Data Quality

Alastair Culham & Chris Yesson
Data sources

1. Climate models and their basis
2. Distribution data
3. Other data
Climate model data

- The Intergovernmental Panel on Climate Change
- IPCC1 - 1990
- IPCC2 - 1995
- IPCC3 - 2001
- IPCC4 - 2007

- IPCC provides consensus on what scientists expect to happen
- IPCC5 is on the way
'Show Your Working': What 'ClimateGate' means

VIEWPOINT

Mike Hulme and Jerome Ravetz

The "ClimateGate" affair - the publication of e-mails and documents hacked or leaked from one of the world's leading climate research institutions - is being intensely debated on the web. But what does it imply for climate science? Here, Mike Hulme and Jerome Ravetz say it shows that we need a more concerted effort to explain and engage the public in understanding the processes and practices of science and scientists.

As the repercussions of ClimateGate reverberate around the virtual community of global citizens, we believe it is both important and urgent to reflect on what this moment is telling us about the practice of science in the 21st Century.

In particular, what is it telling us about the social status and perceived authority of scientific claims about climate change?

We argue that the evolving practice of science in the contemporary world must be different from the classic view of disinterested - almost robotic - humans establishing objective claims to universal truth.

Practising scientists know that they do not simply follow a rulebook to do their science, otherwise it could be done by a robot.

Climate change policies are claimed to be grounded in scientific
Can we rely on future climate models?

- They are models, not predictions
- Sound basis in science
- Real observations – $\text{CO}_2$ Now

![Atmospheric CO$_2$ May 1958 - May 2010](image)
How can we trust climate models for the 2080s when we don’t know what the weather is tomorrow?

• Important to understand the difference between:
  – *Climate* dealing with models of general trends
  – *Weather* dealing with predictions of the specific
Can we rely on past climate models?

• Modelling here relies on:
  – Knowing continental positions
  – Knowing altitudes
  – Knowing sea levels
  – Knowing atmospheric gas concentrations

• This can be validated against fossil evidence
  – Pollen/macrofossils
  – ‘Fossil’ atmospheres – from ice cores
Palaeohistory

• Fossil history
  – Mostly pollen

• Geological record
  – Continental drift
  – Climate

• Computer models
  – Climate
Gathering the evidence

- Fossil history is generally poor and patchy even in the best recorded groups.

- Pollen offers the best fossil record for most flowering plants.
Distribution data

• Many Sources
  – Your own validated points
  – Secondary sources
    • Individually validated
    • Batch processed
Accuracy vs Precision

The differences between accuracy and precision in a spatial context. The red spots show the true location, the black spots, represent the locations as reported by a collector.

a. High precision, low accuracy.
b. Low precision, low accuracy showing random error.
c. Low precision, high accuracy.
d. High precision and high accuracy

Chapman 2005 Principles of Data Quality
• Fabaceae data from GBIF showing patchy geographic coverage
Basic errors

- In the sea
- Lat/Long reversals
- Near Valid
- Lat/Long zero
• Global Legume coverage from GBIF data per TDWG level 4 area
Combining data

- *Cyclamen* – a data rich example
- GPS
- New localities
Other issues

• Taxonomy and checklists

• Misclassified data
  – Synonymy
  – Homonymy
  – Misidentification
More taxonomy

- Correctly determined taxa
- Wrongly databased
- *Lictoria achillae*
  - GBIF – listed as *Rhodophyta*
  - Source database – listed as *Lepidoptera*!
Problems with distribution data

• Accuracy of identification
• Disambiguation
• Accuracy of record
• Context of record
• Geographical pattern to records
• http://data.gbif.org/species/
Good data that appear bad
The challenge

1. To get enough data
2. To get accurate and precise data
3. To get correctly identified records
4. To get even geographic coverage
How Global Is the Global Biodiversity Information Facility?

- Chris Yesson¹, Peter W. Brewer¹, Tim Sutton¹, Neil Caithness¹, Jaspreet S. Pahwa², Mikhaila Burgess², W. Alec Gray², Richard J. White², Andrew C. Jones², Frank A. Bisby¹, Alastair Culham¹
- A review of taxonomic and geographic coverage of the GBIF database