ANDREA SALTELLI
SILVIO FUNTOWICZ

When All Models Are Wrong

More stringent quality criteria are needed for models used at the science/policy interface, and here is a checklist to aid in the responsible development and use of models.

"Beware the rise of the government scientists turned lobbyists," trumpeted the headline on an article by British journalist George Mombiot in the left-leaning newspaper *The Guardian*, adding that "From badgers to bees, government science advisers are routinely misleading us to support the politicians' agendas." The article, published on April 29, 2013, criticized the current chief scientist at the UK's environment department for his assessment of the desirability of culling badgers, and the British government's new chief scientist for his opposition to the European ban on the pesticides blamed for killing bees and other pollinators.

From the other side of the ocean and the political spectrum, Rep. Chris Stewart (R-UT) asked (rhetorically) during a U.S. congressional hearing in July 2013 whether the federal Environmental Protection Agency's study of shale-gas fracking "is a genuine, fact-finding, scientific exercise, or a witch-hunt to find a pretext to regulate."

Wherever one stands on these specific issues, such skepticism seems increasingly common, and increasingly independent of ideological position. Science is facing a question: Does the tone of these sorts of attacks reflect a collapse of trust in the scientific enterprise and in its social and institutional role?

Scientific leaders have long portrayed their enterprise as a self-regulating community bound to a higher ethical commitment to truth-telling than society as a whole. Yet the tone and intractability of controversies ranging from badgers to bees to fracking suggests that society may be less willing to accept such claims than in the past.

Perhaps with good reason. The October 19, 2013, cover article of *The Economist*, a nonspecialist periodical with a centrist, proscience political stance, asserts: "Scientists like to think of science as self-correcting. To an alarming degree, it is not." It goes on to recommend that "checklists . . . should be adopted widely, to help guard against the most common research errors. Budding scientists must be taught technical

skills, including statistics, and must be imbued with skepticism towards their own results and those of others."

The scientific enterprise seems only slowly to be awakening to this problem and its dangers. In 2011, *Science* magazine published a special series of articles on reproducibility problems in several disciplines, while its counterpart *Nature* published an article by a pharmaceutical industry executive suggesting rules to spot suspect work in preclinical cancer papers published in top-tier journals. The journal *Organic Syntheses* accepts only papers whose syntheses can be reproduced by a member of the journal's editorial board, and Science Exchange, a commercial online portal, has launched a Reproducibility Initiative that matches scientists with experimental service providers. Meanwhile, the number of retractions of published scientific work continues to rise.

Against this background of declining trust and increasing problems with the reliability of scientific knowledge in the public sphere, the dangers for science become most evident when models—abstracts of more complex real-world problems, generally rendered in mathematical terms—are used as policy tools. Evidence of poor modeling practice and of negative consequences for society abounds. Best-selling books by Nassim Taleb and Joseph Stiglitz have documented for public consumption the contributions of models to recent financial disasters; just two examples of what seems to be a proliferation of books, reports, and papers that lambast the role of economists and mathematicians in pricing a class of derivatives at the heart of the subprime mortgage crisis. Even the Queen of England got into the act, questioning the London School of Economics' economists on why they did not see the crisis coming.

The situation is equally serious in the field of environmental regulatory science. Orrin Pilkey and Linda Pilkey-Jarvis, in a stimulating small volume titled *Useless Arithmetic: Why Environmental Scientists Can't Predict the Future*, offer a particularly accessible series of horror stories about model misuse and consequent policy failure. They suggest, for example, that the global change modeling community should publicly recognize that the effort to quantify the future at a scale that would be useful for policy is an academic exercise. They call modeling counterproductive in that it offers the illusion of accurate predictions about climate and sea level decades and even centuries in the future. Pilkey and Pilkey-Jarvis argue that given the huge time scales, decisionmakers (and society) would be much better off without such predictions, because the accuracy and value of the predictions themselves end up being at the center of policy debates, and distract from the need and capacity to deal with the problem despite ongoing uncertainties.

