

An intercomparison between four tropical ocean models

Thermocline variability

By CLAUDE FRANKIGNOUL*, SABINE FÉVRIER, NATHALIE SENNÉCHAEL, *Laboratoire d'Océanographie Dynamique et de Climatologie (LODYC), Unité mixte de recherche CNRS-ORSTOM-UPMC, Université Pierre et Marie Curie, 4 place Jussieu, 75005 Paris, France*, JACOB VERBEEK, *Koninklijk Nederlands Meteorologisch Instituut (KNMI), PO Box 201, 3730AE de Bilt, The Netherlands*, and PASCALE BRACONNOT, *Laboratoire de Modélisation du Climat et de l'Environnement (LMCE), Centre d'Etude de Saclay, 91191 Gif sur Yvette, France*

(Manuscript received 7 February 1994; in final form 19 October 1994)

ABSTRACT

A multivariate model testing procedure is used to intercompare several tropical ocean models, using the evolution of the thermocline depth during the 1982–1984 FOCAL/SEQUENTIAL experiment in the equatorial Atlantic as observational basis. Four models of increasing complexity are considered: Cane's linear multimode model, a nonlinear 2-layer model developed at LODYC, the KNMI GCM and the LODYC GCM. Some of the uncertainties in the atmospheric forcing are taken into account by forcing the models with 3 equally plausible windstress fields whose differences are consistent with the wind measurement and sampling errors, and the drag coefficient indeterminacy. Although the resulting uncertainties in model response are substantial, none of the models is fully consistent with the observations, i.e., within the error bars. The more complex models represent the thermocline depth variations significantly better than the simpler ones, in particular when the comparison is done over broad geographical areas. When the whole intercomparison domain (12°N–12°S) is considered, the LODYC GCM performs better than the KNMI GCM, both for the yearly mean conditions and for the variations around the 3-year mean, while the 2 models have comparable performance in the equatorial waveguide (3°N–3°S). In view of its simplicity, the linear model shows much skill, in particular for simulating the long-term mean. The 2-layer model poorly represents the long-term mean position of the thermocline, but represents well the variations around the 3-year mean, outperforming the linear model in the equatorial waveguide. Along meridional sections, model performance may be more variable, stressing the need for a global approach to model validation. Finally, a coarse investigation of the simulations of the thermocline depth anomalies with respect to the mean seasonal cycle suggests that, because of the small signal-to-noise ratio, most models have only little skill in the 12°N–12°S domain, but more in the equatorial waveguide. The linear model, however, does not seem to have predictive skill for these noisy anomalies.

1. Introduction

There is growing evidence that the El Niño–Southern Oscillation (ENSO) phenomenon, which is largely controlled by the interaction of the atmosphere with the tropical Pacific Ocean, is predictable more than a year in advance (Cane et al., 1986). Hence, coupled ocean-atmosphere models

are being developed for climate predictions. However, it has not been established what level of model complexity is required to predict ENSO and other climatic changes regulated by tropical air-sea interactions, like precipitation in Brazil and in the Sahel, which are influenced by sea surface temperature variations in the tropical Atlantic (Hastenrath and Heller, 1977).

A crucial step in coupled model development is the separate validation and tuning of the atmo-

* Corresponding author.

spheric and oceanic components, as this should minimize the need for flux corrections at the coupling stage. In view of the limited amount of observational data available, however, the validation of tropical ocean models is difficult. The problem does not lie so much in the sparsity of the oceanic observations, but in the large uncertainties of the atmospheric forcing. The tropical oceans are primarily wind-driven and could be simulated deterministically for the most part if the atmospheric forcing were accurately known. However, marine observations are inaccurate and concentrated along the main shipping lanes. Furthermore, the bulk formulae used to estimate the air-sea fluxes are major sources of uncertainty, and the synoptic global analysis still needs to be improved in the tropics. For these reasons, the atmospheric forcing is poorly known, and even the predictions of a perfect oceanic model would not be consistent with the observations.

The uncertainties in the atmospheric forcing and the oceanic initial conditions induce uncertainties in model response that have large correlation scales. In addition, the oceanic fields used for validation often are analyzed fields with highly correlated errors. Hence, distinguishing between model inadequacies and data uncertainties requires a multivariate viewpoint, rather than the usual univariate point-by-point comparisons with the observations. Frankignoul et al. (1989) have developed a multivariate model testing procedure which compares the main space and time structures of the model-observation differences to those of the data uncertainties. The method has been used to test and tune several ocean models in the tropical Atlantic ocean, which has a good historical data base (Duchêne and Frankignoul, 1991; Braconnot and Frankignoul, 1993, 1994).

The aim of the present paper is to describe a systematic intercomparison between 4 tropical ocean models of increasing complexity: the linear, multimode model of Cane (1984), the LODYC nonlinear 2-layer model, the KNMI general circulation model (GCM) (Allaart and Kattenberg, 1990) and the high resolution LODYC GCM (Blanke and Delecluse, 1993). Tropical Atlantic data are used as observational basis. The intercomparison is conducted for a primarily wind-driven "dynamical" variable, namely the depth of the thermocline represented by the 20°C isotherm depth, which is reasonably well-documented by

observations. The variations in thermocline depth are linked to the equatorial current system and indicative of vertical motions. In a forthcoming paper, the intercomparison will be extended to a "thermodynamical" variable, the sea surface temperature.

Following Braconnot and Frankignoul (1993, 1994), we use as thermocline depth the analyzed fields derived by Reverdin et al. (1991) from temperature observations during the 1982–1984 FOCAL/SEQUAL experiment. To represent the forcing uncertainties, we force each model with 3 equally plausible wind stress fields whose dispersion is a measure for the uncertainties due to wind sampling and measurement errors, and drag coefficient indeterminacy. The FOCAL/SEQUAL data have been compared to the linear model in Braconnot and Frankignoul (1993), using a larger set of plausible forcing fields. However, it is important in model intercomparisons to use identical settings and the analysis was redone. The three forcing fields have been used by Braconnot and Frankignoul (1994) to test the LODYC GCM, but the results differ slightly as we use here an improved version of the test. Note that a more rigorous validation of the GCMs would require to also represent the heat flux uncertainties, as vertical mixing is affected by the stratification, hence by the surface heat fluxes. However, the effect is expected to be small and has been neglected.

Section 2 describes the tropical Atlantic models included in the intercomparison. Section 3 discusses the observation, the forcing uncertainties, and the simulations of the thermocline depth variations during 1982–1984. The multivariate model testing procedure is used in section 4 to compare the four models to the observations.

