

1 **Climate model independence and agreement**

J. D. Annan,¹

J. D. Annan, Research Institute for Global Change, 3073-25 Showamachi, Yokohama, Japan.
(jdannan@jamstec.go.jp)

¹Research Institute for Global Change,
3173-25 Showamachi, Yokohama, Japan.

2 How to use the multi-model ensemble generated through the CMIP3 project
3 remains an area of active research. One major question of how to interpret
4 this ensemble has centred on whether the models can be regarded as ‘inde-
5 pendent’, this being a critical factor in determining the robustness of the multi-
6 model consensus. Despite the attention that has been given to this question
7 of independence, the literature is often vague and inconsistent regarding the
8 meaning of this term. Here we demonstrate some problems with previous in-
9 terpretations of this term, and present a new analysis based on the statisti-
10 cal definition. Using this approach, we demonstrate and quantify the depen-
11 dence of models which share similar origins, thus estimating the effective sam-
12 ple size of the 25-member CMIP3 ensemble to be around 20 models. We fur-
13 ther provide a framework for evaluating the robustness and reliability of the
14 ensemble of climate models.

1. Introduction

15 The question of how to interpret multi model data sets such as outputs from the World
16 Climate Research Programme's Coupled Model Intercomparison Project phase 3 (CMIP3)
17 has attracted much debate over recent years. One area of particular interest is whether
18 agreement among models regarding the projections of climate change under anticipated
19 anthropogenic forcing, is a robust indication of confidence in the model outputs, or al-
20 ternatively just the result of models sharing similar faults. The discussion is frequently
21 couched in terms of model 'independence', with this property being considered a desir-
22 able and perhaps necessary condition for robustness. Some authors have argued that the
23 models are not independent, which if true, would cast doubt on the value of the ensem-
24 ble. However, the concept of independence has been used in an unclear and inconsistent
25 manner.

26 In this paper, we reconsider the question of independence and how it relates to the
27 interpretation and robustness of the model ensemble. We begin in Section 2 by identifying
28 some problems with how independence has been defined in the recent literature, and
29 present an alternative approach (based on the standard statistical definition) which allows
30 us to quantify the degree of dependence among related models, and thus calculate the
31 effective ensemble size. We then consider in more detail the concept of the robustness of
32 ensemble agreement in Section 3. Our results are summarised in Section 4.

2. Independence

2.1. How has independence been defined in climate science?

33 The question of model independence has its origins in the interpretation of agreement
34 across the model ensemble. We may hope that when a large number of models agree on a
35 particular phenomenon (such as the large-scale warming due to anticipated greenhouse gas
36 emissions) that this is a more secure indication that such an outcome is highly probable,
37 compared to the situation where we only have one or two models which indicate this
38 result. However, it is also clear that if we were to repeat the same numerical experiment
39 numerous times with the same deterministic model, the results would always be identical,
40 and therefore these extra model integrations would provide no additional evidence beyond
41 that of the first. Thus, the question is often formulated in terms of how many truly
42 ‘independent’ models the ensemble really contains. However, independence has frequently
43 not been clearly defined, and the term has been used in a number of quantitatively vague
44 and inconsistent ways. For example, *Abramowitz and Gupta* [2008] proposed to measure
45 the degree of independence according to the dissimilarity of outputs for the same inputs.
46 Such a definition would appear to rule out *a priori* any use of model consensus as an
47 indicator of increased confidence, since it would instead be interpreted as meaning that
48 the models were dependent. *Räisänen* [2006] and *Pirtle et al.* [2010] consider models to
49 be dependent to the extent that they share underlying ideas and assumptions. However,
50 it is surely appropriate that models should share ideas which are firmly established in
51 theory and practice, such as the conservation laws, and the use of the Navier-Stokes
52 equations on a rotating sphere as the appropriate method for representing the dynamics
53 of the atmosphere and ocean. Conversely, methods for parameterising convection and
54 cloud behaviour vary widely across the ensemble, because there is no consensus regarding

55 the best approach. The models may thus be considered to form a sample of the range
56 of beliefs of the modellers regarding the best way of representing the climate system,
57 within the bounds of current knowledge and technology [*Annan and Hargreaves, 2010a;*
58 *Hargreaves, 2010*]. The above examples suggest that for each component of the models,
59 the range of assumptions adopted across the ensemble cannot be judged in isolation but
60 must instead be considered in relation to model errors and/or our own uncertainty as to
61 how that component can reasonably be represented. We will consider this point in more
62 detail in Section 3.

