On the interpretation of constrained climate model ensembles

Benjamin M. Sanderson[1] and Reto Knutti[2]

Received 8 June 2012; revised 18 July 2012; accepted 19 July 2012; published 29 August 2012.

[1] An ensemble of models can be interpreted in two ways. The first treats each model as an approximation of the true system with some random error. Alternatively, the true system can be interpreted as a sample drawn from a distribution of models, such that model and truth are statistically indistinguishable. Both interpretations are ubiquitous and have different consequences for the uncertainty of model projections, but are rarely defended. Here we argue that the two seemingly conflicting views are in fact complementary, and the interpretation of the ensemble may evolve seamlessly from the former to the latter. We show some 'truth plus error' like properties exist for historical and present day climate simulations in the CMIP archive, and that they can be explained by the ensemble design and tuning to observations, although both models and tuning are imperfect. For future projections, structural differences in model response arise which are independent of the present day state and thus the 'indistinguishable' interpretation is increasingly favored. Our inability to define performance metrics that identify 'good' and 'bad' models can be explained by the models having largely exploited the available observations. The remaining model error is largely structural and the observations are often uninformative to further reduce model biases or reduce the range of projections covered by the ensemble. The discussion here is motivated by the use of multi model ensembles in climate projections, but the arguments are generic to any situation where multiple different models constrained by observations are used to describe the same system. **Citation:** Sanderson, B. M., and R. Knutti (2012), On the interpretation of constrained climate model ensembles, *Geophys. Res. Lett.*, *39*, L16708, doi:10.1029/2012GL052665.

1. Introduction

[2] Recent coordinated efforts to produce simulations of past and future climate with many climate models developed at different institutions have provided new opportunities to explore uncertainties in climate projections. Literally hundreds of studies were conducted using data from the recent World Climate Research Program (WCRP) Coupled Model Intercomparison Project Phase 3 (CMIP3) [*Meehl et al.*, 2005], and the next CMIP5 effort is already underway. While the simulations in those intercomparisons are clearly specified, participation by the modeling groups is voluntary. The resulting ensemble is neither sampled randomly nor systematically and is often called an "ensemble of opportunity" [*Tebaldi and Knutti*, 2007]. Model performance, resolution, as well as the degree of complexity by which the different models describe various processes vary substantially [*Gleckler et al.*, 2008; *Reichler and Kim*, 2008]. Models also share whole components or parameterizations and are therefore not independent [*Masson and Knutti*, 2011; *Pennell and Reichler*, 2011], making the sampling distribution of any statistic computed from ensemble difficult to interpret [*Knutti et al.*, 2010a].

[3] For many qualitative or process-based studies, a strict statistical interpretation of the ensemble is unnecessary, indeed models can be used to validate physical arguments, conservation laws or for individual model evaluation without any need for underlying assumptions about the ensemble distribution. However, for any probabilistic evaluation of ensemble projections – one is forced to make a judgment on the ensemble structure and there are two fundamentally different interpretations of the ensemble that might appear mutually exclusive [*Annan and Hargreaves*, 2010; *Knutti et al.*, 2010b; *Pennell and Reichler*, 2011; *Tebaldi and Sanso*, 2009]. The models can be interpreted as 'truth plus error', e.g., random samples from a distribution of plausible models centered about the true climate [*Buser et al.*, 2009; *Furrer et al.*, 2007; *Smith et al.*, 2009; *Tebaldi et al.*, 2005]. Note that for simplicity the observed state is considered as truth in the following discussion. Alternatively, in the statistically 'indistinguishable' interpretation [*Annan and Hargreaves*, 2010], each model can be considered exchangeable with the other members and with the real system [e.g., *Jackson et al.*, 2008; *Murphy et al.*, 2007; *Perkins et al.*, 2007], i.e., the truth (the real climate) and all models are thought to be drawn from the same distribution. Many studies which attempt to integrate simulations from different models necessarily use one or the other interpretation, but with a few exceptions [*Annan and Hargreaves*, 2010; *Lopez et al.*, 2006; *Tebaldi and Sanso*, 2009] usually do not discuss the statistical framework. Uncertainty in projections decreases strongly in the 'truth plus error' view as more models are considered [*Annan and Hargreaves*, 2010; *Knutti et al.*, 2010a; *Lopez et al.*, 2006; *Tebaldi and Sanso*, 2009]. This is because the uncertainty of the model consensus (similar to the uncertainty in the mean for independent measurements) is estimated more precisely as the sample size increases. In contrast, in the 'indistinguishable' interpretation the uncertainty is characterized by the ensemble spread and is largely independent of the sample size (see *Annan and Hargreaves* [2010] for more details). For that reason, the interpretation of the ensemble is not just an academic question, but of direct relevance to quantifying uncertainties in projections.

