. The fluxes were obtained from a series of integrations of the Aries atmospheric model (e.g., Suarez and Takacs, 1995) forced by the same interannually varying sea surface temperatures (SST) and differing only in slight perturbations to the initial atmospheric state. The interannual anomalies in surface stress and heat flux components were added to seasonal forcing estimated from the sources described in the section 2(b). This approach attributes all of the ocean model forecast error to uncertainties in the longer time scale surface flux anomalies, since differences between the ensemble members were due to atmospheric internal variability.

In all, 32 runs were conducted, each 15 years long, corresponding to the 1979-1993 period of the SST data used to force the atmospheric model. Five day averages (pentads) of the model fields were archived. These were subsequently interpolated to the 11 depth levels, coincident with the depths of the TAO observations. All the covariance estimates have been made using these fields. The matrix of ensemble members, $\mathbf{A}$ was formed by selecting at random five years from the 15 year period, then choosing a pentad from each year corresponding to the same date, say, the first of January. The same pentad was used for each ensemble member. This allowed for collection of an ensemble of 160 members. The mean was removed separately for each of the 5 years to remove the influence of interannual variability. The eofs of the matrix **P** were then computed. The properties of the error covariance matrix constructed in such a way are discussed below.

2) Compact support

A persistent problem associated with the empirical model error covariance estimation is the appearance of unphysical large lag correlations that are an artifact of the limited ensemble size - we use an ensemble size of 160, yet the potential numbers of degrees of freedom are $O(10^6)$. To alleviate this problem, the multivariate anisotropic inhomogeneous matrix was modified by a matrix specified by a covariance function that vanishes at large distances, i.e. a Hadamard product of the two matrices was employed, as discussed by Houtekamer and Mitchell (2001). Keppenne and Rienecker (2002) implemented the compact support for the Ensemble Kalman Filter developed by the NASA Seasonal-to-Interannual Prediction Project (NSIPP) for parallel computing architectures, and that implementation is used in the present study. The functional form follows the work by Gaspari and Cohn (1999) who provided a methodology for constructing compactly supported multi-dimensional covariance functions. The characteristic scales of this function were selected in such a way that most of the local features of the empirically estimated error covariance structure are preserved but at large spatial lags the covariance vanishes: 30°, 8° and 100 m in the zonal, meridional and vertical directions respectively.

To visualize the covariance structure, an artificial example is considered with a single observation different from a background field by one non-dimensional unit. The resulting correction reflects the model error correlation structure - it corresponds to a section of a single row of the **P** matrix for the case of a perfect observation. This is also termed the marginal gain since is measures the impacts of processing a single measurement without reference to other data that might be assimilated. The correlation between the temperature observation at several locations across the equatorial Pacific ocean (156°E, 180°W, 155°W and 125°W) at depths roughly corresponding to the position of thermocline, approximately the 20°C isotherm depth in figure 1) and the temperature error elsewhere in the Pacific (submatrix $\mathbf{P}^{T,T}$ in equation (9)) as derived from the ensemble of states with compact support (figure 3) reveals that the long range correlation is eliminated, but the local structure is intact.

3) Multivariate error covariance patterns

The following discussion of the multivariate error covariance model will focus on the thermocline region in the equatorial Pacific Ocean. The shapes of the correlation structure associated with a single point differ between the eastern and western regions (figure 3, top 4 panels). The zonal scale tends to be shorter in the western and central and longer in the eastern part of the basin. Meridional decay scales are similar along the equator, but the vertical correlation (figure 3, middle 4 panels) varies: shorter and symmetrical in the western part, slightly skewed in the central part and symmetrical but more elongated in the eastern part of the equatorial Pacific basin. Zonal sections (figure 3, bottom 4 panels) illustrate the anisotropy associated with the tilt of the thermocline. This example alone demonstrates that the univariate temperature error covariance structure is so complex that a homogeneous error correlation structure is not applicable.

Although to date there have been very few salinity observations, this is changing with the Argo program (http://argo.jcommops.org and Wilson, 2000). Hence, it is of interest to explore corrections associated with salinity observations (figure 4). The decorrelation scales in the western basin are noticeably longer than in the middle and eastern basin, 8 to 10 degrees in zonal and 4 to 6 degrees in meridional direction in the west and 2-4 degrees in zonal and 1-2 degrees in meridional direction in the east. The scales are notably shorter that those for temperature (figure 3) except for the meridional scales in the west.

In a similar fashion one can analyze the temperature-salinity, temperature-velocity and other cross-variable relationships, i.e. the effect of a single unit observation on various fields - components of the ocean state vector. Corrections in S and U fields associated with a T observation and corrections in T and U associated with S observation are displayed for a single location, 155°W at equator (figure 5).

The temperature-salinity covariance, i.e. the effect of a single unit temperature observation on the salinity field (this would correspond to the subsection of the submatrix $\mathbf{P}^{T,S}$) is described next (based on the figure 5 and plots of cross-covariances at various locations not shown here). The structure of the temperature-salinity relationship is complex and irregular. The change in salinity associated with a temperature increment is not necessarily density-compensating. Equatorial temperature and salinity south of the equator in the western region are anticorrelated, while temperature at the equator and salinity immediately to the north are correlated at 150 meters in the western and central Pacific and the scales of influence are short compared with the temperature-temperature relationship. The anticorrelation is consistent with the mean thermohaline (T-S) structure, with fresh water overlying a saline core. In the east, the correlation between T and S is primarily vertical; horizontal scales are very short, on the order of 2-4 degrees. The positive correlations on the equator, as seen on the meridional sections of the central basin, are higher towards the northern hemisphere, and the negative correlations to the south are consistent with higher temperatures straddling the cold tongue with more saline water south of the equator and fresher water north. Thus the covariances are consistent with vertical and meridional variations.

The relationship between temperature and velocity in the western Pacific reflects temperature changes associated with upstream advection/convergence effects. At 156°E and at the dateline, the higher temperatures are associated with a weaker equatorial undercurrent in a broad region to the west. At 155°W, the effects are more local and wavelike with increased temperature associated with a stronger equatorial undercurrent. At 125°W the scales are shorter and also wavelike, with changes in temperature apparently associated with instability waves.

It is possible to infer from the multivariate analysis the effect a single salinity observation would have on temperature and zonal velocity fields at various locations across the equatorial Pacific ocean. The high level of positive correlation between salinity and temperature field in the central and to a lesser degree in the eastern Pacific indicates that the correction of the salinity field may have a significant impact on the temperature. The S-U relationship is weak in the western part of the basin and the correlation patterns are wavelike in the east, strongly pronounced in the north-south direction.

4. Robustness of the model error covariance estimate

In this section, the sensitivity of the covariance structure to the choice made in populating the ensemble, i.e. to seasonal or interannual variations in the atmospheric forcing is explored to evaluate the robustness of the covariance estimates. The robustness is tested by randomly sampling the full suite of integrations. Five years out of 15 (the length of the run) were picked at random, then the same date (e.g., January 1-5 pentad) was taken for each year. As before, the mean across the ensemble was removed for each year. The procedure was repeated ten times allowing us to obtain ten realizations of the covariance matrix **P**. The pentads were chosen so that realizations from the same season and from different seasons could be compared. From visual assessment of figures similar to figures 3-5, the correlation structures represented by the different estimates of **P** were very similar.

One comparison of the robustness of covariance estimates is pointwise covariance sections (figure6) at the same locations as simulated temperature observations as in figures 3-4. The tight distribution of the decorrelation curves from the 10 different **P** realizations (thin lines) indicates good reproducibility of the covariance structure. No significant interannual variability is apparent within this collection of **P** matrices. The over-plotted Gaussian curves show that the decorrelation scales vary at the four locations across the equatorial basin and can hardly be fitted by a single parameter (scale estimate) in a functional covariance model. In the UOI covariance model used for comparison below, the temperature decorrelation scales chosen are consistent with the scales of the empirical error covariance model in the western and central equatorial Pacific.

The difference among the Monte Carlo estimates of $\mathbf{P}$ can also be quantified in terms of the dominant error subspaces spanned by each of the ensemble sets. These subspaces are best described by the orthonormal bases of empirical orthogonal functions (eofs). The use of eofs allows a spatial filtering of the covariance structures by inclusion of only those eofs that are non-noise-like, thus defining the dominant error subspace. This procedure also eliminates problems associated with different levels of variance even though the spatial structures (covariances) are similar.

Consider the projection of an ensemble of ocean state anomalies onto a given set of eofs. An anomalous ocean state vector $\mathbf{a}$ can be expressed in terms of the eof basis $\{\alpha\}$ as

$$\mathbf{a} = \Sigma_i a_i \alpha_i + \delta^\alpha. \tag{15}$$

The set of eofs $\{\alpha\}$ spans the subspace $\mathcal{S}_\alpha$ of the model error space $\mathcal{S}$ and δ^α is the residual lying in the complement of $\mathcal{S}_\alpha$, i.e., subspace $\mathcal{S}_\alpha^c$, not spanned by $\{\alpha\}$. $\mathcal{S}_\alpha^c$ may or may not contain significant model error covariability information. To assess the information content not included in $\mathcal{S}_\alpha$ we examine covariability through the eofs of δ^α. If the eofs of δ^α are noise-like, this would indicate that the eofs $\{\alpha\}$ captured the significant information regarding the model error contained in $\mathbf{a}$. This calculation was repeated for several instances of $\{\alpha\}$ and $\mathcal{S} = \{\mathbf{a}\}$ to assess the invariability of $\mathcal{S}_\alpha$.

The spectra of various ensembles of $\delta^\alpha \subset \mathcal{S}_\alpha^c = \mathcal{S} \backslash \mathcal{S}_\alpha$ are shown in figure 7, where $\{\alpha\}$ are calculated from January pentads and $\{\mathbf{a}\}$ are pentads from July. In every case, the eigenvalues of $\{\alpha\}$ and $\{\delta\}$ are normalized by the variance of the corresponding ensemble $\{\mathbf{a}\}$. The eigencurves of $\{\delta\}$ are almost flat, characteristic of white noise, and are on order of magnitude less than the dominant eigenvalues of α. Thus the error subspace generated from this Monte Carlo simulation appears to be robust.