Wrong but useful

In this light, we wish to revisit statistician George E. P. Box's 1987 observation that "all models are wrong but some are useful." We want to propose a key implication of Box's aphorism for science policy: that stringent criteria of transparency must be adopted when models are used as a basis for policy assessments. Failure to open up the black box of modeling is likely to lead only to greater erosion of the credibility and legitimacy of science as a tool for improved policymaking. In this effort, we will follow *The Economist*'s recommendations and provide a checklist, in the form of specific rules for achieving this transparency.

The specialized literature now makes clear that the more one understands climate, the more model predictions of specific climate futures become uncertain: The Intergovernmental Panel on Climate Change produces larger, as opposed to smaller, prediction uncertainty ranges as more and more processes, scenarios, and models are incorporated and cascading uncertainties make their effect felt in the final estimates.

Climate dynamics are complex and climate science is hard, so it would be a mistake to expect otherwise. Still, the discourse on climate is populated by crisp numbers based on mathematical modeling. An example is the often-mentioned 50% probability that global temperature would not increase more than 2° Celsius (a climate policy target) if humankind succeeds in keeping the greenhouse gas concentration at or below 450 parts per million of carbon dioxide equivalent, which is a measure for describing how much global warming a given type and amount of greenhouse gas may cause. These model-generated numbers are of course nowhere near as crisp as they appear, and even a standard sensitivity analysis would reveal huge uncertainty bounds once the uncertainties associated with each input assumption were propagated through the models. Many small uncertainties multiplied together yield huge aggregate uncertainties.

The challenge becomes even more daunting when modelers turn their attention to the economic consequences of changes in atmospheric composition. For example, the well-regarded *Review on the Economics of Climate Change*, conducted by a team led by British economist Nicholas Stern, quantifies the economic impact of climate change through a cost/benefit analysis that computes fractional losses in gross domestic product 200 years from now. Such an effort is so remote from current predictive capacity as to verge on the irresponsible. What are the uncertainties associated with these predictions? No one has any idea. In this way, the legitimacy of useful tools such as cost/benefit analysis is undermined.

A common indicator of pseudoscience is spurious precision, for example, when a result is given with a number of digits exceeding (at times ludicrously) any plausible estimate of the associated accuracy.

Overreliance on model-generated crisp numbers and targets recently hit the headlines again in the relation to the 90% ratio of public debt to gross domestic product stipulated by Harvard professors Kenneth Rogoff and Carmen Reinhart. Debt ratios above the threshold were considered by these authors as unsafe for a country, but a later reanalysis by researchers from the University of Massachusetts at Amherst disproved this finding by tracing it to a coding error in the authors' original work. This particular instance of error became subject to self-correction, but most aspects of most models will not be subject to such close scrutiny. Critically, the error was corrected too late and much of the damage could not be undone, as the original model results kept austerity-minded economists trading blows with their antiausterity counterparts on the merits and demerits of balanced budgets and austerity policies, a battle that dominated the financial press for months, was in no way defused by the repudiation of the Rogoff-Reihnart results.

Concerns about the usefulness or relevance of modeling are no longer confined to the scientific literature or to expert blogs (such as www.allmodelsarewrong or www.wattsupwiththat.com) but have become part of the public discourse. The beliefs of the public and policymakers about what should be done on climate (or on the economy, or on many other less currently resonant issues) are relying on what models are forecasting about the future, with little if any sensitivity to the limits on what the models are actually capable of forecasting with any accuracy.

Vigilance aided by rules

Enhanced vigilance is hence needed in drawing model-based inferences for policy. To this end, we propose seven rules that together add up to a checklist for the responsible development and use of models. The checklist covers the entire modeling process, and it draws on methods and strategies from two formal processes for assessing uncertainty, one called global sensitivity analysis and the other commonly known by the acronym NUSAP. Methods for applying our rules are taught to European Commission (EC) staff members in charge of carrying out impact assessments to gauge expected costs and benefits of policy initiatives in preparing EC proposals.

