Energy Policy 101 (2017) 211–215

Contents lists available at ScienceDirect

Energy Policy

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

The importance of open data and software: Is energy research lagging behind?

Stefan Pfenninger[a,*], Joseph DeCarolis[b], Lion Hirth[c], Sylvain Quoilin[d], Iain Staffell[e]

[a] Climate Policy Group, Department of Environmental Systems Science, ETH Zürich, Switzerland
[b] Department of Civil, Construction, and Environmental Engineering, North Carolina State University, USA
[c] Neon Neue Energieökonomik GmbH (Neon), Germany
[d] Institute for Energy and Transport, European Commission Joint Research Centre (JRC), Petten, Netherlands
[e] Centre for Environmental Policy, Imperial College London, UK

ARTICLE INFO

Keywords:
Modelling
Data
Open source
Open data
Transparency
Reproducibility

ABSTRACT

Energy policy often builds on insights gained from quantitative energy models and their underlying data. As climate change mitigation and economic concerns drive a sustained transformation of the energy sector, transparent and well-founded analyses are more important than ever. We assert that models and their associated data must be openly available to facilitate higher quality science, greater productivity through less duplicated effort, and a more effective science-policy boundary. There are also valid reasons why data and code are not open: ethical and security concerns, unwanted exposure, additional workload, and institutional or personal inertia. Overall, energy policy research ostensibly lags behind other fields in promoting more open and reproducible science. We take stock of the status quo and propose actionable steps forward for the energy research community to ensure that it can better engage with decision-makers and continues to deliver robust policy advice in a transparent and reproducible way.

1. Introduction

For nearly a century, the global energy system has remained remarkably stable, powered largely by fossil fuel combustion. However, successfully addressing anthropogenic climate change with low-carbon technologies requires that we fundamentally alter energy supply and demand in the 21st century, yet the pathway and outcomes of this transformation are highly uncertain. For example, rapid improvements in solar photovoltaics and batteries coupled with information technology may point towards a more distributed energy system with its design actively shaped by consumers. Alternatively, large-scale technologies like nuclear, biomass, carbon capture and storage or wind may extend the dominance of a centralised power system.

Given the uncertainty and complexity of the energy system, quantitative models are one of the few available tools that allow analysts to explore alternative scenarios and help guide public policy. Quantitative analysis from energy models underpins much of academic research and energy policy-making (Strachan et al., 2009). Yet most models and data relied upon by utilities, consultancies and public research institutes remain inscrutable "black boxes" – whether econometric models with a small number of parameters, or large linear optimisation models with hundreds of thousands of input variables. In contrast to closed models, "open" models imply that anyone can freely access, use, modify, and share both model code and data for any purpose (Open Knowledge Foundation, 2015). Here, we (1) argue why energy data and models urgently need to become open; (2) discuss the key reasons why many are currently not; (3) examine whether energy research is lagging behind other fields in becoming more open; and finally (4) outline specific issues for individuals to consider and propose next steps for the energy research community.

2. Why models and data should be open

Given the critical guidance that energy models and data provide to decision makers, they should be made open and freely available to researchers as well as the general public. There are four specific reasons for this:

1. **Improved quality of science**. Fundamental scientific principles such as transparency, peer review, reproducibility and traceability are almost impossible to implement without access to models and data (DeCarolis et al., 2012; Nature, 2014). Better adherence to these principles leads to higher quality science. Researchers are

* Corresponding author.
E-mail address: stefan.pfenninger@usys.ethz.ch (S. Pfenninger).

http://dx.doi.org/10.1016/j.enpol.2016.11.046
Received 1 July 2016; Received in revised form 20 October 2016; Accepted 27 November 2016
Available online 02 December 2016
0301-4215/

fallible human beings and errors are inevitable under pressure to deliver. Such mistakes can have profound implications. For example, the Reinhart-Rogoff spreadsheet error arguably skewed the international debate on austerity (Herndon et al., 2014). Such incidents serve as warnings against poor programming practices, such as a lack of auditing as well as closed models and data: it was only through sharing the spreadsheet that the errors were discovered.