63 Yet another contrasting perspective on independence is presented by *Tebaldi and Knutti*
64 [2007], who equate this notion with that of the ensemble being centred on the truth.
65 This enables quantitative testing of model independence though examining the pairwise
66 correlations of model errors. However, the underlying premise of a truth-centred ensemble
67 is strongly refuted by analysis of observational data [*Knutti et al., 2010; Annan and*
68 *Hargreaves, 2010a*], so the validity of these calculations is doubtful, as we discuss further
69 in Section 2.2. Thus, it appears that the concept of independence has not yet been
70 addressed in a clear and useful manner.

2.2. Quantifying independence of climate models

71 While usage in climate science has been rather vague, the concept of independence
72 in statistics is straightforward and well-understood. Two or more random events are
73 independent if each one's probability of occurrence does not depend on whether the other
74 event occurred or not, so that their joint probability distribution is given by the product
75 of their marginal distributions [*Wilks, 1995, Section 2.4.3*]. This concept can be directly

76 related to the effective sample size. For example, when n independent samples are drawn
77 from a distribution of standard deviation σ but unknown mean, then the sample mean
78 provides an estimate the true mean with an uncertainty given by σ/\sqrt{n} . If, instead
79 of being independent, the samples are serially correlated with correlation coefficient ρ ,
80 then the error of the sample mean is given by $\sigma/\sqrt{n'}$ where $n' \simeq n \frac{1-\rho}{1+\rho}$ is the number of
81 effectively independent samples [Wilks, 1995, Eqn 5.12].

82 If random variables are independent, then their expected covariance $Cov(x_1, x_2) =$
83 $E(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)$ will be zero. While the converse does not always hold, it does in the
84 particular case of multivariate Gaussian distributions (and some other cases). Therefore,
85 calculations based on this property (or the near-equivalent use of correlations in place of
86 covariances) are a natural first step towards testing and quantifying independence and have
87 been presented by Jun *et al.* [2008] and Knutti *et al.* [2010]. However, those calculations
88 used observational data as the estimate of the mean of the sampling distribution, on the
89 assumption that the models were drawn from some distribution centred on the truth. The
90 resulting strongly positive average correlations have been interpreted as implying that the
91 models are not independent. An alternative and perhaps more obvious interpretation
92 would simply be that the mean of the sampling distribution does not coincide with the
93 truth.

94 It has recently been argued that, as an alternative to assuming a truth-centred distribu-
95 tion, the ensemble can be more naturally interpreted through the paradigm of statistical
96 indistinguishability, that is to say that the truth can be treated as a sample from the same
97 distribution as the models [Annan and Hargreaves, 2010a]. Under this interpretation, we

98 have no direct knowledge of the parameters of the sampling distribution of the models,
99 beyond that provided by the sample itself, and therefore the only available estimate of the
100 mean of the distribution is that provided by the sample mean. The expected covariance
101 of two random samples picked independently from this distribution will automatically be
102 zero, and thus the mean correlation over all model pairs cannot be used as a meaningful
103 test of independence. We can, however, still perform some interesting diagnoses using the
104 pairwise correlations. In particular, we now examine the spread of pairwise correlations,
105 and also test whether specific subsets of models exhibit any dependency.

106 We use all 25 models for which 20C3M simulations are available on the PCMDI
107 database. As in the earlier work of *Annan and Hargreaves* [2010a] (who did not in-
108 clude CSIRO-Mk3.5 in their 24-member ensemble), we consider global climatologies of
109 modelled surface air temperature, precipitation and sea level pressure (SAT, PPT, and
110 SLP respectively), and regrid these data to common regular 5 degree grids to enable cal-
111 culation of pairwise correlations of the model anomalies relative to the sample mean. The
112 CMIP3 ensemble contains some models which have very closely linked pedigrees. There
113 are two pairs of models that essentially differ only in resolution (and perhaps also in some
114 associated resolution-dependent parameters such as gravity-wave drag). These are the
115 Canadian Centre for Climate Modelling and Analysis CGCM at T42 and T63 resolution,
116 and MIROC3.2 at T42 and T106. In addition to these two pairs, there are several in-
117 stitutes that have submitted models with different version numbers or names, which one
118 might still reasonably expect to have more in common than two randomly-selected models
119 from different institutes. These are the three models from GISS, and two models each

120 from CSIRO, GFDL, NCAR and the Hadley Centre. In calculating the sample mean,
121 only the first-listed model from each of the 17 distinct modelling groups was used in the
122 results presented here, so as to avoid the risk of biasing the mean towards models which
123 we expect may be dependent. This decision does not materially affect the results.