5. Assimilation experiments

The effectiveness of the empirical multivariate forecast error covariance estimate is assessed by assimilating the temperature observations from the TAO moorings. The evaluation uses a set of independent (i.e. not assimilated) temperature, salinity and zonal velocity observations from the TAO servicing cruises. The temperature and salinity data are based on Conductivity-Temperature-Depth (CTD) profiles and the velocity data from the Acoustic Doppler Profiler (ADCP). The comparison uses a gridded analysis of these data, as described by Johnson et al (2000).

The assimilation experimental setup is as follows. The model was spun-up for 10 years with climatological forcing and then integrated with time dependent forcing for 1988-1998 in all the experiments. The assimilation began in July 1996. The initial conditions and the forcing were identical in all assimilation experiments. In addition to the data assimilation runs, a forced model integration without assimilation (referred to as the control) serves as a baseline for assessing the assimilation performance. The assimilation run with a simple univariate covariance model is denoted UOI. The run with the empirical multivariate forecast error covariance model is termed MvOI.

In every assimilation experiment, the daily-averaged subsurface temperature data from the TAO moorings was assimilated once a day. To alleviate the effects of the large shock on the model resulting from the intermittent assimilation of imperfectly balanced increments, the incremental update technique was used (Bloom et al, 1996). In this implementation, the assimilation increment is added gradually to the forecast fields at each time step

The figures 8 and 9 show the cross-validation of the simulation (i.e., the control, with no assimilation) and two assimilation tests against the independent (i.e., not assimilated) temperature, salinity and zonal velocity vertical profiles from Johnson et al. (2002). All of the available observed profiles are used and the statistics are separated corresponding to four regions: Niño 4 (160°E-150°W) and Niño 3 (150°W-90°W), further divided into two halves, south and north of the equator (0°-5°N and 5°S-0°). To put the amplitude of the

RMSD in prospective, the mean monthly standard deviation (std) of the model is plotted as well. It is calculated using daily values at the same predefined depth levels on which the analyses are performed. The std represents the level of the internal variability in the model for the submonthly temporal scales which could in part be responsible for the errors in the monthly averaged profiles assessed against single asymptotic ship observations. In general, the RMSD of the control quantities and the data is about twice as large as the model std. The MvOI experiment shows comparable skill in temperature as the UOI with the greatest reduction in RMSD in the thermocline in the Niño 3 region south of the equator. Below 400 meters neither of the assimilation schemes shows smaller RMSD than the control run due to the fact that data for assimilation is only available above 500 meters and at this level the observations are sparse. The MvOI is able, however, to preserve the salinity structure south of the equator and in the Niño 3 region north of the equator. To a lesser extent the MvOI current structure is also improved compared with the UOI, especially south of the equator.

The UOI assimilation improves upon the control case in the representation of temperature, yet the investigation of other model fields, such as salinity, reveals potential problems in a long-term integration. Figure 10 shows time series of the equatorial salinity, averaged between 2°S and 2°N at the thermocline depthmsince the beginning of the assimilation. Where available, the observed salinity is shown by stars. In the UOI experiment it took on the order of 3 to 4 months for the salinity structure to deteriorate significantly. Poor performance of UOI is due to the fact that correcting the temperature field alone introduces artificial and potentially unstable water mass anomalies whose propagation and eventual enhancement destroys model dynamic balances. A method to alleviate this problem, proposed by Troccoli and Haines (1999) relies on the model-derived water mass properties to correct the model salinity commensurate with the temperature corrections made by assimilating temperature observations. The salinity increments are calculated according to the temperature analysis by preserving the model's local T-S relationship. While the proposed method shows improvement in temperature and salinity analyses when tested with Poseidon ocean model (Troccoli et al., 2003), it has some limitations, i.e. the scheme is designed solely for temperature observations and relies on the model maintaining a consistently good T-S relationship.

To test how well the assimilation schemes preserve the water mass properties, we consider, in a manner similar to Troccoli et al. (2003), the T-S relationships in the same subregions as used above. T-S pairs at each observation are compared with model values interpolated to the same locations using a T-S grid of ganularity 0.25°C by 0.1 (figures 11 and 12). At least 5 T-S pairs must be found for a colored circle to be plotted to make sure that the features in the figures are robust. South and north of the equator in both Niño 3 and Niño 4 regions the model without assimilation (top panels) shows good representation of T-S except in the area of warmest water (cyan circles near the top of the plot) and somewhat in the representation of the dense cool saline water (few cyan circles below the main body of red color). The first deficiency is successfully corrected by the MvOI and to a lesser degree by the UOI. Some observed surface warm saline waters in the Niño 3 region north of the equator are not included in any of the model analyses, probably due to errors in surface forcing that the assimilation is not able to rectify. The problem of the lack of dense saline water in the model is slightly overcorrected by MvOI: all cyan circles change to red and some black circles appear in both regions north and south of the equator. The UOI scheme shows gross over-production of this type of water south of the equator and to a lesser degree in the north and it misses the more saline side of the distribution from σ_θ of 22 to 26, north of the equator as well as in the south. Thus, significant problems are apparent in the UOI scheme, while MvOI is able to improve upon the control over almost the entire range of T-S diagram.

Figures 13, 14 and 15 show examples of a meridional cross-section of the temperature, salinity and zonal velocity fields compared to a selection of sections prepared and presented in Johnson et al (2002). The sections are chosen so that approximately simultaneous sections across the Pacific basin can be shown after a long period of integration (about 2 years).

These sections are included in the RMSD statistics of figures 8 and 9. The temperature in the UOI experiment is an improvement over the control, while the salinity structure in the UOI has little resemblance to data. The model by itself is capable of producing good salinity and current fields. The UOI salinity cross sections display no penetration of the saline waters from the south across the equator and erroneous deep extension of high salinity around 2°S in the central and eastern basin. The MvOI salinity cross sections are more similar to the observations, although the salinity near the surface at 155°W north of the equator is somewhat low. The MvOI zonal current is the closest to the observed in the western and eastern Pacific with a better representation of the deeper subsurface maxima and a surfacing of the undercurrent at 165°E. The UOI currents reach too deep. At the dateline the current structure in MvOI is exaggerated compared to observed but the secondary subsurface maximum at about 4°N (the northern subsrface countercurrent) is captured in the assimilation. UOI is again too strong too deep south of the equator and too weak at the equator. It is apparent from these figures that the MvOI corrects the current structure on and close to the equator better than the statistics of figures 8 and 9 might suggest.

6. Conclusion

Two conceptually different forecast error covariance models were considered in the context of the optimal interpolation data assimilation. One is the univariate model of the temperature error which uses Gaussian spatial covariance function with different scales in zonal, meridional and vertical directions. The second is the multivariate error covariance matrix estimated in the dominant error subspace of empirical orthogonal functions (eofs) generated from Monte Carlo simulations. The latter provides an empirical estimate of the covariability of the errors in temperature, salinity and current fields and spatial structure consistent with the governing dynamics. Thus during an assimilation cycle not only the temperature field, but the entire ocean state vector can be updated.

The univariate assimilation scheme brought the temperature field close to observations, yet the structure of the unobserved fields (salinity and currents) deteriorated quickly, precluding long-term integration. The multivariate scheme is more successfully corrects the salinity and currents as verified by independent observations.

The empirical error covariance model presented in this study is an initial estimate of the forecast error covariance, and is used throughout the assimilation under the assumption that the forecast error statistics do not change significantly in time or after prior assimilation. The robustness of such an estimate was investigated and it was found that it does not exhibit significant seasonal or interannual variability, although there are not enough simulation years to distinguish among statistics during El Niño, La Niña and normal years.

The empirical multivariate forecast error covariance model provides important information regarding the error statistics of all the model fields, prognostic or diagnostic. This gives a natural way to include into the state estimation process the observations of different kinds, for example, the sea surface height, which is often a model diagnostic.

Further developments are underway in implementing the MvOI method for the global ocean model configuration, particularly improving the ensemble statistics by including synoptic perturbations to the forcing fields, perturbations to the model parameters and initial conditions. It is more natural, taking into account the Poseidon ocean model formulation, to consider the covariances of the model variables within the quasi-isopycnal layers. Investigations are also underway to make the MvOI scheme more efficient in a reduced space by including only a limited number of leading eofs.

7. Acknowledgments

Computer time was provided by the NASA Center for Computational Sciences at NASA/Goddard Space Flight Center. GCJ was supported by the NOAA Office of Oceanic and Atmospheric

Research, the NOAA Office of Global Programs and the NASA Physical Oceanography Program.

REFERENCES

S. C. Bloom, L. L. Takacs, A. M. D. Silva, and D. Ledvina, 1996: Data assimilation using incremental amalysis updates. *Mon. Wea. Rev.*, **124**.

A. Borovikov, M. M. Rienecker, and P. S. Schopf, 2001: Surface heat balance in the equatorial Pacific Ocean: Climatology and the warming event of 1994-95. *Journal of Climate*, **14(12)**, 2624–2641.

J. A. Carton and E. C. Hackert, 1990: Data assimilation applied to the temperature and circulation in the tropical Atlantic, 1983-1984. *J. of Phys. Oceanogr.*, **20(8)**.

S. E. Cohn, 1997: An introduction to estimation theory. *Journal of the Meteorological Society of Japan.* Special issue dedicated to "Data Assimilation in Meteorology and Oceanography: Theory and Practice".

G. Evensen, 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **C99**, 10143–10162.

G. Evensen, 1997: Application of ensemble integrations for predictability studies and data assimilation. *Monte Carlo Simulations in Oceanography, Proceedings 'Aha Huliko'a Hawaiian Winter Workshop.*

H. P. Freitag, Y. Feng, L. J. Mangum, M. P. McPhaden, J. Neander, and L. D. Stratton, 1994: Calibration procedures and instrumental accuracy estimates of tao temperature, relative humidity and radiation measurements. Erl pmel-104, NOAA.

