Optimal Growth of Antarctic Circumpolar Waves

CHRISTOPHER M. AIKEN AND MATTHEW H. ENGLAND

Centre for Environmental Modelling and Prediction, School of Mathematics, University of New South Wales, Sydney, New South Wales, Australia

CHRISTOPHER J. C. REASON

Department of Oceanography, University of Cape Town, Cape Town, South Africa

(Manuscript received 9 August 2004, in final form 10 August 2005)

ABSTRACT

Generalized stability theory is applied to a simple dynamical model of interannual ocean–atmosphere variability in the southern midlatitudes to determine the perturbations that create the most rapid growth of energy in the system. The model is composed of a barotropic quasigeostrophic atmosphere coupled to a 1.5-layer quasigeostrophic ocean, each linearized about a zonally invariant mean state, and with atmospheric and ocean surface temperature obeying a simple heat balance. Eigenanalysis of the system reveals modes of interannual variability that resemble the so-called Antarctic Circumpolar Wave (ACW), consistent with an earlier analytical study of the system. The optimal excitation of these modes relative to an energy norm is found to be a perturbation almost entirely restricted to the ocean momentum field and is shown to resemble strongly the optimal perturbations in energy for the system. Over interannual time scales most rapid growth is seen in zonal wavenumbers 4–6, despite the fact that the least-damped eigenmodes of the system are of a lower zonal wavenumber. The rapid transient growth in energy occurs by extracting perturbation energy from the mean state through advection of the mean meridional oceanic temperature gradient. This transient growth of high-zonal-wavenumber modes dominates the model’s variability when it is forced by noise that is white in space or time. A dominant low-zonal-wavenumber response, consistent with the observed and modeled ACW, occurs only when the forcing is red in space or time, with decorrelation scales greater than 3 yr or 10 000 km. It is concluded that, if the ACW is a coupled mode analogous to that supported in this simple model, then it is excited by other large-scale phenomena such as ENSO rather than by sources of higher-frequency forcing.

1. Introduction

Over the past two decades there has been a sustained effort toward observing and understanding interannual to decadal climate variability. This has been spurred to some degree by the need for a baseline against which to detect changes in the present climate. The identification of interannual and decadal modes of variability also offers the prospect of increased climate predictability, with that afforded by improved understanding of ENSO being a notable example. As such, the discovery of phase-linked interannual anomalies in a number of atmospheric and oceanic variables in the southern mid- to high latitudes, such as the Antarctic Circumpolar Wave (ACW), may have important ramifications for improving our understanding of Southern Hemisphere climate variability and assessing its predictability. In this study we focus on the ACW, although we recognize there are other important modes in the mid- to high southern latitudes, such as the Antarctic oscillation and the semiannual oscillation, that are also important in the context of climate variability studies.

The ACW was originally characterized as a set of eastward-propagating wavenumber-2 anomalies in sea surface temperature (SST), wind stress, sea level pressure (SLP), sea surface height (SSH), and sea ice extent (White and Peterson 1996; Jacobs and Mitchell 1996). SLP anomalies lead SST anomalies by approximately $\pi/4$, suggesting a coupling mechanism involving geostrophic advection of the mean SST field and a subsequent barotropic atmospheric response to surface heating. The whole pattern appears to be advected by the
Antarctic Circumpolar Current, giving a period between phases of roughly 4 yr at any particular point in the Southern Ocean. The shortness of the observational record from which the ACW was deduced in comparison with the period of the mode called into doubt the robustness of the phenomenon, because of the possibility that the analysis of a dominant coupled mode in the data may be an artifact of the insufficient sampling period, sparse data, and/or analysis methods.

A number of recent analyses of extended datasets have been in general agreement with the original results, although with a number of key differences. Bonekamp et al. (1999) demonstrated that only the SST and not the SLP ACW signal was present in the 6 yr preceding the White and Peterson (1996) data, while Cai and Baines (2001) and Venegas (2003) find that zonal wavenumber 3 also contains substantial variability over longer observational periods in addition to zonal wavenumber 2. On the other hand, a significant number of studies have shown that a large majority of the interannual Southern Hemisphere atmospheric variability exists in wavenumber-1 and -3 quasi-stationary waves, rather than propagating waves (e.g., van Loon and Jenne 1972; Kalnay et al. 1986, Milliff et al. 1999). Park et al. (2004) also find a predominance of standing waves in SST, SSH, and SLP, with a link to ENSO as in Cai and Baines (2001), while the propagating component was found to explain less than 25% of the variance and to not be truly circumpolar.

The existence of an ACW has also received support from a number of modeling studies that have reproduced ACW-like modes. The coupled general circulation models (GCMs) of Christoph et al. (1998) and Cai et al. (1999) simulated wavenumber-3 eastward-propagating modes in SST and sea surface density, but with a standing rather than propagating wave pattern in SLP. Oceanic GCMs forced by observed (Bonekamp et al. 1999) and stochastic (Weisse et al. 1999) surface fluxes also support eastward-propagating low-wavenumber anomalies in SST and sea surface salinity, clearly without the need for explicit ocean-to-atmosphere coupling.

A number of mechanisms have been proposed as being responsible for the ACWs seen in models and observations. The ACW signal is thought by some to indicate the existence of a true coupled mode of the southern midlatitudes, either being self-sustaining (Qiu and Jin 1997) or externally excited by ENSO events (White et al. 1998) or through environmental stochastic forcing (Baines and Cai 2000, hereinafter BC). Analytical models studied by Qiu and Jin (1997), Talley (1999) and BC each support the existence of unstable ACW-like coupled modes, with the propagation characteristics of the original data. Cai et al. (1999) also support the concept of a fully coupled mode, but with a standing wave SLP structure whose period is set through interaction with the propagating SST anomalies. Observation-based studies by Cai and Baines (2001) and Venegas (2003) suggest that the wavenumber-2 and -3 components of the ACW signal have separate origins, with the former being forced by ENSO through the Pacific–South American pattern (PSA), and the latter in the manner suggested by Cai et al. (1999). Park et al. (2004) also find ENSO to modulate the quasi-stationary variability. Christoph et al. (1998), Bonekamp et al. (1999), and Weisse et al. (1999), however, find that the oceanic component of the mode in their models does not require an explicit feedback to the atmosphere, with ACW-like SST anomalies being forced predominantly by independent large-scale variability in wind stress curl. Aiken and England (2005) also showed that a stochastically forced SST advection equation was capable of sustaining an ACW-like SLP response without explicit atmospheric coupling.

