Constraints on Model Response to Greenhouse Gas Forcing and the Role of Subgrid-Scale Processes

BENJAMIN M. SANDERSON,* R. KNUTTI,+ T. AINA,* C. CHRISTENSEN,* N. FAULL,* D. J. FRAME,# W. J. INGRAM,*,@ C. PIANI,& D. A. STAINFORTH,# D. A. STONE,* AND M. R. ALLEN*

*Department of Physics, University of Oxford, Oxford, United Kingdom
+National Center for Atmospheric Research, Boulder, Colorado
#Environmental Change Institute, University of Oxford, Oxford, United Kingdom
@Met Office, Exeter, United Kingdom
&International Center for Theoretical Physics, Trieste, Italy

(Manuscript received 26 January 2007, in final form 3 March 2008)

ABSTRACT

A climate model emulator is developed using neural network techniques and trained with the data from the multithousand-member climate*prediction*.net perturbed physics GCM ensemble. The method recreates nonlinear interactions between model parameters, allowing a simulation of a much larger ensemble that explores model parameter space more fully.

The emulated ensemble is used to search for models closest to observations over a wide range of equilibrium response to greenhouse gas forcing. The relative discrepancies of these models from observations could be used to provide a constraint on climate sensitivity. The use of annual mean or seasonal differences on top-of-atmosphere radiative fluxes as an observational error metric results in the most clearly defined minimum in error as a function of sensitivity, with consistent but less well-defined results when using the seasonal cycles of surface temperature or total precipitation.

The model parameter changes necessary to achieve different values of climate sensitivity while minimizing discrepancy from observation are also considered and compared with previous studies. This information is used to propose more efficient parameter sampling strategies for future ensembles.

1. Introduction

Predictions of future climate are made uncertain in part by lack of knowledge of future greenhouse gas emissions and in the response of the climate system to emissions (Stott and Kettleborough 2002; Hansen et al. 2001). General circulation models (GCMs) are often used to predict the likely response of the climate system, but modeling the earth system at a coarse resolution introduces additional uncertainties in finding the most appropriate values for parameterizations of subgrid-scale physical processes (Johns et al. 2003; Annan et al. 2005a). Advances in available computing resources over recent years have led to the increased use of ensemble techniques for exploring the parameter dependence of physically based climate models (Stainforth et al. 2002).

This paper focuses on the equilibrium response of global mean temperature to a doubling of the atmospheric CO_2 concentration (henceforth referred to as climate sensitivity or S), and how that response is influenced by some of the parameter settings within the GCM. It builds upon the results of an existing ensemble of GCMs (Stainforth et al. 2005).

Minimization of model error

Rougier (2007) described the difference between models and the climate itself to be a sum of two parts: a reducible and an irreducible part. The reducible part may be lessened by a better choice of model parameters, while the irreducible part is a "systematic

Corresponding author address: Ben Sanderson, AOPP, Department of Physics, University of Oxford, Clarendon Laboratory, Parks Road, Oxford OX1 3PU, United Kingdom.
E-mail: sanderson@atm.ox.ac.uk

DOI: 10.1175/2008JCLI1869.1

© 2008 American Meteorological Society

error"—a result of model imperfections that cannot be removed by "tuning" parameters.

The climate*prediction*.net dataset has already been used to produce two predictions of climate sensitivity; both Piani et al. (2005) and Knutti et al. (2006) used the GCM ensemble to find predictors of S from aspects of model climatology. Once established, the predictors were applied to observations of true climate, treating those observations as members of the ensemble. Such approaches make the implicit assumption that the "perfect model" is a tunable state of the GCM, ignoring the irreducible error described in Rougier (2007). To ignore this component of the error means that some degree of extrapolation is required when applying predictors to observations, adding an unknown error to the results. Knutti et al. (2006) already noted that because of the structural biases in the Met Office atmospheric model (HadAM3), the relation between observed and predicted quantities did not always hold in different GCMs.

