Supporting Information

Pierce et al. 10.1073/pnas.0900094106

SI Text

1. Metrics. Let the model pattern of interest be $m(\mathbf{x})$ and the corresponding observed pattern be $o(\mathbf{x})$. The mean squared error (MSE) in $m(\mathbf{x})$ is defined as

$$\mathrm{MSE}(m, o) = \frac{1}{N}\sum_{k=1}^{N}(m_k - o_k)^2 \qquad \textbf{[S1]}$$

where there are N spatial points. We transform this performance measure to a (dimensionless) spatial skill score (SS) by normalizing:

$$\mathrm{SS}=1-\frac{\mathrm{MSE}(m, o)}{\mathrm{MSE}(\bar{o}, o)}. \qquad \textbf{[S2]}$$

A model field identical to observations has a skill score of 1. Our metrics are based on variables with different units; using this skill score allows us to compare them. We normalize by $\mathrm{MSE}(\bar{o}, o)$, where the overbar indicates the spatial mean; as a result, a global model that predicts the correct mean in a limited region, but only as a completely featureless, uniform pattern, yields a spatial skill score of 0. Temporal variability is evaluated by using spatial patterns of temporal behavior.

The skill score can be decomposed as (1):

$$\mathrm{SS} = r_{m,o}^2 - [r_{m,o} - (s_m/s_o)]^2 - [(\bar{m} - \bar{o})/s_o]^2 \qquad \textbf{[S3]}$$

where $r_{m,o}$ is the product moment spatial correlation coefficient between the model and observations, and s_m and s_o indicate the sample standard deviation of the model and observations, respectively. The first term on the right hand side (RHS) of Eq. **S3** is just the square of the pattern correlation between the model and observations. The second term is the "conditional bias," and expresses the degree to which a spatial regression between the model and observed patterns has a slope that differs from unity (1). The third term is the "unconditional bias," and proportional to the square of the mean error normalized by the standard deviation of the observations. The sense of the decomposition is such that the SS has a starting value of the correlation squared, with deductions taken for any conditional or unconditional biases.

We form metrics for seasonal December-January-February (DJF), March-April-May (MAM), June-July-August (JJA), and September-October-November (SON) averages of tas and pr, the amplitude and phase of the annual harmonic, and the temporal standard deviation of the seasonal data averaged into 1-, 5-, and 10-year blocks. To avoid any possible influence of an anthropogenic climate signal on the standard deviations, we detrended the time series before computing the standard deviations. This process gives 32 seasonal metrics (4 per season for tas and pr), plus 4 seasonal cycle metrics (phase and amplitude for tas and pr).

The western U.S. is strongly affected by the El Nino/Southern Oscillation (ENSO) and Pacific Decadal Oscillation (PDO) modes of natural climate variability. Model PDO indices were computed as the standardized principal component (PC) of the leading empirical orthogonal functions (EOF) of cold season November through March surface temperature anomalies. To obtain the "PDO pattern" for each model, we correlated the model's PDO index with ts (limited to the freezing point of seawater) at each point over the North Pacific. We did the same for observations. The metric for the PDO pattern was then obtained by applying the method described in section *Metrics* of the main text to the observed and model PDO patterns. Teleconnections between the PDO and tas and pr in our region of interest (the western U.S.) were evaluated by correlating the PDO index to those variables over the western U.S. This process results in 3 metrics for the PDO. The ENSO index was computed as the standardized PC of the leading EOF of ts in a box from 110° E to the coast of South America, 23° S to 23° N. The "ENSO pattern" and western U.S. teleconnection patterns in tas and pr were calculated as done for the PDO, yielding another 3 metrics for ENSO. All told, we used 42 metrics.

2. Construction of Fig. 3. Fig. 3 in the main text shows Δ_{SS} as progressively more realizations from either the same model (blue) or randomly selected different models (red) are added to the ensemble average. The blue whiskers are computed as in Fig. 1. I.e., if a model has 4 realizations available, 3 estimates of Δ_{SS} for $n = 3$ were computed: the average of runs (1, 2, 3), (1, 2, 4), and (2, 3, 4).

The red whiskers are calculated with different models included in the ensemble average, however, the first (and only the first) realization included is always taken from the model indicated in the title. Consider again the case where a model has 4 realizations available and we are estimating Δ_{SS} for $n = 3$. The first model added to the ensemble average is a randomly selected realization from the model indicated in the title. The second model added to the ensemble average is randomly chosen from the available models, but must be different from the model indicated in the title. A random realization is chosen to use from this model. The third model chosen is randomly selected from the available models, subject to the constraint that it be different from both the first and second models. And again, a random realization from this model is selected to be used. This procedure is repeated 500 times for each N and the results used to construct the red whisker.

3. Subsets of metrics. The model metrics were constructed without any attempt to avoid over-counting similar aspects of model performance. For example, consider DJF pentadal standard deviation of tas, DJF decadal standard deviation of tas, and JJA mean pr. One would imagine that the first 2 of these are considerably more similar to each other than to the last one. If so, Δ_{SS} would be overly influenced by DJF tas variability.

