objectives outlined by President Obama to extend our reach beyond low-Earth orbit and explore into deep space. It's robotic missions like these that will pave the way for future human space missions to an asteroid and other deep-space destinations."

The U.S. National Space Policy, released on 28 June 2010, indicates that NASA will "by 2025, begin crewed missions beyond the Moon, including sending humans to an asteroid."

The asteroid mission was selected over two other candidates, with NASA evaluating science merit, how well the science payload and science implementation would work, and the feasibility of the mission, technically and managerially, according to Paul Hertz, chief scientist of NASA's Science Mission Directorate. The New Frontiers Program conducts scientific investigations of our solar system with medium-class spacecraft missions.

—RANDY SHOWSTACK, Staff Writer

GEOPHYSICISTS

In Memoriam

PAGE 195

Thomas Barrow, 86, 27 January 2011, Ocean Sciences, 1966

Harold Boyne, 80, 26 January 2009, Hydrology, 1980

Wynne Calvert, 72, 10 October 2009, Magnetospheric Physics, 1960

Chen-Wu Chien, 89, 10 July 2010, Magnetospheric Physics, 1961

Justin DeSha-Overcash, 23, 11 January 2011, Study of the Earth's Deep Interior, 2010

John Fett, 76, 14 July 2009, Seismology, 1964

James Fix, 84, 27 January 2011, Seismology, 1965

Siegfried Franck, 59, 16 May 2011, Planetary Sciences, 1999

John Fryling, 78, 4 February 2011, Ocean Sciences, 1987

John Gilliland, 71, 5 May 2010, Geomagnetism and Paleomagnetism, 1974

Henry Healy, 93, 9 December 2010, Hydrology, 1955

Eugene Herrin Jr., 80, 10 November 2010, Seismology, 2003

Leon Knopoff, 85, 20 January 2011, Fellow, Seismology, 1952

George Mars, 67, 9 December 2010, Solar and Heliospheric Physics, 2003

David Murcray, 85, 13 October 2009, Aeronomy, 1964

Jack Oliver, 87, 5 January 2011, Fellow, Seismology, 1949

Patrick Parker, 78, 5 April 2011, Ocean Sciences, 1963

Peter Petrakis, 82, 28 February 2011, Atmospheric Sciences, 2008

George Reid, 81, 6 May 2011, Fellow, Aeronomy, 1961

Stephen Reynolds, 59, 2 January 2010, Ocean Sciences, 1980

Wilfried Schroeder, 65, 12 April 2011, Atmospheric Sciences, 1971

Stanley Schumm, 84, 10 April 2011, Hydrology, 1954

Kensaku Tamaki, 62, 5 April 2011, Marine Geology and Geophysics, 2007

Mizuki Tsuchiya, 81, 24 December 2010, Ocean Sciences, 1965

Noboru Wakai, 81, April 2009, Space Physics and Aeronomy, 1962

Ronald David Wall, 51, 19 May 2010, Hydrology, 1996

FORUM

Should We Assess Climate Model Predictions in Light of Severe Tests?

PAGE 195

According to Austro-British philosopher Karl Popper, a system of theoretical claims is scientific only if it is methodologically falsifiable, i.e., only if systematic attempts to falsify or severely test the system are being carried out [*Popper,* 2005, pp. 20, 62]. He holds that a test of a theoretical system is severe if and only if it is a test of the applicability of the system to a case in which the system's failure is likely in light of background knowledge, i.e., in light of scientific assumptions other than those of the system being tested [*Popper,* 2002, p. 150]. Popper counts the 1919 tests of general relativity's then unlikely predictions of the deflection of light in the Sun's gravitational field as severe.

An implication of Popper's above condition for being a scientific theoretical system is the injunction to assess theoretical systems in light of how well they have withstood severe testing. Applying this injunction to assessing the quality of climate model predictions (CMPs), including climate model projections, would involve assigning a quality to each CMP as a function of how well it has withstood severe tests allowed by its implications for past, present, and near-future climate or, alternatively, as a function of how well the models that generated the CMP have withstood severe tests of their suitability for generating the CMP.