Global sensitivity analysis assesses the relative importance of input factors in terms of the impact of their uncertainty on the model output. For example, suppose an economic analysis of policies aimed at job creation included a number of different uncertain elements, such as discount factors, sector-specific policy compliance rates, elasticity factors, and so on. These make their cumulative effect felt in the uncertainty of the expected outcome of the policy; that is, the increase in the number of jobs. Global sensitivity analysis would ask: Which of those uncertainties has the largest impact on the result? A typical response could be: If one could reduce (for example) the uncertainty around the discount factor, then the uncertainty in the number of new jobs would decrease the most. Discussion then focuses on the appropriate topic—What assumptions about the discount rate are being made? How do practitioners agree or disagree about these factors?—rather than on prematurely injecting the model result itself (a prediction about job creation) into the center of the policy discussion.

NUSAP collectively integrates five qualifiers represented in its name. These are numeral (the numerical value of the claimed quantity), unit (of the numeral), spread (a measure of statistical or measurement error), assessment (of the reliability of the claim made by experts), and pedigree (an assessment of the overall quality of the method itself: modeling, data acquisition, expert elicitation, and so on). Assessment and pedigree in particular reveal how (and by whom) the information was produced and document the contentiousness of any claims. The NUSAP approach has been adopted in the Netherlands as part of the Netherlands Environmental Assessment Agency's Guidance on Uncertainty Assessment and Communication.

Our seven-rule checklist amounts to a process that we call "sensitivity auditing." "Sensitivity," as noted, refers to the effort to understand the different sources of uncertainty and their relative importance. "Auditing" emphasizes the

Simpler models enable scientists and stakeholders alike to understand how assumptions and outputs are linked. Complex and often overparameterized mechanistic models should be used only for more speculative investigations outside of the policy realm.

idea of accountability to a broader audience (in this case, policymakers and the public) and thus demands that the model must be accessible and transparent and that expertise is not narrowly defined to exclude everyone but those who created the model. Sensitivity auditing does not aim to improve the model; rather, like a tax audit, it comes at the end of the process, at the point when the model becomes a tool for policy assessment, when all possible model calibration, optimization, data assimilation, and the like have been carried out by the developers using the tools of their craft. And as with tax audits, sensitivity audits aim to help keep those who run the numbers honest.

Rule 1: Use models to clarify, not to obscure. Who owns the model? What are the owner's interests? Is the model proportionate to the task? Is it used to elucidate or to obfuscate?

Today there is little doubt about the links between the 2008 credit crunch and the mathematical models used to price the financial products at the heart of the crisis. As Joseph Stiglitz noted in his book *Freefall*: "[P]art of the agenda of computer models was to maximize the fraction of, say, a lousy sub-prime mortgage that could get an AAA rating, then an AA rating, and so forth . . ." That is, the models were used not to provide insight into the risks of the subprime mortgage industry, but rather to conceal and propagate those risks. Applying rule 1 would have made it clear that predatory lending and irresponsible risk-taking were concealed and sanitized by the mathematics used to price the financial products by those who would make a profit from them.

Another example of obfuscatory use of mathematical modeling is the Total System Performance Assessment model used for evaluating the safety of nuclear waste disposal, which according to Pilkey and Pilkey-Jarvis includes 286 interacting submodules, thus making it inherently incomprehensible even to experts. Models should illuminate complexity, not create it.

Rule 1 prescribes that questions must be raised about who benefits from the model and what motivations and incentives animate model developers. Further, models must be open to and available for skeptical assessment by prospective sensitivity auditors. If the models are too complex to be assessed by well-informed nonexperts, then there is no way to know what biases they may contain or how plausible their results might be.

Rule 2: Adopt an "assumption hunting" attitude. What are the (possibly tacit) assumptions underlying the analysis? What coefficients or parameters had to be given a numerical value in order to make the model work? What was considered irrelevant?

Models are full of implicit assumptions. These may have been made early in the model construction history and henceforth taken for granted within the community of the developers, meaning that they are unlikely to be recognized by the community of the model's users. Sensitivity auditors must go assumption hunting as a necessary step for model appraisal, for example, by applying the NUSAP methodology. John Kay, a prominent British economist, makes the point vividly by exposing the "making up" of the missing data that is needed to operate policy-related models: "To use Britain's Department of Transport scheme for assessing projects, you have to impute values of time in 13 different activities, not just today, but in 2053. Fortunately, you can download answers to these questions from the official website. And answers to many others you probably did not know you wanted to ask. What will be average car occupancy rates, differentiated by time of day, in 2035?"