One way to address this problem is to construct EOFs of the skill score array shown in Fig. S2. Retaining only the leading modes accounts for most of the variance between the model metrics, while reducing the number of model skill measures from the 42 original, co-varying metrics to a few retained (and orthogonal) EOFs.

A drawback of this approach is that EOF-based techniques work on anomalies. Up to now, our evaluation has been in terms of absolute model errors with respect to the observations. Switching to an anomalous analysis means that results will instead be relative to the average model error. In other words, the EOFs will give the most compact and orthogonal set of metrics for differentiating the models from each other, rather than for describing absolute model skill.

As illustrated in Fig. S7 for an idealized case with only 2 metrics, the total model error is the multimodel mean error (shown by the "X") plus deviations from the mean as described in direction by the EOFs and in magnitude by the associated PCs. Worst case, if an EOF (plus mean) is perpendicular to the direction toward the perfect skill point (e.g., EOF "A" in Fig. S7), that mode would say little about model skill in simulating the observations. In fact, the best model skill would be at PC = 0, and both positive and negative PC values would indicate worse models. In general, if the angle between the EOF and the perfect skill point is not 90 ° (e.g., EOF "B" in Fig. S7), travel along the direction indicated by the EOF will initially result in travel toward the perfect skill point, indicating increasing model skill. It is also possible that continued travel in this direction will result in approaching the perfect skill point as closely as possible, and then in travel away from the perfect skill point. In theory then, no simple mapping exists between results of model skill from an EOF analysis (or any analysis based on relative model errors) and absolute model skill.

The EOFs and associated PCs of our metrics array are shown in Fig. S8. The *Top Left* shows the mean model skill score. Two particular problem areas across the models are the representation of low-frequency temperature variability in spring and the seasonal amplitude of precipitation. As expected, the EOFs reflect these mean errors. The eigenspectrum is shown in the *Bottom Left*, along with the sampling uncertainties (2). Asterisks denote modes that are nondegenerate with the subsequent mode. The first and second modes are distinct, whereas the third, fourth, and fifth modes are degenerate with each other, but separate from the noise tail. Between them, the first 5 modes capture just under 90% of the variance (Fig. S8, *Bottom Right*), consistent with the suggestion that 42 metrics overstates the number of independent measures of model quality.

The leading EOF shows greatest expression in the springtime temperature variability (particularly on the annual and 5-year timescales) and amplitude of the seasonal cycle of precipitation, which suggests that those problems may be causally linked in the models. As noted above, one of the potential problems with the EOF analysis is that the scatter of model errors might be perpendicular (in metrics space) to the direction toward perfect skill. This angle is shown in the title of each EOF. For the first EOF, the angle is nearly 45 °, which indicates model differences contribute significantly toward movement toward or away from perfect skill. The associated PC (Fig. S8 *Right*) shows many models do well on this skill measure, but there are 5 or so models with very poor performance that cause this mode to have the largest variance. The gray contours shown on the PC plot indicate distance to the perfect skill point. Between the best and worst models, the distance varies by 18 units (nondimensional because this is a distance in metrics space), much more than any of the higher modes.

The worst-simulated metric, low-frequency (10-year averaged) springtime temperature variability, is described by EOF 2 (Fig. S8). However, the angle between this EOF and the direction toward perfect skill is nearly 90 ° (87.2 °), so model differences along this EOF have little effect on overall model skill. The PC plot shows that the best skills are associated with a PC value of nearly 0, and model skill worsens in both directions away from the model mean. The distance varies by less than 3 units between the best and worst models.

Modes 3, 4, and 5 are degenerate with each other, so cannot sensibly be interpreted individually. As a group they involve low-frequency temperature variability throughout the year, the amplitude of the precipitation and phase of the temperature seasonal cycles, and the temperature teleconnection of the PDO. ENSO, other aspects of precipitation, and annual temperature variability have uniformly weak loadings.

Overall, the EOF analysis indicates that 5 independent measures of model quality capture the majority of the differences between models, given the strong covariances between the original 42 metrics. The strongest mode that makes a difference to model quality links annual and 5-year averaged springtime temperature variability to the amplitude of precipitation's seasonal cycle. The least-well simulated metric is 10-yr averaged springtime temperature variability; the EOF analysis shows that the spread of errors across models is perpendicular to the direction of increasing model skill, which suggests that some new physics or formulation will be required to make progress in better simulating this phenomenon rather than simple model tuning.

[1] Murphy AH (1988) Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon Wea Rev* 116:2417–2424.

[2] North GR, Bell TL, Cahalan RF (1982) Sampling errors in the estimation of Empirical Orthogonal Functions. *Mon Wea Rev* 110:699–706.

