# Upper Ocean Heat Content Uncertainty in the CCSM3's Ocean using a Large Ensemble

**Robin Tokmakian**[1]: rtt@nps.edu and **Peter G. Challenor**[2]: P.Challenor@noc.ac.uk

1. Dept. of Oceanography, Naval Postgraduate School, Monterey, CA, USA. 2. National Oceanography Centre, Southampton, United Kingdom.

National Oceanography Centre
NATURAL ENVIRONMENT RESEARCH COUNCIL

## Abstract:

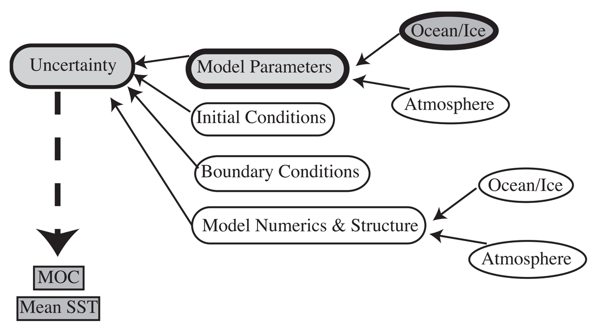
We examine the heat content and uptake in the ocean component (POP2) of the CCSM3.0 model using an emulator created from a large member ensemble. The uncertainty in a climate model is a combination of the uncertainty in the initial conditions, the parameter space and the model structure. "Statistical Analysis of Computer Code Output" (SACCO) methods, based on Bayesian statistics, can be used to explore uncertainties associated with such complex models. This paper discusses the results of a designed experiment to explicitly determine the formal uncertainty in the ocean and ice components of the CCSM3.0 climate model. After an initial spin-up, the ocean/ice system is forced with the NCEP reanalyses (COREv2) with the last 40 years of a 100-year simulation.

A 100 member ensemble of the CCSM3.0 ocean/ice system at a 3 degree resolution was created, of which 89 are used in the final analyses. Each ensemble member used a different set of parameter settings. The values for the parameters were determined using a Sobol sequence. Sobol sequences produce a sampling of a multiple parameter space, such that the sampling is uniform, but sparse, across input space. The parameters we varied relate to such things as ocean mixing and advection.

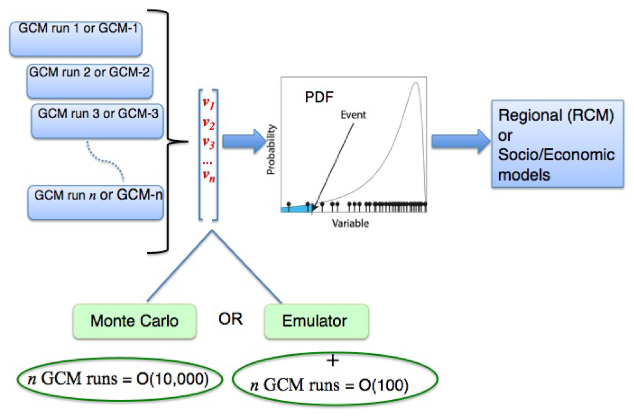
A relatively small member ensemble (100 member, as compared to the 10000 required for determining an accurate uncertainty distribution using Monte Carlo methods) of the complex model is used to create an emulator. The emulator is then used to produce a PDF of the that is equivalent to a 10000 member ensemble based on only 100 model runs. An emulator is a tool to investigate the uncertainty characteristics of the model and its outcomes and is not a replacement for the model itself. The full PDF can then be used to determine uncertainty values associated with a metric determined by or computed from the outputs of the physical model.

We compare the model output to observational data, initially, to show the realism of the model. We discuss how the model output and its emulator can be used to understand intrinsic uncertainties, related to parameter settings, in the uptake of heat in the ocean component. The analysis is shown for the global heat content, as well as for regional areas, such as the North Atlantic, the North Pacific, and the Southern Ocean.

## Background




- Uncertainty in a GCM contains contributions from the parameters, initial conditions, boundary conditions, and underlying model structures.
- We look at the uncertainty in only the parameter space of the ocean and ice system.
- For parameters, the contribution to total uncertainty is about around 20% for all periods.