I. Fukumori, J. Benveniste, C. Wunch, and D. B. Haidvogel, 1993: Assimilation of sea surface topography into an ocean circulation model using a steady-state smoother. *J. of Phys. Oceanography*, **23**.

G. Gaspari and S. E. Cohn, 1999: Contruction of correlatoin functions in two and three dimensions. *Quart. J. Roy. Meteorol. Soc.*, **125**.

M. Ghil and P. Malanotte-Rizzoli, 1991: Data assimilation in meteorology and oceanography. *Advances in Geophysics*, **33**.

P. Houtekamer and H. Mitchell, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129**, 123–137.

M. Ji, A. Leetmaa, and J. Derber, 1995: An ocean analysis system for seasonal to interannual climate studies. *Mon. Wea. Rev.*, **123**, 460–481.

G. C. Johnson, M. J. McPhaden, G. D. Rowe, and K. E. McTaggart, 2000: Upper equatorial Pacific Ocean current and salinity variability during the 1996-1998 El niño-La Niña cycle. *J. of Geophys. Res.*, **105**, 1037–1053.

R. Kalman, 1960: A new approach to linear filtering and prediction problems. *J. Basic Eng.*, **D82**, 35–45.

A. Kaplan, Y. Kushnir, M. A. Cane, and M. B. Blumenthal, 1997: Reduced space optimal analysis for historical data sets: 136 years of Atlantic sea surface temperature. *J. of Geophys. Res.*, **102**, 27835–27860.

C. L. Keppenne and M. M. Rienecker, 2001: Initial testing of a parallel ensemble Kalman filter with the Poseidon isopycnal ocean general circulation model. *Mon. Wea. Rev.*, **130**, 2951–2965.

A. C. Lorenc, 1986: Analysis methods for numerical weather prediction. *Quart. J. R. Met. Soc.*, **112**, 1177–1194.

M. J. McPhaden, A. J. Busalacchi, R. Cheney, J.-R. Donguy, K. S. Gage, D. Halperin, M. Ji, P. Julian, G. Meyers, G. T. Mitchum, P. P. Niiler, J. Picaut, R. W. Reynolds, N. Smith, and K. Takeuchi, 1998: The tropical ocean global atmosphere observing system: A decade of progress. *J. of Geophys. Res.*, **103**, 14169–14240.

R. W. Preisendorfer, 1988: *Principal component analysis in meteorology and oceanography.* Elsevier, New York.

R. D. R., 1991: *Atmospheric Data Analysis.* Cambridge University Press.

M. M. Rienecker and R. N. Miller, 1991: Ocean data assimilation using optimal interpolation with a quasi-geostrophic model. *Journal of Geophysical Research*, **96(C8)**, 15093–15103.

A. Rosati, R.Gudgel, and K.Miyakoda, 1996: Global ocean data assimilation system. *Modern Approaches to Data Assimilation in Ocean Modeling*, Elsevier.

H. Samet, 1990: *Applications of Spatial Data Structures.* Addison-Wesley Publishiing Company, Inc.

P. S. Schopf and A. Loughe, 1995: A reduced-gravity isopycnal ocean model - hindcasts of El Niño. *Mon. Wea. Rev.*, **123**, 2839–2863.

M. J. Suarez and L. L. Takacs, 1995: Documentation of the Aries/GEOS dynamical core Version 2. Technical Memorandum 104606 5, NASA.

A. Troccoli and K. Haines, 1999: Use of the temperature-salinity relation in a data assimilation context. *J. Atmos. Oceanic Technol.*, **16**, 2011–2025.

A. Troccoli, M. M. Rienecker, C. L. Keppenne, and G. C. Johnson, 2003: Temperature data assimilation with salinity corrections: Validation in the tropical Pacific Ocean, 1993-1998. Technical report, NASA GSFC.

P. J. Van Leeuwen and G. Evensen, 1996: Data assimilation and inverse methods in terms of a probabilistic formulation. *Monthly Weather Review*, **124**.

S. Wilson, 2000: Launching the Argo armada. *Oceanus*, **42**, 17–19.

P. P. Xie and P. Arkin, 1997: Global precipitation: a 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *B. Am. Meteorol. Soc.*, **11**, 2539–2558.

S. Yangand, K.-M. Lau, and P. S. Schopf, 1999: Sensitivity of the tropical Pacific Ocean to precipitation-induced freshwater flux. *Climate Dynamics*, **15**, 737–750.

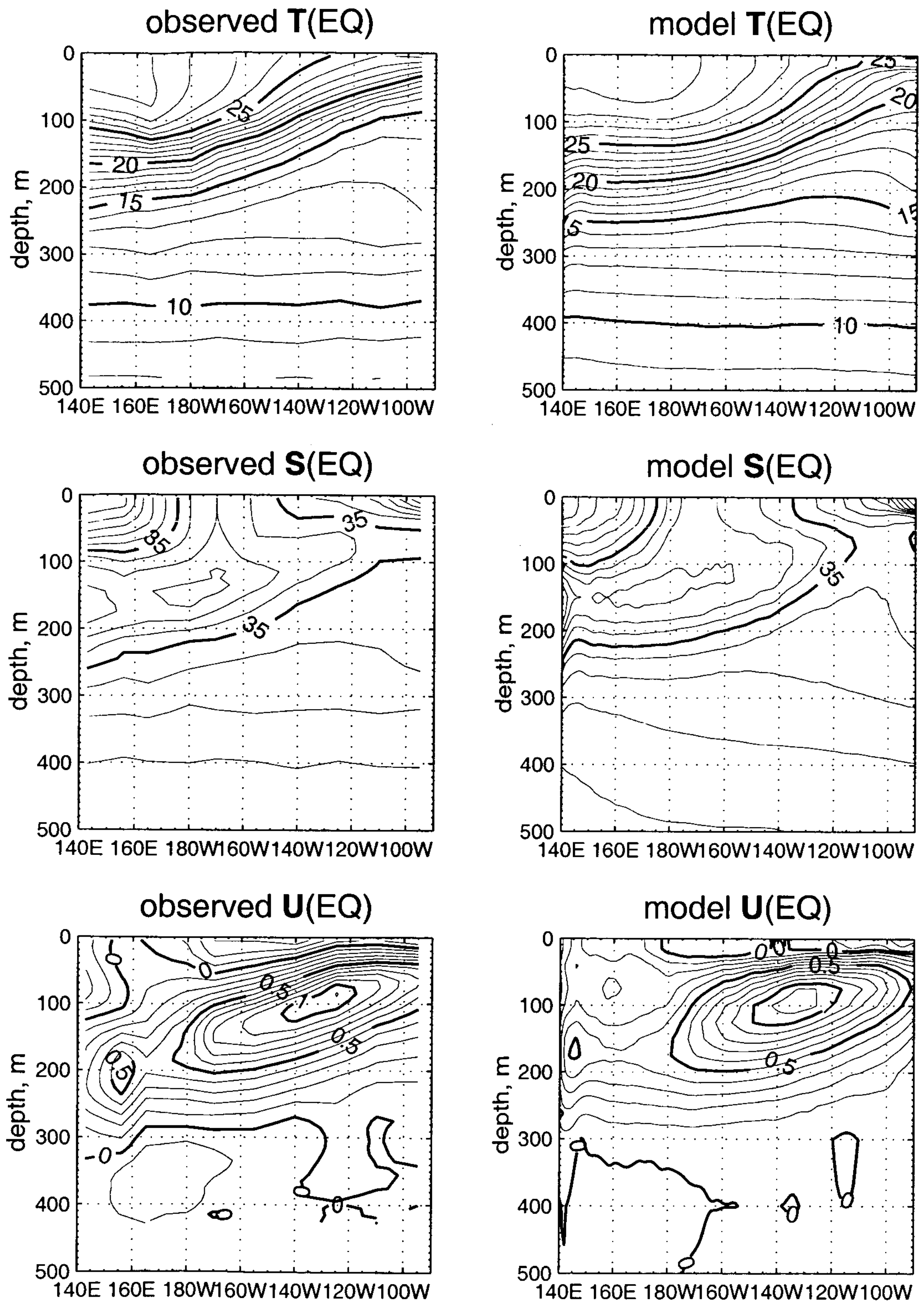

Figure 1: Equatorial cross-section of the Poseidon model means (1988-1997) of temperature, salinity and zonal velocity (right panels) and corresponding data-based estimates (left panels) from Johnson et al.(2002).

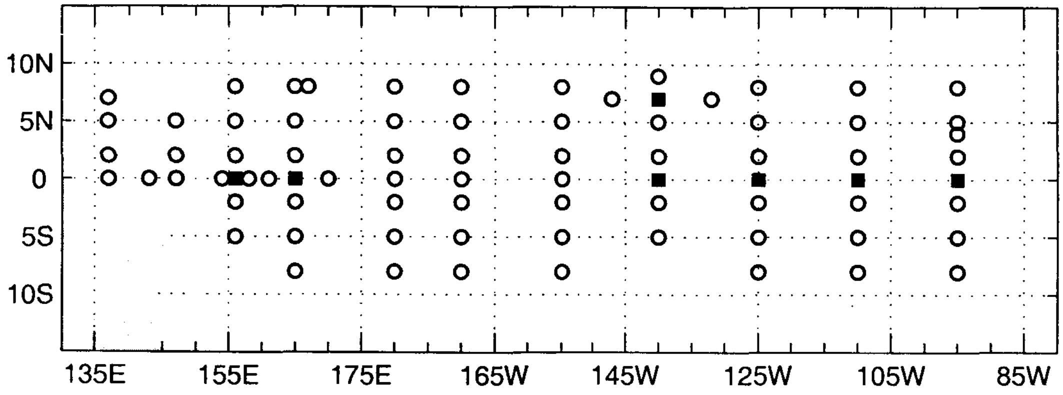

Figure 2: Map of the TAO array, consisting of approximately 70 moored ocean buoys in the Tropical Pacific Ocean. Squares indicate locations of the buoys equipped with current meters.

