Predictor Screening, Calibration, and Observational Constraints in Climate Model Ensembles: An Illustration Using Climate Sensitivity

DAVID MASSON

Institute for Atmospheric and Climate Science, ETH Zurich, and Federal Office of Meteorology and Climatology, MeteoSwiss, Zurich, Switzerland

RETO KNUTTI

Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland

(Manuscript received 16 November 2010, in final form 15 July 2012)

ABSTRACT

Climate projections have been remarkably difficult to constrain by comparing the simulated climatological state from different models with observations, in particular for small ensembles with structurally different models. In this study, the relationship between climate sensitivity and different measures of the present-day climatology is investigated. First, it is shown that 1) a variable proposed earlier that is based on interannual variation of seasonal temperature and 2) the seasonal cycle amplitude are unable to constrain the range of climate sensitivity beyond what was initially covered by the ensemble. Second, it is illustrated how model calibration helps to reveal potentially useful relationships but might also complicate the interpretation of multimodel results. As a consequence, when ensembles are small, when models are neither independently developed nor structurally identical, when observations are likely to have been used in the model development and evaluation process, and when the interpretation of the relationships across models in terms of well-understood physical processes is not obvious, care should be taken when using relationships across models to constrain model projections. This study demonstrates the pitfalls that might occur if emergent statistical relationships are prematurely interpreted as an effective constraint on projected global or regional climate change.

1. Introduction

General circulation models (GCMs) are tools that can be used to understand climate processes and to make climate projections for the next decades to centuries. The discretization of the equations of motion on a grid is subject to several choices (resolution, numerical schemes, hardware and software, etc.) and the need for parameterizations leads to structural differences and uncertainties that are difficult to explore fully by perturbing parameters in a single model (Knight et al. 2007; Stainforth et al. 2007) . Yet such perturbed physics ensembles (PPEs) have undoubtedly been an interesting playground over the past years to test ideas on how model parameters can be constrained by observations and to study methodological issues. The availability of many ensemble members with perturbed parameters has led to a variety of studies exploring, in particular, the relationship between metrics of the present-day climate that can be observed, and climate sensitivity (the global mean equilibrium surface warming for a doubling of the atmospheric carbon dioxide concentration) as a rough proxy for the magnitude of climate change predicted by a model (Knutti et al. 2006; Piani et al. 2005; Sanderson et al. 2008b,a, 2010; Stainforth et al. 2005). However, it is becoming increasingly clear that many (if not all) perturbed versions of a particular base model can share certain structural similarities (Collins et al. 2011; Masson and Knutti 2011), and relationships derived from a PPE may be different in other structurally different models (Sanderson 2011; Sanderson and Shell 2012; Yokohata et al. 2010). Thus, establishing relationships between observables and predictions that are valid across a range of structurally different models is important. At present, the data from phase 3 of the Coupled Model Intercomparison (CMIP3; Meehl et al.

Corresponding author address: David Masson, MeteoSwiss, Krähbühlstrasse 58, 8044 Zurich, Switzerland.
E-mail: massond@phys.ethz.ch

DOI: 10.1175/JCLI-D-11-00540.1

2007), collected for the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4; Solomon et al. 2007), forms the largest and most complete group of structurally different global models. However, this ensemble contains only 24 different models, and these models are not totally independent (Jun et al. 2008a,b; Masson and Knutti 2011). A number of studies have pointed out the difficulty of finding robust and strong relationships between observables and predictions in CMIP3 (Jun et al. 2008a,b; Knutti 2010; Knutti et al. 2010a,b; Raisanen et al. 2010). Notable exceptions, where the relationships are quite clear and the physical processes are well understood, include the high-latitude albedo feedbacks, the climate of the Arctic, and the decline in Arctic sea ice (Boé et al. 2009c,b,a; Mahlstein and Knutti 2011, 2012). Several studies have also attempted to constrain climate sensitivity from the CMIP models (Huber et al. 2011; Shukla et al. 2006; Wu and North 2003; Wu et al. 2008). In most cases, the observationally-constrained range is similar to the climate sensitivities originally covered by the models. One issue in using CMIP-type ensembles is that the effective number of independent models is much smaller than what the ensemble suggests (Annan and Hargreaves 2011; Jun et al. 2008a,b; Masson and Knutti 2011). The difficulty of finding robust relationships in a high-dimensional output space with only a dozen or so models in the sample is obvious, and the danger of screening predictors has been pointed out (DelSole and Shukla 2009). In this study, we analyze relationship between temperature interannual variability, the seasonal cycle, and the mean climate state on the one hand, and climate sensitivity on the other hand in both the structurally different models of CMIP3 and two perturbed physics ensembles of the Hadley Centre model. We demonstrate that such relationships may differ across ensembles, and may be related to whether and how the ensemble is constrained with observations.

2. Models and data

The data consist of simulated and observed monthly surface air temperature fields from preindustrial control experiments with no external forcing. Three sets of fully coupled ocean–atmosphere GCMs are used in this study. The first set belongs to phase 3 (Meehl et al. 2007) of the World Climate Research Program (WCRP) Coupled Model Intercomparison, a coordinated project to gather and compare about 24 different GCMs for the IPCC AR4 (Solomon et al. 2007). The second set is the Atmosphere–Ocean PPE model ensemble with perturbed atmospheric parameters (AO-PPE-A) from the ''Quantifying Uncertainty in Model Predictions'' (QUMP) experiment, which was generated by perturbing the atmospheric parameters of the third climate configuration of the Met Office Unified Model (HadCM3; Collins et al. 2011). This ensemble (simply referred as the ''QUMP ensemble'' in this publication) contains 17 simulations that do not result from a random perturbation of a base model but are constrained by climatology (Murphy et al. 2004; Webb et al. 2006) in much the same way that CMIP3 is designed to agree reasonably well with observations. The third ensemble is known as the climateprediction.net (CPDN) ensemble and consists of several thousand simulations that were designed to explore parametric uncertainty in a single model (Frame et al. 2009; Rowlands et al. 2012). More than 50 parameters of the HadCM3L coupled model (the HadCM3 model with a reduced ocean resolution) were perturbed in a range determined plausible by experts. All preindustrial control runs with at least 160 simulated years were selected. This includes the HadCM3L coupled model experiment and the British Broadcasting Corporation (BBC) Climate Change Experiment (Frame et al. 2009). In total, 4846 CPDN simulations were used. In contrast to CMIP3 and QUMP, this ensemble is not strongly constrained by observations.

The observation-based datasets for surface air temperature are the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40; Uppala et al. 2005), the National Centers for Environmental Prediction (NCEP)–National Center for Atmospheric Research (NCAR) reanalysis (Kalnay et al. 1996), the third revision of the Hadley Centre Climatic Research Unit instrumental temperature records (HadCRUT3) (Brohan et al. 2006), and the Modern-Era Retrospective Analysis for Research and Applications (MERRA) (Rienecker et al. 2011). Because individual models and observations come at different spatial resolutions, the data are bilinearly interpolated to a common T42 grid (i.e., a Gaussian grid associated with spectral truncation having 128 latitudinal by 64 longitudinal grid points).