[4] A perfect ensemble of weather forecasts would aim to be indistinguishable from truth, and with sufficient samples,

[1]Climate and Global Dynamics, National Center for Atmospheric Research, Boulder, Colorado, USA.
[2]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland.

Corresponding author: B. M. Sanderson, Climate and Global Dynamics, National Center for Atmospheric Research, PO Box 3000, Boulder, CO 80307, USA. (bsander@ucar.edu)

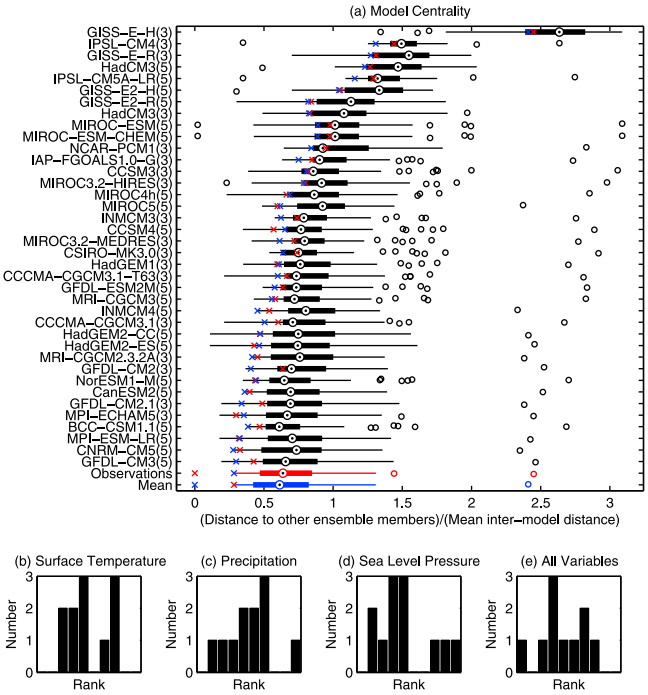

Figure 1. (a) Distributions of inter-model distances calculated over multiple variables for CMIP-3, CMIP-5 and selected observations. For each model, a box-whisker plot is shown where the central, filled circle is the median distance between that model and other models in the combined ensemble. Boxes indicate the 25th to 75th percentiles of the distribution, while the whiskers show the full width of the distribution. Outliers, defined to be greater than 1.5 times the inter-quartile range above the upper quartile, or less than the lower quartile, are shown as unfilled circles. The observations (red) and multi-model mean (blue) are treated as additional ensemble members, and the corresponding distances from each model to the observations and mean are plotted with red and blue crosses on each line respectively. All distances are normalized by the mean inter-model distance in the combined CMIP3/CMIP5 ensemble. Rank histograms for the observations in the combined CMIP3/CMIP5 ensemble. The rank of the observations in the context of the combined ensemble is evaluated for each of a truncated set of EOFs, weighted by the variance associated with that EOF, and collated into 10 equally sized bins. EOFs are calculated for (b) surface temperature, (c) total precipitation, (d) sea level pressure and (e) for the multivariate case. Red vertical bars show rank delimitations for 25th and 75th percentiles of the ensemble, with numbers showing the frequency that the observations fall into each delimitation.

