GEOPHYSICAL RESEARCH LETTERS, VOL. 26, NO. 13, PAGES 1885-1888, JULY 1, 1999

Global mean surface air temperature and North Atlantic overturning in a suite of coupled GCM climate change experiments

Keith W. Dixon and John R. Lanzante

NOAA, Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey

Abstract. The effects of model initial conditions and the starting time of transient radiative forcings on global mean surface air temperature (SAT) and the North Atlantic thermohaline circulation (THC) are studied in a set of coupled climate GCM experiments. Nine climate change scenario experiments, in which the effective levels of greenhouse gases and tropospheric sulfate aerosols vary in time, are initialized from various points in a long control model run. The time at which the transition from constant to transient radiative forcing takes place is varied in the scenario runs, occurring at points representing either year 1766, 1866 or 1916. The sensitivity of projected 21st century global mean SATs and the THC to the choice of radiative forcing transition point is small, and is similar in magnitude to the variability arising from variations in the coupled GCM's initial three-dimensional state.

Introduction

Coupled atmosphere-ocean general circulation models (CGCMs) forced with increasing greenhouse gases (GHGs) and the direct effect of tropospheric sulfate aerosols have been found to simulate 20th century surface air temperature (SAT) records better than CGCM scenario experiments that include only GHG forcing [*IPCC*, 1996]. Historical reconstructions and future projections of GHG and sulfate aerosol levels are prescribed in these transient climate change scenario experiments (referred to as GHG+SUL model runs). In GHG+SUL simulations, the tropospheric sulfate aerosols' direct effect (cooling due to the aerosol's reflection of shortwave radiation back to space) can be parameterized by spatially and temporally varying surface albedos [*Mitchell et al.*, 1995]. Yet, uncertainties remain regarding the actual quantitative impact of aerosols and other atmospheric constituents on radiative forcing and the global climate [*Hansen et al.*, 1997].

Since state-of-the-art CGCMs are computationally demanding, researchers are limited in the number and length of climate change scenario runs that can be performed with their most costly models. Computational constraints lead to experimental design compromises, as a balance is sought between CGCM complexity, spatial resolution, and the duration and number of simulations conducted.

A climate change experiment needs to be ~300 years in length to extend from pre-industrial times to the late 21st century. Computational economies can be realized by limiting the length of a transient climate change simulation. For example, computational costs can be halved if one assumes that an equilibrated control model's state represents a time during the early 20th century (rather than pre-industrial conditions), and initiates a transient forcing experiment from that point. Yet, such an approach neglects the fact that GHG levels actually began to increase more than a century earlier. One expects the climate response of such an experiment to be less than that produced by a similarly configured but longer model run that uses a control model state to represent pre-industrial conditions. Initialization techniques of this type can yield a "cold start error" (*i.e.*, a persistent underestimation of a simulated climate response), the magnitude of which can be estimated in some cases using linear response theory [*Hasselmann et al.*, 1993].

Here, we examine a suite of nine climate change experiments generated using a version of a global coupled ocean-atmosphere GCM developed at NOAA's Geophysical Fluid Dynamics Laboratory (GFDL). Results of the earlier version of this model were presented in the Intergovernmental Panel on Climate Change's (IPCC's) scientific assessments [*IPCC*, 1990; *IPCC*, 1996]. It is now feasible to run a suite of climate change experiments with an updated version of this model. The suite of runs is used to study the sensitivity of simulated global mean surface air temperature (SAT) and the North Atlantic thermohaline circulation (THC) to the choice of initial conditions and radiative forcing time histories (*i.e.*, the length of the transient simulation).

Model Description and Experimental Design

The relatively coarse resolution version of the GFDL coupled climate model used here is very similar to the CGCM of *Manabe et al.* [1991]. The earlier version's control state and its responses to transient climate forcings have been reported previously [*e.g.*, *Manabe et al.*, 1991; *Delworth et al.*, 1993; *Manabe and Stouffer*, 1994; *Haywood et al.*, 1997]. A brief model description is offered below, including mention of some differences between the earlier model version and the one used here.