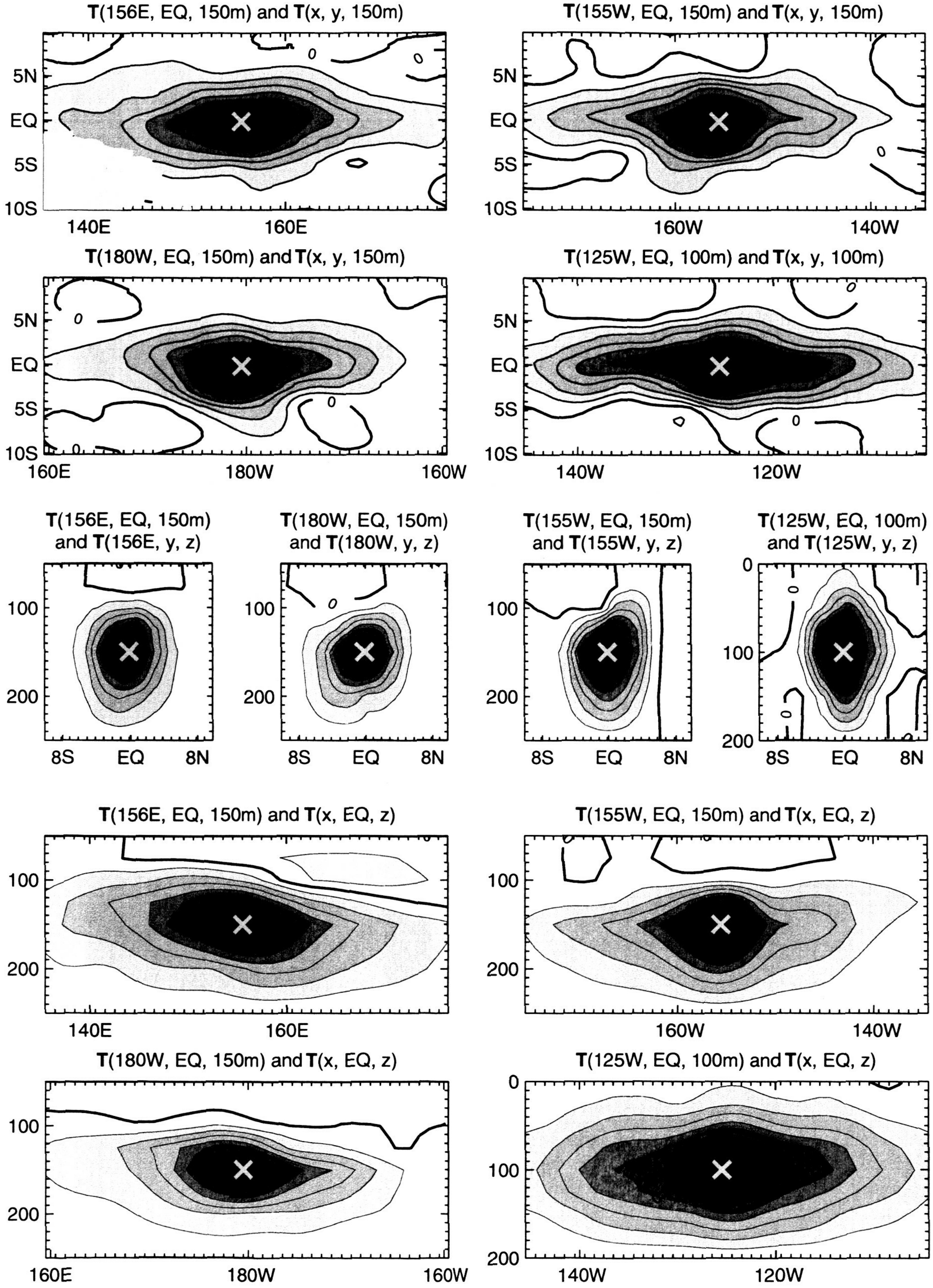

Figure 3: Examples of correlation structure derived from a 160 member ensemble with compact support. Contour interval is 0.1. The cross marks the position of the simulated observation.

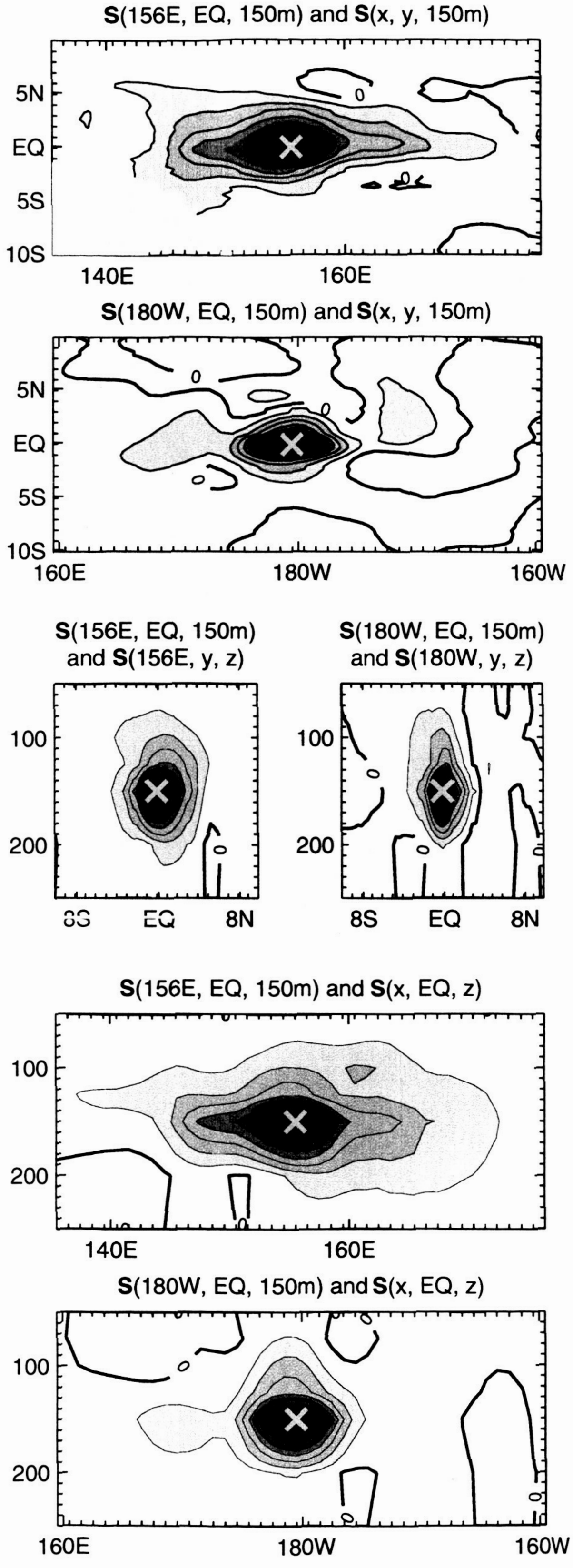

Figure 4: Examples of correlation structure derived from a 160 member ensemble with compact support. Contour interval is 0.1. The cross marks the position of the simulated observation.

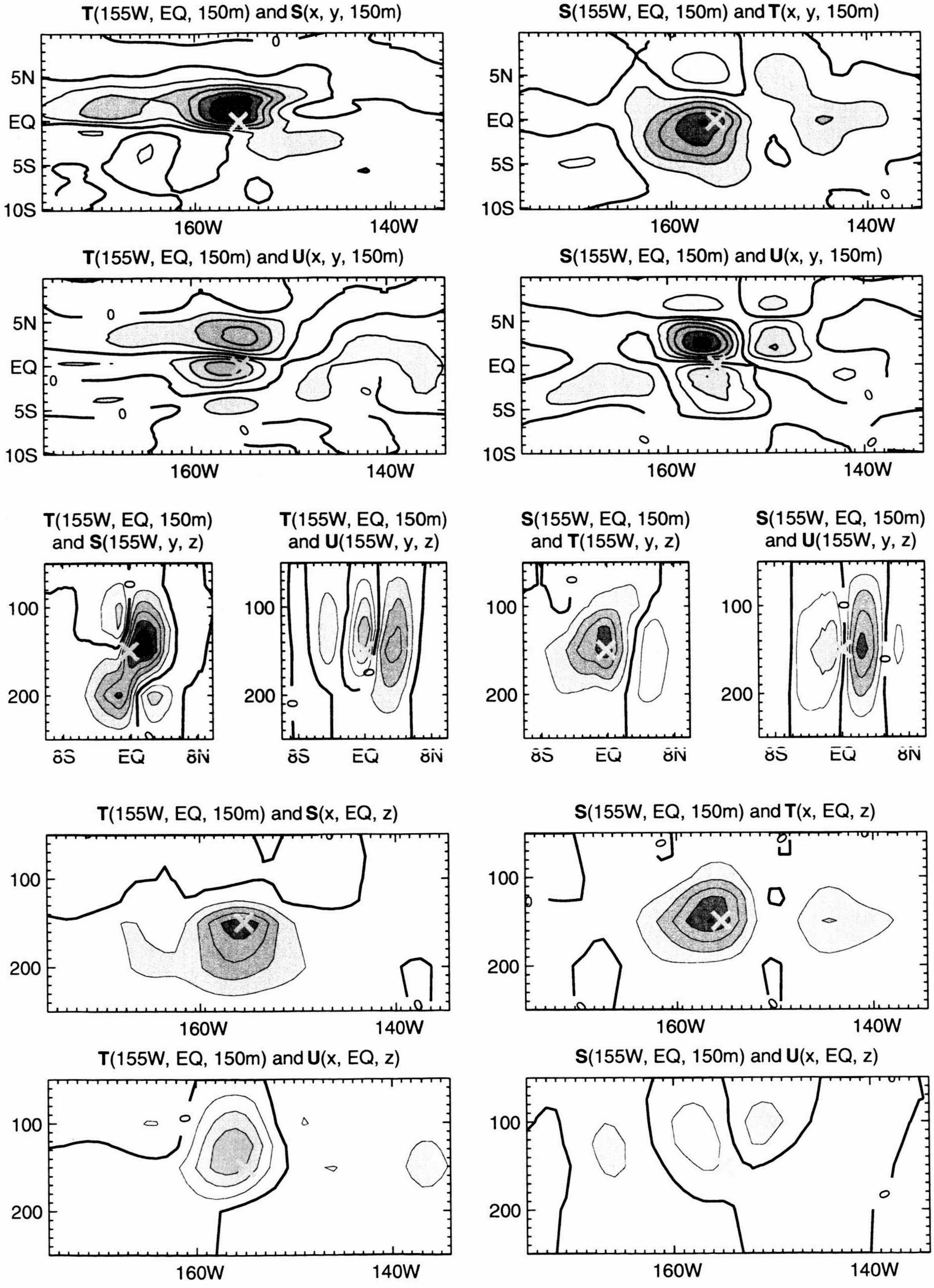

Figure 5: Examples of correlation structure derived from a 160 member ensemble with compact support. Various combinations of observed and updated variables are presented. Contour interval is 0.1. The cross marks the position of the simulated observation.