one can validate whether this goal has been achieved [*Hamill*, 2001]. Since climate projections cannot be verified directly, the arguments for the different interpretations are more circumstantial. The 'truth plus error' paradigm appears to be rooted in how models are developed: each group tries to replicate observations in their model, making somewhat different and ideally independent choices. Support for this view comes from the fact that the multi-model mean is often closer to observations than any individual ensemble member [*Gleckler et al.*, 2008; *Knutti et al.*, 2010a; *Lambert and Boer*, 2001; *Schaller et al.*, 2011]. Part of this effect may arise from the centralising tendency from tuning models towards a particular observation [*Yokohata et al.*, 2011] but another part arises by construction, since the squared error of the ensemble mean (relative to any reference) is never greater than the mean of the squared errors of each member (Cauchy-Schwartz inequality) [*Annan and Hargreaves*, 2011]. In other words, the model mean must perform better than the models *on average.* But in many cases the model mean beats *every* model, i.e., the effect is larger than if the reference was chosen randomly. *Annan and Hargreaves* [2011] find that in an idealized indistinguishable ensemble, the result is dependent on both the relative widths of the distributions from which the 'observations' and models are taken as well as the dimensionality of the space. They find cases where the CMIP3 ensemble mean is closer to observations than any other model to be unexceptional considering similar results when other models are treated as truth. An EOF based analysis of a small number of output fields in the CMIP3 models is used to calculate an effective dimension for the ensemble, by examining the rate of decay of variance in ensemble modes of variability. They obtain a dimensionality of 5 to 7 – which they use to show consistency with the number of nearer neighbors found in their idealized indistinguishable case with similar dimensionality. The authors do note that the analysis of *Gleckler et al.* [2008] with a more comprehensive set of diagnostics implies far fewer nearer neighbors, implying a more over-dispersive ensemble.

[5] However, the idealizations employed in *Annan and Hargreaves* [2011] introduce some conceptual difficulties. The multi-model ensemble contains some models with strong similarities (such as single models at different resolutions), which tend to pair together in any 'nearest neighbor' type analysis – thus biasing any interchangeability argument which treats certain models as truth. Secondly, this type of effect also tends to reduce the number of 'effective models' in the ensemble – since some models are, in effect, near duplicates. If the effective ensemble size is significantly smaller than the apparent size, the sampling error will increase (or worse, the apparent mean will be biased by the error in the replicated models). Finally, as the authors point out, in an idealized indistinguishable ensemble with the effective dimensionality of CMIP3, the 'truth' having no nearer neighbors than the multi-model mean would be unexceptional . However, by definition that makes this metric an inappropriate test of truth-centeredness. It shows that a case where the multi-model mean is closest to observations can appear in an indistinguishable ensemble, and thus does not exclude that interpretation, but also does not exclude a truth-centered ensemble.

2. Evidence for Truth-Centered Behavior in CMIP

[6] We propose instead that a metric of model centrality which relies on the complete distribution of distances from each model to all other models in the ensembles (see auxiliary material) might be less susceptible to the model-interdependency and sampling issues listed above.[1] In Figure 1a, we show that if one jointly considers a sufficiently large number of variables, the CMIP3 and CMIP5 historical

[1]Auxiliary materials are available in the HTML. doi:10.1029/2012GL052665.

simulations appear to be approximately centered on observed values (more so than any other model in the ensemble, but not dramatically so). The analysis, described in detail in the auxiliary material, uses a distance metric based on a multivariate EOF analysis to define inter-model distances as well as model-observation distances. The mean model-observation distance is smaller than the mean distance when any particular model is treated as truth, despite the fact that many of the models have close relatives in the ensemble with very small inter-model distances. The near-neighbor analysis also shows the observations to be closer to the multi-model mean than to any other model. However, as noted by *Annan and Hargreaves* [2011], this is hardly conclusive given that in our analysis, 5 models in the combined CMIP3/5 ensemble are also closer to the multi-model mean than they are to any other model.

[7] Tuning to observations may explain part of this apparent truth-centeredness: each model has limited degrees of freedom with which to match observations, resulting in an irreducible error even after optimal tuning. But to the extent that the structural errors are independent across the ensemble, then the multi-model mean can be closer to the observations than any individual member. Structural errors that are common to all models would remain even for a large ensemble and perfect tuning. Structural errors may be common to all models (e.g., limited resolution) or model specific (e.g., assuming fixed vegetation vs. dynamic vegetation). They are dependent both on the structural form of the model and the implicit choice of model performance metrics [*Knutti et al.*, 2010a] chosen when tuning the model.