In summary, while it appears that ACW-like modes of interannual variability in SST can be sustained without coupling to a dynamical atmosphere, modes involving an atmospheric response to anomalous SST also appear to be possible in the southern midlatitude ocean–atmosphere system and are consistent with observations. In the present study the excitation of such coupled modes is explored through analysis of a simple linearized coupled model. The model is based on that of BC, although it is extended to allow nonsinusoidal zonal variability. Techniques of generalized linear stability theory are employed to determine mechanisms for both transient and long-term growth of these coupled ACW-like anomalies. The response of the model to stochastic forcing is then understood in terms of the excitation of the optimal linear perturbations, but it will be shown that the response only resembles the observed ACW when forcing is dominantly of a large spatial or temporal scale. These results are consistent with the recent observation-based studies by Cai and Baines (2001) and Venegas (2003).

The rest of this paper is organized as follows. The coupled model is presented in section 2, and its eigenmodes are discussed in section 3. Section 4 summarizes some results of generalized stability theory, which are then applied to the model to examine the system’s non-normality in section 5 and to determine the optimal perturbations in section 6. These results are then used to understand the response of the model to stochastic forcing in section 7. We summarize our major conclusions in section 8.
2. A linearized coupled model of Southern Ocean SST anomalies

Interannual variability in the southern midlatitudes is investigated through analysis of the development of small perturbations to the mean state of the atmosphere–ocean. The assumption is made that the amplitude of interannual climate variability is sufficiently small relative to that of the mean climate state so that the evolution of anomalies may be simulated using a tangent linear model (TLM). Note that the anomalies discussed here are identically the perturbation variables more commonly referred to in the generalized stability theory literature. That is, the perturbations to the mean state are also referred to here as anomalies. The TLM is formed by applying standard linearization techniques to the full nonlinear system. We base our TLM on that derived and studied in BC. This model represents a barotropic quasigeostrophic atmosphere coupled to a 1.5-layer quasigeostrophic ocean, each linearized about a zonally uniform mean state. Both atmosphere and ocean are forced by anomalous heating or cooling. Additionally, the ocean is forced by wind stress curl, while the atmosphere experiences surface drag. Atmospheric and oceanic surface temperature anomalies evolve according to linearized advection equations, with a standard parameterization for heat transfer between the two.

The anomalous climate state is given in terms of atmospheric and oceanic streamfunction and temperature anomalies \( \psi_a, \psi_o, T_a, \) and \( T_o, \) while the background state is defined by the mean zonal atmospheric \((\bar{U}_a)\) and oceanic \((\bar{U}_o)\) velocities, and the mean meridional temperature gradient in the atmosphere \((\bar{T}_a)\) and ocean \((\bar{T}_o)\). The TLM is then written as follows:

\[
\left[ (\partial_t + \bar{U}_a \partial_x)(\nabla^2 - \lambda^2_a) + \beta_o \partial_y + C_x \nabla^2 H^2 - \nu_{ho} \nabla^4 \right] \psi_a = -|f| \bar{K}_1 \frac{\bar{U}_a}{H_o \bar{T}_a} (T_o - T_a), \]

(1)

\[
\left[ (\partial_t + \bar{U}_o \partial_x)(\nabla^2 - \lambda^2_o) + \beta_o \partial_y - \nu_{ha} \nabla^4 \right] \psi_o = \bar{K}_3 \frac{\bar{U}_o}{D_o} \nabla^2 \psi_a + K_3 \bar{f} \partial_y \psi_a, \]

(2)

\[
(\partial_t + \bar{U}_a \partial_x - \kappa_{ha} \nabla^2) T_a = -\partial_y \bar{U}_a \partial_x \psi_a + K_2 \bar{U}_a / H_a (T_o - T_a), \]

(3)

\[
(\partial_t + \bar{U}_o \partial_x - \kappa_{ho} \nabla^2) T_o = -\partial_y \bar{U}_o \partial_x \psi_o - K_2 \bar{U}_o / D_a (T_o - T_a). \]

(4)

Here \( x \) and \( y \) are the zonal and meridional coordinates, respectively, \( \partial \) represents the derivative with respect to the subscript variable, and subscripts \( a \) and \( o \) refer to the atmosphere and ocean, respectively; \( \nabla^2_H = \nabla^2 + \nabla^2_o \), \( \lambda \) is the Rossby radius, \( f_o \) is a central value of the Coriolis parameter, \( \beta = \partial f / \partial y \), \( \nu_h \) is a horizontal eddy viscosity, and \( \kappa_o \) is a horizontal eddy diffusivity; \( H_o \) is the atmospheric scale height, \( D_o \) is the thickness of the active ocean layer, and \( \alpha_o \) is the thermal coefficient of expansion for water. The coupling parameters are given by \( K_1 = C_H(1 + 1/B) \), \( K_2 = C_H(1 + 1/B) \), \( K_3 = C_D \), \( C_F \) with \( B \) the Bowen ratio, \( A \) a parameter that is weakly scale dependent, \( C_H = C_H (\bar{U}_a(0)/\bar{U}_a) \) with \( C_H \) a surface heat transfer coefficient, and \( C_D = C_D \bar{U}_a(0)/\bar{U}_a \) with \( C_D \) the surface drag coefficient. A rigorous derivation of the form taken by the coupling parameters \( K_i, i = 1, 4 \) is given in appendix A of BC.

The system (1)–(4) can be summarized as

\[
\frac{du}{dt} = Au, \quad (5)
\]

where \( u \) represents the perturbation state vector \([\psi_a, \psi_o, T_a, T_o]^T\) and \( A \) is the tangent linear operator. For time-independent \( A \) the system has the solution

\[
u(t) = e^{At}u(0) = R(0, t)u(0), \quad (6)
\]

where \( R \) is known as the propagator. In their study, BC presented an analytical normal mode analysis of the TLM (1)–(4), which is essentially equivalent to performing eigenanalysis on the tangent linear operator \( A \) in (5). They assumed solutions with spatial components proportional to \( e^{i(kx + ly)} \), in which case \( A \), and the resulting eigenvalue problem, reduce to dimension 4. We refer to this wavenumber-dependent model as the analytical model in the following.

