Likelihood-based Climate Model Evaluation

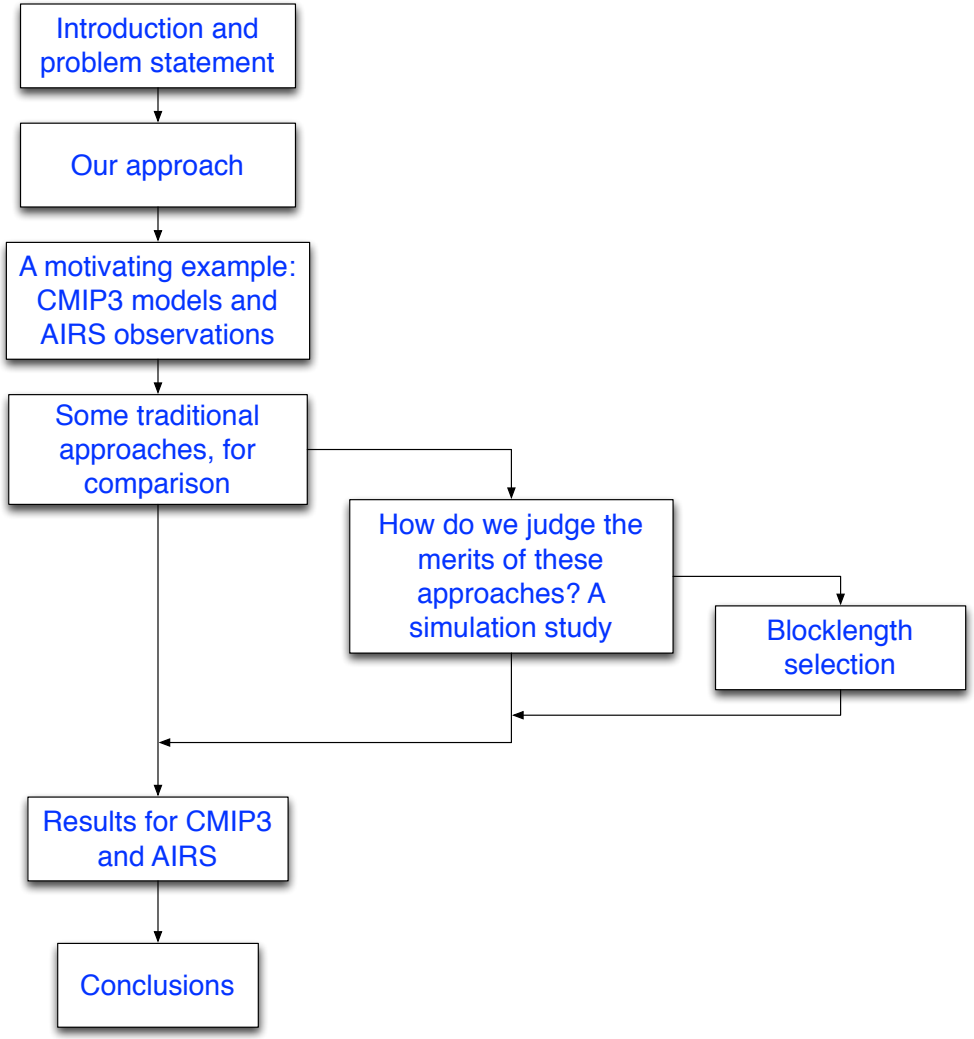
Amy Braverman[1] Noel Cressie[2] Joao Teixeira[1] Greg Huey[1]

[1]Jet Propulsion Laboratory,
California Institute of Technology

[2]Department of Statistics,
The Ohio State University

April 18, 2012

Road map


Introduction and problem statement
Our approach
A motivating example: CMIP3 models and AIRS observations
Some traditional approaches, for comparison
How do we judge the merits of these approaches? A simulation study
Blocklength selection
Results for CMIP3 and AIRS
Conclusions

Introduction

- 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.
- Quantify the likelihood that a (summary statistic computed from a) set of observations arises from a physical system with the characteristics captured by a model-generated time series.
- Given a prior on models, we can go further: posterior distribution of model given observations.

Approach

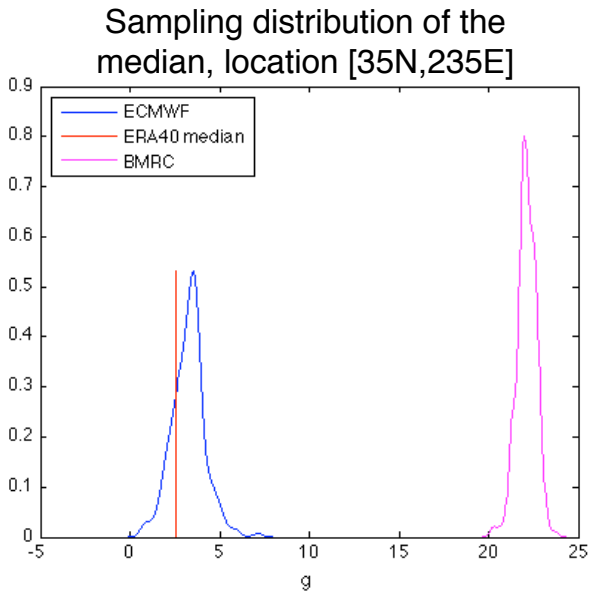
If the atmosphere behaves as the model specifies, then we would expect the observations to look like the model output to within the inherent variability of the model output.

Observations: $\mathbf{Y}_0 = \left(Y_{01}, \ldots, Y_{0N_0}\right)'$.

Output of model j: $\mathbf{Y}_j = \left(Y_{j1}, \ldots, Y_{jN_j}\right)'$.

Statistic: $g(\cdot)$: $g(\mathbf{Y}_0) = g_0$, $g(\mathbf{Y}_j) = g_j$.

Estimate the sampling distribution of g_j by resampling.

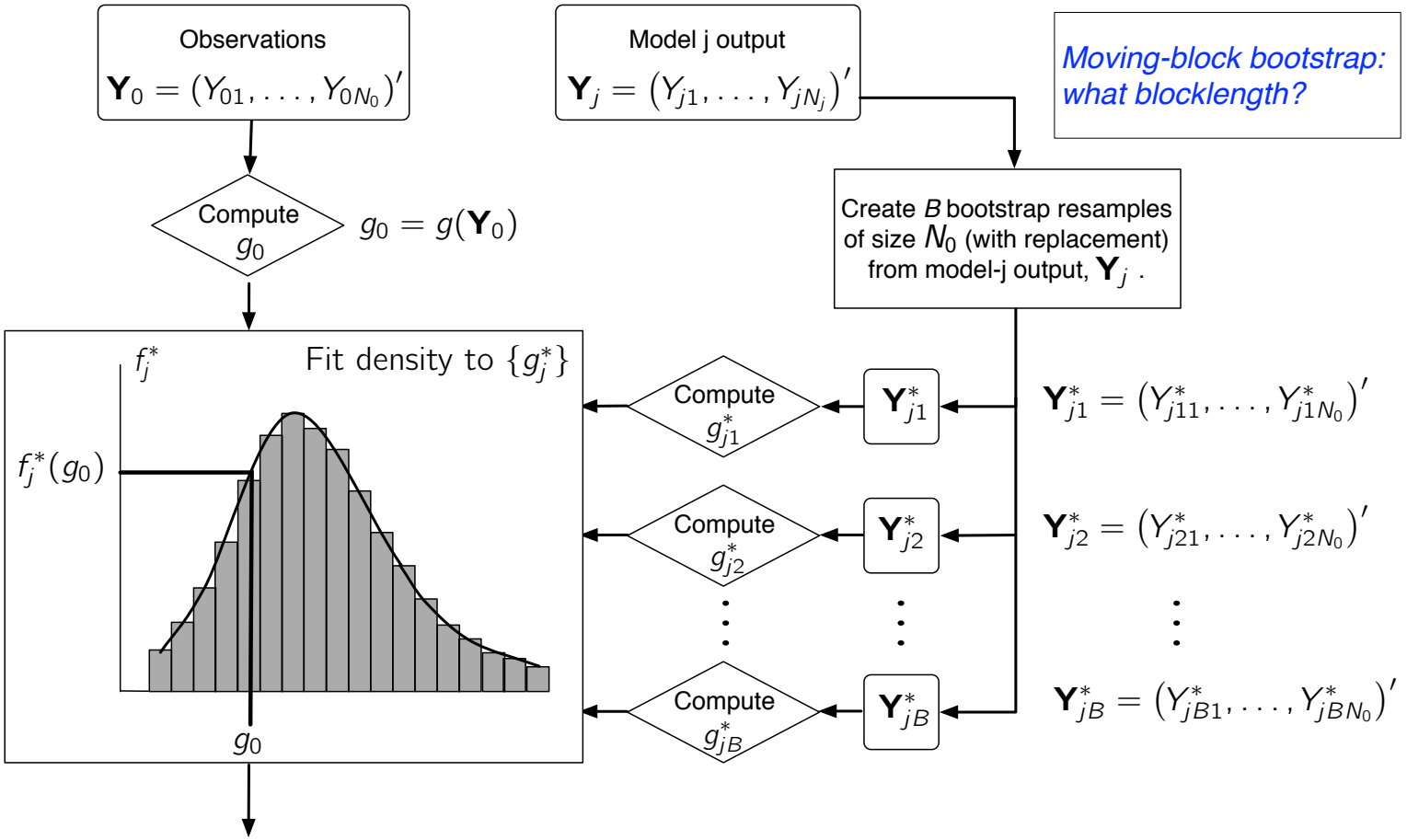


Likelihood of observing g_0 given model j sampling distribution is a figure of merit.

Approach

- Let $A = j$ be the event that model j best represents the physical system.
- Let $g_0 = g(\mathbf{Y}_0)$ be a statistic computed from the time series of observations.
- Let $f(x|A = j)$ be the sampling distribution (density) of that statistic given $A = j$.
- $f(g_0|A = j)$ is the likelihood of g_0 given $A = j$.
- $P(g_0|A = j) = \int_{g_0-\epsilon/2}^{g_0+\epsilon/2} f(x|A = j)dx, \quad \epsilon$ small.
- $P(A = j|g_0) \propto P(g_0|A = j)P(A = j)$.

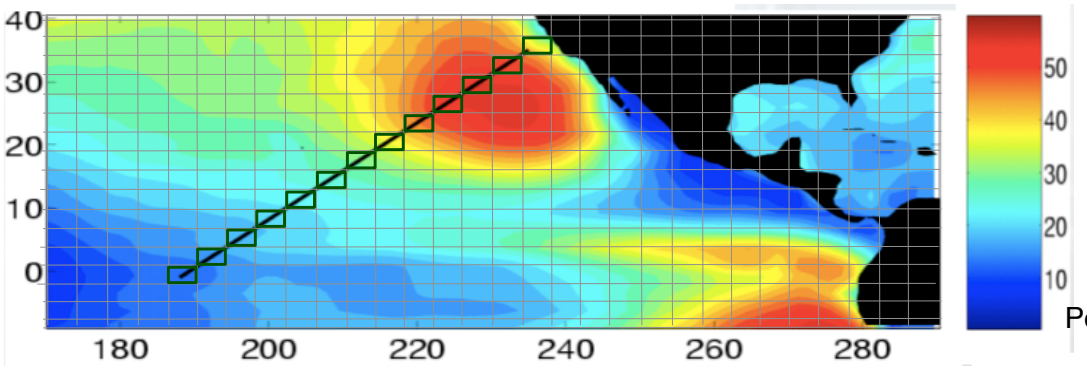
Approach



$\hat{f}(g(\mathbf{Y}_0)|A = j) = f^*(g_0)$ ⟵ Empirical likelihood

A motivating application: GPCI

- "Water vapor changes represent the largest feedback affecting climate sensitivity... Cloud feedbacks remain the largest source of uncertainty..." (IPCC 2007).



GPCI Transect

GCSS = GEWEX Cloud System Study.

Color: International Satellite Cloud Climatology Project low-cloud cover.

GCSS grid cells = $3^\circ \times 4^\circ$

- GCSS Pacific Cross-section Intercomparison Project (GPCI):
  - Study important physical regimes and transitions.
  - Evaluate models and observations in the tropics and sub-tropics in terms of hydrological cycle.
  - Utilize a new generation of satellite data sets.

A motivating application: GPCI

Model time series:

- Coupled Model Intercomparison Project (CMIP3) “AMIP" runs forced with observed sea-surface temperatures.
- Monthly time series of specific humidity from late 1970’s through early 2000’s at varying spatial resolutions.
- Time series range from 228 to 300 months.
- Multiple atmospheric levels- we concentrate on 850 hPa.
- For each model, data for model grid cells entirely contained within GCSS grid cells are averaged to form time series.

