On the observational assessment of climate model performance

J. D. Annan,[1] J. C. Hargreaves,[1] and K. Tachiiri[1]

Received 27 September 2011; revised 9 November 2011; accepted 11 November 2011; published 17 December 2011.

[1] Comparison of model outputs with observations of the climate system forms an essential component of model assessment and is crucial for building our confidence in model predictions. Methods for undertaking this comparison are not always clearly justified and understood. Here we show that the popular approach of comparing the ensemble spread to a so-called "observationally-constrained pdf" can be highly misleading. Such a comparison will almost certainly result in disagreement, but in reality tells us little about the performance of the ensemble. We present an alternative approach, and show how it may lead to very different, and rather more encouraging, conclusions. We additionally present some necessary conditions for an ensemble (or more generally, a probabilistic prediction) to be challenged by an observation. **Citation:** Annan, J. D., J. C. Hargreaves, and K. Tachiiri (2011), On the observational assessment of climate model performance, *Geophys. Res. Lett.*, *38*, L24702, doi:10.1029/2011GL049812.

1. Introduction

[2] One primary method by which we assess models and gain confidence in their predictions, is by the comparison of model outputs with observations of the climate system. This paper considers the issue of assessing model ensembles such as that arising from the CMIP3 coordinated multi-model experiments. Here an important component is not merely the closeness of the models to observations in absolute terms [e.g., *Reichler and Kim*, 2008] but also the reliability of the ensemble spread as an indication of uncertainty. In this context, it has been widely argued that the multi-model ensemble of opportunity which participated in the CMIP3 project is insufficiently broad to adequately represent uncertainties regarding future climate change. *Meehl et al.* [2007, p. 801] summarise the consensus with the sentence: "Those studies also suggest that the current AOGCMs may not cover the full range of uncertainty for climate sensitivity."

[3] This statement is based on a comparison of the spread of the CMIP3 ensemble to a number of probability density functions (pdfs) which were generated through probabilistic analyses of observational evidence. While the equilibrium climate sensitivity is by far the most intensively studied case, similar claims have been made for other properties of the climate system, including for example the transient climate response [*Knutti and Tomassini*, 2008] and efficiency of ocean heat uptake [*Forest and Stone*, 2008]. In this paper, we present an alternative perspective on these model-data comparisons. We start out in section 2 by presenting an idealised version of the type of pdf comparison that forms the basis of the literature summarised above. We then introduce an alternative perspective on assessing probabilistic predictions (and more specifically, the CMIP3 ensemble) with uncertain observations, which directly tests the predictive ability of the ensemble. We show how and why these two approaches may lead to very different conclusions and explain why the latter approach, which is well-established in other fields, is preferable. We find some necessary conditions under which confidence in the ensemble is actually reduced by an observation.

[4] In section 3, we present a more realistic (although still substantially simplified) example based on using observations of recent climate change to constrain parameters in an energy balance model. We show our example to be consistent with the literature in that an observationally-constrained probabilistic estimate of climate system properties has substantially greater uncertainty than an ensemble of models designed to fit the CMIP3 results. However, when instead we consider the predictive ability of the ensemble, we find that the observations actually enhance our confidence in it, albeit weakly. We also show that evaluation of the CMIP3 ensemble with these observations should lead to more positive conclusions than have usually been drawn.

2. Theoretical Example

[5] We first illustrate the principles under consideration via an idealised univariate case. We assume that our ensemble of model outputs can be summarised by the Gaussian $N(M, \sigma_M)$, and we have a single uncertain observation O with associated known Gaussian observational uncertainty σ_O, that is to say, O is considered to have been drawn from the distribution $N(T, \sigma_O)$ where T is the unknown truth. Our likelihood function for the truth, $L(T|O)$, is therefore proportional to a Gaussian centred on the observed value, $N(O, \sigma_O)$.