The coupled model incorporates GCMs of the global atmosphere and ocean with simpler sea ice and land hydrology models. The atmospheric spectral model uses rhomboidal 15 truncation and has 9 vertical levels. The atmospheric model yields an annual mean, global mean, top of the atmosphere net radiative flux imbalance of less than 1 Wm^{-2} when current SST climatologies are imposed as lower boundary conditions. Soil moisture calculations are performed using a 0.15m bucket hydrology model, with each grid point's runoff assigned to a river drainage basin according to observations (the earlier model version routed runoff following the slope of the model's spectral topography).

The global ocean model is based on the GFDL Modular Ocean Model version 1 [*Pacanowski et al.*, 1991]. It has ~4.5° latitude by 3.75° longitude grid spacing, 12 vertical levels and a maximum depth of 5000m. The sea ice model predicts sea ice thickness, but not fractional coverage, and its horizontal grid duplicates the ocean model's. Fourier filtering of sea ice, salinity and oceanic potential temperature occurs along grid rows poleward of ~68° latitude, and affects a much smaller area than in the *Manabe et al.* [1991] model.

The coupled GCM is driven by seasonally varying insolation entering the top of the atmosphere, without a diurnal cycle nor any interannual variations. Once per day, fluxes of heat, freshwater and momentum are communicated between the model's atmosphere-

Paper number 1999GL900382.

land surface component and the ocean-sea ice component. To reduce climate drift, heat and water flux adjustments are added to the ocean model's surface (see *Manabe et al.* [1991] for a description). Although it is preferable not to use flux adjustments, it is encouraging that an intercomparison of CGCMs found similar surface temperature sensitivities to imposed GHG increases in models with and without flux adjustments [*Kittel et al.,* 1998].

In the nine GHG+SUL experiments examined here, historical reconstructions of scaled, time-varying GHG levels are prescribed until year 1990. After 1990, a 1% per year compounded increase in effective CO_2 levels is imposed. Temporal and spatial variations of surface albedos associated with the tropospheric sulfate aerosols' direct effects are implemented as in *Haywood et al.* [1997].

The control model's effective atmospheric CO_2 concentration, representing the combined effects of the GHGs (CO_2, CH_4, N_2O, halocarbons, etc.), is 360ppmv - higher than that used by *Haywood, et al.* and higher than the actual pre-industrial value of ~280 ppmv. That the control model's CO_2 level differs from pre-industrial values is addressed in GHG+SUL model runs by multiplicative scaling of the transient effective CO_2 levels. This scaling is appropriate to the extent that a change in the net radiative flux (ΔF) for a CO_2 change from C_0 to C can be computed after *Shine et al.* [1990] as $\Delta F = 6.3 \ln (C/C_0)$.

The organization of the model experiments is depicted schematically in Figure 1. The three GHG+SUL experiments initialized with radiative conditions representing 1 January 1766 (labelled A1766, B1766 and C1766 in Fig. 1) experience a smooth transition from the control model's forcings. However, abrupt changes in GHGs and tropospheric sulfates are felt at the start of the other six, shorter scenario experiments. Effective CO_2 levels instantaneously increase 6% when going from the control to 1866 conditions, and by 11% when going from the control to 1916 conditions.

In the following analyses, time series of global mean SAT (the temperature of the atmospheric GCM's lowest level) and THC index (maximum oceanic overturning value for the model's North Atlantic) are examined. These large-scale climatic variables are often studied in climate change experiments, and are presumed to have different response time scales (SAT responds faster than the THC). Results from the control and GHG+SUL model runs are compared to one another in various combinations.

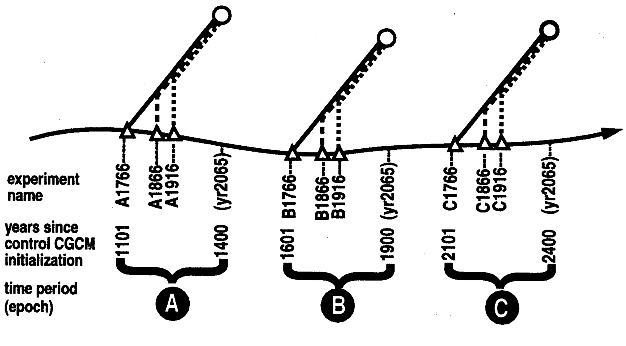

Figure 1. One long running CGCM control run (horizontal line) and nine GHG+SUL runs are depicted schematically. The GHG+SUL experiments are conducted in three groups of three (model epochs A, B, and C). Successive epochs are separated by 500 model years. Within each epoch, scenario runs of 300, 200, and 150 years duration were initialized from the control run, with the transition to prescribed, time-varying GHG and tropospheric sulfate forcings simulating starting points of historical years 1766, 1866 and 1916, respectively (ascending solid, dashed and dotted lines). Initialization points (triangles) were staggered so all runs in each epoch completed year 2065 (circles) on the same time step.

