Global Environmental Change 33 (2015) 23–31

Contents lists available at ScienceDirect

Global Environmental Change

journal homepage: www.elsevier.com/locate/gloenvcha

Approaches to evaluating model quality across different regime types in environmental and public health governance

Brian H. MacGillivray [a,*], Keith Richards [b]

[a] Sustainable Places Research Institute, 33 Park Place, Cardiff University, UK
[b] Department of Geography, University of Cambridge, UK

ARTICLE INFO

Article history:
Received 29 October 2014
Received in revised form 31 March 2015
Accepted 8 April 2015
Available online

Keywords:
Science policy
Model evaluation
Risk regulation
Environmental governance
Civic epistemologies
Quality evaluation

ABSTRACT

A reliance on mathematical modelling is a defining feature of modern global environmental and public health governance. Initially hailed as the vanguard of a new era of rational policy-making, models are now habitually subject to critical analyses. Their quality, in other words, is routinely queried, yet what exactly is quality in this context? The prevailing paradigm views model quality as a multi-dimensional concept, encompassing technical dimensions (*e.g.* precision and bias), value judgments, problem-framing, treatment of “deep” uncertainties, and pragmatic features of particular decision contexts. Whilst those technical dimensions are relatively simple to characterise, the broader dimensions of quality are less easily formalised and as a result are difficult to take account of during model construction and evaluation. Here, we present a typology of governance regimes (risk-based, precautionary, adaptive and participatory) that helps make explicit what these broader dimensions of model quality are, and sketches out how the emphasis placed on them differs by regime type. We show that these regime types hold distinct positions on what constitutes sound evidence, on how that evidence should be used in policy-making, and to what social ends. As such, a model may be viewed within one regime as providing legitimate evidence for action, be down-weighted elsewhere for reflecting a flawed problem-framing, and outright rejected in another jurisdiction on the grounds that it does not cohere with the preferred ethical framework for decision-making. We illustrate these dynamics by applying our typology to a range of policy domains, emphasising both the disconnects that can occur, as well as the ways that modellers have adapted their practices to ensure that their evidence is brought to bear on policy problems across diverse regime types.

1. Introduction

What do the policy responses to swine flu, climate change, and the 2010 eruption of Eyjafjallajökull have in common? All were based in part on the use of mathematical models, which has been one of the defining features of public policy-making in recent decades. This widespread and prominent role for modelling has many drivers. It stems in large part from advances in computational power, mathematical methods, and in our theoretical understanding of a range of social, economic, and physical phenomena. Perhaps most importantly, such modelling techniques purported to usher in a new form of governance, wherein public policies would be developed based on neutral, rigorous evaluations of their likely consequences (Tribe, 1972; Sarewitz and Pielke, 1999). This rather technocratic rationale did not lack opposition, particularly from advocates of deliberative democracy (Dryzek, 1993). But what Porter (1995) called the “pursuit of objectivity” nevertheless held substantial influence on public policy. The idea was that decisions would no longer be driven by vested interests, mere speculation, ideology or horse-trading, but instead would find their basis in objective technical analysis. Of course today this seems a touch utopian, and models with the potential to inform public policy are now routinely subject to rigorous critical analyses by regulators and model developers alike. These analyses (*e.g.* NRC, 2007) focus on issues including the plausibility of modelling assumptions; precision and bias; the adequacy of the treatment of uncertainty; and the value judgments that models may implicitly or explicitly encode (*e.g.* in the choice of impact variables to be included). Model *quality*, in other words, is routinely queried. By quality, we mean the properties that are desirable in a model. We interpret quality

* Corresponding author. Tel.: +44 029 2087 9271.
E-mail addresses: macgillivraybh@cardiff.ac.uk (B.H. MacGillivray), keith.richards@geog.cam.ac.uk (K. Richards).

http://dx.doi.org/10.1016/j.gloenvcha.2015.04.002
0959-3780/

broadly, as relating to the contents of the model (*e.g.* the entities included, the variables that characterise them, and the equations or algorithms that relate them), the model's formal properties (*e.g.* precision and bias), and the processes through which the model is developed. When a model is evaluated as being of sufficient quality, it is often referred to as adequate or sound. As will become clear, we think that model quality is contingent and multi-dimensional. More on this later. To begin with, what does quality in the context of models-for-policy mean? What sort of properties are we dealing with?