[8] We can show some evidence of this behavior by considering the rank of the observations for different variables – some of which are more likely to be tuned due to availability of observations and ease of measurement. To somewhat address the issues of the independence for the rank histogram, we evaluate the rank of the observations in an EOF space, such that each rank represents an independent measure of inter-model spread. We truncate the EOFs to describe 85 percent of the total ensemble variance, which ensures the overall shape of the histogram is most reflective of the leading modes of model differences in the ensemble. The results show that for all variables considered (Figures 1b–1d) the ensemble appears somewhat overdispersive (i.e., partly truth-centered, indicated by a tendency for the rank of the observations to be clustered towards the center of the histogram). A rank histogram produced using a multi-variate EOF (Figure 1e) using all available diagnostics also appears overdispersive. To state this quantitatively, rather than a complex decomposition of the chi-squared statistic, we propose a simple test of truth centeredness based on the rank histogram by counting the frequency of the observations lying between the 25th and 75th percentiles of the distribution and subtracting the frequency outside this range. If the distribution were flat, we would expect this value to be near zero. However, we find that in the all variable case (where there are 13 truncated modes), the value is +9 (Figure 1e). A perfect model study, with each single ensemble member treated as truth, shows that this value is not exceeded within the ensemble (although a single model has an equal score of 9 and the ensemble average, by construction is 0).

[9] The discrepancy between these results and those of *Annan and Hargreaves* [2011], which defend the indistinguishable interpretation, can be explained if one considers the different sensitivities of the two techniques. A near-neighbor analysis such as by *Annan and Hargreaves* [2011] finds the observations to be indistinguishable from truth because there exist other models in the ensemble which are comparably close to the mean of the remaining models. However, the statistics of such an analysis are dominated by a few `good' models which lie close to the observations and is largely unaffected by the behavior of outliers. In the analysis presented above, however, the mean inter-model distance would tend to be dominated by the outliers. Hence, a plausible interpretation is that the observations are largely indistinguishable from a subset of high-performing models within the ensemble, but the ensemble as a whole appears weakly truth centered because there are a significant number of outliers with largely independent model errors.

[10] Although the unweighted multi model mean has been the default choice in the Intergovernmental Panel on Climate Change (IPCC) Assessment Reports, uncertainties in IPCC reports were largely based on a spread of models, irrespective of the size of the ensemble, thus assuming an 'indistinguishable' view for the future. The implicit justification for the latter is that each model is a plausible representation of the system given our incomplete understanding of the processes, limited computational capacity, uncertainty in observations, and noise from natural variability [*Knutti*, 2008; *Knutti et al.*, 2010a; *Parker*, 2006]. If, at least for a very large number of models, uncertainty is dominated by common structural limitations and observation uncertainties, then errors cannot be expected to cancel, making the 'truth plus error' interpretation difficult to defend for projections.

3. Reconciling the Two Paradigms

[11] The above discussion and the recent studies [e.g., *Annan and Hargreaves*, 2010] imply that the two interpretations are mutually exclusive. Here we argue that the CMIP3 and CMIP5 ensembles (and in fact any ensemble constrained with data) may have elements of both interpretations that are not contradictory. Consider a very broad ensemble of independent models (assuming no common structural model error for the moment) where the parameters of each are adjusted to optimize their respective performance relative to an observed `truth'. Because remaining errors are independent, we obtain *by design* a truth centered ensemble. In other words, with independent decisions in model development and performance metrics, the truth-centered interpretation is more appropriate. However, there is structural error common across models, so for a large and optimally tuned ensemble the model average would converge to "truth plus common structural error". Also, the number of models is small, CMIP models are not developed independently [*Masson and Knutti*, 2011] and optimal calibration is difficult due to computational cost. Each of these effects will tend to reduce the truth-centered behavior in the ensemble.

[12] Common structural error may be difficult to separate from intermodel differences, i.e., "truth plus error plus structural error" may have similar properties to "indistinguishable" in simple tests using correlations or root mean square errors. Correlated errors and the bias of a model mean not decreasing quickly with more models thus does not exclusively support the interpretation of an "indistinguishable" ensemble, as argued by *Annan and Hargreaves* [2010], but can just as plausibly be the result of structural errors