To address this issue, we seek to identify the irreducible components of model error by minimizing the model–observation error over a perturbed physics ensemble. To do this, we use the climate*prediction*.net ensemble to fit a surface representing key model output as a function of the model parameter values. This surface can be used to find those models with output closest to observations. By also predicting how climate sensitivity varies with model parameters, we can restrict consideration to models with a specific value of S, thus examining how the systematic error varies with equilibrium response. The relative systematic errors at different values of S provide some constraint on sensitivity.

The surface fitting procedure requires an emulator for the parameter dependence of model climatology. The method used in previous studies (Murphy et al. 2004) was to conduct a set of single-perturbation experiments, each producing a range of observable diagnostics and an estimate of equilibrium response. It was then assumed that observables for any combination of the individually perturbed parameter settings could be estimated by linear interpolation from the single-perturbation simulations (some allowance for nonlinear parameter dependence was made by using some multiply perturbed simulations). Thus they linearly predicted the equilibrium response of a large number of randomly generated parameter settings. A likelihood weighting was predicted for each simulation based on its predicted closeness to observations. The resulting probability density function (PDF) for climate sensitivity was produced by generating a weighted histogram of equilibrium response for the simulated ensemble, each model weighted according to its predicted likelihood, as judged by its Climate Prediction Index (CPI)—the combined normalized root-mean-square error over a number of different mean climate variables.

The use of a linear model to predict the model response as a function of parameter values was tested by Stainforth et al. (2005). In this paper, conducted with the results taken from the climate*prediction*.net ensemble of climate models, it was found that a linear prediction of S made by interpolating the results of single parameter simulations was a poor estimator of the true response of a multiperturbation simulation. The ensemble response landscape was found to be a strongly nonlinear function.

Hence in this paper, we propose the use of an artificial neural network (ANN) whose weightings may be trained to best relate the perturbed parameters of a climate model to both output diagnostics and equilibrium response. The application of nonlinear neural network techniques to analyze climate model output is not new (Hsieh and Tang 1998; Knutti et al. 2006; among others), but the use of a neural network to directly emulate climatological model output from perturbed parameter settings has been suggested (Collins et al. 2006) but not attempted before to the knowledge of the authors.

Rodwell and Palmer (2007) conducted an ensemble using initial value forecasting techniques, where the performance of a model was judged by the rate at which it diverged from an observational state. Their work suggested that some of the discrete parameter settings that lead to high sensitivity in the climate*prediction*.net ensemble may lead to unrealistic atmospheres that quickly diverge from observations. Hence, by using the ANN to emulate an ensemble based on a Monte Carlo–style parameter sampling scheme, we explore the error as a continuous function of the model's parameter space and not just at the extreme values.

We divide the rest of the paper into three sections: in section 2, we discuss the methodology used for the analysis; section 2a describes the climate*prediction*.net dataset, 2b shows the techniques used to compress the climatological data, while 2c describes the neural network architecture and training process that is used to optimize the nonlinear fit.

In section 3a, we present the results of the emulator—its ability to predict a verification set of models, and how well it interpolates between known values. Section 3b is a discussion of how the emulator may be used to produce a Monte Carlo ensemble of simulations, allowing us to predict the most realistic model for

TABLE 1. Definition of perturbed parameters as used in the subset of climate*prediction*.net experiments used in this analysis. Parameters marked * and ** are perturbed together. Ice-type parameters (**) are switches, but for the purposes of this paper, it has been assumed that a continuum exists for the emulated models between the on and off states.