Transient growth may be achieved through the linear interference of a number of eigenmodes of differing wavenumbers and frequencies. To allow this possibility, and hence determine a more complete inventory of anomaly growth in the model, in this study the TLM was solved numerically on a zonal grid. It will be shown that application of generalized stability theory to this model furnishes an understanding of the response of the system to forcing that would not be gained through traditional eigenanalysis. In the numerical model the meridional structure of the form \( e^{iy} \) was assumed as in BC, while 200 east–west grid points were used giving a zonal resolution of approximately 150 km. Spatial derivatives were calculated through centered differences, and periodic boundary conditions were applied at the zonal extremities. The matrix equivalent to \( A \) (in this case of dimension 800) was generated by the perturbation method directly from the numerical model that solves the discretized version of the TLM. The coupling
terms from the analytical model are maintained in this version. The value of $K_i$ is strictly speaking weakly scale dependent in the derivation of BC, because it contains the scale-dependent parameter $A$, while in this model a range of scales are considered simultaneously. For generating the results that follow, a single value was used for $A$, and hence $K_i$, which corresponds to the gravest zonal wavenumber. Sensitivity to this parameter was found to be minimal.

The parameter values used are the following, identical to the “standard set” from BC: $B = 0.5$, $C_D = 0.0013$, $C_H = 0.0013$, $c_{pu} = 1006$ J kg$^{-1}$ K$^{-1}$, $c_{pu} = 3985$ J kg$^{-1}$ K$^{-1}$, $c_s = 330$ m s$^{-1}$, $D_a = 400$ m, $f = -10^{-4}$ s$^{-1}$, $g' = 0.01$ m s$^{-2}$, $H_s = 9300$ m, $T_{py} = T_{py} = 6.8 \times 10^{-6}$ K m$^{-1}$, $U_a = 10$ m s$^{-1}$, $U_o = 0.1$ m s$^{-1}$, $\alpha'_w = 0.0001$ K$^{-1}$, $\beta = 1.63 \times 10^{-11}$ m$^{-1}$ s$^{-2}$, $U_{asy} = -0.61 \times 10^{-11}$ m$^{-1}$ s$^{-1}$, $\lambda = 3300$ km, $\lambda = 20$ km, $\rho_a(0) = 1.2$ kg m$^{-3}$, $\rho_o = 1000$ kg m$^{-3}$, $\nu_a = 10^6$ m$^2$ s$^{-1}$, $\nu_o = 9 \times 10^5$ m$^2$ s$^{-1}$, $\kappa_{hu} = 10^3$ m$^2$ s$^{-1}$.

### 3. Eigenmodes of the TLM

Numerical eigenanalysis was performed on the zonally discretized model outlined above, and also on the analytical model for all zonal wavenumbers $k$ between 1 and 50. This is in essence equivalent to the stability analysis performed in BC. In both cases the results are in close agreement with BC’s analysis, as should be expected. In both models coupled modes of interannual low-wavenumber eastward-propagating anomalies with a similar phase structure to the observed ACW exist, and are the least damped of all eigenmodes in each case. As will be discussed further below, two eigenmodes with the characteristic phase structure of the ACW occur for each zonal wavenumber. These will be referred to in future as ACW modes, although clearly in general the zonal wavenumber will not match that of the observed ACW.

As shown in BC, when viscous and dissipative terms are ignored in the analytical model, each of the first three wavenumbers are found to be slightly unstable, with wavenumber 2 being marginally the fastest growing. Inclusion of the viscosity and dissipative terms in either of the models, however, renders all modes damped, with wavenumber 1 becoming the least-damped eigenmode. That is, we find that the slight dominance of wavenumber 2 is sensitive to the presence of dissipation, as was found by BC. In fact, in the numerical model wavenumber 1 is slightly fastest growing even without explicit viscosity, a result we attribute to the dissipation inherent in the numerical differencing scheme. Note, however, that the eigenvalues of the numerical and analytical models agree closely (to within less than 1% for the first 10 modes). The apparent qualitative difference between the models for the invisible case is in fact because of very small quantitative differences in growth rates. As will be shown below, the small differences in growth/decay rates for the first few wavenumber ACW modes are largely insignificant for explaining the model’s response to continuous forcing. In the following we consider only the viscous case for the numerical model, in which case all eigenmodes are damped. The frequency and e-folding decay time of the least-damped wavenumber-1 ACW mode are approximately 10.1 yr$^{-1}$ and 55.6 yr, respectively, and for wavenumber 2 are approximately 5 yr$^{-1}$ and 48.1 yr, respectively.

The entire spectrum of eigenvalues $\sigma$ of the discretized model is shown in Fig. 1a, with the part of the spectrum close to the origin shown in more detail in Fig. 1b. The modes whose eigenvalues do not appear in Fig. 1b are high frequency, strongly damped, and dynamically unimportant, and so will not be considered further. The eigenvalues in Fig. 1b can be seen to fall on two separate “branches,” identified in the figure by differently shaped markers. The eigenvalues of these two branches correspond to the ACW modes introduced above. On each branch zonal wavenumber increases with distance from the origin, so it may be seen that for each zonal wavenumber there are two distinct ACW modes. Hereinafter these will be referred to as the type-1 and type-2 ACW modes, corresponding to the eigenvalues marked by triangles and dots, respectively, in Fig. 1b. It can be seen that for wavenumbers below 14 the type-1 ACW modes have larger values of Real($\sigma$) and hence are less damped than the type-2 ACW modes. The structure of the type-1 and type-2 wavenumber-1 ACW modes are shown in Fig. 2 by the solid and dashed curves, respectively. The differences seen in Fig. 2 are characteristic of the differences between the two types of ACW modes in general. At each zonal wavenumber, the amplitude and phase of the two ACW modes are almost identical in atmospheric streamfunction and atmospheric and oceanic temperature, but differ for oceanic streamfunction: $\psi_o$ in the type-1 ACW modes leads in phase and has an amplitude two orders of magnitude greater than $\psi_o$ in the type-2 ACW modes. The type-1 ACW mode is also of a slightly lower frequency than the type-2 ACW mode at each zonal wavenumber. Thus, at every zonal wavenumber two very similar lightly damped modes exist with characteristics resembling the observed ACW. It is only the type-1 ACW modes of a low wavenumber, whose eigenvalues are uppermost in Fig. 1a, that traditionally would be expected to dominate the system response. It is shown below, however, that all modes rep-
resented in Fig. 1b have importance, in that linear interference between the two types of ACW modes can produce transient perturbation growth that can explain a significant fraction of the model’s response to forcing.

