660-km density jump), these forces are strong enough to produce partial layering in the mantle. The strength of the buoyancy forces, however, is proportional to the assumed density contrast across the interface. If the density contrast across the 660-km discontinuity is only 5%, as our results suggest, then layered convection is less likely.

References and Notes

1. Many authors, including A. E. Ringwood, *Composition and Petrology of the Earth's Mantle* (McGraw-Hill, New York, 1975); L.-G. Liu, *Nature* **262**, 770 (1976); I. Jackson, *Earth Planet. Sci. Lett.* **62**, 91 (1983); T. Katsura and E. Ito, *J. Geophys. Res.* **94**, 15663 (1989); E. Ito, M. Akaogi, L. Topor, A. Navrotsky, *Science* **249**, 1275 (1990).
2. U. R. Christensen and D. A. Yuen, *J. Geophys. Res.* **90**, 10291 (1985); P. Machetel and P. Weber, *Nature* **350**, 55 (1991); P. J. Tackley, D. J. Stevensen, G. A. Glatzmaier, G. Schubert, *ibid.* **361**, 699 (1993); S. Honda, D. A. Yuen, S. Balachandar, D. Reuteler, *Science* **259**, 1308 (1993); L. P. Solheim and W. R. Peltier, *J. Geophys. Res.* **99**, 15861 (1994); P. J. Tackley, D. J. Stevensen, G. A. Glatzmaier, G. Schubert, *ibid.*, p. 15877; J. Ita and S. D. King, *ibid.*, p. 15919; C. Thoraval, P. Machetel, A. Cazenave, *Nature* **375**, 777 (1995).
3. J. S. Revenaugh and T. J. Jordan, *J. Geophys. Res.* **96**, 19763 (1991).
4. M. P. Flanagan and P. M. Shearer, *ibid.* **103**, 2673 (1998).
5. Y. Gu, A. M. Dziewonski, C. B. Agee, *Earth Planet. Sci. Lett.* **157**, 57 (1998).
6. G. Nolet and M. J. R. Wortel, in *The Encyclopedia of Solid Earth Geophysics*, D. E. James, Ed. (Van Nostrand Reinhold, New York, 1989), pp. 775–788; G. Nolet, S. P. Grand, B. L. N. Kennett, *J. Geophys. Res.* **99**, 23753 (1994).
7. The reflection coefficient is the ratio of the reflected wave amplitude to the incident wave amplitude.
8. *SS* and *PP* are the multiples of the direct *S* and *P* phases that contain a single reflection off Earth's surface midway between source and receiver. The precursors to *SS* and *PP* resulting from underside reflections off a discontinuity at depth *d* are termed *SdS* and *PdP*, respectively. For example, *S410S* indicates the *S* reflection off the bottom of the 410-km discontinuity.
9. P. M. Shearer, *J. Geophys. Res.* **96**, 18147 (1991).
10. C. H. Estabrook and R. Kind, *Science* **274**, 1179 (1996).
11. P. M. Shearer, *J. Geophys. Res.* **101**, 3053 (1996).
12. M. P. Flanagan and P. M. Shearer, *Geophys. Res. Lett.* **26**, 549 (1999).
13. We obtained this factor by computing the effect of time shifts on the amplitude of the *SS* and *PP* reference pulses in the stacks, assuming that the time shifts have a Gaussian distribution characterized by a standard deviation, σ. The time shifts result from differences in the two-way surface-to-discontinuity travel times caused by both discontinuity topography and upper mantle velocity heterogeneity. In general, σ will vary depending on the wave type (for example, *P* or *S*) and the geographic diversity of ray geometries in the stack. For *ScS* waves, $\sigma = 4$ s was obtained for mixed continental and oceanic paths (*3*), whereas $\sigma = 2.5$ s has been estimated for *SdS* waves in purely oceanic regions (*11*). The limited distribution of reflection points contained within each of our source-receiver range bins is likely to decrease σ, whereas the mixture of continental and oceanic points is likely to increase σ. For σ values of 2, 3, and 4 s, *SdS* amplitudes in our stacks are reduced by 8, 16, and 28%, respectively. The time shifts for *PdP* phases are likely to be smaller than those for *SdS*. For σ values of 1.0, 1.7, and 2.5 s, *PdP* amplitudes are reduced by 4, 12, and 23%, respectively. Our amplitude corrections are based on $\sigma = 2.5$ s for *SdS* and $\sigma = 1.7$ s for *PdP*.