Perturbation label	Description	Sample size
CW_SEA*	Cloud water threshold for precipitation over ocean	3
CW_LAND*	Cloud water threshold for precipitation over land	3
Entcoef	Entrainment coefficient	3
RHCrit	Critical relative humidity	3
CT	Accretion constant	3
EACF	Empirically adjusted cloud fraction	2
VF1	Ice fall speed	3
DTHETA	Initial condition ensemble generator	10
ALPHAM	Albedo at melting point of ice	3
DTICE	Temperature range of ice albedo variation	3
I_CNV_ICE_LW**	Type of convective cloud ice crystal used in longwave radiation	3
I_CNV_ICE_SW**	Type of convective cloud ice crystal used in shortwave radiation	3
I_ST_ICE_LW**	Type of stratiform cloud ice crystal used in longwave radiation	3
I_ST_ICE_SW**	Type of stratiform cloud ice crystal used in shortwave radiation	3
ICE_SIZE	Ice crystal size in radiation	3

a given equilibrium response. We compare the constraints imposed by various different observations, using both annual mean and seasonal data.

Finally, in section 4 we discuss the parameter settings suggested by this optimization process for different values of climate sensitivity. We analyze these parameters in the light of previous research and propose a more efficient method of sampling parameters for future ensembles of this type.

2. Methodology

a. *The climate*prediction.*net ensemble*

This work uses a subset of data from the first climate*prediction*.net ensemble, which is the result of a distributed computing experiment that allows interested members of the public to run a perturbed climate model on their own computers. Model diagnostics are returned back to a central server for analysis.

The reference model used is the HadAM3 coupled to a single layer thermodynamic ocean (Pope et al. 2000). The model has a resolution of $3.75° \times 2.5°$, with 19 vertical levels ranging from about 1000 to 10 mb. A total of 15 parameters in the model are perturbed, some of which are always perturbed together. Each parameter is perturbed discretely and may assume one of two or three possible values, which represent estimates of the extremes of the range of current uncertainty in the value of that parameter (which were established through expert solicitation).

Each individual parameter set is simulated several times, with slightly altered initial conditions to produce subensembles of model variability. In this analysis, the results of initial condition ensembles are averaged together. This leaves 10 remaining degrees of freedom in the perturbed parameter space. The perturbed parameters are listed in Table 1.

Each model in the ensemble is divided into three individual simulations: the first is a *calibration* simulation, where the sea surface temperatures (TASs) are constrained to match observations. The resulting heat imbalance at the ocean surface required to maintain these temperatures is measured and is henceforth referred to as the anomalous heat flux convergence field. In the following simulations, *control* and *doubled* CO_2, the convergence field is applied at the ocean surface, and the ocean temperatures are allowed to vary freely.

Most simulations remain stable in the control simulation, but some suffer a drift in global mean temperatures owing to an unrealistic feedback with the models' thermodynamic oceans. We remove drifting simulations in our analysis following the conditions established in Stainforth et al. (2005). In addition, we remove models with missing data and those models with highly unphysical temperatures in the control simulation. The calculation of model climate sensitivity is also calculated using the same exponential fitting algorithm used in Stainforth et al. (2005).

b. *Data preparation*

We seek a smooth fit to the simulated climatology and likely response to greenhouse gas forcing, both as functions of model parameters. The data required to train this emulation are taken from the first climate*prediction*.net ensemble of climate models, those experiments conducted with perturbed atmospheric parameters only.

TABLE 2. Definition of regions as used in the climate*prediction*.net experiment.

No.	Definition of regions as used in the analysis				
	Name	Min lon	Max lon	Min lat	Max lat
1	Australia	110	155	−45	−10
2	Amazon Basin	280	325	−20	10
3	Southern South America	285	320	−55	−20
4	Central America	245	275	10	30
5	Western North America	230	255	30	60
6	Central North America	255	275	30	50
7	Eastern North America	275	300	25	50
8	Alaska	190	255	60	70
9	Greenland	255	350	50	85
10	Mediterranean Basin	350	40	30	50
11	Northern Europe	350	40	50	75
12	Western Africa	340	20	−10	20
13	Eastern Africa	20	50	−10	20
14	Southern Africa	350	50	−35	−10
15	Sahara	340	65	20	30
16	Southeast Asia	95	155	−10	20
17	East Asia	100	145	20	50
18	South Asia	65	100	5	30
19	Central Asia	40	75	30	50
20	Tibet	75	100	30	50
21	North Asia	40	180	50	70
22	Antarctica	0	360	−90	−65
23	NH	0	360	0	90
24	SH	0	360	−90	0
25	NH extratropics	0	360	30	90
26	SH extratropics	0	360	−90	−30

After filtering, an N member subset of models remains for use in this analysis (where N is 6096).

