

TESTING AND EVALUATING ATMOSPHERIC CLIMATE MODELS

Model validation is a crucial process that underpins model development and gives confidence to the results from running models. This article discusses a range of techniques for validating atmosphere models given that the atmosphere is chaotic and incompletely observed.

Any model must be tested to assess the quality of its results and to give guidance on the level of confidence to ascribe to those results. We also must be able to compare alternative formulations of model components and judge whether changes really reduce model error. In this article, we describe the special nature of climate model validation. In general, climate models consist of different model components such as atmosphere, ocean, ice, land, and chemistry model packages, which interact with each other. Because of space limitations, we focus here on atmosphere. Climate-modeling centers are working toward modeling much of the global environment, including some representation of chemistry, aerosols, the biosphere, and carbon cycle interactions with the climate of the atmosphere and ocean.¹

A numerical model consists of a series of approximations, each of which is a potential source of error. Such errors result in a climate model that differs from reality, and we wish to minimize this difference. For example, because the continuous versions of the dynamical equations are discretized, errors occur, particularly near the truncation limit. For instance, aliasing can occur in an Eulerian grid point model. Other physical processes are important parts of the climate system and are needed to represent basic processes

such as the radiative and water balances. We cannot model these processes explicitly because of cost or resolution, or because they are poorly understood. So, we use physical parameterizations to represent them. For example, to reduce computation we approximate radiative calculations. The formation of clouds and precipitation occurs on the microphysical scale and therefore must be parameterized. Convection can only be modeled explicitly at very high resolution, so it also must be parameterized. We can improve the models by improving our understanding and representation of processes and by increasing resolution if the computing resources are available. Individual improvements do not always produce improved results, because errors frequently offset one another.

The atmosphere is chaotic, and we have insufficient observations to describe it accurately. So, we do not know the precise truth against which we need to validate. The validation process therefore must encompass a variety of techniques to show that the models are describing the atmosphere's essential characteristics. W. Lawrence Gates and his colleagues have described a number of validation techniques for climate models.² Such techniques range from using reduced systems that validate the modeling system's individual components, to the validation of the full climate model against analyses of the recent climate. The tests we describe in this article require changing the resolution (horizontally, vertically, or both) and the time step to demonstrate numerical convergence. However, because the physics parameterizations take into account processes not explicitly modeled, they might need

© Crown Copyright 2002

VICKY POPE AND TERRY DAVIES
Met Office

retuning for each resolution or time step. This is often not practical over large resolution changes.

In this article, we outline various forms of validation in a hierarchical order. We start with methods for validating individual components and end with validation of the complete atmospheric climate-modeling system.

Simplified tests

We can test the numerical schemes for solving equations on a variety of simplified test problems that might describe either a particular aspect or a suitable analog of atmospheric behavior. We can do this separately for the model's dynamics or physics components. For example, we can reduce the 3D equations of motion to a 2D horizontal flow and test them with the shallow-water testbed that David Williamson and his colleagues proposed.³ These short-term deterministic tests might possess either an analytical solution or a reference solution against which we can compare our test scheme or model component. We can also determine a numerical scheme's relative accuracy and stability and test its robustness over a wide parameter space. These tests are often the starting point for developing new schemes to apply in climate and numerical weather prediction (NWP) models.

Single-column tests for physics components

We can test individual physical parameterizations in isolation from the rest of the climate model. For example, the Intercomparison of Radiation Codes used in Climate Models (ICR-CCM) used standard atmospheres to compare radiation codes ranging from detailed line-by-line codes to highly parameterized band models.⁴ Resolving differences between models requires accurate spectral measurements such as those that Atmospheric Radiation Measurement sites provide (for an example, see the paper by Gerald Stokes and Stephen Schwartz⁵).

We can test physical parameterizations by running them on a single column of data⁶ with the large-scale horizontal forcing prescribed from either observations or idealized profiles. This lets us study the behavior of particular schemes and their nonlinear interaction between other parameterizations without the cost and complication of running a complete model. These tests have two main limitations. First, they involve no interaction with the dynamics—large-scale ten-

dencies of heat and moisture are prescribed, and winds are usually relaxed to a specified profile. Second, they involve no interaction with adjacent grid points, as there would be in a complete model. Tests might also be limited by the availability of suitable data (as truth) for evaluation. Detailed data are available for some special observing campaigns such as GATE (the GARP Atlantic Tropical Experiment)⁷ or other field data experiments. However, the available examples do not cover the full range of behavior that the parameterization represents in a complete model because the observing campaigns usually cover a limited geographical and time range.

