Statistical Issues Involving (Climate) Computer Models

Jim Berger

Duke University

Statistical and Applied Mathematical Sciences Institute

A Statistical Consensus on Global Warming

October 26, 2007

Outline

- Views on (climate) computer modeling
- Some sources of uncertainty in computer modeling
- A pedagogical example
- The state of consideration of uncertainty in climate modeling
- How can statisticians help?
- How can statisticians become involved?
- What will statisticians be comfortable in concluding?
- Any other issues?

Three views of (climate) computer modeling:

- It is the only scientific way forward, so damn the torpedoes ...
- It holds great promise, but requires careful statistical validation.
- It is too difficult and has too much inherent uncertainty to be useful; one should utilize much simpler models with limited physics (ecology).

Four views on validation and verification of computer models:

- Hard Core Modeler: “Don’t bother me; I need every waking moment to work on the science/math/computation to improve the model.”
- Hard Core Statistician: “Practical use of the model is irresponsible unless all sources of uncertainty have been properly accounted for.”
- Soft Core Modelers:
  - “I’ll talk to statisticians if it help’s me to improve the model;”
  - “I’ll even consign some time and model runs to dealing with uncertainty.”
  - “In private and after a few drinks, I’ll admit that my model isn’t perfect.”
- Soft Core Statisticians: “I’ll do anything to improve the model and uncertainty assessment (with some guilt as to how little can be done).”

Some Sources of Uncertainty in Computer Modeling:

Reality: denote the real process being modeled by $y^R(\boldsymbol{x}, Z)$, where

- $\boldsymbol{x}$ are the model inputs (e.g., initial conditions, control variables, ...),
- $\boldsymbol{Z}$ are stochastic elements of the model.

Computer Model: $y^M(\boldsymbol{x}_1, \boldsymbol{x}_2^*, \boldsymbol{u}, \boldsymbol{Z}^*)$, where

- $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2^*)$, with * indicating those inputs that are uncertain,
- $\boldsymbol{u}$ are unknown parameters of the model (e.g., unknown coefficients of equations or rate constants, or unknown fudge factors),
- $\boldsymbol{Z}^*$ are the modeled stochastic elements (rarely identical to Z),
- $b(\boldsymbol{x}, Z) = y^R(\boldsymbol{x}, Z) - y^M(\boldsymbol{x}_1, \boldsymbol{x}_2^*, \boldsymbol{u}, \boldsymbol{Z}^*)$ is unknown bias (convergence, systematics, discrepancy) of the model.

Field data (if available): $y^F(\boldsymbol{x}, Z) = y^R(\boldsymbol{x}, Z) + \epsilon$, where ϵ could have a complicated error structure, depending on inputs or other features.

Pedagogic Example

- Dynamics of a chemical reaction process $y(t)$ are thought to be

$$dy(t)/dt = -uy(t)\,; \quad y(0) = 5\,,$$

  with the reaction rate u unknown, and known initial chemical concentration of 5.

- Solution $y^M(u,t) = 5\exp(-ut)$ is the 'computer model.'

- Field data from the process, at ten times t_i equally-spaced over the interval $(0.11, 3.01)$, is modeled as

$$y^F(t_i) = y^R(t_i) + \epsilon_i\,,$$

  where $y^R(t)$ is the real process and the ϵ_i are independently $N(0,\sigma^2)$, with σ^2 unknown. Three independent replicate observations $y^F_r(t_i)$ were obtained at each time point.

Standard 'Best Fit' Analysis: Estimate u from the data, say by least squares, i.e., choose u to minimize

$$\sum_{r=1}^{3}\sum_{i=1}^{10}\left[y_r^F(t_i) - 5\exp(-ut_i)\right]^2 .$$

Note that, in essence, this is assuming that the computer model is "truth plus random error."

Answer: $\hat{u} = 0.63$, with the resulting computer model being

$$y^M(\hat{u}, t) = 5\exp(-0.63t) .$$

This is graphed in Figure 1 along with the data.