The nature of the climate*prediction*.net ensemble means that the available data for each model are limited by the bandwidth available to the participants. Thus, from each control simulation in the ensemble, we examine data from regions as defined in Giorgi and Francisco (2000), which are listed in Table 2. The included regions are land based (although, because of the regions being rectangular, there is some ocean area included near coastlines). All zonal mean regions are excluded because they include a large amount of ocean, which is adjusted to climatology in the *calibration* simulation; thus any comparison with observations is an unfair test of model behavior. The total number of remaining regions used in the analysis, R, is 21.

In each region, we take a subset of atmospheric variables from the model's control simulation to represent the model climatology (Table 3). These are compared with climatological means from the National Centers for Environmental Prediction (NCEP) reanalysis (for temperature and precipitation data) and Earth Radiation Budget Experiment (ERBE; for radiative data). These sources are henceforth referred to as observations (although it is recognized that alternative observations are available).

We calculate empirical orthogonal functions (EOFs) of the control climatic states of the ensemble to determine dominant modes. In contrast to conventional EOFs, the temporal dimension is replaced by the ensemble itself, which provides a convenient orthogonal

TABLE 3. Climatological fields measured for comparison to observational datasets. Winter (December–February) and summer (June–August) means over all available data are taken in the regions specified, along with standard deviations to represent interannual variability. ERBE* is used for radiative data where available, and is supplemented with NCEP data for latitudes greater than 67.5° N–S. All fields are sampled for seasonal means over a 15-yr time period in regions 1–21 (Table 2).

Climate fields chosen for analysis	
Climate variable	Dataset(s) used
TAS	NCEP
Shortwave upward radiation at TOA	NCEP/ERBE*
Longwave upward radiation at TOA	NCEP/ERBE*
Longwave clear-sky upward radiation at TOA	NCEP/ERBE*
Shortwave clear-sky upward radiation at TOA	NCEP/ERBE*
Total precipitation	NCEP

