

1 **Reliability of the CMIP3 ensemble**

J. D. Annan,¹ and J. C. Hargreaves¹

J. D. Annan, Research Institute for Global Change, 3073-25 Showamachi, Yokohama, Japan.

(jdannan@jamstec.go.jp)

J. C. Hargreaves, Research Institute for Global Change, 3073-25 Showamachi, Yokohama, Japan. (jules@jamstec.go.jp)

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

2 We consider paradigms for interpretation and analysis of the CMIP3 en-
3 semble of climate model simulations. The dominant paradigm in climate sci-
4 ence, of an ensemble sampled from a distribution centred on the truth, is con-
5 trasted with the paradigm of a statistically indistinguishable ensemble, which
6 has been more commonly adopted in other fields. This latter interpretation
7 (which gives rise to a natural probabilistic interpretation of ensemble out-
8 put) leads to new insights about the evaluation of ensemble performance. Us-
9 ing the well-known rank histogram method of analysis, we find that the CMIP3
10 ensemble generally provides a rather good sample under the statistically in-
11 distinguishable paradigm, although it appears marginally over-dispersive and
12 exhibits some modest biases. These results contrast strongly with the incom-
13 patibility of the ensemble with the truth-centred paradigm. Thus, our anal-
14 ysis provides for the first time a sound theoretical foundation, with empir-
15 ical support, for the probabilistic use of multi-model ensembles in climate
16 research.

1. Introduction

17 The World Climate Research Programme’s Coupled Model Intercomparison Project
18 phase 3 (CMIP3) multi-model dataset of more than 20 global climate models developed
19 around the world has proved to be a valuable resource which has motivated and enabled
20 much research. Since the ensemble was not generated through a coordinated attempt to
21 sample a specific distribution but is instead an “ensemble of opportunity” (that is, a
22 collection of model outputs based solely on availability), there has been extensive discus-
23 sion on how best to interpret its construction and analyse its outputs, in order to make
24 predictions of future climate change [eg *Tebaldi and Knutti*, 2007, and references therein].

25 One paradigm which has formed the basis for many analyses is to consider that the mod-
26 els are “random samples from a distribution of possible models centered around the true
27 climate” [eg *Jun et al.*, 2008; *Tebaldi and Knutti*, 2007]. This truth-centred paradigm
28 appears to have arisen as a post-hoc interpretation of the ad-hoc weighting procedure
29 known as “Reliability Ensemble Averaging” or REA [*Giorgi and Mearns*, 2002; *Nychka*
30 *and Tebaldi*, 2003], and it has since been widely adopted [eg *Tebaldi et al.*, 2005; *Smith*
31 *et al.*, 2009]. When the outputs from the CMIP3 ensemble are analysed, however, they are
32 found to not possess the statistical properties that would be expected of such a sampling
33 distribution. For example, if the models are indeed sampled from a distribution centered
34 on the truth, then the biases of different models should have, on average, near-zero pair-
35 wise correlations. In practice, however, the correlations are for the most part strongly
36 positive [*Jun et al.*, 2008]. An immediate consequence of this is that, as models are added
37 to an ensemble, the multi-model mean does not converge to observations as rapidly as

38 would be expected for independent biases [Knutti *et al.*, 2009]. These disquieting results
39 have lead to claims that the model spread is “likely too narrow” [Knutti *et al.*, 2009] and
40 limit the confidence that we can have in analyses and results that are based on this un-
41 derlying interpretation. Various adaptations and corrections have been proposed to adjust
42 for these problems [eg Jun *et al.*, 2008; Jewson and Hawkins, 2009], but there is no clear
43 consensus on the way forward.