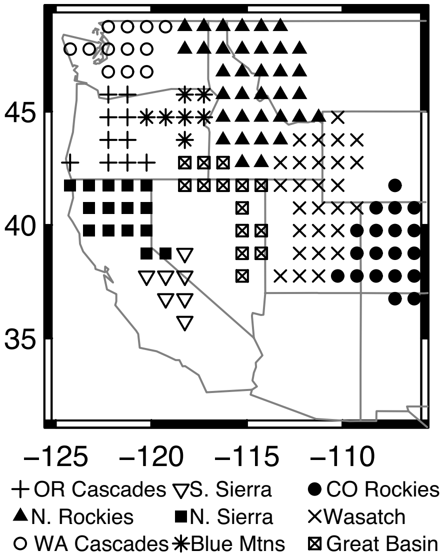

Fig. S1. Our western U.S. domain. Symbols indicate centers of the 1 ° × 1 ° blocks grouped into the 9 mountainous regions analyzed.

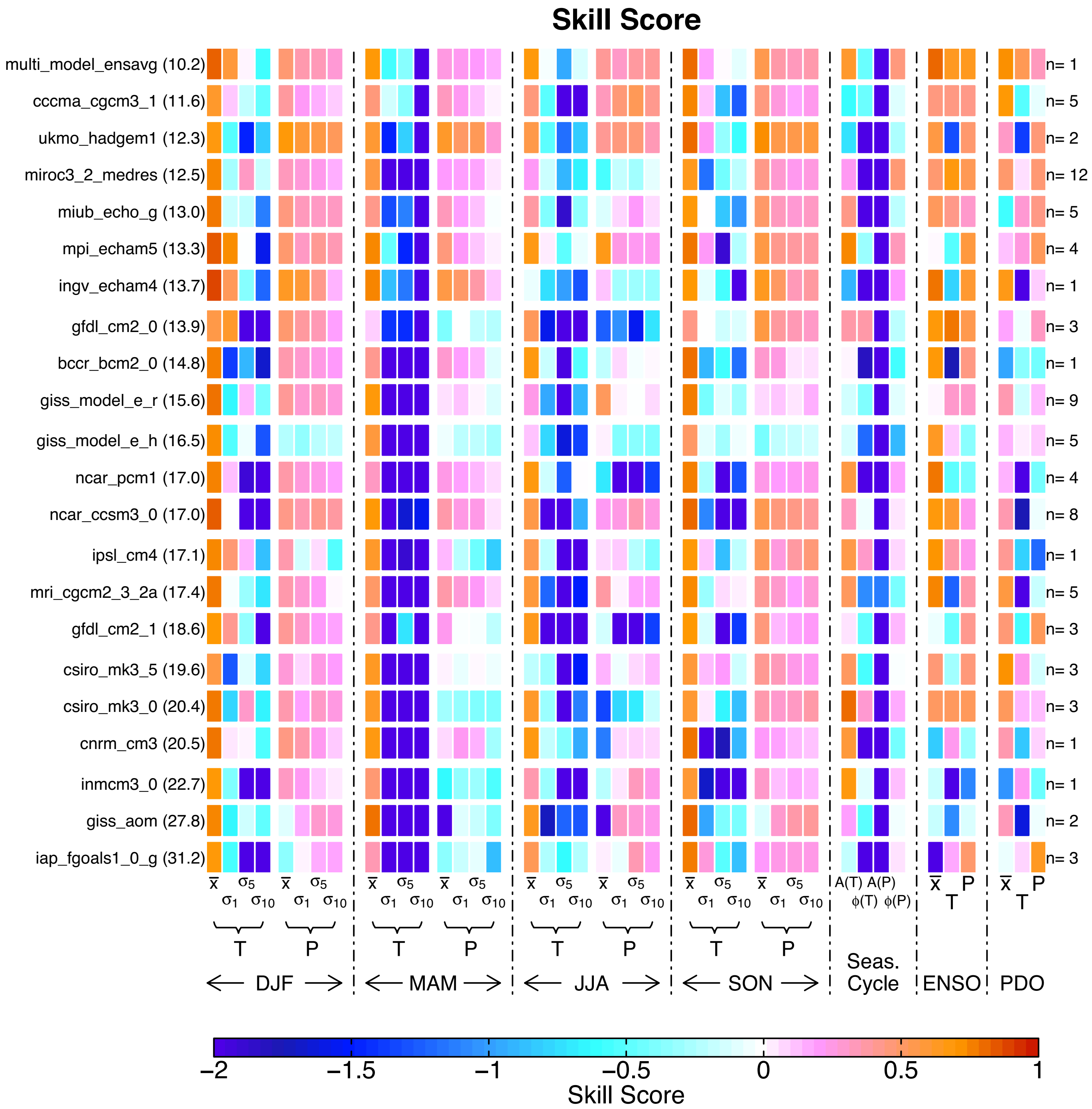

Fig. S2. Portrait plot of model skill scores for each metric. The models are ordered by Δ, the Euclidian distance from perfect skill, point (1, 1, 1, . . . , 1); Δ is noted in the parenthesis (lower values are better). Columns are ordered as described in the text; briefly, the seasonal metrics show skill scores for temperature (T) and precipitation (P), for the climatological mean pattern ($\bar{x}$), and the standard deviations for the data averaged into 1, 5, and 10-year blocks (σ_1, σ_5, and σ_{10}, respectively). The seasonal cycle metrics show the amplitude (A) and phase (ϕ) of the annual harmonic. The ENSO and PDO metrics show the pattern ($\bar{x}$) and the correlation map of the index with temperature (T) and precipitation (P) over the western U.S.

