Probabilistic Climate Model Evaluation

Amy Braverman[1]

Joint work with Snigdhansu Chatterjee[2], Megan Heyman[2], and Noel Cressie[3]

[1]Jet Propulsion Laboratory, California Institute of Technology
[2]School of Statistics, University of Minnesota
[3]NIASRA, University of Wollongong

June 15, 2016

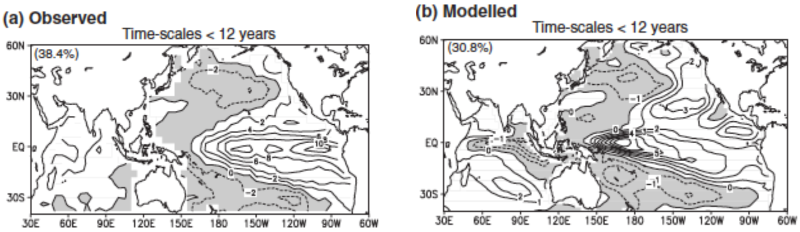
Outline

- Motivation
- Hypothesis testing framework
- Test statistics and hypothesis testing
- Application to CCSM4 vs HadCRUT4
- The null distribution of the test statistic
- Hypothesis test result
- Conclusion
- Acknowledgements
- References

Motivation

- Climate models are deterministic, mathematical descriptions of the physics of climate.
- Confidence in predictions of future climate is increased if the physics are verifiably correct.
- A necessary (but not sufficient) condition is that past and present climate be simulated well.
- How do we judge this?

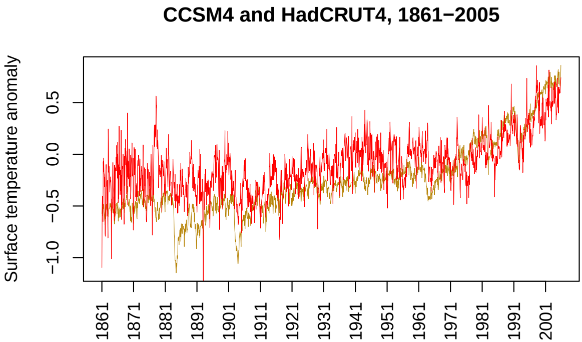
Motivation



Two panels of Figure 8.21 from Chapter 8, Third Assessment Report of IPCC Working Group 1 (2007). Comparison of eigenvectors for the leading EOFs of the SSTs between the ENSO time-scale (<12 years) (a) observation, and (b) the MRI coupled climate model, respectively (Yukimoto, 1999). Numbers in bracket at the upper left show explained variance in each mode.

Are these “the same"?

Motivation



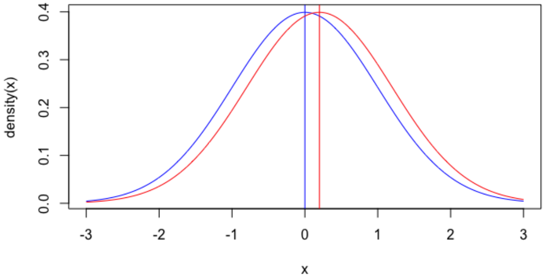
Monthly global average surface temperature anomaly (vs 1961-1990 mean) for CCSM4 (Gent et al., 2011) and HadCRUT4 (Monice et al., 2012 in red).

Motivation

- How similar do the fields or time series have to be to call them “the same"?
- Depends on the inherent variability of the statistic used to measure similarity.
- Hypothesis testing framework:
  - H_0: modelled and observed come from the same population.
  - Test H_0 using the modeled and observed fields or time sequences.
  - Reject $H_0 \longrightarrow$ not the same.
  - Do not reject $H_0 \longrightarrow$ the same?

Hypothesis testing framework

Example: Is the temperature at 12:00 noon in the month of July the same at JPL as it is in Pasadena?



$$X \sim N(\mu_1, \sigma^2), \quad Y \sim N(\mu_2, \sigma^2), \qquad H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_A : \mu_1 \neq \mu_2.$$