2. The tropical Atlantic models

2.1. The linear multimode model

The simplest model that we consider here is that of Cane (1984). The model is based on the linearized equations on an equatorial beta plane with the long-wave approximation; it thus lacks the effects of mean flow and nonlinear interactions. The present version has three vertical modes, which are characteristic of mean tropical Atlantic conditions. The model basin extends from 30° N to

20°S and has a simplified geometry; its resolution is 1° in longitude and 0.5° in latitude. Here we only consider a domain (see Fig. 1) which should not be affected by the model artificial boundaries.

2.2. The LODYC 2-layer model

This model of intermediate complexity is a non-linear reduced-gravity model with a constant-depth surface mixed layer. It is derived from that of Schopf and Cane (1983), and described by Février (1993). The surface layer has a constant depth of 30 m, the second layer a mean thickness of 120 m, and the deep ocean is at rest. There is no thermodynamic feedback on the dynamics. The gravity wave speed is 1.7 m/s, the horizontal mixing coef-

ficient equal to $2 \cdot 10^3 \text{ m}^2/\text{s}$, and there is mixing of momentum between the first two layers. The model has realistic coastlines, but artificial boundaries at 27°N and 20°S. The resolution of the C-grid is variable, ranging in latitude between 0.6° at the equator and 2.6° at the northern boundary, and in longitude between 0.6° near the coasts and 1.5° in the center of the basin.

2.3. The KNMI general circulation model

The KNMI GCM is an Atlantic version of the tropical Pacific GCM of Allaart and Kattenberg (1990). It has 13 vertical levels (10 in the upper 300 m), and no bottom topography. The horizontal resolution of the C-grid is 2° zonally and

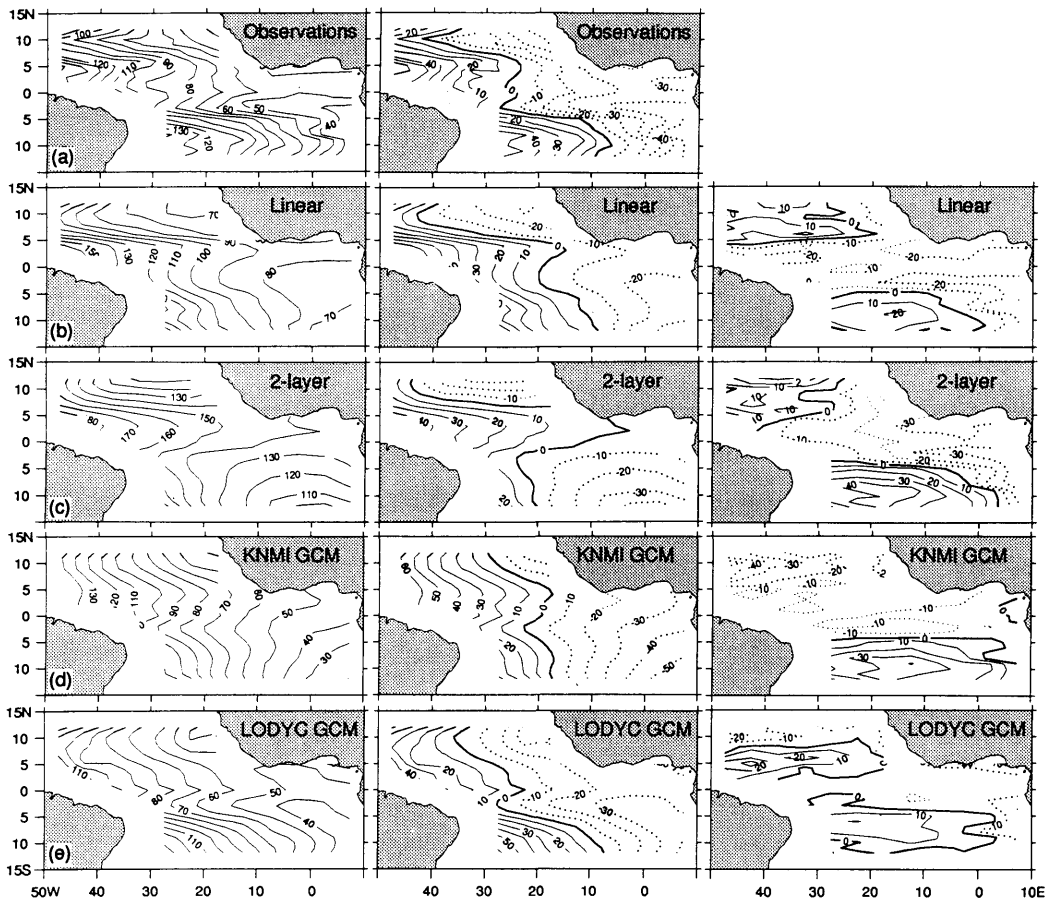


Fig. 1. Left: 3-year mean depth of the 20°C isotherm for (a) the observations, (b) the linear multimode model, (c) the nonlinear 2-layer model, (d) the KNMI GCM, and (e) the LODYC GCM. Center: same, after subtraction of the spatial mean. Right: differences between observation and simulations, after subtraction of the spatial mean.

varies in latitude from 0.5° at the equator to 2° at the north and south boundaries (50°N , 35°S). The horizontal mixing coefficient is equal to $5 \cdot 10^3 \text{ m}^2/\text{s}$. The vertical mixing is Richardson number dependent as in Pacanowsky and Philander (1981). The vertical static stability is ensured by convective adjustment. Integration is based on a semi-implicit scheme with a 2-h time step. The boundary conditions for the walls and the bottom are no-slip for momentum and no-flux for heat and salt.

The surface heat flux is specified from the climatological monthly data Q_{obs} of Oberhuber (1988), together with a linear feedback to the observed SST taken from Levitus (1982):

$$Q = Q_{\text{obs}} - \partial_t Q(T_{\text{obs}} - T), \quad (1)$$

where $\partial_t Q$ is given by Oberhuber (1988). Surface salinity is forced toward the Levitus (1982) climatology with a time constant of 30 days.

2.4. The LODYC general circulation model

The "TKE version" of the LODYC GCM is a primitive equation model with an embedded 1.5 turbulence closure model that is described in Blanke and Delecluse (1993). The model has 16 vertical levels with high resolution in the surface layers (13 levels in the upper 325 m), and a horizontal C-type grid varying in latitude from 0.33° at the equator to 1.5° at 50°N . In longitude, the resolution is about 0.75° . At 50°N and 40°S the ocean is closed by artificial coastlines in the vicinity of which a linear damping is applied on momentum. The horizontal mixing coefficients are equal to $10^3 \text{ m}^2/\text{s}$. The turbulent kinetic energy is defined through a prognostic equation while the turbulent length scales for the vertical mixing are defined by a diagnostic formulation.