Detailed case studies of modeling activity in policy-relevant problems as diverse as climate change, nuclear waste disposal, and beach-erosion assessment show that many model assumptions are themselves the result of a negotiation process among scientists with different perspectives on the problem; that is, the assumptions are value-laden. This is probably unavoidable, but it does not have to be rendered invisible by model complexity.

What are the implications of "discovering" implicit assumptions? One possible outcome is that stakeholders in regulatory science debates may judge a model's assumptions to be either implausible or contentious. This may well simply add to existing debates, but the standard response to

disagreement—do more modeling (with more embedded assumptions)—is not going to resolve these debates anyway. New processes for model development and use are required, in which engaged stakeholders work with disciplinary experts to develop new models that can be used to test various policy options. For example, S. N. Lane of Durham University and coauthors described in a 2010 article in the *Transactions of the Institute of British Geographers* a "coproduction of knowledge model," in which a team of experts and laypeople joined together to investigate the problem of how to deal with recurrent flooding. The collaborative work led the team to discard off-the-shelf hydrogeological models and to co-produce a new model that led to hitherto unthought-of options for reducing flood risks.

The White House Office of Management and Budget, in its *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies*, requires that models used to guide regulatory decisions be made available to a third party to enable assessment of the impact of changing the input assumptions or parameters on the model-based conclusion. This approach would give any stakeholder the possibility of obtaining a different conclusion by just changing the input. Of course, the guidelines have the potential downside of allowing sensitivity audits to be used as a pretext to obstruct the regulatory process. Yet the best approach to combating obstructionism is not to protect models from scrutiny, but to ensure that models are developed and used in an appropriate and open manner to begin with, as our rules stipulate.

Rule 3: Detect pseudoscience. Check that uncertainty has not been over- or underestimated to yield a result that advances the model proponents' preferences.

For sensitivity auditing, we are defining pseudoscience as the practice of ignoring or hiding the uncertainties in model inputs in order to ensure that model outputs can be linked to preferred policy choices. A common indicator of this kind of pseudoscience is spurious precision, for example, when a result is given with a number of digits exceeding (at times ludicrously) any plausible estimate of the associated accuracy. As the great mathematician C. F. Gauss noted, "lack of mathematical culture is revealed nowhere so conspicuously as in meaningless precision in numerical computations." The problem is nicely illustrated by the story of the museum guard who tells a group of visitors that the fossil they are viewing is "60 million and 9 years old." He can be this precise because, when he began working at the museum nine years before, he was informed that the fossil was 60 million years old.

The tendency toward overprecision is not limited to jokes. *The Review on the Economics of Climate Change*, for example, declares that "By 2100, in South Asia and sub-Saharan Africa, up to 145-220 million additional people could fall below the $2-a-day poverty line, and every year an additional 165,000-250,000 children could die compared with a world without climate change." But how can one reasonably believe that over such a time scale, a set of models can predict the number of people living in poverty, or the death rates of children, with a precision of roughly 20%, given that these variables are influenced by so many independent factors and that long-term regional impacts of climate change are unpredictable? This is what Nassim Taleb calls the delusion of uncertainty.

Rule 4: Find sensitive assumptions before they find you. Do not publish your inferences without having done a careful sensitivity auditing.

One of the 10 commandments of applied econometrics according to a popular handbook by the late Peter Kennedy is: "Thou shall confess in the presence of sensitivity. Corollary: Thou shall anticipate criticism." The reason is that once an unacknowledged assumption is publicly exposed, the entire model is effectively falsified for public purposes, even if it has valuable insights to impart.

We turn again to the *Review on the Economics of Climate Change* for an example. Here, the sensitivity analysis was performed only after economist William Nordhaus published a critique of the implausible discount factors used in the original models. When after-the-fact sensitivity analysis reveals problematic or arguable assumptions, a stakeholder might reasonably then ask: What were the motives behind the use of such implausible assumptions? Such questions can badly damage the credibility of an otherwise useful modeling exercise. Of course, economists and other scientists studying complex systems have the right to use model-based narratives. But when these narratives feed into the policy process, the standard of quality for models must be high, lest model use falls into disrepute and stakeholders reject the use of models altogether, as has happened in the arenas of toxic chemical regulations and food and nutrition policy.