2. **More effective collaboration across the science-policy boundary.** Better and more transparent science ought to enable better policy outcomes, but the issue is more complex than that. Academic peer review routinely does not (and cannot) check model arithmetic and data validity, just that the analytical approach is appropriate. A separate process of quality assurance (QA) is required to verify and validate model mechanics and output. While mostly absent from academic practice, this is often implemented as a formal procedure in government (DECC, 2015). The reason for this is that unlike academics, governments, private companies and NGOs often model for numbers rather than insight. The specific numbers can be of great societal importance, such as the level at which to set subsidies or the cost of specific policies. Thus, in many cases, the most important aspect is the quality or transparency of input data, rather than the novelty of the modelling methodology. In large datasets used in government decision-making, traceability and referencing can become major problems, as civil servants developing models and data are often not trained scientists. Openly available, collaboratively developed datasets and reference models would allow the burden of this work to be shared more widely, and across both academia and government. There is a growing sense that the link between energy modelling and policy needs fundamental rethinking (Strachan et al., 2016), and opening up models and data will play a crucial role in enabling the transparency and better quality assurance necessary for this to happen.
3. **Increased productivity through collaborative burden sharing**. Collecting data, formulating models and writing code are resource-intensive. Research funding is limited and researcher time is a scare resource. Society as a whole saves time and money if researchers avoid unnecessary duplication and learn from one another. Individual researchers gain more time to spend on pressing research questions rather than redundant work on model or dataset development. Furthermore, research only matters if it is seen and used, and open-access publishing has been shown to increase readership and citations (McCabe and Snyder, 2014). Since openly shared code or data is more likely to be known to others, it is more likely to be used and further improved. Not only does this benefit the original researcher through peer recognition and academic credit, but moves the research community as a whole forward.
4. **Profound relevance to societal debates**. Reengineering the energy landscape will affect everyone, producing winners and losers. A balanced societal and political debate requires transparent arguments based on scientific justifications, but escalating concern about reproducibility in some fields is shaking public confidence in scientific research (Goodman et al., 2016). Finally, besides the practical considerations outlined above, there remains the ethical argument that research funded by public money should be available to the public in its entirety.

3. Why models and data are mostly not open

Despite these arguments, we see four main reasons why closed models and data may remain attractive and rational in some cases:

1. There is a range of valid ethical and security concerns, particularly in the case of data. Researchers may have access to sensitive commercial data or to data containing personal information (particularly relevant when moving towards more decentralised smart grids with their focus on individual households). The aspiration to open up as much data as possible may give way to a more regulated approach to open data if individual researchers increasingly cross ethical boundaries, as in the recent release of personal data about users of a major online dating website (Resnick, 2016). Setbacks in the wider open data movement could also have repercussions on the use of information perceived as sensitive in the energy modelling context, e.g. data on energy consumer behaviour or on grid infrastructure.
2. Openly sharing details of models, analysis and data can create unwanted exposure. Flawed code or data can discredit research results and cause embarrassment to their authors, but only if they are visible. Indeed, a reluctance to share data was shown to be associated with weaker evidence (Wicherts et al., 2011). Furthermore, there may be a fear that inexperienced researchers use an open model or open data to produce flawed analysis that reflects poorly on its original authors. There is also a policy dimension: government departments may choose to keep information closed precisely because of the potentially serious impact it may have on a country's economy and society, rather than opening the models and data to enable a more transparent political and societal discussion. For example, while the UK Department of Energy and Climate Change (DECC) are working with University College London to develop an open source UK TIMES energy system optimisation model (UCL, 2014), political sensitivities mean that code and data will not be released until its use in a major policy analysis (the UK's 5th carbon budget) is complete (Sargent, 2016).
3. It is time-consuming to write legible and reusable code, track data provenance and processing steps, document models and data and respond to feature requests or bug reports. Because model and dataset development are large investments, it is often rational for researchers and institutions to maintain "trade secrets" to compete in consulting work and third-party research funding. On the one hand, this can be seen as a classical collective action problem where individual actors are trapped in a suboptimal non-cooperative equilibrium. But, as discussed further below, the incentive structure that gives rise to this bargaining problem is also linked to institutional issues within academia, particularly the unrelenting pressure to publish ever-greater quantities of high-quality publications which underlies most academic career incentives and impact metrics (Sarewitz, 2016). A significant share of energy modelling is done in the private sector, in utilities, consulting firms, and financial institutions, where the need to protect the intellectual property within models and data is certainly more pressing than in academia. Nevertheless, where private sector modelling is used to inform public policy and/or where it is funded by public money, we believe the long-term goal should be for models and data to be open, even if this would challenge consultancies' established business practices. While examples of successful open-source businesses exist (e.g. RedHat or Canonical in the Linux world), it is clear that working business models can be difficult to find, especially in the energy field with the added difficulty of balancing commercial and academic principles. The "share alike" clause in licenses like the GPL (see below) may offer opportunities for companies here. Furthermore, private companies and consulting firms are also selling their expertise: energy models must be adapted for specific analyses, and the real value arguably comes from the application of judgement and expertise to adapt and apply the models in a way that produces useful insight.
4. Finally, there is simple institutional and personal inertia, often alongside complex and uncoordinated institutional setups. For example, energy models and datasets are developed and applied by