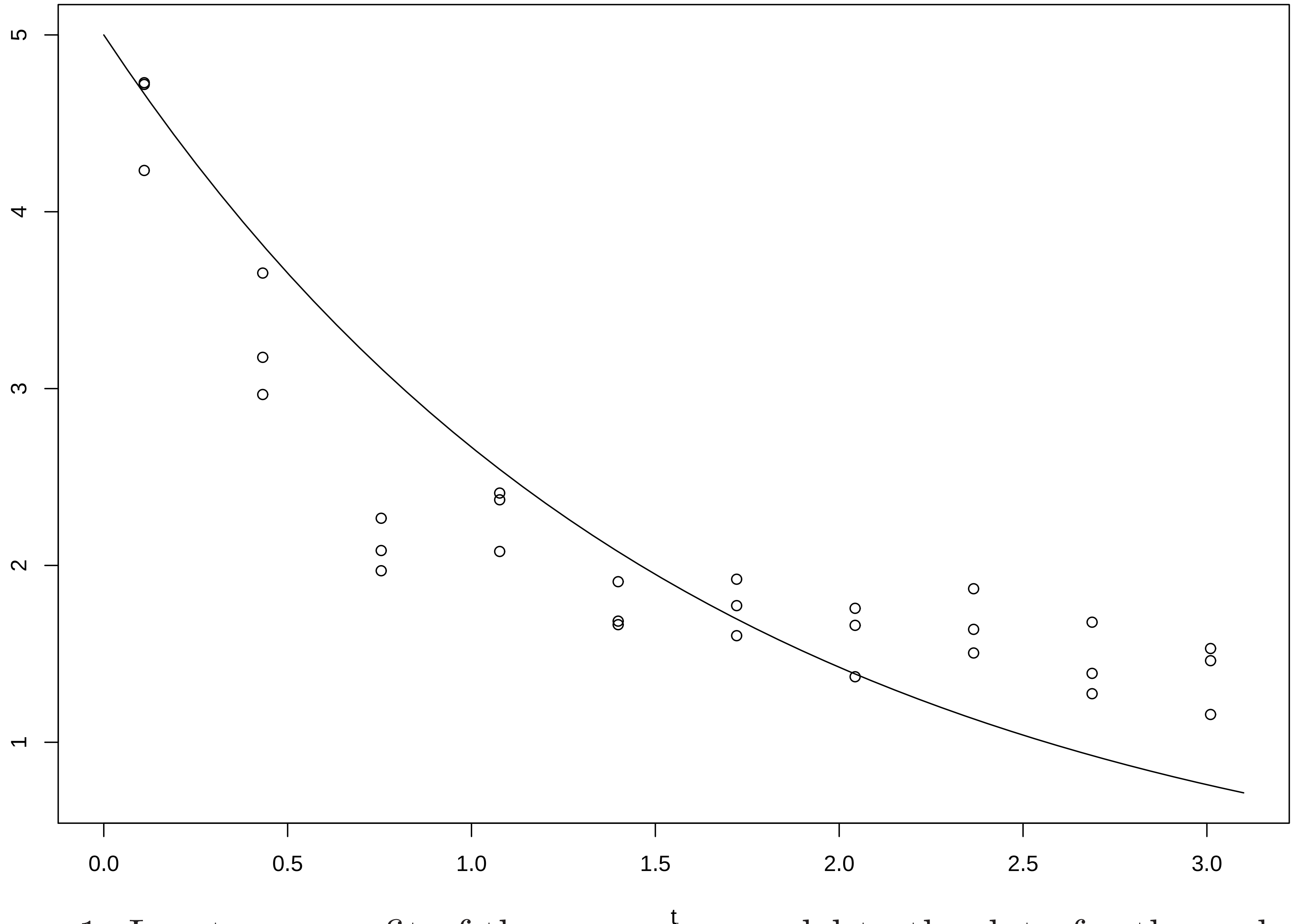


Figure 1: Least squares fit of the computer model to the data for the pedagogic example.