Any model-based inference that is introduced into the policy environment unaccompanied by a technically sound sensitivity analysis should be regarded as suspicious. In the absence of a prior sensitivity analysis, model developers and proponents of the inference are accountable for explaining why it was dispensable.

Rule 5: Aim for transparency. Stakeholders should be able to make sense of and, if possible, replicate the results of the analysis.

Lack of transparency of procedures in science is another source of legitimate public distrust, well displayed in the recent "Climategate" row, where emails among climate scientists (mischievously exposed by critics of climate science) revealed scientists discussing the "craft-skill" assumptions and shortcuts they used to help manage the huge complexities involved in climate modeling, and the tactics they used to keep their critics at bay. Of course, this is a problem only because scientists encourage their enterprise to be portrayed as an overidealized Elysium of objectivity and personal rectitude, rather than as a human and social endeavor. The error was compounded when scientists who had sent the emails refused to provide further access to their data because they were not legally required to do so. The UK Royal Society formed a working party that issued a strong criticism in a June 2012 report titled *Science as an Open Enterprise*, which said that the use of the UK Freedom of Information Act to justify blocking access to the data "reflects a failure to observe what this report regards as . . . a crucial tenet for science, that of openness in providing data on which published claims are based."

A corollary of the transparency rule is that simple or parsimonious model representations are better than more "sophisticated" or complex models, when they are being used for policy impact assessments. Simpler models enable scientists and stakeholders alike to understand how assumptions and outputs are linked. Complex and often overparameterized mechanistic models should be used only for more speculative investigations outside of the policy realm.

Transparency also demands comprehensibility. A good sensitivity analysis should be understandable by those with a stake in the results and communicated in plain English, with minimal jargon. As an example of what to avoid: "If we could get rid of the uncertainty in the ingestion rate, then the overall uncertainty of the predicted health effect (the "variance," in statistical parlance) would be reduced by 40%."

Rule 6: Don't just "Do the sums right," but "Do the right sums." When relevant stakeholder viewpoints are neglected, modelers may focus on or address the wrong uncertainties.

In an impact assessment study, a type-one error is a false positive: A practice is determined to be unsafe when it is safe, or an intervention nonbeneficial when it is beneficial. A type-two error is the reverse: a false negative. A type-three error, in contrast, is one where the analysis itself is framed incorrectly and thus the problem is mischaracterized.

In modeling, as in everyday life, type-three errors are the most dangerous because they can leave the real problem unaddressed, waste resources, and impede learning.

When performing an uncertainty and sensitivity analysis, one may easily fall into what we can call "lamp-posting," whereby the uncertainties or parameters that are more carefully scrutinized are those that are the least relevant but easiest to analyze. (The term refers to Mullah Nasruddin's story about the drunkard looking for his lost keys not where he lost them but under a street lamp because that is where it was light enough to look.)

Type-three errors may arise when modelers worry about the elegance of their models but fail to appreciate the way in which stakeholders may understand the problem and its context. For example, in the flood modeling case mentioned under rule 2, years of modeling stream flow and cost/benefit ratios for flood protection structures had failed to consider an alternative intervention—upstream storage of flood waters—until local stakeholders were brought into the modeling process. One specific reason for this neglect is worth highlighting. According to Lane and colleagues, upstream storage was neglected in the models because of the "use of a pit-filling algorithm that made sure that all water flows downhill"!

Rule 7: Focus the analysis. Do not do perfunctory sensitivity analyses, merely changing one factor at a time.

Sensitivity is often omitted in modeling studies, or it is executed in a perfunctory fashion. Many sensitivity analyses seen in the literature are run without a statistical design, moving just one input factor at a time away from a pre-established baseline and always using that baseline as starting point. But this move just scratches the model's uncertainty; for example, in a system with 10 uncertain factors, moving just one at a time risks exploring only a tiny part of the total potential input uncertainty.

