Empirical Methods for Testing Climate Models Against Observations

Ross McKitrick
Professor of Economics
University of Guelph

Presented to
World Federation of Sciences Meeting
Erice Italy, August 2016

Model Evaluation

- About 2 dozen General Circulation Models (GCMs) are in use
  - Relied on for major policy decisions as well as scientific research
- As indicated in IPCC reports, field of climatology has not developed consistent statistical methodologies for model evaluation
- Methods usually borrowed from other fields and applied at relatively simple levels
- Econometrics is an example of a field where specific methodologies developed to an advanced level for model evaluation and testing

Model Evaluation

- Sensitivity Analysis *vs* Model Testing

- Sensitivity: Across the range of reasonable parameter values does the model give a single coherent answer?
  - Example: Integrated Assessment Models (IAMs) for studying the economics of climate policy
  - Within the range of mainstream estimates of climate sensitivity and discount rates, IAMs tell us that the optimal policy regarding CO2 emissions is somewhere between subsidizing them and banning them.

Model Evaluation

- Model Testing is necessary to reduce the range of models by weeding out false components, if possible

- Tests do not seek to answer: *Is this model true?*

- Tests seek to answer:
  - Is this model significantly different than the world it is supposed to represent?
  - Is this model better than random guesses?
  - Do rival models based on alternative hypotheses do better than this one?

- Tests done so far have revealed many problems regarding how well GCMs compare against observations and whether they are sufficiently reliable for policy advice

Example 1: HAC-Robust Trend Modeling

- *Is this model significantly different than the world it is supposed to represent?*

- GCM application: comparison of trends projected by models vs trends in observations
- Statistical question: How to compare trends across different data sets that have unknown autocorrelation structures?
- Econometric approach: Heteroskedasticity and Autocorrelation-Robust (HAC) estimators applied to OLS trend coefficients

Example 1: HAC-Robust Trend Modeling

- Climate application: Tropical troposphere
  - GCMs predict very strong warming pattern in response to rising GHG levels
  - Observations from weather balloons and satellites apparently do not show this, but models and observations contain autocorrelated randomness
  - Confusion in climate literature over how to account for autocorrelation in conducting comparisons of trend coefficients
  - This topic was extensively studied in econometrics

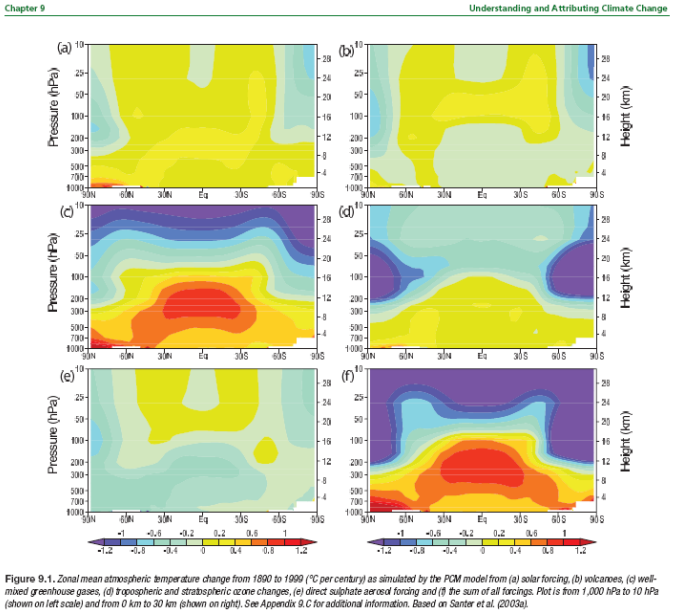
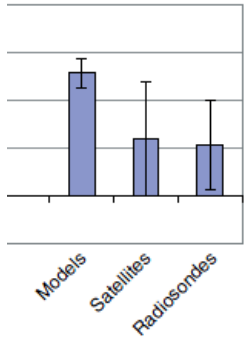


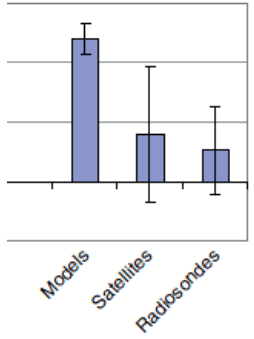
Figure 9.1. Zonal mean atmospheric temperature change from 1890 to 1999 (°C per century) as simulated by the PCM model from (a) solar forcing, (b) volcanoes, (c) well-mixed greenhouse gases, (d) tropospheric and stratospheric ozone changes, (e) direct sulphate aerosol forcing and (f) the sum of all forcings. Plot is from 1,000 hPa to 10 hPa (shown on left scale) and from 0 km to 30 km (shown on right). See Appendix 9.C for additional information. Based on Santer et al. (2003a).

