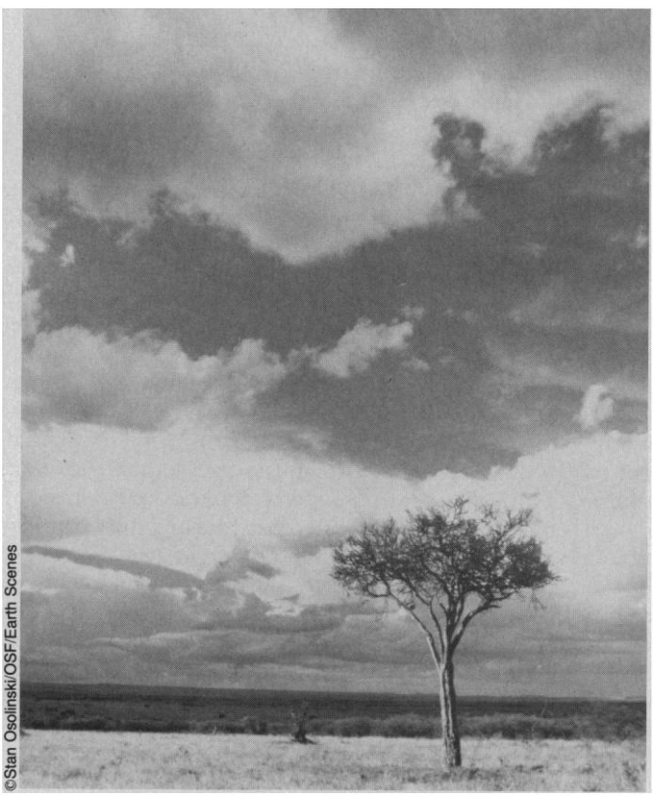
Telling Weathermen When to Worry

Research meteorologists are coming up with ways to alert weather forecasters when they have a bad prediction or a particularly good one; the result should be more credible forecasts

IF RESEARCHERS have their way, the weather forecast on the evening news will one day have a different tone about it: "Well, folks, my forecast for tomorrow and Friday is sunny, clear, and warm, highs in the low 80s, lows in the upper 50s. For the Fourth of July weekend, well . . . frankly, I can't say much about that. The latest computer analysis says I shouldn't be making a forecast that far out today, so I won't. Tune in tomorrow when we're a bit closer to the weekend."

Such frankness might not be popular with news directors who care more that their weathermen appear infallible than that their viewers be confronted with the limits of weather forecasting. But atmospheric modeler David Baumhefner of the National Center for Atmospheric Research in Boulder, Colorado, among others, thinks that such candor must become more common. "There are times when the atmosphere is so unstable you can't say what's going to happen. I've been arguing that there are times that a forecaster should say, 'I'm not sure.' "



©Stan Osolinski/OSF/Earth Scenes

Whither the weather? *Knowing how stable the atmosphere is helps.*

Whether doubt sells well on TV or not, researchers are hard at work finding ways to warn forecasters about dubious days, as well as to alert them to times when they can have exceptional confidence in their forecasts. Such prediction of the accuracy of forecasts is a new, major direction in forecasting research. It will do nothing to improve traditional forecasts themselves. Instead, this new effort, called skill prediction, will increase the usefulness of forecasts produced in the traditional way. If skill prediction comes into wide use, consumers of forecasts, such as a family planning a picnic or a farmer planting his crop, would have some idea of how good each forecast is, something that heretofore remained a gut feeling of the weather forecaster.

The driver of this new field of skill prediction is meteorologists' increasing frustration with the usual approaches to improving forecasts. Data pour in from satellites and ground-based sites; new, more powerful supercomputers come online every few years to process the data; but forecast accuracy barely creeps upward. And this is holding true for all types of forecasting. For example, short-range forecasting—looking out only a few days—has long depended on computer predictions spruced up with some human insights, an approach that is yielding only small increases in accuracy (*Science*, 10 May 1985, p. 704). Ten years of research and several generations of supercomputers have served to push the limits of medium-range forecasting, which is wholly dependent on computers, from 5 days out to just 7 days.