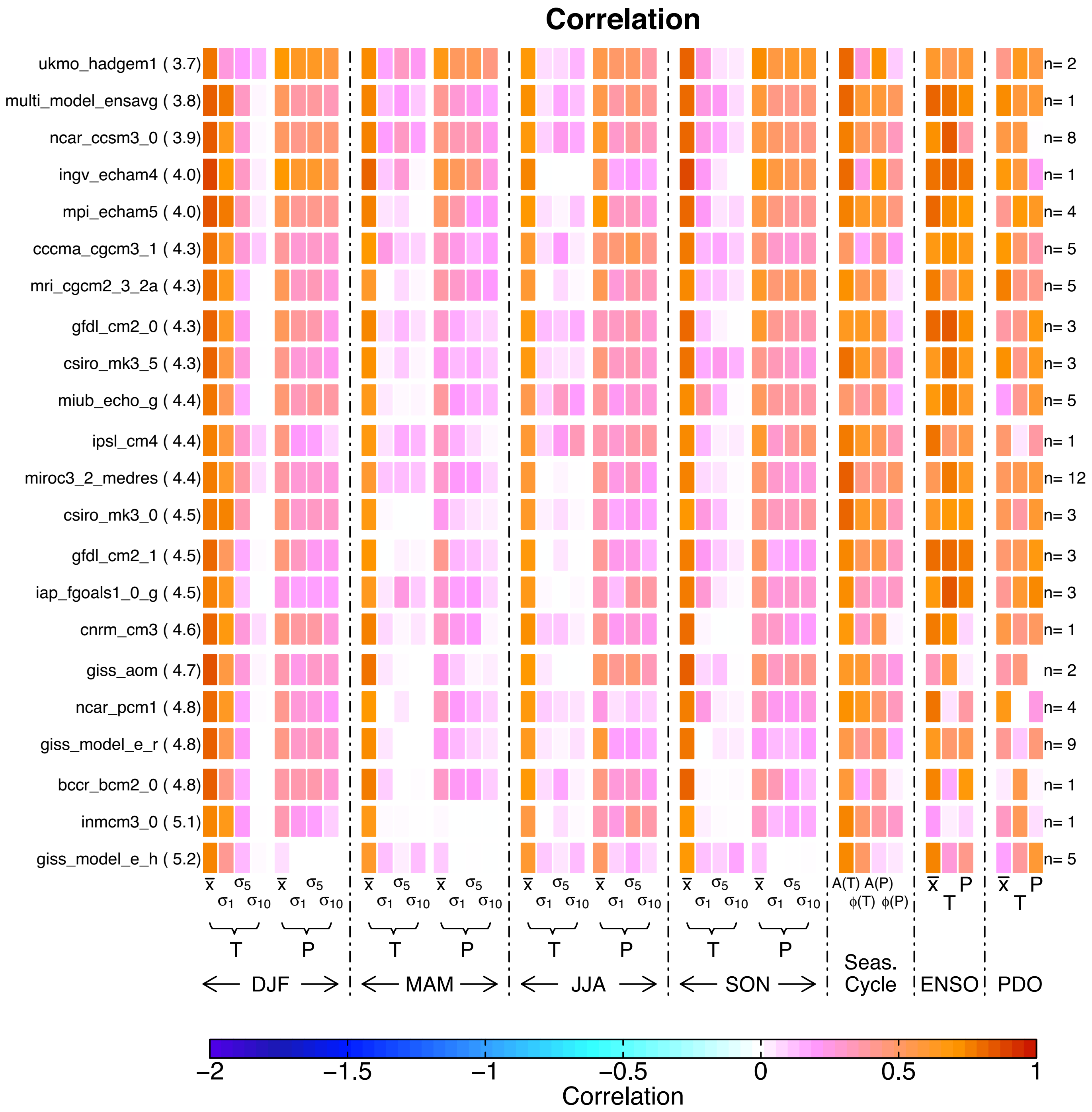

Fig. S3. As in Fig. S2, but for the correlation-squared component of the skill score.