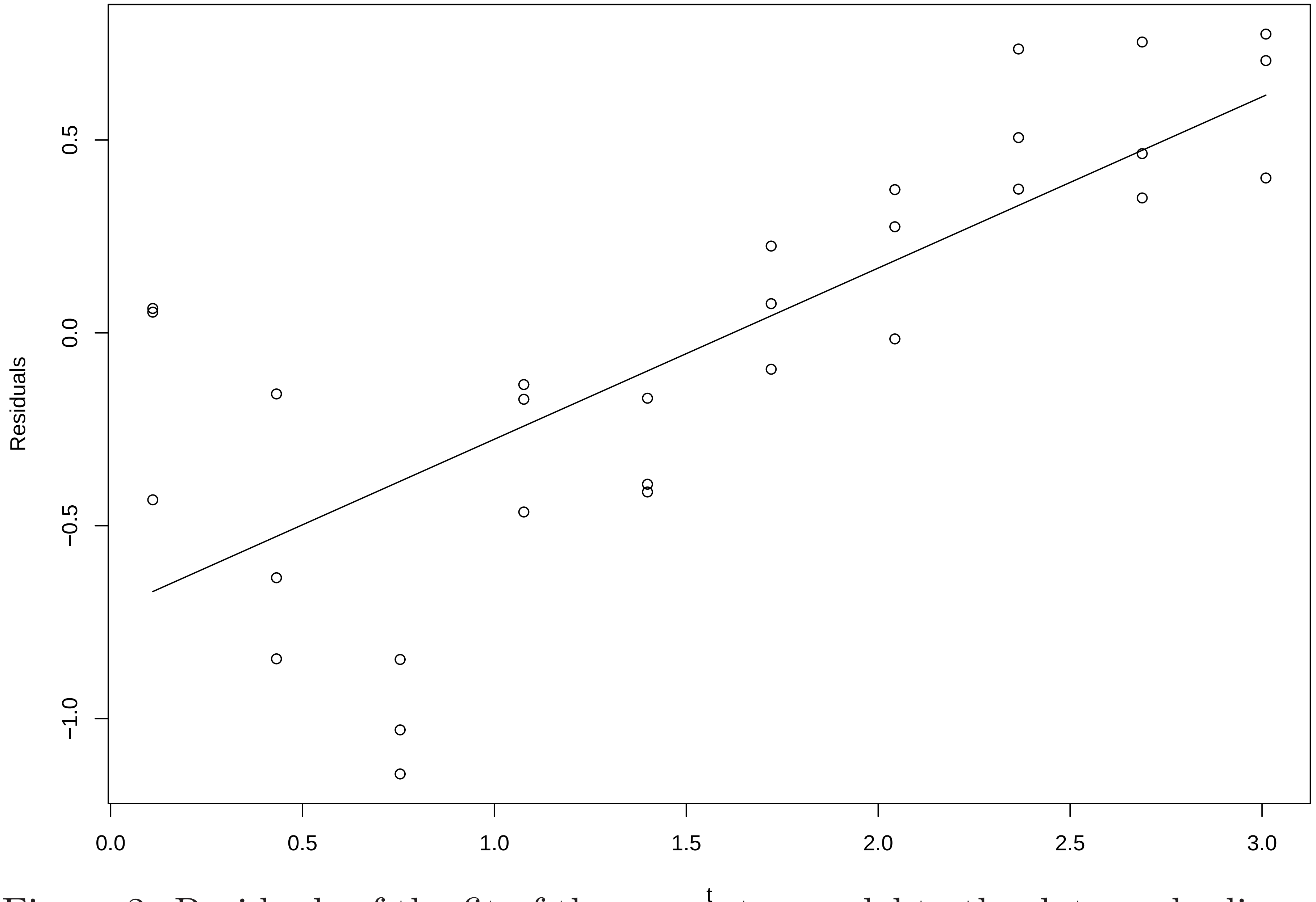


Figure 2: Residuals of the fit of the computer model to the data and a linear fit to the residuals.

Issues in proceeding:

1. We are pretending that $u = \hat{u}$. The uncertainty in this should be taken into account.

2. If the computer model is incorrect, i.e. has a systematic *bias* or *discrepancy*

   $$b(t) = y^R(t) - y^M(u, t)$$

   from the real process, 'over-fitting' will typically have occurred;
   the fit tries to 'make up' for the model inadequacy by over-shifting u to compensate (an especially serious problem if the model will be used to extrapolate)

3. This over-fitting makes it problematical to believe any structure found in the residuals, e.g. the linear structure in Figure 3.

Bayesian approach:

1. **Without bias:** Acknowledge the uncertainty in u (and σ^2), by viewing them as random with *prior distribution* $\pi(\mu, \sigma^2)$; Bayes theorem then gives the posterior distribution of (μ, σ^2), given the data $\boldsymbol{x}$, as

$$\pi(\mu, \sigma^2 \mid \boldsymbol{x}) \propto \; \pi(\mu, \sigma^2) \frac{1}{\sigma^{30}} \exp\left(-\frac{1}{2\sigma^2} \sum_{r=1}^{3} \sum_{i=1}^{10} \left[y_r^F(t_i) - 5\exp(-ut_i) \right]^2 \right) .$$