2.1. Method 1: Direct Comparison of pdfs

[6] In principle, performing an observational analysis to generate a posterior pdf for the truth requires the used of Bayes' Theorem and a prior on T. In practice, it is widespread to adopt a uniform prior (perhaps implicitly) and effectively interpret the likelihood function $N(O, \sigma_O)$ directly as this posterior pdf. Clearly this will differ from the distribution arising from the ensemble, $N(M, \sigma_M)$, except in the special case where $O = M$ and $\sigma_O = \sigma_M$, and since observational error $O-T$ is generally expected to be independent of model error $M-T$, such agreement is vanishingly improbable. Even in the case of some other prior being used, a mismatch will occur with probability 1. Thus, we see that a direct comparison of model-based and observationally-based pdfs, as presented by *Meehl et al.* [2007], will in general always result in disagreement. For a concrete example, we consider the specific case where the ensemble mean and spread are

[1]Research Institute for Global Change, Yokohama, Japan.

0094-8276/11/2011GL049812

respectively $M = 0$ and $\sigma_M = 1$, and the observation takes the value $O = 1.5$ with uncertainty $\sigma_O = 2$. The model ensemble $N(0, 1)$ appears somewhat biased and rather narrow compared to the observational analysis of $N(1.5, 2)$. Note, however, that the ensemble will appear narrow compared to the observational analysis *irrespective of the observed value itself*, since this property is simply due to the observational uncertainty being greater the ensemble spread.

2.2. Method 2: Evaluation of Predictive Performance

[7] We now present an alternative approach to ensemble validation, in which we directly address the question of whether the observation supports or challenges the ensemble in terms of matching the models' prediction. In order to assess the GCM ensemble, we must first briefly consider how to interpret it. The most natural hypothesis, that underpins the widespread use of ensembles in probabilistic prediction, is that the ensemble members are considered to be equally plausible samples of our collective beliefs about the climate system, and so can be interpreted as a finite sample from an underlying probability distribution which we may hope to generate a "reliable" prediction [*Annan and Hargreaves*, 2010]. While it is of course unlikely that this hypothesis will be precisely true, this approach provides a simple starting point for interpreting and using the ensemble, and at the very least provides a benchmark for assessing ensemble performance. That is, we can consider the predictive distribution arising from the ensemble, and check as to whether the observations are in fact compatible with it.

2.2.1. Frequentist Hypothesis Testing

[8] The hypothesis under consideration is that the ensemble provides a reliable prediction, and thus (for the general Gaussian ensemble) that truth is also sampled from $N(M, \sigma_M)$. We only have a noisy observation rather than the truth with which to test this hypothesis, but this can be accounted for by adding observational uncertainty onto the model estimates [*Anderson*, 1996], so as to generate the predictive distribution for the observation itself. The question then becomes whether the observation is compatible with our predictive distribution for it, which we can test in the standard frequentist manner. That is, should we reject the hypothesis that the observation was sampled from our predictive distribution?

[9] In the univariate Gaussian case considered here, the observational and modelling uncertainties add in quadrature to give the predictive distribution $N(M, \sqrt{\sigma_M^2 + \sigma_O^2})$ for the observation. It is immediate that in order for the observation to lie outside the X% probability range of the predictive distribution, for a given X, it has to be at least outside the X% range of the original ensemble (since $\sqrt{\sigma_M^2 + \sigma_O^2} > \sigma_M$). Moreover, and at least as importantly, the mean of the model distribution has to lie outside the natural X% confidence interval of the observation. It is crucial to recognise this latter point, especially when observational uncertainty is high. In the specific numerical case presented in Section 2.1, the observed value of $O = 1.5$ is well within the predictive 90% interval of $[-3.67, 3.67]$ and thus provides no reason to doubt the models even at the moderate $p < 0.1$ level.