A direct calculation of the CMIP3 climate sensitivity values is computationally expensive because the climate system has to reach a thermal equilibrium. Rather than simulating a fully coupled transient evolution of the ocean, which takes several centuries to equilibrate with the surface forcing, a more efficient method is to use an atmospheric GCM coupled to a slab ocean, which yields a slab ocean equilibrium climate sensitivity. Another method regresses the radiative flux at the top of the atmosphere against the global average surface temperature change to produce an ''effective climate sensitivity''(Gregory et al. 2004). To maximize the number of available climate sensitivity values, the climate sensitivity definition used here for the CMIP3 ensemble is the mean value of the two methods (or one

TABLE 1. List of the GCMs used, the length of their control simulations, and their climate sensitivity. See the text for more details.

Model	Length (yr)	Climate sensitivity (K)
Canadian Centre for Climate Modelling and Analysis (CCCma) Coupled General Circulation Model, version 3.1 (cccma_cgcm3_1)	1001	3.21
CCCma CGCM 3.1, T63 resolution (cccma_cgcm3_1_t63)	350	3.4
Centre National de Recherches Météorologiques (CNRM) Coupled Global Climate Model, version 3 (cnrm_cm3)	500	2.45
Commonwealth Scientific and Industrial Research Organisation (CSIRO) Mark, version 3.0 (csiro_mk3_0)	380	2.65
Geophysical Fluid Dynamics Laboratory (GFDL) Climate Model, version 2.0 (gfdl_cm2_0)	500	2.62
GFDL Climate Model, version 2.1 (gfdl_cm2_1)	500	2.84
Goddard Institute for Space Studies (GISS) Model E-H (giss_model_e_h)	400	2.87
GISS Model E-R (giss_model_e_r)	500	2.63
Institute of Atmospheric Physics (IAP) Flexible Global Ocean–Atmosphere–Land System (FGOALS) Model, gridpoint version 1.0 (iap_fgoals1_0_g)	350	2.13
Institute of Numerical Mathematics Coupled Model (INCM), version 3.0 (inmcm3_0)	330	2.19
L'Institut Pierre-Simon Laplace (IPSL) Coupled Model, version 4 (ipsl_cm4)	320	4.11
Model for Interdisciplinary Research on Climate (MIROC) 3.2, high-resolution version (miroc3_2_hires)	100	5.08
MIROC 3.2, medium-resolution version (miroc3_2_medres)	500	3.96
Meteorological Institute University of Bonn (MIUB) ECHAM and the global Hamburg Ocean Primitive Equation (miub_echo_g)	341	3.10
Max Planck Institute (MPI) ECHAM5 (mpi_echam5)	506	3.63
Meteorological Research Institute (MRI) Coupled General Circulation Model (CGCM), version 2.3.2a (mri_cgcm2_3_2a)	350	3.08
NCAR Community Climate System Model (CCSM), version 3 (ncar_ccsm3_0)	230	2.53
NCAR Parallel Climate Model (PCM) (ncar_pcm1)	350	1.99
Met Office (UKMO) Hadley Centre Unified Model (HadCM), third configuration (ukmo_hadcm3)	341	3.18
Met Office Hadley Centre Global Environmental Model (HadGEM), version 1 (ukmo_hadgem1)	240	3.51
climateprediction.net (CPDN) (HadCM3L)	160	0.13–9.27
Quantifying Uncertainty in Model Predictions (QUMP) (HadCM3)	80	2.2–6.04

of the methods if only one is available). The results do not depend on this choice. Finally, the CPDN equilibrium climate sensitivity values were estimated for 152 physics perturbation categories from a separate ensemble slab model experiment (i.e., using the atmospheric GCM coupled to a slab ocean; Stainforth et al. 2005). Hence, the number of possible climate sensitivity values is much less than the ensemble size. This is due to the setup of the BBC Climate Change Experiment, which contains, for example, several ocean and forcing versions for the same atmospheric perturbation set. While these permutations do not affect the value of the climate sensitivity much, they impact other properties such as interannual variability. Table 1 summarizes which GCMs have been used, the length of their control simulations, and their climate sensitivity values. The CPDN dataset is not used in all calculations, since some output fields are not available at the required resolution. Note that the uncertainties in estimating climate sensitivity from the short CPDN simulations are significant.

3. Constraining climate sensitivity from observed interannual variability

We start with a constraint on future climate change that was based on temperature variability and was suggested a few years ago. Natural variability is a crucial variable for detection and attribution studies of climate change (Barnett et al. 2005; Wigley et al. 1998), and climate model development is often focused as much on

getting an adequate representation of variability as getting the mean state. Wu and North (2003) have reported a relationship between interannual variability and climate sensitivity in an earlier set of GCMs. The idea behind Wu and North (2003) relies on the cyclo-stationarity of interannual variability (Kim and Wu 2000)—the fact that in a control climate each calendar month has its own interannual variability that can be considered constant despite being different from the 11 other months. Monthly mean surface temperatures are considered first at the gridpoint scale. The interannual variability is individually calculated at each grid point and for each month. Then, the global average of monthly interannual variability is calculated for each CMIP3 model and the observational and reanalysis datasets. Note that the globally averaged variability is higher during the boreal winter than during the boreal summer. The larger winter variability is due to different landmass distributions in the Northern and the Southern Hemispheres. Because land regions have a smaller heat capacity than the ocean regions, synoptic weather systems can generate larger temperature variations over the Northern Hemisphere where more land exists compared to the Southern Hemisphere. Another factor that increases variability at high latitudes is the albedo change caused by snow cover (Kumar and Yang 2003) and sea ice variability. The models with the highest and lowest climate sensitivity values show interannual variability that is quite different from the observations, leading to the hypothesis that the shape of the globally averaged variability, as a function of the calendar month, could be used to evaluate some aspects of climate models. Wu and North (2003) proposed quantifying the asymmetry of the calendar month variances σ^2 by the σ_s^2/σ_w^2 ratio, where σ_s^2 is the global average of the smallest variance for the summer months (from June to August) and σ_w^2 is the global average of the largest variance for the winter months (from December to February).

a. Linear prediction

An approximately linear relationship is empirically found in the CMIP3 ensemble between the σ_s^2/σ_w^2 ratio and climate sensitivity and is shown in Fig. 1. The Pearson correlation coefficient is 0.78 and is statistically significant according to a two-sided t test (p value of 4.8×10^{-5}). If the CMIP3 model with the highest climate sensitivity is removed from the analysis, the correlation is still 0.71. This relationship is consistent with the results described for an earlier set of models (Wu and North 2003). Using the linear relationship above, it seems possible to constrain climate sensitivity by using the σ_s^2/σ_w^2 ratio derived from the observational references. Several sources of uncertainty contribute to the width of the final confidence interval of climate sensitivity and are included in the calculation. First, the scatter of the data points makes the optimal linear regression uncertain. The uncertainty of the linear regression is represented in Fig. 1 by the two black dashed curves that correspond to the 95% prediction confidence interval. Second, the real climate is always subject to changing external forcings (anthropogenic, volcanic, and solar) that are not represented in the control simulations used in this study. Alternatively, the σ_s^2/σ_w^2 ratio derived from the CMIP3 transient simulation could be used instead of the control runs. However, this inflates the uncertainty because not all the models simulate the natural forcing. Also, the control simulations are longer and provide a more precise estimate of σ_s^2/σ_w^2. Third, the observations are uncertain because of measurement uncertainty, model uncertainty in the reanalysis procedure, and sampling uncertainty caused by the relatively short length of observational records. The effects of multidecadal changes in modes of natural variability (Hurrell 1996) and nonlinear changes due to external aerosol forcings are neglected. The uncertainty of the observational references σ_s^2/σ_w^2 ratio (marked by horizontal lines in Fig. 1) is estimated in two steps. First, the observational data are detrended for each individual month to