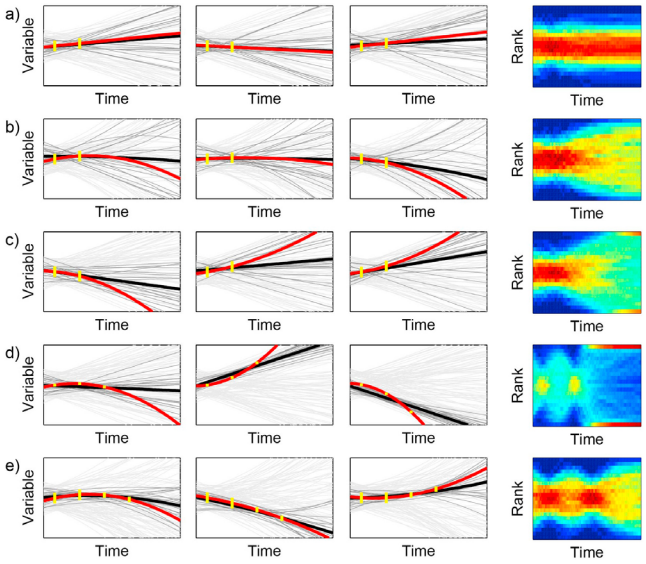

Figure 2. Illustrations of an unconstrained (light grey) and a constrained (dark grey, mean in black) ensemble of a toy model. Observational uncertainties are marked by yellow ranges. Columns one to three show randomly chosen realizations. Column four shows the density of a rank histogram of the model mean, calculated at every time step and averaged over many realizations. Yellow to red colors in the center indicate high probability for the truth (red) to be near the center of the dark grey ensemble, ('truth plus error' interpretation), while uniform light blue color indicates that the ensemble member equally likely to be anywhere in the range of the individual responses ('indistinguishable' interpretation). Red colors near the edges imply that the truth is outside the model range, i.e., then ensemble is overconfident. Truth/models are (a) linear (L)/linear (L), (b) quadratic (Q)/quadratic (Q), (c, d) Q/L and (e) Q/Q.

common to most models. Hence we argue that conceptually, an ensemble of historical simulations constrained with observations *by design* is truth centered to some degree for variables which is used in the tuning process, although it might *look like* "indistinguishable" in a statistical analysis due to structural error, a small sample, and limited model calibration and degrees of freedom to tune to observations. For long term climate prediction, where the observations of the present no longer strongly constrain the response, each ensemble is a plausible future and the indistinguishable interpretation is appropriate. Constraints of the present day climate on future projections are often weak (see below) [*Knutti et al.*, 2010a], thus the 'indistinguishable' interpretation to infer model uncertainty as in IPCC is probably useful for future projections, in agreement with the arguments made by *Annan and Hargreaves* [2010].

4. An Illustrative Toy Model

[13] In Figure 2 we show several variations of a toy model to illustrate the argument. A large ensemble (thin grey lines) is produced, 'truth' is randomly selected from the ensemble, and observations (yellow bars) are used to select a subset of 20 simulations "consistent" with observations (bold grey lines). The likelihood for accepting a simulation for the "consistent" subset is proportional to $\exp(-x^2)$, where x^2 is the sum of squared distances from the observation mean divided by the observation uncertainty. In Figure 2a all models are linear and truth is always close to the ensemble mean. Three cases are shown for illustration. The rightmost column shows a density plot formed by a rank histogram at every time step, based on many samples. Yellow to red colors near the center imply that the truth is much more likely to be near the ensemble mean, i.e., the ensemble is largely truth centered, while a uniform density means that the truth is equally likely to be anywhere in the ensemble, as it would be in an 'indistinguishable' paradigm. Because models and truth in case (a) are linear, it is constrained for all forecast lead times by the observations and the ensemble remains largely truth-centered. In contrast, in Figure 2b the models are quadratic and underconstrained, and the rank histogram shows that the ensemble is truth centered in the beginning, but then gradually transitions to 'indistinguishable'. In the third case (Figure 2c) the observations are quadratic but the models are linear. This is a simple analogue to the models being underconstrained and structurally wrong, i.e., we have a "truth plus error plus structural error" as would occur if a feedback process existed in reality, but was absent from the models. The fourth case is similar to the third but the models are overconstrained. This makes the ensemble look similar to 'indistinguishable', but in reality it is the result of structural error preventing perfect tuning. For a short time the ensemble becomes 'indistinguishable' but then we observe a third state, the 'overconfidence', in which truth is mostly outside the model range as a result of structural error. In this case the ensemble becomes simply uninformative. Finally, it is important to note that the transition from 'truth plus error' to 'indistinguishable' only occurs if the future response is unconstrained by the observations. In Figure 2e, starting from Figure 2b two additional observations (yellow bars) are added. The ensemble now remains truth centered for a longer time. Therefore, at least in principle, if more observations become available which enable parameters of the model to be directly constrained, the ensemble can remain truth centered, if no significant structural error is present.