But the frustrations that have led most directly to the new thrust in skill prediction have come in two very different but equally ambitious types of forecasting. One is the attempt to peer 30 days ahead, using the power of the computer, to sketch in some detail the weather during a period of several days or a week. And the second spans the next 90 days, this time using the power of the human brain, to predict average conditions of an entire season. Neither of these efforts has proven consistently reliable. But on certain, infrequent occasions, these longer range forecasts are successful. Can forecasters know in advance which times are propitious for them and which look so bad that they should doubt the reliability of their forecasts?

Forecasters at the U.S. Weather Service think they can, using their long track record as a guide. For 25 years they have been trying to wrest from an infuriatingly variable atmosphere even the occasional useful forecast of average conditions during the next 90 days. Their forecast tools have included subjective decipherings of recent atmospheric behavior around the globe, such as broad wind patterns, that are used to foreshadow future behavior. Unfortunately, their long-range record so far is a paltry 8% improvement over sheer chance in forecasting temperature (*Science*, 7 April, p. 30).

Things are not that bad all the time, however, which is what gives hope of predicting forecast skill. By studying the rare occasions when they did better than average, long-range forecasters discovered some guidelines that could suggest when they would have a better chance of being right. Forecasters consistently do better in some seasons than others, for example. In winter, forecasting skill is an encouraging 18%, but the fall is a dead loss, with skill plummeting to zero percent.

Seasonal forecast skill also varies from place to place. Even in winter, temperature forecasts for a broad band running along the Rockies are useless, on average, while skill in the Southeast soars to better than 40%. And skill seems to vary by the kind of weather. Predicting near-normal temperatures appears to be hopeless, but the skill in predicting winter temperatures well below nor-

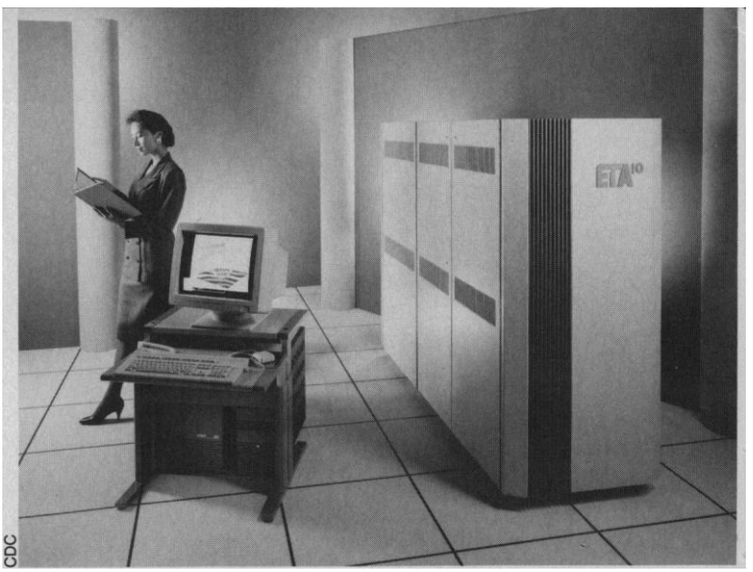
mal—the bitter winter that killed your camellias—or well above normal—the one for which your natural gas supplier did not need extra reserves—is not so bleak. Skill in forecasting such extremes exceeds 60% in the Southeast.

No one can say in detail why forecast skill varies from case to case this way, but it has something to do with stability. The longer-lived the regimes controlling the weather of a particular season or region, the easier the forecast.

One situation in which a long-lived regime stabilizes the weather over parts of North America and allows the anticipation of exceptionally good 90-day forecasts is El Niño. Recently Chester Ropelewski and Michael Halpert of the National Meteorological Center (NMC) in Camp Springs, Maryland, pointed out that the regional weather effects of El Niño, which is the appearance of exceptionally warm waters in the tropical Pacific, actually appear in their opposite sense when El Niño becomes La Niña, a time of unusually cool tropical waters. Based on this observation, official forecasters last fall called for a drier than normal Southeast in the coming winter, which turned out to be the one bright spot in that season's prediction of precipitation.