For example, a severe testing assessment of a CMP generated by a member of the ensemble of global climate models on which the fifth assessment report of the Intergovernmental Panel on Climate Change (IPCC) will rely might involve assessing how well the member has done at simulating data that are both relevant to determining its suitability for generating the CMP and unlikely in light of the ensemble of global climate models on which the IPCC fourth assessment report relied. Data capturing global mean surface temperature trends during the second half of the twentieth century are relatively well simulated by, and thus not unlikely in light of, the ensemble of global climate models on which the IPCC fourth assessment report relied. These data would, accordingly, not be expected to challenge global climate models developed since the fourth report and are thus unsuitable for severely testing models that will be relied on in the fifth report. Data capturing the positive global mean surface temperature trend during the late 1930s and early 1940s are not well simulated by the ensemble relied on in the fourth IPCC report [*Solomon et al.,* 2007, p. 61]. These data will thus better serve to test severely models used in the fifth IPCC report.

Of course, scientists might not always manage to devise adequate severe tests, especially for long-term CMPs. But this merely means that some CMPs will be assigned a low quality when assessed from a severe testing perspective.

An important question is whether Popper's injunction should be applied in assessing CMP quality. As we will see, performance at severe tests currently plays a limited role in such assessment. I argue that this should change.

Severe Testing Assessment of CMPs: Current Situation

The scientific community has placed little emphasis on providing assessments of CMP quality in light of performance at severe tests. Consider, by way of illustration, the influential approach adopted by *Randall et al.* [2007] in chapter 8 of their contribution to the fourth IPCC report. This chapter explains why there is confidence in climate models thus: "Confidence in models comes from their physical basis, and their skill in representing observed climate and past climate changes" [*Randall et al.,* 2007, p. 601].

The focus in this quote, and elsewhere in the chapter, is on what model agreement with physical theory as well as model simulation accuracy confirm. Supposedly, better grounding in physical theory or increased accuracy in simulation of observed and past climate means increased confirmation of CMPs. In this vein, the question addressed in the section on metrics of how reliable models are at generating projections is just: "What does the accuracy of a climate model's simulation of past or contemporary climate say about the accuracy of its projections of climate change?" [*Randall et al.,* 2007, p. 594].

CMP quality is thus supposed to depend on simulation accuracy. However, simulation accuracy is not a measure of test severity. If, for example, a simulation's agreement with data results from accommodation of the data, the agreement will not be unlikely, and therefore the data will not severely test the suitability of the model that generated the simulation for making any predictions (see comment 1 in the online supplement to this *Eos* issue (http://www.agu.org/eos_elec)).

Another important and commonly used approach to assessing CMP quality is the Bayesian approach [see, e.g., *Hegerl et al.,* 2006]. Let us consider one of its simple, but sufficiently representative, applications [*Frame et al.,* 2007]. In this application the posterior probability distribution function, P(*F*|data)—which specifies the probabilities of values of a model parameter *F* in light of data—is calculated using Bayes' rule, P(*F*|data) = P(data|*F*) P(*F*)/P(data). P(data|*F*) captures the likelihood that the data would be simulated by model simulations and does so as a function of values of *F*. P(*F*) is the probability distribution function assigned to *F* prior to consideration of the data. P(data) is taken to be a normalizing constant required to ensure that the probabilities yielded by *P(F*|data) sum to 100%.

For example, we can use Bayes' rule to calculate a posterior probability distribution for the equilibrium climate sensitivity parameter of a simple energy balance model on the basis of an expert estimated prior probability distribution for the parameter and the likelihood, as a function of the parameter's values, that the model gives to paleoreconstructions of global mean annual temperature over the past millennium.

The Bayesian approach does not consider whether data appealed to in calculating P(*F*|data) provide severe tests of estimates of *F*. This approach does take into account the extent to which model simulations agree with data—something that is captured by P(data|*F*)—but we have seen that degree of agreement with data is not itself a measure of test severity. This approach also takes the prior probability distribution assigned to *F*, P(*F*), into account. P(*F*) is correlated with test severity but in the wrong way: Bayes' rule is such that the less probable values from a range of values of *F* are prior to testing, the lower the posterior probabilities of the values will be. On a severe testing approach, however, confidence in values of *F* should increase with the severity of the tests at which the values have succeeded and thus also increase with how unlikely the values were prior to these successes. As to P(data), it is the same for all values of *F*. It cannot, accordingly, be an indicator of the severity with which data test different estimates of *F* (see comment 2 in the online supplement).