[14] Clearly the distribution for both the constrained period and the future is somewhat dependent on the prior distribution of possible models , and a prior which was not uniform in the observed domain would results in an posterior which was not exactly truth centered [*Yokohata et al.*, 2011]. However, the toy example provides a simple visualization of the argument that two seemingly irreconcilable statistical interpretations can apply to the same ensemble, although for different lead times or for different variables.

5. Interpreting the Spread of CMIP

[15] An important related question is whether the CMIP ensemble spread is too narrow (i.e., drawn from a smaller distribution than the hypothetical spread of possible futures), about right (drawn from an appropriate distribution), or too broad. Although there is no way to properly test this for the future due to the lack of a large verification ensemble, analysis of the model spread for the present day as in Figure 1 or by *Annan and Hargreaves* [2010] is interesting but of limited use, because present day spread is likely dominated by structural limitations (e.g., imperfect physics,

limited resolution) which prevent perfect tuning to observations, whereas model spread in the future arises due to different representations of physical processes and feedback mechanisms. Indeed, correlations between the simulated current and future climate are generally weak [*Knutti et al.*, 2010a]. The ensemble may be adequate for reproducing current climate but overconfident for projections if key uncertainties are not adequately sampled. Second, if the CMIP spread was too large, we should be able to reduce it by conditioning on observations. Except for a few cases (e.g., for the Arctic) [*Boe et al.*, 2009; *Hall and Qu*, 2006; *Mahlstein and Knutti*, 2011] such constraints are rarely found, implying that any outcome in the ensemble range is as plausible as the others, often referred to as 'model democracy' [*Knutti*, 2010]. Third, in some cases the ensemble is known to underestimate spread, for example due to unresolved processes and feedbacks (e.g., dynamic vegetation, ice sheets, methane hydrates), or because some known uncertainties are not sampled (e.g., the carbon cycle). Simplifications and parameterizations common to many models (e.g., limited resolution, imperfect numerical schemes) also favor the interpretation of an overconfident ensemble. Clearly, the conclusions outlined here are applicable to the current generation of climate models. The structural limitations in models, the resolution, complexity and adequacy of parameterizations can reasonably be expected to improve in the future as it has done in the past [*Reichler and Kim*, 2008].

[16] In summary, model spread in CMIP3 or CMIP5 could be interpreted as an emerging property of an ensemble that has been calibrated and constrained with observations. In that case, the resulting distribution of the ensemble of opportunity would be the posterior distribution given the data produced from a poorly planned Bayesian experiment. Such an interpretation is conceptually interesting but probably overly optimistic, because model calibration in this high-dimensional parameter space is difficult, models are dependent and the ensemble is small. We argue that the lack of correlation between observables and predictions in CMIP3 and CMIP5 could partly arise because the currently available observations have already largely been exploited in the tuning and model development process. When combined with structural errors in each model which prevent perfect tuning, this does explain the difficulty in further reducing the model spread in CMIP3 and CMIP5.

6. Conclusion

[17] Here we offer a solution to reconcile two seemingly inconsistent interpretations of model ensembles that have caused debate in the climate modeling community. Conceptually, we argue that in a constrained ensemble there can be elements of the 'truth plus error' and 'indistinguishable' paradigms. The tuning process should introduce a tendency for models to be centered on the observations and we show that the CMIP ensembles tend to show some truth-centered behavior in their historical simulations for variables which are commonly used as tuning metrics. Structural errors and limited degrees of freedom prevent the models from being tuned to match the observations exactly, and any structural errors (e.g., limited resolution) which are common amongst models will tend to shift the ensemble mean to 'truth plus common structural error'. Such effects can potentially introduce 'indistinguishable' features when simulating observed climate, but do not justify the use of this paradigm to explain the spread of models in historical simulations which we know have been somewhat tuned to best replicate the past climate. It is thus expected that the multi-model mean of historical simulations might lie closer to observations than would be expected in a truly indistinguishable ensemble, but not so close that the ensemble can be described as perfectly truth-centered. Hence, in reality neither framework is entirely representative of the weakly truth centered CMIP ensembles that we appear to see.

