A Statistical Approach to Testing Equatorial Ocean Models with Observed Data*

CLAUDE FRANKIGNOUL AND CHRISTINE DUCHENE

Laboratoire d'Océanographie Dynamique et de Climatologie, Universite Pierre et Marie Curie, Paris, France

MARK A. CANE

Lamont Doherty Geological Observatory, Palisades, New York

(Manuscript received 5 July 1988, in final form 1 February 1989)

ABSTRACT

A model testing procedure based on multivariate statistical analysis has been developed to provide an objective measure of the fit between ocean model simulations and observations, taking into account the uncertainties in the atmospheric forcing and the inaccuracies in the oceanic data.

The method is applied to the seasonal variations in an ocean model of the tropical Atlantic. A wind-driven linear multimode model with a simple mixed-layer is tested against surface currents estimated from ship drifts. The uncertainties in the observations, and in the model response due to random errors in the wind stress and its interannual variability create a substantial indeterminacy in the available sample, but it is shown that they do not explain the large discrepancies that are found between observed and modeled seasonal surface currents. Nor are the uncertainties in the wind stress bulk formulation sufficient to account for the model-data differences. These can then be attributed unambiguously to the oversimplification of the model physics.

The use of the method in model tuning is illustrated by determining the vertical resolution that provides an optimal fit to the observed surface currents. The linear model works best with only one vertical mode, and a mixed-layer depth of 40 m. However, the discrepancies with the observations remain too large for the improvement in model performance to be statistically significant.

1. Introduction

Much progress has been made in the modeling of the equatorial oceans, which respond primarly to changes in the wind stress. "Rather successful" simulations of the variations in thermocline depth, sea level and sea surface temperature have been achieved with simple models (e.g., Busalacchi and O'Brien 1981; Cane 1984; Du Penhoat and Treguier 1985), and the whole equatorial current system has been "reproduced realistically" with general circulation models (e.g., Philander and Pacanowski 1986). Usually, the degree of success is assessed subjectively, based on visual comparisons with observations. While such comparisons can reveal obvious differences, they become less effective as the models fidelity increases, and, more importantly, they are inadequate for separating the effects of model inadequacies from the large uncertainties in the atmospheric forcing and the oceanic data. Correlations between model output and observations also fail to distinguish between model and data errors. Yet, as there is considerable interest in developing realistic ocean models for climate studies, it is necessary to test the models objectively, in a way that takes into account the observational uncertainties.

The wind stress is usually estimated from the historical ship reports, which are inaccurate, gappy, and unequally distributed. There may be enough wind data to document the mean seasonal variations of the wind stress above the tropical oceans (Hellerman and Rosenstein 1983), but the data base is barely adequate to describe its interannual variability. Gridded fields of monthly averaged wind stress for the last few decades can only be constructed by using subjective or objective analysis to eliminate obvious errors and fill data gaps (e.g., Goldenberg and O'Brien 1980; Servain et al. 1985). There also remain large uncertainties in the bulk formulations used to estimate wind stress from wind data (Blanc 1985), and information on the high frequencies are limited. Even less reliable information is available on the surface heat exchanges, since they are estimated with larger errors and biases. Uncertainties in the meterological forcing thus limits the accuracy to be expected in the model simulations, even if the physics were perfectly represented. Furthermore, the oceanic initial conditions are usually unknown, and

* Lamont Doherty Geological Observatory Contribution Number 4503.

Corresponding author address: Dr. Claude Frankignoul, Université Paris VI, LODYC, 4 Place Jussieu, Tour 14, 2eme Etage, 75232 Paris Cedex 05, France.

the data used for model assessment are mostly sparse and noisy and are affected by motions not represented in the models. More extensive observations are becoming available in the tropics and better initial conditions can be provided by data assimilation, but the data will remain incomplete and noisy.

There seems to be no general methodology for taking account of the data uncertainties in ocean model testing, although some work has been done on the effects of observational errors on model predictions. Using a Monte-Carlo approach, Adamec and Elsberry (1984) have studied the effects of erroneous initial conditions and random errors and biases in the atmospheric forcing on mixed-layer evolution predicted with a one-dimensional model. Schroeter and Wunsch (1986) have used nonlinear programming to study the effect of uncertainties in the wind forcing of simple steady-state ocean models and have discussed how dynamical and data constraints can influence the "optimal" solution. This latter approach is elegant and promising, but it requires extensive programming and it may not be applicable to complex models because of the computational difficulties. This holds for Wunsch's (1989) attempts to deal with the ill-conditioning of the ocean modeling problem caused by boundary conditions uncertainties. Blumenthal and Cane's (1989) study is principally concerned with effects of model parameter uncertainties and makes less thorough use of the information in the data than the method proposed here. Inverse modeling and data assimilation studies can also provide some information on how good models are, but further developments are needed for their use in model testing, and there are large computational requirements for data assimilation.

Here we develop a new data sensitivity procedure for model testing in the presence of inaccurate oceanic and atmospheric data, which can be used with existing models and involves little programming. The method, which is based on multivariate statistical analysis, involves the construction of detailed error models for the atmospheric forcing and the oceanic observations.

Although the response of the equatorial ocean to atmospheric forcing is primarly deterministic, the large uncertainties in the atmospheric input and in the oceanic initial conditions create a corresponding uncertainty in the model output. Given the uncertainties in the oceanic observations, the model-reality intercomparison consists of comparing noisy multidimensional fields with large correlation scales. Model testing then reduces to estimating if the null hypothesis that models and observations are identical is rejected at a given level of statistical significance. In practice, the dimension of the fields is large, while the error models are only approximate and the sample size is limited. Hence, as in climate experiments with general circulation models of the atmosphere (Hasselmann 1979; Preisendorfer and Barnett 1983; Hannoschöck and Frankignoul 1985), only a limited number of details can be considered in the assessment of significance. The models can thus only be tested after performing a data compression, which greatly reduces the dimension of the intercomparison space.

Although the testing procedure is general and can be applied to any set of observations, its detailed formulation depends on the nature of the data used to verify the model predictions. In this paper, we develop the method for the case of seasonal variations in a tropical ocean. It is illustrated by testing the linear, multimode model of Cane (1984), which is coarse but economical, and has been used successfully in ocean–atmosphere studies (Zebiak and Cane 1987). We consider the seasonal changes of surface currents in the tropical Atlantic, using the ship-drift data of Richardson and McKee (1984) as an observational basis for the model testing.

The applications of the testing procedure are numerous. In particular, model intercomparisons and tuning can be done in a more objective and effective manner, or an optimal model resolution easily established, as illustrated here for the vertical resolution of the linear model. The statistical procedure should also be useful in model development, showing for example whether the available data is extensive and accurate enough to reveal shortcomings and improvements in model performance.

The plan of the remainder of the paper is as follows. In section 2, the testing procedure is described in a general context. Section 3 discusses the surface current data and its compression, while the error model is constructed in appendix A. Section 4 discusses the wind stress data and its uncertainties. Section 5 tests a "standard" version of the multimode model. Section 6 shows how the testing procedure may be used in model tuning. Conclusions are given in the final section.

2. The multivariate model testing procedure

To test a model, it is appropriate to work in the whole space where observations are sufficiently reliable, and where the model is believed to be realistic. Let the observed time series (the seasonal cycle in our case) at the N points of such a region be denoted by the N-dimensional vector $\mathbf{d}^t$, where the time $t = 1, T$ is denoted by an upper index. Because of measurement and sampling errors, including that due to interannual variability and high frequency motions, the observed signal can be considered to be the sum of the true seasonal signal $\langle \mathbf{d}^t \rangle$ and an error $\delta\mathbf{d}^t$. If the normality assumption can be made, the observational uncertainties are described by the lagged error covariance matrix $\langle \delta\mathbf{d}^t \delta\mathbf{d}^{q\prime} \rangle$, where the prime denotes a vector transpose. An estimate $\mathbf{D}^{tq}$ of this matrix can generally be constructed by an error analysis of the observations, as illustrated in appendix A, or it can be directly estimated when sufficiently long time series are available. This leads to the definition of a probability region (an ellip-

soid) in which there is, say, a 95% chance that the true seasonal cycle lies.

For comparison purposes, the model simulations are considered in the same space, which may require suitable averagings and interpolations. Let $\mathbf{m}^t$ denote the model sesaonal response, which is a linear or nonlinear function of the atmospheric forcing and the oceanic initial conditions, denoted as $\mathbf{f}$. Since the input fields are not known perfectly, $\mathbf{f}$ can also be considered to be the sum of the true forcing $\langle \mathbf{f} \rangle$ and an error $\delta\mathbf{f}$. The model seasonal response can then be written as the sum of the model response to the true forcing and that resulting from the forcing error, $\mathbf{m}^t = \langle \mathbf{m}^t \rangle + \delta\mathbf{m}^t$. For the simplest linear models, the error covariance matrix $\langle \delta\mathbf{m}^t \delta\mathbf{m}^{q'} \rangle$ can be estimated analytically from the error covariance matrix of the input fields $\langle \delta\mathbf{f}\delta\mathbf{f}' \rangle$. For more complex models, it must be estimated from numerical experiments. Suppose that different sets of plausible forcing fields and initial conditions are available or can be constructed by estimating their likely errors, for instance using a Monte-Carlo approach. Then, different plausible model outputs can be produced. An estimate $\mathbf{M}^{tq}$ of the error covariance matrix due to forcing uncertainties can be calculated, or a probability region for $\langle \mathbf{m}^t \rangle$ constructed.

In principle, we can test the null hypothesis that the model has no errors due to misrepresentation of the physics, so that its response to the true forcing should represent the true seasonal signal perfectly, $\langle \mathbf{m} \rangle = \langle \mathbf{d} \rangle$, where $\mathbf{d}' = (\mathbf{d}^{1'}, \mathbf{d}^{2'}, \cdots \mathbf{d}^{T'})$ and $\mathbf{m}' = (\mathbf{m}^{1'}, \mathbf{m}^{2'}, \cdots \mathbf{m}^{T'})$. Of course, no model is perfect, so the null hypothesis per se is not of interest. However, the test is equivalent to asking whether the available data is good enough to reveal shortcomings in the model, which is of interest. For example, to decide upon attempting further model improvement: if the data is not good enough, it cannot be used to decide if a change has in fact made the model better.