44 In this paper, we reconsider paradigms for ensemble generation and interpretation. In
45 the following section, we contrast the truth-centred paradigm described above, with an
46 alternative interpretation which is well-established in numerical weather prediction and
47 other fields: that of an exchangeable or *statistically indistinguishable* ensemble, that is,
48 one where the truth is drawn from the same distribution as the ensemble members, and
49 thus no statistical test can reliably distinguish one from the other. A simple method
50 of evaluating ensemble performance under this interpretation is presented. We contrast
51 the two paradigms through the analysis of some idealised examples in Section 3. In
52 Section 4 we evaluate outputs from the CMIP3 database in the context of the statistically
53 indistinguishable paradigm. We discuss the implications of our findings and conclude with
54 some suggestions for future research directions.

2. Ensembles and reliability

55 The paradigm of a statistical indistinguishable ensemble has been widely adopted in
56 numerical weather prediction [eg Toth *et al.*, 2003] but less commonly in climate change
57 research [Räisänen and Palmer, 2001]. A fundamental distinction can easily be made be-
58 tween the truth-centred approach described above, and the statistically indistinguishable

59 interpretation; in the latter case the mathematical expectation of the truth is still given
60 by the mean of the sampling distribution, but we no longer expect the truth to be at, or
61 even close to, this location. For example, if the ensemble members and the truth are all
62 drawn independently from the 40-dimensional standard Normal distribution (Z_1, \dots, Z_{40}) ,
63 then even though the expected value of the truth is $(0, \dots, 0)$, the probability of it being
64 found in the bounded box $[-2, 2]^{40}$ is less than 16%, and the probability of it lying in
65 $[-0.5, 0.5]^{40}$ is 2×10^{-17} .

66 There is a natural probabilistic interpretation of such ensembles based on simple count-
67 ing arguments. When a proportion p of the ensemble members have a particular property
68 Q (such as the temperature at a particular time and place exceeding a specified thresh-
69 old) then this is interpreted as assigning probability p to the event Q . Although such
70 a single-event probability cannot be directly validated against the observation, it is, in
71 principle, straightforward to check over a large number of similar forecasts whether events
72 that are predicted to occur with probability p actually do occur on a proportion p of the
73 occasions. A system is called “reliable” if the forecast and observed frequencies do in fact
74 agree [*Murphy*, 1973; *Toth et al.*, 2003].

75 The rank histogram or Talagrand diagram [*Anderson*, 1996; *Talagrand et al.*, 1997] is
76 a common method for evaluating the reliability of ensemble forecasts. It is based on the
77 histogram of the rank of each of the observations in the ordered set formed by the union
78 of the n predictions of the individual ensemble members together with the one observed
79 value. If the truth and ensemble members are drawn from the same distribution, then
80 the ranks of the observations should be uniformly distributed in $\{1, \dots, n + 1\}$. If the

81 ensemble spread is too low, then observations will frequently lie close to or outside the
82 edges of the ensemble, resulting in a u-shaped rank histogram. Conversely, if the ensemble
83 spread is too broad, then the rank histogram will have a central dome. Computing the
84 histogram of the ranks of the observations, and checking for consistency with uniformity,
85 therefore provides a necessary condition for an ensemble prediction system to be reliable.

3. Idealised analysis

86 We now examine the properties of statistically indistinguishable ensembles in more de-
87 tail, in particular investigating how they perform when analysed using methods developed
88 for the truth-centred paradigm.

89 We perform three experiments, in each of which we draw 24 ensemble members indepen-
90 dently from the 40-dimensional standard Normal (Z_1, \dots, Z_{40}) . The truth for the three
91 experiments is sampled from the three distributions $0.5 \times (Z_1, \dots, Z_{40})$, (Z_1, \dots, Z_{40}) and
92 $2 \times (Z_1, \dots, Z_{40})$ respectively. Thus, the second ensemble is statistically indistinguish-
93 able from the truth, and the first and third are sampled from broader, and narrower,
94 distributions, respectively.