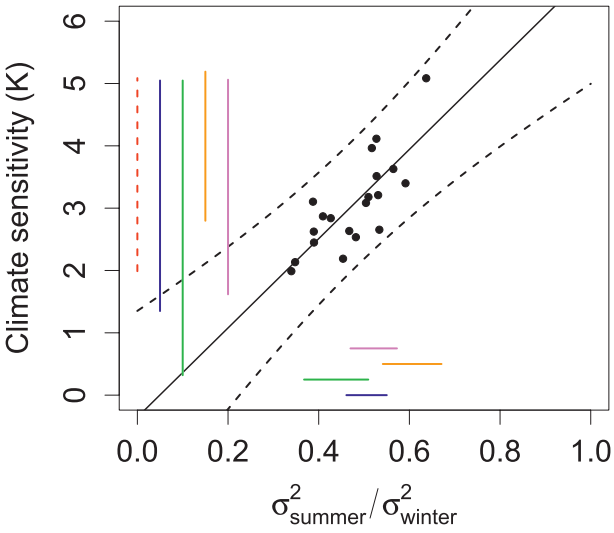

FIG. 1. Scatterplot of CMIP3 climate sensitivity vs the summer-to-winter interannual variability σ_s^2/σ_w^2 ratio. The correlation is 0.78, and the dashed lines represent the 95% prediction confidence interval (CI) for the linear regression. The horizontal colored lines correspond to the 95% CI calculated for the references [ERA-40 (blue), HadCRUT3 (green), NCEP–NCAR (orange), and MERRA (purple)]. The vertical dashed red line is the original CMIP3 minimum to maximum range, and the other vertical colored lines are the 95% CI for the climate sensitivity constrained by the corresponding reference dataset. The climate sensitivity uncertainty range takes into account both the regression and observational uncertainty.

TABLE 2. Constraint on climate sensitivity from the observed interannual variability σ_s^2/σ_w^2 ratio, using the linear relationship within the CMIP3 ensemble shown in Fig. 1. The columns indicate the time period covered, the length of the observational datasets, the best estimate of the predicted climate sensitivity, and the 95% confidence interval (CI).

Dataset	Time period covered (yr)	Length (yr)	Climate sensitivity (K)	95% CI (K)
ERA-40	1958–2001	44	3.25	1.35–5.05
HadCRUT3	1850–2009	159	2.79	0.32–5.05
NCEP–NCAR	1948–2009	61	3.97	2.8–5.2
MERRA	1979–2008	29	3.41	1.62–5.06
Combined	—	—	3.36	1.52–5.09

remove the twentieth-century transient warming. Second, the unforced natural variability of each observational dataset is approximated by bootstrapping 30 individual years (20 for MERRA, since MERRA only has 29 years of data) from the total number of years. The interannual variability is then calculated for the 12 calendar months at each grid point. This procedure is repeated 1000 times to derive the 95% confidence interval for the observational σ_s^2/σ_w^2 ratio. The results are shown as colored horizontal lines in Fig. 1.

The ERA-40 and MERRA reference datasets show σ_s^2/σ_w^2 ratios of similar magnitude and width. The HadCRUT3 dataset uses geographically scattered instrumental data that increase the uncertainty of σ_s^2/σ_w^2. The marine (sea surface temperature) and land (1.5-m temperature) data are blended together, and the station data are interpolated to a common grid. This homogenization artificially brings more variability over grid boxes with fewer observations (Brohan et al. 2006). We speculate that, because fewer observations exist over oceans, the winter variability is overestimated and the HadCRUT3 σ_s^2/σ_w^2 ratio is shifted toward smaller values. The σ_s^2/σ_w^2 confidence interval based on the NCEP–NCAR reanalysis is larger and higher than for the other reference datasets.

In a final step, the uncertainty of the observational reference is combined with the regression uncertainty to derive the 95% confidence interval for climate sensitivity by using a resampling method: For each observational dataset, 1000 possible σ_s^2/σ_w^2 ratios are sampled to predict the climate sensitivity. Then, every sampled σ_s^2/σ_w^2 ratio is used in the linear regression and leads to a distribution of predicted climate sensitivity. Assuming a Gaussian distribution for the predicted values, 1000 climate sensitivity values are sampled for each predictor. Finally, the 95% confidence interval is estimated over an ensemble of 10^6 predicted values. The result is given in Table 2. All reference datasets are assumed to be equally likely, but the conclusions do not depend strongly on this assumption. The best estimate for the climate sensitivity value is the average of the four predictions derived from the four observational references and is equal to 3.4 K. The corresponding 95% confidence interval is 1.5–5.1 K, which is comparable to the ''likely'' range of 2–4.5 K given in the recent IPCC AR4 (Solomon et al. 2007) and a recent review summarizing multiple constraints on climate sensitivity (Knutti and Hegerl 2008).

Despite the fact that a statistical test ensures that the linear relationship is not by coincidence, a significant correlation is not a strong argument for a reliable constraint unless the correlation is extremely strong. To test the robustness, the same experimental design is applied to the QUMP ensemble (shown in Fig. 2) and a significant anticorrelation of -0.66 is found. The anticorrelation makes sense from a physical point of view: a larger albedo feedback implies a larger snow and sea ice variability (Kumar and Yang 2003) and therefore a smaller σ_s^2/σ_w^2 value. It also implies a greater climate sensitivity (Soden and Held 2006). While this interpretation holds for the QUMP ensemble, it does not apply to the CMIP3 case, where a linear relationship with a slope of opposite sign was found. This contradiction challenges the relevance of σ_s^2/σ_w^2 as a consistent predictor. If the apparent correlations are still tied to a physical process, then one would expect a stronger correlation over regions of the globe or over time periods for which the feedback is most relevant. According to Fig. 3, even if the relationship in both ensembles is stronger for land (Fig. 3d) than for ocean (Fig. 3c), as expected, the strength of the relationship in other regions [Northern Hemisphere, high latitudes (Fig. 3b) or the globe without the tropics (Fig. 3e)] never exceeds the original value found for the global case (see Fig. 1). On the other hand, if the relationship has a global character and is related to the seasonal cycle, then shifting the Southern Hemisphere by 6 months to align boreal and austral winter and summer should amplify the correlation. As a matter of fact, no increase is apparent (see Fig. 3f). A large number of additional diagnostics were tested, but we were unable to clearly associate a physical process to the apparent correlations. This leaves the conclusion that some of these correlations occur by chance. The argument that the CMIP3 correlation is significant and similar to an earlier ensemble is not very strong, since models not only share biases and parameterizations over time but also between modeling centers (Masson and Knutti 2011), which reduces the effective sample size from around 20 to probably more like 10 or so.

b. Nonlinear prediction

The same analysis was repeated for the CPDN ensemble in which the σ_s^2/σ_w^2 ratio was computed for ~5000 control simulations. Because of the large number of simulations, one would expect to see a more detailed