The problem of comparing two sample vectors $\mathbf{d}$ and $\mathbf{m}$ with unequal dispersion matrices is a classical one. If the data are multinormal, and the error covariance matrices can be considered to be known (e.g., because the sample size is very large), an appropriate test statistic is (e.g., Seber 1984)

$$\rho^2 = (\mathbf{d} - \mathbf{m})'(\langle \delta\mathbf{d}\delta\mathbf{d}' \rangle + \langle \delta\mathbf{m}\delta\mathbf{m}' \rangle)^{-1}(\mathbf{d} - \mathbf{m}), \quad (1)$$

If the null hypothesis holds and the covariance matrix has full rank, ρ^2 has the chi-squared distribution with $T \times N$ degrees of freedom. In practice, however, the sample is limited, and the error models are only approximate, so that the covariance matrix is of reduced rank, or, at best, it differs sufficiently from the true one for its smallest eigenvalues to be unreliable. Since the latter dominate the inverse in (1), if it even exists, incorrect results are to be expected. The dimensionality of the system must be strongly reduced, minimizing the role of the troublesome uncertain components.

Data compression can be performed in a number of ways; while in a particular problem other methods may recommend themselves, a general efficient method is to use principal component analysis. First, we determine the space of the major sources of variations in the observations $\mathbf{d}^t$. Assume that the mean has been removed from $\mathbf{d}^t$, and let $\mathbf{D}$ be an unbiased estimate of the data covariance matrix; $\mathbf{D}$ has dimension $N \times N$, but in the usual case where $N > T$, it has only $T - 1$ nonzero eigenvalues. The smallest eigenvalues generally correspond to minor, discardable sources of variations, or to data noise, and they need not be considered. If $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, \cdots \mathbf{e}_k)$ is the matrix of the first K orthonormal eigenvectors (referred to here as empirical orthogonal functions or EOFs) that we select, then for each observation $\mathbf{d}^t$ we can define a K-dimensional vector of sample principal components

$$\mathbf{y}^t = \mathbf{E}'\mathbf{d}^t. \quad (2)$$

Our choice of the EOF base provides the most efficient way to represent the observed seasonal changes, but, in general, it does not suffice to represent the model response, since a part of the latter may be in the null space of the observations. The truncated basis EOFs for the observations may be extended by using the first few, say R, EOFs of the part of the model response that does not project on the observation EOFs. Since these EOFs are orthogonal to the previous ones, we have constructed a more complete basis of $K + R = Q$ orthonormal vectors $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, \cdots \mathbf{e}_Q)$. To give a fair weight to the two subspaces, R may be determined by only keeping the EOFs with eigenvalues at least as large as the smallest one retained in the observations. The increase in dimension should be small if the model is realistic (to include the annual mean in the analysis, a further extension of the truncated base would often be necessary). The reduced data matrix is now

$$\mathbf{Y}' = \mathbf{E}'(\mathbf{d}^1, \mathbf{d}^2, \cdots \mathbf{d}^T). \quad (3)$$

If only the low frequencies are generally well-resolved and energetic, a further data compression can be performed in the time domain, say from dimension T to L:

$$\mathbf{Z}' = (\mathbf{z}^1, \mathbf{z}^2, \cdots \mathbf{z}^L) = \mathbf{Y}'\mathbf{A}, \quad (4)$$

where $\mathbf{A}$ is a $T \times L$ matrix describing the (linear) filtering or averaging process. We will use weighted seasonal means as a time filter. The observational data is now reduced to $Q \times L$ matrix, written in vector form as $\mathbf{s}' = (\mathbf{z}^{1'}, \mathbf{z}^{2'}, \cdots \mathbf{z}^{L'})$.

The error covariance matrix $\langle \delta\mathbf{s}\delta\mathbf{s}' \rangle$ of the seasonal data $\mathbf{s}$ can be partitioned into L^2 submatrices $\langle \delta\mathbf{z}^k \delta\mathbf{z}^{j'} \rangle$ of dimension $Q \times Q$, given by

$$\langle \delta\mathbf{z}^k \delta\mathbf{z}^{j'} \rangle = \sum_{t=1}^{T} \sum_{q=1}^{T} \mathbf{E}' A_{tk} \langle \delta\mathbf{d}^t \delta\mathbf{d}^{q'} \rangle A_{qj} \mathbf{E}, \quad (5)$$

and can be estimated from $\mathbf{D}^{tq}$.

The model prediction $\mathbf{m}^t$ is projected similarly on

the truncated EOF base. The reduced model signal is the data matrix

$$\mathbf{X}' = (\mathbf{x}^1, \mathbf{x}^2, \cdots \mathbf{x}^L) = \mathbf{E}'(\mathbf{m}^1, \mathbf{m}^2, \cdots \mathbf{m}^T)\mathbf{A}, \tag{6}$$

given in vector form by $\mathbf{p}' = (\mathbf{x}^{1'}, \mathbf{x}^{2'}, \cdots \mathbf{x}^{L'})$. The error covariance matrix associated with the forcing uncertainties can be estimated as above from $\mathbf{M}^{fq}$.

If the data compression is large enough, the sample vectors $\mathbf{s}$ and $\mathbf{p}$ can be compared objectively, as the statistic

$$\rho^2 = (\mathbf{s} - \mathbf{p})'(\langle \delta\mathbf{s}\delta\mathbf{s}' \rangle + \langle \delta\mathbf{p}\delta\mathbf{p}' \rangle)^{-1}(\mathbf{s} - \mathbf{p}), \tag{7}$$

should not be affected by small inaccuracies in the error models. If the null hypothesis holds, ρ^2 has the chi-squared distribution with $L \times Q$ degrees of freedom, and the usual acceptance and rejection rules can be applied.

If the error covariance matrices are not known but estimated from samples, we have a multivariate generalization of the so-called Behrens-Fisher problem of comparing two means with different sample variances, for which various solutions have been proposed (e.g., Seber 1984). Suppose that $\mathbf{s}$ and $\mathbf{p}$ have been estimated from n_s and n_p independent years (samples), respectively. Then, if $\mathbf{Q}$ and $\mathbf{P}$ are estimates of the dispersion matrices, a natural test statistic for the null hypothesis is

$$(\mathbf{s} - \mathbf{p})'(\mathbf{Q}/n_s + \mathbf{P}/n_p)^{-1}(\mathbf{s} - \mathbf{p}). \tag{8}$$

Approximate distributions are available for (8), and they can be calculated from the data, providing somewhat more conservative rejection criterions than the chi-squared distribution. The statistic (8) is expected to be robust to normality, since when n_s and n_p are very large, (8) tends to (7) (Seber 1984). Note that observational errors should be normal, but that the interannual variations may not be. However, the EOF projection decreases departures from normality for the principal components, by the central limit theorem applied to linear combinations.

The oceanic case may fall between the two extremes above. In many cases, the observation error covariance matrix $\mathbf{D}^{fq}$ will be prescribed from an error analysis of the observations, and it should be considered as "known" from the statistical viewpoint. On the other hand, meteorological fields can be used to drive the ocean models, and the sample error covariance matrix $\mathbf{M}^{fq}$ estimated from the model results. However, if the uncertainties in the meteorological fields and the oceanic initial conditions have also been simulated, for example, by Monte-Carlo techniques, then their form has been prescribed, rather than computed from a sample, making the choice of an appropriate test statistic difficult. It should be adapted to the case at hand, and to the level of sophistication needed.

3. Seasonal variations of surface currents in the tropical Atlantic

a. *The ship drift data*

Historical ship drifts are the only available data capable of resolving the long-term seasonal variations of surface currents over broad regions. Richardson and McKee (1984) have used ship drift to describe the long-term monthly mean surface currents in the tropical Atlantic. The observations were made over a 100-year period from 1875 to 1976, but most of the data with a known year (66%) were collected during the years 1920–40. The data are nearly evenly distributed seasonally. For 2° × 5° boxes (Richardson and Walsh 1986), the average number of measurements is about 180, but the spatial distribution is not homogeneous as observations are clustered along major shipping routes. Figure 1 shows the data density in the region considered in this study, which is limited to the domain of validity of the numerical model (see below). Each ship-drift measurement of surface current is an average over time (usually 12 to 24 h), over a path of a few hundred kilometers, and over the depth of the ship's hull. There are many sources of errors in the individual velocity observations; Richardson (1983) suggests that the error is primarly random, of the order of 20 cm s^{-1}. There may also be a small bias in the direction of the mean winds due to wind drag on the ships, since the trade winds are very steady.

The analyzed fields consist of mean velocity components u and v, variances $\overline{u'^2}$, $\overline{v'^2}$ and $\overline{u'v'}$, and number of observations n. Using the formalism of section 2, we describe the mean velocity for each month t at the N boxes by the $2 \times N$-dimensional vector $\mathbf{d}^t = (u_1{}^t, v_1{}^t, u_2{}^t, \cdots v_N{}^t)$, $t = 1, 12$. For the region of interest (Fig. 1), $2 \times N$ is equal to 214. However, it was found when performing the data reduction described below that boxes with small data density introduce large uncorrelated noise in the dominant patterns of seasonal variability, so that they should be eliminated. Sensitivity experiments showed that a reasonable cutoff criterion is to retain boxes with more than 180 measurements, which reduces $2 \times N$ to 192.

The observed seasonal current departures from the yearly mean are illustrated in Fig. 2 (left). The yearly mean signal is not considered in this study, as it is governed by physics and global scale processes not represented in the linear model considered below. The dimension of the observation vector is 192 × 12. This is too large for model testing, so we next reduce the dimensionality.

b. *Data compression for the observations*

Principal component analysis was used to reduce the dimension of the observations in an effective way, while filtering out some of the data noise. The first four principal components have eigenvalues that account

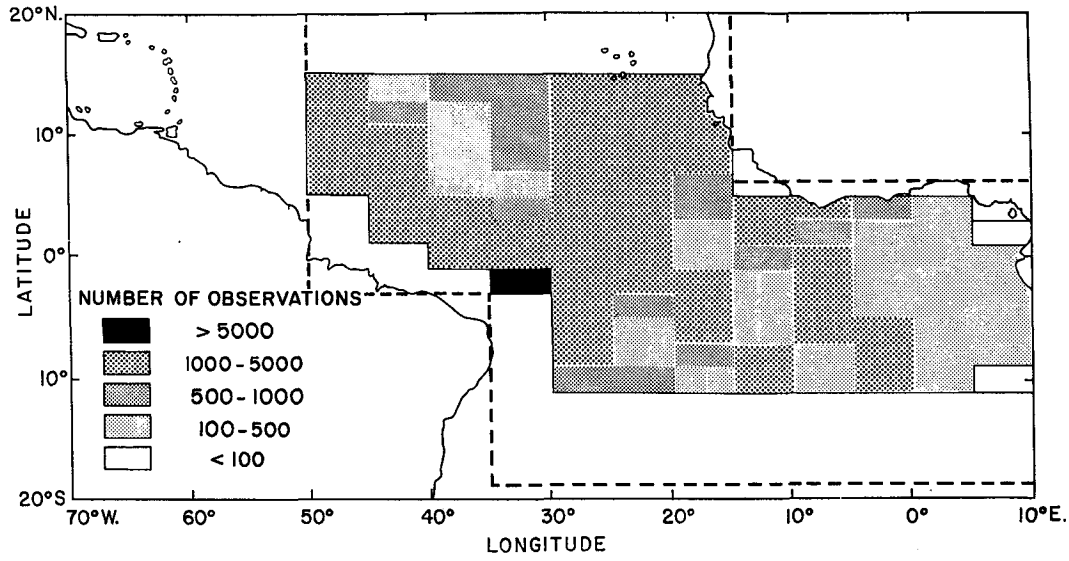

FIG. 1. Distribution in 2° × 5° boxes of the ship-drift observations in the region of interest to this study. The dashed line represents the boundaries of the linear model.