14. G. Helffrich and B. J. Wood, *Geophys. J. Int.* **126**, F7 (1996); L. Stixrude, *J. Geophys. Res.* **102**, 14835 (1997).
15. For linear velocity and density gradients, our observed *SdS* and *PdP* amplitudes are reduced by 1 to 2% for a 10-km interval (compared with a sharp discontinuity), 3 to 7% for a 20-km interval, 6 to 13% for a 30-km interval, and 11 to 22% for a 40-km interval. The effect is largest where the velocities near the discontinuity are the smallest (that is, for *S410S*) and smallest where the velocities are the largest (that is, for *P660P*), although the *S* versus *P* difference is lessened in our case by the higher frequency content in the *PP* stack (the *SS* data have a dominant period of 30 s; the *PP* data are peaked at a 21-s period). Nonlinear gradients, which may occur for mantle phase transitions (*14*), will produce smaller amplitude reductions for reflected pulses than linear gradients over the same depth interval. Thus, these numbers represent upper bounds; the actual amplitude reductions may be smaller.
16. B. Efron and R. Tibshirani, *Science* **253**, 390 (1991).
17. A. M. Dziewonski and D. L. Anderson, *Phys. Earth Planet. Inter.* **25**, 297 (1981).
18. Values of the incoherent stacking parameter ranging from $\sigma = 0$ s (perfect coherence) to 1.6 times larger than our preferred values (for example, $\sigma = 4$ s for *SdS* waves and $\sigma = 2.7$ s for *PdP* waves) resulted in adjusted amplitudes that varied by less than 20% from our computed amplitudes.
19. The shear impedance is the product of the shear velocity and density; thus, the shear impedance contrast is given by the sum of the *S* velocity and density jumps.
20. T. S. Duffy and D. L. Anderson, *J. Geophys. Res.* **94**, 1895 (1989).
21. J. Ita and L. Stixrude, *ibid.* **97**, 6849 (1992).
22. H. Fujisawa, *ibid.* **103**, 9591 (1998).
23. J. B. Gaherty, Y. Wang, T. H. Jordan, D. J. Weidner, *Geophys. Res. Lett.* **26**, 1641 (1999). This paper contains the changes in density and *S* velocity at 410 km for pyrolite and piclogite models. Corresponding values for the changes in *P* velocity were obtained from J. B. Gaherty (personal communication).
24. D. J. Weidner and Y. Wang, *J. Geophys. Res.* **103**, 7431 (1998).
25. Weidner and Wang (*24*) computed density and velocity profiles through the 660-km discontinuity for several different models of mantle composition (with differing Al content) and temperature (1700, 1900, and 2100 K). The points labeled "Pyr" in Fig. 3B are for their 5% Al model at 1700 K; for this model, the velocity and density jumps occur across an interval of less than 20 km. Some of the other models have more gradual gradients that cannot be directly converted to points on Fig. 3. We tested to see if these models might fit our data by computing synthetic seismograms based on the profiles in Fig. 4 of Weidner and Wang (*24*). All of these models grossly overpredict our observed *S660S* and *P660P* amplitudes and lie outside the error bars plotted in Fig. 2.
26. P. M. Shearer, *Nature* **344**, 121 (1990).
27. S. M. Ridgen *et al.*, *ibid.* **354**, 143 (1991).