Dynamical core tests

We can test the role of dynamics in a climate model in isolation by using a *dynamical core test*. Such tests typically replace the physics parameterization package with an idealized forcing in the form of simplified physics such as that proposed by Isaac Held and Max Suarez.⁸ This forcing involves running a model with no topography or moisture, a uniform surface, and no seasonal or diurnal cycle. The thermal forcing (simplified radiation) consists of a slow relaxation toward an equilibrium temperature symmetric about the equator. These tests apply Rayleigh friction in the lower part of the troposphere, mimicking the planetary boundary layer's frictional effects. The model is initialized with zero winds, and the validation does not use the initial spin-up period of 200 or more days. One-thousand-day means give a coherent signal, although evidence exists that variability on all timescales might be an issue.

These tests provide a means of assessing the numerical convergence of the dynamics (including numerical diffusive effects) together with the added diffusion usually used in the full-model runs. In addition, we can assess the impact of different numerical techniques on the mean circulation. Results are typically compared against a high-resolution reference run, owing to the lack of an analytical solution.

In contrast to this long-term model assessment, short-term deterministic test cases can be applied to the dynamics component. One example is the breaking polar vortex experiment suggested by Lorenzo Polvani and Ramalingam Saravanan.⁹

The idealized aquaplanet

The simplest test of the full climate model uses

the complete system of dynamics and physical parameterizations but with simplified boundary conditions. Richard Neale and Brian Hoskins proposed a standard set of experiments using the *aquaplanet model*, which retains a full atmospheric general circulation model (GCM) but simplifies the surface boundary as an idealized sea surface everywhere with specified idealized sea-surface temperature (SSTs).¹⁰ These tests can be used to investigate numerical convergence and time-step sensitivity without the additional complication from mountains and land–sea contrasts. We can also study processes that are driven mainly by SST variations in the tropics.

Realistic climate regimes

In full atmospheric climate tests, we run the model over multiple years with prescribed sea surface temperatures, sea–ice boundary conditions, the appropriate solar forcing, and quantities such as the ozone distribution. If we use quantities based on observations for the boundary forcing, we can use other observations and climatologies derived from the observations from the same period to validate the model’s climate. We normally average results for each of the four seasons and average each season’s results over several years or more to produce an appropriate sample from a large number of events. We can assess both the mean climate and the variability.

Statistical tests must establish how accurately the model represents climate. Some simple statistical tests are insufficient if they do not account for internal or interannual variability properly. Hans Von Storch and Francis Zwiers¹¹ and David Rowell and his colleagues¹² have described suitable methods.

An important standardization of this atmospheric climate model test is the *Atmospheric Model Intercomparison Project*.^{13,14} The AMIP makes available additional input fields such as a standard land–sea mask, topography data, and greenhouse gas and aerosol concentrations. Using the framework model, we can compare runs with different versions of the same model or even with models from different modeling centers. These thorough model assessments require a series of sophisticated diagnostic tools and experimental techniques, which have been developed during the course of the AMIP program.

Evaluation techniques

At the Met Office, we use AMIP integrations

to benchmark model changes. AMIP integrations with and without the changes are compared with each other and with the relevant observed climatologies (which we describe in the next section). This lets a wide range of researchers evaluate a series of improvements relative to each other. It can also help us understand feedback in the model. This is not always straightforward because the parameterization changes interact with one another and therefore do not add together linearly.

We have used AMIP integrations to test particular parts of new parameterizations or to look at the impact of changing tuneable parameters in the parameterizations. One problem of analyzing the impact of a model change is that other changes might be necessary before we can implement it. For example, if we increase resolution, we will need to change both the time step and diffusion to maintain numerical stability. Because both these changes affect the results, they require separate evaluation.

AMIP integrations must be long enough to accurately assess the change that we are testing. We typically use five- or 10-year runs. Seasonal means for any shorter period would not reflect the change’s impact accurately, because the impact would be indistinguishable from interannual variability. Five-year means are suitable for evaluating seasonal mean and zonal mean winds and temperatures, top-of-the-atmosphere fluxes, and major features in precipitation. Ten-year means are needed to get a robust signal for mean sea level pressure and mid-latitude storm tracks because of the high degree of variability in extra-tropical meteorology.

We can understand some differences between new and old versions of the same model by doing a series of AMIP tests, adding each change in turn. However, understanding the differences between models developed at different centers is much more difficult, because there are so many differences in the way atmospheric processes are represented. The best way to do this is to take parameterizations from one model and include them in another model.