Attention to stabilizing influences such as El Niño during certain seasons and places "should have a huge impact in a few [seasonal] forecasts," says meteorological statistician Robert Livezey of NMC. "Where we want to make a killing is in such individual cases. We're going after a lot of skill in a few circumstances." The overall average skill will not jump much, he concedes, but when forecasts go out that are billed as more trustworthy than usual, the farmers, energy suppliers, and others who should act on them will be more likely to take them seriously.

Human forecasters looking 90 days ahead find skill prediction helpful, but researchers attempting to extend computer forecasting from its present range of a week ahead to 30 days ahead have no choice. They have hit a brick wall in this range (see box); they have to know how reliable they are going to be on any given occasion. Computer forecasting this far ahead is called dynamical extended range forecasting or DERF, dynamical referring to computer calculations of atmospheric motions. Ninety-day human forecasting has a meager but real overall skill of 8%, but DERF turns out to have no useful skill at all in making 30-day forecasts.



Smaller is faster. *This compact ETA 10 supercomputer is fast enough to predict the reliability of some forecasts.*

But even with a technique that is not useful in routine use, some forecasts are better than others. Says Steven Tracton, the lead DERF researcher at the National Meteorological Center, "there is a potential for extracting useful information [from DERF] if you can tell in advance good cases from poor ones." Indeed, when Tracton and his colleagues ran a computer forecast model out to 30 days on 108 consecutive days to test the DERF concept, they discovered that about 20 forecasts scattered among the useless forecasts actually had useful skill.

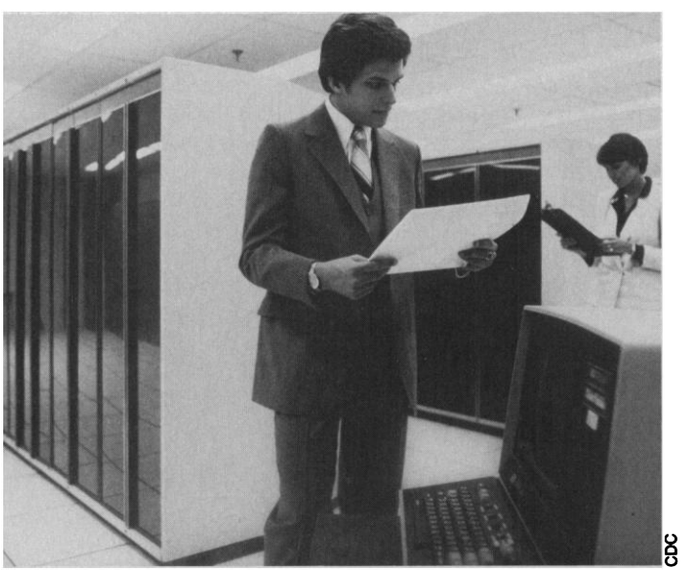
And once again, the key factor was stability of the atmosphere. Instead of using a long historical record of forecast success and failure to identify more stable periods, DERF forecasters can turn to the immediate past. They can use their computer to ask the same questions about recent computer forecasts as human forecasters ask themselves. A forecaster might think: Is the medium-range model's forecast for next Tuesday pretty much the same today as it was yesterday? Do the model forecasts for Tuesday from NMC, the European Center for Medium Range Forecasts, and the United Kingdom's Meteorological Office agree in general? If the forecasts tend to agree, then the atmosphere is probably stable enough to trust the model forecasts, the forecaster concludes.

When Tracton and his colleagues applied a computer version of this forecast agreement technique to their 30-day cases, they found that the correlation between forecast agreement and forecast skill explained a modest 17% of the variability of the atmosphere on a hemispheric basis. So, this forecast agreement technique might not demand much computer time, but it obviously is not going to be the sole answer to the skill prediction problem.