The radiative heat fluxes are given by the Esbensen and Kushnir (1981) climatology, and the surface turbulent fluxes computed from bulk formulae using model sea surface temperature but climatological air temperature, taken from Esbensen and Kushnir (1981). In addition, the model temperature and salinity fields are relaxed towards the Levitus (1982) climatology, using restoring coefficients that increase with latitude θ and decreases with depth as (for temperature):

$$(T - T_{\text{obs}})(1 - \cos \theta)(k_1 + (k_2 - k_1) \times \exp(-z/H)), \quad (2)$$

where T is the simulated temperature and T_{obs} the climatological value, with $k_1 = (720 \text{ days})^{-1}$, $k_2 = (4 \text{ days})^{-1}$ and $H = 200 \text{ m}$. At the surface, the relaxation terms start to be significant 10° away from the equator.

3. Thermocline depth variations during 1982–1984

3.1. Observations

The low frequency variability of the 20°C isotherm depth in the tropical Atlantic was derived from temperature observations during the FOCAL/SEQUAL experiment in 1982–1984 by Reverdin et al. (1991). The data coverage was sparse, especially in 1982, so the fields were mapped monthly on a $1^\circ \times 5^\circ$ grid between 12°N and 12°S , using a function fitting algorithm with a cut-off period of 3 months. Because the functions are not local, there can be remote influences on data, especially in regions of inadequate sampling. However, Reverdin et al. (1991) estimated the error covariance matrix of the analyzed fields, so that the non-locality of the interpolation procedure is accounted for in the error fields.

In the tropical Atlantic, the prevailing northeast and southeast trade winds converge on the inter-tropical convergence zone (ITCZ), which moves meridionally with the season, with its increased rainfall and cloudiness. The ITCZ normally extends furthest south from February to April, with weak winds along the equator. It migrates northward in summer, bringing rapidly intensifying equatorial winds that are most intense in August and September, remaining strong until December. The atmospheric conditions during 1983–1984 have been described by Horel et al. (1986). The ITCZ was weak and slightly further north in spring 1983, and vigorous and further south in spring 1984. There was a gradual build-up of the trade winds during mid-1982, possibly associated with the 1982–1983 ENSO event, and the winds remained stronger than normal until a strong relaxation between November 1983 and January 1984.

The oceanic observations have been discussed by Reverdin et al. (1991). The 3-year mean thermocline depth shows the strong shoaling from west to east and, in the meridional plane, the troughs and ridges associated with the mostly

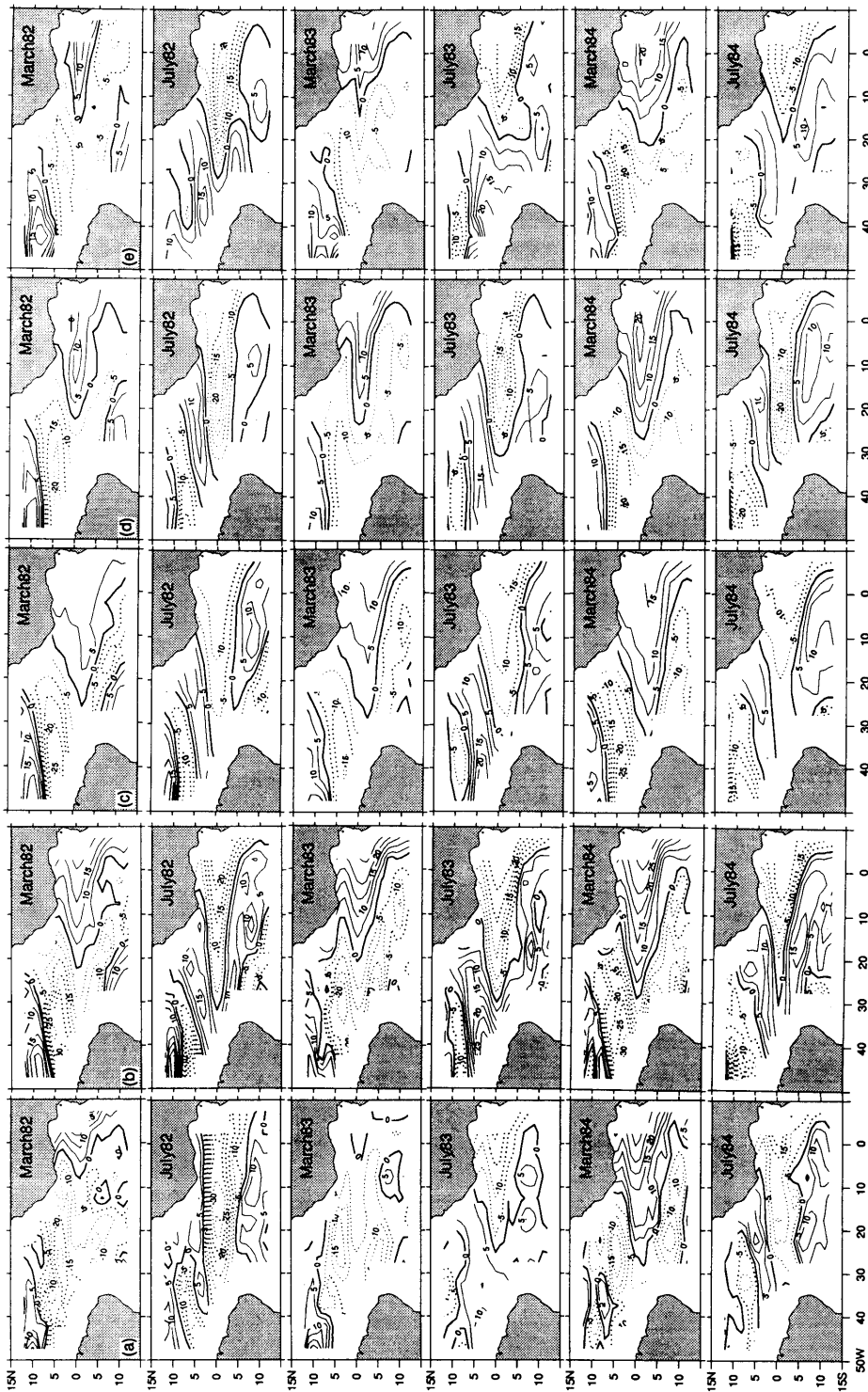


Fig. 2. (a) Observed departures (in meter) of the 20°C isotherm depth from the 1982–1984 annual mean for March and July, from 1982 to 1984. Corresponding simulations with (b) the linear multimode model, (c) the nonlinear 2-layer model, (d) the KNMI GCM, and (e) the LODYC GCM.

zonal tropical currents (Fig. 1, top). The seasonal and interannual variations around the 3-year annual mean are illustrated for March and July in Fig. 2 (left). The seasonal cycle dominates the variability. The zonal slope at the equator and the north-south slopes vary nearly in phase with the winds, so the thermocline is flatter during the first part of the year. The North Equatorial Countercurrent (NECC) disappears seasonally, flowing only eastward between July and December.