2.2.2. Bayesian Updating

[10] We can also consider a more overtly Bayesian approach in which we consider the ensemble as providing a prior pdf for the truth, and update this distribution with the observational evidence. Following the numerical example, our prior pdf for T arising from the ensemble is $N(0, 1)$, and updating with the observation through Bayes' Theorem results in the posterior $N(0.3, 0.9)$. If we integrate this posterior pdf over the 90% probability range of the prior $[-1.64, 1.64]$, we find that it actually assigns a slightly increased probability of 92% to this range. That is, while our posterior belief has been nudged towards slightly higher values, the observation has also marginally enhanced our belief that the truth lies in the original 90% probability range derived from the models. In Figure 1, we present an analysis of how the posterior probability integrated over the prior 90% probability interval will vary, depending on the observed value and its uncertainty. We find that in order for the posterior probability assigned to the prior 90% interval to decrease (so that the observation weakens our faith in the models) not only must the observation lie outside the prior 80% interval, but also the prior mean must lie outside the one standard deviation (68%) confidence interval of the observation. These results differ quantitatively from the frequentist hypothesis-testing approach described in section 2.2.1, but have a similar overall flavour. That is, the observation must not only be distant, but also precise, in order to reduce confidence in the models.

2.2.3. Interpretation

[11] Any doubts about the performance of the ensemble arising through the direct comparison of ensemble with "observationally-constrained pdf" as outlined in section 2.1, may now be understood as being due to the use of a prior for the truth which assigns very little probability to the ensemble being adequate in the first place. If an unbounded uniform prior is used for the observational analysis, this technically assigns zero probability to the ensemble's 90% probability range of $[-1.64, 1.64]$. Even if the uniform prior is bound to the moderate range $U[-10, 10]$, the prior probability assigned to the model range is still as low as 16.4%. The "observationally-derived pdf" of $N(1.5, 2)$ actually assigns a substantially increased probability of 47% to this range. In other words, rather than the observation telling us that the model range is inadequate, the observational analysis actually started from a strong implicit presumption in this direction, and while the observation greatly increased our confidence in the models, it was insufficiently precise to overcome that initial prejudice.

3. Practical Example

[12] We now consider a more realistic application illustrating the practical impact of these alternative perspectives. The presentation here is necessarily greatly simplified compared to a detailed analysis, but should adequately represent the essential elements. We use a two-box energy balance model of the climate system which includes a surface layer with low heat capacity, and a deep ocean with a greater heat capacity [*Gregory*, 2000]:

$$C_1 dT_1/dt = F - \lambda T_1 - k(T_1 - T_2) \quad (1)$$

$$C_2 dT_2/dt = k(T_1 - T_2) \quad (2)$$

where $C_{1,2}$ are the respective heat capacities of the upper and lower layers (fixed at 8 and 300 W a m^{-2} K^{-1} respectively, corresponding to uniform well-mixed layers of 60 and