Evaluation climatologies

We extensively use the ECMWF (European Centre for Medium Range Weather Forecasts) reanalysis climatology (ERA).¹⁵ An alternative is the NCEP/NCAR (National Centers for Environmental Prediction/National Center for Atmospheric Research) reanalysis data.¹⁶ Re-

analysis data sets use all available observations collected by an up-to-date NWP data assimilation system. Such a system uses variational techniques, combining data from observations with a first guess provided by a short-range (usually six hours) forecast. The forecast–analysis cycle repeats every six hours, and the model plays an important role in propagating information forward in both space and time into data-sparse regions.

The reanalysis climatologies have become indispensable for evaluating climate models for the AMIP period because

- They use the same analysis method throughout
- They provide self-consistent fields
- They cover the globe
- They span long time periods and therefore cover the AMIP assessment intervals
- The ERA climatology has a particular advantage for humidity analyses because it assimilates satellite radiances

However, the reanalysis data have certain pitfalls. First, the large variation in the accuracy and range of observation types in different regions is not obvious in the analyzed fields. In northern mid latitudes over land, radiosonde and aircraft measurements are combined and supplemented with satellite measurements to give accurate analyses. Elsewhere, there are fewer observations; in some regions, satellite observations might be the only source of information, and these have limited vertical resolution. Model biases might dominate fields that are not measured directly, such as cloud water.

We use other data sets to evaluate specific fields, particularly surface variables—for example, the *CPC Merged Analysis of Precipitation*.¹⁷ CMAP is a global, monthly precipitation data set covering the 17-year period of 1979 to 1995. It incorporates gauge observations, estimates inferred from a variety of satellite observations, and the NCEP/NCAR reanalysis. David Legates and Cort Willmott produced a climatology of screen temperatures (at 1.5 m above the surface) over land.¹⁸ This is based on observations for the period from 1920 to 1980. ERBE (the Earth Radiation Budget Experiment) provides satellite measurements of radiative fluxes at the top of the atmosphere for the period from 1985 to 1990.¹⁹

Double-call tests

Another method of understanding the impact

of changes to the model is *double-call tests*. This method runs a new or changed scheme alongside the old scheme. It uses fields directly from the rest of the model but does not feed back to the model. With this method, we can assess the change's direct impact. Normally, if we add a new scheme, indirect effects through feedbacks complicate the signal. We used this method to assess the impact of a new radiation scheme in the Met Office model.²⁰

Spin-up tendencies

To evaluate the contribution of individual physical parameterizations and the dynamics scheme to systematic errors in the model and to changes in model climatology, we can use mean *spin-up tendencies*. This involves running a series of one- to five-day integrations starting from operational analyses. An ensemble of approximately 60 integrations gives statistically significant results for one-day tendencies. To get the spin-up tendencies, we take the accumulated increments for each basic-model field from the dynamics and physical-parameterization schemes and average them for all the runs.

NWP often uses spin-up tendencies to evaluate model biases.²¹ However, climate integrations rarely use this technique.^{20,22} This is primarily because the technique works best if the model used to produce the operational analysis (using data assimilation) is the same as the model being tested. Otherwise, disentangling model differences from differences between the model and observations is difficult. At the Met Office, we have the advantage of using basically the same “unified model” for both weather forecasting and climate research. The forecast model has some differences in its parameterizations and has much higher horizontal resolution, but the dynamics and most of the physical parameterizations are the same.

For the technique to give us direct information about the causes of model errors, the total spin-up tendency should qualitatively resemble the biases in the full simulation. The spin-up tendency then gives the initial model drift toward the model climatology, and the individual components give each model scheme's contribution. We can also analyze mean tendencies for the full AMIP integrations. However, the model schemes balance in the mean, and the mean tendencies do not indicate how model errors arise. Differences between spin-up tendencies in models that have changed can give us information about the mechanisms that led to the change.

However, some signals evolve slowly and therefore might not appear in the spin-up tendencies.

Even where we can discover associations, the link between climate and parameterization changes might not always be direct. For example, we found that including convective momentum transport in our model²⁰ had a large direct impact on winds in the tropics. This altered the global circulation, indirectly affecting the extra-tropical circulation.

NWP tests


Climate models of the atmosphere are based on the same principles and equations as models used in NWP. NWP models are used to produce deterministic weather forecasts, which predict the atmospheric conditions up to several days ahead starting from an initial state based on the observed state at analysis time. Deterministic forecasts longer than a week or so are rarely accurate, primarily because of the growth of errors in the initial conditions. However, we can model the atmosphere's mean state and statistical variability on longer timescales, provided we recognize that we are dealing with a chaotic system that can take a variety of states constrained by external forcing such as the Sun's heating and the long-term Earth-ocean-atmosphere interactions.