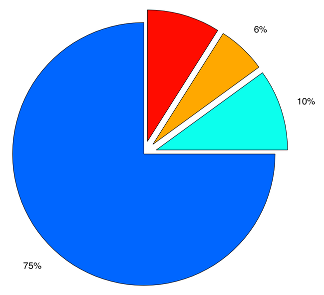

- We know there is some level of uncertainty in our imperfect models.
- To determine the uncertainty in ***v***, we need to examine the outcomes from all possible inputs so that we can produce a complete PDF for ***v***.
- We have a choice to create a very large Monte Carlo GCM ensemble or to create a much smaller GCM ensemble and combine the outcomes with an emulator.
- Once we know the uncertainty, we can use the information to inform subsequent analyses such as informing regional models or risk in socio/economic models.

## GCM Model Details

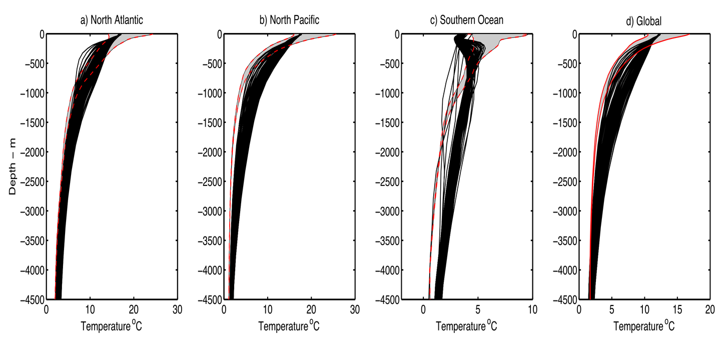
- CCSM3 x3 (~ 3° resolution) (Community Climate System Model)
- Active ocean/ice components (POP2 & CICE)
- NCEP inter-annual reanalyzes forcing (COREv2 )
- 9 parameters (x or inputs)

## Experiment Setup

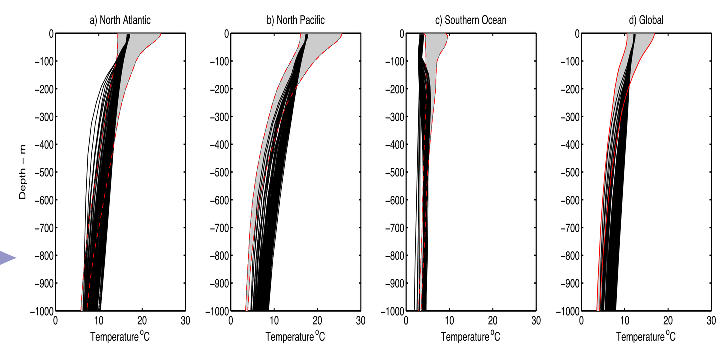
A Sobol sequence was used to set parameter values to use for each simulation.



100 runs
- 85 for building emulator (training or design locations) $Y_{emul}$
- 47 as independent validation points (multiple tests) $Y_{val}$
- 9 runs failed to run to competion
- 6 not used (outcomes invalid)



- Compared average temperature profiles (black) of each simulation to the Levitus data set (Locarnini et al. 2006) (gray/red)
- 6 Southern Ocean profiles indicate invalid outcomes - eliminated



## Emulator Details

$$F(x) \sim f(x) = GP(m_0, k(x,x')) = m_0(x) + GP(k(x,x'))$$

Truth (metric) — $F(x)$; estimate — $f(x)$

$$m_0(x) = h(x)^T b$$

Mean process — $m_0(x)$; vector of q regression functions — $h(x)$; q coefficient vector — $b$

$$GP(k(x,x')) = \sigma^2 e^{(-(x_1-x_2)^T B (x_1-x_2))}$$

Gaussian process — $GP(k(x,x'))$; Variance — $\sigma^2$; A covariance function, in this case, a Gaussian covariance which assumes stationarity

- The emulator **f(x)**
  - approximates the outcomes from the GCM **F(x),**
  - is a **Gaussian Process** (GP): a function with a
    - **mean process**
    - **covariance process, k(x,x')**, where **x** is a vector of inputs.
- The **mean process**:
  - composed of a set of regression functions, contain some prior information about how the outcomes are related to the inputs.
  - $\mathbf{h(x)^T}$ is a prior defined here as a linear function.
- The **Gaussian process**,
  - a quantity, **σ**, related to the outcome variance
  - a quantity defined as a covariance function may have different forms (e.g. a Mate´rn function)
  - the emulator is fairly smooth defined by **B**, a scaling factor.
- An emulator, **f(x),** should have the following attributes given a true model **F(x) = Y:**
  - reflect the true value of $Y$ *(the outcomes for a metric)* at input points $\boldsymbol{x}$ *(Design points)*
  - at other points, distribution of f(x) should give a mean value for F(x) that is plausible
  - the probability distribution should be a realistic view of the uncertainty