28. This work was supported by NSF grants EAR93-15060, EAR95-07994, EAR96-14350, and EAR96-28020. M.P.F. was supported by an NSF Postdoctoral Fellowship and the Cecil H. and Ida M. Green Foundation. J. Gaherty and J. Vidale provided constructive reviews of an earlier version of this report.

28 April 1999; accepted 13 July 1999

Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble

T. N. Krishnamurti,[1] C. M. Kishtawal,[1] Timothy E. LaRow,[1] David R. Bachiochi,[1] Zhan Zhang,[1] C. Eric Williford,[1] Sulochana Gadgil,[2] Sajani Surendran[2]

A method for improving weather and climate forecast skill has been developed. It is called a superensemble, and it arose from a study of the statistical properties of a low-order spectral model. Multiple regression was used to determine coefficients from multimodel forecasts and observations. The coefficients were then used in the superensemble technique. The superensemble was shown to outperform all model forecasts for multiseasonal, medium-range weather and hurricane forecasts. In addition, the superensemble was shown to have higher skill than forecasts based solely on ensemble averaging.

Sophisticated numerical models used in operational and research centers throughout the globe routinely make short-term (1 to 7 days in advance) weather and seasonal (one to several seasons in advance) climate forecasts. Individually each modeling group tracks the forecast skill of their model. Within recent years, the use of model ensembles has become an important forecasting component. The methodology of how to generate the ensemble is the focus of many forecasting centers. Here we show that a multimodel superensemble can more accurately predict weather and seasonal climate. The superensemble is developed by using a number of forecasts from a variety of weather and climate models. Along with the benchmark observed (analysis) fields, these forecasts are used to derive simple statistics on the past behavior of the models. These statistics, combined with multimodel forecasts, enable us to construct a superensemble forecast.

Given a set of past multimodel forecasts, we used a multiple regression technique (for the multimodels), in which the model forecasts were regressed against an observed (analysis) field. We then used least-squares minimization of the difference between the model and the analysis field to determine the

[1]Department of Meteorology, Florida State University, Tallahassee, FL 32306, USA. [2]Center for Atmospheric and Oceanic Sciences, Indian Institute of Science, Bangalore, India.

weights. We carried out this minimization at all vertical levels, at all geographic locations (the grid points of the multimodels), and for all model variables.

The motivation for this approach came from construction of a multimodel superensemble from a low-order spectral model (*1*). In this low-order model, it is possible to introduce various (proxy) versions of cumulus parameterization (or model physics) by simply altering a forcing term. Time integration of this system showed that the multiple regression coefficients of these multimodels (regressed against the nature run) showed a marked time invariance. This time invariance is a key element for the success of the proposed method.

We used many models at various horizontal and vertical resolutions. Most of the models had a horizontal resolution of <250 km and a vertical resolution of about 1 km. Model output was interpolated to a common grid of 2.5° and 100 hPa vertically. These global models include parameterizations of physical processes; effects of ocean, snow, and ice cover; and treatment of orography. We divided the run timeline into a control and a forecast part. The observed (or the analysis) fields are used only during the control period to determine the weights.

The Atmospheric Model Intercomparison Project (AMIP) data set (*2*) was used to test our procedure for seasonal forecasting. This data set contains a 10-year integration with 31 global atmospheric general circulation models, all beginning from a prescribed initial time of 1 January 1979 and subject to identical prescribed sea surface temperature (SST) and sea ice boundary forcings (*3, 4*). The control and forecast periods for the superensemble forecasts were arbitrarily derived from these 10-year-long histories of eight selected models at two different timelines: Bureau of Meteorology Research Center (BMRC), Melbourne; Council of Scientific and Industrial Research Organization (CSIRO), Melbourne; European Center for Medium Range Weather Forecasts (ECMWF), London; Geophysical Fluid Dynamics Laboratory (GFDL), Princeton; Laboratory Meteorologic Dynamique (LMD), Paris; Max Planck Institute (MPI), Hamburg; National Center for Environmental Predictions (NCEP), Washington, D.C.; and United Kingdom Meteorological Office (UKMO), London. The first timeline, January 1981 through December 1988, was the control, and January 1979 through December 1980 was used for the forecast. The monthly means of the forecasts and the observed (analysis) fields of the past 8 years were used to generate the statistical weights. The most typical results for this timeline for the root-mean-square (rms) error of the south-north component of winds at 850 hPa (about 1.5 km over the ocean) averaged over an Asian monsoon domain bounded between 50°E and 120°E and between 30°S and 35°N are shown in Fig. 1A. The rms errors for the superensemble and its regression are comparable to typical analysis errors. Examination of the forecast period shows that the superensemble outperforms all models.