Nevertheless, NWP systems that include the GCM and data assimilation processes are playing a growing role in monitoring the environment. They are the most effective way to produce the analyses needed to describe the present-day climate. Furthermore, they are the basis for the reanalysis data discussed earlier. The analyses are essentially the same as those required for weather forecasting, although a wider range of quantities can be analyzed.

Changes to NWP systems also must be evaluated before their routine use in operations. The data assimilation cycles run continually over a trial period, which might last several weeks, and forecasts run either daily or twice daily. Results from the data assimilation and the forecasts can be compared directly against observations, and the model forecasts can be compared against the analyses. The verification process typically involves calculating mean and root-mean-square errors for a wide range of parameters in different regions. To some extent, verification against observations and against analyses complement one another. The southern hemisphere in particular, and to some extent the tropics and the stratosphere, lack sufficient conventional observations

to construct a large enough sample for verification. Verification against analyses is also imperfect because the analysis will contain errors, which have their source from the model or in the data assimilation process.

This type of trial can also serve to evaluate short-lived phenomena in climate models. Essentially, it is a more sophisticated version of the spin-up tests we outlined previously. However, results can be more difficult to interpret because errors might depend on the assimilation process as well as the subsequent evolving model errors. Also, most climate-modeling groups do not have an assimilation system for their model, so this technique is not readily accessible to them.

These techniques we've described have played a major part in improving the quality and accuracy of both climate and NWP systems. However, as we mentioned before, the atmosphere forms only one (albeit important) part of the climate system. Further testing is required when the climate system's different components are combined. For example, the Coupled Model Intercomparison Project is a framework for comparing and evaluating coupled ocean-atmosphere models.²³ 

References

1. *Climate Change 2001: The Scientific Basis*, J.T. Houghton et al., eds., Cambridge Univ. Press, Cambridge, UK, 2001.
2. W.L. Gates, P.R. Rowntree, and Q.-C. Zeng, "Validation of Climate Models," *Climate Change IPCC Scientific Assessment*, J.T. Houghton, G.J. Jenkins, and J.J. Ephraums, eds., Cambridge Univ. Press, Cambridge, UK, 1990, pp. 93–130.
3. D.L. Williamson et al., "A Standard Test Set for Numerical Approximations to the Shallow Water Equations in Spherical Geometry," *J. Computational Physics*, vol. 102, no. 1, Sept. 1992, pp. 211–224.
4. R.G. Ellingson and Y. Fouquart, "The Intercomparison of Radiation Codes Used in Climate Models: An Overview," *J. Geophysical Research*, vol. 96, no. D5, May 1991, pp. 8925–8927.
5. G.M. Stokes and S.E. Schwartz, "The Atmospheric Radiation Measurement (ARM) Program: Programmatic Background and Design of the Cloud and Radiation Test Bed," *Bull. Am. Meteorological Soc.*, vol. 75, no. 7, July 1994, pp. 1201–1221.
6. D.A. Randall et al., "Single Column Models and Cloud Ensemble Models as Links between Observations and Climate Models," *J. Climate*, vol. 9, no. 8, Aug. 1996, pp. 1683–1697.
7. J.P. Kuettner, "General Description and Central Program of GATE," *Bull. Am. Meteorological Soc.*, vol. 55, no. 7, July 1974, pp. 712–719.
8. I.M. Held and M.J. Suarez, "A Proposal for the Intercomparison of the Dynamical Cores of Atmospheric General Circulation Models," *Bull. Am. Meteorological Soc.*, vol. 75, no. 10, Oct. 1994, pp. 1825–1830.
9. L.M. Polvani and R. Saravanan, "The Three-Dimensional Structure of Breaking Rossby Waves in the Polar Wintertime Stratosphere," *J. Atmospheric Sciences*, vol. 57, no. 21, Nov. 2000, pp. 3663–3685.
10. R.B. Neale and B.J. Hoskins, "A Standard Test for AGCMs Includ-

ing Their Physical Parameterizations: I: The Proposal," *Atmospheric Science Letters*, vol. 1, no. 2, 1 July 2000, pp. 101–107.