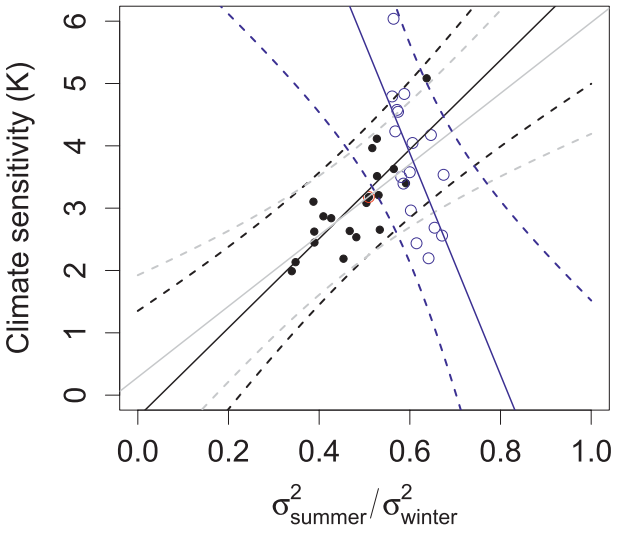

FIG. 2. As in Fig. 1, but the CMIP3 models are shown by black dots, and the QUMP models are shown by blue circles. The anticorrelation between the climate sensitivity and σ_s^2/σ_w^2 is -0.66 for the QUMP ensemble. The gray color stands for the CMIP3 experiment but without the model with the highest climate sensitivity.

picture of the correlation between σ_s^2/σ_w^2 and the CPDN climate sensitivity values. Note that the CPDN simulation output is not available at the full gridpoint resolution but is regionally aggregated and 32 regions were chosen to cover most of the globe surface. As a consequence, the global average of the interannual variability is necessarily less accurate than if computed directly from the gridpoint scale but should still provide useful information. The small gray dots in Fig. 4a show climate sensitivity versus σ_s^2/σ_w^2 for the CPDN ensemble. To make a comparison between CMIP3 and CPDN possible, the CMIP3 and the QUMP ensembles were in turn regionally aggregated for the same regions to ensure consistency, and they are represented by black and blue circles, respectively. The global average of the regional σ_s^2/σ_w^2 ratios is shifted toward smaller values because part of the gridpoint information is lost, but the correlation is still high for both CMIP3 and QUMP. Surprisingly, no linear correlation exists for the CPDN ensemble.

This result might seem disturbing at first, but it does not imply that no relationship exists. To seek a possible relationship, a pattern recognition algorithm is used to predict the CPDN climate sensitivity values with the regional σ_s^2/σ_w^2 ratios as input. The algorithm is called ''random forest'' and is based on multiple regression trees (Breiman et al. 1984). Similar to neural networks, random forest belongs to supervised statistical learning techniques. The random forest is trained to match climate sensitivity values with 60% of the data and the remaining 40% of the data is used for validation. To minimize the risk of overfitting, the training dataset consists of independent simulations chosen in the 152 climate sensitivity categories. The true climate sensitivity versus the sensitivity predicted from the 32 regional σ_s^2/σ_w^2 ratios is shown in Fig. 4b. The prediction explains 69% of the variance in climate sensitivity across the ensemble, with a root-mean-square (RMS) prediction error of 1.24 K.

Apparently, the nonlinear relationship is strong enough to constrain climate sensitivity from the pattern of interannual variability. Such a technique has already been used in the past with the regional seasonal cycle amplitudes as in input to constrain climate sensitivity Knutti et al. (2006). This study could explain 77% of the variance in the set of 40% data not used during the training phase through a neural network algorithm.

Despite high explained variance values, how relevant are such nonlinear predictions when attempting to reduce the uncertainty of future climate change? To address this question, we repeated the experiment done by Knutti et al. (2006). In this experiment, a physical relationship was expected between the seasonal cycle amplitude and climate sensitivity on the basis of previous evidence found in Covey et al. (2000). In contrast to σ_s^2/σ_w^2, no linear relation was observed on a global scale in either the CMIP or the CPDN ensembles (see Fig. 5a). We followed the definition of the seasonal cycle amplitude $|A|$ given in Covey et al. (2000); that is, $|A|$ is the globally averaged absolute value of the July minus January surface air temperature. In the current experimental design, the random forest algorithm explains 72% of the climate sensitivity (see Fig. 5b) with an RMS prediction error of 1.15 K. Thus far, relationships between climate sensitivity and two different measures of variability seem to exist, even if they may be complex and nonlinear. While these predictors are potentially useful, the consistency of the nonlinear relationships should be tested in other ensembles, as done with the linear relationship above. Therefore, the CMIP3 climate sensitivity values are in turn predicted using the nonlinear relationship found in CPDN. The predicted CMIP3 values are indicated in Figs. 4b and 5b as black circles. As expected, the unperturbed HadCM3 simulation belonging to the CMIP3 ensemble model is correctly predicted by the random forest algorithm. But for the rest of the CMIP3 ensemble, it is not clear whether the nonlinear relationship is able to correctly predict the climate sensitivity value. Some simulations lie well outside 1 CPDN standard error, suggesting that these models are structurally too different from the HadCM3L model. The predictive skill of the method can be quantified by using the general formulation of a skill score (SS; Stevenson 2006). The skill is defined by