## Emulator Validation

- After the emulator has been built, it is necessary to evaluate its quality.
- We use the Mahalanobis distance diagnostic (Bastos & O'Hagan, 2009): $D^{MD}(Y) = [Y - f(X)]^T V(f(X))^{-1}[Y - f(X)]$
- A subset of outcomes are used for validation (10 in this case) $Y_{val}$, with 75 outcomes for the emulator design points: $Y_{emul}$.
- A number of different emulators are created (20 for the global case) ***f(x)*** - with an associated variance: ***V(x).***
- The quantity, $D^{md}$ has a $\chi^2$ distribution. with 10 degrees of freedom; values between 8.29 and 10.47 indicate a reasonable emulator.

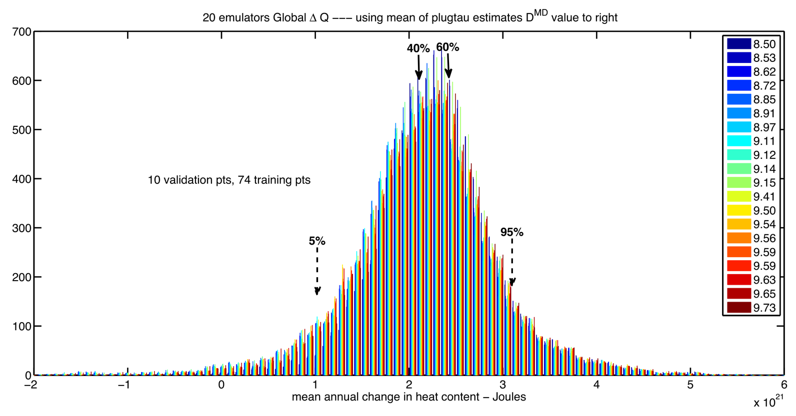
## Climate Metric - Upper Ocean Heat Content Change

$$\Delta Q(t) = c_p \rho_0 \left(\int_i \int_j \left[\int_{700}^{0} T(t)\,dz - \int_{700}^{0} T(t-1)\,dz\right] dx\,dy\right)$$

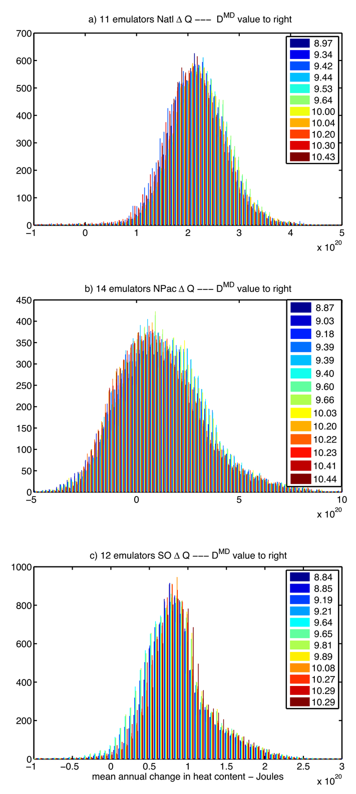
$$\overline{\Delta Q} = \sum_N \Delta Q(t) / N$$

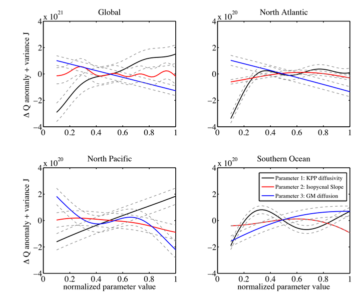
- The metric or outcome is the average annual change in heat content, ΔQ, the upper ocean (0 - 700 m).
- Four areas: North Atlantic, the North Pacific, the Southern Ocean, and the total global heat content.
  - $c_p$ is the specific heat
  - $\rho_0$ is an average density
  - $T$ is the potential temperature at time $t$
  - grid cell (i; j)
  - $t$ annual values. A 36-year time series is used
  - $N$ is the number of years
- $\overline{\Delta Q}$ is the average annual change - a measure of heat uptake in the model over this period.