Hypothesis testing framework

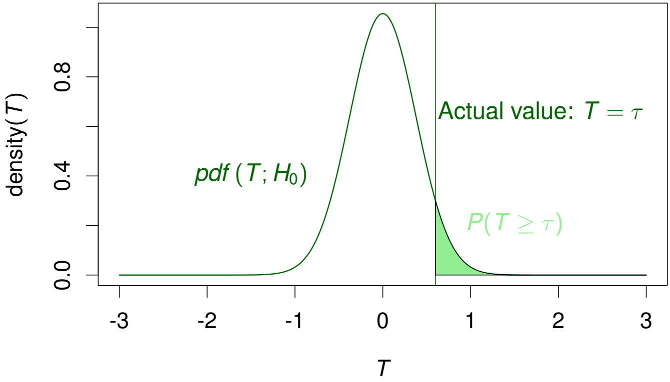
$$X \sim N(\mu_1, \sigma^2), \quad Y \sim N(\mu_2, \sigma^2), \qquad H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_A : \mu_1 \neq \mu_2.$$

1. Collect data: $X_1, X_2, \ldots, X_N$ from population 1, $Y_1, Y_2, \ldots, Y_M$ from population 2.
2. Choose test statistic: $T = (\bar{X} - \bar{Y})$.
3. Obtain distribution of test statistic under assumption of the H_0, $pdf(T; H_0)$.
4. Locate T in the distribution $pdf(T; H_0)$ and determine how extreme T is.

If T is “extreme" then we reject H_0 because T is “inconsistent" with it.

Hypothesis testing framework

Traditionally, reject H_0 if $P(T \geq \tau) < \alpha$, $\alpha = .05$ (say).



Hypothesis testing framework

Remarks:

- To respect uncertainty, it is useful to model data with probability distributions even if they are produced by deterministic mechanisms.
- Choice of the test statistic is up to us.
- Choice of how we obtain $pdf(T; H_0)$ is up to us (analytically, via simulation, etc.)
- $P(T \geq \tau)$ is called the *p*-value of the test. It is a scaled “distance" between τ and the expected value of T under the assumption that the null hypothesis is true.
- Threshold α is called the significance level of the test, and it is our choice.

Application to CCSM4 vs HadCRUT4

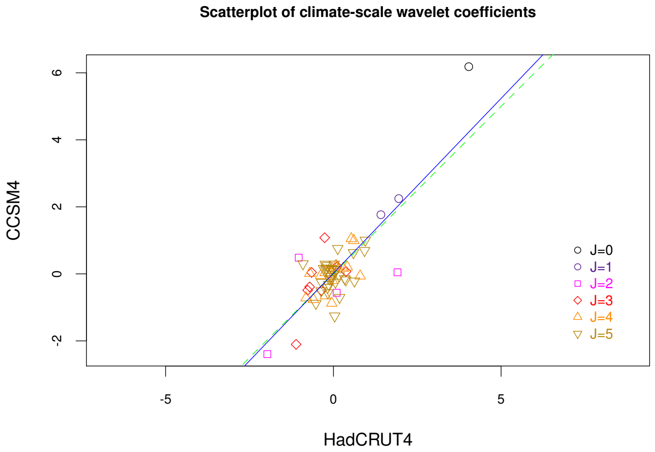
For CCSM4 vs. HadCRUT4 monthly surface temperature anomalies (relative to the mean 1960–1991)

1. Collect data: 1739 monthly values (1861–2005) for CCSM4 (**X**) and HadCRUT4 (**Y**).

2. Choose test statistic:
   - regress “climate-scale" wavelet coefficients* of **X** on those of **Y**,
   - obtain slope, β_1, and intercept, β_0,
   - test statistic is $T = [(\beta_1, \beta_0) - (1, 0)]\,\mathbf{K}^{-1}\,[(\beta_1, \beta_0) - (1, 0)]'$.
   - **K** is an estimate of the covariance matrix of (β_1, β_0).

* Climate-scale defined as coarsest six (of 11 total) wavelet coefficient levels.

Test statistic



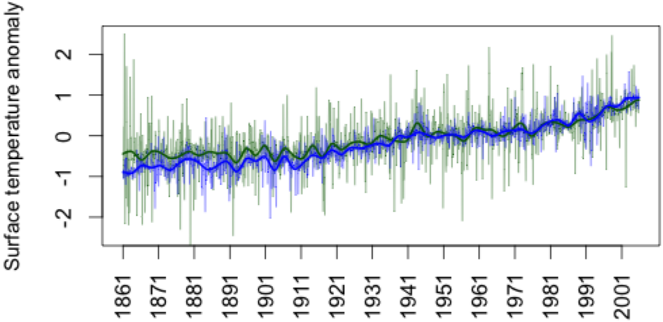
The null distribution of the test statistic