In the second timeline, the period January 1979 through December 1986 defines the control and January 1987 through December 1988 defines the forecast. A time history of precipitation forecast skill with the second

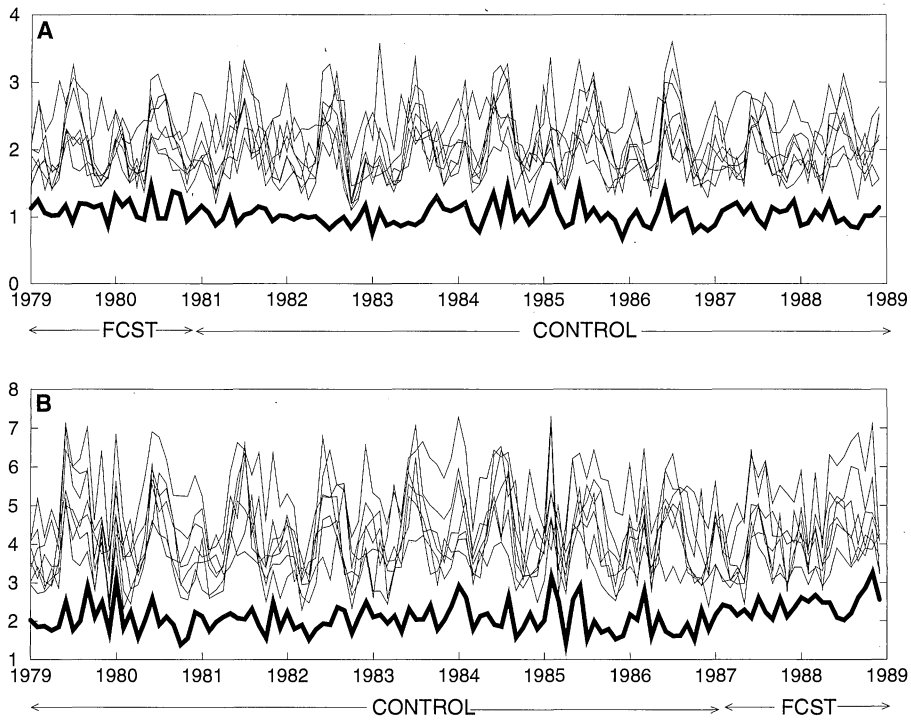

Fig. 1. Asian monsoon domain average rms error for the superensemble (heavy line) and the selected AMIP models (thin lines) for 850-hPa meridional wind (**A**) and precipitation (**B**). Units in (A) are ms^{-1} and units in (B) are mm day^{-1}.

Table 1. The 850-hPa wind rms error (ms^{-1}) for 3-day prediction.

	ECMWF	RPN	UKMO	NCEP	NRL	BMRC	JMA	Ensemble mean	Super-ensemble
Globe	4.1	4.7	5.8	5.8	4.2	4.7	4.8	4.0	3.5
Tropics	2.7	3.5	3.4	4.5	3.1	3.4	3.5	2.7	2.2
Monsoon	2.6	3.4	2.9	4.6	3.1	3.4	3.5	2.7	2.0
Europe	2.0	2.2	2.9	3.3	2.2	2.2	2.0	2.0	1.7
United States	2.6	3.1	3.8	4.5	2.9	3.0	3.1	2.9	2.5
Northern Hemisphere	3.0	3.7	3.8	4.8	3.2	3.6	3.9	3.3	2.8
Southern Hemisphere	4.8	5.4	7.2	6.6	5.0	5.5	5.6	4.6	4.2