The thermocline was shallow in the whole equatorial band in summer 1982, and deep during the second half of 1983. In the eastern equatorial Atlantic, the thermocline deepened less during the beginning of the year in 1983, but more in 1984, as also illustrated in Fig. 3 (left). The 1984 deepening, which was strong from January to April, was followed by an anomalous SST warming off the coast of southwestern Africa, peaking in summer 1984.

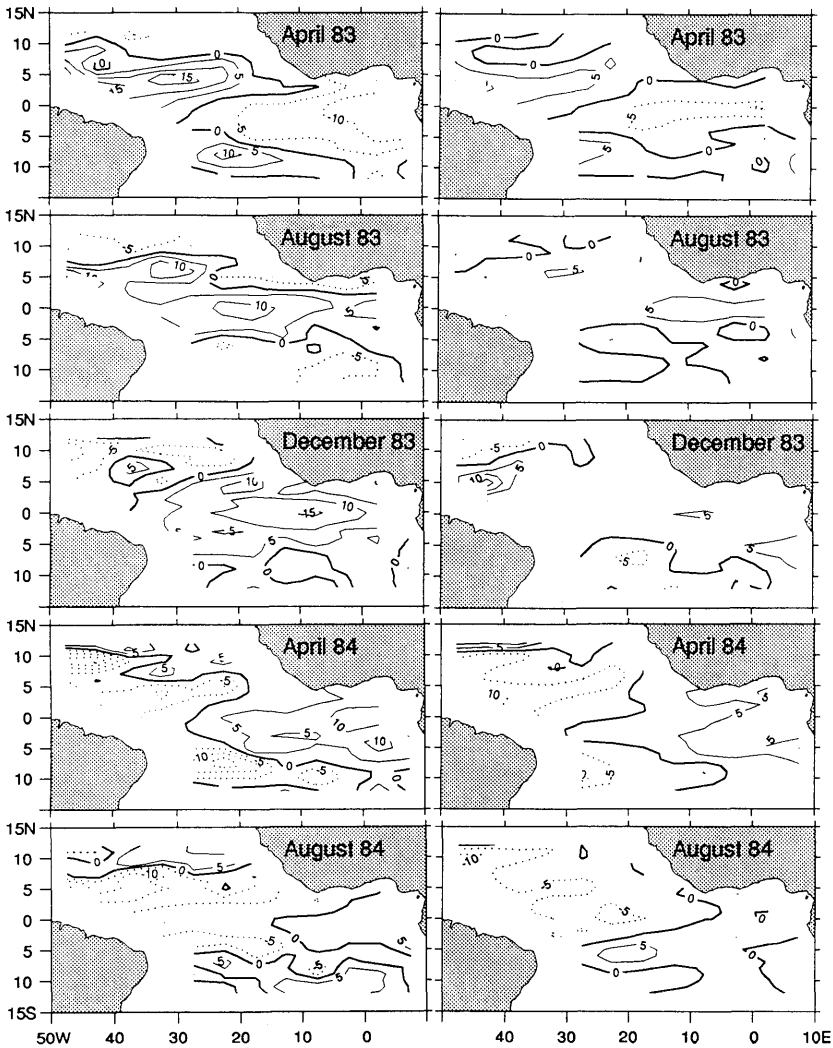


Fig. 3. Left: observed anomalies (in meter) of the 20°C isotherm depth from the 1982–1984 mean seasonal cycle for April, August and December from April 1983 to August 1984. Right: corresponding simulations with the LODYC GCM.

3.2. Forcing uncertainties

The main source of forcing uncertainty in model simulations of the tropical thermocline is the wind stress. For the 1981–1984 period, Braconnot and Frankignoul (1993) used a Monte Carlo approach and several stability-dependent drag coefficients to construct statistically independent, equally plausible monthly wind stress fields whose dispersion was representative of wind sampling, measurement errors and drag coefficient indeterminacy. Here we use the three fields selected in Braconnot and Frankignoul (1994), which are based on the bulk parameterizations of Large and Pond (1981), Smith (1988) and Wu (1986), and suffice to satisfactorily simulate these wind stress uncertainties. The corresponding thermocline depth response uncertainty has a rms of several meters, nearly as large as the predicted interannual variability (Braconnot and Frankignoul, 1994). Note that monthly wind stress fields were only available between 20°S and 30°N, so that they were smoothly merged (over 8°) with the climatology of Hellerman and Rosenstein (1983).

Additional forcing uncertainties ideally ought to be taken into account, even though their effects may be expected to be small. The errors associated with the limited spatial resolution should have been represented: the wind stress was calculated on a $5^\circ \times 2^\circ$ grid, so that the wind stress curl is strongly smoothed, possibly resulting in shifts in position and strength of narrow wind-driven currents such as the NECC. Nonlinear models are affected by day-to-day changes in the forcing, and the high frequencies should have been represented, particularly in the GCM runs. The uncertainties in surface heat exchanges need not be considered for dynamical variables in simple models with prescribed vertical structure. However, vertical mixing in the GCMs is strongly controlled by the stratification which in turn depends on the surface heat flux, so that the entire structure of the upper layers is affected by surface heating. Hence, its uncertainties should have been taken into account when testing the GCMs.

3.3. Simulations

The three independent wind stress fields have been used to drive the four tropical ocean models. To represent the effects of the unknown initial conditions, the simulations started from seasonal

equilibrium in 1981, except for the LODYC GCM which started from seasonal equilibrium in 1982.

The simulated mean thermocline depths (averaged over three years and three forcings) are shown in Fig. 1, left. Since the vertical resolution of the two intermediate resolution models is limited, the spatial means are removed for the intercomparison (Fig. 1, middle). Although the agreement with the observations varies, seeming best for the LODYC GCM and poorest for the 2-layer model, there are similarities in the model-observation differences (Fig. 1, right) which may be attributed in part to the lack of resolution of the wind stress curl (Subsection 3.2). In particular, the simulated meridional gradients between 4 and 10°N are smoother than observed, which should shift and spread the NECC, and the trough seen in the observations near 10°S is not reproduced. On the other hand, the models behave differently near the equator: the mean zonal gradient is realistically reproduced by the two GCMs, but underestimated by the 2-layer model. The crest near 3°S is also reproduced with varying degree of success, being most realistic with the LODYC GCM, and least realistic with the two simple models. Away from the equator, the KNMI lacks meridional structure and overestimates the zonal gradients.