There is a better, if not perfect, answer. The skill of 30-day forecasts proved to be highest when a single atmospheric pattern called the Pacific–North American pattern was present. The pattern consists of a chain of high-altitude highs and lows arcing over the North Pacific Ocean. For forecasts centered about 2 weeks ahead, an impressive 50% of the variability of the atmosphere is explained by the Pacific–North America pattern.

In his role as lead researcher in DERF at the European Center, Timothy Palmer finds that such forecast agreement and pattern recognition methods can be useful sometimes. He recalls that computer forecasts of the storm that downed 15 million trees in the south of England in October 1987 varied considerably from day to day and from model to model. That helped alert French forecasters to the hazards of trusting the computers that day, a lesson lost on British forecasters (*Science*, 11 March 1988, p. 1238).

"These indicators were useful at times," says Palmer, "but the problem is that they are not consistent enough for routine prediction of forecast skill on a day-to-day basis. The basic direction we want to go is the Monte Carlo approach." In the Monte Carlo



Computer forecasting the old way. *This Cyber 205 supercomputer is too slow to allow the prediction of forecast accuracy.*

Forecasting Pushed Too Far

Research meteorologists have been using computers to predict the behavior of the atmosphere since the days of vacuum tubes and room-sized computers. As computers became more powerful, researchers dreamed up more complex forecast models to run on them, allowing them to predict the weather farther into the future. But now the reach of forecast computers has greatly exceeded their grasp. The problem is chaos.

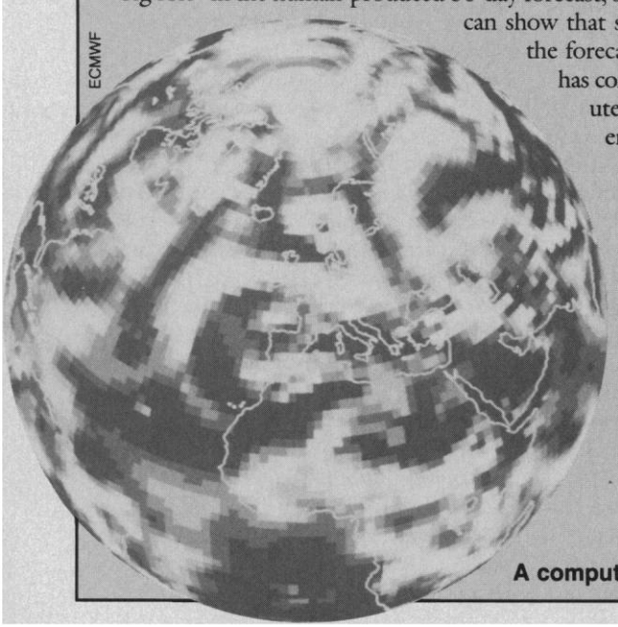
The atmosphere is all too chaotic. Models run out to 10 days on the latest supercomputers turn out daily forecasts of weather patterns that on average are useful out to about 6 or 7 days. Beyond then, chaos usually takes over. The inevitable snowballing of small errors in the model's initial picture of the atmosphere usually destroys any detailed resemblance between the forecast and reality.

The hope had remained, however, that there might be times when the atmosphere was not so sensitive to errors, when some added stability might make forecasts out to, say, 30 days practical with today's computer models. At the Geophysical Fluid Dynamics Laboratory in Princeton, where new modeling techniques are developed for forecasting, Kikuro Miyakoda and his colleagues found one case in 1983, a forecast beginning 1 January 1977, in which a model predicted that an existing, unusually stable weather pattern would persist for 30 days, just as happened in the atmosphere (*Science*, 6 May 1983).