CMIP3 models:

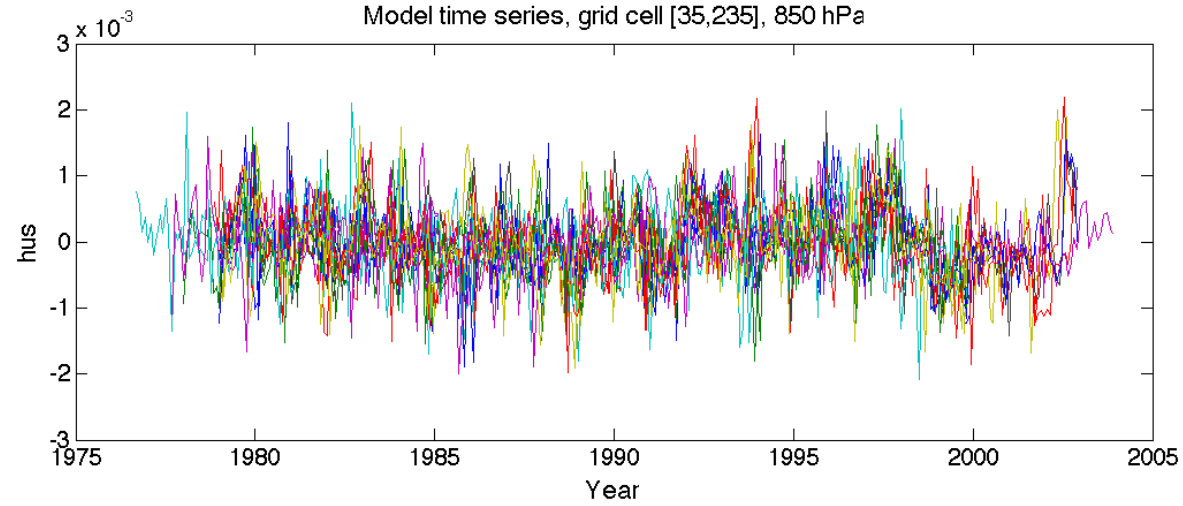
- CMRM_CM3
- GFDL_CM2_1
- GISS_MODEL_E_R
- IAP_FGOALS1_0_G
- INMCM3_0
- IPSL_CM4
- MIROC3_2_HIRES
- MIROC3_2_MEDRES
- MPI_ECHAM5
- MRI_CGCM2_3_2A
- NCAR_CCSM3_0
- NCAR_PCM1
- UKMO_HADGEM1

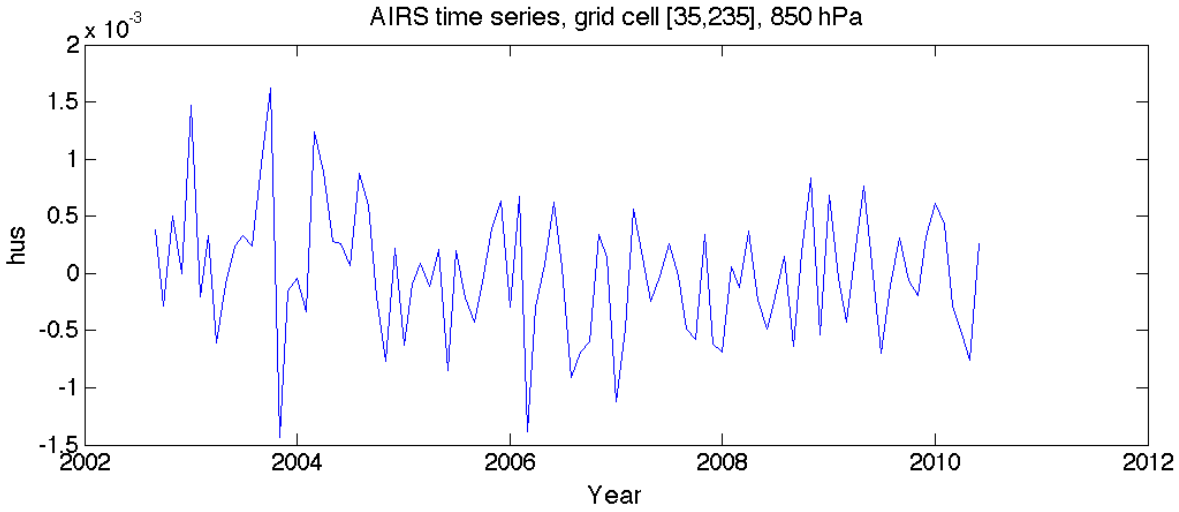
A motivating application: GPCI

Observational time series:

- Specific humidity from NASA's Atmospheric Infrared Sounder (AIRS) instrument, using the AIRS "IPCC" data set.
- Monthly time series from September 2002 through June 2010 (94 months) at $1^\circ \times 1^\circ$ spatial resolution.
- Multiple atmospheric levels interpolated to match model levels. We use 850 hPa.
- Data for AIRS grid cells entirely contained within GCSS grid cells are averaged to form observational time series.

A motivating application: GPCI





- Working with anomalies (annual cycle removed).
- No one-to-one match-up of time points.
- Here, model runs and observations barely overlap.
- Which statistics are important?

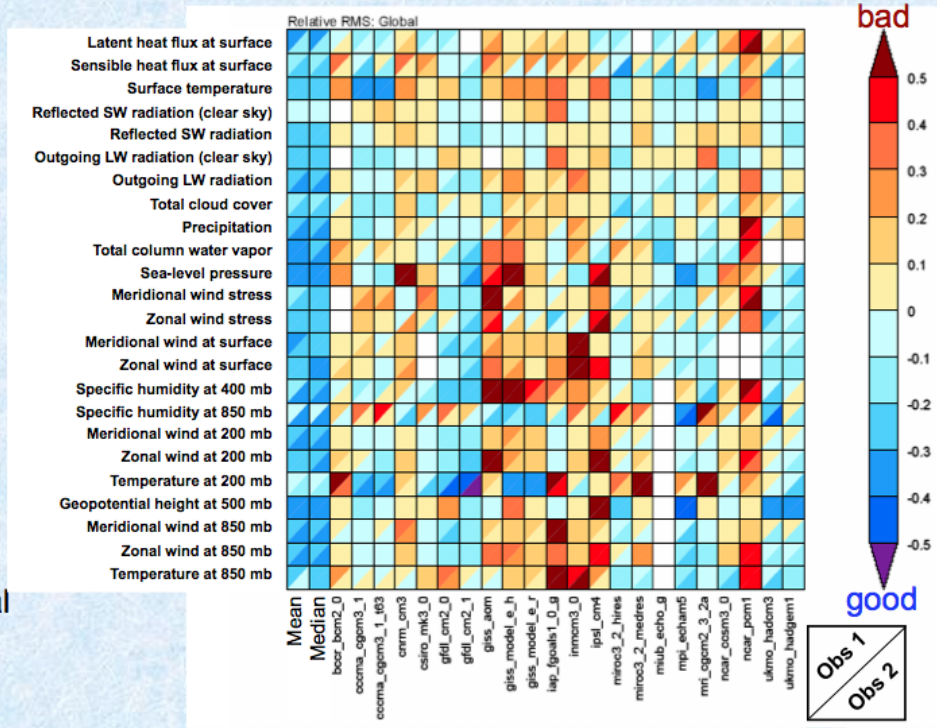
Other approaches

Annual cycle of global fields: Assessment of the relative skill (S) of individual CMIP3 models.

E_{vm} = RMS error in simulating the spatial pattern of the climatological annual cycle of variable *v* by model *m*

$$S_{vm} = \frac{E_{vm} - \hat{E}_v}{\hat{E}_v}$$

where $\hat{E}_v$ is the median of the individual error measures, E_{vm}



Gleckler, et al., J. Geophys.Res., 2008

- Model “metrics" based on simple descriptive statistics of discrepancies between model output and observations.
- Heritage from weather forecast verification.

Other approaches

Two specific examples:

- Mean squared error:

$$d_1(\mathbf{Y}_j, \mathbf{Y}_{j'}) = \frac{1}{M}\|\mathbf{Y}_j - \mathbf{Y}_{j'}\|^2,$$

  where $\mathbf{Y}_j$ and $\mathbf{Y}_{j'}$ are two time series of length M. (Pierce et al., 2009).

- Scaled difference of means:

$$d_2(\mathbf{Y}_j, \mathbf{Y}_{j'}) = \frac{|\bar{\mathbf{Y}}_j - \bar{\mathbf{Y}}_{j'}|}{3s_j},$$

  where s_j is the standard deviation of the elements of $\mathbf{Y}_j$. (Waugh and Eyring, 2008).

A simulation study

How will we know how well this works? How do we judge?

A simulation study:

Consider six moving-average models with orders of dependence $\omega = 0, 2, 4, 6, 8$, and 10, respectively. Index the models by $j = 1, 2, 3, 4, 5, 6$ so that $\omega(j) = 2(j-1)$.

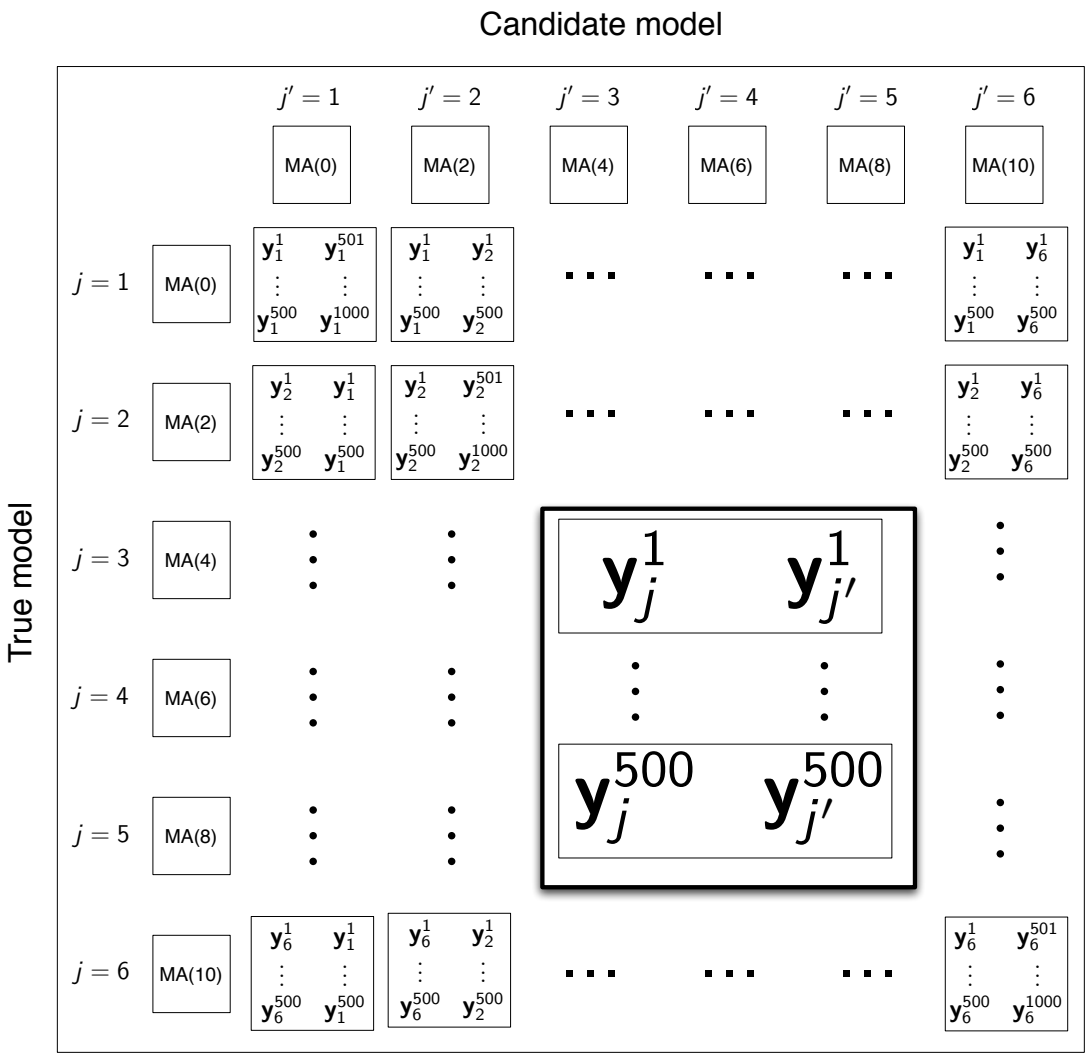
A generic realization from model j is $\mathbf{Y}_j = \{Y_{jn} : n = 1, \ldots, 1000 - \omega(j)\}$,

$$Y_{jn} \equiv \frac{1}{\sqrt{\omega(j)+1}} \sum_{i=n}^{n+\omega(j)} e_i, \quad e_i \sim \chi^2(1) - 1, \ \ iid.$$

Our experiment: let each model $j = 1, \ldots, 6$ successively represent the “true" model, and evaluate all six models $j' = 1, \ldots, 6$ against it.

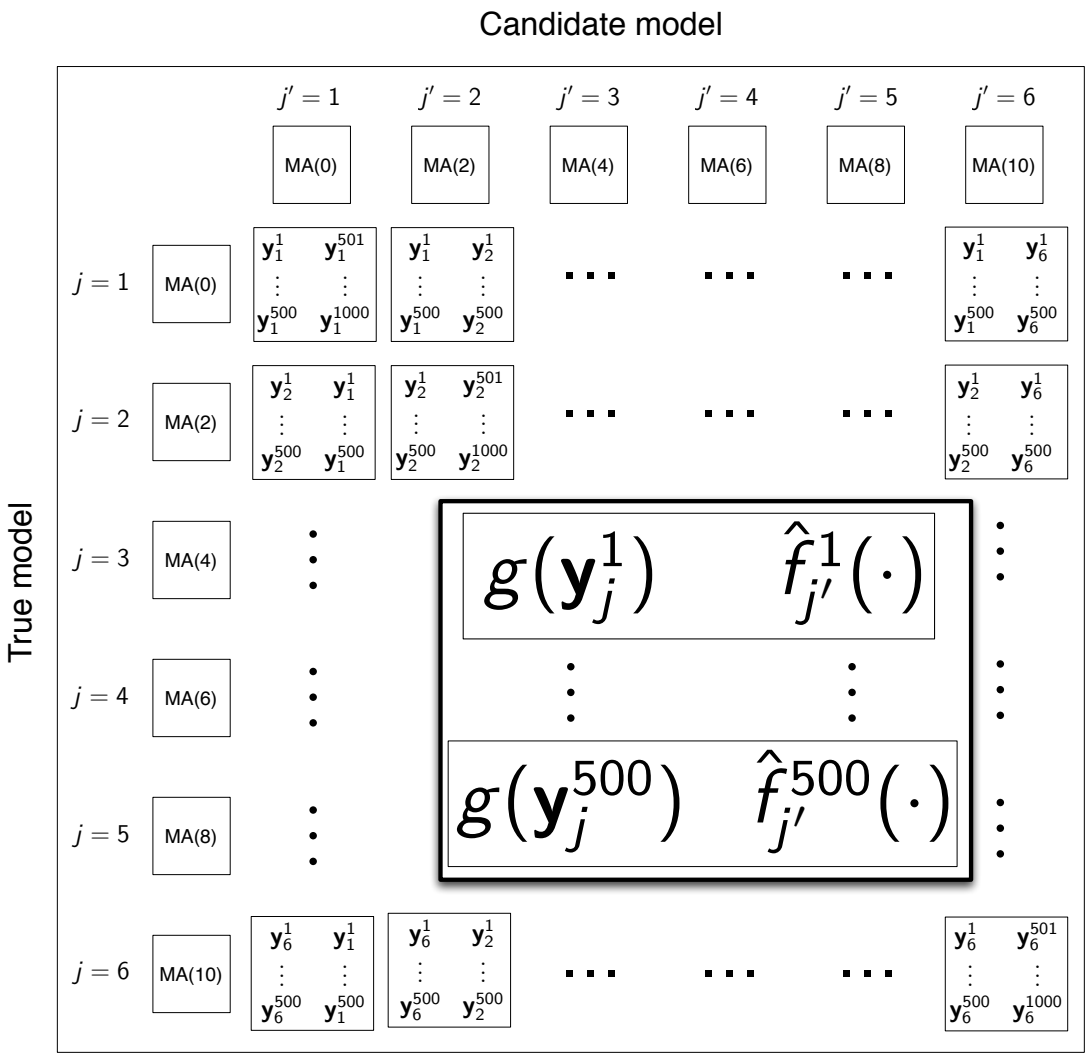
For each (j, j') combination, base the evaluation on $K = 500$ realizations from $\mathbf{Y}_j$ and $K = 500$ independent realizations form $\mathbf{Y}_{j'}$.