for a total of 86% of the field variance, and they correspond to coherent space–time structures, as illustrated in Fig. 3 and 4. The first two principal components describe well-defined large scale patterns, with a primarily annual or semiannual period. They are very similar to those computed for a broader region by Richardson and Walsh (1986, Fig. 9). The pattern of EOF 3 and 4 are a bit noisy and have a broader range of time scales, but they still represent coherent structures. On the other hand, the higher EOFs and principal components (not shown) are much more noisy and do not seem to correspond to any significant pattern of seasonal variability. Thus, they are presumed to be data noise, and are not included in the analysis, which is performed in the subspace of the four EOFs in Fig. 3. Examination of the eigenvalues shows that this truncation is not in the middle of a multiplet, and is thus acceptable.

In the truncated EOF subspace, the seasonal signal is given by the four principal component time series in Fig. 4 (error bars are discussed below). They are not normalized, so that the larger surface current changes have more weight in the model-data intercomparisons. Reconstruction on the original grid (not shown) confirms that the seasonal currents have not been noticeably altered by the EOF filtering, as expected from the large percentage of variance retained. A large data compression has thus been achieved without significant loss of information.

Finally, a further reduction in dimensionality is performed in the time domain, by considering the yearly cycle on the seasonal scale. Weighted averages, ¼–½–¼, centered on February, May, August and November are used as in Fig. 2, to avoid too much smoothing. This defines the matrix **A** in (4).

c. *Error analysis*

We need a quantitative estimate $\mathbf{D}^{tq}$ of the observation lagged error covariance matrix, so that the error covariance matrix of the reduced data, say **S**, can be calculated from (5). The construction of a realistic error model for the observations is a crucial, but time-consuming, step in the model testing procedure. Details of our error model for the surface currents are presented in appendix A. It accounts for measurement and sampling errors, as well as the dominant space and time scales of the current fluctuations, but neglects possible biases and the small contribution of errors in the mean. The error covariance matrix $\mathbf{D}^{tq}$ is calculated from the sample covariances $\overline{u'^2}$, $\overline{v'^2}$ and $\overline{u'v'}$ provided by Richardson and Walsh (1986), and it depends on an estimate of the effective number of independent observations. Though $\mathbf{D}^{tq}$ is calculated from sample estimates, **S** is considered hereafter as the true error covariance matrix of the reduced data vector. While this is not statistically rigorous, the high degree of data compression applied to obtain **S** suggests that it is a satisfactory approximation: the uncertain details of $\mathbf{D}^{tq}$ have been eliminated.

To illustrate the observational uncertainties, the 95% confidence intervals for the principal components time series of the four EOFs are given in Fig. 4. They are derived from the diagonal elements of **S**, under the assumption of normality, and should be understood as defining a probability region for the time evolution of each of the patterns considered separately (thus they are not appropriate to making simultaneous inferences on the signal). The first two principal components (primarly the annual and semiannual signals) clearly stand above the noise, whereas the other two are not

always distinguishable from zero. For the higher EOFs not retained in the analysis, the principal components become increasingly noisy (not shown), consistent with our assumption that they are dominated by data noise.

4. The tropical Atlantic wind stress

To drive the ocean model, we use a new wind stress product derived from the historical ship reports (Duchene 1989). The data are constructed from monthly averaged fields of pseudo wind stress (the surface wind vector $\mathbf{u}^a$ multiplied by its magnitude) on a $2° \times 5°$ grid, kindly provided by J. Servain for the period 1964–84. As described by Servain et al. (1985), the monthly averages were constructed from the individual ship reports, after application of stringent quality control tests. The data density was generally sufficient to estimate mean monthly values, but there were regions with gaps or very sparse data. Servain et al. then used the objective successive correction method to produce regular fields at 2°-resolution. As this procedure is not tailored to fields as inhomogeneous as those encountered in the tropics, it may introduce spurious scales of variability, and eliminate real ones.

Duchene (1989) has designed a new procedure based on a principal component analysis of the original gappy data, thus preserving the correlation scales inherent in the data. Filtering is via truncation to a principal component subset that can be distinguished from white noise, as determined by the method of Preisendorfer et al. (1981). The resulting pseudostress field is rather smooth, and is believed to contain no systematic errors other than those due to the measurements themselves, and to the limited spatial resolution, which certainly affects the wind stress curl. A comparison with independent measurements made during the FOCAL/SEQUAL experiment in 1983–84 shows that the new product is more accurate than that of Servain.

From the model response to a 21-year dataset, sample estimates of its mean seasonal cycle and error covariance matrix can be computed. The uncertainties in the model seasonal response that are associated with the interannual variability of the atmospheric forcing are then taken into account explicitly, together with those due to random errors and persistent gaps in the wind data. Indeed, unless the pseudostress uncertainties introduce systematic long term biases, they contribute to the variance in the sample in the same way that the interannual variability does. On the other hand, we neglect the influence of the nonsimultaneity of the wind with the surface current data, which were mainly collected between 1920 and 1940, in the belief that the interdecadal differences are much smaller than the shorter term interannual variability we do include.

To calculate the wind stress from the pseudostress, we follow the approach used by Cardone et al. (unpublished manuscript) to construct the FOCAL/SEQUAL wind product. The drag coefficient C_D is given by Large and Pond (1981), and stability effects are taken into account by assuming mean tropical Atlantic conditions, namely an air minus sea temperature dif-

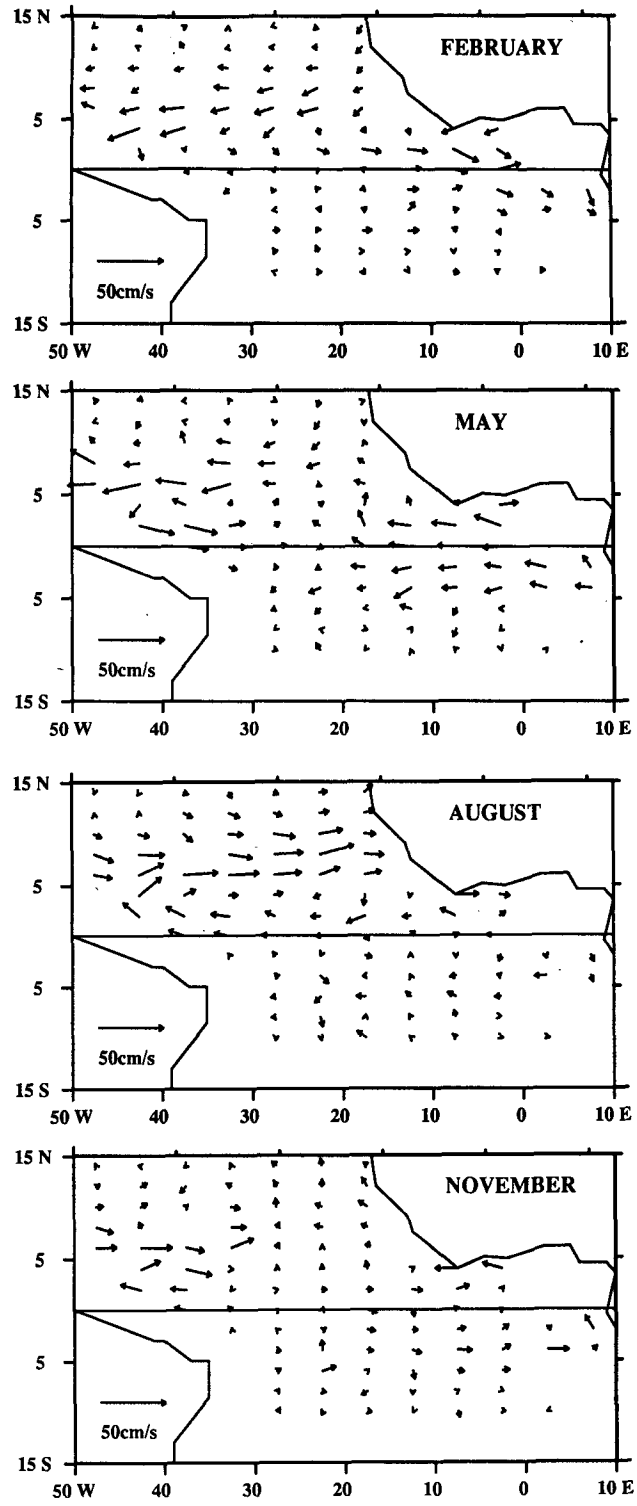

FIG. 2. Left: observed seasonal surface current fluctuations for February, May, August and November; the annual mean has been removed, and weighted averages used (see text).