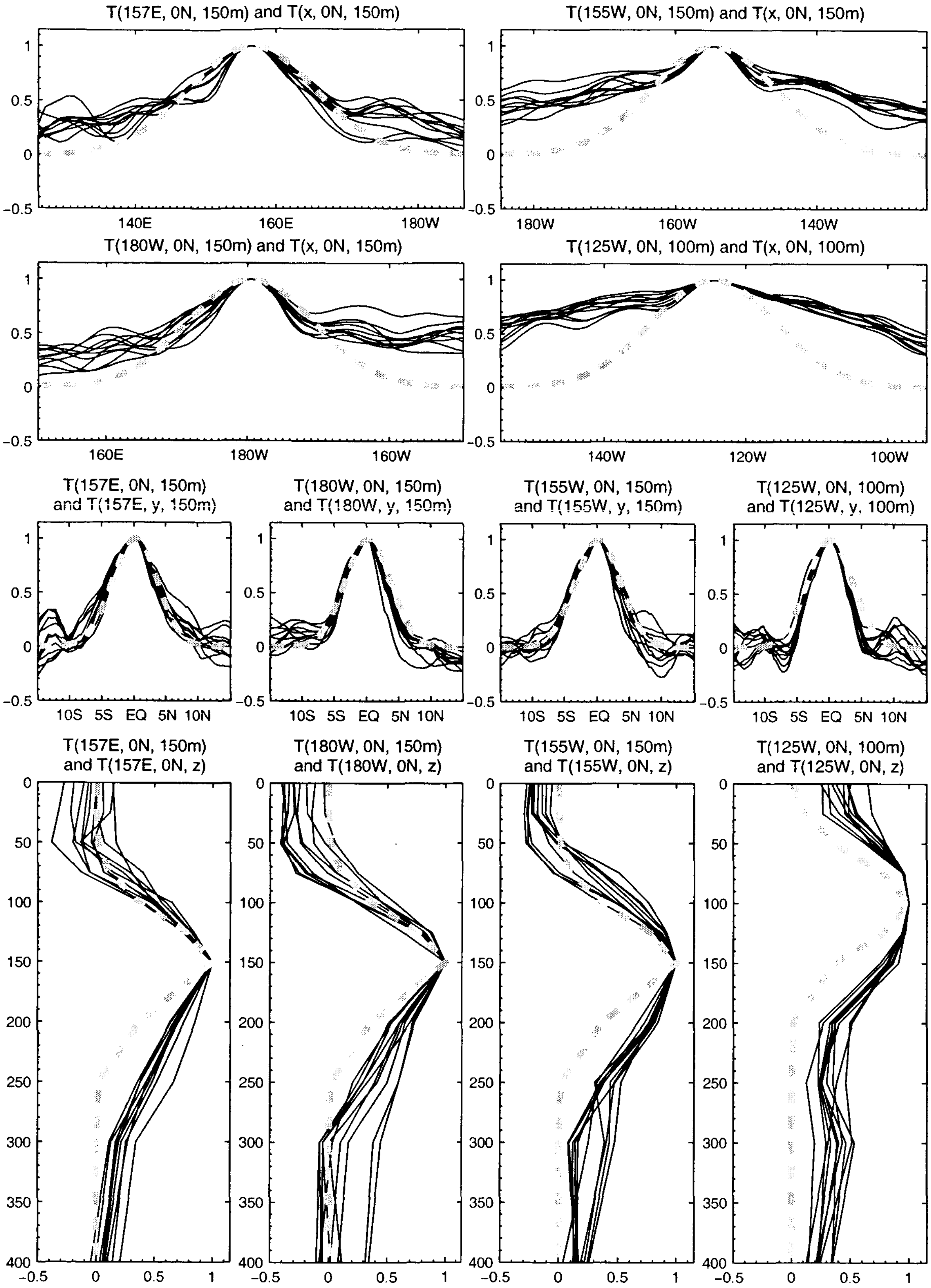

Figure 6: One-dimensional decorrelation curves (zonal, meridional and vertical directions) corresponding to the simulated observation at the specified locations. Each thin solid line produced by a different realization of the error covariance matrix. Dashed grey lines show the Gaussian functional error covariance model used in UOI.

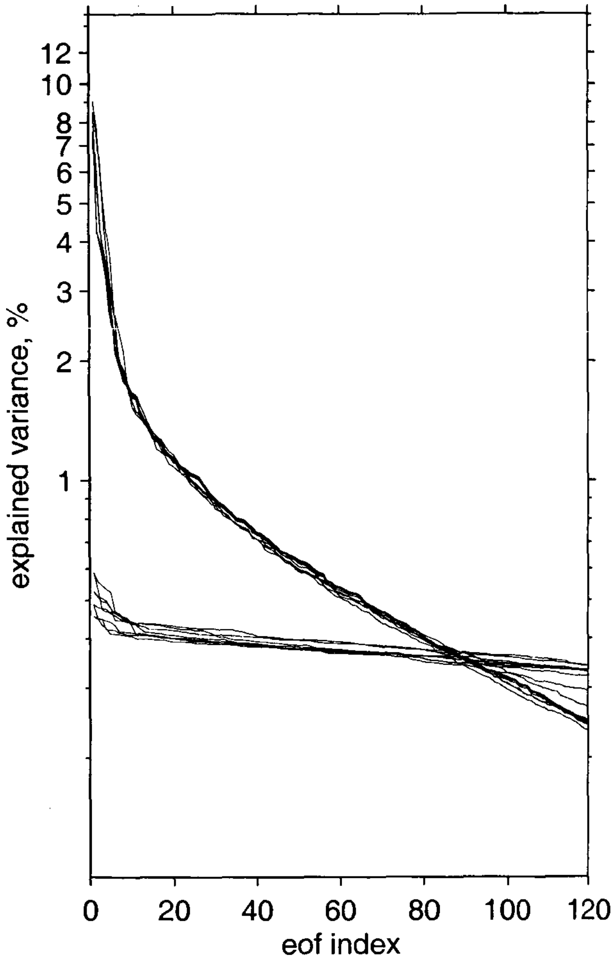

Figure 7: Eigenvalues for several realizations of the matrix **P** (marked α) and the eigenvalues for ensembles of δ's - the residuals of the projections of an arbitrary collection of anomalous ocean states onto a basis of eofs.