Example 1: HAC-Robust Trend Modeling

- McKitrick, McIntyre & Herman (2010) applied HAC-robust method (Vogelsang-Franses 2005) and showed models significantly over-predict warming trends in satellite era



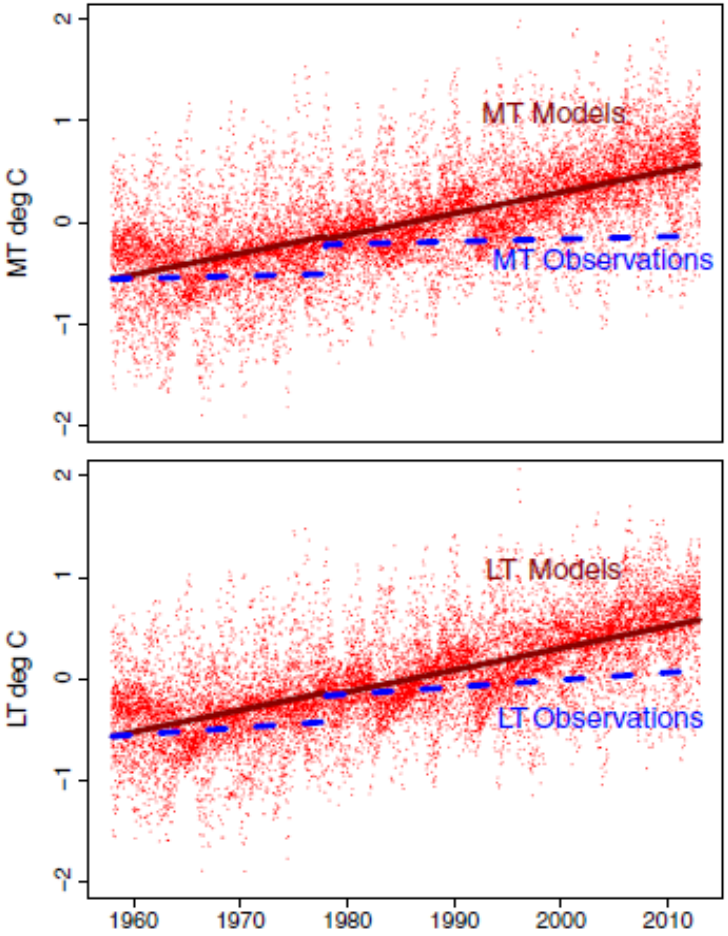
tropics, LT layer. 95% confidence interval



tropics, MT layer. 95% confidence interval

Example 1: HAC-Robust Trend Modeling

- McKitrick, McIntyre & Herman (2010) applied HAC-robust method (Vogelsang-Franses 2005) and showed models significantly over-predict warming trends in satellite era
- McKitrick & Vogelsang (2014) extended HAC methods to allow for step-changes at unknown date, showed models significantly over-predict since 1958



Example 2: Testing against randomness

- With 24 GCMs available, even if they are no better than Gaussian noise, at least 1 may appear to yield a “significant” fit
- The problem is that test must take into account the prior search over model specifications
- One approach: Bayesian Model Averaging (BMA)
  - Assesses all (2^{24}) combinations of models and assign a probability weight to each one based on its support in the data, then compute posterior distribution around coefficients
  - Robust to model selection effect
  - McKitrick and Tole (2012) used BMA to show that most GCMs had no ability to explain the spatial pattern of trends over land, and some models were worse than random numbers

Example 3: Rival, non-overlapping models

$$(1) \quad y = \mathbf{X}\beta + e$$
$$(2) \quad y = \mathbf{Z}\gamma + v$$

- Principle of "Encompassing": If (1) is the true model, it should not only explain *y* in terms of **X**, but any correlation between **X** and **Z** should account for the apparent explanatory power of **Z** in (2)
- Can be formulated in terms of an F test
- McKitrick (2013) used encompassing tests to show that standard testing methods to validate surface temperature data did not show what authors claimed they did
- Could be applied to any situation where *failure to observe an effect* is taken as proof of a result

Other examples

- Trend ratios $\frac{\beta_1}{\beta_2}$ e.g. troposphere versus surface
- Persistency and long memory in temperature data
- Statistical forecasting methods (ARIMA) versus structural models