2. **With bias:** Recognize that the computer model may have a bias (Kennedy and O'Hagan, 2001 JRSSB), $b(t)$, which is then viewed as an unknown function and assigned a prior distribution $\pi(b)$, leading to the joint posterior distribution

$$\pi(\mu, \sigma^2, b \mid \boldsymbol{x}) \propto \; \pi(\mu, \sigma^2)\pi(b) \frac{1}{\sigma^{30}} \exp\left(-\frac{1}{2\sigma^2} \sum_{r=1}^{3} \sum_{i=1}^{10} \left[y_r^F(t_i) - 5\exp(-ut_i) - b(t_i) \right]^2 \right) .$$

Note: This is very different than the ubiquitous scientific modeling

$$\textit{field observation} = \textit{computer model} + \textit{random error}$$

Major difficulties:

- There is a severe lack of identifiability of u and $b(\cdot)$.
- There is severe practical confounding of σ^2 and $b(\cdot)$ if there are no replicate field observations.

Bayesian analysis (in principle) can accommodate this:

- In the computer model scenario, u may have physical meaning (e.g., the reaction rate in the example) or, at least, physical limits, so that experts may be able to construct a fairly tight prior distribution for u.
- The prior distribution for the bias 'encourages' $b(\cdot)$ to be zero, allowing a correct computer model to emerge with little bias if supported by the data.

But there are severe computational issues, due to the confounding of uncertain quantities, and special computational techniques are often required (MCMC with modularization, cutting feedback, ...)

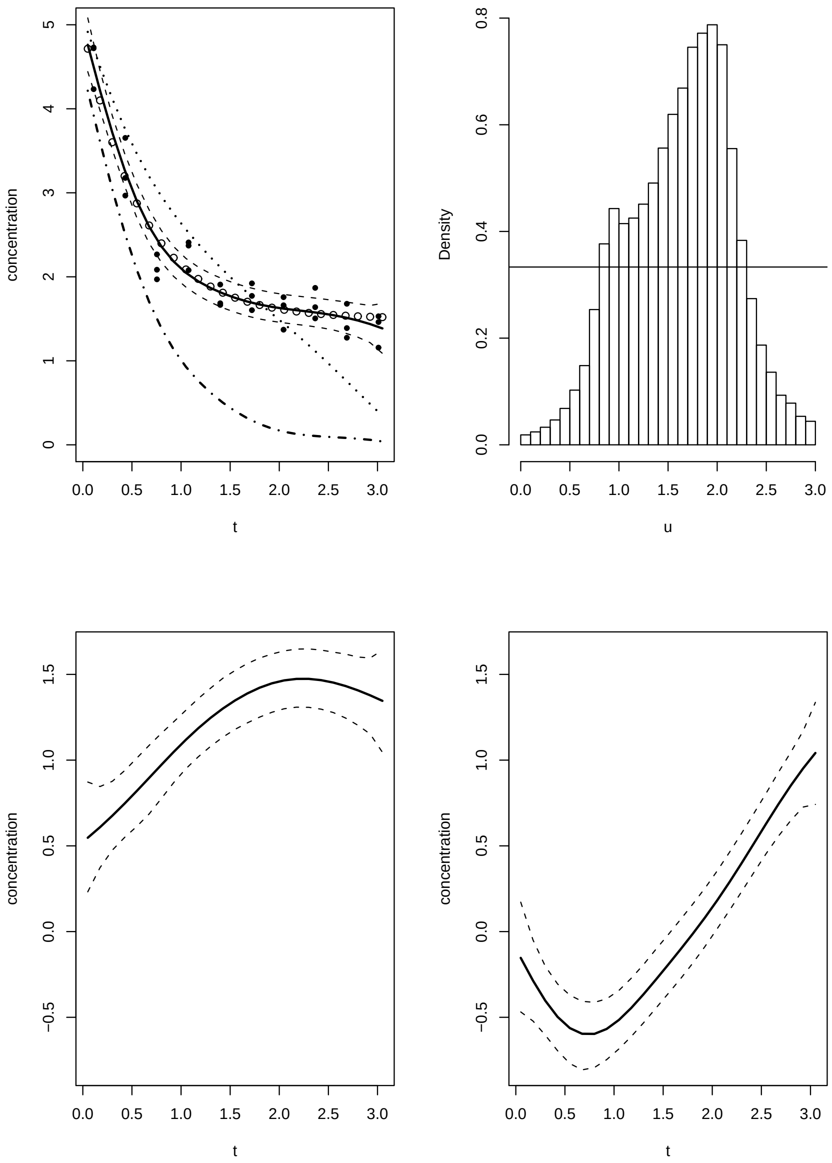


Figure 3: *Upper left:* bias-corrected predictions (solid) with 90% confidence bands (dashed). *Upper right:* marginal posterior distribution of u. *Bottom:* posterior distribution of bias, given the posterior mean *(left)*, given the least squares estimate *(right)*.