95 A simple but intuitively appealing investigation into the correlation of errors across the
96 CMIP3 ensemble has been presented by *Knutti et al.* [2009, Figure 3]. We replicate this
97 method of analysis in the top two rows of Figure 1, and obtain strikingly similar results.
98 That is, (i) the errors of ensemble members are generally positively correlated, (ii) as
99 the sample size increases, the RMSE of the ensemble mean converges to a value that is
100 substantially greater than zero, and (iii) a few “good” ensemble members can be found, the
101 mean of which outperforms the average of the larger set. It therefore appears that these

102 properties are not directly informative regarding the reliability of the ensemble (in the
103 technical sense introduced above). In fact, in the case of the statistically indistinguishable
104 ensemble, the RMSE ensemble mean will typically converge to a value which is only a
105 factor $1/\sqrt{2}$ smaller than the average RMSE of the individual models. However, the
106 precision with which this ratio is attained will depend on the sampling variability not only
107 of the ensemble, but also of the true value itself, which might by chance lie somewhat closer
108 to or further from the ensemble mean, compared to its expected distance. The central
109 plot in Figure 1 provides some indication of these effects, in the discrepancy between the
110 observed (solid red) and theoretically predicted (dashed black) lines. The latter was, in a
111 minor divergence from the analysis of *Knutti et al.* [2009], not fitted to the results of the
112 specific samples but directly calculated from the underlying sampling distributions.

113 In contrast to these analyses, the rank histograms contained in the bottom row of
114 Figure 1 exhibit (in order from left to right) the characteristic domed, flat, and u shapes
115 that we should expect given the sampling distributions, and thus enable us to correctly
116 classify the experiments. The χ^2 statistic based on the contents of the 24 bins is a common
117 statistical test of uniformity, but it provides a rather poor test of reliability since it is
118 insensitive to order and thus does not directly consider issues such as bias or the overall
119 spread of the ensemble. We therefore decompose the χ^2 statistic according to the method
120 proposed by *Jolliffe and Primo* [2008], using components to evaluate both a bias (linear
121 trend across the histogram) and spread (approximated by a v-shape). The contributions
122 of these two components to the total χ^2 statistic are also presented in the Figure. If the
123 rank histogram was generated from a uniform distribution, both of these statistics should

124 be distributed according to the χ^2 distribution with one degree of freedom. Therefore,
125 we correctly find no signs of any significant biases in these experiments, but the spreads
126 of the first and third experiments are both found to be significantly non-uniform at the
127 $p < 1\%$ level ($\chi^2 > 6.64$).

4. Analysis of CMIP3 models

128 We now investigate the reliability of the CMIP3 ensemble using the rank histogram
129 approach. We follow the approach of previous authors in evaluating the mean climatic
130 state against modern observational data. We analyse fields of three climatic variables:
131 surface air temperature, with data obtained from *Brohan et al.* [2006], precipitation,
132 using the data of *Adler et al.* [2003], and sea level pressure against the data of *Allan and*
133 *Ansell* [2006]. All model and observational data sets are firstly averaged onto 5 degree
134 global grids and over the years 1961-1990 (temperature and sea level pressure) or 1979-
135 1999 (precipitation). We only present results from annual mean values here, but using
136 seasonal averages (DJF, JJA) or the magnitude of the seasonal cycle (JJA minus DJF)
137 give broadly similar results.

138 Rank histograms of the three sets of observations are shown in Figure 2. These are
139 calculated on an area-weighted basis, with the totals normalised to 40. Since neighbouring
140 grid points are highly correlated, the number of effective degrees of freedom of the data
141 (which determines the precision with which the empirical rank histograms should match
142 the uniform distribution) is not entirely clear. Semi-variograms of the inter-model (and
143 model minus data) differences suggest a decorrelation distance of around 1000-2000km.
144 This would imply at least 40 degrees of freedom for each data field, which motivates our

145 choice of this value both here and for the idealised tests presented in Section 3. *Jolliffe*
146 *and Primo* [2008] suggest using 25 degrees of freedom for a single hemisphere, which is
147 similar to our choice. Changing the number of degrees of freedom alters the statistical
148 significance of our results, but not their qualitative nature.