This single tantalizing success in dynamical extended range forecasting, or DERF, eventually prompted several research centers to get into the DERF business. At the National Meteorological Center (NMC)in Camp Springs, Maryland, nine cases were run as a quick test of the practicality of 30-day forecast experiments. The results "were sufficiently encouraging to pursue further," says Steven Tracton of NMC. Then the arrival of a new Cyber 205 supercomputer provided a fleeting opportunity when enough computer time would be available to run NMC's medium-range forecasting model out to 30 days rather than the usual 10 days. One hundred eight 30-day cases were run on consecutive days during the winter of 1986–87. After that, routine forecasting chores and other research gobbled up the new computer power for good.

These 30-day forecasts, says Tracton, illustrate "one of the basic laws of meteorology—the first case you look at is the best, and it's downhill from there." In the case of the 108 30-day forecasts, notes Tracton, "there is little if any useful information in the runs beyond 10 days." The average skill for all the cases was 0.39 on a scale on which a perfect forecast scores 1.00. A score of 0.5 to 0.6 or higher is considered an indication of useful skill. In short, the average forecast was useless. Somehow the forecaster would have to determine which forecasts would be distinctly better than average.

Computer forecasts do seem to be making some contribution to the routine prediction of next month's weather. The human forecasters who produce forecasts of average conditions during the next 30 days from a melding of weather observations also look at the computer forecast for the next 10 days. This peek ahead "plays a very big role" in the human-produced 30-day forecast, says Robert Livezey of NMC. "You can show that somewhere around the late 1970s, the forecast started to show useful skill and has consistently since. You can only attribute that to the advent of the computer-generated 10-day forecast." Some researchers believe that the improvement is just a faint echo of the high skill in the first few days of the computer forecast, but Livezey sees it as more complicated than that. He believes the forecaster uses the 10-day forecast to make an intelligent guess about how stable the atmosphere will be over the next 30 days, high stability being a key to a more accurate forecast. ■ R.A.K.



ECMWF

A computer's view of clouds.

technique, the model is run at least ten times, instead of once, for each forecast. After the first run, each additional run begins with an initial picture of the atmosphere that has been altered imperceptibly to mimic the real errors in the observations. The less resemblance among the model forecasts after 10 days, the more sensitive to errors the forecast is that day. Only now is the necessary computer power becoming available for such Monte Carlo runs, even to those rich in computer time.

Last winter the European Center ran a series of Monte Carlo experiments every other week on its Cray X-MP-48 supercomputer. "We're pretty encouraged," says Palmer. "With one or two interesting exceptions, this scheme can give you indications of the reliability of the forecast, but it's still early days."

The Meteorological Office's recent acquisition of a Control Data Corporation ETA 10 has let the British also throw big computer power at the forecast skill problem. Every 2 weeks the medium-range model, set at a slightly reduced spatial resolution, is run out to 30 days every 6 hours as new observations arrive, until there have been nine runs. This is the lagged average approach, a sort of poor man's Monte Carlo. These 30-day forecasts are now being provided to paying customers on an experimental basis.

U.S. researchers, meanwhile, are stuck, at least for the moment. "We can't do further significant work in DERF without the next generation computer," says Tracton. The National Meteorological Center is traditionally short on computer power. Two CDC Cyber 205's there, which are 1980 hardware 20 times slower than the ETA 10, run a daily menu of short-range, medium-range, and aviation weather models as well as research runs. A new machine such as the Cray Y-MP, which would be ten times faster than a 205, may be installed before the end of 1990. By then the European Center should have its new supercomputer, perhaps something like the Cray 3. It would be a newer design, run a bit faster, and have a larger memory. In the race to predict forecast skill, as has been the case in the race to improve forecasts themselves, U.S. researchers could again be handicapped.

Even as researchers apply skill prediction at the limits of forecasting, it shows promise at the shorter range of a few days to a week. Stay tuned and you may soon hear your weatherman conceding that "I'm not sure."

■ **RICHARD A. KERR**

ADDITIONAL READING

S. Tracton, K. Mo, W. Chen, E. Kalnay, R. Kistler, G. White, "Dynamical extended range forecasting (DERF) at the National Meteorological Center," *Mon. Weath. Rev.*, in press.