124 The results are presented in Table 1 and Figure 1. As explained above, the mean of the
125 pairwise correlations over the full ensemble is close to zero by construction, so the table
126 only provides the standard deviation of the pairwise correlations to indicate the typical
127 magnitude of the correlations. *Annan and Hargreaves* [2010b] used an EOF analysis of
128 model anomalies to estimate the effective dimensionality of the climatological fields of the
129 CMIP3 ensemble to be approximately 6, 11 and 4 for the three data types respectively.
130 The histograms obtained from the pairwise correlations of the CMIP3 models are com-
131 pared in Figure 1 to synthetic results generated by randomly sampling pairs of artificial
132 data points from distributions with these effective dimensions. The correspondence of
133 the distributions obtained by both models and theory is consistent with the models be-
134 ing broadly independent. In particular, the generally slightly narrower results obtained
135 for the model data argues against there being widespread dependencies (and thus high
136 correlations) among subgroups of similar models, other than the small number of cases
137 we describe below. It should be noted, however, that this may be a rather weak test of
138 independence since using the sample mean as the mean of the distribution will tend to
139 minimise the correlations. The estimate of effective dimension is also rather uncertain,
140 and this strongly influences the idealised results, as can be seen in the differences between
141 the results for the three data fields.

142 For the pairs of models where we have prior knowledge of a relationship, the mean
143 pairwise correlations are strongly positive, indicating dependence. In contrast to roughly
144 half of the random pairs exhibiting a negative correlation, only one one of these special
145 pairs exhibits a negative correlation, and that only for one climatic field. For the models
146 which differ only in resolution, the correlations are very high indeed, frequently exceeding
147 that of every other model pair. This should be considered an encouraging sign, as it
148 suggests that the model behaviour is determined largely by its parameterisations and
149 physical approximations rather than by the details of the numerical solution. For model
150 pairs which are related more loosely though their origins at a single institute, there is still
151 a substantial positive correlation on average, but at a markedly lower level.

152 These results can be used to estimate the effective sample size. The comparison with
153 the synthetic data suggests that the original 17 models can be assumed independent. It is
154 straightforward to derive the effective ensemble size of $2/(\rho+1)$ for a pair of models which
155 are correlated at the level ρ , which means that adding a new model which is correlated
156 to one of the existing sample increases the effective ensemble size by $(1-\rho)/(1+\rho)$.
157 According to this expression, creating a new model merely by changing the resolution
158 of an existing one, for which we expect the new model to be correlated to its parent at
159 the $\rho = 0.76$ level, will only increase the effective ensemble size by 0.14. For the typical
160 correlation of 0.42 that arises from successive model versions, the new model increases the
161 effective sample size by 0.41. Thus the effective sample size of this 25 member ensemble
162 can be roughly estimated as $17+6\times 0.41+2\times 0.14 \simeq 20$. Therefore, while the dependency

163 identified here is a significant issue for the specific related models, these form a sufficiently
164 small proportion of the total ensemble that the effective sample size remains substantial.

165 The correlations across three generations of models as presented in the second, third and
166 fourth IPCC assessment reports can also be tested. Model outputs are readily available
167 for 6 institutes or consortia which have reliably provided data to the IPCC process over
168 this interval, these being CCMA (Canada), CCSR/NIES/FRCGC (Japan), CSIRO (Aus-
169 tralia), MPI (Germany), GFDL (USA) and HC (UK), making 18 models in all. Results
170 from the pairwise correlations of two fields of variables (SAT and PPT) are presented in
171 Table 2. The results appear very similar to those for the CMIP3 ensemble. That is, the
172 standard deviation of the pairwise correlations of the full ensemble is similar at 0.31, and
173 the average correlation of pairs of consecutive models from each institute is about the
174 same level as for the institute-related models in the CMIP3 data set, for the equivalent
175 variables. There is also no evidence for the generations themselves being particularly de-
176 pendent, that is to say, the correlations between those models within a single generation
177 are not high compared to those of random pairs.