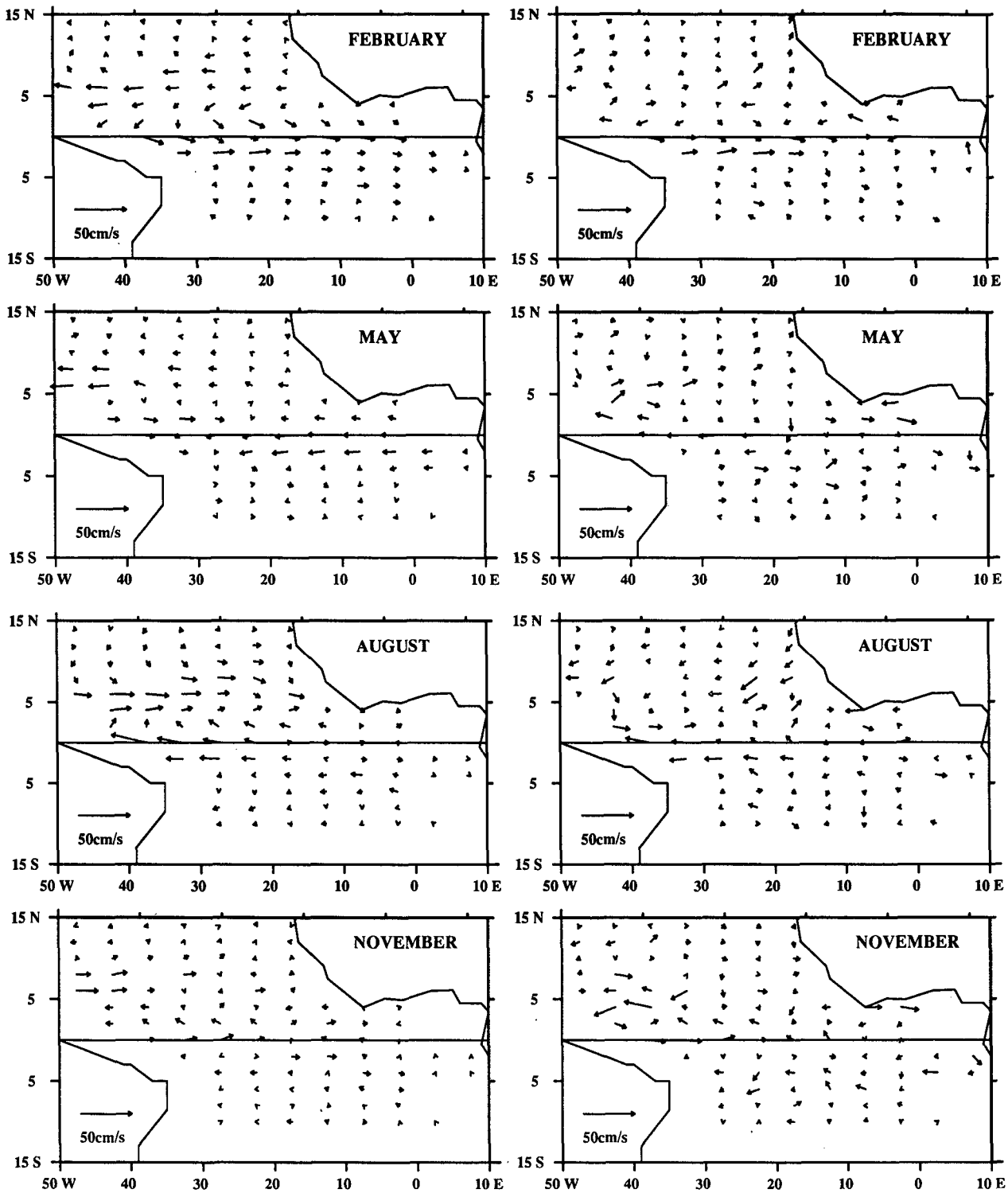

FIG. 2. (*Continued*) Middle: corresponding model currents for the standard case. Right: corresponding differences between modeled and EOF-filtered observed currents.

ference of −1°C, and a relative humidity of 80%. This increases the wind stress at low wind speed.

Although this drag formulation is very appropriate to the tropical Atlantic, its accuracy remains questionable, as is the case for all bulk formulations. Blanc (1985) has studied the uncertainty associated with dif-

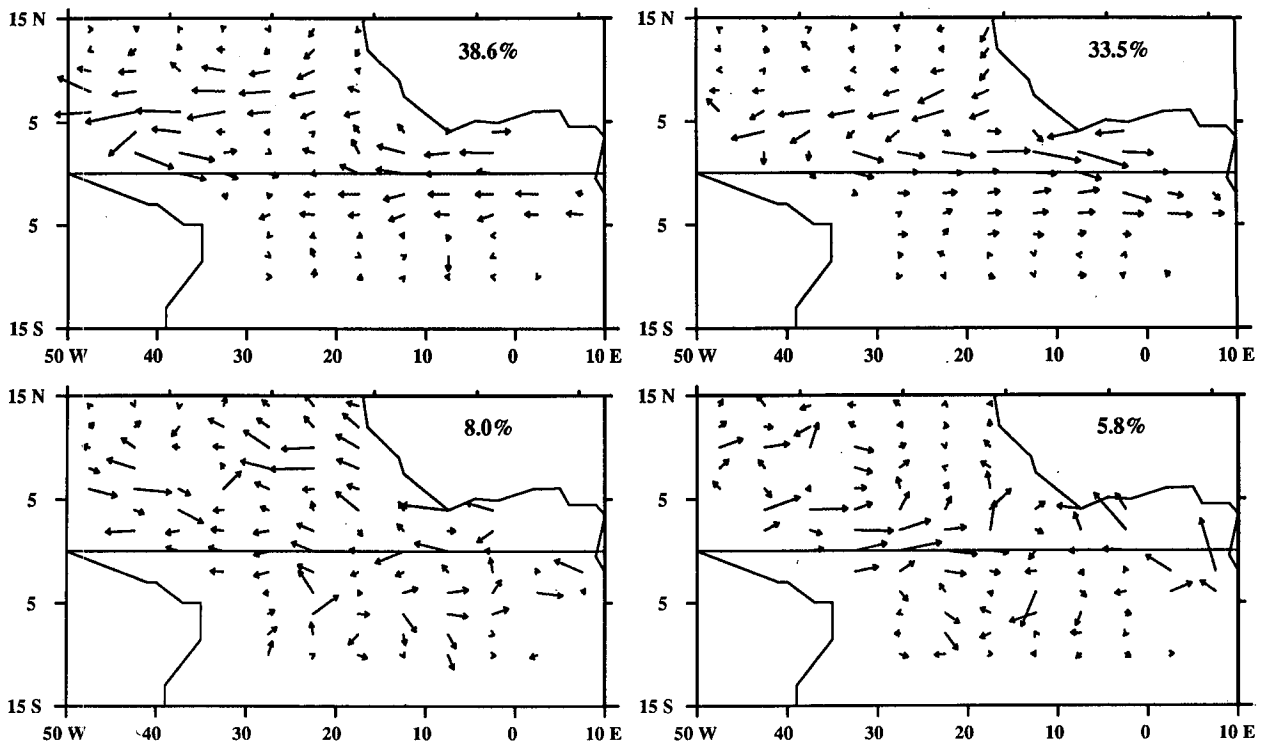

FIG. 3. First four orthonormal EOFs of the observed seasonal surface currents, which account for 38.6%, 33.5%, 8% and 5.8% of the total variance, respectively.

ferent widely used drag formulations: in the wind stress range encountered here, the average uncertainty (rms) due to scheme variation is about 20% of the stress value. In the following, we shall use this figure as a measure of the uncertainty in the bulk formulation, assuming blithely that the adopted law is the "true" one. The uncertainty in the drag formulation will be modeled by assuming

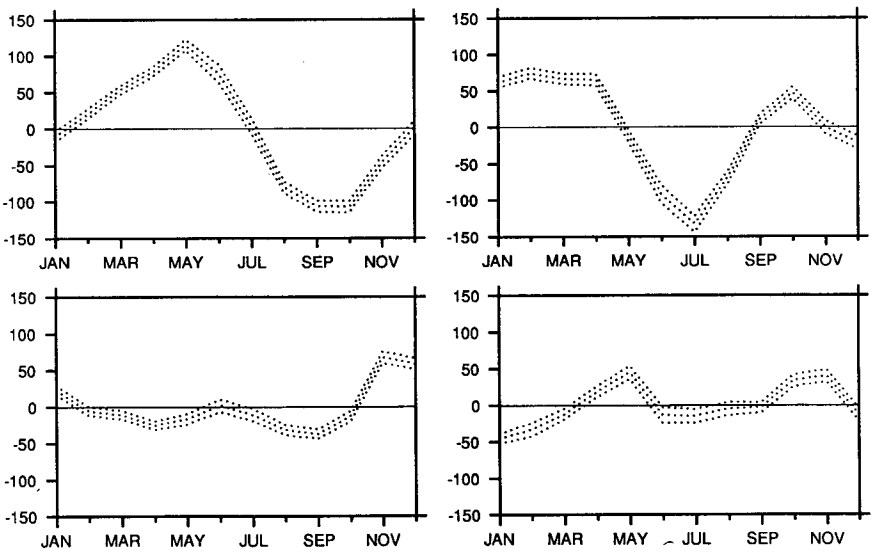

FIG. 4. Principal component time series of the observed seasonal surface currents in the EOF subspace of the observations (relative units). The error bars are estimates of the 95% confidence intervals for each value considered separately.

$$\tilde{C}_D = C_D(1 + c), \tag{9}$$

where C_D is the drag coefficient calculated as indicated above, and c a normal variate with zero mean and standard deviation $\sigma = 0.2$. Since the ocean model is linear, its mean seasonal response is not affected by the drag uncertainty, but its error covariance matrix is increased (cf. section 5). In a nonlinear model, the effects of the drag uncertainty could be estimated from Monte Carlo simulations made with various drag coefficients derived from (9).

Here, we shall also perform Monte Carlo simulations where a 20% random uncertainty is added to the drag coefficient at each time step and grid point. This crudely represents wind measurement errors, aliaising, mesoscale motions, high frequencies, as well as space–time changes in the drag coefficient.

In summary, we consider the uncertainties in the mean seasonal response of the equatorial model that are due to the limited wind stress sample and its interannual variability, and to nonsystematic wind errors and data gap effects. We also model simply the uncertainties in the drag formulation. Systematic errors remain that we do not consider due for example to windage, to the long term variability in the wind field, or to the lack of spatial resolution. It is assumed that these errors are smaller than those considered here.

5. Model testing

a. *Model description*

We consider the linear model of Cane and collaborators (e.g., Cane 1984; Du Penhoat and Tréguier 1985); it is an equatorial β-plane model with M vertical modes, and a surface mixed layer of fixed depth H. The multimode model is well documented, and its extension to represent surface currents given for the $M = 1$ version in Zebiak and Cane (1987); the extension to the multimode case is presented in Blumenthal and Cane (1989). In the course of this study, M and H will be varied, but in this section we only consider a standard version with $M = 2$, $H = 35$ m.

The model, which cannot handle all the irregularity of the Atlantic basin, covers the region limited by a dashed line in Fig. 1. As it neglects the short, eastward propagating Rossby waves which are important at the western boundary, but propagate too slowly to penetrate into the interior, it is inappropriate for simulation of currents near the western boundary. In the following, we limit the testing of the model to the region delineated by the continuous line in Fig. 1; it is in this region that the model surface currents will be compared to the observations.

b. *The standard case*

The numerical model is first spun up to a seasonally varying equilibrium state, which is reached within a 10-year integration. It is then forced by the 21-year wind stress dataset, using the reference drag coefficient C_D. To eliminate the effects of the unknown initial conditions, the first year of the 21 is ignored. The response time is short in the near equatorial zone of interest, so this delay is sufficient for the model adjustment. The monthly surface current for month t of year i is given on the model validation grid defined in section 3 by the $2 \times N$ dimensional vector $\mathbf{m}^t(i)$, $i = 1, 20$. Its sample mean over the 20 years is our estimate of the model seasonal response.