A simulation study



y_j^v is the vth time series generated by model j.

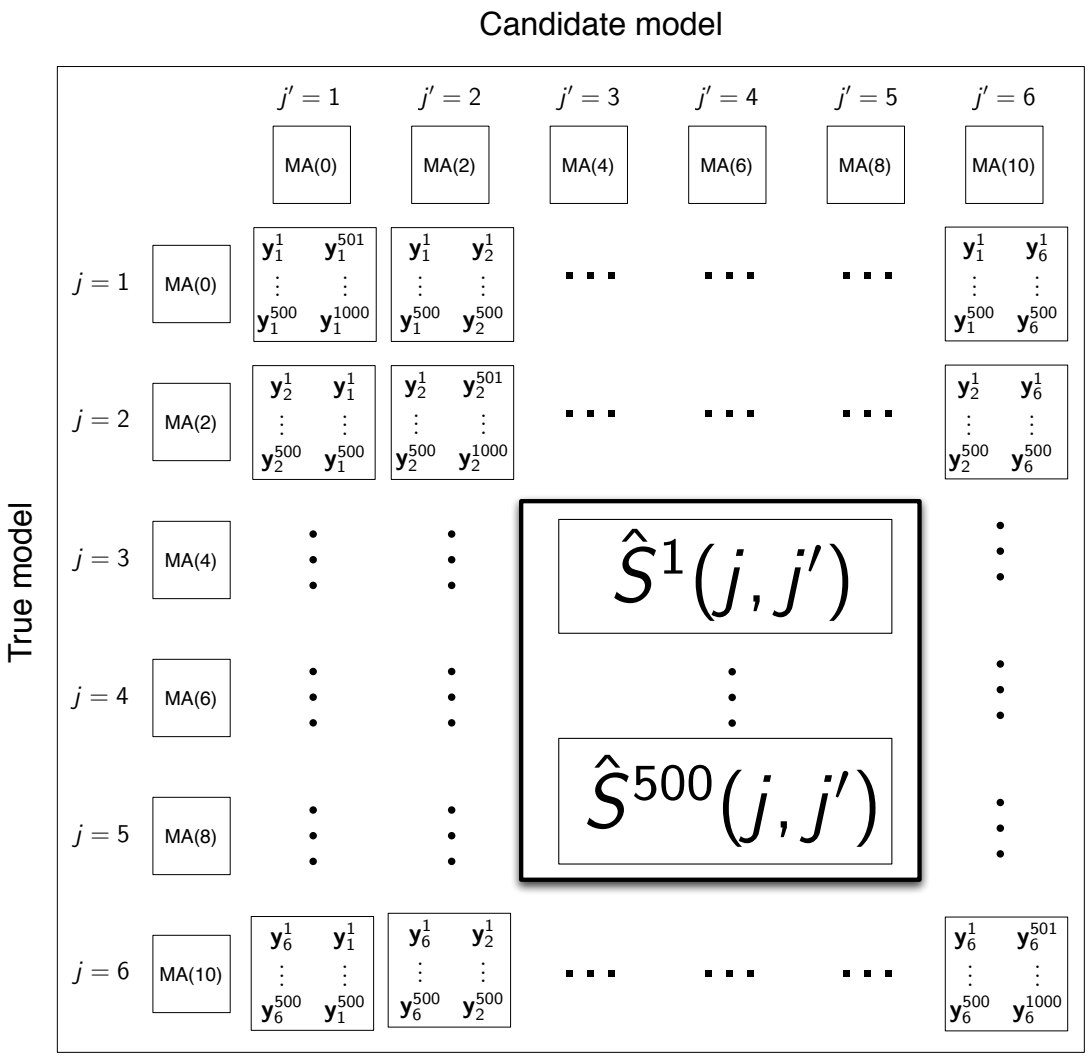
A simulation study



Evaluate:

$$\hat{S}^v(j, j') \equiv \hat{f}_{j'}^v(g(y_j^v)),$$
$$v = 1, \ldots, 500,$$
$$j = 1, \ldots, 6.$$

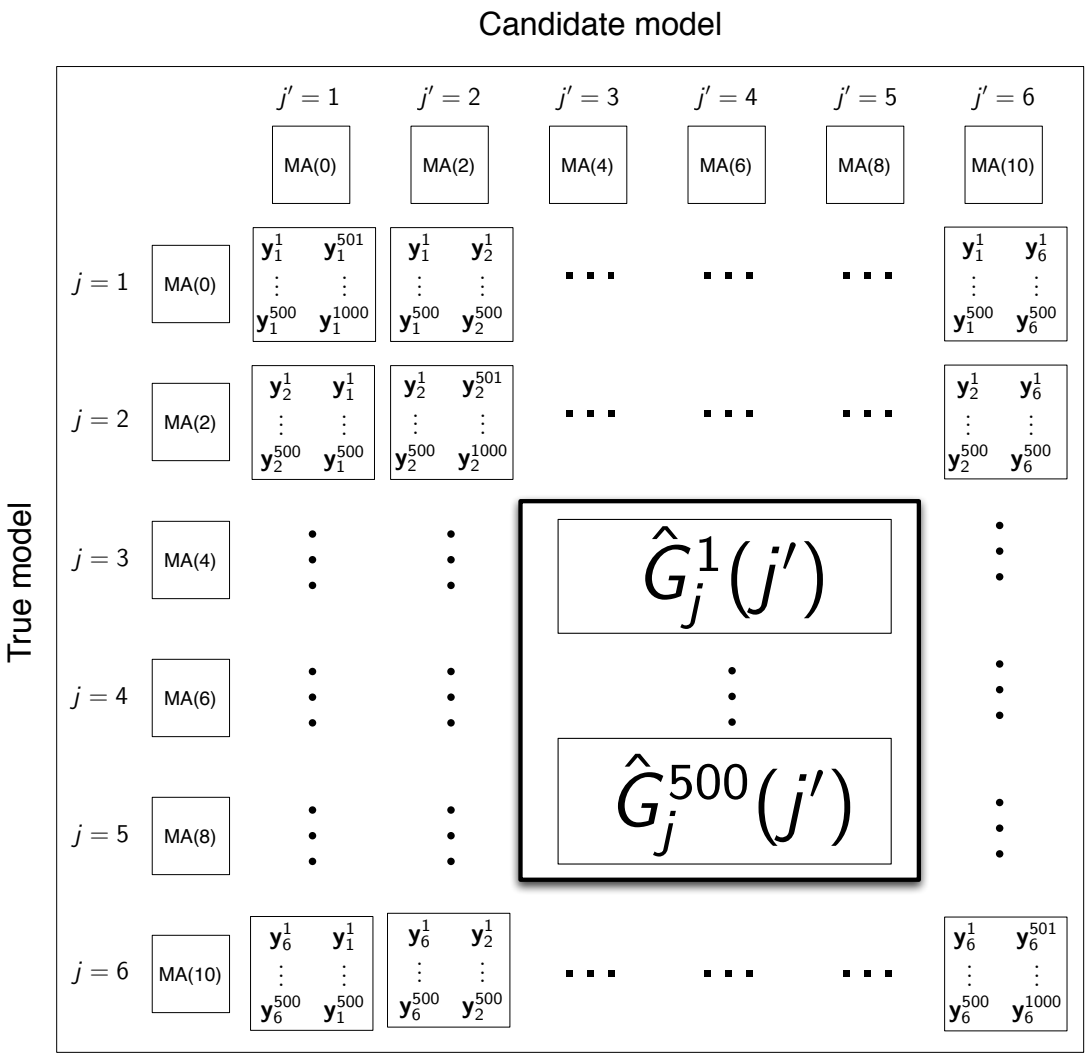
A simulation study



Evaluate:

$$\hat{S}^v(j, j') \equiv \hat{f}^v_{j'}(g(y^v_j)),$$
$$v = 1, \ldots, 500,$$
$$j = 1, \ldots, 6.$$