Although all modes decay exponentially in this model, BC suggest the possibility that, because the ACW modes are only lightly damped, they may be sustained through stochastic forcing. This point will be discussed in section 7, in particular with reference to the optimal excitation of these modes that is discussed in the following sections.

As noted in BC, in the analytical model without the viscous damping terms growth/decay rates of the modes at each zonal wavenumber are sensitive to some model parameters, in particular the mean atmospheric flow speed $U_a$, meridional wavenumber $l$, and the oceanic Rossby radius $R_o$. Similar dependence does not occur for the discretized model considered here in the presence of viscosity, however. The model displays some sensitivity to the choice of $R_o$, with the upper (least damped) branch of ACW modes being damped more strongly with increasing wavenumber as $R_o$ is increased, but modal decay rates are found to be relatively insensitive to the choice of other parameters such as $A$, $U_a$, $D$, and $l$.

4. Generalized stability analysis

Historically, stability analysis of linearized systems has been restricted to a search for exponentially growing or decaying solutions to the system, that is, to the calculation of the system’s eigenmodes. However, it is now appreciated that system variance is not always determined by the properties of the fastest growing or least-damped eigenmode. This is true in particular of geophysical systems for two reasons. First, the ubiquitous presence of environmental noise in geophysical systems may sustain damped eigenmodes that otherwise, in a controlled environment, would contain no system variance. Second, the tangent linear operators that govern the development of small perturbations in geophysical systems are often found to possess the property of nonnormality (where the operator does not commute with its adjoint operator), with the implication that transient growth of nonmodal solutions is possible despite uniform modal decay. In particular, a stochastically forced nonnormal system is capable of sustaining variability at higher levels and of a different character than would be predicted by a traditional eigenmode analysis. Generalized stability analysis furnishes the characteristics of perturbation growth over all time scales, including both nonmodal and modal solutions, and therefore gives a more complete picture of the response of linear systems to forcing. Excellent reviews of generalized stability theory may be found in Farrell and Ioannou (1996) and Moore et al. (2004).

Transient growth of perturbations is possible in nonnormal systems because of the possibility for linear interference between the eigenmodes. Even when all eigenmodes decay exponentially, a sum of nonorthogonal eigenmodes can grow for a transient period of time if the sum is initially in a destructive linear interference that subsides as the component modes evolve. For example, Aiken et al. (2003) have shown that such linear interference effects can sustain a vortex street in the wake of an island even when the vortex street eigen-
The mode itself is strongly damped. It will be shown for the model considered here that the linear interference between the two sets of ACW modes similarly can result in stronger-than-expected anomaly growth.

Nonnormality also influences the way in which a particular eigenmode may be optimally excited. Here a perturbation is considered optimal if it maximizes growth in a particular norm of interest when allowed to evolve in the TLM. While in a normal system the optimal means of exciting any mode is by perturbing with the mode itself, this result does not generalize to nonnormal systems. In general, a given mode is optimally excited by perturbing the system with the mode’s biorthogonal, that is, the perturbation orthogonal to all other modes (Farrell and Ioannou 1996). The desired biorthogonality relation exists with the eigenmodes of the adjoint operator defined in the chosen norm, commonly called the adjoint modes. Thus, maximum growth of a particular mode as measured by the chosen norm is achieved by perturbing the system with the corresponding adjoint mode in that norm. While in a normal system the eigenmode and corresponding adjoint eigenmode coincide, in nonnormal systems they may differ substantially.

The concepts regarding nonnormality and perturbation growth discussed above require the definition of a norm, which is usually quadratic (\( \|u\| = u^*Xu \), positive definite, where the superscript * represents the complex conjugate transpose), with which to quantify perturbation amplitude. That is, the system’s nonnormality and potential for supporting transient perturbation growth depends upon the way in which it is measured. For example, Moore et al. (2002) give an example of a geophysical system that is normal with respect to perturbation energy, but nonnormal with respect to enstrophy. For time-independent operators it is always possible to find a norm that renders the eigenmodes orthogonal, but most interest lies in whether growth occurs in physically meaningful norms. Typical quadratic norms chosen to measure geophysical state vectors include energy, enstrophy, and SST squared. The system considered here contains both dynamic and thermal components, so we define a norm that combines the two. For this purpose it is necessary to define

![Fig. 2. Phase structure and relative amplitudes of the two wavenumber-1 ACW modes from the discretized system: (a) atmospheric streamfunction (m² s⁻¹), (b) oceanic streamfunction (m² s⁻¹), (c) atmospheric temperature (°C), and (d) SST (°C). The solid and dashed curves correspond to the type-1 and type-2 ACW modes, respectively. Note that \( \psi_o \) for the type-2 ACW mode has been increased by a factor of 100 in order for the phase to be apparent.](image-url)
a quadratic version of the thermal energy \( c_p \rho HT^2 / \dot{T} \), where \( c_p \) is the specific heat capacity for constant temperature, pressure, and salinity, \( \rho \) is the density, \( H \) is the layer thickness, and \( \dot{T} \) is a scale for the temperature \( T \). Hence, a quadratic energy norm can be defined as

\[
\|u\|_E = \sqrt{\left[ (\partial_x \psi_u)^2 + (\partial_y \psi_u)^2 \right] \rho_o / 2 + \left[ (\partial_x \psi_\omega)^2 + (\partial_y \psi_\omega)^2 \right] \rho_o / 2 + c_{po} \rho_o \dot{H} T^2 / \dot{T} u^2 + c_{po} \rho_o D_o T^2 / \dot{T} o dx.}
\]

This will be referred to throughout simply as the energy norm, although it should be noted that it is a sum of the kinetic energy and the size of the heat anomaly in the atmosphere and ocean, and thus differs from the standard energy norm that sums kinetic and potential energy components.