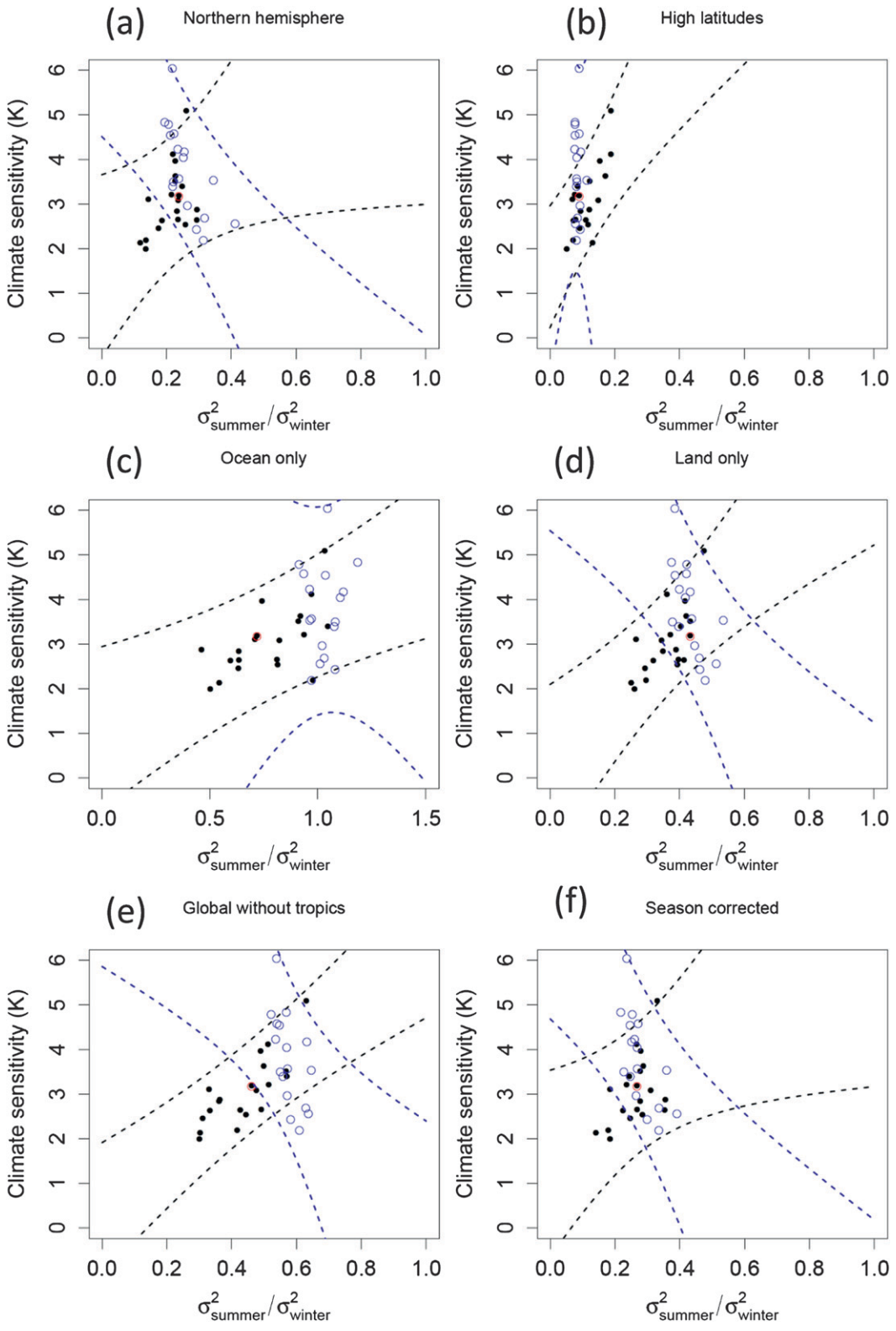

FIG. 3. As in Fig. 2, but for limited regions: (a) Northern Hemisphere, (b) high latitudes, (c) ocean only, (d), land only, and (e) the globe without the tropics. (f) The entire globe is considered, with the Southern Hemisphere shifted by 6 months to match the season occurring in the Northern Hemisphere. The CMIP3 models are shown by black dots, and the QUMP models are shown by blue circles.

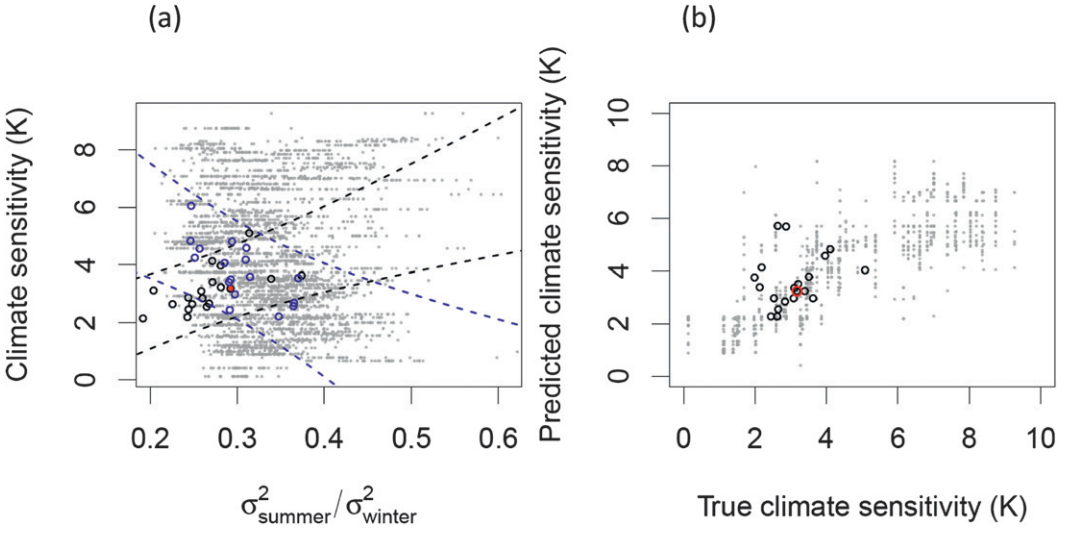

FIG. 4. (a) Scatterplot of climate sensitivity vs the globally averaged regional σ_s^2/σ_w^2 ratios for the CPDN (gray dots), CMIP3 (black dots), and QUMP ensembles (blue dots). The red circle is the CMIP3 HadCM3 unperturbed model. The dashed lines correspond to the prediction CI based on a linear regression. (b) Climate sensitivity predicted with a random forest algorithm using the regional variability σ_s^2/σ_w^2 ratios as an input vs true climate sensitivity. The scatterplot represents a subset of 40% CPDN data not used during the training phase (~2000 simulations; gray dots). The explained variance within CPDN is 69%. The black circles represent the CMIP3 values predicted using the relationship found using the CPDN ensemble; the red circle is the CMIP3 HadCM3 unperturbed model.

$SS = 1 - MSE_{pred}/MSE_{ref}$, where MSE_{pred} is the mean square error of the predicted CMIP3 climate sensitivity value, and MSE_{ref} is the mean square error obtained by using a constant value equal to the median CMIP3 climate sensitivity value. The method has skill if $0 < SS \leq 1$ and has no skill if $SS \leq 0$. Neither the prediction based on the regional σ_s^2/σ_w^2 ratios ($SS = -1.63$) nor that based on the regional seasonal cycle amplitudes $|A|$ ($SS = -1.24$) has predictive skill when applied to the CMIP3 models. In contrast, the prediction applied to the CPDN ensemble is skillful for σ_s^2/σ_w^2 ($SS = 0.69$) and $|A|$ ($SS = 0.76$). As a consequence, the nonlinear relationships are not applicable to different ensembles, consistent with results found by Sanderson and Shell (2012). This is a strong

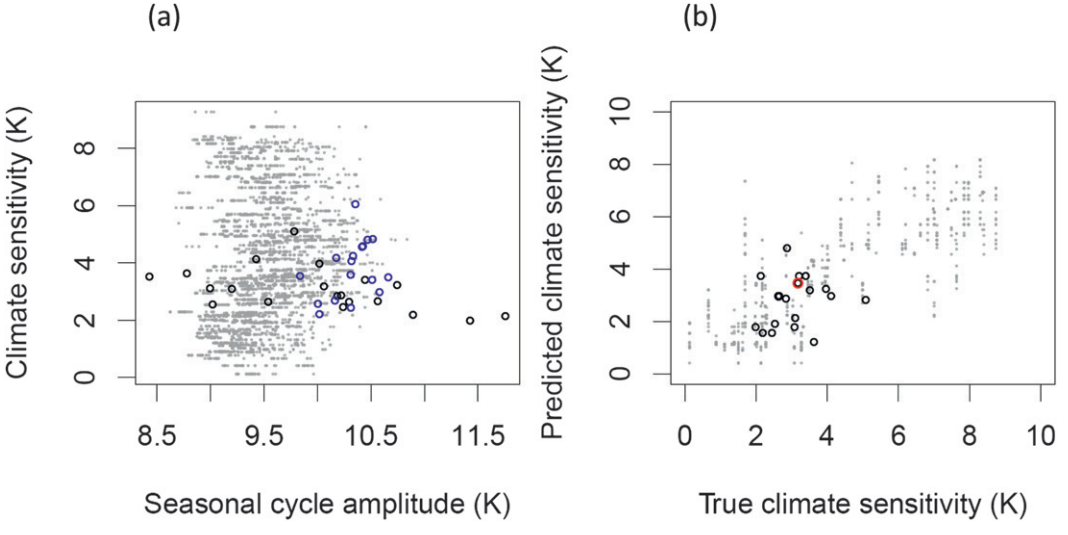

FIG. 5. As in Fig. 4, but using the seasonal cycle amplitude $|A|$ instead. The explained variance within CPDN is 72%.