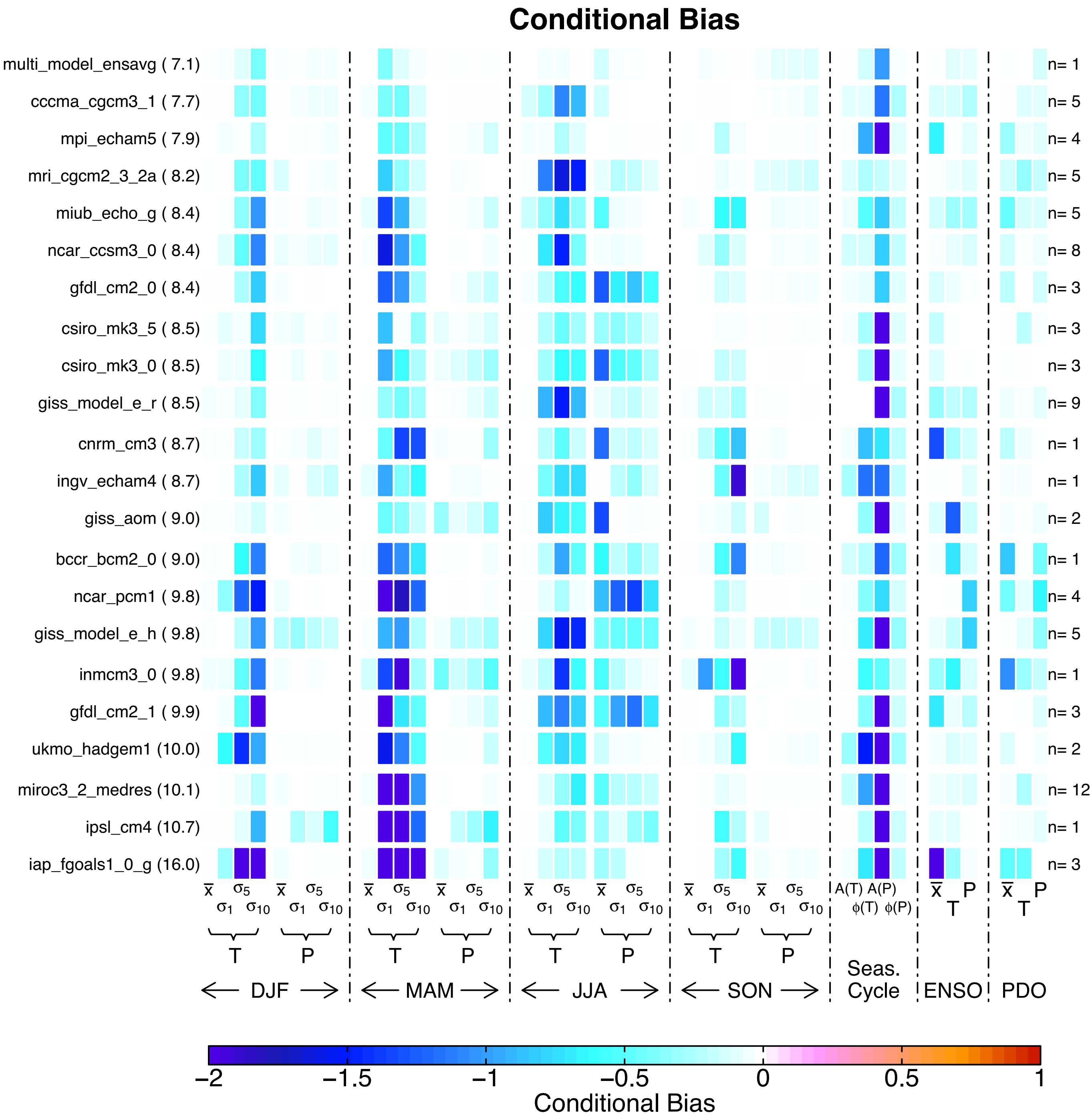

Fig. S4. As in Fig. S2, but for the conditional bias part of the skill score.