Figure 2 (middle) shows the model seasonal cycle, after subtraction of the annual mean. As is typically the case with visual displays, some features compare well with the observations (left), while others do not. The differences between modeled and observed currents (right) emphasize that the model seems unable to reproduce the amplitude of the currents south of the equator, and in the region of the North Equatorial Countercurrent, in particular during summer. All the discrepancies cannot be attributed to the oversimplification of the model, however, in view of the data uncertainties discussed above.

To distinguish between model and data errors, we apply the testing procedure of section 2. Only about half of the variance of the model seasonal response projects onto the truncated EOF base defined from the observations in section 3b. Thus, as anticipated, the validation subspace has to be extended. This is done by estimating the EOFs of the part of the mean model response that does not project on the four EOFs in Fig. 3, and keeping those with eigenvalues at least as large as the smallest one retained in the observations. Only two additional EOFs, shown in Fig. 5, are needed, which brings the spatial dimension to 6. Note the similarity in pattern between Fig. 2 (right) and Fig. 5. Now 90% of the model and 86% of the observed variance of the mean seasonal changes, respectively, project onto the truncated base.

Using weighted time averages as before, and reordering, we represent the model response at year i by the reduced signal $\mathbf{p}(i)$, $i = 1, 20$. The vectors have dimension $6 \times 4 = 24$. In the notation of section 2, the mean model response is estimated by

$$\bar{\mathbf{p}} = \frac{1}{20} \sum_{i=1}^{20} \mathbf{p}(i), \tag{10}$$

and its sample covariance matrix by

$$\mathbf{P} = \frac{1}{19} \sum_{i=1}^{20} (\mathbf{p}(i) - \bar{\mathbf{p}})(\mathbf{p}(i) - \bar{\mathbf{p}})'. \tag{11}$$

Monthly anomalies at yearly intervals are basically uncorrelated, so we assume that the data are independent and estimate the error covariance matrix of $\bar{\mathbf{p}}$ by $\mathbf{P}/20$, with 19 degrees of freedom.

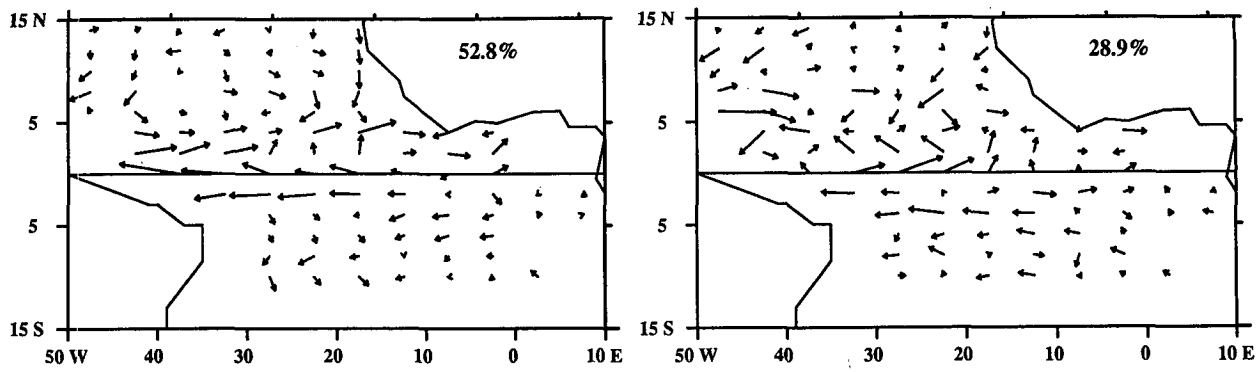

FIG. 5. First two EOFs of the residual currents (see text) which account for 52.8% and 28.9% of their variance, respectively.

The observed and the model seasonal variations are represented at montly intervals in the expanded EOF space in Fig. 6, together with univariate 95% confidence intervals estimated as in Fig. 4. This representation provides a much more synthetic evaluation of the model performance than the comparisons in the original space, and the error bars give some insight into the extent of the data uncertainties. The main features of the first two principal component times series, which correspond to the least noisy EOFs (Fig. 3), are fairly well reproduced by the linear model: the phase seems correct, but the amplitude of the model response is weaker than in the observations. On the other hand, the agreement for the third and fourth principal components, which correspond to more noisy EOFs, is very poor. The model response in the model EOF sub-

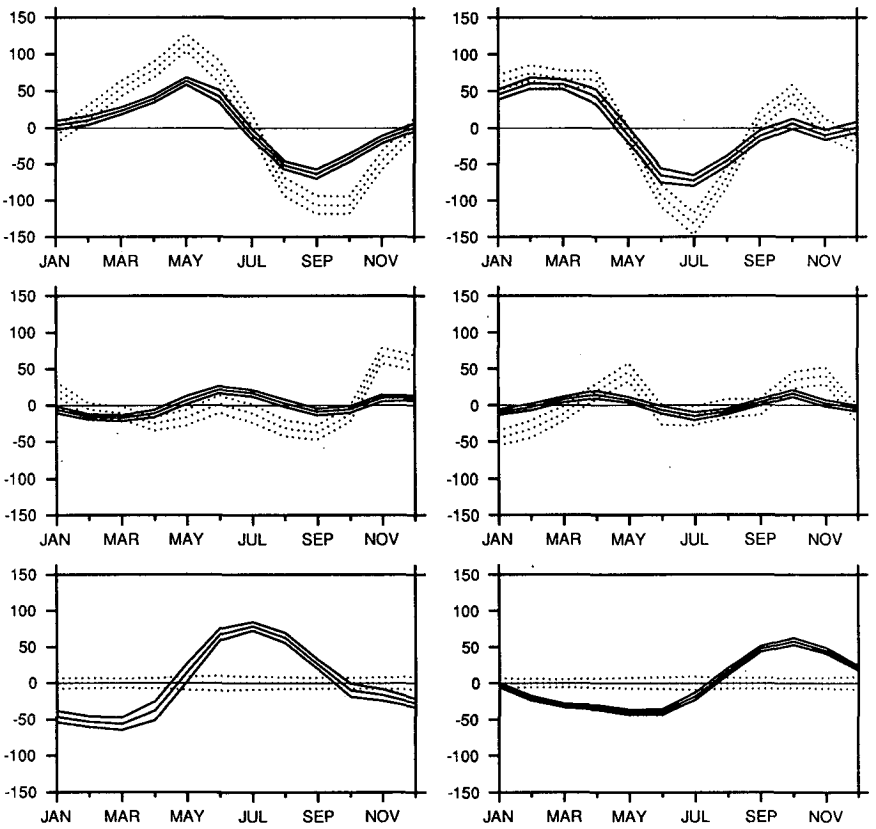

FIG. 6. Principal component time series of the observed (dotted line) and modeled (continuous line) seasonal surface currents in the extended EOF subspace. The error bars are estimates as in Fig. 4.

spaces clearly stands above the noise. Even though the error bars should not be used to make inferences on how close the true values are, as discussed earlier, Fig. 6 suggests that the data uncertainties will be unable to explain the discrepancies between model and observations.

More quantitative information is provided by the multivariate test of the null hypothesis that the model response is consistent with the oceanic observations (section 2). As mentioned before, we assume that we know the true error covariance matrix of the observations, but we estimate the model response uncertainties from the sample. An appropriate version of the test statistic (7) and (8) is then

$$T^2 = (\bar{\mathbf{s}} - \bar{\mathbf{p}})'(\mathbf{S} + \mathbf{P}/20)^{-1}(\bar{\mathbf{s}} - \bar{\mathbf{p}}), \quad (12)$$

where $\mathbf{P}$ is given by (11) and $\mathbf{S}$ by (5). Following Yao (1965), we approximate the distribution of (12), if the null hypothesis is true, by Hotelling's T^2, with 24 and ν degrees of freedom, where ν is estimated by (e.g., Seber 1984)

$$\nu^{-1} = (20^3 - 20^2)^{-1} \times [(\bar{\mathbf{s}} - \bar{\mathbf{p}})'\mathbf{S}_T^{-1}\mathbf{P}\mathbf{S}_T^{-1}(\bar{\mathbf{s}} - \bar{\mathbf{p}})/T^2]^2 \quad (13)$$

with $\mathbf{S}_T = \mathbf{S} + \mathbf{P}/20$. The Hotelling distribution is related to the F distribution with 24 and $\nu - 23$ degrees of freedom by (Morrison 1976)

$$F = \frac{\nu - 23}{24\nu} T^2, \quad (14)$$

For the standard run, the value of T^2 is 1732, and the degrees of freedom are 953. For such large ν, T^2 behaves approximately as a chi-square variable, and the critical value is 36, at the 5% level of significance. The null hypothesis is rejected, which confirms that the data uncertainties are insufficient to explain the discrepancies between observed and model surface currents. The uncertainty in the drag formulation has not been included, however, so we next take it into account.

c. *Drag coefficient uncertainty*

First, we consider the effects of purely random changes in $\tilde{C}_D$ on the model seasonal response, by adding a 20% random uncertainty in the forcing at each time step and grid point, as described in section 4. The results of Monte Carlo experiments (not shown) indicate that there is no noticable effects on the model seasonal response, nor on its error covariance matrix. The lack of sensitivity to uncorrelated noise in the forcing can be understood easily, considering that the low frequency, large scale ocean dynamics investigated here effectively integrates over many of these uncorrelated input.