[18] Structural model error may be difficult to separate from intermodel differences in simple statistical tests, even for the present day, but it is even harder for a prediction where no direct verification is available, and skill must be established indirectly [*Knutti*, 2008; *Tebaldi and Knutti*, 2007]. The transition from a 'truth plus error' to 'indistinguishable' or 'overconfident' interpretation reflects the increasing dominance of model responses which are uncorrelated to observable quantities within the ensemble, or to initial model perturbations (i.e., the prediction or projections). We argue that it is not obvious how the interpretation of CMIP for the present can be transferred to projections. Ensemble spread in the future is increasingly dominated by different representations of physical processes and feedback mechanisms so although the future simulations are truly indistinguishable, the spread of present day simulations tells us little about whether the ensemble is under- or overdispersive in the future.

[19] Only in the special case in which a relationship exists between observables and future response can the ensemble remain truth-centered in the future. Otherwise, the observations have already been used and are no longer useful to distinguish between the individual members. The above conceptual arguments are illustrated here for climate model ensembles and a toy models, but apply to any set of different numerical models calibrated to observations.

[20] Model spread in ensembles like CMIP3 or CMIP5 may be too large if data is not fully used to tune each member, or if model structural errors are large. It may be too small if all models are structurally similar but incomplete, or if models do not sample the uncertainty in observations. If uncertain processes are missing, the spread is also likely to be underestimated (e.g., carbon cycle uncertainties). If the ensemble spread was much too large, we should be able to reduce it by weighting models or selecting a subset of them [*Knutti*, 2010]. The fact that we are unable to reduce the spread, and largely unable to agree on appropriate metrics to do so suggests that the ensemble spread in CMIP3 is not strongly overestimated, and the interpretation of the posterior as some sort of lower bound on model uncertainty is not unreasonable. So our inability to find constraints within CMIP3 should maybe not be interpreted as a failure but as a success in converging to an ensemble of (almost) equally plausible models given limitations in observations, our understanding, and computational capacity.

[21] **Acknowledgments.** This research was supported by the Office of Science (BER), U.S. Department of Energy (DOE). We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

[22] The Editor thanks the two anonymous reviewers for assisting in the evaluation of this paper.

References

Annan, J. D., and J. C. Hargreaves (2010), Reliability of the CMIP3 ensemble, *Geophys. Res. Lett.*, *37*, L02703, doi:10.1029/2009GL041994.

Annan, J. D., and J. C. Hargreaves (2011), Understanding the CMIP3 multimodel ensemble, *J. Clim.*, *24*, 4529–4538, doi:10.1175/2011JCLI3873.1.

Boe, J. L., A. Hall, and X. Qu (2009), September sea-ice cover in the Arctic Ocean projected to vanish by 2100, *Nat. Geosci.*, *2*(5), 341–343, doi:10.1038/ngeo467.

Buser, C. M., H. R. Kunsch, D. Luthi, M. Wild, and C. Schar (2009), Bayesian multi-model projection of climate: Bias assumptions and interannual variability, *Clim. Dyn.*, *33*(6), 849–868, doi:10.1007/s00382-009-0588-6.

Furrer, R., R. Knutti, S. R. Sain, D. W. Nychka, and G. A. Meehl (2007), Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis, *Geophys. Res. Lett.*, *34*, L06711, doi:10.1029/2006GL027754.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux (2008), Performance metrics for climate models, *J. Geophys. Res.*, *113*, D06104, doi:10.1029/2007JD008972.

Hall, A., and X. Qu (2006), Using the current seasonal cycle to constrain snow albedo feedback in future climate change, *Geophys. Res. Lett.*, *33*, L03502, doi:10.1029/2005GL025127.

Hamill, T. M. (2001), Interpretation of rank histograms for verifying ensemble forecasts, *Mon. Weather Rev.*, *129*(3), 550–560, doi:10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

Jackson, C. S., M. K. Sen, G. Huerta, Y. Deng, and K. P. Bowman (2008), Error reduction and convergence in climate prediction, *J. Clim.*, *21*(24), 6698–6709, doi:10.1175/2008JCLI2112.1.

Knutti, R. (2008), Should we believe model predictions of future climate change?, *Philos. Trans. R. Soc. A*, *366*, 4647–4664, doi:10.1098/rsta.2008.0169.