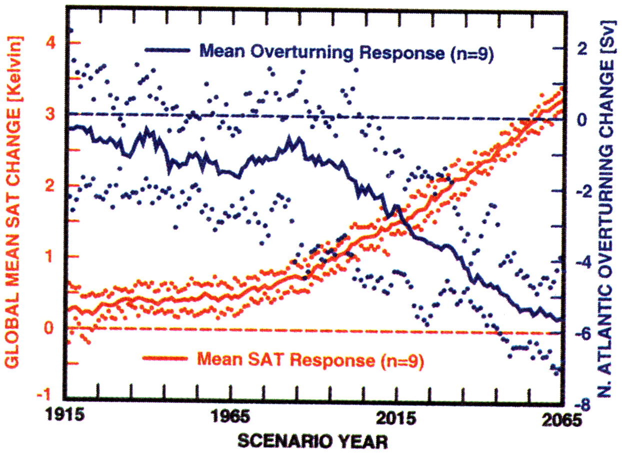

Figure 2. Time-dependent responses of annual mean, global mean SAT (red; scale on left) and North Atlantic THC index (blue; scale on right). Solid lines show the mean response of the nine GHG+SUL model runs. Color coded circles represent year-by-year maxima and minima of the nine simulations. Dashed horizontal lines represent zero response relative to the control run.

While both the number of experiments and variables examined here are limited, the experimental design permits investigation of two questions relevant to the design and analysis of coupled GCM climate change experiments. First, to what extent do the results of shorter, less computationally expensive GHG+SUL model runs differ from those produced by longer runs? And secondly, how might the model's response to GHG+SUL forcings depend upon the choice of the coupled GCM's initial three-dimensional state?

Results and Discussion

The control model exhibits little climate drift during the period studied. Least-squares regression lines have been fit to the control run's 1300 year period extending from model years 1101 to 2400 (the period encompassing all three epochs of GHG+SUL runs). The line fit to the annual mean, global mean SAT time series has a slope of only -0.009K century^{-1} (cooling occurs in both the Northern and Southern Hemispheres). The trend in the annual mean THC index time series is -0.093 Sv century^{-1} ($1\text{Sv} = 10^6 \text{m}^3\text{s}^{-1}$). If large trends were present in the control run, the assumption that the nine CGCM states used to initialize the GHG+SUL simulations represent the same climatic regime, and therefore could be expected to yield similar responses to transient climate forcings, would have been in doubt. Although the overall climate drift is slight, sizable amounts of internally generated variability are evident in the model at many time scales, including multi-decadal to century time scales.

Figure 2 depicts the ensemble mean SAT and THC responses computed from all nine GHG+SUL model runs. The range of responses found among the ensemble elements is also indicated. The time varying "response" of a GHG+SUL simulation is computed by subtracting a time series of a control run segment's 30-year centered running means from a corresponding GHG+SUL experiment's time series of annual means. By year 2065, the GHG+SUL experiments' global mean SATs average ~3.3K warmer than the control and the North Atlantic THC has weakened on average by ~40%. Compared to their respective climate change responses (signals), more unforced variability (noise) is evident in the THC index time series than in the SAT time series.

Non-parametric tests are used to determine if the projected climatic response (SAT and THC) of certain experiments or groups of experiments differ significantly from one another. The method of *Ebisuzaki* [1997] is used to generate random time series for use in Monte Carlo testing. The control model's SAT and THC time series from years 1101 to 2400 are used as input to generate 500 random time series that retain the same amplitude (*i.e.*, spectral power) at each Fourier frequency but are randomized in phase. Monte Carlo tests performed by sampling these synthesized time series provide estimates of statistical significance levels, assuming that the frequency characteristics of variability in the scenario experiments are similar to those found in the control integration.