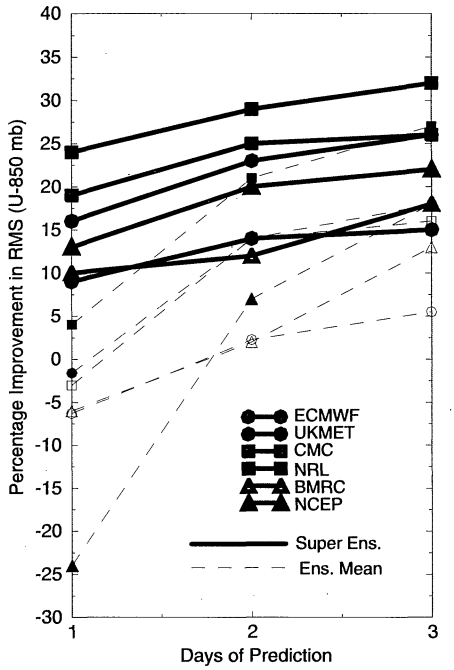

Fig. 2. Percentage improvement of rms error of selected models against the superensemble (solid line) and that of the ensemble mean (dashed line) for numerical weather prediction forecasts during August 1998.

timeline shows a persistent low rms error for the superensemble through about 20 months of integration (Fig. 1B). Because of the length of the forecast, the multimodel's skills were not expected to be high compared with monthly mean observed amounts. After 18 months of forecasts the superensemble retains a high degree of skill for all the dependent variables of the global models. This skill is beyond that of most current atmospheric general circulation models. These two timelines yield similar results for the superensemble. In contrast to the first timeline, all models used in the second timeline do not have identical initial conditions because of the systematic errors and biases of each model. These differences in the initial states do not contribute to any major differences in the overall performances of the multimodels, the ensemble mean, or the superensemble. In the second timeline, the models had already been integrated for 8 years. Thus, we attribute the results (Fig. 1B) to the robustness of the initial states of the ensemble in January 1987 and less so to the resilience of the statistical weights derived from the past (January 1979 through December 1986) behavior of the multimodels and prescribed SST and sea ice.

A number of weather services around the world use multilevel global models for daily numerical weather prediction. These models vary in their resolution, treatment of physics and dynamics, representation of orography, and initial data assimilation procedures. Collectively, they provide a useful multimodel data set. Thus, it is possible to explore the usefulness of the proposed multimodel superensemble for weather prediction. Using real-time data sets from the NCEP; ECMWF; Japan Meteorological Agency (JMA), Tokyo; BMRC; Research Provision Numerique (RPN), Montreal; and UKMO weather services, we examined the global data analyses and forecasts for June, July, and August 1998. These included the initial state and predictions through day 3 of the forecasts. The data sets for the first 61 days were designated as the control timeline. The multimodel forecasts and the observed analysis for this period determined the statistical weights. These were handled separately for days 1, 2, and 3 of the multimodel forecasts. The results, averaged over August 1998, are shown for selected domains (*5*) in Table 1. The superensemble daily skill is similar to that obtained in the seasonal climate forecasts. The superensemble outperformed all other models, including the ensemble average. In the tropics and the monsoon region, errors were reduced by 25% over the ensemble average and skill increased up to 100% over some of the models. Errors are larger for the Southern than for the Northern Hemisphere. The tropical regions generally have a low predictability (as shown in the results from the models). Consequently, the improvements shown from the performance of the superensemble in the tropics and Asian monsoon domains are quite promising.

The superensemble also exhibits a higher skill compared with that of the ensemble mean for the entire global domain (Fig. 2). The evaluation of the skill relies on the observed (analysis) fields against which the respective forecasts are compared. That skill may be biased somewhat toward the group whose analysis is being used for the evaluation of skill. To avoid this bias, we verified the performance of the ensemble mean and the superensemble against each forecasting center's analyses. The superensemble's forecast skill of the 850-hPa meridional wind stands out regardless of the analysis used as a benchmark (Fig. 2).