The energy norm (7) may be summarized in matrix notation as

\[
\|u\|_E = u^* Xu = u^* M^* Mu = e^* e.
\]

where the norm-defining positive definite matrix \( X = M^* M \), and the “generalized velocity” \( e = Mu \). Thus, the energy norm of \( u \) is equivalent to the L2 norm of \( e \), and as such it is convenient to transform (6) to be in terms of \( e \),

\[
e(t) = Mu(t) = MRu(0) = MRM^{-1} e(0) = Be(0).
\]

Perturbation growth is then determined by

\[
\frac{\|u(t)\|_E}{\|u(0)\|_E} = \frac{\|e(t)\|}{\|e(0)\|} = \frac{e^*(0) B^* Be(0)}{e^*(0) e(0)}.
\]

From this it may be seen that growth of some anomaly \( u \) over time \( t \) depends on the operator \( B^* B \). The leading eigenvectors of this operator are known as the optimal perturbations because they yield the perturbations that grow most rapidly over the time period \( t \). For the norm defined above they represent the perturbations that undergo the largest change in total thermal plus kinetic energy over the chosen time period.

5. Nonnormality

Effects of nonnormality upon a system are also apparent in its response to harmonic forcing. The size of this response, as measured by the energy norm, is bounded above by the norm of the resolvent \( R(\omega) = (i \omega - MAM^{-1})^{-1} \), where \( \omega = \sqrt{-1} \), \( \omega \) is a complex forcing frequency, and \( I \) is the identity matrix. Contours of \( \|R(\omega)\| \) also define the boundaries of \( e \) pseudospectra (Trefethen et al. 1993). It may be shown that

\[
\frac{1}{\text{dist}(\omega, \Sigma)} \leq \|R(\omega)\| \leq \frac{\kappa}{\text{dist}(\omega, \Sigma)},
\]

where \( \text{dist}(\omega, \Sigma) \) represents the least distance between the forcing frequency \( \omega \) and the set \( \Sigma \) of eigenvalues \( \sigma \), and \( \kappa \) is the condition number of the matrix with the eigenmodes as its columns. For normal systems \( \kappa = 1 \), and (11) reduces to the equality giving the familiar resonant response. Nonnormality, however, implies \( \kappa > 1 \), with the result that a greater response to harmonic forcing is possible in nonnormal systems. Thus, the difference between contours of \( R(\omega) \) and \( \text{dist}(\omega, \Sigma)^{-1} \) highlight in which regions of frequency space the enhancement of the response resulting from nonnormality is greatest.

The solid curve in Fig. 3a shows \( \|R(\omega)\| \) as a function of imaginary \( \omega \) for \( \text{Real}(\omega) = 0 \), representing the maximal growth factor in the energy norm for harmonic forcing of the system at frequency \( \text{Imag}(\omega) \). The \( \text{dist}(\omega, \Sigma)^{-1} \) resonant response is plotted in Fig. 3a as the dashed curve, although it is multiplied by \( 10^4 \) in order to be visible on the same scale. The difference in shape between the solid and dashed curves demonstrates that maximum growth does not correspond to the frequencies of the least damped of the ACW modes, which were of low wavenumber and frequency, but instead to the frequencies in the approximate band \( 1 \times 10^{-7} \) to \( 2 \times 10^{-7} \) s\(^{-1}\), corresponding to ACW mode wavenumbers 4–10. The ratio of \( R(\omega) \) to \( \text{dist}(\omega, \Sigma)^{-1} \) is plotted in Fig. 3b for the same domain as Fig. 3a, representing the factor by which the response can increase above pure modal resonance because of the system’s nonnormality as a function of forcing frequency. It can be seen that the response is a number of orders of magnitude greater than can be explained by modal resonance for all frequencies, but is greatest in particular for frequencies corresponding to the higher-zonal-wavenumber ACW modes.

The enhanced response in energy at higher wavenumbers can be linked to the fact that these modes have a greater degree of nonnormality in the energy norm than the lower-wavenumber least-damped ACW modes. The degree of nonnormality of each eigenmode may be quantified by recalling that each adjoint mode \( (r_k) \) represents the structure that is orthogonal in the defined norm to all but its corresponding mode \( (s_k) \). Therefore, a mode’s projection onto its adjoint mode quantifies its orthogonality with respect to all other modes, with the result that modal nonnormality may be measured through

\[
v_k = \frac{|r_k^* s_k|}{r_k^* X s_k}
\]

(12)
(Farrell and Ioannou 1999; Moore et al. 2002). If the \( r_k \) are normalized by \( r_k^2 X_k \), then \( \nu_k \) has the interpretation of being equal to the secant of the angle between each mode and its corresponding adjoint mode. The adjoint modes for the norm \( X = M^T M \) may be calculated through eigenanalysis of the operator \( C = M^{-1} A M^T \), where \( A^T \) represents the adjoint operator in the L2 norm. In this case, because the tangent linear operator \( A \) is a real matrix, \( A^T = A^\dagger \).

In Fig. 3c \( \nu_k \) is plotted for the set of ACW modes shown in Fig. 1b as a function of modal frequency \( \text{imag}(\sigma_k) \) for the energy norm. This figure shows that the ACW modes with the greatest nonnormality are those with higher frequencies corresponding to zonal wavenumbers 11–16. Comparison with Fig. 3b shows that the enhanced response to harmonic forcing correlates closely with the nonnormality of the mode closest to the particular forcing frequency. Both maximum

Fig. 3. (a) The norm of the resolvent \( R(\omega) \) (solid) and \( 1 / \text{dist}(\omega, \Sigma) \) (dashed) for harmonic forcing at frequency \( \text{imag}(\omega) \). The latter has been multiplied by \( 10^4 \) for comparison on the same axes. (b) Ratio of \( \| R(\omega) \| \) to \( 1 / \text{dist}(\omega, \Sigma) \) for the domain in (a). (c) Value of \( \nu \) vs eigenfrequency for the modes whose eigenvalues are shown in Fig. 1b. (d) Projection of global optimal onto the ACW modes as a function of eigenfrequency.
nonnormality and maximum nonresonant response occur for the ACW mode of wavenumber 14, corresponding to the point at which the type-1 and type-2 branches in Fig. 1b cross. However, these very high wavenumber, most-nonnormal ACW modes are kept from dominating the response to harmonic forcing shown in Fig. 3a by their relatively heavy damping. The frequencies corresponding to ACW modes of wavenumbers 4–10 dominate $|R(\omega)|$ in Fig. 3a because they combine relatively high nonnormality (Fig. 3c) with relatively low damping (Fig. 1b).