Testing is performed looking at the entire 150 years (1916-2065) for which all nine GHG+SUL model runs provide output, and by dividing the time series into two 75 year long periods. The SAT (or THC index) response values of all nine scenario experiments are ranked from one (lowest value) to nine (highest value) for each year. The average rank over the period of interest is then calculated and compared to similarly computed Monte Carlo test results. In utilizing the ranks, rather than the actual SAT or THC response values, we diminish the effects of outlying values.

For purposes of statistical testing, the nine simulations are grouped into three groups of three experiments in two different ways. First, by grouping the experiments by radiative forcing scenario (*i.e.*, the three *x*1766 model runs vs. *x*1866 runs vs. *x*1916 runs) we can determine whether the computational savings associated with the shorter simulations result in significantly different climate responses. Secondly, grouping the simulations by the part of the control run they were initialized from (*i.e.*, epoch A vs. epoch B vs. epoch C) tests the sensitivity of model results to the period of the control run used for initial conditions. Results of the global mean SAT average rank tests computed using these two methods of grouping the simulations appear in Table 1. Similarly, North Atlantic THC index results are presented in Table 2.

Having experienced 150 years of slowly increasing radiative forcing, it is perhaps not surprising that the trio of longest running scenario experiments (*x*1766) have the highest (warmest) SAT average rank during the first 75 years of comparison (significant at the 90% confidence level). The shortest runs (*x*1916) have the lowest average rank, and the *x*1866 runs' average ranks are intermediate during this period. However, none of SAT time series' average ranks differ significantly from 5.0 during the final 75 years (scenario years 1991 to 2065) when grouped by radiative forcing. The *x*1766 model runs' ensemble mean SAT response for the latter period is only 0.04K warmer than that of the *x*1866 model runs, and 0.02K warmer than the *x*1916 runs. (For comparison, the *x*1766 ensemble mean warmed ~0.1K by year 1866 and ~0.3K by 1916.) These results suggest that biases introduced by initiating the transient forcing at 1866 or 1916 rather than 1766 become indistinguishable from the noise of internal variability when considering this model's 21st century global mean SAT projections.

Table 1: Global Mean SAT Average Rank Results

Group (n=3 in each)	Avg. Rank 1916-1990	Avg. Rank 1991-2065	Avg. Rank 1916-2065
Grouped by forcing			
*x*1766	6.00*	5.51	5.75*
*x*1866	4.80	4.44	4.62
*x*1916	4.20	5.05	4.63
Grouped by epoch			
Epoch A	5.02	3.79	4.40
Epoch B	5.28	4.66	4.97
Epoch C	4.70	6.55**	5.63

Null hypothesis is that avg. rank = 5.0. Significance levels of two-tailed tests indicated as: * = 0.10; ** = 0.05; *** = 0.01. In the Monte Carlo tests, epoch group variances are greater than forcing group variances. This is because the same control segment is subtracted from all 3 runs in an epoch group, while a different control segment is subtracted from each of the forcing group runs. So, to be statistically significant, an epoch group's avg. rank must differ from 5.0 more so than is the case for a forcing group.

Table 2: North Atlantic THC Average Rank Results

Group (n=3 in each)	Avg. Rank 1916-1990	Avg. Rank 1991-2065	Avg. Rank 1916-2065
Grouped by forcing			
*x*1766	5.09	5.24	5.16
*x*1866	4.29	3.83	4.05
*x*1916	5.62	5.93	5.78
Grouped by epoch			
Epoch A	3.61	3.53	3.57
Epoch B	7.06*	4.98	6.02
Epoch C	4.33	6.48	5.41

Same null hyphothesis and significance level indicators as in Table 1.

During the period 1916 to 1990, little difference exists among the SAT responses for different epochs. But, a pattern emerges during the final 75 years studied - a period of more rapidly changing radiative forcing and SAT warming. The average rank results show that the later the initialization epoch, the greater the warming response (epoch C warms 0.13K more than epoch A and 0.08K more than epoch B). One may speculate that the control model's cooling trend leads to an increased SAT response as one moves from epoch A to epoch B to epoch C, due to enhanced snow and ice albedo feedbacks; however, other temporal changes in the control model also could play a role in generating this apparent sensitivity.

To achieve statistical significance, the North Atlantic THC time series' average rank results shown in Table 2 must deviate further from 5.0 than is the case for the SAT average ranks. This is due to the greater amount of internally generated low frequency variability present in the control model's THC time series. Note that when interpreting the THC results, the lower the rank the greater the response to the transient climate forcing (more THC weakening).