11. H. Von Storch and F.W. Zwiers, *Statistical Analysis in Climate Research*, Cambridge Univ. Press, Cambridge, UK, 1999.
12. D.P. Rowell et al., "Variability of Summer Rainfall over Tropical North Africa (1906–92): Observations and Modelling," *Quarterly J. Royal Meteorological Soc.*, vol. 121, no. 523, Apr. 1995, Part A, pp. 669–704.
13. W.L. Gates, "AMIP: The Atmospheric Model Intercomparison Project," *Bull. Am. Meteorological Soc.*, vol. 73, no. 12, Dec. 1992, pp. 1962–1970.
14. W.L. Gates et al., "An Overview of the Results of the Atmospheric Model Intercomparison Project (AMIP I)," *Bull. Am. Meteorological Soc.*, vol. 80, no. 1, Jan. 1999, pp. 29–55.
15. J.K. Gibson et al., *ERA Description, ECMWF Re-Analysis, Project Report Series, No. 1*, European Centre for Medium-Range Weather Forecasts, Reading, UK, 1997.
16. E. Kalnay et al., "The NCEP/NCAR 40-Year Reanalysis Project," *Bull. Am. Meteorological Soc.*, vol. 77, no. 3, Mar. 1996, pp. 437–472.
17. P. Xie and P.A. Arkin, "Global Precipitation: A 17-Year Monthly Analysis Based on Gauge Observations, Satellite Estimates, and Numerical Model Outputs," *Bull. Am. Meteorological Soc.*, vol. 78, no. 11, Nov. 1997, pp. 2539–2558.
18. D.R. Legates and C.J. Willmott, "Mean Seasonal and Spatial Variability in Global Surface Air Temperature," *Theoretical and Applied Climatology*, vol. 41, 1990, pp. 11–21.
19. B. Barkstorm et al., "Earth Radiation Budget Experiment (ERBE) Archival and April 1985 Results," *Bull. Am. Meteorological Soc.*, vol. 70, no. 10, Oct. 1989, pp. 1254–1262.
20. V.D. Pope et al., "The Impact of New Physical Parameterizations in the Hadley Centre Climate Model—HadAM3," *Climate Dynamics*, vol. 16, nos. 2–3, Jan. 2000, pp. 123–146.
21. E. Klinker and P.D. Sardeshmukh, "The Diagnosis of Mechanical Dissipation in the Atmosphere from Large-Scale Balance Requirements," *J. Atmospheric Sciences*, vol. 49, no. 7, Apr. 1992, pp. 608–627.
22. V.D. Pope and R.A. Stratton, "The Processes Governing Resolution Sensitivity in a Climate Model," to be published in *Climate Dynamics*, 2002.
23. C. Covey et al., "An Overview of Results from the Coupled Model Intercomparison Project (CMIP)," to be published in *Global and Planetary Change*, 2002.

Vicky Pope is the manager of climate model development and validation at the Met Office. Her recent work includes papers on the processes affecting sensitivity and convergence when horizontal resolution is increased in climate models, the simulation of water vapor and its dependence on vertical resolution, and the evaluation of the improved physical parameterizations. She received her MA in mathematics from Cambridge University and her PhD in stratospheric dynamics from Reading University. She is a fellow of the Royal Meteorological Society. Contact her at Met Office, London Rd., Bracknell, Berks, RG12 2SZ, UK; vicky.pope@metoffice.com.

Terry Davies is a manager in dynamical research at the Met Office. Early in his career, he spent some time as a weather forecaster, but he has spent most of his career developing models for numerical weather prediction and climate research. He received his BSc (Hons.) in applied mathematics and computing science from Sheffield University and his PhD in heat and mass transfer in fluid flow from Liverpool University. He is a fellow of the Royal Meteorological Society. Contact him at Met Office, London Rd., Bracknell, Berks, RG12 2SZ, UK; terry.davies@metoffice.com.

For more information on this or any other computing topic, please visit our Digital Library at <http://computer.org/publications/dlib>.

How to Reach CiSE

Writers

For detailed information on submitting articles, write to cise@computer.org or visit <http://computer.org/cise/edguide.htm>.

Letters to the Editors

Send letters to

Jenny Ferrero, Contact Editor
jferrero@computer.org

Please provide an email address or daytime phone number with your letter.

On the Web

Access <http://computer.org/cise> or <http://ojps.aip.org/cise> for information about *CiSE*.

Subscription Change of Address (IEEE/CS)

Send change-of-address requests for magazine subscriptions to address.change@ieee.org. Be sure to specify *CiSE*.

Subscription Change of Address (AIP)

Send general subscription and refund inquiries to subs@aip.org.

Subscribe

Visit <http://ojps.aip.org/cise/subscribe.html> or <http://computer.org/subscribe>.

Missing or Damaged Copies

If you are missing an issue or you received a damaged copy (IEEE/CS), contact membership@computer.org. For AIP subscribers, contact kgentlie@aip.org.

Reprints of Articles

For price information or to order reprints, send email to cise@computer.org or fax +1 714 821 4010.

Reprint Permission

To obtain permission to reprint an article, contact William Hagen, IEEE Copyrights and Trademarks Manager, at whagen@ieee.org.