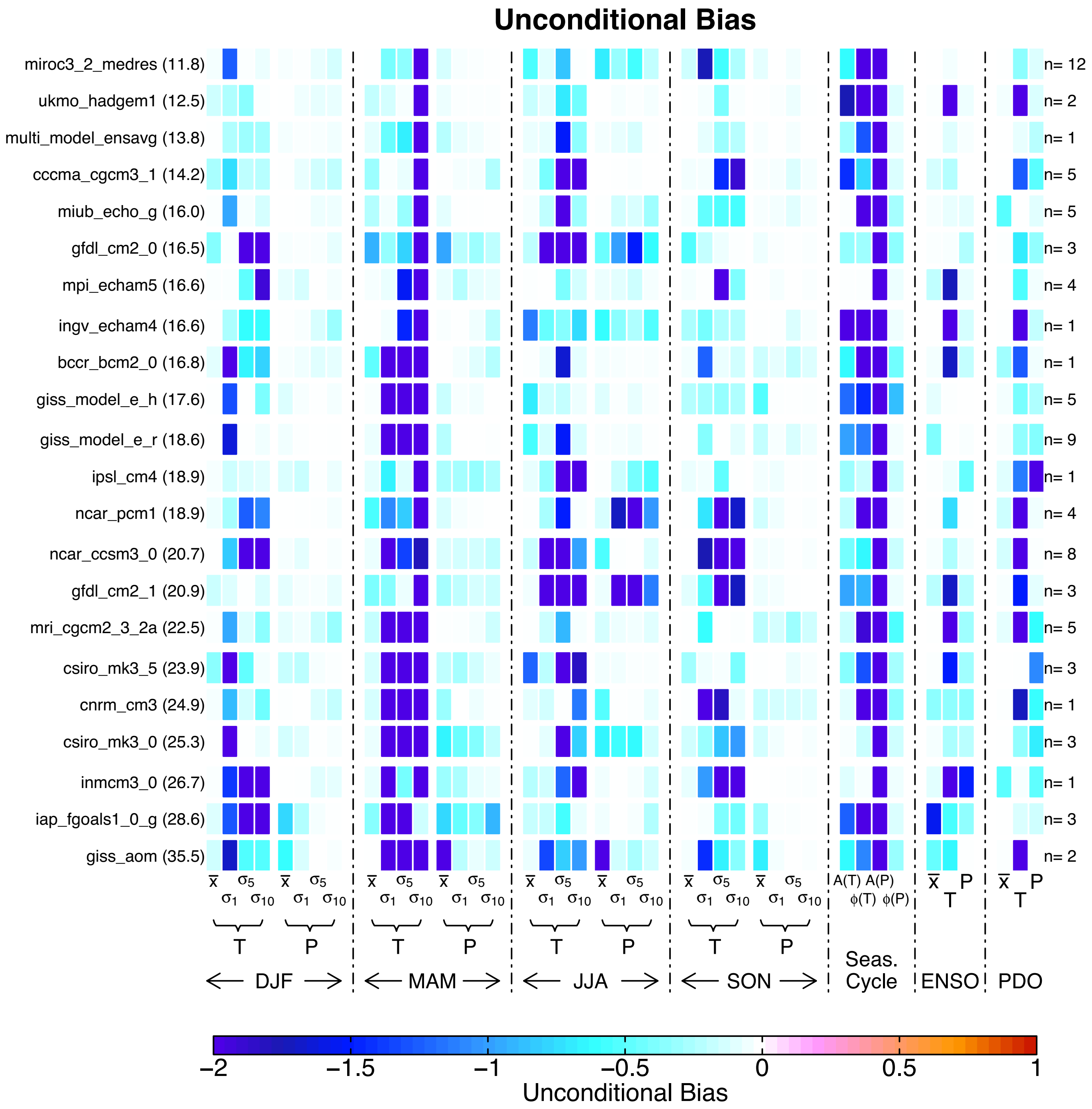

Fig. S5. As in Fig. S2, but for the unconditional bias (mean error) part of the skill score.

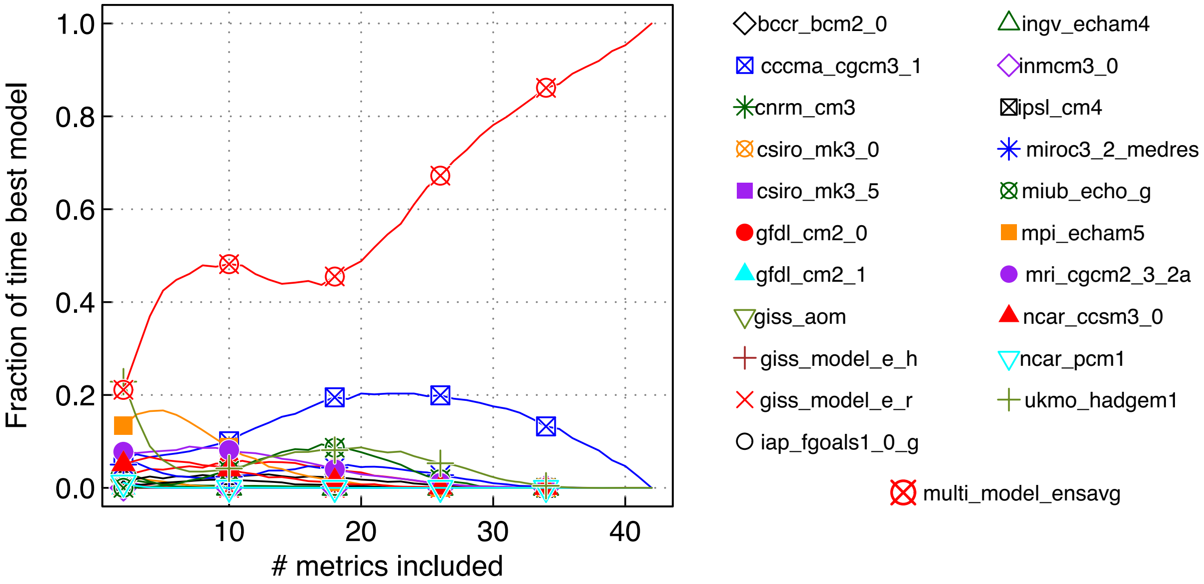

Fig. S6. The fraction of the time the indicated model is the best when different numbers of metrics are included in the evaluation, from $N_{inc} = 2$ to 42 (*x* axis). For each value of N_{inc}, 10,000 random picks of N_{inc} metrics were made, and the models evaluated on those metrics.