## Emulator outcomes




- The $\overline{\Delta Q}$s for our simulator ensemble (n =85) were determined: $Y_{emul}$
- We create a set of emulators: $f(x)_i$ for ΔQ,
  - a Mate´rn correlation function is used for $\chi(x_1; x_2)$
  - $i$ = *1:M, M* = number of validation tests
  - a different 10 $Y_{val}$ & 75 $Y_{emul}$ for each test
- 20 emulators satisfy our "success"criteria for $D^{MD}$ (global)
- The most-likely value of ΔQ is 2.32e21 +/- 0.12e21 Joules (J).
- 90% of the time, $\overline{\Delta Q}$ is between 0.99e21 J and 3.38e21 J.
- 80% of the time, the model's $\overline{\Delta Q}$ is between 2.04e21 J and 2.34e21 J. (solid arrows)






- Using the same process as above, resulting histograms for the 3 regions are given
- The normalized response curves for 3 input parameters are also shown.
- The North Pacific and North Atlantic are similar to the global histogram; Southern Ocean response distinctly different.

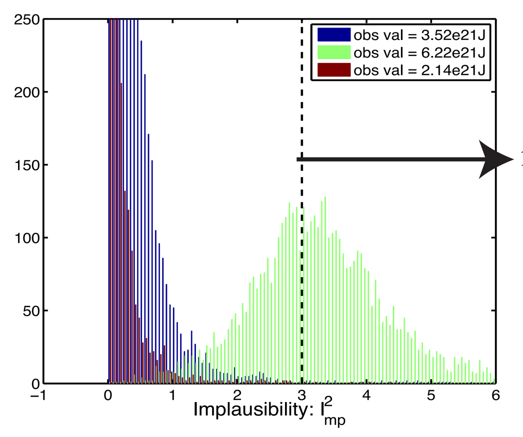
## Global emulator comparison to observations

- An implausibility score ($I^2_{mp}$; Vernon et al. 2010) determines if our emulator outcomes are implausible or not.
  - $Y_{obs}$ is some metric calculated from observations
  - ***f(x)*** is the emulator outcome at input location $\boldsymbol{x}$
  - ***V(x)*** is the emulator variance associated with ***f(x)***
  - $\sigma^2_{obs}$ is the variance of the observation
  - $\sigma^2_{disc}$ is a discrepancy term based on expert knowledge (=0 for these examples)
  - A score > 3 implies an implausible outcome

$$I^2_{mp}(x) = \frac{(Y_{obs} - f(x))^2}{(V(x) + \sigma^2_{obs} + \sigma^2_{disc})}$$

- 4 observational estimates of $\overline{\Delta Q}$ derived from these data sets
  - Levitus et al. (2009) 2.24e21 J
  - Palmer et al. (2010) 6.22e21 J
  - Ishii and Kimoto (2009) 2.14e21 J
  - Domingues et al. (2008) 3.49e21 J
  - Mean 3.52e21 J

- Three example histograms of implausibility scores are given comparing emulator $\overline{\Delta Q}$ to observed $\overline{\Delta Q}$.
  - The blue histogram sets $Y_{obs}$ = 3.52e21 J (the observational mean) as the observed estimate of $\overline{\Delta Q}$. < 1% of the outcomes are implausible, i.e. all ***f(x)*** are consistent with the observation
  - $Y_{obs}$ = to high observational estimate of 6.22e21 J ~ 60% of the outcomes are implausible (green histogram).
  - $Y_{obs}$ = the low end of our observational estimates: 2.14e21 J almost no implausible outcomes (red histogram).




