Scientific Programming 18 (2010) 221–223
DOI 10.3233/SPR-2010-0304
IOS Press

Book Review

A Vast Machine, Computer Models, Climate Data, and the Politics of Global Warming, by Paul N. Edwards, April 2010, MIT Press, ISBN 978-0-262-01392-5.

Paul Edwards explains how we measure and predict weather and climate, why this is a complex but necessary process, and why we should trust this difficult and highly technical procedure. The book is detailed, thorough, and is of interest both to those who care about weather and climate specifically and to those who worry about the more general problem of how to get reliable data for predicting or understanding scientific issues in general. Personally, I valued it mostly as an argument for why data curation – the process of absorbing scientific observations and making them easy to use by others – is a significant and vital task. The book does describe informally the way climate models are made and evaluated, but it is not a technical treatise on modeling and contains no equations. It is instead a higher-level explanation of how scientific modeling and prediction proceeds and how we need to combine computing, data analysis, and the theory of an application area to get a basis for policy.

Large scale data involves far more complicated measurement and processing than is commonly realized. We think perhaps that observers look at a thermometer and write down a number, but there are thousands of such observers, scattered all over the globe, whose predecessors may have been doing this for a century. They work on different time schedules, they may or may not have shielded their thermometers from the wind and sun, and they miss taking readings from time to time. All their observations must be merged with data from ships, satellites and balloons. Many of those instruments depend not on columns of mercury but on electronic equipment which needs complex calibration and gets improved and replaced regularly. As a result, the underlying climate data set is produced by a symbiotic process in which atmospheric models are developed at the same time that observations are corrected and joined to the data.

This process is not without risk. The most famous example is probably the widely-repeated claim that NASA missed the ozone hole because software had been designed to filter out unusually low values as experimental errors; when their data from 1978 to 1985 was reprocessed after some British scientists reported the ozone hole in 1985, the ozone hole was evident; NASA denies this. What is clear from both sides is that the scientific process was working, the ozone loss had been eventually noticed in the satellite data, and at worst there would have been only a few months delay in public recognition of the ozone depletion in Antarctica.

As another example of the detailed data complexity (not in the book), I found myself a few months ago reading a discussion of why Darwin, Australia, was reported as slightly hotter between 1900 and 1940 than it has been since. Climate change skeptics had seized on this example to refute global warming. More careful scholars picked through the history. In early 1941 the Royal Australian Air Force built a new airfield near Darwin (remember there was a war on), and asked the meteorologists to relocate the weather station from the town center to the airport. Towns are "heat islands" thanks to human activity and pavement, and one would expect the readings out in the country to be lower; in addition, the old thermometer had not been properly shielded. We do not have, and cannot get, comparative sequences of measurements at the two locations since in January 1942 the Japanese bombed Darwin and destroyed the center of the city, including the post office building that provided the "micro-climate" around the pre-1940 thermometer. But there is enough information to know that the Darwin temperature decline is spurious, and the city will not have to worry about glaciation any time soon.

Prof. Edwards explains carefully why we need to develop models and datasets together. Modeling equations (or the more modern equivalent, machine learning systems) typically want data uniformly distributed across a regular grid or some other possible set of input variables. The data any geographic system rarely comes in such a neat form. Thus, we imagine a model for the data, fit the model as best we can, and use that both to identify erroneous data readings and to further improve the model. For example, if we wish to base temperature estimates on the dates of the wine-grape harvest or bird migrations, we need to have a model

1058-9244/10/$27.50 © 2010 – IOS Press and the authors. All rights reserved

which proposes how the natural phenomena depend on temperature, precipitation, length of day, or whatever; we then have to see what we can learn by fitting this model to the data we have.

The need to develop both models and data implies two different sorts of problems, which the author calls data friction and computational friction. Data friction is the difficulty of using raw observations in later research. It may be as simple as reconciling distances in inches with those in centimeters, or as complicated as relating upper atmosphere measurements taken at different altitudes. Computational friction is the difficulty of actually computing results from different models: we may find ourselves lacking theoretical understanding, mathematical skill, programming techniques or packages, computing power, or the data needed to run the particular programming packages. If your problem is to relate temperatures in Philadelphia to those in Pittsburgh, but somebody wrote down dawn temperature in Philadelphia and sunset temperature in Pittsburgh, that's data friction. If your problem is that you need a program to compute the correlation, but you don't know where to find it, or you don't have the skill to write it, or you can't afford the processor power to run the program, that's computational friction. Data friction affects climate change severely since we want to use observations taken a century or more in the past, when nobody understood what we would be doing with these numbers or how it would be best to gather them.

Computational friction can also involve the tension between algorithms, computer programming, and computing machines. Modern climate models (and also models such as those for earthquakes, protein folding or evolution) use our very largest computers. Those tend to be improving themselves through changes in architecture, so that the software of today may not be the best adapted to run the same model on the machine of tomorrow.

However these software systems are also large and difficult to rewrite, so that the research teams must choose between effort spent changing code written for vector machines (SIMD) to run on multiprocessors (MIMD) and effort spent armtwisting manufacturers to produce better machines in the old architecture. Increasing specialization means that no single person can think of a new model, create the best algorithm to solve it, implement that in computer code, and design the computer to run it. No research group is even likely to have all the necessary skills. Thus, we must again somehow resolve the "friction" though a mixture of negotiation, compromise and effort in multiple areas, including compromises with other research groups in different countries using the same computer designs for entirely different purposes.