Knutti, R. (2010), The end of model democracy?, *Clim. Change*, *102*(3–4), 395–404, doi:10.1007/s10584-010-9800-2.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl (2010a), Challenges in combining projections from multiple climate models, *J. Clim.*, *23*(10), 2739–2758, doi:10.1175/2009JCLI3361.1.

Knutti, R., G. Abramowitz, M. Collins, V. Eyring, P. J. Gleckler, B. Hewitson, and L. O. Mearns (2010b), Good practice guidance paper on assessing and combining multi model climate projections, in *Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections*, edited by T. F. Stocker et al., pp. 1–15, Univ. of Bern, Bern.

Lambert, S. J., and G. J. Boer (2001), CMIP1 evaluation and intercomparison of coupled climate models, *Clim. Dyn.*, *17*, 83–106, doi:10.1007/PL00013736.

Lopez, A., C. Tebaldi, M. New, D. A. Stainforth, M. R. Allen, and J. A. Kettleborough (2006), Two approaches to quantifying uncertainty in global temperature changes, *J. Clim.*, *19*, 4785–4796, doi:10.1175/JCLI3895.1.

Mahlstein, I., and R. Knutti (2011), Ocean heat transport as a cause for model uncertainty in projected Arctic warming, *J. Clim.*, *24*(5), 1451–1460, doi:10.1175/2010JCLI3713.1.

Masson, D., and R. Knutti (2011), Spatial-scale dependence of climate model performance in the CMIP3 ensemble, *J. Clim.*, *24*(11), 2680–2692, doi:10.1175/2011JCLI3513.1.

Meehl, G. A., C. Covey, B. McAvaney, M. Latif, and R. J. Stouffer (2005), Overview of the coupled model intercomparison project, *Bull. Am. Meteorol. Soc.*, *86*, 89–93, doi:10.1175/BAMS-86-1-89.

Murphy, J. M., B. B. B. Booth, M. Collins, G. R. Harris, D. M. H. Sexton, and M. J. Webb (2007), A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles, *Philos. Trans. R. Soc. A*, *365*, 1993–2028, doi:10.1098/rsta.2007.2077.

Parker, W. (2006), Understanding model pluralism in climate science, *Found. Sci.*, *11*, 349–368, doi:10.1007/s10699-005-3196-x.

Pennell, C., and T. Reichler (2011), On the effective number of climate models, *J. Clim.*, *24*(9), 2358–2367, doi:10.1175/2010JCLI3814.1.

Perkins, S. E., A. J. Pitman, N. J. Holbrook, and J. McAneney (2007), Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions, *J. Clim.*, *20*(17), 4356–4376, doi:10.1175/JCLI4253.1.

Reichler, T., and J. Kim (2008), How well do coupled models simulate today's climate?, *Bull. Am. Meteorol. Soc.*, *89*(3), 303–311, doi:10.1175/BAMS-89-3-303.

Schaller, N., I. Mahlstein, J. Cermak, and R. Knutti (2011), Analyzing precipitation projections: A comparison of different approaches to climate model evaluation, *J. Geophys. Res.*, *116*, D10118, doi:10.1029/2010JD014963.

Smith, R. L., C. Tebaldi, D. Nychka, and L. O. Mearns (2009), Bayesian modeling of uncertainty in ensembles of climate models, *J. Am. Stat. Assoc.*, *104*(485), 97–116, doi:10.1198/jasa.2009.0007.

Tebaldi, C., and R. Knutti (2007), The use of the multi-model ensemble in probabilistic climate projections, *Philos. Trans. R. Soc. A*, *365*, 2053–2075, doi:10.1098/rsta.2007.2076.

Tebaldi, C., and B. Sanso (2009), Joint projections of temperature and precipitation change from multiple climate models: A hierarchical Bayesian approach, *J. R. Stat. Soc.*, *172*, 83–106, doi:10.1111/j.1467-985X.2008.00545.x.

Tebaldi, C., R. W. Smith, D. Nychka, and L. O. Mearns (2005), Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multi-model ensembles, *J. Clim.*, *18*, 1524–1540, doi:10.1175/JCLI3363.1.

Yokohata, T., J. D. Annan, M. Collins, C. S. Jackson, M. Tobis, M. J. Webb, and J. C. Hargreaves (2011), Reliability of multi-model and structurally different single-model ensembles, *Clim. Dyn.*, *39*, 599–616.