A baseline is customarily selected so that all uncertain factors have their "best" or reference value; hence, the reluctance of modelers to depart too severely from it. Different studies will have different input variables and different baseline values, for example, for a discount rate, for the permeability of a geological formation, or for the maximum sustainable level at which a crop can be harvested. These baselines, in turn, may be the average of a set of measurements, the opinion of one or more experts, or a number generated by another model. That is, the choice of a baseline is itself an assumption-laden process and thus itself subject to criticism and sensitivity analysis. Perhaps the discount rate is too sensitive, or not sensitive enough, to the needs of future generations; or the rock permeability too high (or low) because the existence of faults has been neglected (or overplayed); or the role of pest infestations (or innovation to reduce infestations) in agricultural practice has been neglected.

A credible sensitivity audit must not be anchored to baselines that are themselves subjective; they must evaluate the

When relevant stakeholder viewpoints are neglected, modelers may focus on or address the wrong uncertainties.

effect of any input while all other inputs are allowed to vary as well.

Improving society's choices

The checklist we offer should aid developers and users of mathematical models as they pursue applications ranging from engineering projects to environmental regulation to economic performance. The rules become cogent at the moment when the model is used to inform policy; for example, in the context of an impact assessment study. The rules cannot by themselves ensure that the model will be a reasonable and relevant representation of the situation being modeled. But if they have not been followed, people with an interest in the results of the model have good reason to be skeptical of both the motives of the modelers and the plausibility of model outputs.

Our view is that current modeling practices, in their development and use, are a significant threat to the legitimacy and the utility of science in contested policy environments. The commitment to transparency and parsimony that the rules demand will encourage modelers themselves to focus on parameters, inputs, assumptions, and relationships that are well constrained and understood. Further, the assumption-laden aspects of the system should be clearly spelled out. The result will be greater credibility for models and greater clarity about what aspects of difficult policy choices can appropriately be constrained by modeling, and what aspects need to be turned over to democratic political institutions. It is these institutions, after all, to which society appropriately assigns the task of making decisions in the face of irresolvable uncertainties; and it is the process of open debate about such decisions, not technical arguments about models, that can make clear the links between political interests and policy preferences.

Recommended reading

S. O. Funtowicz and J. R. Ravetz, *Uncertainty and Quality in Science for Policy* (Dordrecht, Netherlands: Springer, 1990).

S. N. Lane, N. Odoni, C. Landström, S. J. Whatmore, N. Ward, and S. Bradley, Doing Flood Risk Science Differently: An Experiment in Radical Scientific Method. *Transactions of the Institute of British Geographers* 36, 15–36 (2011).

M. Maslin and P. Austin, Climate Models at Their Limit? *Nature* 486, 183 (2012).

Office of Management and Budget, *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies*, Friday, February 22, 2002. Notice; republication http://www.whitehouse.gov/omb/fedreg_final_information_quality_guidelines/, last accessed December 2012.

N. Oreskes and E. M. Conway, *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming* (London: Bloomsbury Press, 2010).

O. H. Pilkey and L Pilkey-Jarvis, *Useless Arithmetic. Why Environmental Scientists Can't Predict the Future* (New York: Columbia University Press, 2007).

A. Saltelli, A. Guimarães Pereira, J. P. van der Sluijs, and S. Funtowicz, What Do I Make of Your Latinorum? Sensitivity Auditing of Mathematical Modeling (http://arxiv.org/abs/1211.2668). To appear in *Foresight and Innovation Policy*, Special Issue on Plausibility.

A. Saltelli, B.D'Hombres, Sensitivity Analysis Didn't Help. A Practitioner's Critique of the Stern Review. *Global Environmental Change* 20, 298–302 (2010).

Trouble at the Lab, Scientists Like to Think of Science as Self-Correcting. To an Alarming Degree, It Is Not. *The Economist*, October 19, 2013, 21–24.

Andrea Saltelli (andrea.saltelli@jrc.ec.europa.eu) is at the Joint Research Centre's Institute for the Protection and Security of the Citizen at the European Commission. Silvio Funtowicz is at the Centre for the Study of the Sciences and the Humanities (SVT) at the University of Bergen, Norway.