The variations of the thermocline depth around the 3-year annual means (averaged over the three forcings) are shown for the four models in Fig. 2. The signal is dominated by the mean seasonal cycle and is, in general, reasonably well simulated by the models. The interannual variability with respect to the mean seasonal cycle tends to be underestimated, however, and some of the events seen in the analyzed fields are poorly reproduced, as illustrated in Fig. 3 for the LODYC GCM. There are many similarities in the model behaviors. Some features are not well reproduced by any of the models, like the interannual variations of the equatorial thermocline during July. During summer, the equatorial upwelling is too strong and shifted westward in the KNMI GCM, while the 20°C isotherm is deeper and the upwelling more eastward in the LODYC GCM, as observed. These differences between the two GCMs are likely to be due to the vertical mixing, which is intensified in the LODYC GCM by the use of a prognostic turbulent kinetic energy equation (Blanke and Delecluse, 1993); interestingly, the LODYC GCM behaved like the KNMI

GCM when its vertical mixing was similarly based on the local Richardson number (see Fig. 15 in Braconnot and Frankignoul, 1993). The strong 1983 shallowing and the 1984 deepening in the Gulf of Guinea during the early part of 1984 are strongly underestimated by the 2-layer model, but reasonably well-reproduced by the others, even if somewhat underestimated. As expected from the poor simulations of the NECC, its variability is poorly reproduced.

Because of our limited knowledge of the forcing fields, some of the discrepancies in Fig. 1 to 3 are due to forcing uncertainties, while others are due to errors in the analyzed maps, especially in data poor regions, which probably explains why the models may more closely resemble each other than the observations (as during March 1983). A more precise intercomparison must be based on signal-to-noise ratios, thereby taking into account the systematic or random nature of the model-observation differences.

4. Model intercomparison

4.1. Method

To try distinguishing between data uncertainties and systematic model errors, we use the model testing procedure of Frankignoul et al. (1989). Let us consider the simulation of the thermocline depth, denoted by $\mathbf{m}$, in the space-time domain where the model is supposed to be realistic and the observations, denoted by $\mathbf{d}$, are reliable. This is a high dimensional space where the uncertainties in the observations are characterized by an error covariance matrix $\mathbf{D}$, given by Reverdin et al. (1991), which defines a probability region for the (unknown) true oceanic state. The true forcing is not known, but the three different, equally plausible forcing fields also define a probability region for the model response. If the multinormal assumption holds, the model response is described by the sample mean $\bar{\mathbf{m}}$ over the $n = 3$ forcings and the error covariance matrix $\mathbf{M}$, defined by

$$\mathbf{M} = \frac{1}{n(n-1)} \sum_{j=1}^n (\mathbf{m}_j - \bar{\mathbf{m}})^t (\mathbf{m}_j - \bar{\mathbf{m}}), \quad (3)$$

where t indicates vector transpose and the index j identifies the drag law. The testing problem

reduces to that of comparing two multidimensional vectors with unequal error covariance matrices.

The agreement between observations and simulations is characterized by the misfit

$$T^2 = (\bar{\mathbf{m}} - \mathbf{d})^t (\mathbf{M} + \mathbf{D})^{-1} (\bar{\mathbf{m}} - \mathbf{d}), \quad (4)$$

which measures the differences between the two fields relative to the data accuracy, and can be interpreted as a multivariate signal-to-noise ratio (squared). If the null hypothesis that there is no model error holds, the test statistic (4) is approximately distributed as Hotelling's T^2 with p (the dimension of the intercomparison space) and q (the degrees of freedom of $\mathbf{M} + \mathbf{D}$) degrees of freedom. Since only part of the data uncertainties have been represented and no model is expected to be perfect, T^2 will normally be superior to its critical value at a reasonable level of significance and the null hypothesis rejected. However, the test statistic indicates how closely the model is able to reproduce the observations, and is thus handy for model intercomparison. In general, the misfit does not directly characterize model adequacy, as it depends on the data accuracy: a good model will fare poorly if the observations are sufficiently accurate, and a bad one, well, if the data are inaccurate. Finally, note that, in this paper, the error bar on T^2 should only be considered once when comparing the misfits of different models, since it represents the same observational and forcing uncertainties.

In practice, the dimension of the intercomparison space must first be strongly reduced. Indeed, the observational error covariance matrix $\mathbf{D}$ is provided by a limited error analysis and its details cannot be trusted. Furthermore, $\mathbf{M}$ is estimated from a small sample and is of much reduced rank. Thus, the inverse in (4), even if it exists, is dominated by unreliable information, and a strong data compression is needed for the problem to be well-posed.

4.2. Yearly means

The testing of the 3-year means is done for each model by working in the normalized basis of the mean model-observation differences, $\langle \bar{\mathbf{m}} - \mathbf{d} \rangle$, where the brackets indicate 3-year mean (Fig. 1, right). For each year and wind stress, the model-

observation differences are projected onto this subspace, yielding an amplitude $y_{i,j}$

$$y_{i,j} = (\mathbf{m}_{i,j} - \mathbf{d}_i)^t \mathbf{e},$$

where i denotes the year, j the forcing field and $\mathbf{e}$ the basis vector, so the misfit (4) is simply calculated in one dimension by comparing the mean difference $\langle \bar{y} \rangle$ to its sample variability. The latter is estimated by assuming independence between each of the $m = 3$ years (which is acceptable for the tropical Atlantic) and each of the $n = 3$ forcings. The test statistic reduces to

$$t = \frac{\langle \bar{y} \rangle}{\sqrt{s_f^2 + s_y^2}}, \quad (5)$$

with

$$s_f^2 = \frac{1}{n(n-1)} \sum_{j=1}^n (\langle y \rangle_j - \langle \bar{y} \rangle)^2,$$

$$s_y^2 = \frac{1}{m(m-1)} \sum_{i=1}^m (\bar{y}_i - \langle \bar{y} \rangle)^2$$

If the null hypothesis holds, it is approximately (because of the unequal variances) distributed like a t variable with degrees of freedom (e.g., Seber, 1984)

$$q = 2 \frac{(s_y^2 + s_f^2)^2}{(s_y^2)^2 + (s_f^2)^2}. \quad (6)$$

Together with the critical values at the 5% level, the misfits for the four models are given in Table 1 for the whole 12°N–12°S domain and for the equatorial waveguide between 3°N and 3°S; the western boundary currents are not included in the intercomparison since they cannot be represented well by the simpler models. Remembering that (5) is a signal-to-noise ratio for the difference patterns in Fig. 1 (right), the interpretation of the results in Table 1 is that the model-observation differences would need to be about t/t_{crit} time smaller for the models to be consistent with the observations at the 5% level, at least if all the data errors had been represented. In the best cases, the differences are three times too large. Although the forcing uncertainties are somewhat underestimated (Subsection 3.2), this suggests that the FOCAL/SEQUAL

Table 1. *Model-observation misfits for the yearly mean thermocline depth, with the critical values at the 5% level; the norm of the model-observation differences is given in parentheses*

Models	$t/\text{critical value}$	
	12°N–12°S	3°N–3°S
Linear	23.9/2.8 (15.3)	21.6/3.3 (18.3)
2-layer	64.1/3.4 (23.2)	46.0/2.8 (22.9)
KNMI GCM	50.5/4.3 (19.2)	8.7/2.9 (12.1)
LODYC GCM	13.2/3.4 (10.6)	10.7/3.6 (7)

data are sufficiently accurate to compare the models and detect some of their deficiencies.