6. Optimal growth of ACWs

As discussed in section 4, the optimal way to excite a particular eigenmode is by perturbing the system with the corresponding adjoint mode as defined in the norm of interest. The adjoint modes in the energy norm of interest here were found to have the same zonal wavenumber and phase as their corresponding ACW mode, but with amplitude predominantly confined to the ocean streamfunction. That is, the optimal means to achieve maximal growth in the energy norm for each of the ACW modes is to perturb ocean streamfunction predominantly.

Perturbing the system with the adjoint mode ensures maximal growth of the corresponding eigenmode as measured by the energy norm. However, in general perturbations that grow more rapidly in this norm over transient time scales can exist. The maximum perturbation growth possible in the energy norm $|R(0, t)|_E$ is given by the largest singular value of $MR(0, t)M^{-1}$, corresponding to the growth rate of the optimal perturbation in energy for time $t$. In Fig. 4 $|R(0, t)|_E$ is plotted for optimization times up to 100 yr. The figure reveals that, despite the fact that all modes decay, perturbations exist that can grow by up to four orders of magnitude and over interannual time scales in the energy norm.

The optimal perturbation that yields maximal growth over any time scale, called the global optimal, occurs for $t \approx 2$ yr. For all optimization time scales the structure of the optimal perturbation is always largely confined to the ocean streamfunction, that is, representing the assignment of almost all energy to perturbing the oceanic momentum field. It may be recalled from Fig. 3 that the largest response to harmonic forcing coincided with the most nonnormal ACW modes, which were of zonal wavenumbers 4–10. Consistent with this, maximal transient growth in the energy norm is found to occur by forcing with high zonal wavenumbers. For example, the global optimal has zonal wavenumber 6, which is the wavenumber of the ACW mode that has the maximum response to harmonic forcing in Fig. 3a.

It can be noted that the forms of the optimal perturbations closely resemble that of the adjoint ACW modes, in that both to a large degree represent perturbations to the ocean streamfunction. This suggests that the optimal growth corresponds to the optimal excitation of individual ACW modes. To illustrate this, Fig. 3d shows the projection of the global optimal $p$ onto the ACW modes $s_k$, given by $p^T X s_k / ||p|| ||s_k||$. The figure shows that the global optimal projects onto two separate eigenmodes in nearly equal measure in the energy norm. This result is true of the optimal perturbations for all optimization times. The two eigenmodes in each case are in fact the type-1 and type-2 ACW modes discussed in section 3 of the same zonal wavenumber as the perturbation.

The rapid growth of the optimal perturbations in the energy norm may be understood geometrically as resulting from the linear interference between these two modes. The fact that these two modes are very similar a priori suggests the possibility for linear interference between them; a perturbation composed of the differ-
ence of these two modes will have a very small amplitude, but can experience rapid growth over time as they evolve out of phase and the more greatly damped of the modes decays to leave the less damped ACW mode. Recalling from section 2 that the main difference between the type-1 and type-2 ACW modes was in the amplitude of ocean streamfunction, it may be visualized how a difference of the these modes may be concealed in an ocean streamfunction perturbation such as the optimal perturbations.

The form of the optimal perturbations in the energy norm may also be readily understood by considering the sources of perturbation energy in the system. Moore et al. (2002) demonstrate that a system’s non-normality relative to the energy norm may be linked to the presence of sources of perturbation energy in the system. In this model, extraction of thermal energy through advection of the mean temperature fields in the atmosphere and ocean represents the only net source of perturbation energy. There is no mechanism for extraction of mean state kinetic energy, so growth in perturbation kinetic energy can occur solely through the transfer of the thermal energy released from the mean state into the perturbation temperature fields. The large heat capacity of the ocean ensures that advection of the mean ocean temperature field is the dominant mechanism for altering perturbation energy over all time scales. An initial perturbation to the ocean streamfunction can induce a large increase in the perturbation energy through the creation of SST anomalies, signifying the extraction of thermal energy from the mean temperature field. The anomalous SST then in turn generates perturbation kinetic energy in the atmosphere and ocean through their response to thermal expansion. The generalized stability analysis suggests that this process is most efficient at zonal wavenumbers 4 and higher. Wavenumber 4 is often prominent in mid-latitude Southern Hemisphere variability (Kiladis and Mo 1998) and has been linked to interannual SST variability in the South Atlantic, South Indian, South Pacific, and Tasman Sea (Fauchereau et al. 2003; Hermes and Reason 2005).

The above demonstrates that large transient increases in anomalous SST are possible in the model, despite the absence of exponentially growing anomalies. It is also of interest to know whether energy growth is possible in all variables, not just SST. As discussed earlier, the energy norm used here is strongly weighted toward variations in SST because of the large heat capacity of water. However, it is possible to modify the optimization problem to search for perturbations that maximize the growth in energy in one variable relative to the perturbation’s initial energy. This may be done by applying a projection operator $H$ (Buizza and Palmer 1995) to the left side of the expression for $e$ in (9). That is, the optimization problem becomes the calculation of the leading singular vectors of $HMR(0, t)M^{-1}$. In this case $H$ is a matrix with 1 on diagonals corresponding to the chosen variable and 0 elsewhere. Performing this process for each variable revealed that significant and sustained perturbation energy growth is possible in all variables except for ocean streamfunction. No perturbations exist that result in a net growth of ocean kinetic energy over any time scale. For each variable, the optimal way to increase its energy is with a perturbation confined largely to the ocean. That is, perturbing the ocean circulation is the optimal way to increase both perturbation kinetic and thermal energy, even though oceanic kinetic energy never grows in the process.

It is similarly possible to alter the optimization problem to maximize the final perturbation energy relative to the initial energy of a single variable by applying the projection operator to the right side of (9). Calculation of the optimals in this case demonstrates that energy growth is only possible by initially perturbing the flow field in the ocean or atmosphere. Perturbing either temperature field results in decay in the energy norm on all time scales. Atmospheric perturbations, however, generate only a modest increase in energy, and this occurs only on time scales of days.

The above demonstrates that perturbations to the oceanic momentum field can induce rapid growth in SST and SAT anomalies over time scales of 4 yr in the circumpolar system. This is a result of the fact that a small addition of kinetic energy to the ocean can extract a large amount of thermal energy from the mean SST field. The large thermal energy anomaly, however, does not result in the transfer of even greater kinetic energy to the ocean momentum field, reflected in the fact that all modes are damped and initial perturbations in thermal energy decay monotonically. In practice, the magnitude of growth seen will depend on the characteristics of the forcing, and whether or not they encourage the transient growth mechanism. This point is considered in the following section.