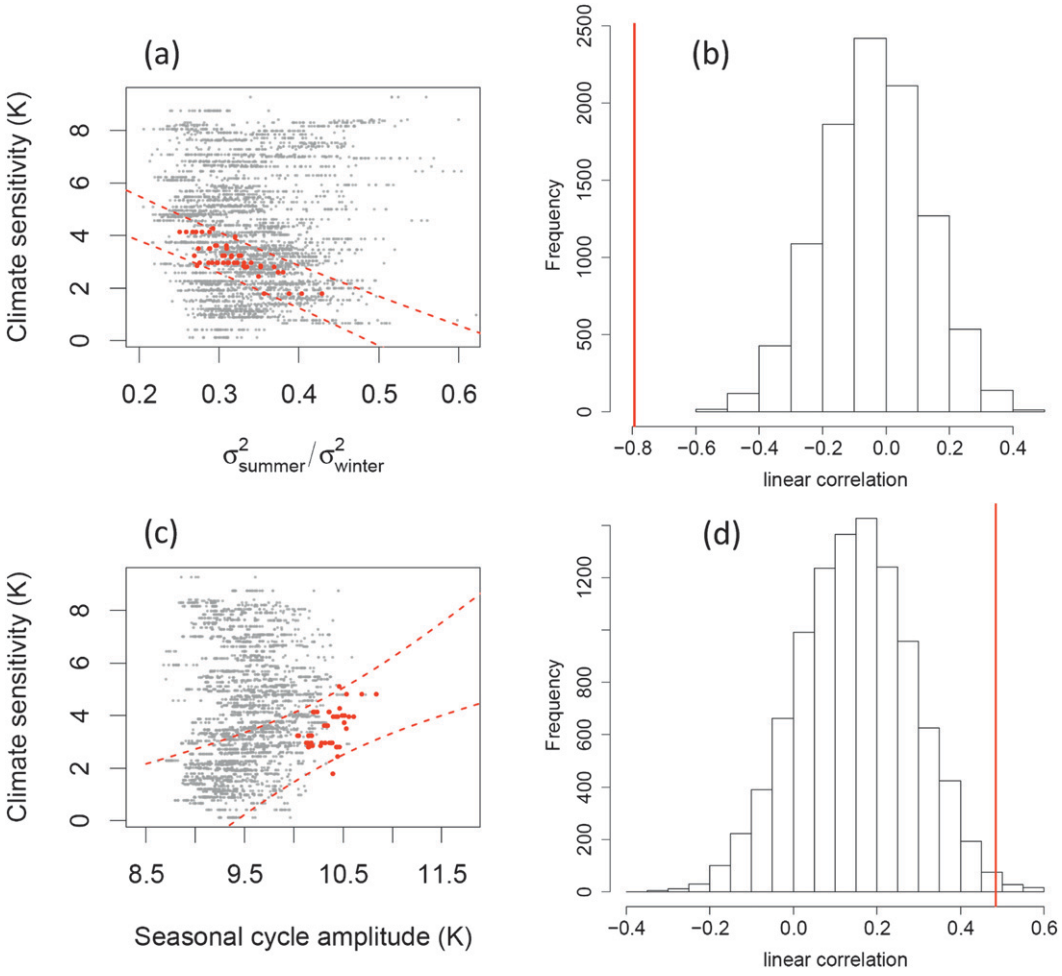

FIG. 6. (a) Scatterplot of climate sensitivity vs the globally averaged regional σ_s^2/σ_w^2 ratios. The gray dots represent ~5000 CPDN perturbed simulations. The red dots represent a subset of 50 CPDN simulations that are both close to the regional temperature mean state and the seasonal cycle of the unperturbed HadCM3 model as a reference. The dashed lines represent the 95% prediction CI of a linear regression. The Pearson correlation coefficient is significant (−0.79). (b) Frequency distribution of the linear correlation coefficients obtained when randomly sampling 50 simulations 10 000 times. The correlation found on the calibrated subset is indicated by a vertical red line. (c) Scatterplot of climate sensitivity vs the seasonal cycle amplitude $|A|$. The red dots represent 50 simulations that are close to the regional temperature mean state of the HadCM3 model. The Pearson correlation coefficient is significant (0.48). (d) As in (b), but for the seasonal cycle vs climate sensitivity relationship.

argument to rule out these two constraints as valid predictors for climate sensitivity, at least when the CPDN ensemble serves as a training set.

c. *Model calibration and uncertainty*

An interesting question is how a significant correlation can emerge without an obvious physical reason, and not be by chance. We focus on illustrating how ''tuning'' or ''calibrating'' model parameters can play a central role. For this question, the CPDN ensemble is the best choice because of its large ensemble size and the possibility to access the parameter values of each simulation. Here ''calibration'' is the selection of a subset of simulations according to some performance metric. It is assumed that the design of the CPDN experiment did not try to reproduce the observations and that the complete ensemble represents a largely uncalibrated state. This is not entirely true, since the CPDN HadCM3L ensemble is already a subset of earlier atmospheric Hadley Centre model (HadAM3) simulations that performed reasonably well, but the range of responses covered by the CPDN ensemble is still very broad, and strong observational constraints were not explicitly placed on any set of parameters.

Figure 6a shows the scatterplot of climate sensitivity versus σ_s^2/σ_w^2 for the CPDN ensemble (gray dots). The

effect of calibration is illustrated by selecting a subset of 50 CPDN simulations according to their proximity to a reference dataset. In this case, a simulation belongs to the calibrated subset if it is close (in terms of RMS error) to the regional temperature mean state and to the regional seasonal cycle value of the unperturbed HadCM3L model as reference. The sample size of 50 is a subjective choice, but it is a compromise between a sufficient number of models to get a robust result and a restrictive criterion for the calibration. The subset is indicated by red dots in Fig. 6a. While no linear relationship exists for the entire CPDN ensemble, a statistically significant linear anticorrelation of -0.79 emerges out of the calibrated subset. When 50 models are randomly chosen instead, the frequency of finding such a high linear correlation is very small. Figure 6b shows the frequency distribution of the correlation calculated in 10 000 randomly chosen simulation subsets of size 50 and demonstrates that the correlation obtained by calibration is not due to chance. A similar approach is repeated in Fig. 6c showing climate sensitivity versus the seasonal cycle amplitude $|A|$. In this case, the selection criterion is the proximity to the regional temperature mean state only. Although lower (0.48), the linear correlation in the subset is still statistically significant. As before, Fig. 6d shows that this correlation is unlikely to be the result of chance, since more than 99% of the correlation coefficients found in the 10 000 random ensembles lie below 0.48. The calibration toward an observational dataset instead of the unperturbed HadCM3L model was also done but resulted in lower correlations because of the structural differences between observations and models.