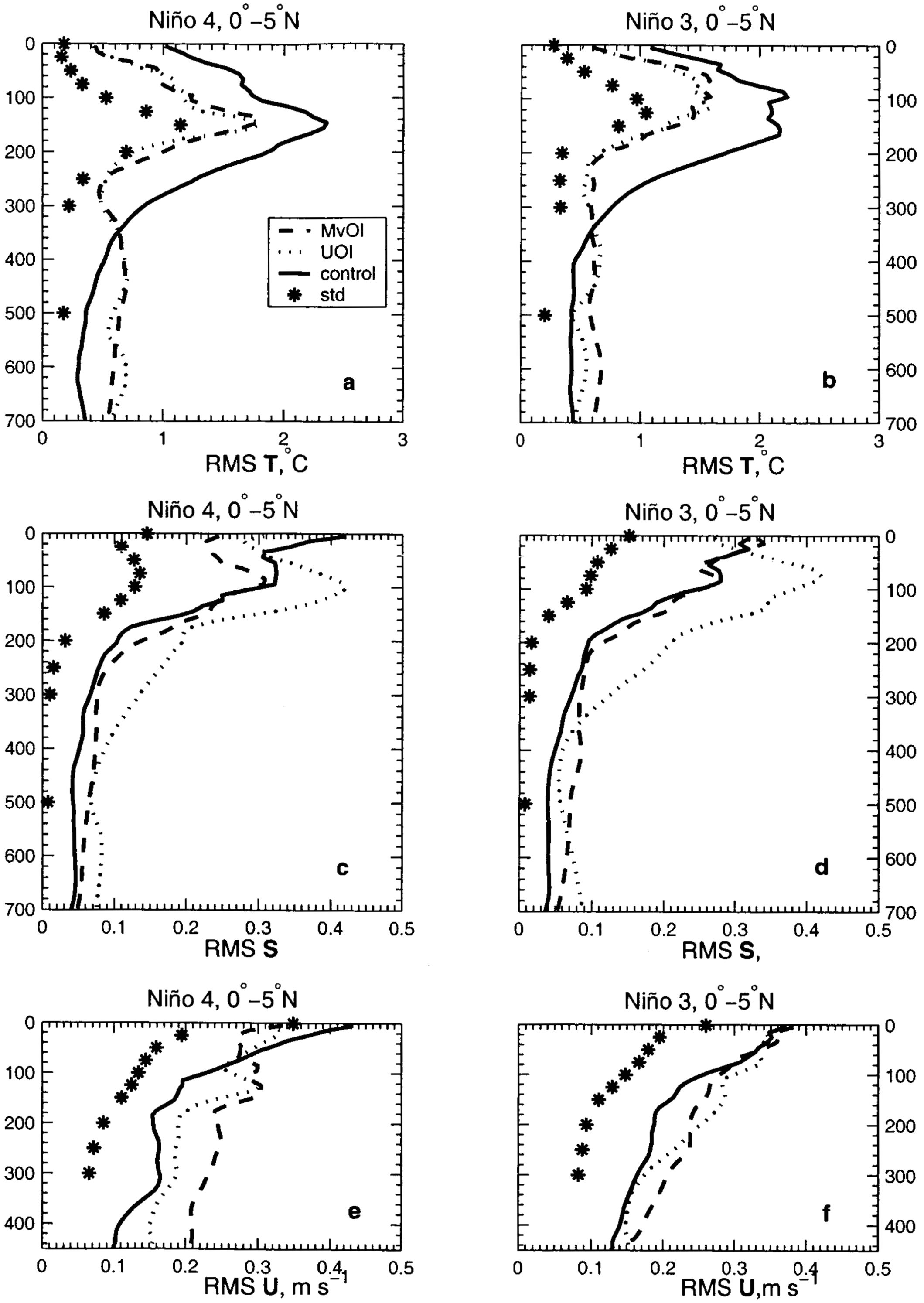

Figure 8: RMSD between the three model runs (UOI, MvOI and control) and the observations as a function of depth for the 35 transects. Statistics are grouped by Niño 4 (160°E-150°W) and Niño 3 (150°W-90°W) regions, and each area is further divided into two halves, south and north of the equator (0°-5°N) shown here). Temperature RMSD (a-b), salinity RMSD (c-d) and zonal velocity RMSD (e-f) are shown. The mean monthly standard deviations of the corresponding model fields for the same regions are shown by stars.

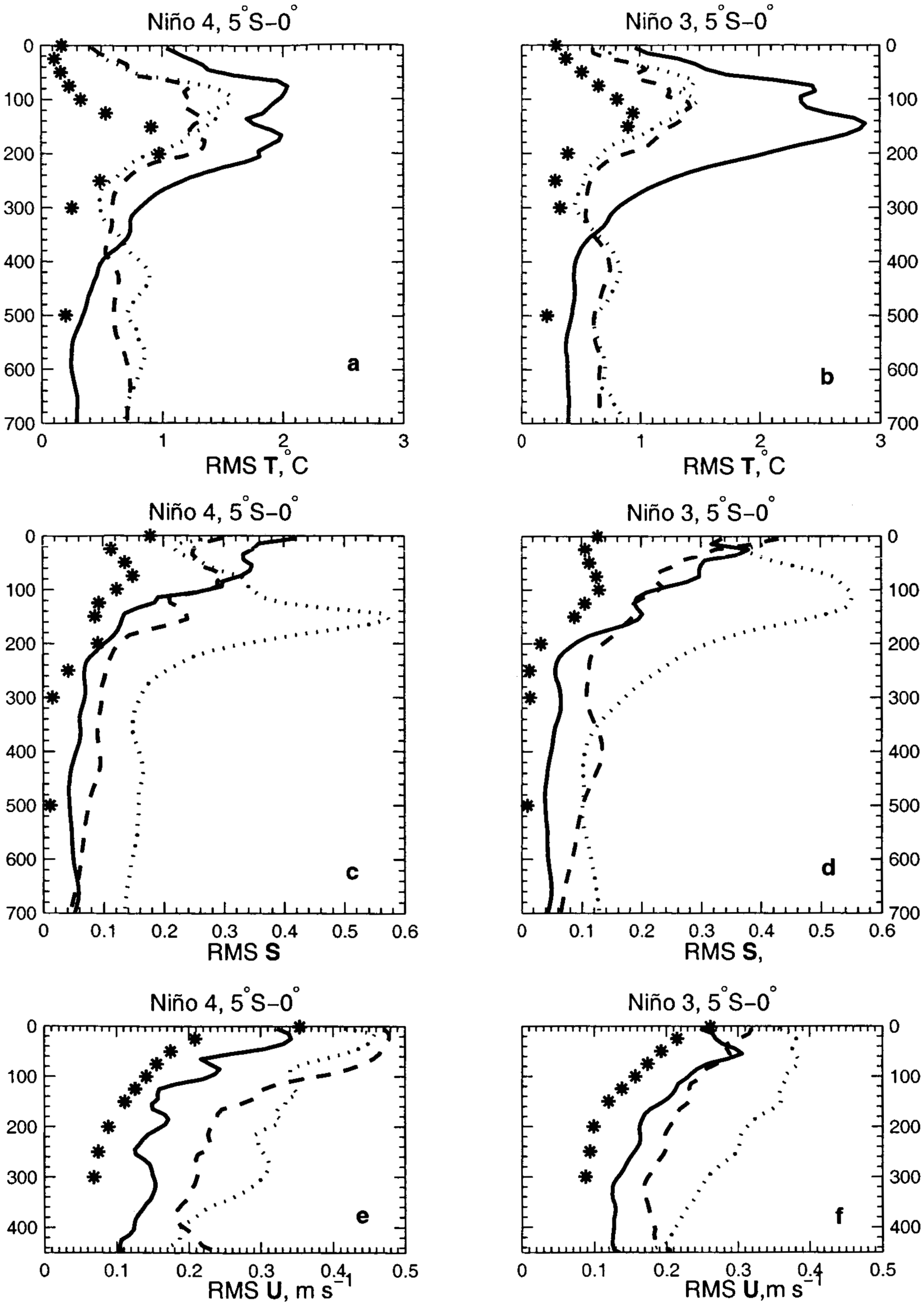

Figure 9: RMSD between the three model runs (UOI, MvOI and control) and the observations as a function of depth for the 35 transects. Statistics are grouped by Niño 4 (160°E-150°W) and Niño 3 (150°W-90°W) regions, and each area is further divided into two halves, south and north of the equator (5°S-0° shown here). Temperature RMSD (a-b), salinity RMSD (c-d) and zonal velocity RMSD (e-f) are shown. The mean monthly standard deviations for the corresponding model fields for the same regions are shown by stars.

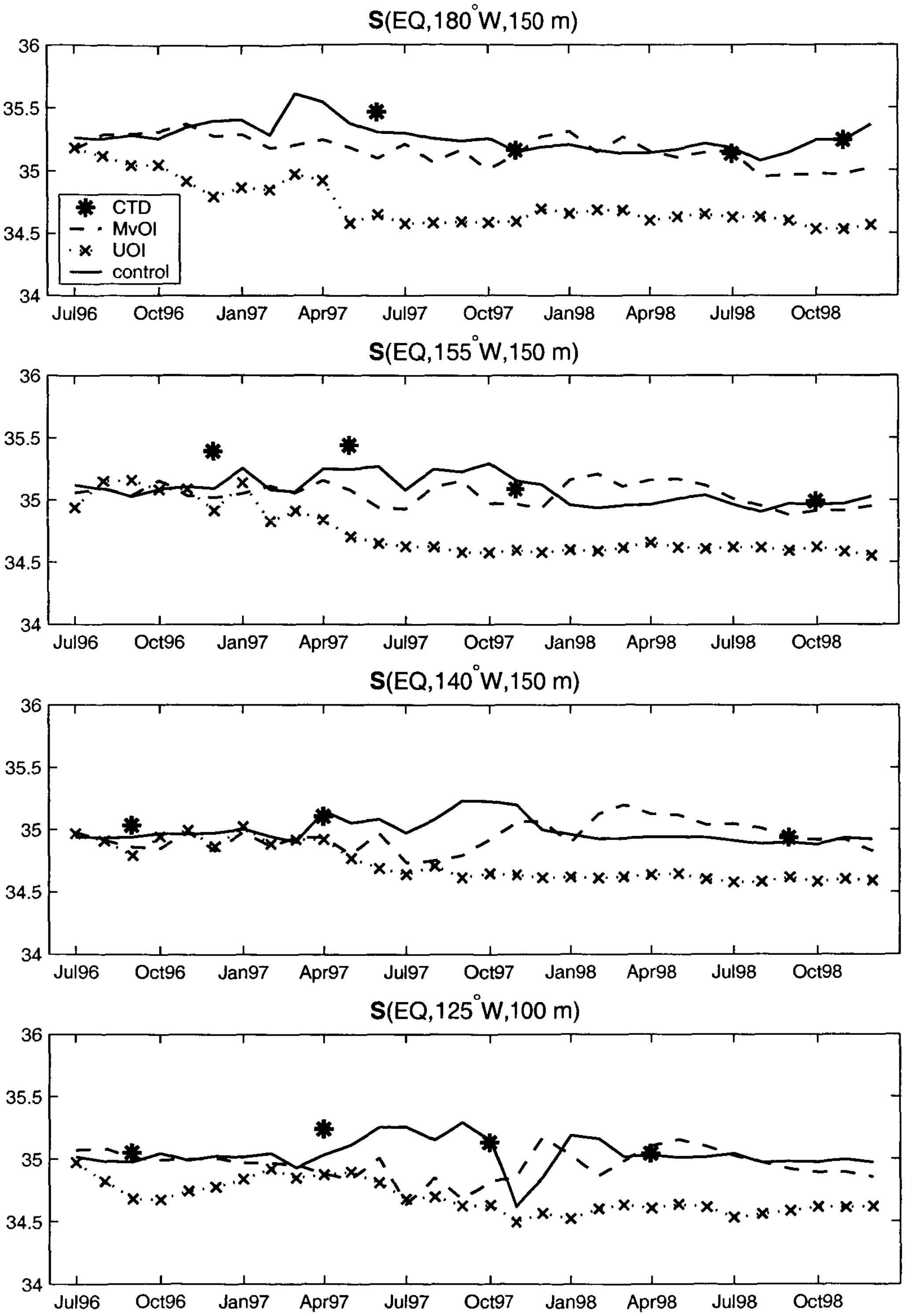

Figure 10: Salinity time series for the control, UOI and MvOI integrations. CTD observations are shown where available. Values are averaged between 2°S-2°N at the specified longitudes.