3. Obtain distribution of T under assumption H_0 is true ($pdf(T; H_0)$):

- we use a resampling method called the Wild Scale-Enhanced Bootstrap (WiSEBoot)
- create 1000 resampled time series pairs $(\mathbf{X}^*, \mathbf{Y}^*)$ with common (HadCRUT4) climate-scale coefficients and perturbed “noise".
- $T_i^* = [(\beta_{i1}^*, \beta_{i0}^*) - (1, 0)]\,\mathbf{K}^{-1}\,[(\beta_{i1}^*, \beta_{i0}^*) - (1, 0)]', \quad i = 1, 2, \ldots, 1000.$
- $\mathbf{K}$ is the empirical covariance matrix of $(\beta_{i1}^*, \beta_{i0}^*), \quad i = 1, 2, \ldots, 1000.$
- Histogram of $\{T_i^*\}, \quad i = 1, 2, \ldots, 1000$ is an approximation of $pdf(T; H_0)$.

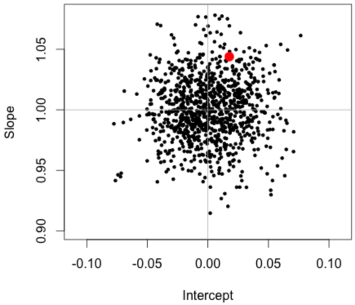
The null distribution of the test statistic

One resampled pair of HadCRUT4 (green) and CCSM4 (blue) time series.

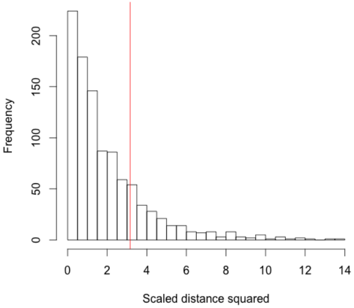


Thick lines are climate-scale reconstructions, and thin lines are full reconstructions.

The null distribution of the test statistic



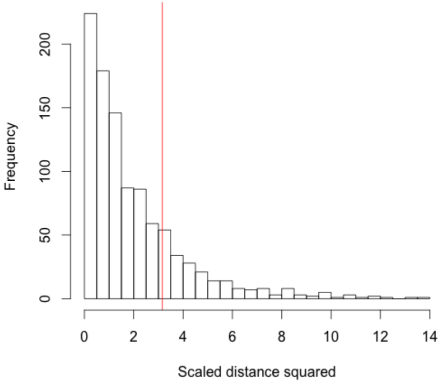
$(\beta_{i1}^*, \beta_{i0}^*)$, $i = 1, 2, \ldots, 1000$



Histogram of $\{T_i^*\}$, $i = 1, 2, \ldots, 1000$ and actual value of T for CCSM4 and HadCRUT4 (red line).

Hypothesis test result

4. Locate T in the distribution $pdf(T; H_0)$ and determine how extreme T is.



- $P(T^* \geq T) = .199$.
- We do *not* reject the null hypothesis that the two series share the same climate signal at $\alpha = .05$.

Conclusion

If CCSM4 and HadCRUT4 really did share the same climate signal (as we have defined it), then we would obtain values of the test statistic T as larger or larger than that computed from the original CCSM4 and HadCRUT4 time series with probability $p = 0.199$.

Moreover, p determined in this way is a quantitative measure of the compatibility of the data (CCSM4 and HadCRUT4 time series) with the null hypothesis. CMIP5 models can be compared using this measure.

See Braverman, A., Chatterjee, S., Heyman, M., and Cressie, N.C. (2016) for details.

Acknowledgements

This work was partially performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. Government sponsorship acknowledged.

Contact information: `Amy.Braverman@jpl.nasa.gov`.

We would like to thank NASA's Earth Science Data Records Uncertainty Analysis Program, NSF, and the Australian Research Council for supporting this research.

References

Monice, C.P., Kennedy, J.J., Rayner, N.A. and Jones, P.D., 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: the HadCRUT4 dataset. J. Geophys Res., 117, D08101.

Gent, P.R., Danabasoglu, G., Donner, L.J., Holland, M.M., Hunke, E.C., Jayne, S.R., Lawrence, D.M., Neale, R.B., Rasch, P.J., Vertenstein, M., Worley, P.H., Yang, Z-L., and Zhang, M., 2011: The Community Climate System Model Version 4. J. Clim., 24, pp. 4973–4991. DOI: 10.1175/2011JCLI4083.1.

Braverman, A., Chatterjee, S., Heyman, M., and Cressie, N.C., 2016: Probabilistic evaluation of competing climate models. Submitted to J. Clim.