So far, a physical explanation for the original correlation between climate sensitivity and the σ_s^2/σ_w^2 ratio has not been discovered. While a physical link cannot be fully excluded if some mutual dependences between feedback and forcing exist (e.g., involving aerosol properties), such connections are difficult to trace. In contrast, examples of correlations introduced by observational constraints have already been demonstrated. They might seem surprising because there is no obvious process or mechanism that causes them, yet they are entirely plausible. For example, Kiehl (2007) and Knutti (2008) showed that climate feedback (or climate sensitivity)is correlated with the aerosol forcing across models, because many models try to reproduce the twentieth-century warming. A higher sensitivity in a model can be compensated with a weaker forcing to match the observations. Therefore, it is likely that choices are made in the model development process that introduce correlations between forcings and feedbacks, even if they are not physically related. Such correlations should not be interpreted as being problematic; they simply reflect that different choices in a model are possible given a set of observations. Huybers (2010) similarly showed correlations between feedback processes in CMIP3 that are not obviously related. They likely appear because all models have to close the global energy balance at the top of the atmosphere. Another interesting consequence of model calibration can be the absence of predictors within some ensembles. Because calibration is an iterative process, repeated selections of the best parameters to match the observations reduce the phase space of an ensemble of parameters. This causes the differences between the ensemble members to vanish. Making use of the available information from the observations can improve the reliability of the climate models (Johns et al. 2006). But, as a consequence, if the CMIP3 models are tuned to match the twentieth-century warming trend, the temperature mean state, or seasonal cycle, it is not surprising that these variables are no longer available to constrain future climate change because the observations are no longer independent from the models. This argument is consistent with the fact that few predictors are known to constrain the CMIP3 climate sensitivity values. Even in the case where a valid predictor has been found, the posterior confidence interval for climate sensitivity is often close to the original prior range of the original CMIP3 ensemble (Huber et al. 2011). In general, note that calibration does not necessarily suppress ensemble diversity, as multivariate sets of calibrations are often used to calibrate the parameters of a GCM (Jackson et al. 2008) and many combinations of compensating errors can provide a model that reproduces observations similarly well (Sexton et al. 2012).

4. Summary and discussion

With the availability of coordinated multimodel experiments (Tebaldi and Knutti 2007) and perturbed physics experiments, ideas and methods that relate metrics of model quality to projections have been getting a lot of attention recently. While some of these studies are promising and are likely to improve the accuracy of projections or reduce the spread between models, many have demonstrated that the use of observational constraints on models is far from straightforward. Here, we have shown that an earlier proposed relationship between interannual variability and climate sensitivity across models, even if statistically significant, is unlikely to be robust. In a different model ensemble, the sign of the correlation is reversed, which is difficult to reconcile with the argument that the relationship would be caused by a physical process. Apart from a structural

problem in the HadCM3 model, for which we have no evidence, this only leaves the conclusion that the apparent correlation is by chance, possibly influenced by the screening of many possible predictors. As defined by DelSole and Shukla (2009), screening is ''any procedure for choosing variables that preferentially includes or excludes certain characteristics of the joint relation between predictor and predictand.'' With this meaning, an illustration of screening is provided in Fig. 3 where a lot of configurations were tested and only the case with the highest correlation was kept. This underscores the argument for process understanding as an important component of such an analysis. The case for such an observational constraint is clearly much stronger if the relationship across models can be explained by a known and well-quantified physical process (as in Knutti et al. 2006). In a second part, we have demonstrated that relationships across models can also appear as a result of observational constraints imposed on an ensemble of uncalibrated models.

While model calibration can help to discover potentially useful relationships, the use of the same observations as those assimilated in the calibration to constrain future climate projections could be problematic. Along with small ensembles of the size of CMIP3 and interdependencies between models (Masson and Knutti 2011), such relationships complicate the interpretation of multimodel results and the use of observations in model evaluation and model selection (Knutti 2010; Knutti et al. 2010b; Tebaldi and Knutti 2007; Weigel et al. 2010). In summary, three hurdles need to be overcome before constraining future climate change. First, a relationship has to be found. At this stage, calibration could either help or hinder the process, depending on whether or not the relationship has a physical basis. Second, the relationship needs to be consistent across different types of ensemble. Third, the relationship needs to be physically understandable. Unless a relationship meets these criteria, it is of limited value.

Acknowledgments. We acknowledge the international modeling groups for providing their data for analysis, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) for collecting and archiving the model data, the WCRP/CLIVAR Working Group on Coupled Models (WGCM), and their Coupled Model Intercomparison Project (CMIP) and Climate Simulation Panels for organizing the model data analysis activity. The CMIP3 Multi-Model Data Archive at Lawrence Livermore National Laboratory is supported by the Officeof Science, U.S. Department of Energy.
We also thank all the individuals who have contributed their computer time for CPDN, Oxford University for providing the data, and the Met Office Hadley Centre for the QUMP data.

REFERENCES

Annan, J., and J. Hargreaves, 2011: Understanding the CMIP3 multimodel ensemble. *J. Climate,* **24,** 4529–4538.

Barnett, T., and Coauthors, 2005: Detecting and attributing external influences on the climate system: A review of recent advances. *J. Climate,* **18,** 1291–1314.

Boé, J., A. Hall, and X. Qu, 2009a: Current GCMs' unrealistic negative feedback in the Arctic. *J. Climate,* **22,** 4682–4695.

——, ——, and ——, 2009b: Deep ocean heat uptake as a major source of spread in transient climate change simulations. *Geophys. Res. Lett.,* **36,** L22701, doi:10.1029/2009GL040845.

——, ——, and ——, 2009c: September sea-ice cover in the Arctic Ocean projected to vanish by 2100. *Nat. Geosci.,* **2,** 341–343.

Breiman, L., J. Friedman, C. Stone, and R. A. Olshen, 1984: *Classification and Regression Trees.* Chapman and Hall, 368 pp.

Brohan, P., J. Kennedy, I. Harris, S. Tett, and P. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *J. Geophys. Res.,* **111,** D12106, doi:10.1029/2005JD006548.

Collins, M., B. Booth, B. Bhaskaran, G. Harris, J. Murphy, D. Sexton, and M. Webb, 2011: Climate model errors, feedbacks and forcings: A comparison of perturbed physics and multi-model ensembles. *Climate Dyn.,* **36,** 1737–1766.

Covey, C., and Coauthors, 2000: The seasonal cycle in coupled ocean–atmosphere general circulation models. *Climate Dyn.,* **16,** 775–787.

DelSole, T., and J. Shukla, 2009: Artificial skill due to predictor screening. *J. Climate,* **22,** 331–345.

Frame, D., and Coauthors, 2009: The climateprediction.net BBC climate change experiment: Design of the coupled model ensemble. *Philos. Trans. Roy. Soc. London,* **A369,** 855–870.

Gregory, J., and Coauthors, 2004: A new method for diagnosing radiative forcing and climate sensitivity. *Geophys. Res. Lett.,* **31,** L03205, doi:10.1029/2003GL018747.

Huber, M., I. Mahlstein, M. Wild, J. Fasullo, and R. Knutti, 2011: Constraints on climate sensitivity from radiation patterns in climate models. *J. Climate,* **24,** 1034–1052.

Hurrell, J., 1996: Influence of variations in extratropical wintertime teleconnections on Northern Hemisphere temperature. *Geophys. Res. Lett.,* **23,** 665–668.

Huybers, P., 2010: Compensation between model feedbacks and curtailment of climate sensitivity. *J. Climate,* **23,** 3009–3018.

Jackson, C. S., M. K. Sen, G. Huerta, Y. Deng, and K. P. Bowman, 2008: Error reduction and convergence in climate prediction. *J. Climate,* **21,** 6698–6709.

Johns, T., and Coauthors, 2006: The new Hadley Centre Climate Model (HadGEM1): Evaluation of coupled simulations. *J. Climate,* **19,** 1327–1353.

Jun, M., R. Knutti, and D. Nychka, 2008a: Local eigenvalue analysis of CMIP3 climate model errors. *Tellus,* **60A,** 992–1000.