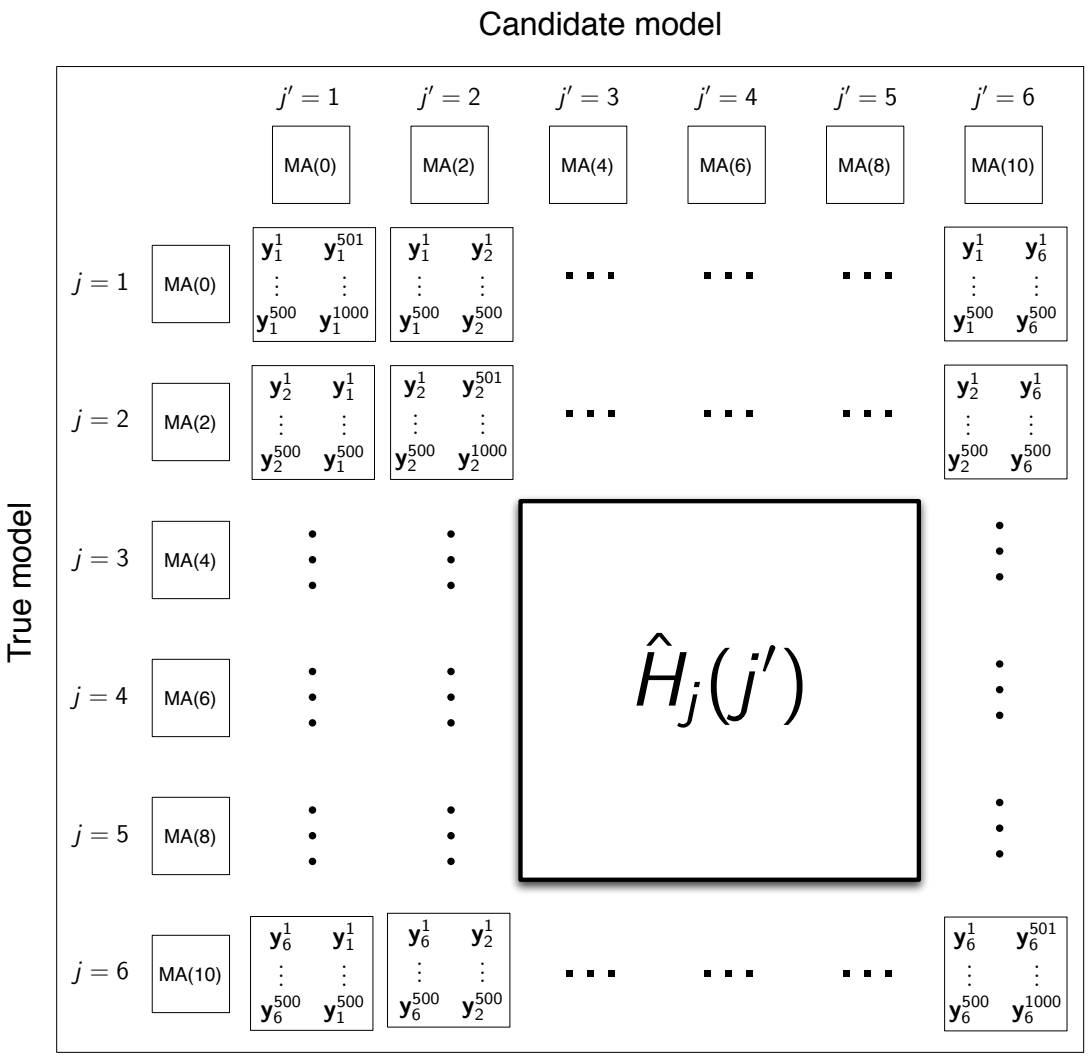
A simulation study



$$\hat{G}_j^v(j') = \frac{\hat{S}^v(j,j')}{\max_{\{m\}}\{\hat{S}^v(j,m)\}},$$

$$m = 1, 2, 3, 4, 5, 6.$$

A simulation study



$$\hat{H}_j(j') = \text{median}_{\{v\}}\{\hat{G}_j^v(j')\}.$$

A simulation study

Candidate model

True model

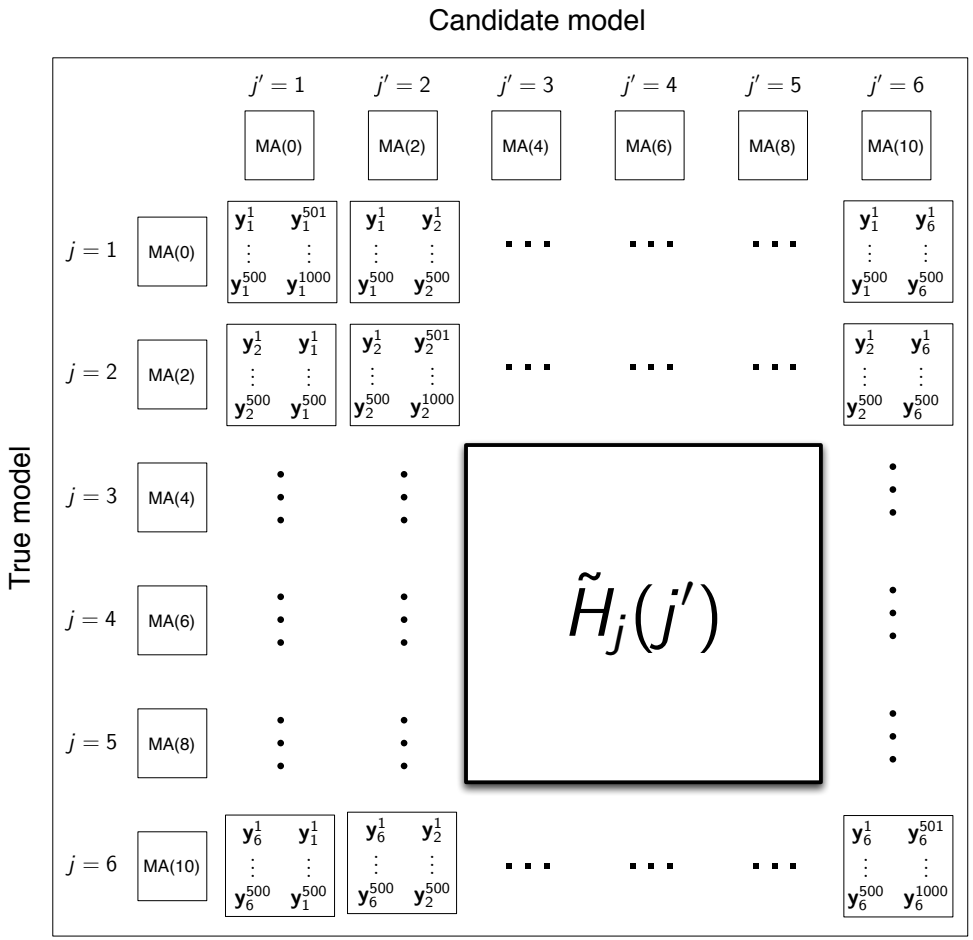
		$j' = 1$ MA(0)	$j' = 2$ MA(2)	$j' = 3$ MA(4)	$j' = 4$ MA(6)	$j' = 5$ MA(8)	$j' = 6$ MA(10)
$j = 1$	MA(0)	$\hat{H}_1(1)$	$\hat{H}_1(2)$	$\cdots$	$\cdots$	$\cdots$	$\hat{H}_1(6)$
$j = 2$	MA(2)	$\hat{H}_2(1)$	$\hat{H}_2(2)$	$\cdots$	$\cdots$	$\cdots$	$\hat{H}_2(6)$
$j = 3$	MA(4)	$\vdots$	$\vdots$	$\ddots$	$\cdots$	$\cdots$	$\vdots$
$j = 4$	MA(6)	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\cdots$	$\vdots$
$j = 5$	MA(8)	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$j = 6$	MA(10)	$\hat{H}_6(1)$	$\hat{H}_6(2)$	$\cdots$	$\cdots$	$\cdots$	$\hat{H}_6(6)$

$$\hat{H}_j(j') = \text{median}_{\{v\}}\{\hat{G}_j^v(j')\}.$$

Repeat for three summary statistics:

$$g = q_{.25},$$
$$g = q_{.50},$$
$$g = q_{.75}.$$

A simulation study



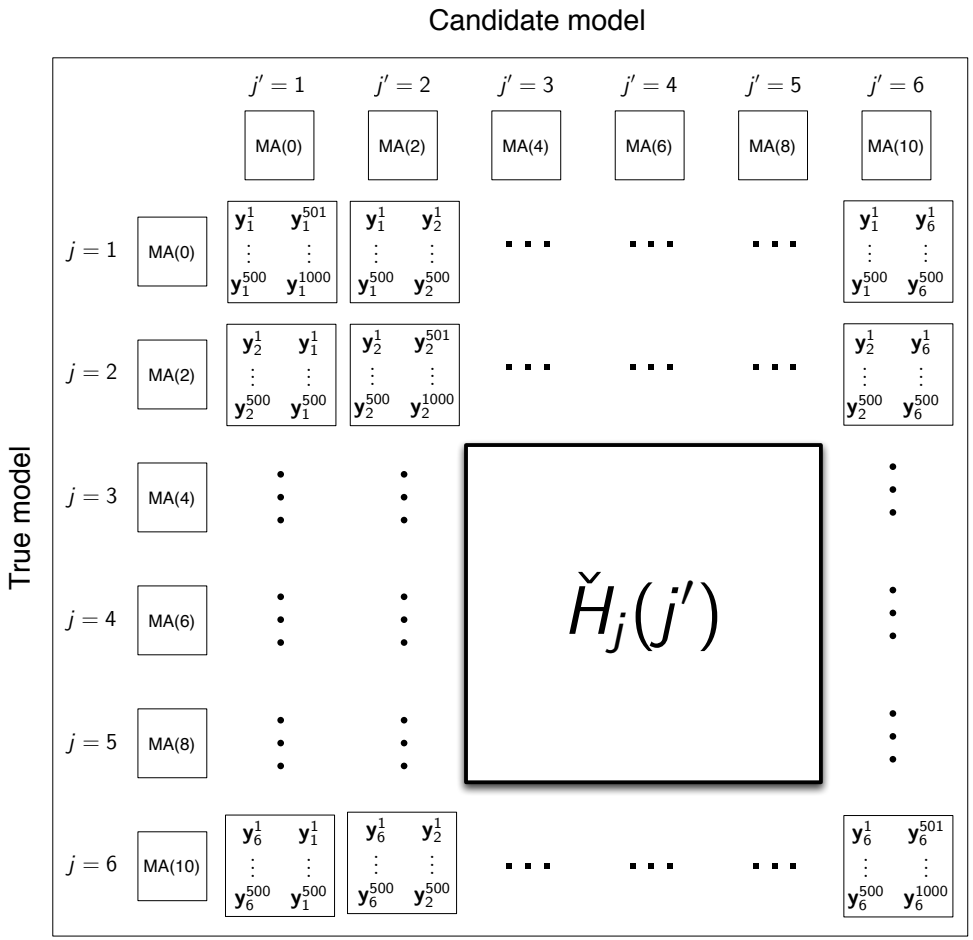
$$\tilde{S}^v(j,j') = d_1(\mathbf{Y}_j, \mathbf{Y}_{j'}) = \frac{1}{M}\|\mathbf{Y}_j - \mathbf{Y}_{j'}\|^2,$$

$$\tilde{G}_j^v(j') = \frac{\tilde{S}^v(j,j')}{\max_{\{m\}}\{\tilde{S}^v(j,m)\}},$$

$$m = 1, 2, 3, 4, 5, 6,$$

$$\tilde{H}_j(j') = \text{median}_{\{v\}}\{1 - \tilde{G}_j^v(j')\}.$$

A simulation study



$$\check{S}^v(j,j') = d_2(\mathbf{Y}_j, \mathbf{Y}_{j'})$$

$$= \frac{|\bar{\mathbf{Y}}_j - \bar{\mathbf{Y}}_{j'}|}{3s_j},$$

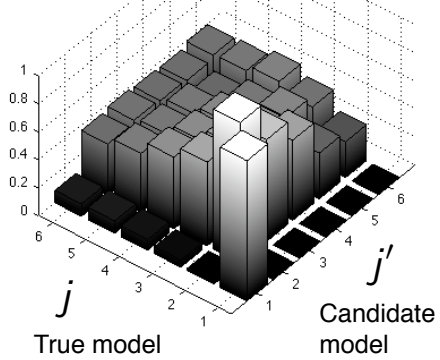
$$\check{G}_j^v(j') = \frac{\check{S}^v(j,j')}{\max_{\{m\}}\{\check{S}^v(j,m)\}},$$

$$m = 1, 2, 3, 4, 5, 6,$$

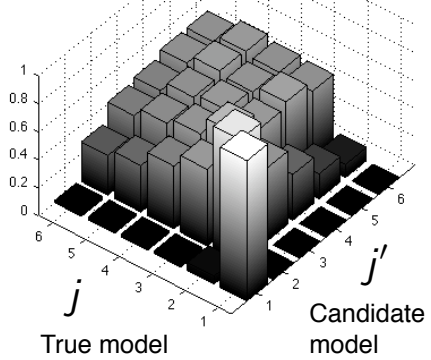
$$\check{H}_j(j') = \text{median}_{\{v\}}\{1 - \check{G}_j^v(j')\}.$$

A simulation study

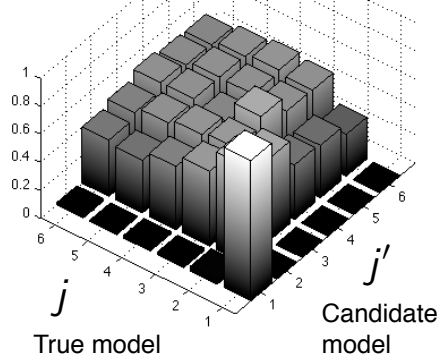
$\hat{H}_j(j'),\ g = q_{.25}$



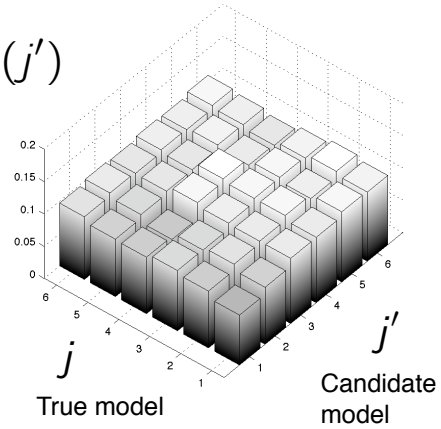
$\hat{H}_j(j'),\ g = q_{.50}$



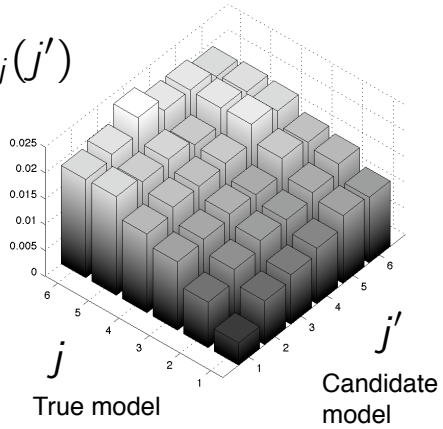
$\hat{H}_j(j'),\ g = q_{.75}$



$\tilde{H}_j(j')$



$\check{H}_j(j')$



A simulation study

Are these results significant?

A perfect result has high values on the diagonal, and zeros elsewhere.

Measure departure from the perfect result by

$$D = \sum_{j=1}^{6} \sum_{j'=1}^{6} \left(w_{jj'} r_{j'|j} \right)^2.$$

$w_{jj'}$

j	6	5	4	3	2	1
	5	6	5	4	3	2
	4	5	6	5	4	3
	3	4	5	6	5	4
	2	3	4	5	6	5
	1	2	3	4	5	6
				j'		

$r_{j'|j}$

j	3	1	6	4	2	5
	5	3	2	1	6	4
	1	5	2	6	4	3
	6	4	3	1	2	5
	4	1	2	6	3	5
	3	1	2	5	4	6
				j'		

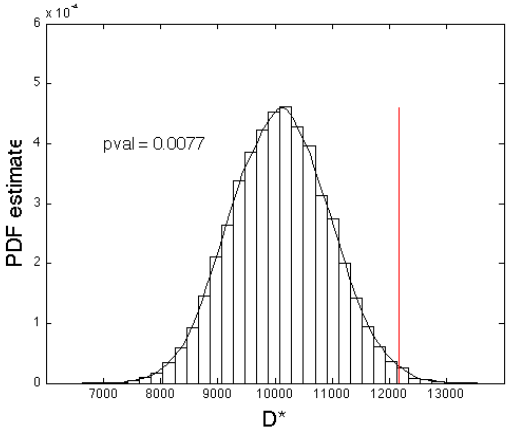
$r_{j'|j}$ is within-row rank.

$w_{jj'}$ is index-difference from the diagonal.

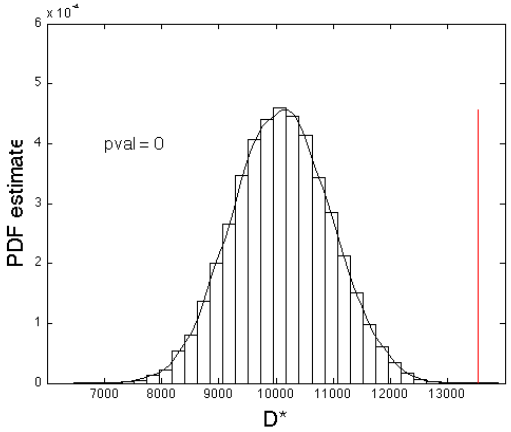
Null distribution of D obtained by permuting $r_{j'|j}$ 20,000 times.