178 In summary, the study of pairwise correlations may not be a very powerful test of
179 independence., but the result here certainly do not suggest any substantial dependence,
180 other than with a few groups of models which are known to share similar origins. By far
181 the most similar models are those which vary only in resolution, with multiple models
182 originating from the same institute also noticeably similar but to a much lesser degree.
183 Based on these calculations, it is questionable to assign equal weights to all models, since
184 some pairs are closely related through their shared origins. However, so long as these

185 dependent models only form a small proportion of the full ensemble, this is unlikely to
186 have a major effect on the results of ensemble analyses. The effective ensemble size is
187 reduced from 25 to 20 effectively independent models, when this dependency is accounted
188 for.

189 It is, however, not clear how this type of analysis can address the underlying question of
190 whether the consensus of the model ensemble can be trusted, since the use of the sample
191 mean guarantees that the correlations are negligible on average, and the power of the test
192 may therefore be rather limited. In the following section, we consider this problem in
193 more detail.

3. Robustness

194 Although it is not always clearly stated in these terms, many questions concerning
195 the robustness of ensemble agreement can be concisely and quantitatively rephrased in
196 terms of whether the truth lies inside or outside of the ensemble range, and henceforth
197 we explicitly adopt this interpretation. For simplicity, observational errors are ignored,
198 and thus truth and observations are considered synonymous. A useful concept when
199 discussing ensemble robustness is that of the (axis-parallel) ‘bounding box’ [*Smith, 2000*].
200 For a one-dimensional target variable, the bounding box is just the range spanned by the
201 ensemble members. In higher dimensions, this generalises to the hypercuboid (in the state
202 space) with pairs of opposite faces defined by the maximum and minimum values of the
203 ensemble for each variable of interest. Thus, the ensemble can be said to provide a robust
204 prediction when its bounding box contains the truth. *Weisheimer et al.* [2005] present

205 some idealised calculations outlining how the performance of bounding boxes depends on
206 the size and sampling characteristics of the ensemble.

207 In the case of a statistically indistinguishable ensemble, the natural probabilistic in-
208 terpretation of the ensemble through relative frequencies is ‘reliable’ [*Toth et al.*, 2003;
209 *Annan and Hargreaves*, 2010a] and it is not without reason that an ensemble system with
210 this property is frequently referred to as ‘perfect’ [eg *Toth et al.*, 2003]. For a perfect,
211 statistically indistinguishable ensemble of size n , the probability of a single observation
212 lying in the one-dimensional bounding box is $(n - 1)/(n + 1)$. For even a modest number
213 of observations, it is therefore likely that some will lie outside of their one-dimensional
214 bounding boxes. For example, a perfect ensemble of 19 members, which will bound any
215 single observation with 90% probability, will probably fail to bound at least one of seven
216 independent observations. Increasing the ensemble size to 99 improves the probability
217 of bounding single observations to 98%, but such an ensemble will still probably fail to
218 bound one or more of 35 independent observations. It is trivially true that the probability
219 of bounding can be increased by artificially inflating the ensemble width. However, in that
220 case the frequency-based probabilistic interpretation of the ensemble would no longer be
221 reliable. Therefore, in any real application where ensembles are intended to provide a
222 probabilistic representation, and the samples are themselves plausible states, we should
223 not be at all surprised to find substantial numbers of observations falling outside the
224 ensemble range.

225 From the perspective of an observation, all that is required in order for it to be bounded
226 is that at least one of the models lies on either side of it. Broadly speaking, this is achieved

227 when the ensemble spread is at least as large as the error in the ensemble mean. So, as
228 was argued in Section 2.1, we cannot determine the ensemble robustness by examining
229 the ensemble spread or inter-model differences in isolation, but instead we must consider
230 whether these differences are an adequate reflection of the likely error in the ensemble
231 mean. The ensemble will bound the truth with high probability when when our un-
232 certainties (as embodied in the ensemble spread) are well-calibrated, or in other words,
233 commensurate with the error in the ensemble mean. It must be emphasised that all of
234 the uncertainties considered here are purely epistemic in nature, that is they relate to
235 our lack of knowledge, and there is no “true range of uncertainty” associated with the
236 climate system and its future changes [*Hargreaves, 2010*]. It is, of course, legitimate to
237 question whether the current model range reliably covers our range of uncertainty. This
238 can be naturally tested by directly comparing observations to the equivalent outputs from
239 the ensemble of models, as demonstrated by *Annan and Hargreaves [2010a]*. While the
240 reliability of future predictions cannot be directly evaluated in this way, comparisons with
241 a wide range of present and historical observations should be of value in evaluating and
242 calibrating the ensemble and building confidence in the behaviour of the models.