7. Stochastically forced anomalies

The analysis presented above suggests the manner in which the system will respond to being impulsively perturbed. In practice, however, the southern ocean–atmosphere system is continuously externally forced with a broad spatial and temporal spectrum whose exact character may determine which anomaly growth mechanisms contribute most to the system variance. In this
section the system’s response to various stochastic forcing scenarios is investigated, with the goal of determining whether realistic interannual variability can be maintained in the model under realistic forcing.

The forced system is given by

\[ x_{n+1} = x_n + A x_n \delta t + \xi_n (\delta t)^{1/2}, \]

where \( x_n \) is the state vector at time step \( n \), \( \delta t \) is the time step, taken here to be 6 h, and \( \xi_n \) contains the spatial structure of the forcing at time step \( n \). Note that the multiplication by \( (\delta t)^{1/2} \) in place of \( \delta t \) in (13) is necessary in order to preserve the variance properties of the time-integrated white-noise (Weiner) process (Gardiner 1985). In each forcing experiment \( x_0 \) was the zero vector, and an ensemble of 100 simulations were run, each of 1000-yr duration. No temporal or spatial filter was applied to the time series. As a first estimate of a realistic forcing, \( \xi \) was taken to be both spatially and temporally uncorrelated. This forcing has a normal distribution with zero mean and unit standard deviation. Note that because the system is linear the response is qualitatively independent of forcing amplitude.

Figure 5 displays the zonal wavenumber spectra of the response of \( T_o \) (modeled SST anomaly) to adding the stochastic forcing to each variable independently as described above. Only results for \( T_o \) are shown, with this being the model variable that is most easily comparable with the ACW observations. Figure 5 shows that, for forcing of \( T_a \) or \( T_o \), the power spectral density (PSD) decreases with zonal wavenumber, but when \( \psi_a \) or \( \psi_o \) are forced the PSD is distributed among higher zonal wavenumbers, predominantly wavenumbers 4–8. Spectra for the other variables are qualitatively similar under each forcing, with the key differences occurring for atmospheric temperature and streamfunction, which respond at even higher wavenumbers when either one of them is forced. It may be noted that neither any of the distributions shown in Fig. 5, nor any sum of them, suggest dominance of wavenumber 2 or 3, as characterizes the ACW.

The distribution of model variance also differs from what would be expected for a stochastically forced “normal” system (one with orthogonal eigenmodes), where the variance maintained in each zonal wavenumber would be expected to be proportional to the decay rate of the corresponding ACW mode. For this system, this would suggest a rapid decay in power spectral density with increasing zonal wavenumber. While the response of \( T_a \) to the stochastic forcing of \( T_o \) does resemble this expected distribution, \( T_o \) displays a broader response, and forcing of \( \psi_a \) or \( \psi_o \) results in SST variance residing at higher zonal wavenumbers than would be expected of a normal system. In the case of \( T_o \) (Fig. 5d), the broadening of the spectrum may be understood as being a sum of the dominantly low-wavenumber normal response and the broadband direct forcing of SST.
The SST response seen to the forcing of $\psi_h$ and $\psi_v$, (Figs. 5a and 5b), however, demonstrates that higher-zonal-wavenumber modes are being sustained through non-normal processes. As a result, the model’s response to forcing of atmospheric or oceanic momentum fields may be understood from the generalized stability analysis presented in sections 5 and 6, where the most rapid growth in the energy norm was seen to involve the higher-zonal-wavenumber ACW modes. In particular, the global optimal was seen to be of zonal wavenumber 6, which can be seen to coincide with the peak of power spectral density for forcing with $\psi_h$ or $\psi_v$. This suggests that it is the transient nonmodal perturbation growth that sustains the higher-than-expected variance in the higher-wavenumber ACW modes. The reason that the high-wavenumber dominance is seen most strongly for the forcing of $\psi_h$ and $\psi_v$ can be linked to the finding that the optimal perturbations are structures largely confined to $\psi_h$. Because coupling of either temperature field back to $\psi_h$ is weak, forcing the temperature field excites the optimals only weakly, and as a result the distribution of variance under such forcing is determined largely by the modal decay rates. In contrast, the forcing of $\psi_v$ can generate nonmodal growth because its coupling to $\psi_h$ is relatively strong.

We find that the response resulting from a realistic level of forcing of the ocean momentum field is an order of magnitude greater than that from realistic oceanic surface heat fluxes. Forcing $\psi_v$ at a realistic amplitude, corresponding to rms wind stress curl variations of the order of $10^{-7}$ N m$^{-2}$, sustains dominantly high-wavenumber SST anomalies of the order of 1°C. A stochastically varying oceanic surface heat flux forcing, corresponding to advection of the mean SST gradient by meridional velocity fluctuations of the order of 1 cm s$^{-1}$, however, induces dominantly low-wavenumber SST anomalies of the order of 0.1°C in the model. Thus, the high-wavenumber nonmodal transient growth is the dominant source of variance under realistic forcing amplitudes.

The forcing of $\psi_h$ discussed above may be interpreted as a first-order approximation to forcing the ocean with synoptic-scale “weather” events, having small decorrelation space and time scales relative to the ACW. For example, there exists a number of alternative potential sources of external forcing for the ACW, however, characterized by longer time and space scales. There exists substantial evidence of an atmospheric teleconnection between the ACW (or at least a component of it) and ENSO via the PSA (Cai and Baines 2001), while many studies have investigated the role of the dominant standing mid- to high-latitude southeast Pacific and South Atlantic via atmospheric Rossby wave propagation from the Tropics, and this pattern projects strongly in the Amundsen/Bellinghausen Sea region (Yuan and Martinson 2001) and the South Atlantic (Mo and Paegle 2001; Colberg et al. 2004). There remains some uncertainty, however, as to whether the well-documented standing wavenumber-3 SLP pattern (Karoly 1989) is an independent process to the ACW, or is reinforced to some extent by the propagating SST anomalies; Cai et al. (1999) provide evidence that the standing SLP pattern may be in fact forced by the propagating interannual SST anomalies. The analytical model of BC studied here does not permit the mechanism discussed by Cai et al. (1999), and hence cannot produce atmospheric standing wave anomalies. Other plausible sources of independent forcing include synoptic-scale heat flux variations and anomalous advection of the mean SST gradient by mesoscale eddies, each more properly modeled as having longer time or space scales than considered thus far.