There were 63 forecasts of tropical systems available from each of the multimodels during 1998 and there were 21 forecasts for 1997. These were provided from four diverse models: Florida State University (FSU), Tallahassee, model; Naval Oceanographic Global Predictions System (NOGAPS), Monterey, California, model; UKMO model; U.S. National Weather Service's GFDL model; and the official forecast from the National Hurricane Center. We applied the regression procedure by using the observed tracks (Table 2) and maximum sustained winds (Table 3) for each of these storms, and we examined the multimodel forecasts every 12 hours for 3-day forecasts. We used the resulting statistics to assess the errors of the superensemble for this control period of 1998. The superensemble shows the smallest track errors compared with all other models. We also used a cross-validation procedure in which, successively, all but one of the storm's forecasts were included to evaluate the statistical weights. These forecast results are as robust as those of the test, illustrating the superior performance of the superensemble. The results from the ensemble averages are superior to all models except those of the superensemble. These results are better than current state-of-the-art methods for hurricane forecasting. Model changes are not desirable during the test and the forecast timelines (*6*). Any major model changes to the multimodels invalidate the use of this procedure. This limits the sources of available model data sets that can be used to test our scheme.

The superensemble described here is far superior, in terms of forecasts, to an ensemble mean. Six or seven models are needed to reduce the multimodel errors. Removing the bias of models (one at a time) and performing the ensemble mean of such (bias removed) multimodels are far less accurate than performing a collective bias removal as we have done.

Establishment of a multimodel superensemble forecast center for weather and seasonal climate forecasts would be a natural proposition from the outcome of these results. Given fast data communications by satellites and computers, it is possible for such a center to receive the multimodel forecasts from various global weather centers and to disseminate the superensemble forecasts back to the modelers and the user community in a timely manner.

Table 2. Hurricane track rms errors (degrees).

Model	NHC	NOGAPS	UKMO	GFDL	FSU	Ensemble average	Cross validation	Super-ensemble
Day 1	1.5	1.7	1.6	1.7	1.7	1.2	1.2	0.9
Day 2	2.8	3.4	2.6	3.0	3.4	2.4	1.9	1.5
Day 3	3.4	3.8	3.9	5.5	4.8	2.6	2.6	1.9

Table 3. Hurricane intensity forecast rms errors (ms^{-1}).

Model	NHC	GFDL	FSU	Ensemble average	Cross validation	Super-ensemble
Day 1	6.6	10.0	11.0	6.5	5.7	5.1
Day 2	9.9	10.3	11.7	8.8	9.0	7.7
Day 3	14.0	12.1	14.5	12.3	12.0	9.6

References and Notes

1. E. J. Lorenz, *J. Atmos. Sci.* **20**, 130 (1963).
2. W. L. Gates *et al.*, *Bull. Am. Meteorol. Soc.* **80**, 29 (1991).
3. T. J. Philips, *PCMDI Report 18* (Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, Livermore, CA, 1994).
4. T. J. Philips, *Bull. Am. Meteorol. Soc.* **77**, 1191 (1996).
5. The domains include the globe (87°S to 87°N), tropics (30°S to 30°N), Asian monsoon (30°E to 150°E, 30°S to 40°N), United States (70°W to 125°W, 25°N to 55°N), Europe (0° to 50°E, 30°N to 55°N), the Northern Hemisphere (0° to 87°N), and the Southern Hemisphere (87°S to 0°).
6. During the 2 years 1997 and 1998, three of these models—FSU, GFDL/NCEP, and NOGAPS—incorporated major changes; thus, we could not apply the statistics derived from 1998 to test the performance of the proposed method on the storms of 1997.

7. Supported by the following weather services and research institutions, to whom we are most indebted: AMIP, BMRC, CSIRO, ECMWF, FSU, GFDL, JMA, LMD, MPI, NCEP, NOGAPS, RPN, and UKMO. This research was supported by NASA grant NAG-1199, NASA grant NAG-4729, NOAA grant NA86GP0031, NOAA grant NA77WA0571, NSF grant ATM-9612894, and NSF grant ATM-9710336.

26 April 1999; accepted 15 July 1999