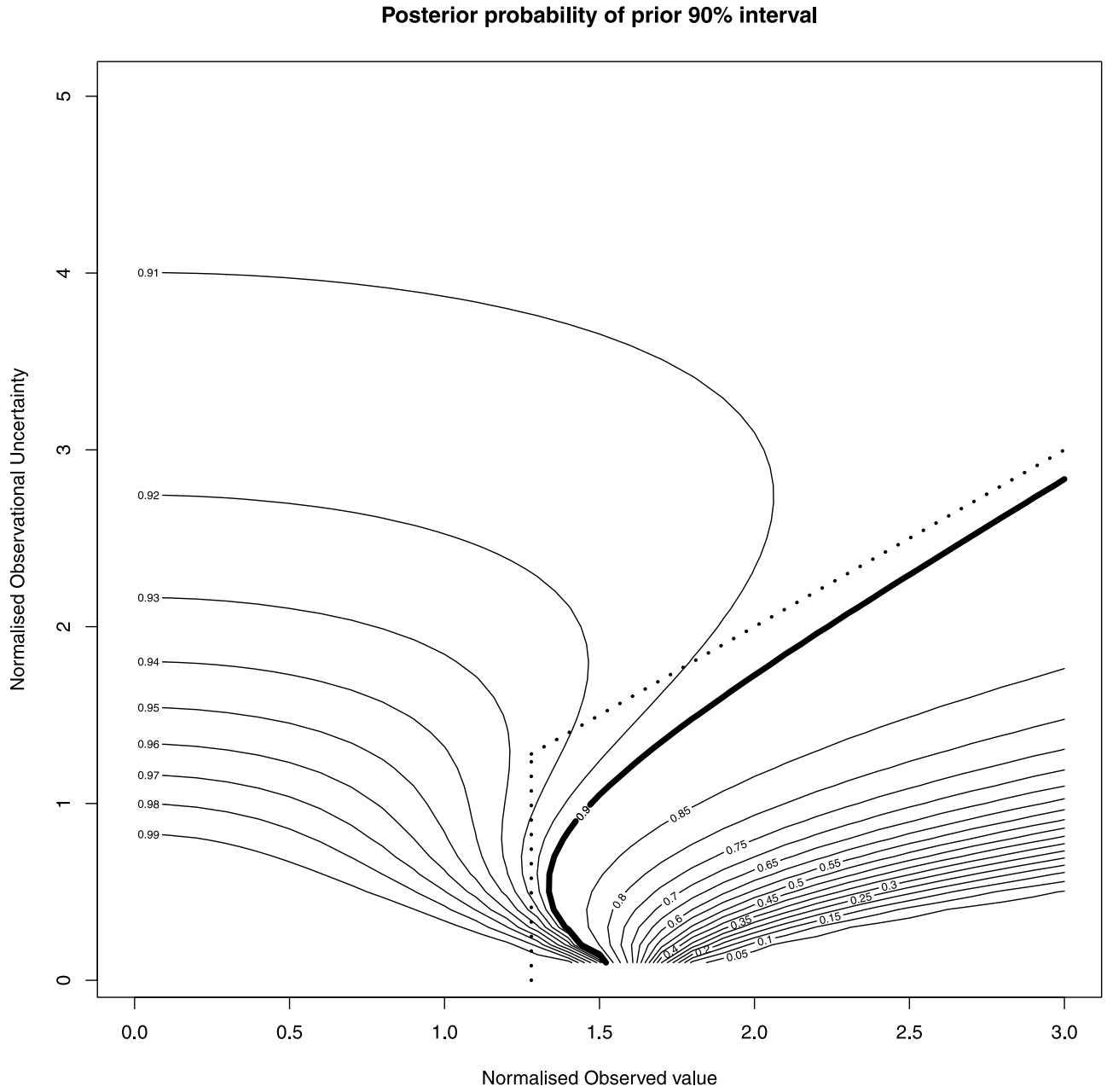

Figure 1. Posterior probability of prior 90% interval, as a function of observed value and uncertainty. Region where posterior probability is lower than 90% — and so the observation reduces belief in the models — is confined to a region of distant but precise observations in the lower right corner of the plot, strictly bounded by the dotted line indicating where both $O > 1.28$ (being the 90th percentile of the model distribution) and $O > \sigma_O$.

2300 m of water), $T_{1,2}$ are the respective temperature anomalies, F is the external radiative forcing anomaly, λ is the radiative feedback (inverse of the equilibrium sensitivity), and k a flux coefficient which is basically equal to the ocean heat uptake efficiency. Forcing is taken from *Crowley* [2000], with uncertainty in aerosol forcing represented by applying a scaling factor. For observations, we use the surface warming trend over the 20th century of 0.65 ± 0.12 K, and oceanic heat gain over the past 50y. Since observational estimates of the latter available in the literature differ enough to substantially affect inferences [*Sokolov et al.*, 2010], we use a representative value of 12.5 ± 2.5 W a m^{-2} representing a mean planetary imbalance of 0.25 ± 0.05 W m^{-2} over this interval (equating to $\simeq 20 \pm 4 \times 10^{22}$ *J* of total heat gain over the 50 year period), and test sensitivity to this choice. The uncertainties here are taken to be one standard deviation of independent Gaussians, and thus we take the likelihood function proportional to $e^{-0.5[(\frac{\Delta T-0.65}{0.12})^2 + (\frac{\Delta OHC-12.5}{2.5})^2]}$ where ΔT and ΔOHC are the modelled changes in surface air temperature and ocean heat content for a given parameter set.

[13] As a proxy for the CMIP3 ensemble of GCMs, we use an ensemble of energy balance models in which $S = 1/\lambda \sim N(3, 0.75)/3.7$ K (ie 3 ± 0.75 K for a doubling of CO_2) and $k \sim N(0.69, 0.12)$Wm^{-2} K^{-1} [*Dufresne and Bony*, 2008].