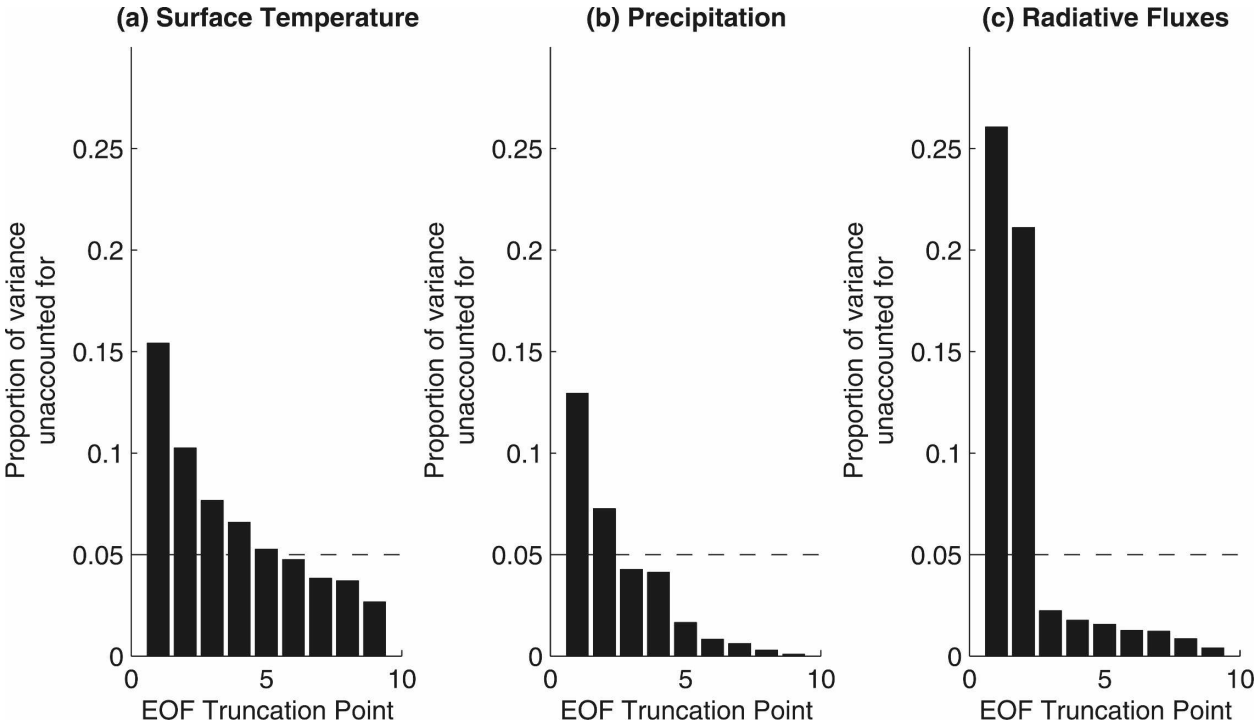

FIG. 1. Plots indicating EOF truncation point for each of three input fields. The series of EOFs is truncated when mean RMSE from observations over the ensemble can be estimated within 5%. Truncation points for temperature, precipitation, and radiation fields are 6, 3, and 3 modes, respectively.

basis to compact the ensemble variance in the control climate. This compacted climate vector allows us to simplify the structure and increase the computational efficiency and reliability of the neural network by decreasing the number of required outputs. The resulting EOFs are spatial patterns, while their principal components are the expansion coefficients showing the amplitude of the EOF in each ensemble member.

The EOFs are taken over three types (s) of model output: surface temperature, radiative fluxes, and precipitation. Thus the input matrices for the temperature and precipitation EOF analyses have $R \times N$ elements, where each element is the annual mean anomaly from the control mean for region r in model n, weighted by the area of region r. The input matrix for the EOF taken over the radiative fluxes is size $4R \times N$ to include clear-sky and cloudy-sky fluxes in shortwave and longwave bands.

The resulting set of EOFs is truncated to the first K modes when 95% of the ensemble variance has been accounted for (see Fig. 1). The truncation is conducted for computational efficiency only, as emulating a large number of outputs with the neural network is computationally expensive. Results are not highly sensitive to a further increase in truncation length.

Model error as compared to observations is calculated by comparing the EOF amplitudes in each simulation to the projection of those EOFs onto observational or reanalysis datasets for each observation type s. The projection is calculated by first removing the climate*prediction*.net mean state from the observational dataset, and then calculating the scalar product with each EOF. Each model's error, E_{is}, is then calculated by taking the root-mean-square error across all truncated modes:

$$E_{is} = \sqrt{\sum_{k=1}^{K} (w_{isk} - o_{sk})^2}, \tag{1}$$

where w_{isk} is the amplitude of mode k in model i for observation type s and o_{sk} is the projection of mode k onto the observational dataset.

To combine the different observational errors, they must first be normalized. This may be achieved by using some estimate of natural variability for the observation in question (Piani et al. 2005). Thus for each observation type s, the error is normalized by the variance of the projection of the leading EOF, e_{s1}, onto a 500-yr control simulation of HadCM3 (the Met Office HadAM3 model coupled to a fully dynamic ocean). The errors are now dimensionless and the root-mean-

square combination of these gives a total error for each model. No further weighting over observational type is applied.

In this case, we have chosen to average the quadratic cost functions rather than adding them. Whereas a sum–squares combination of model errors may be appropriate if the variables used are mutually independent and identically distributed, the quantities in this case are correlated and thus to combine errors in this way would be incorrect. The mean-squared approach described above is used to emphasize the structural model differences between the reanalysis data and the ensemble models by allowing the same scale to be used in each case, regardless of the number of constraints used. In the case that the introduction of a new variable places the observations in a region outside of the ensemble-sampled space, then a systematic error is present that would be apparent by examining the RMSE associated with the best-performing ensemble model. The combined cost term should not, however, be interpreted as a log-likelihood scale. To produce a probability density function from the function E(S) requires further assumptions on how one should deal with the presence of an irreducible error, and this question is left to future work.