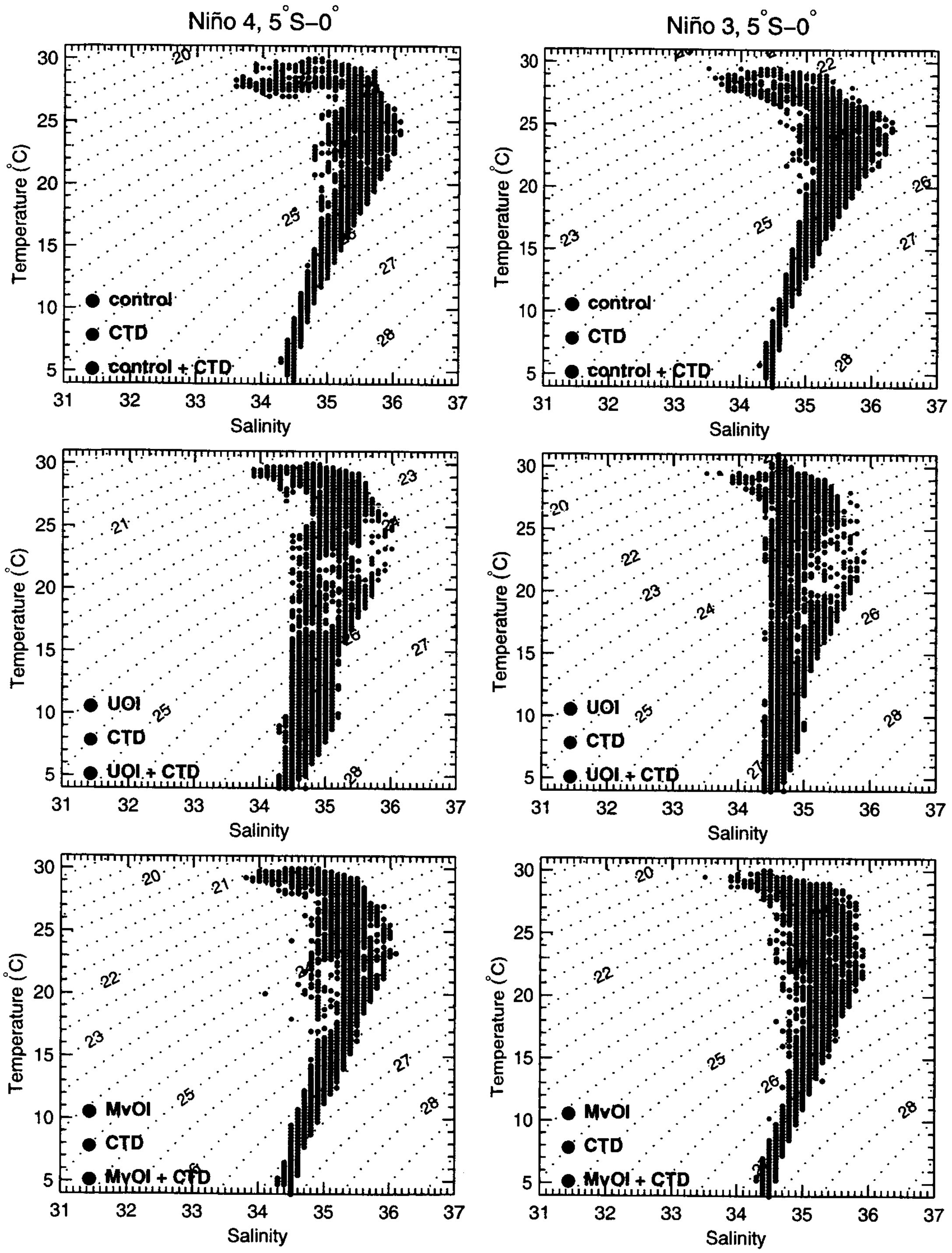

Figure 11: Temperature-Salinity diagram for UOI, MvOI and control experiments for Niño 4 and Niño 3 regions south of the equator. Black dot is plotted for values present only in the model, cyan - only in observations and points where the model and observations agree are shown in red.

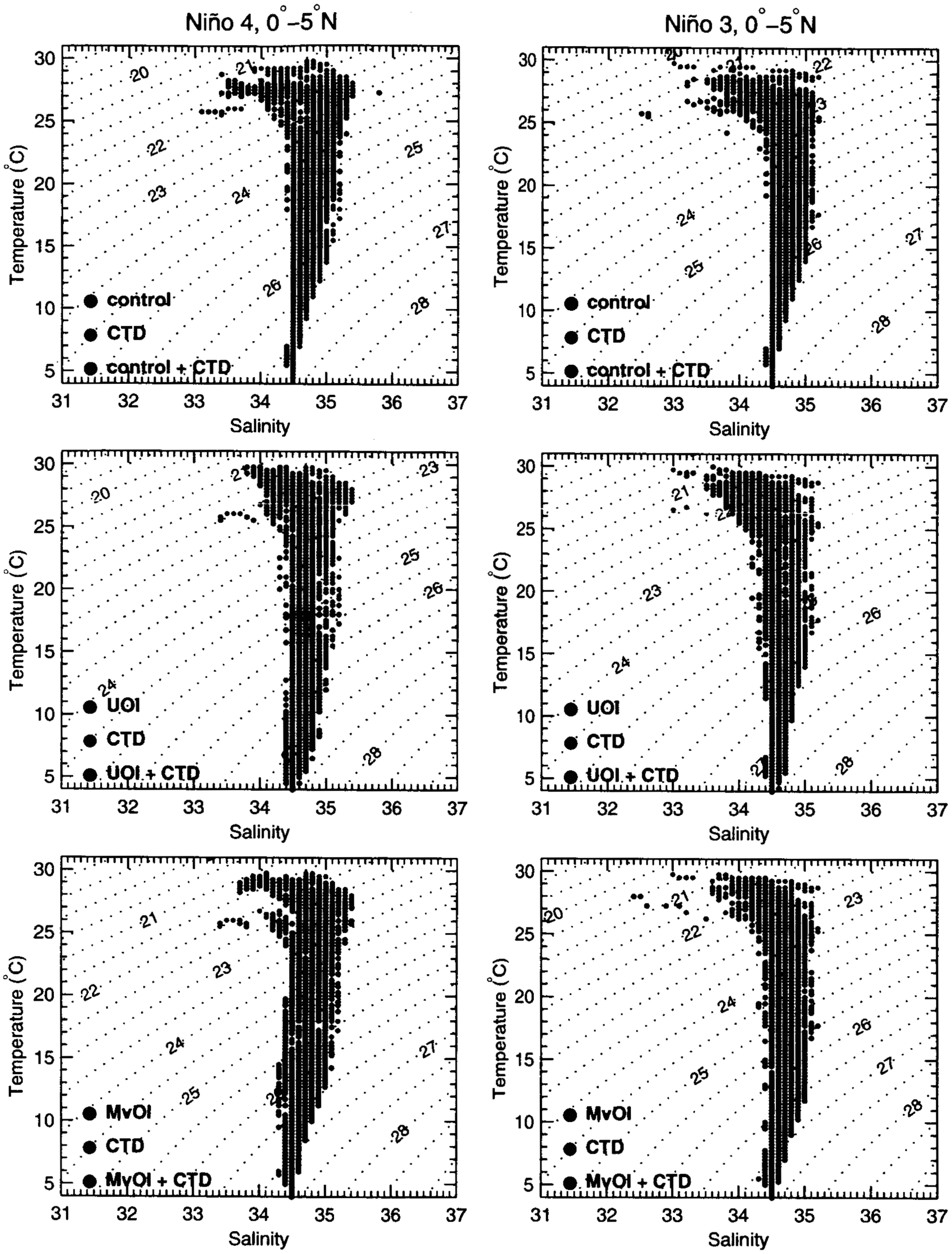

Figure 12: Temperature-Salinity diagram for UOI, MvOI and control experiments for Niño 4 and Niño 3 regions north of the equator. Black dot is plotted for values present only in the model, cyan - only in observations and points where the model and observations agree are shown in red.