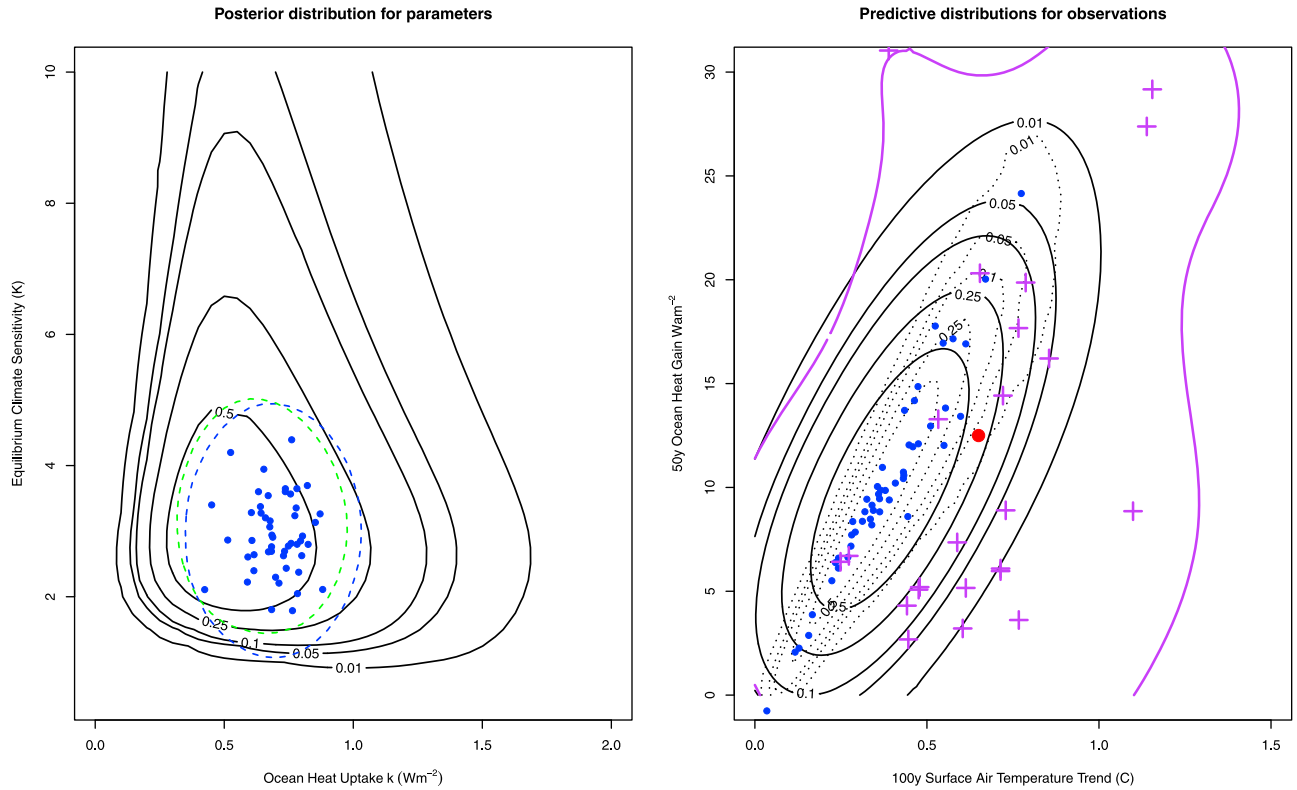

Figure 2. (left) Posterior joint pdf for equilibrium sensitivity and ocean heat uptake parameter, conditioned on observations of 100y temperature trend and 50y oceanic warming. Black contours indicate probability levels as labelled. Blue dots represent energy balance models with parameters sampled from the CMIP3-inspired distribution, and blue dashed line indicates 90% probability contours of this distributions. Green dashed line indicates 90% probability contour of the posterior after updating with observational likelihood. (right) Predictive distributions from energy balance model based on CMIP3 parameters. Solid lines indicate predictive pdf for observations, when accounting for their uncertainty. Dotted contours indicate probability distribution without accounting for observational uncertainty. Red dot indicates observation. Purple crosses indicate outputs of the CMIP3 simulations, and purple line marks their 90% probability level.