## Using emulator outcomes with multi-model outcomes

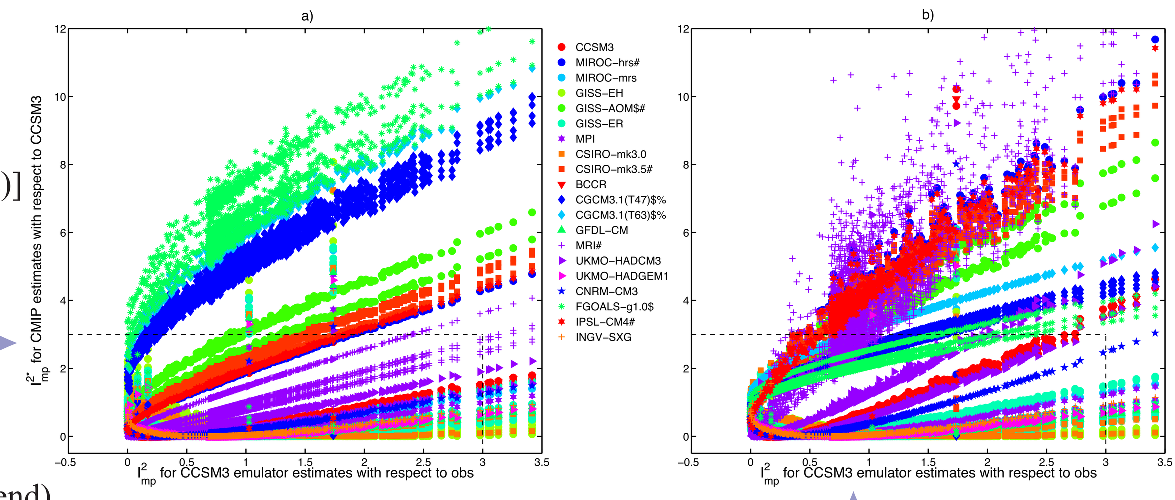
TABLE 2. CMIP3 20c3m Models

| Model | No. | resolution; vertical levels | $\overline{\Delta Q}$ *1e21 J | $\sigma$ * 1e21J |
|---|---|---|---|---|
| CCSM3 | 2 | 320x384; 40 | 2.55 | 0.02 |
| MIROC3.2hrs | 1 | 320x320; 47 | 4.54 | 0.12+ |
| MIROC3.2mr | 3 | 256x192; 43 | 1.60 | 0.21 |
| GISS-EH | 5 | 360x180; hybrid level/isopycnal | 0.59 | 0.14 |
| GISS-AOM | 2 | 90x60; 16 | 5.23 | 0.10 |
| GISS-ER | 9 | 72x46; hybrid level/isopycnal | 1.58 | 0.20 |
| MPI-OM | 3 | 360x180; 40 | 1.73 | 0.10 |
| CSIRO-Mk3.0 | 3 | 192x189; 31 | 1.07 | 0.22 |
| CSIRO-Mk3.5 | 3 | 192x189; 31 | 4.79 | 0.05 |
| BCCR | 1 | 360x180; 35 | 2.56 | 0.12+ |
| CGCM3.1(T47) | 5 | 192x96; 29 | 6.67 | 0.05 |
| CGCM3.1(T63) | 1 | 256x192; 29 | 6.56 | 0.12+ |
| GFDL-CM | 2 | 360x200; 29 | 1.56 | 0.05 |
| MRI | 5 | 144x111; 23 | 3.63 | 0.12 |
| UKMO-HADCM3 + | 2 | 288x144; 20 | 2.66 | 0.15 |
| UKMO-HADGEM1 | 1 | 360x216; 40 | 1.57 | 0.12+ |
| CNRM-CM3 | 1 | 180x170; 31 | 2.31 | 0.12+ |
| FGOALS-g1.0 | 3 | 360x170; 33 | 7.42 | 0.13 |
| IPSL-CM4 | 1 | 180x170; 31 | 4.58 | 0.12+ |
| INGV-SXG | 1 | 360x180; 33 | 1.73 | 0.12+ |

$+\sigma^2$ = mean of all models with more than one simulation

- Determine how a set of multi-model simulations relate to this ensemble (CCSM3-L); i.e. which of the multi-model simulations would be implausible.
  - Using 20th century members of the World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model dataset (Meehl et al. 2007).
  - Calculated a $\overline{\Delta Q}$ for each model; summarized in table.
  - Column 2 in table is the number of simulation runs for a particular model (n)
  - Column 5 is the standard deviation for a model set, when n = 1, σ = average σ across models.
  - The mean ΔQ across the full set of models is 3.09e21 J, with σ = 2.18e21 J.
- General assumptions
  - Feedback of heat between ocean and atmosphere is relatively small
  - All the runs are 20th century runs; thus similar forcing
  - 2 CMIP3/CCSM3 have $\overline{\Delta Q}$ similar to the CCSM3-L ensembple outcomes
  - As in many analyses; CMIP3 model structure is not considered
  - The parameter uncertainty is represented by the CCSM3-L ensemble