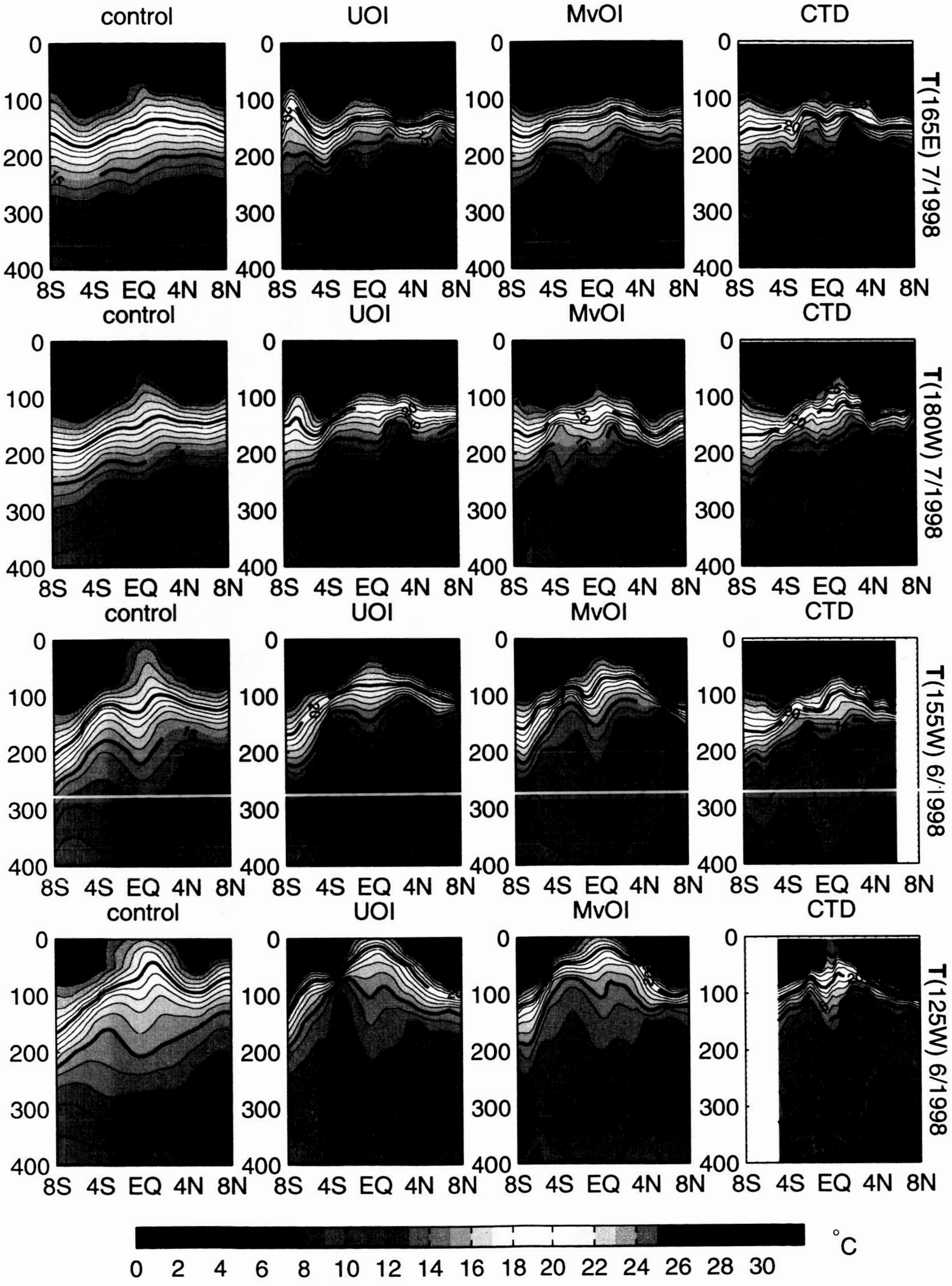

Figure 13: Meridional profiles of the model and observed temperature. Model fields are averaged over 1 month, whereas the observations are from individual quasi-synoptic meridioanl CTD/ADCP sections (following Johnson et al., 2000).

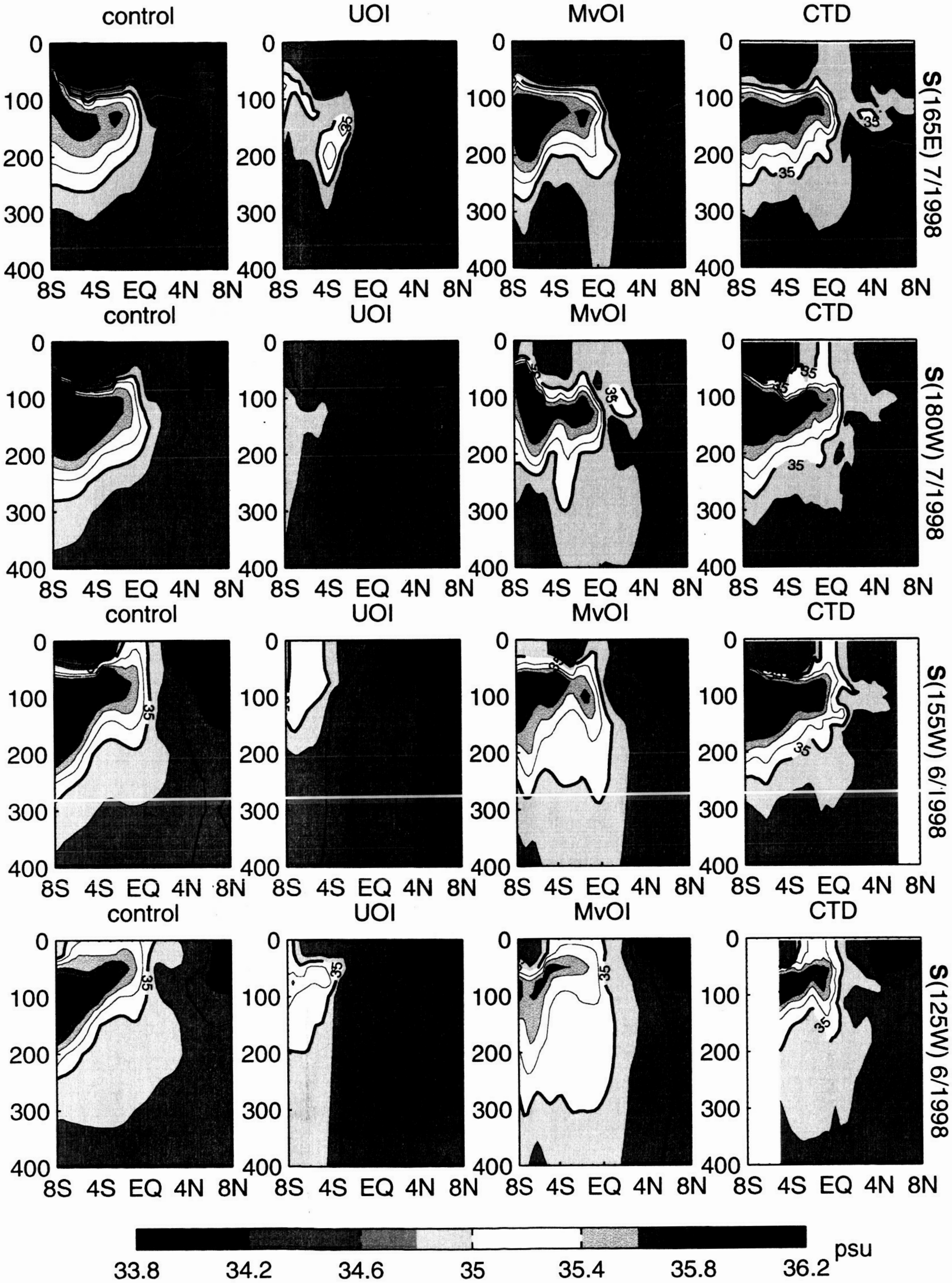

Figure 14: Meridional profiles of the model and observed salinity. Model fields are averaged over 1 month, whereas the observations are from individual quasi-synoptic meridioanl CTD/ADCP sections (following Johnson et al., 2000).

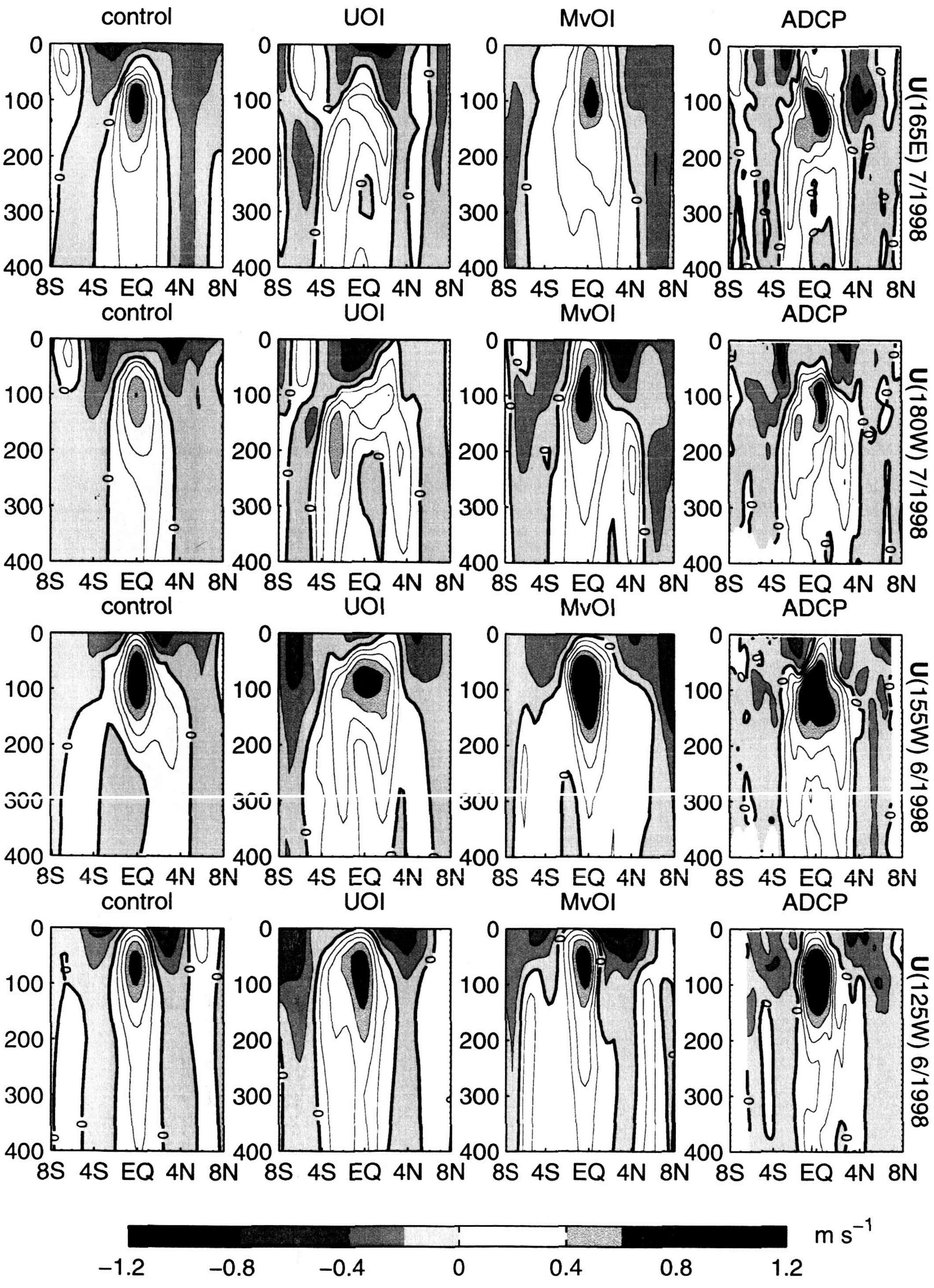

Figure 15: Meridional profiles of the model and observed zonal current. Model fields are average over 1 month, whereas the observations are from individual quasi-synoptic meridioanl CTD/ADCP sections (following Johnson et al., 2000).