A simulation study

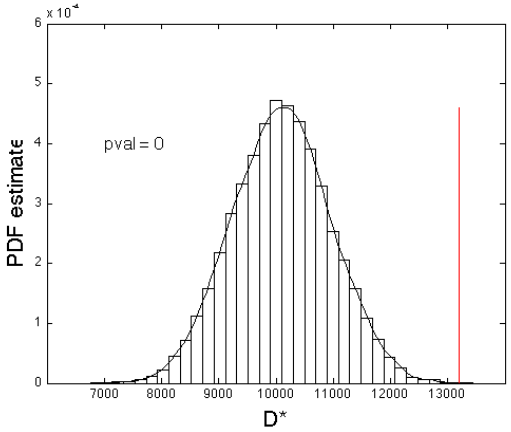
$\hat{H}_j(j'),\ g = q_{.25}$



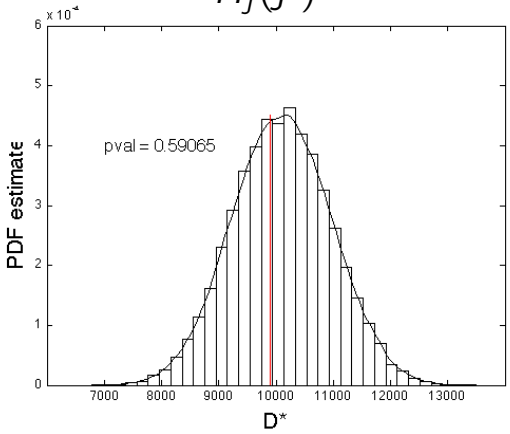
$\hat{H}_j(j'),\ g = q_{.50}$



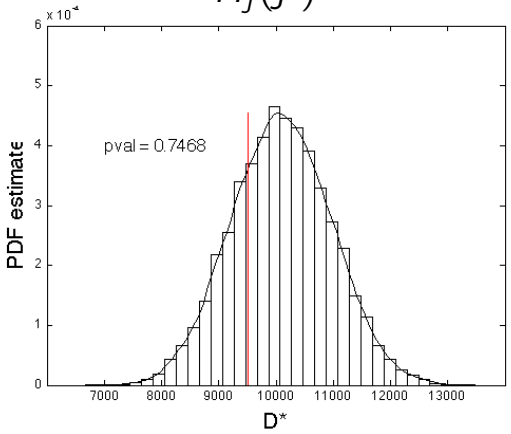
$\hat{H}_j(j'),\ g = q_{.75}$



$\tilde{H}_j(j')$



$\check{H}_j(j')$

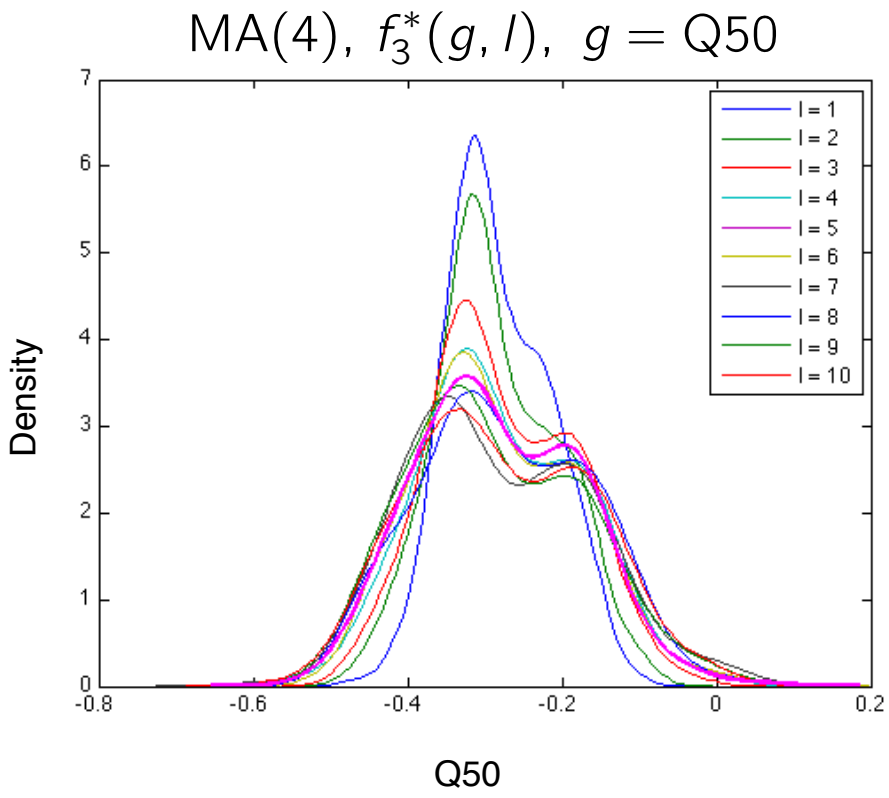


Blocklength selection

- Many approaches to blocklength selection for the MBB in the literature, e.g., Hall, Horowitz, and Jing (1995), Buhlmann and Kunsch (1999), Politis and White (2004), Bickel and Sakov (2008).
- Rely heavily on asymptotics and did not work well in our simple simulation experiments.
- Heuristic: for time series with temporal dependence, choosing blocklength too large is less problematic than choosing it too small.

Blocklength selection

- Sampling distribution of g as a function of blocklength, l, converges.
- Not to the true sampling distribution, but to something.
- Since “too long" is less problematic than “too short", we seek the smallest value of l beyond which the sampling distribution doesn’t change significantly.
- Call this “acceptable" blocklength.



Blocklength selection

- For the simulation study we tested blocklengths $1, 2, \ldots, 15$.

Acceptable Blocklength						
	MA Order (ω)					
g	0	2	4	6	8	10
$q_{.25}$	2	6	6	8	9	9
$q_{.50}$	2	5	7	8	9	10
$q_{.75}$	2	5	7	8	7	8

AIRS and CMIP3 models

Statistics:

- $g = q_{.05}$
- $g = q_{.25}$
- $g = q_{.50}$
- $g = q_{.75}$
- $g = q_{.95}$

Blocklengths tested:

$1, \ldots, 24$ (two year lag).

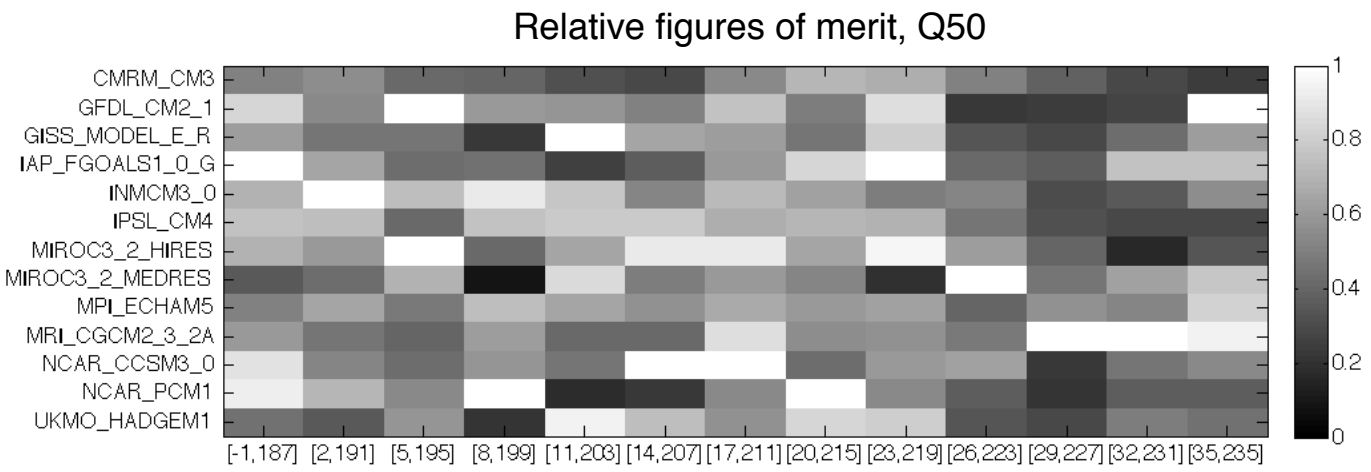
GPCI locations:

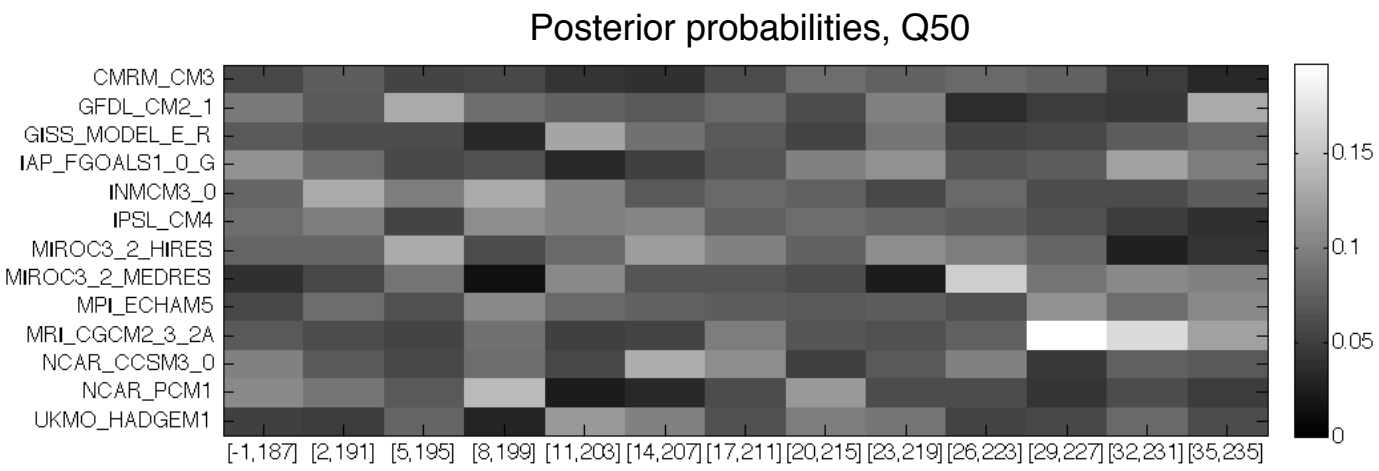
- [35,235]
- [32,231]
- [29,227]
- [26,223]
- [23,219]
- [20,215]
- [17,211]
- [14,207]
- [11,203]
- [8,199]
- [5,195]
- [2,191]
- [-1,187]

CMIP3 models:

- CMRM_CM3
- GFDL_CM2_1
- GISS_MODEL_E_R
- IAP_FGOALS1_0_G
- INMCM3_0
- IPSL_CM4
- MIROC3_2_HIRES
- MIROC3_2_MEDRES
- MPI_ECHAM5
- MRI_CGCM2_3_2A
- NCAR_CCSM3_0
- NCAR_PCM1
- UKMO_HADGEM1

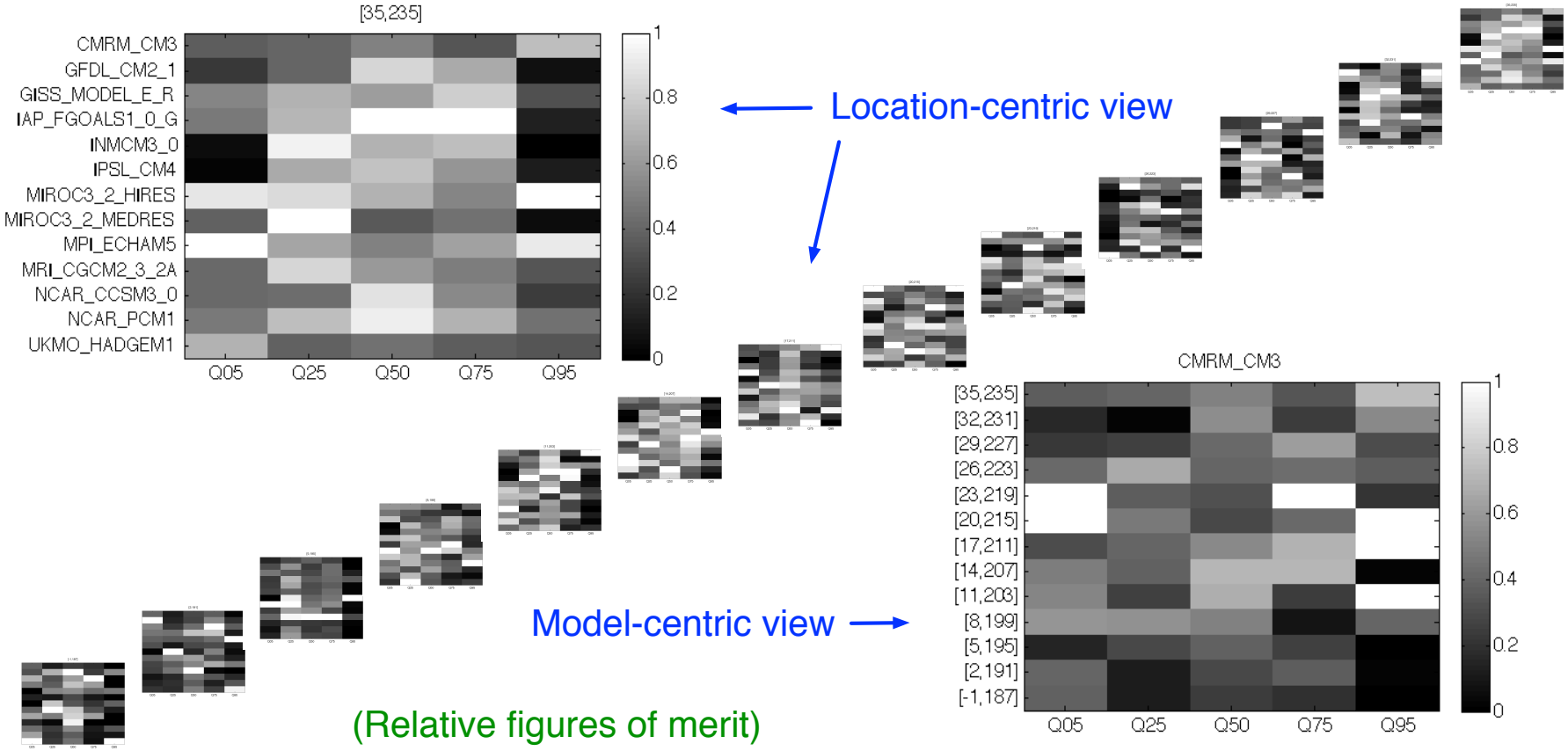
AIRS and CMIP3 models





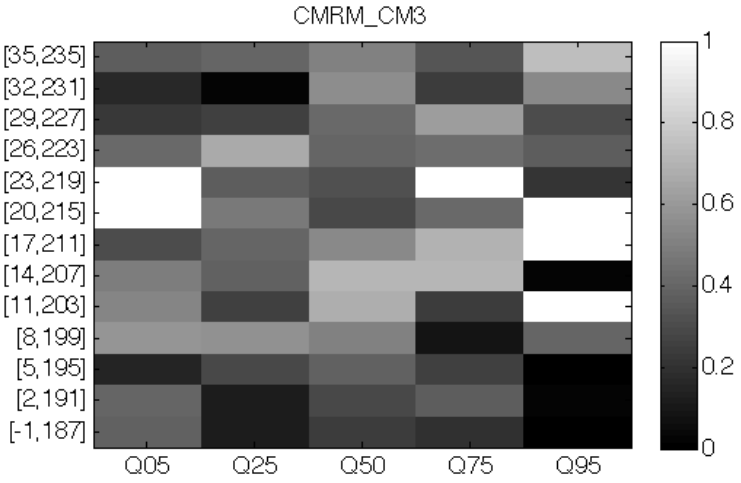
AIRS and CMIP3 models

Other views:

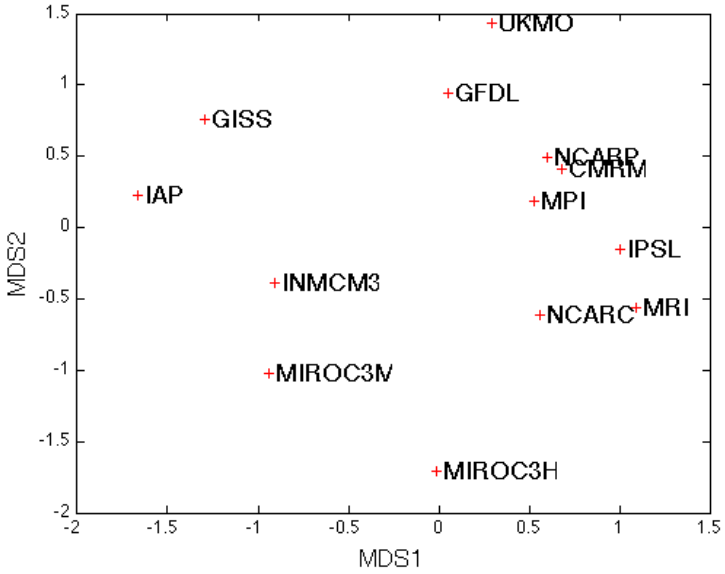


(Relative figures of merit)

AIRS and CMIP3 models



$\mathbf{X}_{\mathrm{CMRM_CM3}}$



- $\mathbf{X}_m$, $m = 1, \ldots, 13$ is the location-by-statisic matrix for model m.
- Compute D, the 13×13 distance matrix between $\mathbf{X}_{m_i}$ and $\mathbf{X}_{m_j}$ for all i, j, using the Froebenius norm.
- Perform multidimensional scaling on D. (Eigenvalues: 9.87, 8.87, 6.56, 6.05, 5.25, 4.51, 4.34, 2.99, 2.41, 2.10, 1.93, 1.48, 0.)

Conclusions (1)

- Statistical:
  - Method evaluates models according to how likely a statistic computed from observations is, given that models represent the system.
  - No Gaussian assumptions, but blocklength selection is crucial. Subject of ongoing work.
  - Express results as relative likelihoods or posterior probabilities. Likelihoods make comparisons easier, but probabilities are more interpretable.
  - Computation is not a problem as long as each grid box treated separately.
  - Next: evaluate models' ability to capture spatio-temporal statistics. Computational issues?

Conclusions (2)

- Scientific:
  - Which statistics are important?
  - Are the observational and the model variables really comparable?
  - What about uncertainty in the observations?
  - What is the overall objective?
    - Improve process representation?
    - Weights for multimodel ensembles? (“The end of model democracy?", Knutti (2010).)

The end

Questions, comments?

Contact Amy.Braverman@jpl.nasa.gov.

Support for this research is provided by NASA's Earth Science Data Records Uncertainty Analysis Program.

This work was performed partially at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. Government sponsorship acknowledged. Copyright 2012, all rights reserved.

References (1)

P. J. Gleckler, K. E. Taylor, and C. Doutriaux (2008). Performance metrics for climate models, Journal of Geophysical Research, 113, D06104.

D. W. Pierce, T. P. Barnett, B. D. Santer, and P. J. Gleckler (2009). Selecting global climate models for regional climate change studies, Proceedings of the National Academy of Science, 106, 21, 8441-8446.

D. W. Waugh and V. Eyring (2008). Quantitative performance metrics for stratospheric- resolving chemistry-climate models, Atmospheric Chemistry and Physics, 8, 5699-5713.

P. Hall, J.L. Horowitz, and B-Y. Jing (1995). On blocking rules for the bootstrap with dependent data, Biometrika, 82, 561-574.

P. Buhlmann and H.R. Kunsch (1999). Block length selection in the bootstrap for time series, Computational Statistics and Data Analysis, 31, 295-310.

References (2)

D.N. Politis and H. White (2004). Automatic block-length selection for the dependent bootstrap, Econometric Reviews, 23, 53-73.

P.J. Bickel and A. Sakov (2008). On the choice of m in the m out of n bootstrap and confidence bounds for extrema, Statistica Sinica, 18, 967- 985.

R. Knutti (2010). The end of model democracy? Climatic Change, 102, 3-4, 395-404.

Backup Slides

Reference

For more information, see:

A. Braverman, N. Cressie, and J. Teixeira (2011). A likelihood-based comparison of temporal models for physical processes, *Statistical Analysis and Data Mining*, Volume 4, Number 3, pp. 247- 258, doi: 10.1002/sam.10113.

The moving-block bootstrap:

Original series: $\mathbf{y}_j^1 = (y_{j1}^1, \ldots, y_{jN}^1)'$.

Create an MBB series:

1. Let $l =$ blocklength, $Q = \lfloor N/l \rfloor$, and $S = N - l + 1$.
2. Sample integers from the set $\{1, \ldots, S\}$ Q times with replacement to obtain $\{t_1^*, \ldots, t_Q^*\}$.
3. $\mathbf{y}_{jt}^{1*}(l) = (y_{jt}^1, y_{j(t+1)}^1, \ldots, y_{j(t+l-1)}^1)$.
4. $\mathbf{y}_j^{1*}(l) = \left(\mathbf{y}_{jt_1^*}^1(l), \ldots, \mathbf{y}_{jt_Q^*}^1(l)\right)'$.

Estimate sampling distribution of g:

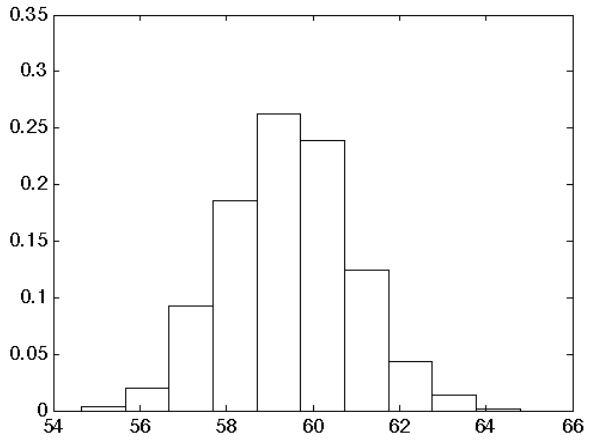
1. Create B MBB series, $\mathbf{y}_j^{1*}(l, b), \quad b = 1, \ldots, B$.
2. Fit kernel density estimate to $\{g(\mathbf{y}_j^{1*}(l, 1)), \ldots, g(\mathbf{y}_j^{1*}(l, B))\}$ to produce $f_j^*(g, l)$.

Moving-block bootstrap example

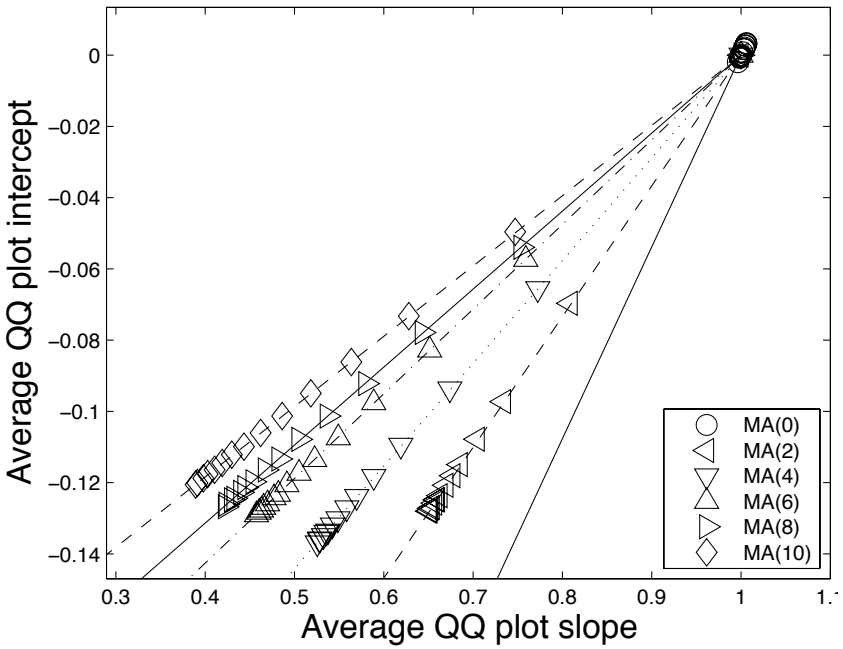
data indices	1	2	3	4	5	6	7	8	9	10
data values	**84**	**105**	**12**	**67**	**24**	**117**	**38**	**89**	**2**	**55**
block indices	1	2	3	4	5	6	7	8	9	

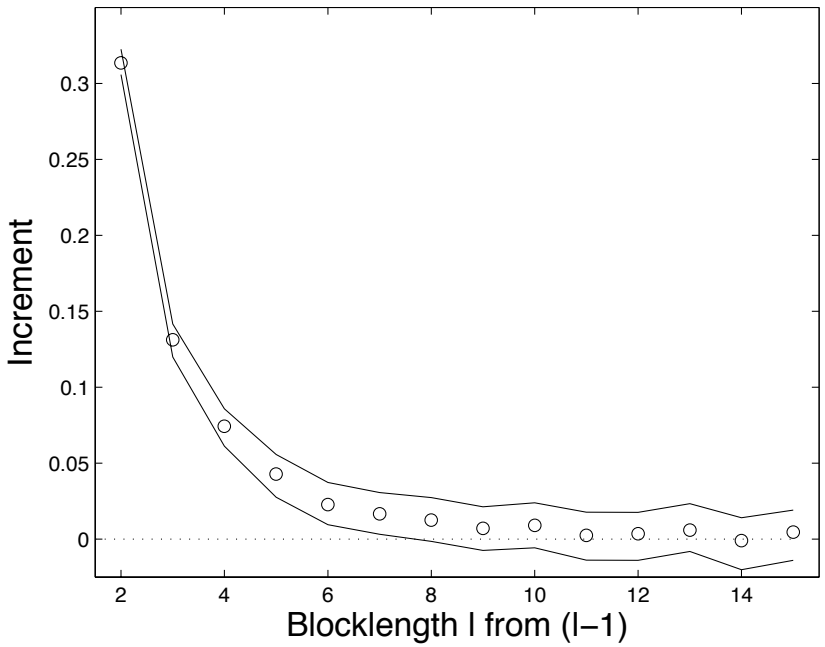
block indices in resample 1	5		9		1			2		3	
data values in resample 1	**24**	**117**	**2**	**55**	**84**	**105**	**105**	**12**	**12**	**67**	mean = 58.3
block indices in resample 2	8		6		2		6		4		
data values in resample 2	**89**	**2**	**117**	**38**	**105**	**12**	**117**	**38**	**67**	**24**	mean = 60.9
⋮											
block indices in resample B	9		7		8		1		2		
data values in resample B	**2**	**55**	**38**	**89**	**89**	**2**	**84**	**105**	**105**	**12**	mean = 58.1