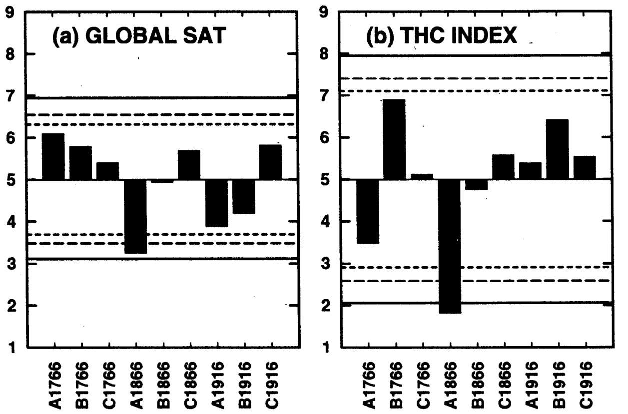

Figure 3. Average ranks for each of the nine scenario experiments' (a) global mean SAT and (b) THC index time series computed over the period 1916 to 2065. Confidence levels for two-tailed tests: short dashed line = 90%, long dashed line = 95%; solid line = 99%.

When grouped by radiative forcing, no statistically significant differences are found among the THC average rank tests. When grouped by epoch, one of the THC average rank results differs significantly from 5.0 at the 90 percent confidence level. During years 1916 to 1990 of epoch B, average ranks are higher than expected (stronger THC). This may be the result of multi-decadal variability that produced unusually weak meridional overturning during this segment of the control experiment - exposing a potential sensitivity to the choice of defining the response of the climate change experiments as the deviation from 30-year running means, as opposed to using a longer term measure of the control. That ensemble responses can depend upon the method of computing a climatic response has been discussed by *Cubasch et al.* [1994].

Results of the average rank tests computed over 150 years for each of the nine scenario experiments are presented in Figure 3. Experiment A1866 is the only simulation that yields results that are statistically significant at the 95% confidence level. In the A1866 model run, the North Atlantic THC is significantly weaker and the global mean SAT also warms more slowly than in the other simulations. Had the A1866 run been the only experiment conducted, one might surmise that the onset of rapid THC weakening occurs approximately two decades prior to that simulated by the other eight experiments performed using the same CGCM.

Conclusions

One motivation for this study is to determine whether it is important to start greenhouse scenario experiments from a "pre-industrial" state, or if shorter integrations using a control model's equilibrated state to represent mid-19th or even early 20th century initial conditions, are adequate. In the suite of nine coupled GCM experiments studied here, simulations of 21st century global mean surface air temperature and the strength of the North Atlantic thermohaline circulation are found to be relatively insensitive to the time when the transition to transient radiative forcing occurs (either 1766, 1866 or 1916). Model-projections for these two large-scale climatic variables indicate that the influence of the choice of radiative forcing transition point is small, and is similar in magnitude to the variability arising from initializing the scenario experiments from different portions of a long, relatively drift-free, control model run. After 75 years, the signal introduced by using different radiative forcing transition points becomes lost in the noise of the model's internally generated variability.

Of course, these results are specific to the model used and variables examined here. Since both the climatic response (the signal) and the amount of low frequency, internally generated variability (the noise) affect the analysis, more significant differences may be found in other model variables with longer response times (*e.g.*, sea level) and less internally generated variability. *Keen and Murphy* [1997] have studied some aspects of this problem. Our results are consistent with those of *Fichefet and Tricot* [1992], who found little difference in 21st century SAT projections produced when starting coupled model GHG scenario runs as late as 1960.

The existence of an outlier simulation (experiment A1866) highlights the value of producing an ensemble of model simulations, rather than relying on a single model run to characterize a given model's response to transient climate forcing. Ensembles also provide additional value for conducting regional analyses, climate change fingerprinting detection studies, etc.

To the extent that these results are applicable to global mean SAT and the North Atlantic THC in other models, two points regarding the design of climate change experiments are suggested: (1) sufficient sensitivity to the choice of the control simulation's ocean-atmosphere initial conditions exists so that interpretations, such as when a climate change signal emerges from the noise, deduced from a single experiment should be viewed with caution, and (2) performing two climate change experiments with transient forcing starting in the early 20th century can be a computationally and scientifically viable alternative to conducting a single, longer experiment that has a completely smooth transition from pre-industrial radiative forcing.