——, ——, and ——, 2008b: Spatial analysis to quantify numerical model bias and dependence: How many climate models are there? *J. Amer. Stat. Assoc.,* **103,** 934–947.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471.

Kiehl, J., 2007: Twentieth century climate model response and climate sensitivity. *Geophys. Res. Lett.,* **34,** L22710, doi:10.1029/2007GL031383.

Kim, K., and Q. Wu, 2000: Optimal detection using cyclostationary EOFs. *J. Climate,* **13,** 938–950.

Knight, C., and Coauthors, 2007: Association of parameter, software, and hardware variation with large-scale behavior across 57,000 climate models. *Proc. Natl. Acad. Sci. USA,* **104,** 12 259–12 264.

Knutti, R., 2008: Why are climate models reproducing the observed global surface warming so well? *Geophys. Res. Lett.,* **35,** L18704, doi:10.1029/2008GL034932.

——, 2010: The end of model democracy? *Climatic Change,* **102,** 395–404.

——, and G. Hegerl, 2008: The equilibrium sensitivity of the Earth's temperature to radiation changes. *Nat. Geosci.,* **1,** 735–743.

——, G. Meehl, M. Allen, and D. Stainforth, 2006: Constraining climate sensitivity from the seasonal cycle in surface temperature. *J. Climate,* **19,** 4224–4233.

——, G. Abramowitz, M. Collins, V. Eyring, P. J. Gleckler, B. Hewitson, and L. Mearns, 2010a: Good practice guidance paper on assessing and combining multi model climate projections. Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections, IPCC, 1–14.

——, R. Furrer, C. Tebaldi, J. Cermak, and G. Meehl, 2010b: Challenges in combining projections from multiple climate models. *J. Climate,* **23,** 2739–2758.

Kumar, A., and F. Yang, 2003: Comparative influence of snow and SST variability on extratropical climate in northern winter. *J. Climate,* **16,** 2248–2261.

Mahlstein, I., and R. Knutti, 2011: Ocean heat transport as a cause for model uncertainty in projected Arctic warming. *J. Climate,* **24,** 1451–1460.

——, and ——, 2012: September Arctic sea ice predicted to disappear near 2°C global warming above present. *J. Geophys. Res.,* **117,** D06104, doi:10.1029/2011JD016709.

Masson, D., and R. Knutti, 2011: Climate model genealogy. *Geophys. Res. Lett.,* **38,** L08703, doi:10.1029/2011GL046864.

Meehl, G., C. Covey, K. E. Taylor, T. Delworth, R. J. Stouffer, M. Latif, B. McAvaney, and J. F. B. Mitchell, 2007: The WCRP CMIP3 multimodel dataset—A new era in climate change research. *Bull. Amer. Meteor. Soc.,* **88,** 1383–1394.

Murphy, J., D. Sexton, D. Barnett, G. Jones, M. Webb, M. Collins, and D. Stainforth, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature,* **430,** 768–772.

Piani, C., D. Frame, D. Stainforth, and M. Allen, 2005: Constraints on climate change from a multi-thousand member ensemble of simulations. *Geophys. Res. Lett.,* **32,** L23825, doi:10.1029/2005GL024452.

Raisanen, J., L. Ruokolainen, and J. Ylhaisi, 2010: Weighting of model results for improving best estimates of climate change. *Climate Dyn.,* **35,** 407–422.

Rienecker, M., and Coauthors, 2011: MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications. *J. Climate,* **24,** 3624–3648.

Rowlands, D. J., and Coauthors, 2012: Broad range of 2050 warming from an observationally constrained large climate model ensemble. *Nat. Geosci.,* **5,** 256–260.

Sanderson, B. M., 2011: A multimodel study of parametric uncertainty in predictions of climate response to rising greenhouse gas concentrations. *J. Climate,* **24,** 1362–1377.

——, and K. M. Shell, 2012: Model-specific radiative kernels for calculating cloud and noncloud feedbacks. *J. Climate,* **25,** 7607–7624.

——, and Coauthors, 2008a: Constraints on model response to greenhouse gas forcing and the role of subgrid-scale processes. *J. Climate,* **21,** 2384–2400.

——, C. Piani, W. Ingram, D. Stone, and M. Allen, 2008b: Towards constraining climate sensitivity by linear analysis of feedback patterns in thousands of perturbed-physics GCM simulations. *Climate Dyn.,* **30,** 175–190.

——, K. Shell, and W. Ingram, 2010: Climate feedbacks determined using radiative kernels in a multi-thousand member ensemble of AOGCMs. *Climate Dyn.,* **35,** 1219–1236.

Sexton, D. M. H., J. M. Murphy, M. Collins, and M. J. Webb, 2012: Multivariate probabilistic projections using imperfect climate models. Part I: Outline of methodology. *Climate Dyn.,* **38,** 2513–2542.

Shukla, J., T. DelSole, M. Fennessy, J. Kinter, and D. Paolino, 2006: Climate model fidelity and projections of climate change. *Geophys. Res. Lett.,* **33,** L07702, doi:10.1029/2005GL025579.

Soden, B., and I. Held, 2006: An assessment of climate feedbacks in coupled ocean–atmosphere models. *J. Climate,* **19,** 3354–3360; Corrigendum, **19,** 6263.

Solomon, S., D. Qin, M. Manning, M. Marquis, K. Averyt, M. M. B. Tignor, H. L. Miller Jr., and Z. Chen, Eds., 2007: *Climate Change 2007: The Physical Science Basis.* Cambridge University Press, 996 pp.

Stainforth, D., and Coauthors, 2005: Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature,* **433,** 403–406.

——, M. Allen, E. Tredger, and L. Smith, 2007: Confidence, uncertainty and decision-support relevance in climate predictions. *Philos. Trans. Roy. Soc. London,* **A365,** 2145–2161.

Stevenson, M., 2006: Forecast verification: A practitioner's guide in atmospheric science. *Int. J. Forecast.,* **22,** 403–405.

Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. Roy. Soc. London,* **A365,** 2053–2075.

Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. Soc.,* **131,** 2961–3012.

Webb, M., and Coauthors, 2006: On the contribution of local feedback mechanisms to the range of climate sensitivity in two GCM ensembles. *Climate Dyn.,* **27,** 17–38.

Weigel, A., R. Knutti, M. Liniger, and C. Appenzeller, 2010: Risks of model weighting in multimodel climate projections. *J. Climate,* **23,** 4175–4191.

Wigley, T., R. Smith, and B. Santer, 1998: Anthropogenic influence on the autocorrelation structure of hemispheric-mean temperatures. *Science,* **282,** 1676–1679.

Wu, Q., and G. North, 2003: Statistics of calendar month averages of surface temperature: A possible relationship to climate sensitivity. *J. Geophys. Res.,* **108,** 4071, doi:10.1029/2002JD002218.

——, D. Karoly, and G. North, 2008: Role of water vapor feedback on the amplitude of season cycle in the global mean surface air temperature. *Geophys. Res. Lett.*, **35,** L08711, doi:10.1029/2008GL033454.

Yokohata, T., M. J. Webb, M. Collins, K. D. Williams, M. Yoshimori, J. C. Hargreaves, and J. D. Annan, 2010: Structural similarities and differences in climate responses to CO_2 increase between two perturbed physics ensembles. *J. Climate,* **23,** 1392–1410.