Estimated sampling distribution of the mean

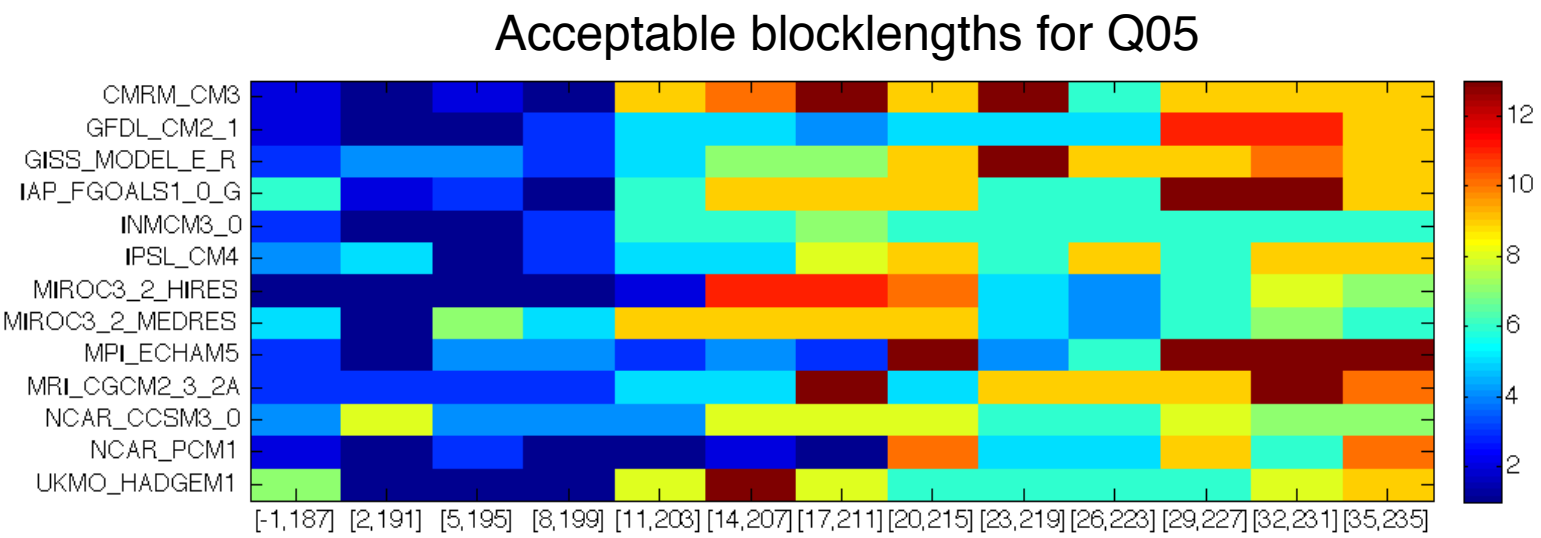


Blocklength selection

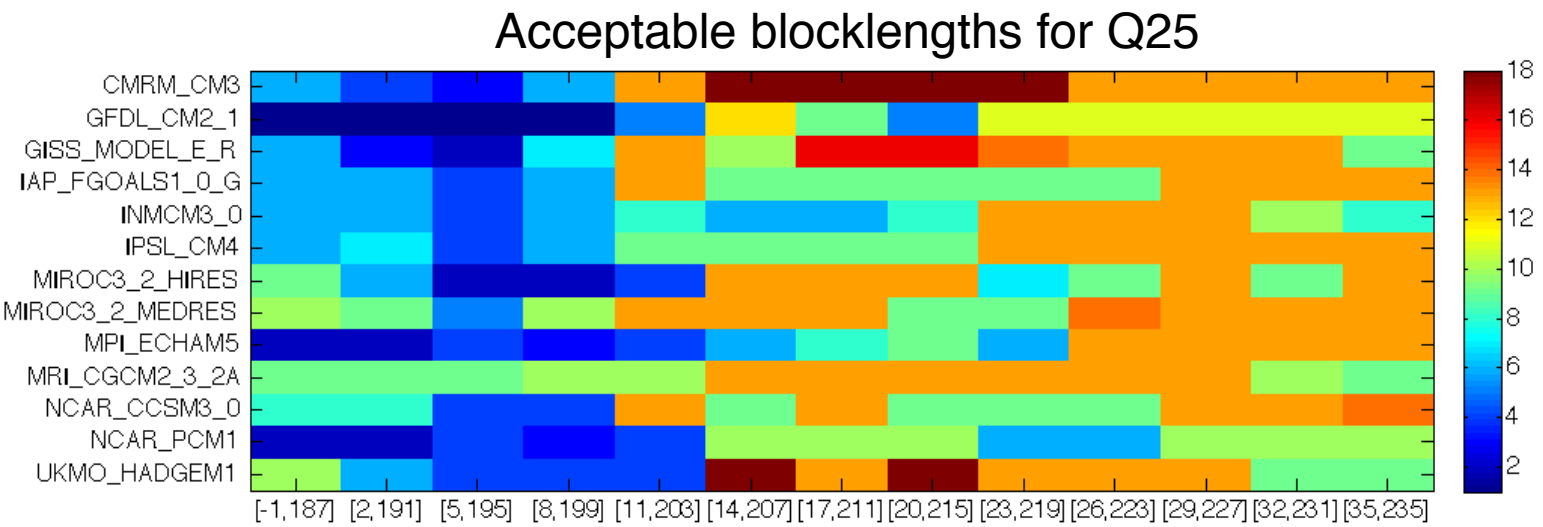




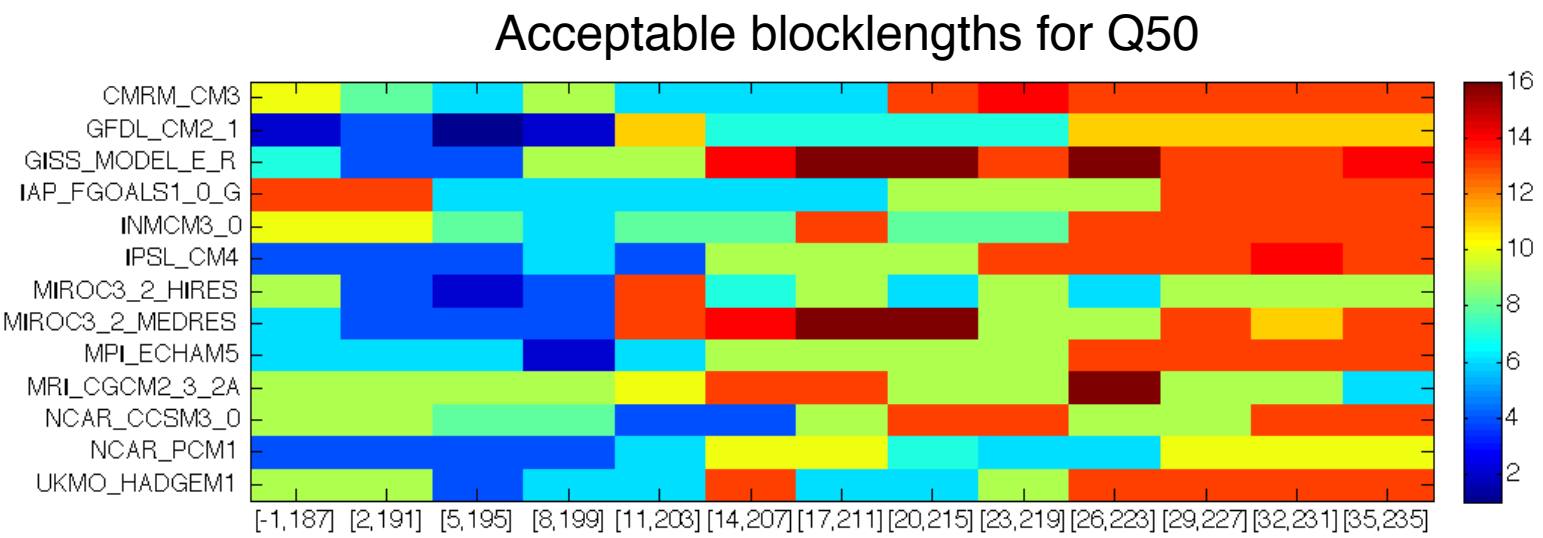
Blocklength selection



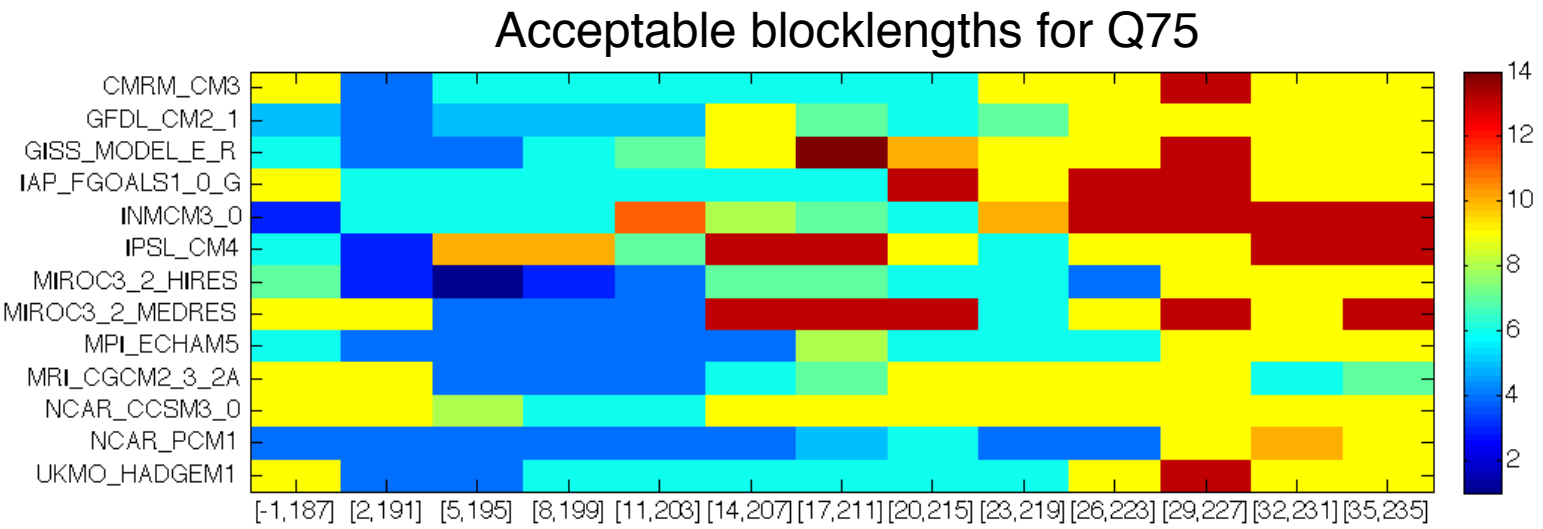
Blocklength selection



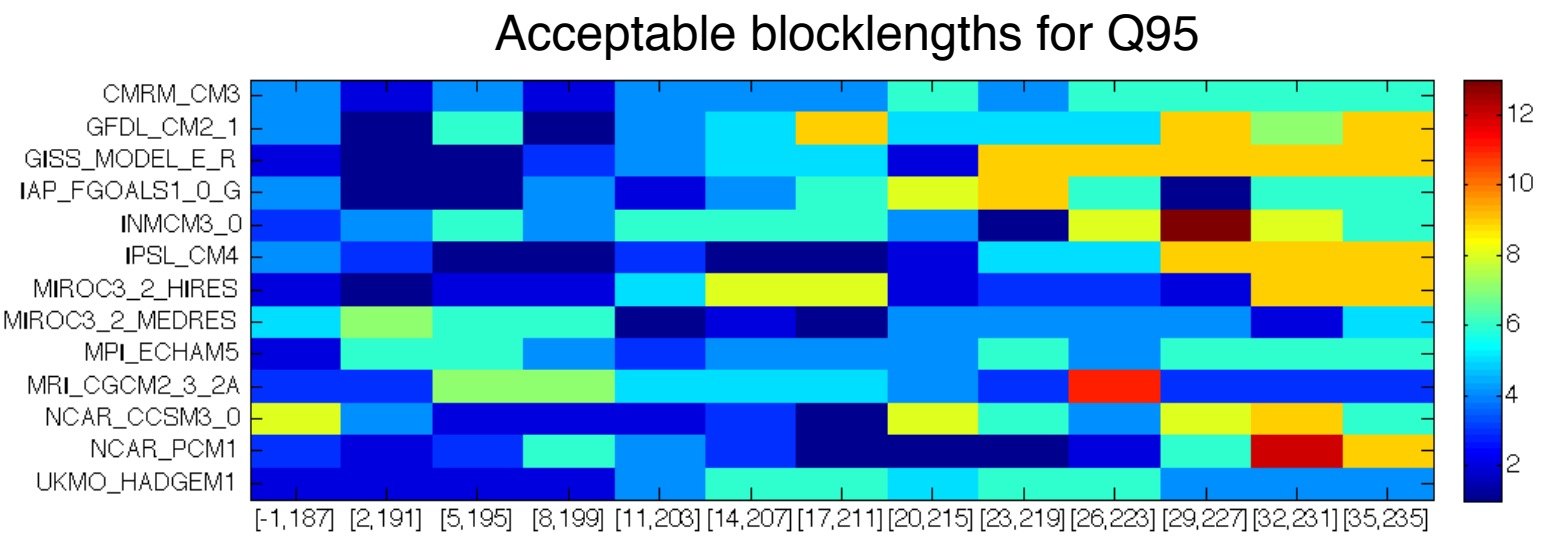
Blocklength selection



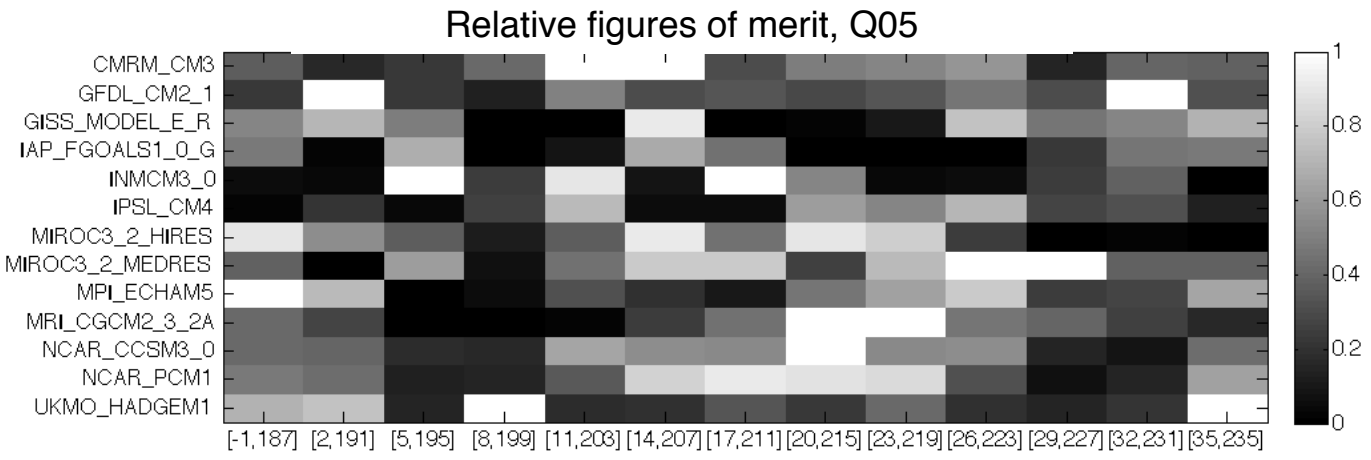
Blocklength selection