On the other hand, coherent changes in $\tilde{C}_D$ have an influence on the model response. Indeed, in the presence of a drag uncertainty of the form (9), and given that the model is linear, the model response at year i can be written $(1 + c)\mathbf{p}(i)$, where $\mathbf{p}(i)$ is the response for the reference drag coefficient C_D. The mean seasonal response should now be considered to be both a sample mean over time and an ensemble mean over c, the latter being denoted by angle brackets. As the two processes are statistically independent and c has zero mean, the mean of $\bar{\mathbf{p}}$ does not vary, i.e.

$$\langle (1 + c)\bar{\mathbf{p}} \rangle = \bar{\mathbf{p}}, \quad (15)$$

but its error covariance matrix does. Indeed, its true value is given by

$$\begin{aligned} \mathrm{cov}[(1 + c)\bar{\mathbf{p}}] &= \langle (1 + c)\bar{\mathbf{p}}(1 + c)\bar{\mathbf{p}}' \rangle \\ &\quad - \langle (1 + c)\bar{\mathbf{p}} \rangle \langle (1 + c)\bar{\mathbf{p}}' \rangle \\ &= \langle (1 + c)^2 \rangle \langle \bar{\mathbf{p}}\bar{\mathbf{p}}' \rangle \\ &\quad - \langle 1 + c \rangle^2 \langle \bar{\mathbf{p}} \rangle \langle \bar{\mathbf{p}}' \rangle \\ &= (1 + \sigma^2)\,\mathrm{cov}[\bar{\mathbf{p}}] + \sigma^2 \bar{\mathbf{p}}\bar{\mathbf{p}}', \end{aligned} \quad (16)$$

and it can be estimated by

$$(1 + \sigma^2)\mathbf{P}/20 + \sigma^2 \bar{\mathbf{p}}\bar{\mathbf{p}}'. \quad (17)$$

For $\sigma^2 = 0.04$, the main effect of the drag uncertainty is to increase the model response variance when the signal $\bar{\mathbf{p}}$ is large; when $\bar{\mathbf{p}}$ is small, the changes are negligible.

Note that the random variable $(1 + c)\mathbf{p}(i)$ is not multinormal, being the product of two normal variables. However, by the central limit theorem, one can still assume that $(1 + c)\bar{\mathbf{p}}$ is approximately multinormal, if the sample is large enough, so that one can calculate approximate confidence intervals as before. The results in Fig.7 show that the increase in the error bars can be substantial when $\bar{\mathbf{p}}$ is large, for the assumed 20% uncertainty in $\tilde{C}_D$. However, the increase in model response uncertainty seems insufficient to explain the discrepancies with the observations. This is confirmed by the multivariate test, where T^2 only reduces to 1374, with $\nu = 1997$. The relative insensitivity of the test reflects the very high correlation of the model response changes due to coherent changes in $\tilde{C}_D$. Note that in this case the T^2 statistic should be used with caution, as (12) is distributed as Hotelling's T^2 only if the covariance matrix has the Wishart distribution (e.g., Morrison 1976). This only holds approximately at best when (17) is used.

To verify our model of the wind stress uncertainties, we have forced the numerical model with the mean seasonal wind stress data of Hellerman and Rosenstein (1983), which were constructed with the drag coefficient of Bunker (1976) that may yield slightly overestimated values. Figure 7 (solid line) shows the model response in the EOF subspace of the observations. Although the seasonal response is much larger than with our wind stress, it stays mostly within the estimated

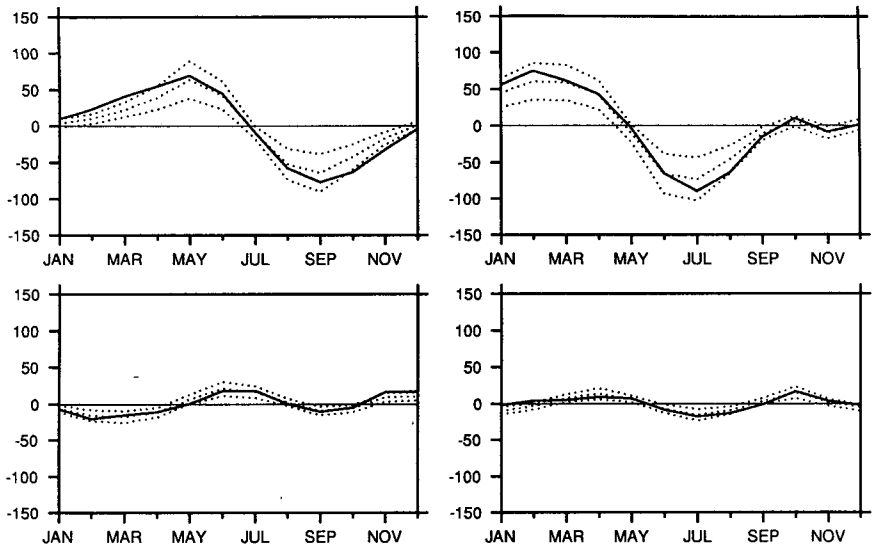

FIG. 7. Principal component time series of the model seasonal surface currents with the effect of drag coefficient uncertainty, in the EOF subspace of the observations (dotted line). The heavy line represents the model response to the climatological forcing of Hellerman and Rosenstein (1983).

95% confidence intervals corresponding to the combined effect of interannual variability and drag uncertainty. This suggests that our wind stress uncertainty model is basically correct.

In summary, the test objectively shows that the uncertainties in the atmospheric forcing and the oceanic observations do not account for the model-reality discrepancies. These can then be attributed unambiguously to deficiencies in the equatorial ocean model or, at least, in the standard version used in this section.

6. Model tuning

Even in the most realistic cases, model development involves some arbitrary choices. Consider for example the choice of an optimal model resolution. Physical analysis often suggests higher resolution than is really needed, as many model parameters are poorly known, and various simplifications may have been made that limit the model accuracy. If sufficient computational resources are available, the ultimate choice should be based on how well the model performs in representing the oceanic observations. For this purpose, the method developed in the present paper is particularly useful, in particular since data uncertainties are often too large to allow for an easy rating between different model versions.

The T^2 statistic is a convenient measure of model-reality agreement, even if the null hypothesis is rejected, and it shows how closely the observations can be reproduced. It is well adapted to model tuning, and its minimum will show when an optimal fit is achieved. It can also be used to evaluate the statistical significance of the improvement in model performances.

The method is illustrated by tuning the vertical resolution of the linear multimode ocean model to the surface current data. Only two model parameters are varied here: the number of vertical modes M and the thickness of the mixed-layer H.

The number of vertical modes that should be included for best model performances has not been established. Linear theory suggests that the model should be more realistic with many modes. However, in the presence of mean shear currents and nonlinearities, the very notion of decoupled vertical eigenmodes becomes more and more questionable as the number of modes increases. Previous choices in simulations of observations vary between $M = 1$ (e.g., Zebiak and Cane 1987) and up to $M = 9$ (Du Penhoat and Tréguier 1985), while more theoretical studies have included as many as 60 (McCreary 1981). Work with sea level data suggests $M \geqslant 2$ (Cane 1984), but the optimal value is difficult to establish without an objective measure of model-data agreement.

In Fig. 8, we have calculated the changes in T^2 as the number of vertical modes is varied between 1 and 5, for the standard 35 m mixed-layer depth; for comparison, we have also run the model without Ekman currents. In all cases, the spatial dimension is 6; the structure of the two additional EOFs is similar to that in Fig. 5, and a comparable percentage of the total variance of the observed and modeled currents is represented in the EOF subspace. Following the discussion in section 5c, the effect of the drag uncertainty are not included in this section, even though they would lower the T^2 values. The results show that our a priori choice of two modes is not the optimal one to represent the

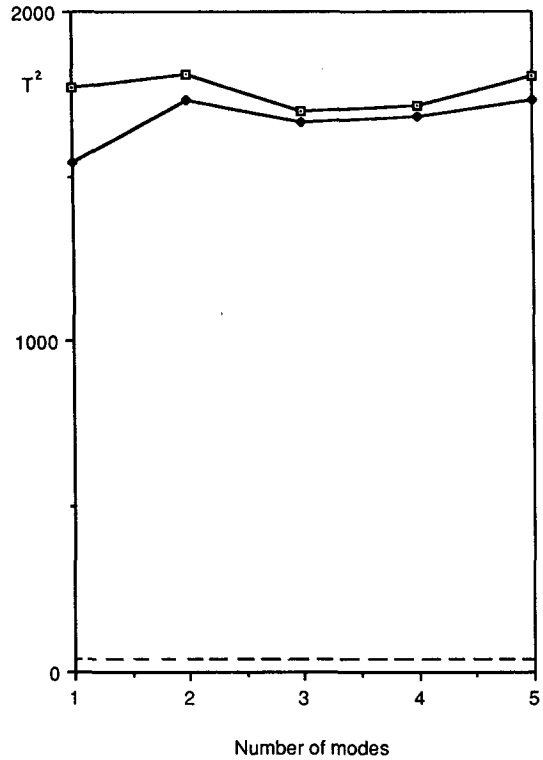

FIG. 8. Value of T^2 for various number of modes, for H = 35 m (full squares), and for the baroclinic current only version (open squares). The dashed line gives the approximate critical value for T^2, at the 5% level.

surface currents. A better fit is obtained with just one vertical mode; including the third mode brings some improvement to the 2-mode version, but not enough to reach the agreement provided by the simplest version. The importance of the Ekman currents is stressed by the large T^2 obtained with just the baroclinic currents.

The choice of the mixed layer depth H is also somewhat arbitrary, though 35 m is representative of the tropical Atlantic. Indeed, there are large zonal changes in the mixed layer depth, and there may be current shears in the mixed layer. Thus, a slightly shallower or deeper value may well be more appropriate for currents. Figure 9 shows the changes in model performance with the value of H, for M = 1 to 3, and for an Ekman current-only version (no baroclinic currents). As expected from theory (Blumenthal and Cane 1989), the optimal mixed-layer thickness decreases with the number of vertical modes, while the relative importance of the Ekman currents increases. The optimal solution is little altered, however, being reached for M = 1 and H = 40 m.

To determine whether the improvement with respect to the standard version (M = 2, H = 35 m) is statistically significant, consider T^2. As it is approximately distributed as a noncentral χ^2 variable, the 95% confidence interval in the standard case is about (1569, 1895), which alone encompasses most of the results in Figs. 8 and 9. Thus, the tuned version minimizes the (weighted) model–data discrepancies, but it seems only marginally superior to the standard version, and it is not significantly better than the three-mode one. The differences with the observations are just too large to distinguish unambiguously between the various versions of the model. It will be shown elsewhere (Duchene 1989) that more definite conclusions on the optimal model resolution can be reached from a comparison with the observed dynamic height, better reproduced by the multimode model.