As expected from the visual agreement in Fig. 1, the LODYC GCM performs best in the whole domain. In the equatorial waveguide, the two GCMs perform best but are not statistically distinguishable, even though the norm of the model-observations differences, $|\langle \bar{m} - d \rangle|$ (in parenthesis in Table 1), which corresponds to the visual agreement, is smaller for the LODYC GCM. This is due to the larger interannual variability of the KNMI GCM, which decreases the signal-to-noise ratio. The linear model shows moderate misfits, even outperforming the KNMI GCM when considered over the whole domain, which confirms that linear dynamics lead to reasonable thermocline simulations. On the other hand, the 2-layer model shows the poorest agreement with the observations (largest differences and small variability).

4.3. Variations around the 3-year annual mean

After subtraction of the 3-year annual mean, the data compression is done in two steps. For each model, an orthonormal base is first defined where the main features of simulations and observations are well represented. Then, a time compression is performed in this subspace. The spatial reduction is done using Common Principal Component Analysis (e.g., Flury, 1988). It only varies slightly from model to model and is illustrated in Fig. 4 for the 2-layer model. The first four common empirical

orthogonal functions (EOFs) account for more than 80% of the variance of the observed and simulated fields (left). Observations and simulations are represented in this subspace by the time series in Fig. 4 (right), where the 95% confidence intervals are derived from the diagonal terms of the error covariance matrices in reduced space, providing an estimate of the magnitude of the observational and forcing uncertainties. Note that we have used the error covariance matrix D of the full signal (Subsection 4.1), thereby neglecting the (small) influence on the observational errors of removing the 3-year mean. Models and observations are always well correlated in the subspace of the first and second vectors (correlation of 0.91 and 0.84 in Fig. 4, respectively), although the

data uncertainties that have been considered are insufficient to explain all the model-observation discrepancies. On the other hand, models and observations are poorly correlated or even anti-correlated in the subspace of the 3rd and 4th common EOFs (correlation of -0.15 and 0.16 in Fig. 4, respectively). This is to be expected as patterns that vary out-of-phase are well-represented by common EOFs.

In this paper, the time compression is done in a more efficient fashion than in Frankignoul and Braconnot (1993, 1994), who used common principal component analysis to determine the main temporal behaviors that are common to both observations and simulations in reduced space, and then tested whether model and observations

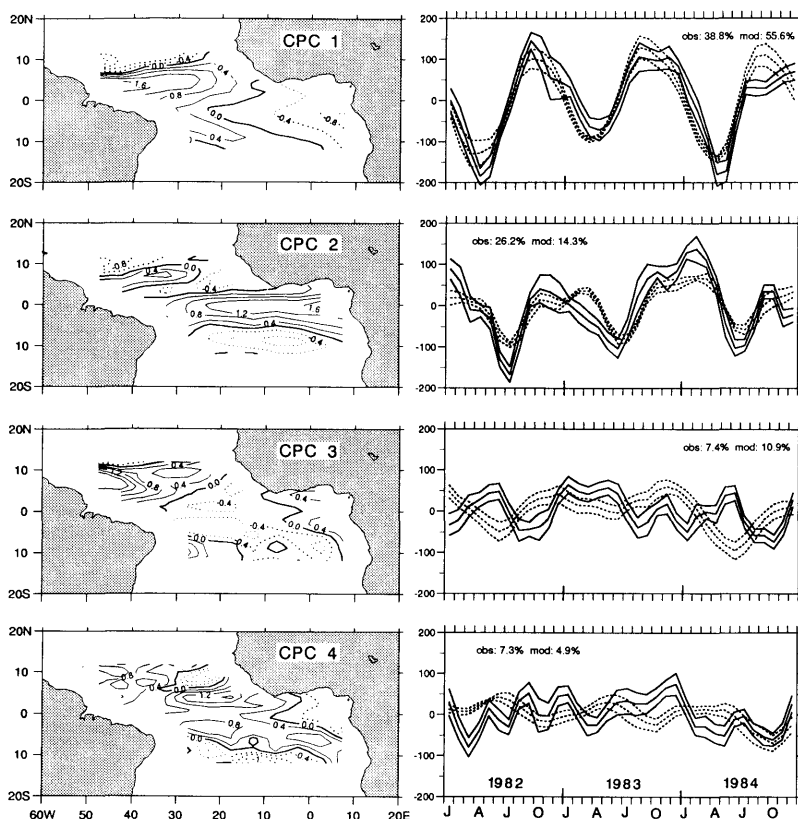


Fig. 4. Spatial reduction for the 2-layer model simulation over the 12°N – 12°S domain. Left: basis orthonormal spatial vectors. Right: projections (time series) of observed (continuous middle line) and modelled (dotted middle line) fields onto the spatial base. Time unit is months, and the percentage of represented variance indicated. The univariate 95% confidence intervals are estimated from the diagonal terms of the error covariance matrix in the reduced space.

were equal in the doubly reduced space. Here we work directly with the differences between the observed and model principal components: for each of the four common EOFs, we only consider the differences between the observed and simulated time series in Fig. 4 (right). Although the time dimension is 36 (number of months), 4 EOFs in the time domain (orthonormal vectors of dimension 36) suffice to completely represent the four difference time series, so the slight truncation that would result from using four common EOFs in the time domain is avoided. The dimension of the reduced space is $4 \times 4 = 16$, and we now test whether the model-observation differences are equal to zero. As model and observation errors are independent, the error covariance matrices simply add, and the degrees of freedom are estimated from a multivariate generalization of (5) (Seber, 1984, Frankignoul et al., 1989). Due to the large number of degrees of freedom, in all cases T^2 behaves approximately as a chi-squared variable with 16 degrees of freedom and the critical value at the 5% level is 26. Note that the yearly variability also enters the misfit for the yearly means, so that the two analysis are not entirely independent.

The misfit of the four models is shown in Fig. 5 for the 12°N – 12°S domain and the 3°N – 3°S equatorial wave guide, as well as for two well-sampled meridional “sections” (along 28°W and 4°W , as shown in Fig. 5, bottom). All the misfits are larger than the critical value at the 5% level, and there are statistically significant differences between the various models; model performances also exhibit significant regional differences. Since the misfit is the square of a multivariate signal-to-noise ratio, the model-observation differences should be, in the best cases, about 5 or 6 times smaller ($(T^2/T_{\text{crit}}^2)^{1/2}$) for the modelled variations to be consistent with the observed ones, at least if the data uncertainties had been well represented.