4. Conclusions

243 The concept of ‘independence’, as it has previously been defined with reference to
244 climate modelling, is of limited value in assessing the robustness of ensemble agreement.
245 The use of pairwise correlations of model errors (relative to observations) as a diagnostic is
246 misleading, as the observations do not lie at the ensemble mean and thus these correlations
247 will always be strongly positive on average. Conversely, use of the sample mean ensures

248 that the average correlation will be very small. When using this approach, however, the
249 distribution of correlations is consistent with the models forming a largely independent
250 sample, and this method also reveals and quantifies the strong dependency between models
251 separated only by resolution changes and the weaker dependency between models with
252 shared institutional origins. The effective sample size of the 25-member CMIP3 ensemble
253 is estimated at 20 models, a value which is substantially higher than that previously
254 estimated by *Jun et al.* [2008]. However, this analysis cannot directly inform on the
255 robustness of the ensemble in bounding reality.

256 In order for the ensemble to bound a scalar truth with high probability while remaining
257 informative, it is necessary that the uncertainties embodied in the ensemble are well-
258 calibrated, so that its spread is commensurate with the error in the ensemble mean. If
259 this is the case, then the ensemble will be reliable [*Annan and Hargreaves, 2010a*]. A
260 perfectly reliable ensemble of 20 samples will bound roughly 90% of (scalar) observations,
261 and thus we should not be surprised or disappointed to find a substantial number of
262 observations outside of the ensemble range. Indeed an absence of such would imply that
263 the ensemble spread is unrealistically large. The testing presented here is only a necessary,
264 and not sufficient, demonstration of independence. Therefore, further critical evaluation
265 of ensemble performance is strongly encouraged.

266 **Acknowledgments.** I am grateful to Julia Hargreaves and three anonymous reviewers
267 for many valuable comments and suggestions. This work was supported by the S-5-1
268 project of the MoE, Japan and by the Kakushin Program of MEXT, Japan. I acknowledge
269 the modeling groups, the Program for Climate Model Diagnosis and Intercomparison

270 (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their
271 rôles in making available the WCRP CMIP3 multi-model dataset. Support of this dataset
272 is provided by the Office of Science, U.S. Department of Energy.

References

- 273 Abramowitz, G., and H. Gupta (2008), Toward a model space and model independence
274 metric, *Geophysical Research Letters*, *35*(5), L05,705.
- 275 Annan, J. D., and J. C. Hargreaves (2010a), Reliability of the CMIP3 ensemble, *Geophys-*
276 *ical Research Letters*, *37*(2), L02,703.
- 277 Annan, J. D., and J. C. Hargreaves (2010b), Understanding the CMIP3 multi-model
278 ensemble, *Journal of Climate*, submitted.
- 279 Hargreaves, J. C. (2010), Skill and uncertainty in climate models, *Wiley Interdisciplinary*
280 *Reviews: Climate Change*, DOI: 10.1002/wcc.58.
- 281 Jun, M., R. Knutti, and D. Nychka (2008), Spatial analysis to quantify numerical model
282 bias and dependence: how many climate models are there?, *Journal of the American*
283 *Statistical Association*, *103*(483), 934–947.
- 284 Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl (2010), Challenges in
285 combining projections from multiple climate models, *Journal of Climate*, *23*, 2739–2758.
- 286 Pirtle, Z., R. Meyer, and A. Hamilton (2010), What does it mean when climate models
287 agree? A case for assessing independence among general circulation models, *Environ-*
288 *mental Science & Policy*, *13*(5), 351–361.
- 289 Räisänen, J. (2006), How reliable are climate models?, *Tellus A*, *59*(1), 2–29.

- 290 Smith, L. (2000), *Disentangling uncertainty and error: On the predictability of nonlinear*
291 *systems*, pp. 31–64, Birkhauser, Boston.
- 292 Tebaldi, C., and R. Knutti (2007), The use of the multi-model ensemble in probabilistic
293 climate projections, *Philosophical Transactions of the Royal Society A: Mathematical,*
294 *Physical and Engineering Sciences*, 365(1857), 2053.
- 295 Toth, Z., O. Talagrand, G. Candille, and Y. Zhu (2003), Probability and ensemble fore-
296 casts, *Forecast Verification: A Practitioners Guide in Atmospheric Science*, pp. 137–
297 163.
- 298 Weisheimer, A., L. Smith, and K. Judd (2005), A new view of seasonal forecast skill:
299 bounding boxes from the DEMETER ensemble forecasts, *Tellus. Series A: Dynamic*
300 *Meteorology and Oceanography*, 57(3), 265–279.
- 301 Wilks, D. S. (1995), *Statistical methods in the atmospheric sciences*, Academic Press,
302 London.