149 The total χ^2 statistics of the rank histograms are all insignificant, but the decomposi-
150 tion into bias and spread components does reveal some problems, in that the modelled
151 temperatures are biased low and the spreads of temperature and sea level pressure ap-
152 pear too large, relative to the observations (albeit the error in the spread of temperature
153 does not quite reach the $p < 5\%$ significance threshold of $\chi^2 > 3.84$). We should note,
154 however, that these errors are in fact relatively small compared to the ensemble ranges
155 themselves. The surface temperature histogram can be effectively flattened by both sub-
156 tracting a mean bias of 0.5C, and adding random noise of magnitude 1C to the data to
157 increase their overall spread. These figures only amount to 7% and 13% respectively of
158 the typical ensemble range of 7.7C at each gridpoint. The sea level pressure histogram
159 can be flattened by adding random noise of magnitude 1hPa to the data, which is less
160 than 8% of the average ensemble spread at each gridpoint. It is also worth noting that
161 this analysis finds the ensemble width to be generally too wide, rather than too narrow
162 as has been previously claimed. This suggests that the direct probabilistic interpretation
163 of the ensemble will err on the side of of caution, rather than having a high probability
164 of excluding the truth.

165 We have also performed a similar analysis for each model in turn, generating its rank
166 histogram among the remaining 23 models. The χ^2 tests for bias and spread frequently fail

167 (eg in 30% of cases at the $p < 5\%$ level), with a similar degree of non-uniformity as found
168 for the observational data. If we assume 5 degrees of freedom rather than 40, the failure
169 rate is reduced to around 5% at the 5% level, and the divergence from uniformity noted
170 for the rank histogram of observational data is no longer statistically significant. Thus,
171 the models appear to be about as reliable at predicting reality as they are at predicting
172 each other, further supporting the hypothesis of exchangeability between the models and
173 true climate system.

5. Discussion

174 The CMIP3 ensemble has arisen through a process of large numbers of researchers
175 making numerous diverse decisions according to their beliefs about the climate system.
176 It should hardly be surprising that these beliefs are biased in their mean, and that the
177 resulting ensemble of models is not centred on the truth. So long as the range of choices
178 made is commensurate with the errors that exist, however, this in no way precludes the
179 ensemble members from forming a sample which is statistically indistinguishable from the
180 truth.

181 Given the various contingencies relating to the creation of the CMIP3 ensemble of
182 opportunity, it would be unreasonably optimistic to expect it to be *perfectly* reliable in all
183 respects. However, our analysis here suggests, at least for the data considered here, that
184 this assumption is in fact not far from the truth, although there are weak indications that
185 the model spread may be a little too broad.

186 An important issue, that we do not address here, is the relationship between past and
187 future performance [eg *Whetton et al.*, 2007; *Abe et al.*, 2009]. Arguably, the CMIP3

188 climate models have already been tuned to some extent to the recent climate data that
189 we have used here. It is not immediately clear that this should affect the reliability of the
190 ensemble, as this would require not just that the biases on individual models are reduced,
191 but that the biases change sign, and do so preferentially in one direction. However, this
192 issue is worthy of further investigation. Additionally, it would be interesting to test further
193 the reliability of the ensemble in other ways, for example considering simulations of other
194 epochs, or other climatic observations that are less widely used during model construction
195 and tuning.

196 The analysis presented here implicitly assigns equal weight to each ensemble member,
197 which would be appropriate if we believe them to be equally good models of the climate
198 system. Such an approach is normal for single-model ensembles generated by perturbing
199 initial conditions, but may not be so sensible when samples arise heterogeneously, as is the
200 case with the CMIP3 ensemble. Therefore, the quality of probabilistic predictions in terms
201 of both reliability and resolution may be improved by some non-uniform weighting. A va-
202 riety of approaches are in common use, including both heuristic re-weighting methods [*Zhu*
203 *et al.*, 1996; *Krishnamurti et al.*, 2000] and formal Bayesian methods such as Bayesian
204 Model Averaging [*Raftery et al.*, 2005]. The paradigm of a statistically-indistinguishable
205 ensemble provides an appropriate theoretical foundation for the exploration of these ideas.
206 There is also a wide range of established analysis techniques which may be applicable for
207 evaluating ensemble performance [*Toth et al.*, 2003].