7. Conclusions

Because of the increasing degree of realism of equatorial ocean models and the sparseness of reliable oceanic and atmospheric data in the tropics, there is a need for a model–data intercomparison method that can distinguish between model errors and data uncertainties. To tackle this problem, we have adopted a statistical approach, and have developed a model validation procedure that provides an objective measure of the fit between model simulations and observations, taking into account the data uncertainties. The method is based on multivariate statistical analysis, and it relies on a detailed error analysis of both the oceanic observations and the atmospheric input. The model–data intercomparison is done in the whole space where the model is valid and there are sufficient observations, but the statistical testing is conducted after reduction to a subspace of much smaller dimension. The data reduction used principal component analysis, but other data compression choices may be as efficient. The method has computational advantages when compared with the nonlinear programming approach (Schroeter and Wunsch 1986), because the simplicity of the calculations and the absence of program developments

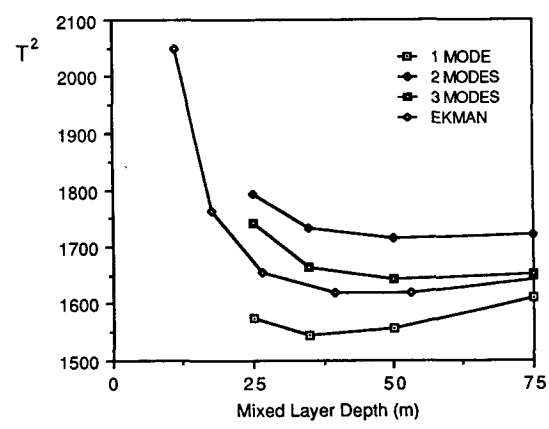

FIG. 9. As in Fig. 8, but for T^2 as a function of the mixed-layer thickness H; Ekman corresponds to the Ekman current only version.

more than compensate for the increase in computer time needed to simulate the forcing uncertainties.

Although the model validation method is general and applicable to any ocean model, its implementation must be adapted to the case at hand. Here we have considered the testing of an equatorial ocean model, where most of the variability is due to the atmospheric forcing, and little to the internal ocean dynamics. We have considered the case of seasonal variations, which does not depend on the initial conditions. Cases influenced by initial conditions will be discussed in a forthcoming paper.

The procedure has been applied to validation of the linear, multimode model of Cane and Collaborators, using the surface currents in the tropical Atlantic estimated from ship drifts by Richardson and Walsh (1986) as an observational basis for the model testing. The errors in the observed surface currents have been modeled, and the uncertainties in the ocean model response due to random errors in the wind stress and its interannual variability have been considered. Although these observational inadequacies create a substantial indeterminacy in the available samples, they are unable to account for the large discrepancies that were found between observed surface currents and model simulations.

Uncorrelated random forcing errors were also considered in Monte Carlo simulations, but they have no influence on the large scale response of the ocean model. This result is expected in a linear model, and we anticipate that it will hold even in the most complex model, because nonlinearities will introduce only a small rectified effect at large scale from small scale, high frequency forcing. On the other hand, the uncertainty in the bulk formulation for the wind stress contributes significantly to the uncertainty in the ocean model response. Still, it is insufficient to explain the model–data differences, so that they must be attributed to the many oversimplifications in the ocean model physics. In subsequent studies, more realistic numerical models will be evaluated, and additional oceanic data will be used for the validation.

In the method, the overall fit between model and observed data is quantified by a statistic that provides a measure of their differences, weighted by the data uncertainties. This statistic is very handy for model tuning, or model intercomparisons. It has been used here to determine the number of vertical modes and the mixed layer depth that provide an optimal fit to the observed surface currents. The model works best with only one vertical mode and a mixed layer depth of 40 m, though the discrepancies remain too large for the improvement to be statistically significant.

The result that the model cannot be remedied by tuning may surprise those readers who believe that models, especially simple models, are limitlessly adjustable. Having established beyond doubt that the model will need more than a parameter change, it is worthwhile to consider the reasons for its failure before beginning the difficult process of model improvement. First, we note that the ship-drift currents and the integrated surface-layer current specified by the model need not be the same thing. Windage effects aside, the ships sample only the top few meters of the ocean while the model layer is 40 m or so in depth. The bulk mixed layer assumption that the current is well mixed throughout the surface layer may be too far from reality for present purposes. Alternatively, the well-mixed assumption may be adequate while the assumption of a constant mixed layer is not.

Perhaps the failure in the North Equatorial Countercurrent stems in part from the smoothing of the wind in the analyzed product and the consequent underestimate of the wind stress curl. However, there may not be enough wind data in the historical ship reports to detect meridional variations on scale finer than the 2 degree resolution of the product used in this study. Failure near the equator is most likely due to the omission of nonlinearities, especially the upwelling of eastward momentum from the undercurrent. However, this does not account for the discrepancies being a maximum just south of the equator. The elaborations of the model suggested here may be investigated and then systematically evaluated using the testing procedure developed here.

There is nothing in the model validation procedure which restricts it to simple models. It is easy to implement, the principal labor being the unavoidable task of establishing the magnitude and structure of the observational errors. This attention to observational errors directs the modelers to areas where the data is adequate to test model changes. No such procedure can replace looking at the data and understanding the causes of the model–data discrepancies, as essential steps in diagnosing model shortcomings. What the model validation does provide, however, is an objective framework for evaluating model changes, an essential step in model improvement.

Acknowledgments. We would like to thank Phil Richardson for providing the ship drift data, and for useful discussions, and Jacques Servain for providing wind stress data. Some of the data were processed by David Walsh, Christine Wooding and Terry McKee, who are gratefully acknowledged. Comments by Gilles Reverdin and the reviewers are also acknowledged. This research was supported in France by a grant from the P.N.E.D.C. and in the U.S.A. by NOAA Grant NA86AA-D-AC108 with the Massachusetts Institute of Technology, and NSF Grant OCE-86-08386 and NOAA Grant NA-87-AA-D-AC081 with Columbia University.

APPENDIX A

An Error Model for the Surface Currents

Since the ship drift data primarly comes from the period 1920–40, the long-term monthly means have

uncertainties reflecting the interannual variability of the currents, as well as the measurement and aliasing errors (possible biases are assumed to be small). If these errors could be considered as statistically independent, the error $\delta \mathbf{d}^t$ on the monthly means would simply be described by a normal distribution with variance given by the data variance divided by the number of observations, and there would be no error covariance between different points and times. However, the individual observations are not independent. Thus, we define an effective number of independent observations, and estimate the dominant space and time scales of $\delta \mathbf{d}^t$.

To get quantitative information on these scales and the effective number of independent observations, we use data time series constructed by Richardson and Walsh (in preparation) for the 12 sufficiently large boxes shown in Fig. A1 to document the interannual variability of the surface currents during the 23-year period 1919–41. The data grouping has been chosen to insure reasonable homogeneity within each box. The monthly values are well-sampled, except in June, July and December 1919–34 when there is no data in the Northern Hemisphere. After linear interpolation of the missing values and removal of the seasonal signal, the lagged correlation matrix was calculated for the 12 anomaly time series, and the results used as follows.

Let us first consider the effective number of independent observations, n_e, for a given box and month. When the observations are very sparse, they can be considered as independent. They are averages over paths comparable to the box size, so a ship only samples once, and it is unlikely that several observations were made during the same year, in view of the relative homogeneity of the data density over the period (see Fig. 2 in Richardson and McKee 1984). When the data density is large, however, measurements may have been made at close intervals by different ships, and the n observations are not independent. To relate n_e to n, consider the long term mean value of, say a zonal velocity in January, $\bar{u}$. Its variance can be written as the sum of the variance due to the interannual variability, a, and the variance due to errors and aliasing, that we assume for simplicity to be given by b/n, with b constant:

$$\overline{u'^2}/n = a + b/n. \tag{A1}$$

As argued above, for small n, $n_e \approx n$, which yields $b \approx \overline{u'^2}$. For very large n, n_e should tend to some constant maximum $n_m \approx \overline{u'^2}/a$. The simplest form for $n_e(n)$ consistent with the two limits is

$$n_e^{-1} = n_m^{-1} + n^{-1}. \tag{A2}$$

The parameter n_e can be estimated directly. Indeed, the sample variance of the long-term mean $\bar{u}$ is given by the variance of the successive *Januarys* $u(i)$, divided by the number of years, if one neglects for simplicity the very weak data correlation at yearly intervals. Using the variance $\overline{u'^2}$ when all *Januarys* are considered simultaneously (as in the 2° × 5° boxes), we get

$$\overline{u'^2}/n_e = \Big(\sum_{i=1}^{23} (u(i) - \bar{u})^2/22\Big)/23 \tag{A3}$$

where an unbiased estimate of the variance has been used. A scatter plot of n_e versus n for the 12 boxes and all months (except the 3 months with many missing data) is shown in Fig. A2. We found no systematic dependence on geography or time of year, hence only the combined data is considered. Although the dispersion is large, the data are roughly consistent with the form (A2). The optimal fit, obtained by least squares estimation, is for $n_m = 1081$. Relation (A2) will be applied to the error calculation in the 2° × 5° boxes, using $n_m = 500$ to coarsely take into account the smaller box dimensions (measurements are more likely to have been made at close intervals). Sensitivity studies show that the errors are insensitive to n_m, except when it becomes small (the rms error on the first principal component increases by 20% for $n_m = 100$).

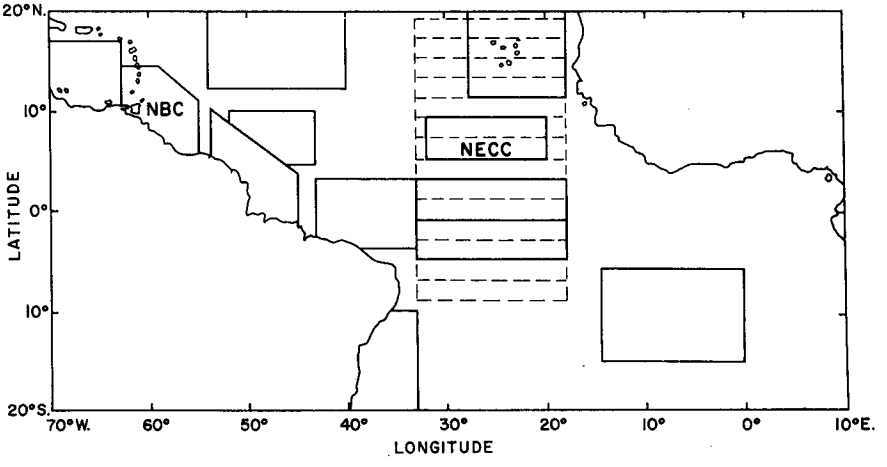

FIG. A1. Location of the "large boxes" used in the interannual variability study.