Figure 1. Pairwise correlations of CMIP3 model fields of surface air temperature (SAT), precipitation (PPT) and sea level pressure (SLP). Upper plots show correlations of all model pairs (solid) and results from idealised testing described in text (dashed), Lower plots show correlations of the same models run at different resolution (solid) and all model pairs from the same institutes (dashed).

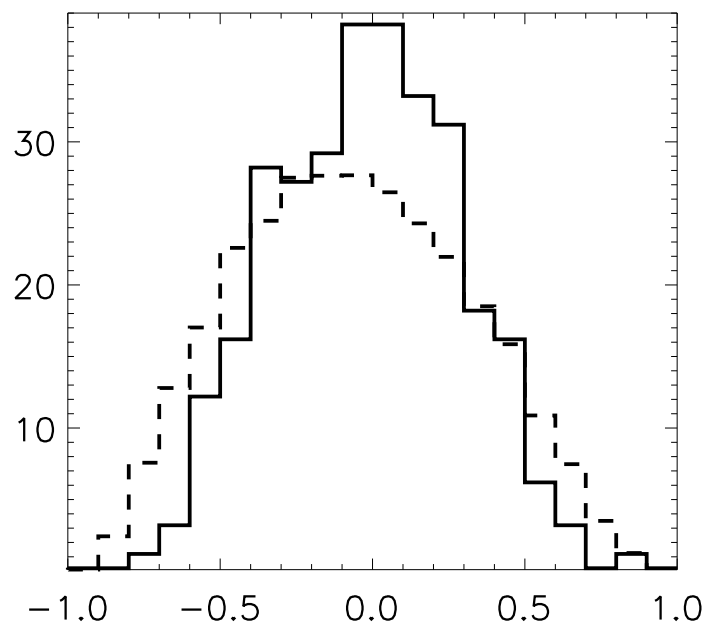
	Idealised SD	CMIP3 SD	Related mean	Resolution mean
SST	0.37	0.30	0.31	0.63
PPT	0.29	0.23	0.36	0.80
SLP	0.45	0.48	0.60	0.85
Average	0.38	0.35	0.42	0.76

Table 1. Analysis of pairwise correlations over various ensembles showing, in order, the standard deviations of the correlations for the idealised example and the full CMIP3 ensemble, the mean correlation for the related model pairs and mean correlation of the models that differ only in resolution.

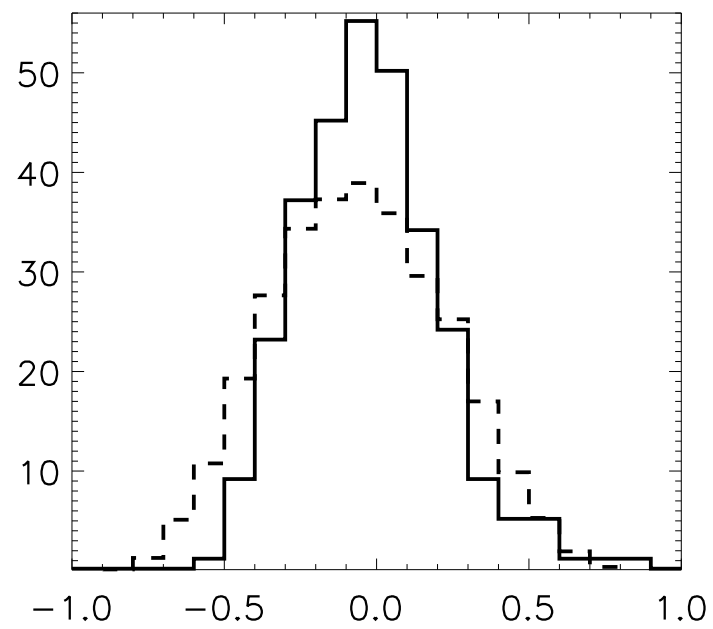
	All SD	Generation mean
SAT	0.31	0.38
PPT	0.30	0.25
Average	0.31	0.32

Table 2. Analysis of pairwise correlations over three generations of climate models. First column shows standard deviation of correlations over all models, second column shows mean correlation between successive models from the same institutes.

SAT



PPT



SLP