Scaling on the aerosol forcing is sampled from $N(0.7, 0.3)$ in order to generate a range of net forcing from around 1 to 2 W over the 20th century [*Kiehl*, 2007], although we do not impose the correlation with sensitivity which he observed in the real models. While it is not our intention that these analyses based on the energy balance model should be considered quantitatively precise, the parameters are selected so as to give a range of behaviour similar to that of the CMIP3 ensemble. In particular, the spreads of both equilibrium sensitivity and transient climate response (1.85 ± 0.3 K) are close to those of the CMIP3 GCMs.

3.1. Method 1: Direct Comparison of pdfs

[14] In order to generate an observationally-constrained analysis of the climate system parameters, we use the following priors; the Gaussian N(1,0.3) for the heat uptake coefficient k, the Cauchy distribution with location 2.5 and scale $\sqrt{3}$ for the equilibrium sensitivity $S = 1/\lambda$ [*Annan and Hargreaves*, 2009], and the Gaussian N(0.8,0.5) for the scaling on the aerosol forcing. These priors are chosen to be reasonably broad while avoiding the pathological behaviour noted by *Annan and Hargreaves* [2009], but our results are generally rather insensitive to them. In particular, uniform bounded priors give qualitatively similar results.

[15] The results from our observational analysis are presented in the left hand panel of Figure 2, together with 50 samples from the 'CMIP3' proxy ensemble of energy balance models. It is clear that the posterior of this analysis has a far greater spread than the ensemble, with a 90% probability range of 1.82–7.62 K (range of 5.80 K) for equilibrium sensitivity and 0.30–1.14 W m^{-2} K^{-1}(range of 0.83 W m^{-2} K^{-1}) for k. The equivalent ranges for the ensemble are 1.77–4.23 K (range of 2.46 K) for sensitivity and 0.49–0.89 W m^{-2} K^{-1} (range of 0.39 W m^{-2} K^{-1}) for k. Thus, the posterior pdf has a substantially greater range than the ensemble, just as in much of the published literature. It might appear to some that our ensemble "may not cover the full range of uncertainty".

[16] Repeating this exercise with observations ranging from 0.4–1 K for the surface trend and 5–20 W am^{-2} for the ocean heat gain we find that for every case the 90% range for climate sensitivity remains greater than 5 K, and the range for ocean heat uptake efficiency is similarly high except in the cases where k is constrained to small values by a large surface warming and low ocean warming. Therefore, if this analysis was valid, we would always conclude that the CMIP3 ensemble does not span an adequate range of uncertainty irrespective of the observations, and indeed we could state this result prior to even making the observations. Of course such a conclusion is nonsensical. In fact the real problem is that our prior only assigns a probability of around 28% to the models' joint 90% range over these two parameters. While

this probability doubles to 56% in the posterior, the observational evidence is insufficiently precise to overwhelm the prior spread.

3.2. Method 2: Evaluation of Predictive Performance

[17] The alternative perspective, in which we evaluate the predictive performance of the models (as in section 2.2), generates strikingly different results. Figure 2 (right) presents the predictive distribution for the observations, based on the model ensemble. If observational uncertainty is ignored, then the observations do appear rather unlikely, lying outside the 5% probability threshold at which we might well reject the models under the common frequentist hypothesis testing paradigm of section 2.2.1. However, once observational uncertainty is correctly accounted for, the values observed appear entirely unremarkable, and the hypothesis is not close to being rejected. As in section 2.2.2, we can also perform a Bayesian analysis using the model ensemble as a prior (together with the possibly dubious assumption that the observations are independent of this prior), and these results are presented in Figure 2 (left). The posterior distribution has been shifted to marginally higher sensitivity and lower ocean heat uptake, but the integrated probability over the prior 90% region has been very slightly increased to around 91%. Thus we conclude that the observations, if considered independent of the models, would act to enhance our confidence (albeit marginally) in the latter. Figure 2 (right) also shows the outputs of the CMIP3 ensemble of GCMs, calculated from the 20th century simulations. The models represent a broader spectrum of behaviours than can be represented by the energy balance model, possibly due to forcing differences, as well as more sophisticated dynamical behaviour. However, the observations of climate change fall well within the ensemble spread.