As a result, the dependence of the model response to forcing decorrelation time and space scales was investigated. Red-noise series were generated for a broad range of decorrelation scales using an autoregressive lag-1 model, and were used to force the model ocean streamfunction. Each time series had zero mean and identical root-mean-square variability. An ensemble of fifty 100-yr simulations was performed for each decorrelation space and time scale, and the resulting zonal wavenumber spectra were averaged over the ensemble. The zonal wavenumber power spectral density as a function of decorrelation time and space scale is shown in Fig. 6a and Fig. 6b, respectively. This figure confirms that increasing either the decorrelation time or space scale of the forcing lowers the zonal wavenumber response of the model. In particular, a dominant zonal wavenumber-2 or -3 response in SST, reminiscent of the observed and modeled ACW, occurs for the forcing of $\psi_h$ with decorrelation space scales greater than 10 000 km or for decorrelation time scales greater than 3 yr. This result may be understood from the fact that such a forcing preferentially excites the zonal wavenumber-2 and -3 ACW modes at the expense of the higher-wavenumber modes responsible for the transient growth that was seen to dominate in the pure white-noise forcing case. The necessity of such large-scale forcing in order to produce a dominantly low-wavenumber response in the model suggests the importance of large-scale atmospheric patterns such as ENSO in driving the ACW, consistent with the results of Cai and Baines (2001) and Venegas (2003). The high levels of local smaller-scale variability found in southern mid-
latitudes, resulting from oceanic eddies and atmospheric weather, do not produce dominant interannual modes in this coupled model. This is illustrated further in Fig. 7, which shows Hovmöller diagrams of SST under stochastic forcing of $\psi_k$ with decorrelation time scales of 1 week (Fig. 7a) and 4 yr (Fig. 7b). The data plotted in each panel are unfiltered and amplitudes are arbitrary. Power spectral densities from Figs. 7a and 7b are shown in Figs. 7c and 7d, respectively, as a function of zonal wavenumber, and in Figs. 7e and 7f, respectively, as a function of frequency. Figure 7c confirms that the response shown in Fig. 7a has variability spread across a broad range of wavenumbers, which is a result of the fact that the high-frequency forcing excites the transient nonnormal growth mechanism discussed in section 6. Only when the forcing does not excite these high-wavenumber modes do the low wavenumbers become dominant. This is seen in Fig. 7d, where the 4-yr decorrelation time scale forcing results in the dominance of the low wavenumbers, consistent with the observed ACW.

8. Conclusions

The ability of a simple linearized coupled model to reproduce observed features of Southern Hemisphere interannual variability has been investigated. An earlier analytical study of the model (BC) revealed that its least-damped/fastest-growing eigenmode corresponded well to the observations of the ACW as given in White and Peterson (1996). Here the same model has been implemented on a zonal grid and solved numerically. Eigenanalysis of the discretized model confirms BC’s findings, suggesting the long-term dominance of a gradually decaying ACW-like mode in the system’s response to an initial perturbation, but also uncovers the existence of two distinct ACW-like modes at each zonal wavenumber. It was found, however, that rapid growth of nonmodal perturbations was possible over time scales of years, and that the fastest growing perturbations were of a high zonal wavenumber. This is despite the fact that the corresponding high-zonal-wavenumber eigenmodes are not the least-damped modes in the system. The perturbation growth, as measured in an energy norm combining perturbation kinetic and thermal energies, was understood geometrically as resulting from the linear interaction of the two slightly different ACW-like modes of the same zonal wavenumber. Through calculation of the norm of the resolvent for imaginary frequencies spanning those of the least-damped eigenvalues, this process was shown to significantly affect the frequency response of the system. White-noise stochastic forcing in space and time of either the atmospheric or oceanic momentum field was found to predominantly excite the transient growth mechanism, resulting in a response dominated by zonal wavenumbers 4–8 and frequencies less than 1 yr. Although sharing the ACW phase relationship, these higher-zonal-wavenumber disturbances do not resemble the spatial or temporal character of the ACW. Thermal forcing of the atmosphere or ocean does produce a dominantly low-wavenumber response, but of a smaller magnitude than that resulting from momentum field forcing of typical magnitudes. Thus, an ACW-like response does not result from this coupled model under purely white stochastic forcing of momentum, despite the fact that the least-damped eigenmodes do closely resemble the observed ACW.

To produce a dominant ACW-like response, the forcing had to preferentially excite the least-damped
modes. Stochastic forcing with decorrelation space scales greater than 10,000 km or decorrelation time scales greater than 3 yr, did result in a response with the general wavenumber and frequency characteristics of the observed and modeled ACW. Thus, while forcing of coupled modes by high-frequency, essentially stochastic variability cannot explain the observation of dominant low-zonal-wavenumber anomalies, low-frequency noise forcing can. The purely white forcing projects equally onto all modes, hence yielding maximal response in the most rapidly growing perturbations, which are of a high zonal wavenumber. This is despite the fact that the high-zonal-wavenumber modes are more greatly damped, and can be understood in terms of the higher degree of nonnormality of these eigenmodes in an energy norm than the least-damped, low-zonal-wavenumber modes. The low-wavenumber-coupled ACW modes only dominate when the forcing preferentially

Fig. 7. Hovmöller plots of model SST for stochastic forcing of $\phi$, with decorrelation time scales of (a) 1 week and (b) 4 yr. Data have been scaled to have unit maximum amplitude. Also shown are the corresponding normalized PSDs as a function of (c), (d) zonal wavenumber and (e), (f) frequency.
projects onto them, as is the case for perturbations with equally low-zonal-wavenumbers or low frequencies. This may be contrasted with the case for an uncoupled model of Southern Ocean SST anomalies (Aiken and England 2005), where white-noise stochastic forcing could excite a predominantly low-wavenumber ACW-like signal. It is concluded that, if the ACW is a true propagating coupled mode as simulated by this simple model, it is likely forced by other large-scale low-frequency phenomena such as ENSO and its teleconnections to higher latitudes.

Acknowledgments. This research was supported by the Australian Research Council. The comments of two anonymous reviewers greatly improved the manuscript and are gratefully acknowledged.

REFERENCES