In the whole 12°N – 12°S domain, the more complex LODYC GCM again reproduces the variations around the 3-year annual mean more realistically than the other models. However, the KNMI GCM is also significantly superior to the two simpler models, which cannot be distinguished by the data set. In the equatorial waveguide, the two GCM perform equally well and are also more realistic than the 2-layer model, which itself outperforms the linear model.

Along the meridional section at 4°W , the two

GCMs are also undistinguishable, and slightly more consistent with the observations than the linear model. The multimode model is superior to the 2-layer model, possibly because the higher vertical modes play an important role in the Gulf of Guinea. Along 28°W , the KNMI GCM outperforms the LODYC GCM, and the two simple models are less realistic, but comparable. Note that the hierarchy between the models is more variable when the intercomparison is performed over small domains, which stresses that model-reality comparisons should be done *globally* in time and space, not locally.

4.4. Thermocline depth anomalies

As both the mean seasonal cycle and its anomalies are included in the analysis above, it is of interest to compare how well the models simulate the thermocline depth anomalies with respect to the climatological monthly means, even if the latter are estimated from only three years. Unfortunately, as the observational error covariance matrix was not available for the anomaly fields, the models could not be tested rigorously.

As a coarse approximation, we have calculated the misfits using the observational error covariance matrix of the full signal, which can be shown to lead to an overestimation of the errors by a factor of about 3/2, lowering an already small signal-to-noise ratio. In the whole 12°N – 12°S domain, the misfits for the different models are difficult to distinguish, except for the linear model which performs more poorly. A comparison with a “zero anomaly misfit” case obtained by setting $\bar{m} = 0$ in (4) (no model prediction) suggests that there is generally little skill, and none for the linear model. This confirms the limited agreement visible in Fig. 3. On the other hand, in the equatorial domain (3°N – 3°S) most models are significantly better than no prediction at all, except again for the linear model. Note that the LODYC GCM shows slightly smaller differences with the observations, but this is not reflected significantly in the tests, which take into account the data uncertainties. The calculation is too approximate and the data too noisy to allow distinguishing further between the performance of the models, and a better data set is clearly needed for a rigorous study.

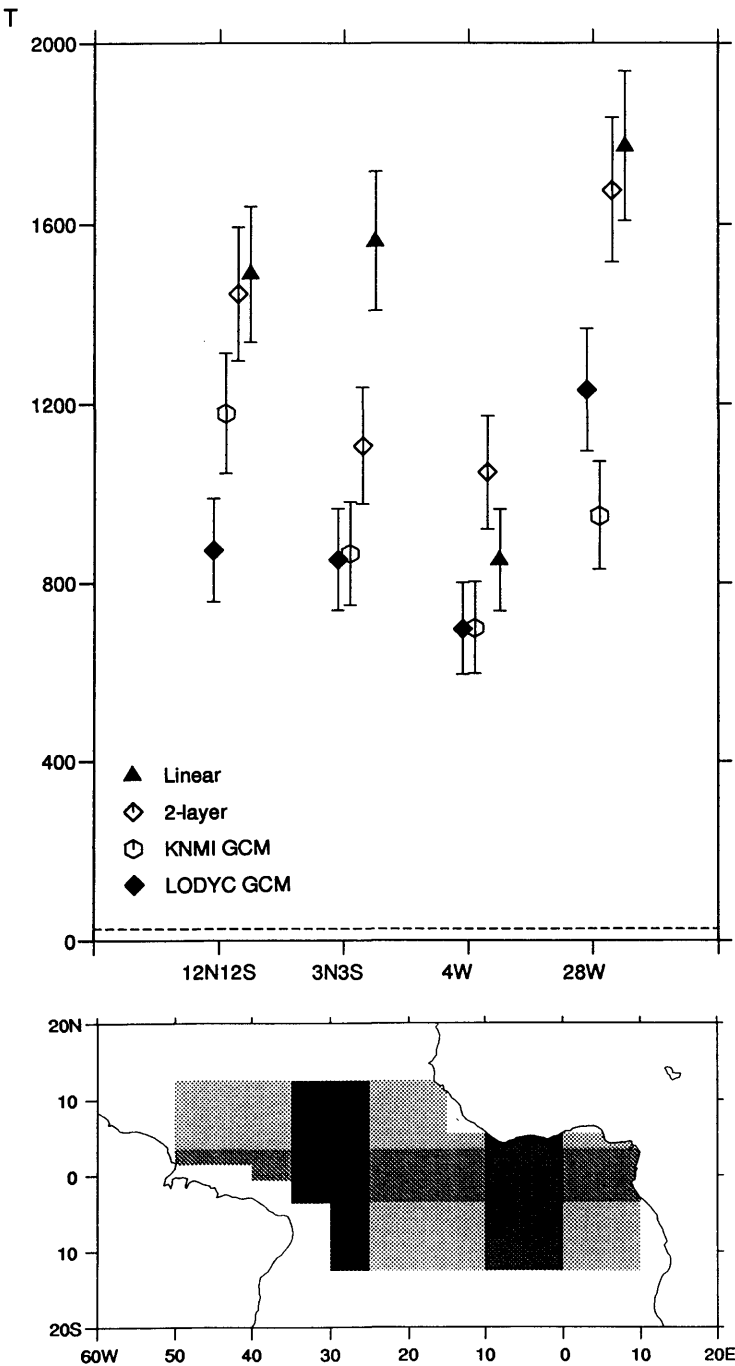


Fig. 5. Misfit between the 4 models and observations in the 4 testing regions indicated on the map below, for the thermocline depth variations around the 3-year mean. The error bars represent the estimated 95 % confidence intervals, and the dotted line the critical value for rejecting the null hypothesis of no model error at the 5 % level.

5. Summary

The multivariate model testing procedure of Frankignoul et al. (1989) has been used to investigate the simulation with four ocean models of the evolution of the thermocline depth in the tropical Atlantic during 1982–1984. To take some of the forcing uncertainties into account, the models were forced with three independent, equally plausible windstress fields whose dispersion is representative of the wind sampling and measurement errors, and the drag coefficient indeterminacy. The observational errors in estimating the depth of the 20°C isotherm were taken from the objective analysis of Reverdin et al. (1991).