Severe Testing Assessment of CMPs: Why Do It?

It appears, then, that a severe testing approach to assessing CMP quality would be novel. Should we, however, develop such an approach? Arguably, yes (see also comment 3 in the online supplement). First, as we have seen, a severe testing assessment of CMP quality does not count simulation successes that result from the accommodation of data in favor of CMPs. Thus, a severe testing assessment of CMP quality can help to address worries about relying on such successes, worries such as that these successes are not reliable guides to out-of-sample accuracy, and will provide important policy-relevant information as a result (see comment 4 in the online supplement).

Second, assessing CMP quality using a severe testing approach would assist in assessing the maturity of the science underlying CMPs. This is because the more mature a body of knowledge is, the easier it is to specify severe tests for its claims. Assume that we want to test a prediction severely. The prediction will have testable implications only when conjoined with a set of additional assumptions, including basic theory and quasi-empirical generalizations. So if we are severely to test the prediction, and not just the conjunction of the prediction and the additional assumptions, then the additional assumptions will have to be established independently of the prediction. Only then might the potential falsity of an implication of the conjunction of the prediction and the additional assumptions constitute a real potential challenge to assuming the truth of the prediction, as opposed merely to a challenge to the conjunction of the prediction and the additional assumptions. The more mature a science is, the more such independently established claims tend to be in place and the easier it is to specify severe tests (for an illustration, see comment 5 in the online supplement).

Although severe testing is not typically used in existing assessments of CMP quality, some severe testing of models and CMPs may already occur. Still—and this brings us to a third reason for using a severe testing approach to assessing CMP quality—applying such an approach would increase the extent to which severe testing is used, which, in turn, might help us to develop better CMPs. According to Popper, severe testing is the way in which science progresses and thus the way in which to uncover better predictions. Even if we don't accept that a methodology based on severe testing is the only way in which we learn about the world, it is clearly one important way of doing so.

Toward Improved Assessments of CMP Quality

Assessing CMP quality is an ongoing challenge, with existing approaches to such assessment facing real difficulties [*Frame et al.,* 2007]. Developing a severe testing approach should assist in handling some of these difficulties. Furthermore, a severe testing approach would facilitate assessing the maturity of the science that underlies CMPs and might assist in developing better CMPs. We should therefore develop a severe testing approach to assessing CMP quality.

Acknowledgment

I thank Henk Dijkstra, Jos de Laat, and this article's four referees for their helpful comments.

References

Frame, D. J., N. E. Faull, M. M. Joshi, and M. R. Allen (2007), Probabilistic climate forecasts and inductive problems, *Philos. Trans. R. Soc. A., 365*(1857), 1971–1992, doi:10.1098/rsta.2007.2069.

Hegerl, G. C., T. J. Crowley, W. T. Hyde, and D. J. Frame (2006), Climate sensitivity constrained by temperature reconstructions over the past seven centuries, *Nature, 440,* 1029–1032, doi:10.1038/nature04679.

Popper, K. R. (2002), *Conjectures and Refutations: The Growth of Scientific Knowledge,* Routledge, London.

Popper, K. R. (2005), *The Logic of Scientific Discovery,* Routledge, London.

Randall, D. A., et al. (2007), Climate models and their evaluation, in *Climate Change 2007: The Physical Science Basis—Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change,* edited by S. Solomon et al., chap. 8, pp. 589–662, Cambridge Univ. Press, New York.

Solomon, S., et al. (2007), Technical summary, in *Climate Change 2007: The Physical Science Basis—Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change,* edited by S. Solomon et al., pp. 19–91, Cambridge Univ. Press, New York.

—Joel Katzav, Eindhoven University of Technology, Eindhoven, Netherlands; E-mail: j.k.katzav@tue.nl