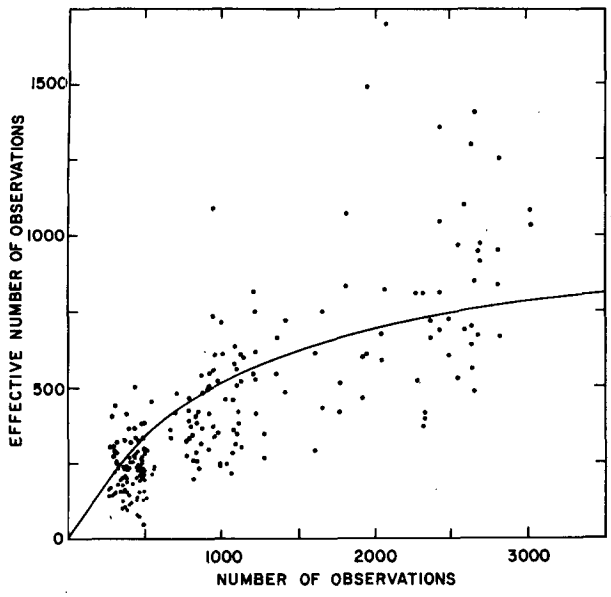

FIG. A2. Scatter plot of n_e versus n. The smooth curve is the best fit relation of the form (A2).

Let us estimate now the space and time scales of the errors $\delta\mathbf{u}^t$. For this purpose, we have considered the lagged correlations between the 12 large box time series. The results are noisy, but have simple dominant features. The averaged one-month autocorrelation is 0.21, but values as high as 0.42 are found; the average two-month autocorrelation is 0.09, with maximum at 0.27. The spatial correlations are of the same order, but show more scatter reflecting large-scale patterns. Studies with 2° × 15° boxes in the central region (dashed lines in Fig. A1) suggest furthermore that u and v in adjacent boxes are weakly correlated (correlation of about 0.3), but not cross-correlated, presumably due in part to the averaged nature of the individual observations.

On the basis of these results, we assume that the time dependence of the errors on the climatological currents in the 2° × 5° boxes can be modeled using an exponential decay-scale of 0.25 (to fit both one- and two-month lagged correlations), and set

$$D_{ij}^{tq} = 0.25^{|t-q|}(D_{ij}^{tt} + D_{ij}^{qq})/2. \qquad \text{(A4)}$$

To model their space dependence, we have to distinguish between the correlations due to small-scale fluctuations and the averaged nature of the individual observations (over path lengths comparable to the box size), and the correlations associated with the large-scale interannual variability. For the former, we assume as a crude rule of thumb an exponential decay-scale of 0.3 between u or v in adjacent boxes (note the implicit anisotropy), and no cross-correlations between u and v, consistent with the 2° × 15° results. For the latter correlations, we simply assume that the main patterns of interannual variability are the same as for the seasonal signal, i.e., that they can be described by the four EOFs in Fig. 3. A similarity between the main patterns of seasonal and interannual variability is commonly observed in large-scale near-surface data (e.g., Weare et al. 1976). It is not inconsistent with an EOF analysis of the large box data (not shown), though it does suggest that the interannual variance, which is as large as the annual, would have to be much more equally distributed among the patterns. We will therefore assume equal energy partition among the four seasonal EOFs for the interannual variability in the 2° × 5° boxes, and finally set

$$D_{ij}^{tt} = 0.5a(i,j)(D_{ii}^{tt} + D_{jj}^{tt})0.3^{d(i,j)} + \alpha \sum_{k=1}^{K}\sum_{l=1}^{K} E_{ik}E_{jl} \qquad \text{(A5)}$$

where $\mathbf{E}$ is the matrix of the first K eigenvectors introduced in section 2, α is the variance due to the interannual variability, $d(i,j)$ is the distance between the boxes in units of 2° and 5° in the meridional and zonal directions, respectively, and $a(i,j)$ is 1 if the velocity components are the same, and zero otherwise. For simplicity, $a(i,j)$ is also taken to be zero if $d(i,j)$ is larger than 2.5. To estimate α, we take the total seasonal variance in the first four EOFs divided by 20, the typical number of years with high sampling, since we model the error of the means. Experiments show little sensitivity to the parameters in (A4) and (A5), for reasonable choices.

One last correction is applied, for there should be a loss of error correlation between different boxes or times where the data density is low. Indeed, it is then very unlikely that the measurements were taken during the same month. This is done by multiplying D_{ij}^{tq} by the product $q(n_i^t) \times q(n_j^q)$, where q models the correlation loss due to poor sampling. Random subsampling experiments (courtesy of J. Walsh) at boxes NBC and NECC (see Fig. A1) suggest little change in the one-month lagged autocorrelation as long as the number of samples per month in the time series is larger than about 10, and a progressive decrease when it becomes less. Thus, we simply set

$$q(n) = \min(1, n/200), \qquad \text{(A6)}$$

since there are approximately 20 years with high sampling. Error bars are sensitive to this last correction, particularly for the first EOF: its omission would increase the rms error on its amplitude by as much as 30%. The effect on the higher principal components is smaller.

With all these simplifying assumptions, the error covariance matrix D^{tq} may be modeled from Richardson's original data $\overline{u'^2}$, $\overline{v'^2}$, $\overline{u'v'}$ and n. Note that, to reduce the algebra, we have neglected the influence of the annual mean in the calculation of $\mathbf{D}^{tq}$. Using the error model, this can be shown to induce only a small

overestimation of $\mathbf{D}^{tq}$, 20% at most, which is clearly within the accuracy of our coarse estimates.

REFERENCES

Adamec, D., and R. L. Elsberry, 1984: Sensitivity of mixed layer predictions at ocean station Papa to atmospheric forcing parameters. *J. Phys. Oceanogr.*, **14,** 769–780.

Blanc, T. V., 1985: Variation of bulk-derived surface flux, stability, and roughness results due to the use of different transfer coefficient schemes. *J. Phys. Oceanogr.*, **15,** 650–669.

Blumenthal, M. B., and M. A. Cane, 1989: Accounting for parameter uncertainties in model verification: an illustration with tropical sea surface temperature. *J. Phys. Oceanogr.*, **19,** 815–830.

Bunker, A. J., 1976: Computation of surface energy flux and annual air–sea interaction cycles in the North Atlantic ocean. *Mon. Wea. Rev.*, **104,** 1122–1142.

Busalacchi, A. J., and J. J. O'Brien, 1981: Interannual variability of the equatorial Pacific in the 1960's. *J. Geophys. Res.*, **86,** 10 901–10 907.

Cane, M. A., 1984: Modeling sea level during El Niño. *J. Phys. Oceanogr.*, **14,** 1864–1874.

Duchêne, C., 1989: Test de la validité des modèles océaniques équatoriaux à l'aide d'observations. Thèse de doctorat, Université Paris 6.

Du Penhoat, Y., M. A. Cane and R. J. Patton, 1983: Reflections of low frequency equatorial waves on partial boundaries. *Hydrodynamics of the Equatorial Ocean* J. C. J. Nihoul, Ed., Elsevier, 237–258.

——, and A. M. Tréguier, 1985: The seasonal linear response of the tropical Atlantic Ocean. *J. Phys. Oceanogr.*, **15,** 316–329.

Goldenberg, S. B., and J. J. O'Brien, 1981: Time and space variability of tropical wind stress. *Mon. Wea. Rev.*, **109,** 1190–1207.

Hannoschöck, G., and C. Frankignoul, 1985: Multivariate statistical analysis of a sea surface temperature anomaly experiment with the GISS general circulation model 1. *J. Atmos. Sci.*, **42,** 1430–1450.

Hasselmann, K., 1979: On the signal-to-noise problem in atmospheric response studies. *Meteorology of the Tropical Ocean,* D. B. Shaw, Ed., Roy. Meteor. Soc., 251–259.

Hellerman, S., and M. Rosenstein, 1983: Normal monthly wind stress over the world ocean with error estimates. *J. Phys. Oceanogr.*, **13,** 1093–1104.

Large, W. G., and S. Pond, 1981: Open ocean momentum flux measurements in moderate to strong winds. *J. Phys. Oceanogr.*, **11,** 324–336.

McCreary, J. P., 1981: a linear stratified ocean model of the equatorial undercurrent. *Phil. Trans. Roy. Soc. London,* **A298,** 603–695.

Morrison, D. F., 1976: *Multivariate Statistical Methods.* McGraw-Hill, 415 pp.

Philander, S. G. H., and R. C. Pacanowski, 1986: A model of the seasonal cycle of the tropical Atlantic ocean. *J. Geophys. Res.*, **91,** 14 192–14 206.

Preisendorfer, R. W., and T. P. Barnett, 1983: Numerical model-reality intercomparison tests using small-sample statistics. *J. Atmos. Sci.*, **40,** 1884–1896.

——, F. W. Zwiers, and T. P. Barnett, 1981: Foundations of principal component selection rules. S.I.O. Ref. Ser. 81-4, Scripps Institution of Oceanography, 192 pp.

Richardson, P. L., 1983: Eddy kinetic energy in the North Atlantic from surface drifters. *J. Geophys. Res.*, **88,** 4355–4367.

——, and T. K. McKee, 1984: Average seasonal variation of the Atlantic equatorial currents from historical ship drift. *J. Phys. Oceanogr.*, **14,** 1226–1238.

——, and J. Walsh, 1986: Mapping climatological seasonal variations in surface currents in the tropical Atlantic using ship drifts. *J. Geophys. Res.*, **91,** 10 537–10 550.

Schroeter J., and C. Wunsch, 1986: Solution of nonlinear finite difference ocean models by optimization methods with sensitivity and observational strategy analysis. *J. Phys. Oceanogr.*, **16,** 1855–1874.

Seber, G. A. F., 1984: *Multivariate Observations.* J Wiley and Sons, 686 pp.

Servain, J., J. Picaut and A. J. Busalacchi, 1985: Interannual and seasonal variability of the tropical Atlantic ocean depicted by 16 years of sea surface temperature and wind stress. *Coupled Ocean-Atmosphere Models,* J. C. J. Nihoul, Ed., Elsevier, 211–237.

Storch, H. v., and H. A. Kruse, 1985: The extra-tropical atmospheric response to El Nino events—a multivariate significance analysis. *Tellus,* **37A,** 361–377.

Weare, B. C., A. R. Navato and R. E. Newell, 1976: Empirical orthogonal function analysis of Pacific ocean surface temperature. *J. Phys. Oceanogr.*, **6,** 671–678.

Wunsch, C., 1989: Tracer inverse problems. *Proceedings of the School on the Ocean Circulation and Geochemical tracer Transport,* Les Houches, France, Feb. 1988. D. L. T. Anderson and J. Willebrand, Eds., North-Holland.

Yao, Y., 1965: An approximate degrees of freedom solution to the multivariate Behrens-Fisher problem. *Biometrika,* **52,** 139–147.

Zebiak, S. E., and M. A. Cane, 1987: A model El Niño–Southern Oscillation. *Mon. Wea. Rev.*, **10,** 2262–2278.