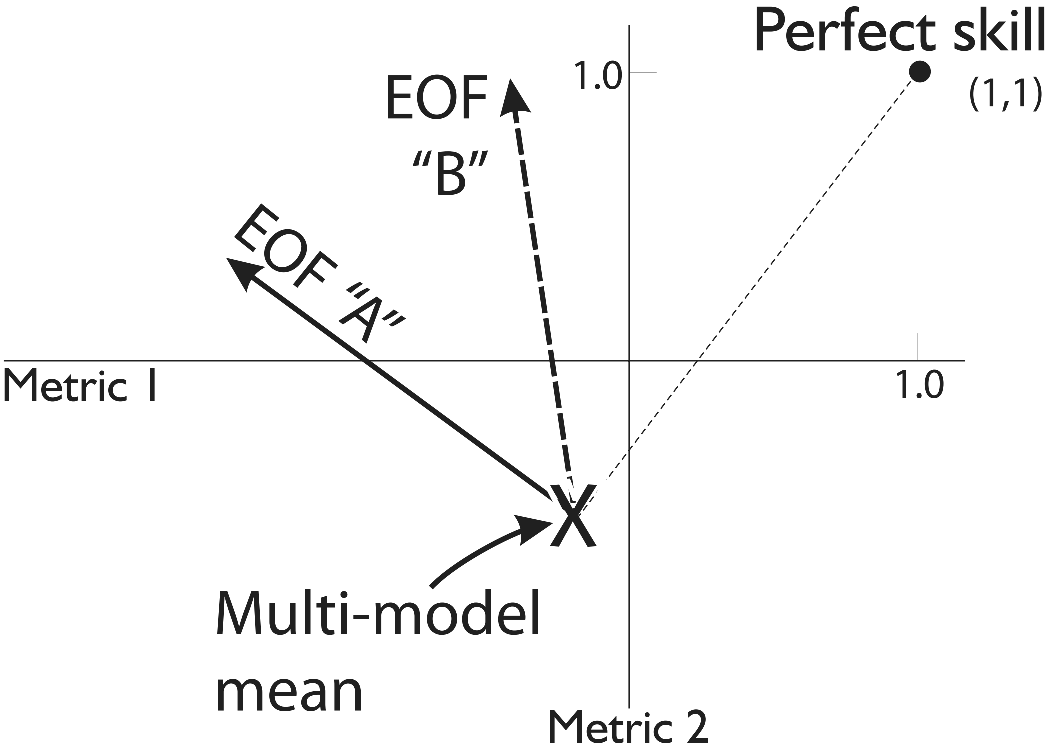

Fig. S7. Schematic illustrating the behavior of the EOFs of the metrics array, for a simplified case with 2 metrics. EOFs are relative to the multimodel mean (indicated by the 'X'). Travel along an EOF that is perpendicular in direction to the perfect skill point (EOF ''A'') has relatively little effect on overall model skill. Moving the same distance along EOF ''B'' has a much larger effect on absolute model skill.