USE OF SEASONAL DATA

The original analysis of climate*prediction*.net data shown in Stainforth et al. (2005) found a weak observational constraint on S; models with climate sensitivities of up to 11 K were shown to perform comparably in a simple test of root-mean-square error as measured from the observations used. Annan et al. (2005b) suggested that the relatively high performance of these models may be due to the omission of seasonal information when comparing each model to observations.

To address this issue, we conduct two additional experiments. The first replaces the annual mean values with a June–August minus December–February (JJA − DJF) difference in each observational field to construct the input EOFs. A second analysis includes both JJA and DJF seasonal means as separate dimensions of the input vector, allowing the seasonal cycle of each region to influence the resulting EOFs. The EOFs using seasonal information are treated identically to the analysis for the annual mean data shown above.

c. *Neural network architecture*

We employ an artificial neural network to emulate the response of the climate model output. We will summarize the theory of neural network architecture and training here, but a full discussion of the topic is given in Hagan et al. (1996).

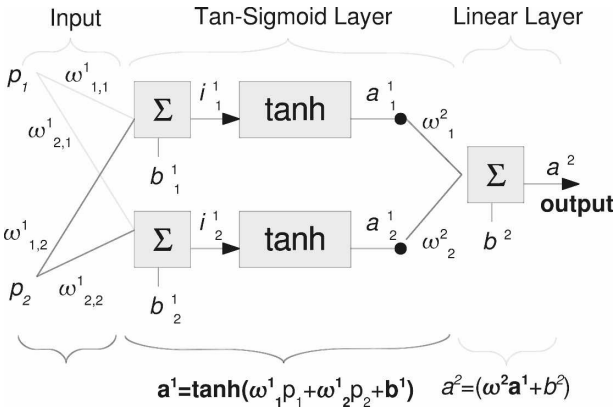

FIG. 2. A diagram illustrating the flow of information through a two-layer feed-forward neural network. The input vector has two elements. The first layer is a "tansig" type, with two neurons, the second layer is a "purelin" type with one neuron.

The network employed is a two-layer, feed-forward ANN (illustrated in Fig. 2). The elements of the input vector, p_{il}, consist of the independent perturbed parameter set associated with each model, i. The parameters are listed in Table 1. Where parameters are perturbed together in the model, only one of the values is used in the analysis. Where parameters are defined on model levels, the average value of the parameter over all model levels is used. The result is a vector of 10 elements that defines the parameter set for any given model in the ensemble.

The output vector is the quantity we wish to predict. In the first instance, this may be a single value: the model's climate sensitivity, S_i. However, later we extend the analysis to predict the set of EOF amplitudes, w_{ik} defined earlier, which define the model's climatology.

For the ANN to best approximate a relationship between these quantities, it is separated into layers. The input to the network is the set of model parameters, p_{il}, which are combined with scale and offset before being passed to the first "hidden" layer. The weights and biases are set iteratively during the training process. The hidden layer is a set of nonlinear functions, arbitrary in principle [in this case, a function closely approximating a hyperbolic tangent is used (Vogl et al. 1988)]. The output of the hidden layer is again weighted and biased to produce the elements of the output vector.