[18] An additional advantage of evaluating predictive performance in this way is that it takes place entirely in observation space. Therefore, we do not need to assert or infer "true" parameter values for intangible properties of the climate system such as equilibrium climate sensitivity or ocean heat uptake efficiency. The epistemic role of such parameters is somewhat obscure, since they represent simplifications that are undoubtedly false in the real world (for example, that these properties are fixed in time). We merely test whether the models provide an adequate prediction of observable attributes, which are the only properties of the real world that are actually accessible to us in any case.

4. Conclusions

[19] We have shown that the popular evaluation of ensemble spread, based on a direct comparison with a pdf based on observational constraints, can be highly misleading. Disagreement between these two pdfs is virtually certain, but such an analysis does not directly address the fundamental question of whether the ensemble provides reliable predictions. Therefore, we recommend that researchers should avoid presenting such unhelpful comparisons in future. An alternative approach, which directly addresses the question of predictive performance, enables us to evaluate to what extent the observations support or contradict the ensemble. In many cases, we expect that this approach will result in the conclusion that observations actually enhance our confidence in the models. We have demonstrated that an ensemble can only be meaningfully challenged by observations which are both distant from it, and precise.

[20] **Acknowledgments.** This work was supported by the S-5-1 project of the MoE, Japan and by the Kakushin Program of MEXT, Japan.
[21] The Editor thanks the two anonymous reviewers for their assistance in evaluating this paper.

References

Anderson, J. (1996), A method for producing and evaluating probabilistic forecasts from ensemble model integrations, *J. Clim.*, *9*(7), 1518–1530.
Annan, J. D., and J. C. Hargreaves (2009), On the generation and interpretation of probabilistic estimates of climate sensitivity, *Clim. Change*, *104*, 423–436, doi:10.1007/s10584-009-9715-y.
Annan, J. D., and J. C. Hargreaves (2010), Reliability of the CMIP3 ensemble, *Geophys. Res. Lett.*, *37*, L02703, doi:10.1029/2009GL041994.
Crowley, T. J. (2000), Causes of climate change over the past 1000 years, *Science*, *289*, 270–277.
Dufresne, J., and S. Bony (2008), An assessment of the primary sources of spread of global warming estimates from coupled atmosphere-ocean models, *J. Clim.*, *21*(19), 5135–5144.
Forest, C., P. Stone, and A. Sokolov (2008), Constraining climate model parameters from observed 20th century changes, *Tellus, Ser. A*, *60*(5), 911–920.
Gregory, J. (2000), Vertical heat transports in the ocean and their effect on time-dependent climate change, *Clim. Dyn.*, *16*(7), 501–515.
Kiehl, J. T. (2007), Twentieth century climate model response and climate sensitivity, *Geophys. Res. Lett.*, *34*, L22710, doi:10.1029/2007GL031383.
Knutti, R., and L. Tomassini (2008), Constraints on the transient climate response from observed global temperature and ocean heat uptake, *Geophys. Res. Lett.*, *35*, L09701, doi:10.1029/2007GL032904.
Meehl, G. A., et al. (2007), Global climate projections, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, chap. 10, pp. 747–845, Cambridge Univ. Press, Cambridge, U. K.
Reichler, T., and J. Kim (2008), How well do coupled models simulate today's climate?, *Bull. Am. Meteorol. Soc.*, *89*(3), 303–311.
Sokolov, A., C. Forest, and P. Stone (2010), Sensitivity of climate change projections to uncertainties in the estimates of observed changes in deep-ocean heat content, *Clim. Dyn.*, *34*(5), 735–745.

J. D. Annan, J. C. Hargreaves, and K. Tachiiri, Research Institute for Global Change, 3073-25 Showamachi, Yokohama, Kanagawa 236-0001, Japan. (jdannan@jamstec.go.jp; jules@jamstec.go.jp; tachiiri@jamstec.go.jp)