- 3 examples of determining multi-model implausibility
- Compare $I^2_{mp}$ (see above) verse $I^{*2}_{mp}(x) = \dfrac{(Y_{cmip} - f(x))^2}{(V(x) + \sigma^2_{cmip} + \sigma^2_{disc})}$
- 1) Truth centered (no plot)
  - $Y_{obs}$ = mean of [observational estimates + (CMIP3 +CCSM3-L outcomes)]
  - $Y_{cmip}$ = CMIP3 model $\overline{\Delta Q}$
  - $\sigma^2_{disc}$ = 0; $\sigma^2_{cmip} = \sigma^2$ over all models;
  - Most scores fall within 'not implausible' space; assumptions too broad
- 2) Exchangeable
  - same as above, except
  - $Y_{obs}$ = 3.52e21 J (the observational mean)
  - $\sigma^2_{cmip}$ = average variance from all the CMIP models.
  - 4 of the CMIP3 models have 5% $I^{*2}_{mp}$ scores greater than 3 ($ in the legend).
  - The CMIP3 models that are "implausible" are those with the highest ΔQ values
- 3) No expected value assumptions
  - same as above, except
  - $\sigma^2_{cmip}$ = from individual CMIP model; $\sigma^2_{disc}$ = bias² between obs and CMIP
  - CMIP3 models with $I^{*2}_{mp}$ greater than 3 ( # and % in the legend).
  - The size of the bias and a model's own variance that determines $I^{*2}_{mp}$ score




- Multi-model comparisons: definitions (see Knutti et al. 2010)
  - The expected value over all the models is the truth: 'truth centered'
  - An ensemble member can be exchanged with another member or the observation; 'exchangeable'

- Using the third example
  - The mean $\overline{\Delta Q}$ for the "not implausible' CMIP3 models = 2.41e21 J.
  - A 63% probability that the "true" ensemble outcome < CMIP3 mean
  - A 37% probability that it will be greater.
  - The $\overline{\Delta Q}$ uncertainty due to parameter and structural uncertainty: between -.95e21 J and 5.78e21 J (i.e. 2.41e21 +/- (1.27e21 + 2.09e21) ).

**References:** Bastos, L., & O'Hagan, A., 2009, Diagnostics for gaussian process emulators. Technometrics, 51 , 425-438.
Challenor, P.G, McNeall, D, Gattiker, J., 2009 Estimating the Probability of Rare Climate Events in The Handbook of Applied Bayesian Analysis, Editor A O.Hagan and M. West , Oxford University Press
Higdon, D., J. Gattiker, B. Williams, and M. Rightley, 2008, Computer Model Calibration Using High-Dimensional Output, J. of the American Statistical Association, 103,482, Applications and Case Studies
O'Hagan, A., 2006, Bayesian analysis of computer code outputs: a tutorial. Reliability Engineering and System Safety, 91, 1290–1300.
Knutti, R., G. Abramowitz, M. Collins, V. Eyring, P. Gleckler, B. Hewitson, and L. Mearns, 2010: Good practice guidance paper on assessing and combining multi model climate projections. Rpt. of the IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections; [Stocker, T.F.,D. Qin, G.-K. Plattner, M. Tignor, and P.M. Midgley (eds.)].
Tokmakian R., P. Challenor, & Andrianakis, 2010, An extreme non-linear example of the use of emulators with simulators using the Stommel Model, submitted
Vernon, I., M. Goldstein, and R. G. Bower, 2010: Galaxy formation: a bayesian uncertainty analysis. Bayesian Analysis, 05 (04), 619–670.

**Acknowledgements:** NCAR High Performance Computing Resources (Bluefire); NPS Hamming High Performance Computer; CCSM community; COREv2 (Large and Yeager, 2008); NSF Grant #0851065; I. Andrianakis: National Oceanography Centre, UK for emulator software (MUCM toolkit: http://bit.ly/cAfLMY), Isaac Newton Institute for Mathematical Sciences, U. Cambridge, UK; We acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multi-model dataset. Support of this dataset is provided by the Office of Science, U.S. Department of Energy.