6. Conclusions

208 An ensemble which is statistically indistinguishable from the truth will appear to be
209 biased and non-convergent when analysed under the assumption that it is a truth-centred
210 ensemble. We have shown that the CMIP3 ensemble appears fairly reliable when tested
211 against recent observations, and if anything tends towards being over-broad, in contrast
212 to recent claims. Thus, our analysis supports the direct probabilistic interpretation of the
213 ensemble, although we expect its reliability (and resolution) could be further improved
214 by non-uniform weighting. We suggest that in place of the truth-centred approach, future
215 research into the use of the CMIP3 and other multi-model ensembles of opportunity
216 should be based on the paradigm of a statistically indistinguishable ensemble, as this is
217 both intuitively plausible and reasonably compatible with observational evidence.

218 **Acknowledgments.** We are grateful to two reviewers for helpful comments. This work
219 was supported by the S-5-1 project of the MoE, Japan and by the Kakushin Program of
220 MEXT, Japan. We acknowledge the modeling groups, the Program for Climate Model
221 Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled
222 Modelling (WGCM) for their roles in making available the WCRP CMIP3 multi-model
223 dataset. Support of this dataset is provided by the Office of Science, U.S. Department of
224 Energy.

References

225 Abe, M., H. Shiogama, J. Hargreaves, J. Annan, T. Nozawa, and S. Emori (2009), Cor-
226 relation between Inter-Model Similarities in Spatial Pattern for Present and Projected
227 Future Mean Climate, *SOLA*, 5(0), 133–136.

- 228 Adler, R., et al. (2003), The version-2 global precipitation climatology project (GPCP)
229 monthly precipitation analysis (1979–present), *Journal of Hydrometeorology*, 4(6),
230 1147–1167.
- 231 Allan, R., and T. Ansell (2006), A new globally complete monthly historical gridded
232 mean sea level pressure dataset (HadSLP2): 1850–2004, *Journal of Climate*, 19(22),
233 5816–5842.
- 234 Anderson, J. (1996), A method for producing and evaluating probabilistic forecasts from
235 ensemble model integrations, *Journal of Climate*, 9(7), 1518–1530.
- 236 Brohan, P., J. Kennedy, I. Harris, S. Tett, and P. Jones (2006), Uncertainty estimates
237 in regional and global observed temperature changes: a new dataset from 1850, *J.*
238 *Geophys. Res.*, 111(D12).
- 239 Giorgi, F., and L. Mearns (2002), Calculation of average, uncertainty range, and reliability
240 of regional climate changes from AOGCM simulations via the “reliability ensemble
241 averaging” (REA) method, *Journal of Climate*, 15(10), 1141–1158.
- 242 Jewson, S., and E. Hawkins (2009), CMIP3 ensemble spread, model similarity, and climate
243 prediction uncertainty, <http://arxiv1.library.cornell.edu/abs/0909.1890>.
- 244 Jolliffe, I., and C. Primo (2008), Evaluating Rank Histograms Using Decompositions of
245 the Chi-Square Test Statistic, *Monthly Weather Review*, 136(6), 2133–2139.
- 246 Jun, M., R. Knutti, and D. Nychka (2008), Spatial analysis to quantify numerical model
247 bias and dependence: how many climate models are there?, *Journal of the American*
248 *Statistical Association*, 103(483), 934–947.