The intercomparison shows that there are significant differences in model performance. Even though thermocline depth is often believed to be primarily governed by linear dynamics, the GCMs represent its changes more realistically than the models of intermediate resolution. When the whole intercomparison domain (12°N–12°S) is considered, the higher resolution LODYC GCM performs better than the KNMI GCM, but the performance of the two models cannot be distinguished statistically in the 3°N–3°S equatorial waveguide. The nonlinear 2-layer model poorly represents the long-term mean position of the thermocline. However, it simulates well its variations around the mean, outperforming the linear model in the equatorial waveguide. Along well-sampled meridional sections, the models show more variable skill and are more difficult to differentiate, stressing that model validations should be done over large domains, rather than locally as often encountered in the literature. Finally, a coarse investigation of the simulations of the thermocline depth anomalies with respect to the mean seasonal cycle suggests that, because of the small signal-to-noise ratio, most models have only little skill in the 12°N–12°S domain, but more in the equatorial waveguide. The linear model, however, does not seem to have any predictive skill for these noisy anomalies.

The present results generalize those of Braconnot and Frankignoul (1994), who showed that the LODYC GCM was simulating the evolution of the 20°C isotherm in the tropical Atlantic significantly better than the linear model. Here it is found that, in the sample of four models that were considered, performance generally increases with model complexity (and cost), even though the linear model represents better the long-term mean than the 2-layer model. An intercomparison between these models (except the KNMI GCM) was recently conducted for the seasonal cycle of surface dynamic topography in the tropical Atlantic (Février, 1993). Although the wind stress uncertainties were more coarsely represented, the results were similar, but the 2-layer model also outperformed the linear model for the annual mean. However, the 2-layer model was tuned to the latter variable.

It will be of interest to verify whether the more complex models also show the highest skill at simulating the sea surface temperature, especially in view of the good performance of the thermodynamical version of Cane's (1984) model in the tropical Pacific recently discussed by Miller et al. (1993). This will be investigated in a forthcoming paper.

6. Acknowledgements

Part of this paper was presented at the workshop "Forecasting Short-range Climate Variations" in Hamburg, 28–30 April 1993. The KNMI GCM was adapted to the Atlantic with the help of Marc Allaart and Arie Kattenberg, and the LODYC GCM runs were kindly provided by Pascale Delecluse. Thanks are due to a reviewer for a useful suggestion. This research was made possible in France by a grant from the PNEDC and by the Commission of the European Communities (EV4C006). N.S. was supported by a grant from the DRET. Support for the LODYC model runs was provided by the Centre de Calcul Vectoriel pour la Recherche (CCVR).

REFERENCES

- Allaart, M. A. F. and Kattenberg, A. 1990. *A primitive equation model for the equatorial Pacific*. Technical Report TR-124, KNMI, P.O. Box 201, 3730AE de Bilft, the Netherland.
- Blanke, B. and Delecluse, P. 1993. Low frequency variability of the tropical Atlantic ocean simulated by a general circulation model with mixed layer physics. *J. Phys. Oceanogr.* **23**, 1363–1388.

- Braconnot, P. and Frankignoul, C. 1993. Testing model simulations of the thermocline depth variability in the tropical Atlantic from 1982 through 1984. *J. Phys. Oceanogr.* **23**, 626–647.
- Braconnot, P. and Frankignoul, C. 1994. On the ability of the LODYC GCM at simulating the thermocline depth variability in the equatorial Atlantic. *Climate Dynamics* **9**, 221–234.
- Cane, M. A. 1984. Modeling sea level during El Niño. *J. Phys. Oceanogr.* **14**, 1864–1874.
- Cane, M. A., Zebiak, S. E. and Dolan, S. C. 1986. Experimental forecasts of El Niño. *Nature* **321**, 827–832.
- Duchêne, C. and Frankignoul, C. 1991. Seasonal variations of surface dynamic topography in the tropical Atlantic: observational uncertainties and model testing. *J. Mar. Res.* **49**, 223–247.
- Esbensen, S. K. and Kushnir, Y. 1981. *The heat budget of the Global ocean: an Atlas based on estimates from marine surface observations*. Climatic Research Institution, Report 29, Oregon State University, Corvallis, 27 pp.
- Février, S. 1993. *Modèle à deux couches de l'océan atlantique tropical*. Thèse de doctorat. Université Pierre et Marie Curie, Paris.
- Flury, B. 1988. *Common principal components and related multivariate models*. Wiley and Sons, 258 pp.
- Frankignoul, C., Duchêne, C. and Cane, M. 1989. A statistical approach to testing equatorial ocean models with observed data. *J. Phys. Oceanogr.* **19**, 1191–1208.
- Hastenrath, S. and Heller, L. 1977. Dynamics of climatic hazards in north-east Brazil. *Quart. J. Roy. Meteor. Soc.* **103**, 77–92.
- Hellerman, S. and Rosenstein, M. 1983. Normal monthly wind stress over the world ocean with error estimates. *J. Phys. Oceanogr.* **13**, 1093–1104.
- Horel, J. D., Kousky, V. E. and Kagano, M. T. 1986. Atmospheric conditions in the Atlantic sector during 1983–1984. *Nature* **322**, 248–251.
- Large, W. G. and Pond, S. 1981. Open ocean momentum flux measurements in moderate to strong winds. *J. Phys. Oceanogr.* **11**, 324–336.
- Levitus, S. 1982. *Climatological atlas of the world ocean*. NOAA Prof. Pap., 13, 173 pp, US Govt. Print. Office, Washington, DC.
- Miller, A. J., Barnett, T. P. and Graham, N. E. 1993. A comparison of some tropical ocean models: hindcast skill and El Niño evolution. *J. Phys. Oceanogr.* **23**, 1567–1591.
- Oberhuber, J. M. 1988. *An atlas based on the COADS data set*. Report 15, Max-Planck-Institut für Meteorologie, Hamburg.
- Pacanowski, R. and Philander, S. G. H. 1981. Parameterization of vertical mixing in numerical models of tropical oceans. *J. Phys. Oceanogr.* **11**, 1443–1451.
- Reverdin, G., Delecluse, P., Levi, C., Morlière, A. and Verstraete, J. M. 1991. The near surface tropical Atlantic in 1982–1984. Results from a numerical simulation and data analysis. *Progress in Oceanogr.* **27**, 273–340.
- Schopf, P. S. and Cane, M. A. 1983. On equatorial dynamics, mixed layer physics and sea surface temperature. *J. Phys. Oceanogr.* **13**, 917–935.
- Seber, G. A. F. 1984. *Multivariate observations*. Wiley and Sons, 686 pp.
- Smith, S. D. 1988. Coefficients for sea surface wind stress, heat fluxes and wind profiles as a function of wind speed and temperature. *J. Geophys. Res.* **93**, 15467–15472.
- Wu, J. 1986. Stability parameters and wind stress coefficient under various atmospheric conditions. *J. Atmos. Oceanic Technol.* **3**, 333–339.