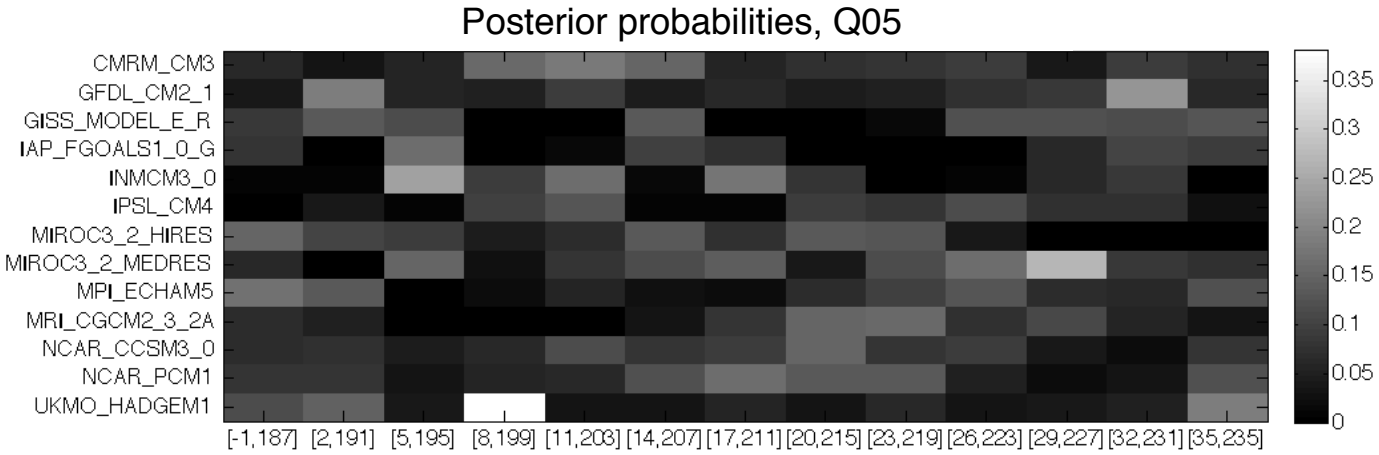


Blocklength selection

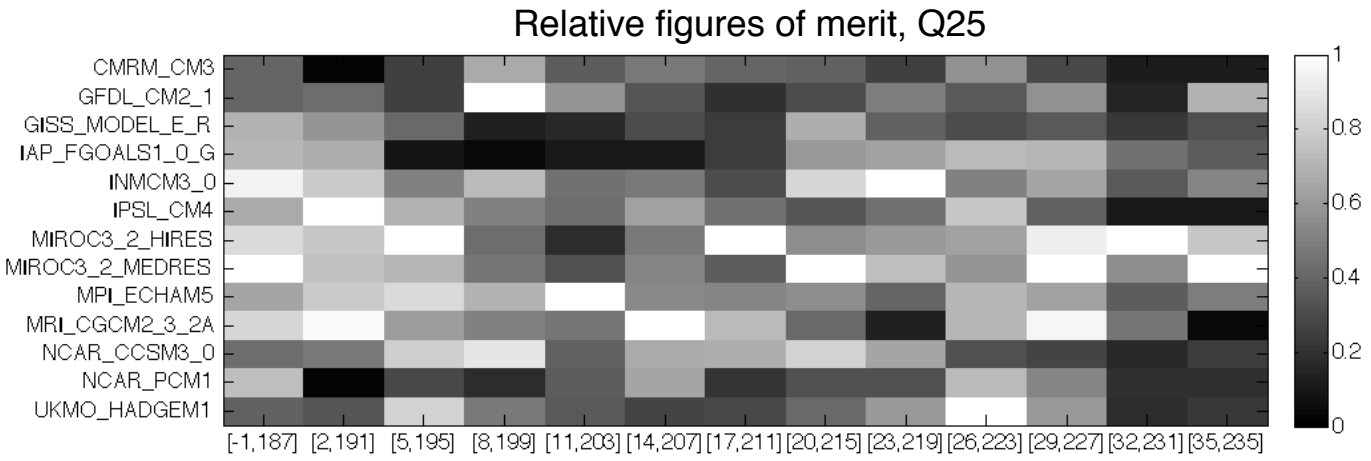


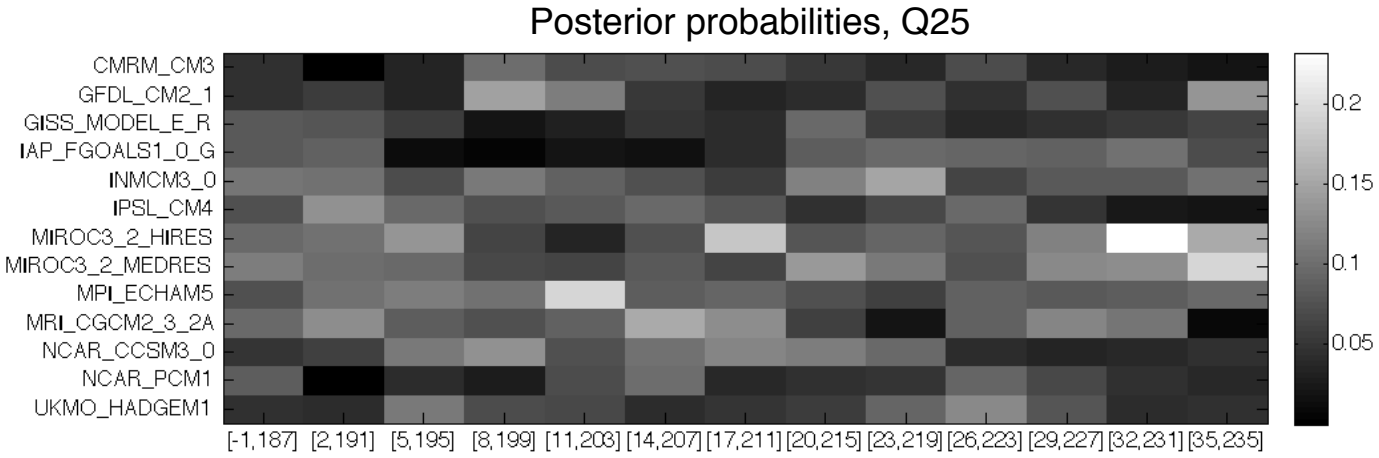
AIRS and CMIP3 models



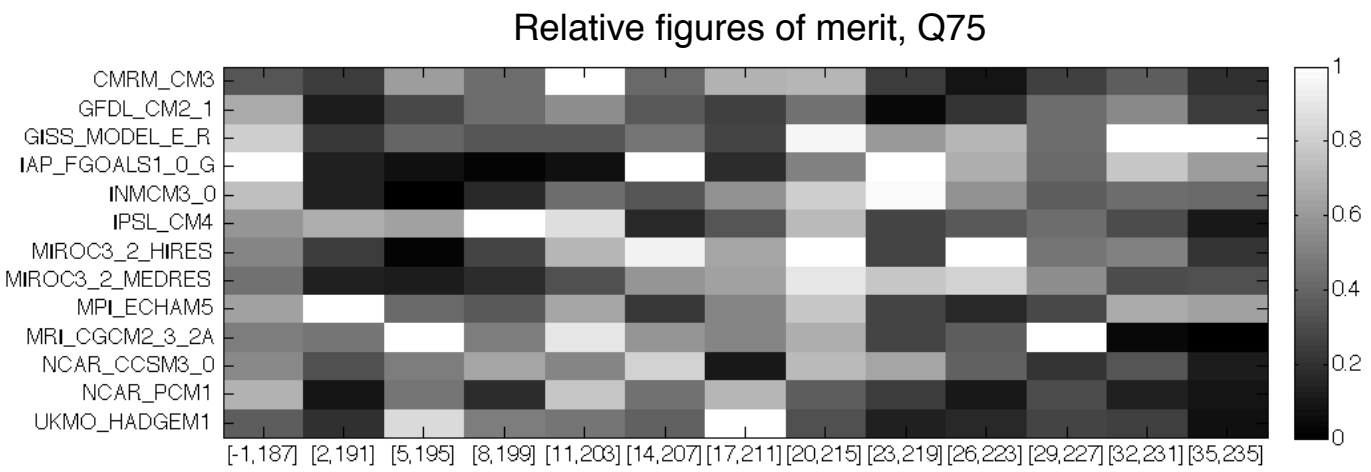


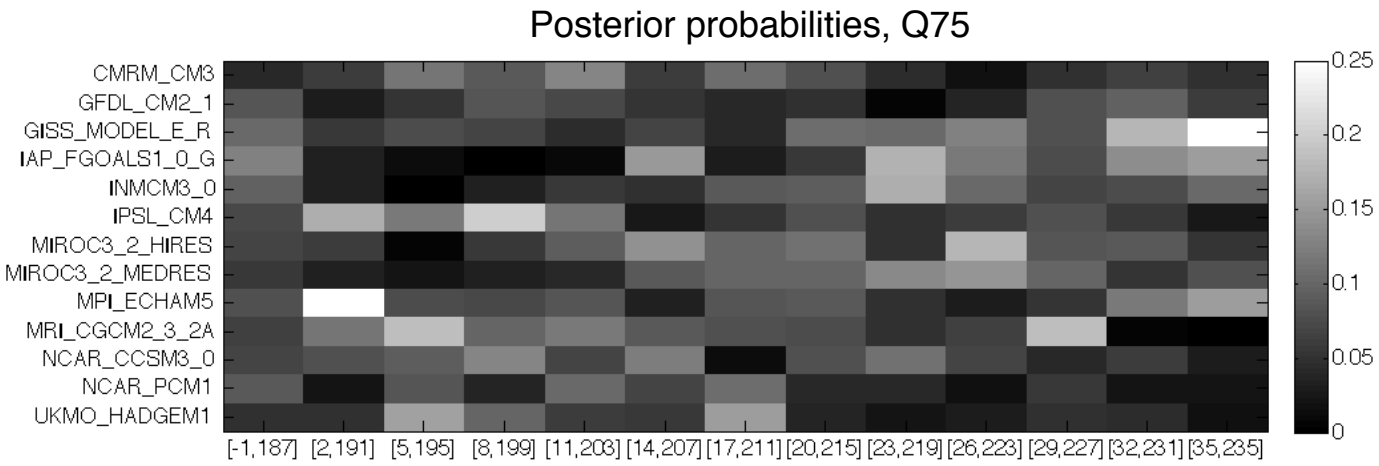
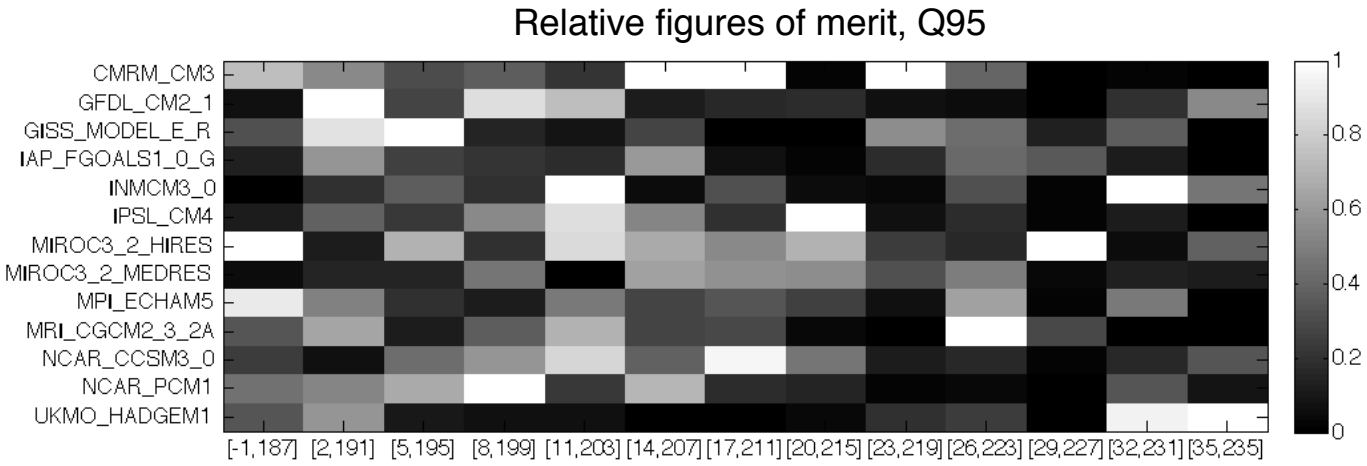
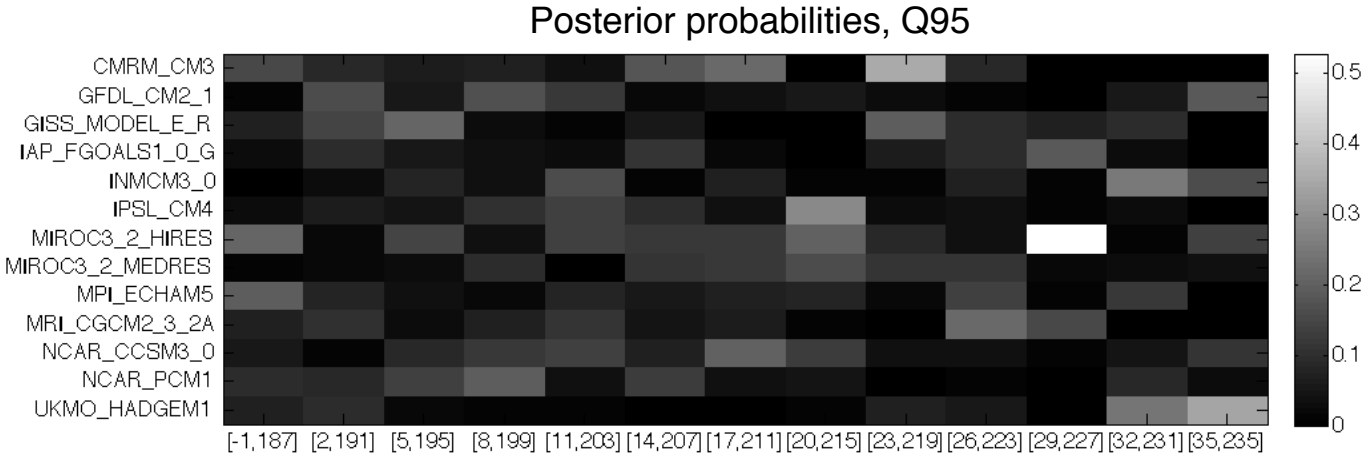
AIRS and CMIP3 models





AIRS and CMIP3 models





AIRS and CMIP3 models