Truth: The true model used to generate the data was

$$y^F(t_i) = (3.5)\exp(-1.7t_i) + 1.5 + \epsilon_i \,,$$

with $\sigma^2 = (0.3)^2$. (The initial chemical concentration was indeed 5, but the reaction had a residual of 1.5 units unreacted.)

- So $\hat{u} = 0.63$ was indeed a severe over-fit.
- There is indeed a systematic bias,

$$b(t) = y^R(t) - y^M(1.7, t) = [(3.5)e^{-1.7t} + 1.5] - 5e^{-1.7t} = 1.5(1 - e^{-1.7t})\,.$$

The State of Consideration of Uncertainty in Climate Modeling:

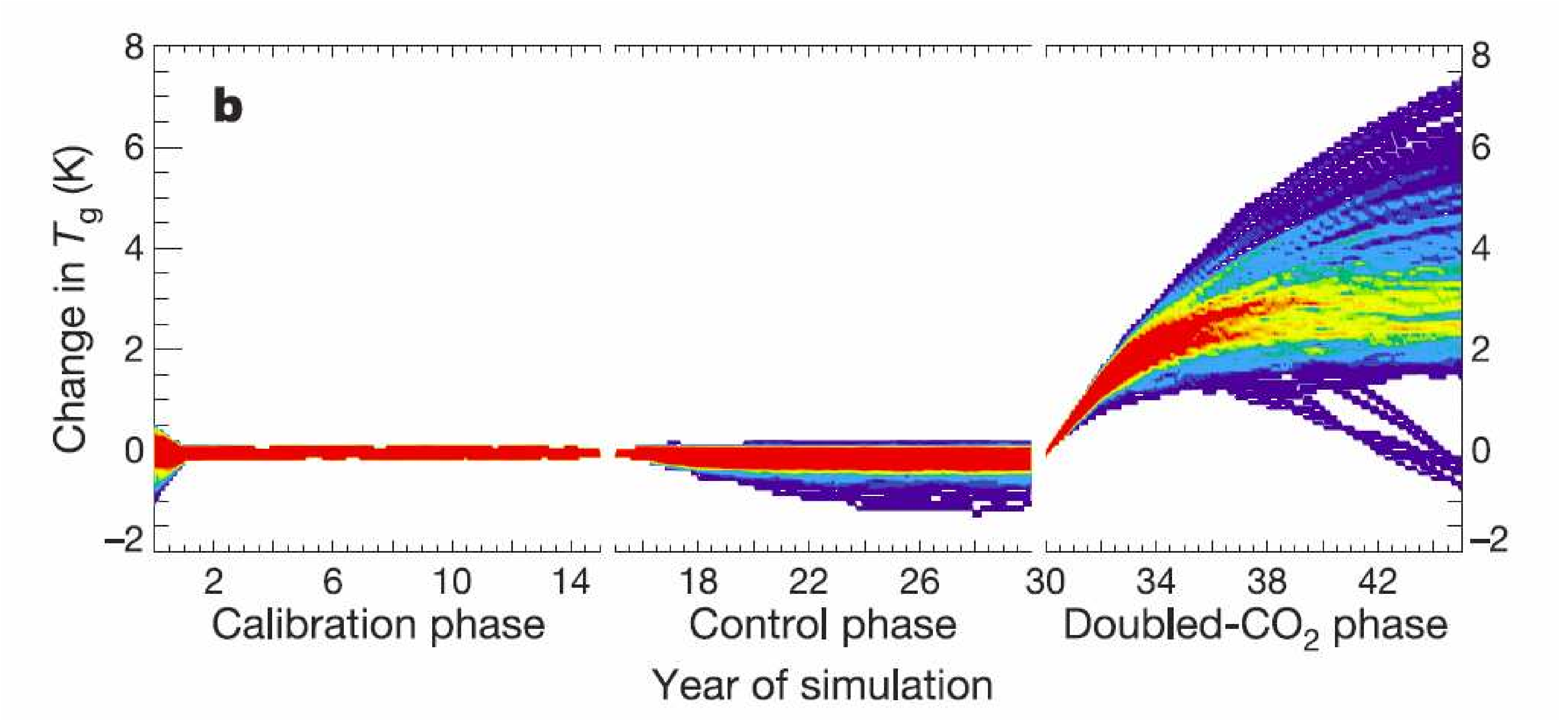
Stochastic parameterizations Z are increasingly being dealt with (e.g., stochastic parameterizations of clouds in WRF-1D).