The interaction between models and data also produces a choice of where to spend effort. Should we focus on better data or on better modeling? The book describes the history of both efforts, showing how they are complementary rather than competitive. Some improvements in modeling, such as the use of progressively smaller mesh sizes, have required both more data and also more computing power and better algorithms. Where you spend your effort also depends on what you choose as your goal. This book largely aims at better long-term climate modeling and specifically measuring the effect of human activity on the climate. Other users of weather observations care more about short-term storm forecasting. Some of the model improvements that are reducing the variance of climate models have little impact on surface forecasts. Even within the group that cares about local forecasts, goals may differ: farmers care a great deal about precipitation and little about clouds, while airplane pilots care a great deal about clouds at different altitudes and less about precipitation.

One aspect of the data-model symbiosis is less relevant for climate, and that is the determination of which data points would be most useful for learning more. Sometimes, to achieve more accurate predictions, you can ask the model "what data point would most reduce uncertainty?" and then go measure it. For example, Jaime Carbonell has built language translation systems which try to improve by posing a small number of questions to a bilingual speaker, recognizing that large amounts of bilingual text including a rare language will rarely be available. For climate modeling, unfortunately, we have no way of asking for more observations of the upper atmosphere in 1800, much as we would like them. Part of data friction is not having the data you most need, or not being able to convert what you have to what you want.

Similar problems affect datasets beyond weather and climate, such as economic data, as mentioned in the book. Economic data is even harder to deal with, since in addition to improvements in measurement technique and random measurement errors, there are more often motivations to lie about economic measures. For example, running a balance of payments surplus is considered "bad" while running a deficit makes a country deserving of sympathy. Decades ago Oskar Morgenstern observed (*The Validity of International Gold Movement Statistics*, Princeton University Press, 1955)

that if you compared all the reported gold transfers in the world, they didn't match up; similarly, the balance-of-trade numbers that are published around the world do not add up to zero.

Either trade goods are appearing out of thin air or some governments are "editing" their statistics. For example, it is common to report exports with FOB prices and imports with prices including transportation. Medical data have similar problems, plus additional ethical issues preventing us from measuring some things we would like to know. So medical researchers are constantly using correlates of the data that would be more relevant (e.g., studies in mice rather than humans) and then they must judge how far health or policy recommendations can be based on these indirect data. Indeed, some writers who once challenged the danger of second-hand cigarette smoke continued on to become climate change skeptics (the late Frederick Seitz and S. Fred Singer are examples).

Prof. Edwards feels that these problems are unavoidable. The best we can do is to get the most accurate data we can, understand the data and how it was gathered as well as possible, have the results reviewed by as many competent people as possible, and then see what conclusions we can draw. This is not going to be satisfactory to many. It still leaves us with studies that a typical person or policy maker will not understand, and almost always with a plethora of studies that do not agree perfectly (although the agreement on global warming is very good). I cannot imagine any Congressmen or Senators reading this entire book themselves, so they will be back to "who can we trust?" That has resulted in a variety of ad hominem attacks on the people who analyze climate data, since the process is not and never can be purely mechanical. We end up trusting scientists, not science.

The uncertainty of all studies, and the reality that research is never finished, also produces an opening for denialists. If I say something like "global warming is caused by vampires" and somebody calls that statement obvious nonsense, I demand to be shown a randomized double-blind trial proving that it's false; lacking such an experiment, I continue arguing that nothing else should be done until proper research has eliminated the vampire theory. If you replace vampires above by sunspots, cosmic rays or volcanoes, you'll get claims that have been proposed seriously.

People then try to argue from anecdote. Doubt global warming? The Iditarod has abandoned its traditional starting point, since there has not been enough snow to start in Wasilla since 2002. In 2008 an old woman in Cape Dorset (latitude 64N), when asked about changes, said that there were now robins around; they never used to have robins. Although impressive, the use of anecdotes is risky. Perhaps there is less snow in Alaska, but in Washington, DC, where normally snow is as rare as an unindicted Illinois governor, there was record snowfall in 2010. For something as variable as weather and climate, we need to try to do better than a few examples. Prof. Edwards explains how you do that, but it isn't simple.

The book does persuade the reader that the result is reliable. So many different people from different institutions have looked at so many scattered data sources that we cannot credibly fear a grand conspiracy extending from tree-ring scholars to radio telemetry experts. A lot of effort has been needed to process climate data; the book mentions 250 papers per year on data re-analysis. Multiple international groups review all of this data and decide on its quality. All of this has been going on for decades.

The models also are important to support the data. Our ability to predict surface weather convinces people that the global circulation models must be valid; and that suggests that the climate models are also valid. Data without models is often not enough. For example, continental drift was dismissed despite the obvious fit of the east and west shores of the Atlantic until plate tectonics provided an explanation of the mechanism. Similarly, the models of global climate enable us to interpret the data that we are gathering and give them practical meaning.

So this book is really an argument for "data curation" – for an increase in the effort spent learning how to gather and unify data for later use. This will never be wholly successful with climate. Inherently, to measure a small effect in the presence of a large random variation is always going to be difficult, even if your data and your models are perfect. We will be left with a difficult national and international political problem, and we're making very little progress on that. I can only say that whatever the outcome of the discussions about climate change, there are so many other examples of what the author calls "computational friction" and "data friction" that learning more about data curation will help in many areas of knowledge.

Michael Lesk
E-mail: lesk@acm.org