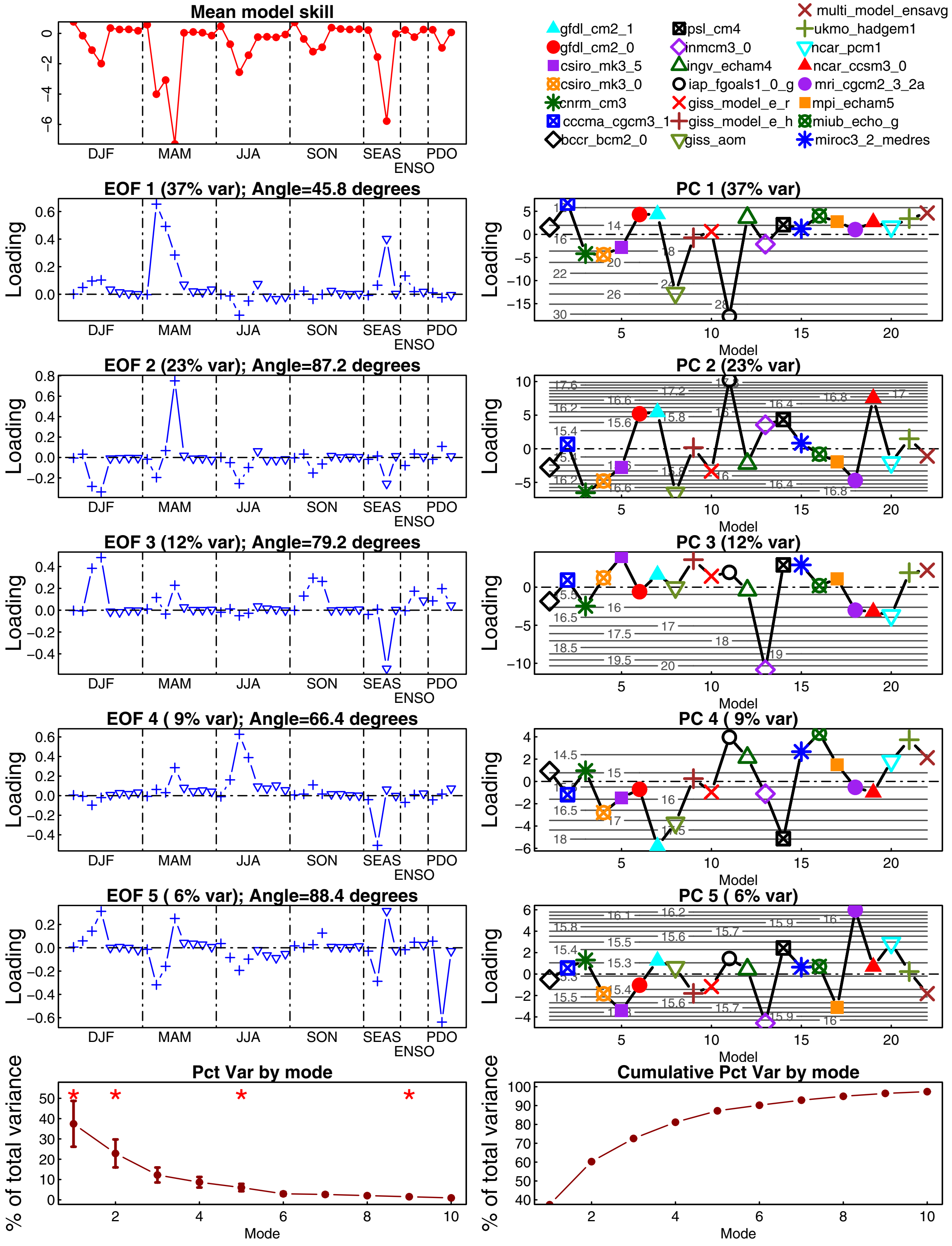

Fig. S8. Leading EOFs (*Left*) and PCs (*Right*) of the model skill score array. EOFs are plotted with the metrics ordered the same way as in Figs. S2–S5, where plus symbols indicate temperature metrics and triangles indicate precipitation metrics. Contours on the PC plots indicate distance from perfect skill. The angle between each EOF and the direction to the perfect skill point is shown in the titles (see also Fig. S7).

Table S1. Models used in this study, a brief indication of their origin (only first institute shown in the case of multiple institutions), the number of ensemble members with ts/pr and tasmin data (respectively), and the ensemble mean JFM tasmin trend in C/year

Name	Origin	No. w/ts & pr	No. w/tasmin	JFM tasmin trend, C/yr
bccr_bcm2_0	Bjerknes Centre Clim. Res., Bergen, Norway	1	1	−0.020
cccma_cgcm3_1	Canadian Centre, Victoria, B.C., Canada	5	5	0.171
cnrm_cm3	Meteo-France, Toulouse, France	1	1	0.124
csiro_mk3_0	CSIRO Atmos. Res., Melbourne, Australia	3	3	0.043
csiro_mk3_5	CSIRO Atmos. Res., Melbourne, Australia	3	3	0.090
gfdl_cm2_0	Geophys. Fluid Dyn. Lab, Princeton, NJ	3	1	−0.051
gfdl_cm2_1	Geophys. Fluid Dyn. Lab, Princeton, NJ	3	3	0.024
giss_aom	NASA/Goddard Inst. Space Studies, NY	2	2	0.036
giss_model_e_h	NASA/Goddard Inst. Space Studies, NY	5	1	0.037
giss_model_e_r	NASA/Goddard Inst. Space Studies, NY	9	1	−0.013
iap_fgoals1_0_g	Inst. Atmos. Physics, Beijing, People's Republic China	3	3	0.103
ingv_echam4	Inst. Geophys. Volcanol., Bologna, Italy	1	1	−0.041
inmcm3_0	Inst. Num. Mathematics, Moscow, Russia	1	1	0.213
ipsl_cm4	Inst. Pierre Simon Laplace, Paris, France	2	2	0.039
miroc3_2_medres	Center Climate Sys. Res., Tokyo, Japan	12	12	0.036
miub_echo_g	Meteor. Inst. U. Bonn, Bonn, Germany	5	3	0.023
mpi_echam5	Max Planck Inst. Meteor., Hamburg, Germany	4	2	0.086
mri_cgcm2_3_2a	Meteor. Res. Inst., Tsukuba, Ibaraki, Japan	5	5	0.008
ncar_ccsm3_0	Nat. Center Atmos. Res., Boulder, CO	8	6	0.070
ncar_pcm1	Nat. Center Atmos. Res., Boulder, CO	4	6	0.068
ukmo_hadgem1	UK Met Office, Exeter, Devon, UK	2	1	−0.009