Weber, N., Thomer, A., & Strand, G. (2013). Complications in Climate Data Classification: The Political and Cultural Production of Variable Names. Advances In Classification Research Online, 23(1), 57-58. doi:10.7152/acro.v23i1.14265

Complications in Climate Data Classification: The Political and Cultural Production of Variable Names

Nicholas M Weber[1], Andrea K. Thomer[1], Gary Strand[2]

[1]*The Center for Informatics Research in Science Scholarship.*
University of Illinois, Graduate School of Library and Information Science.

[2] *Climate Change Prediction Group, Climate and Global Dynamics division, National Center for Atmospheric Research.*

Introduction

Model intercomparison projects are a unique and highly specialized form of data-intensive collaboration in the earth sciences. Typically, a set of pre-determined boundary conditions (scenarios) are agreed upon by a community of model developers that then test and simulate each of those scenarios with individual 'runs' of a climate model. Because both the human expertise, and the computational power needed to produce an intercomparison project are exceptionally expensive, the data they produce are often archived for the broader climate science community to use in future research. Outside of high energy physics and astronomy sky surveys, climate modeling intercomparisons are one of the largest and most rapid methods of producing data in the natural sciences (Overpeck et al., 2010).

But, like any collaborative eScience project, the discovery and broad accessibility of this data is dependent on classifications and categorizations in the form of structured metadata – namely the Climate and Forecast (CF) metadata standard, which provides a controlled vocabulary to normalize the naming of a dataset's variables. Intriguingly, the CF standard's original publication notes, "…conventions have been developed only for things we know we need. Instead of trying to foresee the future, we have added features as required and will continue to do this" (Gregory, 2003). Yet, qualitatively we've observed that this is not the case; although the time period of intercomparison projects remains stable (2-3 years), the scale and complexity of models and their output continue to grow – and thus, data creation and variable names consistently outpace the ratification of CF.

Proposal

This paper describes the use (and lack-of-use) of classification standards in the fifth Climate Model Intercomparison Project[1] (CMIP5). In a convenience sample of CMIP5 data (n ~ 500), we observe that only ~5% of variables are CF-compliant; although, we estimate that up to 12% could be crosswalked to the CF standard.

[1] http://cmip-pcmdi.llnl.gov/cmip5/

ISSN: 2324-9773

Additionally, we've qualitatively observed that there are formal naming conventions in use that, though not CF-compliant, reflect the cultural norms of climate modeling and hence make data 'usable' by those inculcated in the same or similar modeling practices. For instance, 'Q' is the traditional short hand for 'specific_humidity'; 'meridonal' and 'zonal' are often used instead of 'northward' and 'eastward.'

Further complicating the application of standardized naming conventions is that though variables may be sampled in different ways (e.g., surface air temperature vs. upper air temperature), these names still refer to the same physical quantity (Meehl et al. 2007); one is not correct over the other, but rather, they are ontologically equivalent 'things' with different political and epistemological implications.

Our general thesis is that a lack of compliance with CF's standardized naming conventions can be partially attributed to an increased amount of data generated in a relatively small amount of time, – CMIP5 produced ~3380 TB, versus CMIP3's 'mere' 35 TB (Strand, 2011) – but also , as Bowker and Star note, because classification schemes can always be read as political and cultural productions (1999, p. 331).

Citations

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. Cambridge, Mass: MIT Press.

Gregory, J. (2003) *The CF metadata standard 1.0.* Centre for Global Atmospheric Modeling, University of Reading, UK, November 2003, http://cf-pcmdi.llnl.gov/documents/other/cf_overview_article.pdf

Meehl, G., Covey, C., Latif, M., McAvaney, B., Mitchell, J., and R. Stouffer. (2007) IPCC Standard Output from Coupled Ocean-Atmosphere GCMs. WGCM Climate Simulation Panel, retrieved from: http://www-pcmdi.llnl.gov/ipcc/standard_output.html

Overpeck, J. T., Meehl, G. a, Bony, S., & Easterling, D. R. (2011). Climate data challenges in the 21st century. *Science. 331*(6018), 700-2. doi:10.1126/science.1197869

Strand, G. (2011). Community Earth System Model Data Management: Policies and Challenges. *Procedia Computer Science*, *4*, 558-566. doi:10.1016/j.procs.2011.04.058

ISSN: 2324-9773