These questions have crucial implications for the development, evaluation, and use of models in the policy context. Quality, after all, is perhaps the fundamental goal that researchers pursue in the model development stage; is the basic arbiter of whether a model is accredited by peers or regulators; is used to discriminate between competing plausible models; and shapes the level of confidence that policy-makers hold in model outputs. In what follows, we first sketch out a brief history of conceptions of model quality. We describe a shift from the originally dominant statistical paradigm, to a present approach that considers quality to be both a multi-dimensional concept and a function of how the model relates to the decision-making task it was designed to fulfil. We then argue that the question of what constitutes a good model is conditional upon the nature of the governance regime in which the model is to be applied. We identify a small set of governance types that resonate with different uses of scientific evidence: risk-based, precautionary, adaptive, and participatory. These governance regimes hold distinct norms about what constitutes valid evidence, how evidence should be used, and what constitutes the proper ethical framework for decision-making (*e.g.* means-ends *vs.* communicative rationality). As such, a model may be viewed within one regime as providing legitimate evidence for action, be down-weighted elsewhere for reflecting a flawed ontology (*e.g.* privileging universal over contextual knowledge), and be rejected outright in another jurisdiction on the grounds that it fails to align with the preferred ethical framework for decision-making. This is a blindspot of the prevailing quality evaluation paradigm (*e.g.* NRC, 2007), and one which diminishes the capacity to understand and improve the use of science in policy-making

And so our argument is that model quality is not independent of, but rather is intertwined in complex ways with the types of governance regimes that models seek to inform. The corollary is that there can be no unitary set of criteria by which all models can or indeed should be evaluated. This has implications for how models are built, how they might be scrutinised, and how policy-makers use them. Our contribution is twofold. The first is primarily theoretical: our typology of governance regimes has significant explanatory power when applied to the science policy interface across various jurisdictions and policy domains. That is, it helps us understand the ways that models have been used – and sometimes neglected – in particular cases. The second relates to the practical aspects of model building and evaluation. Although some policy domains have their own detailed rules covering the model-building process (and so there is less chance of mismatch between the models that are built and the models that are desired), there are many exceptions to this. This means that general, but tractable guidance on the sorts of model building practices that are favoured (or not) by different sorts of regulators might be of practical use to modellers. More concretely, as well as highlighting mismatches between distinct modelling practices and different regime types, we also discuss several examples where modellers have adapted their methodological approaches to ensure that their evidence was brought to bear on policy questions across a range of regime types, without sacrificing technical quality.

2. Scope, concepts, and definitions

Our argument about the relationship between different governance types and what is perceived as model quality is in part a logical one; however, we discuss a range of examples from several policy domains, necessarily in a somewhat schematic way. This ensures that our argument is empirically grounded, and helps to flesh out its implications and nuances. There are, of course, many different classes of model, and the definition we provide below is not the only way of thinking about models (*cf.* Hastrup, 2013), but is introduced to clarify our scope and to make our argument tractable. Here, we therefore define a class of formal models as purposeful mathematical representations of some real world phenomenon of interest (see Grimm and Railsback, 2005). These are composed of equations, statistical relationships, algorithms, or some combination therein (NRC, 2007). Such models inevitably contain numerous simplifications, approximations, and exclusions, and hence they are never perfect representations of the systems that they aim to characterise (Winsberg, 2014). Moreover, their development is inevitably conditioned by methodological paradigms, computational capacities, and path dependency. Thus, we conceive of these models as unavoidably imperfect decision-making aids, rather than truth machines (Winsberg, 2014). In the policy context with which we are concerned, mathematical models have various functions. They are often concerned with the task of extrapolating beyond known observations, such as predicting or projecting a future (*e.g.* the potential impacts of climate change on species distribution), or answering "what-if" style questions about proposed policy interventions. Models are also applied for the purpose of classifying objects (*e.g.* is this chemical carcinogenic?), or for simply describing relationships amongst variables (*e.g.* statistical models in flood frequency analysis).

Two additional concepts require some definition for the purposes of our exposition: "evidence" and "knowledge." We understand evidence to be some property or material that speaks to the state, mechanics, or future conditions of a phenomenon (*e.g.* a model output). Put another way, evidence makes a difference to what it might be reasonable or justified to believe (Kelly, 2014). Knowledge, by contrast, is often thought of as justified true belief (Steup, 2012). We use the term only when distinguishing between abstract and contextual knowledge. Abstract or general forms of knowledge hold true across time and place (*e.g.* the Navier–Stokes equations). Contextual knowledge is contingent, local, and particular (*e.g.* knowledge of the particular form that general causal mechanisms take in a specific catchment; knowledge of local practices of disposing of radioactive waste, *etc.*).

3. A brief history of quality evaluation paradigms

Models-for-policy were initially viewed as tools that would allow decision-makers to determine the rational course of action in the face of environmental and public health hazards. Reflecting this mindset, early approaches to quality evaluation focussed upon the degree to which a model corresponded with reality. That is, the question of model quality was largely understood in terms of predictive accuracy (bias and precision) and, less significantly, fit to existing datasets. This approach was embodied in categorical tests for "validating" particular models (*e.g.* hypothesis testing), in measures of the difference between observed values and the values predicted by the model (*e.g.* mean square error), and in methods for discriminating between a range of plausible models (*e.g.* "goodness of fit"). These technical dimensions of quality are clearly still remarkably important (see Cox, 2013 for an overview), but a sole reliance on them has encountered criticism from various quarters. For example, Petersen (2006) has argued that quality is not just a question of predictive accuracy, but also of the rigour of the