Acknowledgments: We benefited from discussions with Jeff Anderson, Alex Hall, Isaac Held, Tom Knutson, Suki Manabe and Ron Stouffer. The comments of external reviewers were greatly appreciated.

References

Cubasch, U., B.D. Santer, A. Hellbach, G. Hegerl, H. Höck, E. Maier-Reimer, U. Mikolajewicz, A. Stössel and R. Voss, Monte Carlo climate change forecasts with a global coupled ocean-atmosphere model, *Clim. Dyn., 10,* 1-19, 1994.

Delworth, T.L., S. Manabe, and R.J. Stouffer, Interdecadal variations of the thermohaline circulation in a coupled ocean-atmosphere model, *J. Climate, 6,* 1993-2011, 1993.

Ebisuzaki, W., A method to estimate the statistical significance of a correlation when the data are serially correlated, *J. Climate, 10,* 2147-2153, 1997.

Fichefet, T. and C. Tricot, Influence of starting date of model integration on projections of greenhouse-gas-induced climatic change, *Geophys. Res. Let.*, 19, 1771-1774, 1992.

Hansen, J., M. Sato and R. Ruedy, Radiative forcing and climate response, *J. Geophys. Res., 102D,* 6831-6864, 1997.

Hasselmann, K., R. Sausen, E. Maier-Reimer and R. Voss, On the cold start problem in transient simulations with coupled atmosphere-ocean models, *Clim. Dyn., 9,* 53-61, 1993.

Haywood, J.M., R.J. Stouffer, R.T. Wetherald, S. Manabe and V. Ramaswamy, Transient response of a coupled model to estimated changes in greenhouse gas and sulfate concentrations, *Geophys. Res. Let., 24,* 1335-1338, 1997.

IPCC (Intergovernmental Panel on Climate Change), *Climate Change: The IPCC Scientific Assessment.* edited by J.T. Houghton, G.J. Jenkins and J.J. Ephraums, Cambridge Univ. Press, Cambridge, UK, 365pp., 1990.

IPCC, *Climate Change 1995: The Science of Climate Change,* edited by J.T. Houghton, L.G. Meira Filho, B.A. Callander, N. Harris, A. Kattenberg and K. Maskell, Cambridge Univ. Press, Cambridge, U.K., 572 pp., 1996.

Keen, A.B., and J.M. Murphy, Influence of natural variability and the cold start problem on the simulated transient response to increasing CO_2, *Clim. Dyn., 13,* 847-864, 1997.

Kittel, T.G.F., F. Giorgi, and G.A. Meehl, Intercomparison of regional biases and doubled CO_2-sensitivity of coupled atmosphere-ocean general circulation model experiments, *Clim. Dyn., 14,* 1-15, 1998.

Manabe, S., R.J. Stouffer, M.J. Spelman and K. Bryan, Transient responses of a coupled ocean-atmosphere model to gradual changes of atmospheric CO_2. Part I: Annual mean response, *J. Climate, 4,* 785-818, 1991.

Manabe, S., and R.J. Stouffer, Multiple-century response of a coupled ocean-atmosphere model to an increase of atmospheric carbon dioxide, *J. Climate, 7,* 5-23, 1994.

Mitchell, J.F.B., T.C. Johns, J.M. Gregory, and S.F.B. Tett, Climate response to increasing levels of greenhouse gases and sulphate aerosols, *Nature, 376,* 501-504, 1995.

Pacanowski, R., K. Dixon and A. Rosati, The GFDL Modular Ocean Model Users Guide version 1, *GFDL Ocean Group Technical Report No. 2,* NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, NJ, 1991.

Shine, K.P., R. G. Derwent, D. J. Wuebbles, and J.-J. Morcrette, Radiative forcing of climate, in *Climate Change: The IPCC, Scientific Assessment,* edited by J. T. Houghton, G. J. Jenkins, and J. J. Ephraums, pp. 41-68, Cambridge Univ. Press, 1990.

K. W. Dixon and J. R. Lanzante, NOAA/GFDL, PO Box 308, Princeton, New Jersey 08542 USA (e-mail: kd@gfdl.gov; jrl@gfdl.gov)

(Received March 25, 1999; revised May 15, 1999;
accepted May 18, 1999.)