- 249 Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl (2009), Challenges in
250 combining projections from multiple climate models, *Journal of Climate*, (Accepted).
- 251 Krishnamurti, T., C. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford,
252 S. Gadgil, and S. Surendran (2000), Multimodel ensemble forecasts for weather and
253 seasonal climate, *Journal of Climate*, *13*(23), 4196–4216.
- 254 Murphy, A. (1973), A new vector partition of the probability score, *Journal of Applied*
255 *Meteorology*, *12*(4), 595–600.
- 256 Nychka, D., and C. Tebaldi (2003), Comments on “Calculation of Average, Uncertainty
257 Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the
258 “Reliability Ensemble Averaging” (REA) Method”, *Journal of Climate*, *16*(5), 883–884.
- 259 Raftery, A., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005), Using Bayesian model
260 averaging to calibrate forecast ensembles, *Monthly Weather Review*, *133*(5), 1155–1174.
- 261 Räisänen, J., and T. Palmer (2001), A probability and decision-model analysis of a multi-
262 model ensemble of climate change simulations, *Journal of Climate*, *14*(15), 3212–3226.
- 263 Smith, R., C. Tebaldi, D. Nychka, and L. Mearns (2009), Bayesian modeling of uncer-
264 tainty in ensembles of climate models, *Journal of the American Statistical Association*,
265 *104*(485), 97–116.
- 266 Talagrand, O., R. Vautard, and B. Strauss (1997), Evaluation of probabilistic prediction
267 systems, in *Proc. ECMWF Workshop on Predictability*, pp. 1–25.
- 268 Tebaldi, C., and R. Knutti (2007), The use of the multi-model ensemble in probabilistic
269 climate projections, *Philosophical Transactions of the Royal Society A: Mathematical*,
270 *Physical and Engineering Sciences*, *365*(1857), 2053.

- 271 Tebaldi, C., R. Smith, D. Nychka, and L. Mearns (2005), Quantifying uncertainty in pro-
272 jections of regional climate change: a Bayesian approach to the analysis of multimodel
273 ensembles, *Journal of Climate*, 18(10), 1524–1540.
- 274 Toth, Z., O. Talagrand, G. Candille, and Y. Zhu (2003), Probability and ensemble fore-
275 casts, *Forecast Verification: A Practitioners Guide in Atmospheric Science*, pp. 137–
276 163.
- 277 Whetton, P., I. Macadam, J. Bathols, and J. O’Grady (2007), Assessment of the
278 use of current climate patterns to evaluate regional enhanced greenhouse response
279 patterns of climate models, *Geophysical Research Letters*, 34(14), L14,701, doi:
280 10.1029/2007GL030025.
- 281 Zhu, Y., G. Iyengar, Z. Toth, M. Tracton, and T. Marchok (1996), Objective evaluation
282 of the NCEP global ensemble forecasting system, in *CONFERENCE ON WEATHER*
283 *ANALYSIS AND FORECASTING*, vol. 15, pp. 79–82, American Meteorological Soci-
284 ety.

Figure 1. Analysis of three idealised ensembles. Top row: histograms of the pairwise correlation coefficients of the biases of the ensemble members. Second row: RMSE of the mean of random subsets of the ensemble members, plotted as a function of subset size. Solid red lines show the mean over repeated subsampling, and the dashed red lines give the upper and lower ranges obtained. Black dashed lines gives the theoretically-expected result of $\sqrt{1/n + \sigma^2}$ where $\sigma=0.5, 1, 2$ is the standard deviation of the true variables, and the dotted lines show the $\sqrt{1/n}$ convergence that would be expected for a truth-centred ensemble. Bottom row: rank histograms for the 40 true values of the variables in each experiment, with contributions of bias and spread to the χ^2 statistic (see text for details).

Figure 2. Rank histogram analysis of outputs of CMIP3 models versus observational data. Sea surface temperature, precipitation and sea level pressure are shown from top to bottom. χ^2 statistics for bias and spread are also presented in each subplot (see text for details).