Initial conditions $\boldsymbol{x}^*$: unknown or uncertain, are dealt with

- ideally by use of ensembles or particle filtering, but how realistic here?
- by running the computer model until 'steady state', and then changing the forcings (e.g., doubling the amount of CO_2) to determine the change in this steady state.
  - Is 'change from steady state' arguably the same as 'change from current conditions' globally (locally)?
  - Is 'natural climate variation' on a different time scale, and hence ignorable?

Model parameters $\boldsymbol{u}$: are increasingly being recognized as being uncertain, with simple 'best fit' analyses questioned.

- Formal Bayesian uncertainty analysis has the problem of typically either requiring linearization and data assimilation (e.g., for WFF-1D) or requiring a fast model emulator (e.g., calibration of the TIEGCM model)



Uncertainty in the predictions of the climate response to rising levels of greenhouse gases

(Lenny Smith and many others) Nature (2005)

Model bias (discrepancy) b:

- *Recognition that bias exists* is rather rare (outside of NCAR)
  - partly for sociological reasons ("my model is as good as I could make it, i.e. is current *best science*, so don't call it biased!");
  - partly because the field data about reality is often too limited for sensible estimation of bias (or the field error may contain bias);
  - partly because extrapolation of bias away from the input region where there were field measurements is fraught with peril;
  - partly because bias is severely confounded with other uncertainties.
- *Determining bias* is
  - best done with field data (how frequently is it available here?);
  - might be addressed through surrogates as *fingerprinting* (is that sufficient to eliminate worries about bias?)
  - might be addressed by looking at competing climate models, versions of the models run at different scales, etc.

Other uncertainties in climate modeling?

How can statisticians help?

- Design of computer model experiments
- Modeling parameterizations or other ill-specified parts of the model.
- Creating emulators
- Dealing with calibration and associated uncertainty (even in the face of severe confounding)
- Investigating bias and detecting model inadequacies
- Analyzing importance of inputs
- Modeling input (or other distributions)
- Handling high dimensional outputs or inputs through PCA, POD, EOF
- Utilization of the computer model in prediction, risk analysis, or in combination with other components/systems (e.g., climate models with forests).
- Other ways?

Example: Risk Maps for Pyroclastic Flow

Key elements:

- *Computer model* (unvalidated) of pyroclastic flow over a specified terrain.
- *Data* about past flow volumes and directions

Goal: Preparation of a risk map for a catastrophic event happening at each location over a specified time.

Statistics used to

- Model the data to create the needed input distribution of flows, recognizing the key is good modeling of the extreme events.
- Do the extreme event risk analysis, requiring a combination of importance sampling and adaptive emulator development together with judicious computer model runs.

How can statisticians become involved?

The Key: Becoming involved in a ‘team environment’ with scientists.

Facilitating infrastructure:

- NCAR, where teams operate
- SAMSI (and NPCDS), where teams can be formed
- National labs (both LANL and LLNL have climate/stat teams)
- Large interdisciplinary grants available today

Barriers:

- Statistics cannot generally fund involvement of statisticians in other disciplines which, in turn, rarely have much money for statistics.
- Shortage of statisticians
- The time needed for a statistician to get deeply involved with another science and to also learn the *statistics* needed for it.
- Scientists often have a hard time judging what they can do themselves and when they should seek statistical help.

What will statisticians be comfortable in concluding?

- Existence of humanity-forced global warming?
  - Could we “approve” the methodology used in establishing this, as being ‘best available.’
- Predictions from computer models?
  - We would certainly want these to come with at least some uncertainties stated and handled correctly.
  - If we are utilizing a single model, stated predictions will almost certainly be biased and the stated uncertainties too small; do we always add that caveat to our conclusions? (Such caveats are always ignored or used for the wrong purposes.)
  - Should we do multiple climate model or other adjustments so that we are at least not sure if stated uncertainties are too small?
  - What do we publicly say about all the impossibly certain climate forecasting advice that is about to start appearing?

Other issues?