To train the network to emulate the output of the model ensemble, we employ the Levenberg–Marquardt learning algorithm (Hagan and Menhaj 1994). This back-propagation algorithm is a gradient descent optimization, where the algorithm is provided with a set of

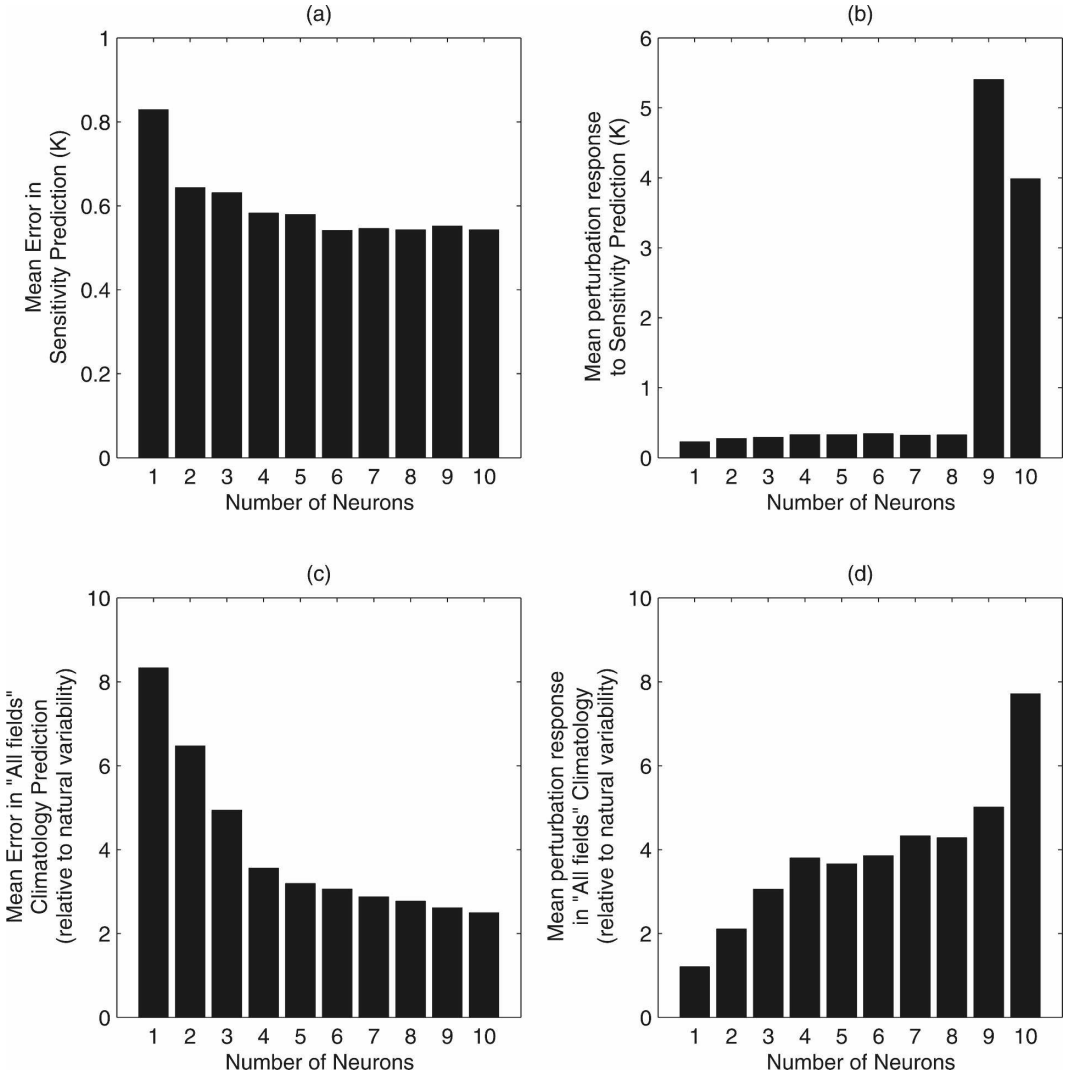

FIG. 3. (a) The mean error in predicting climate sensitivity of the verification set, as a function of the number of neurons in the hidden layer. (b) The mean absolute difference in predicted sensitivity between 1000 random emulations, and a second set of slightly perturbed emulations (the perturbed emulations take a weighted average of input parameters, with a 90% weighting to the original values, and a 10% perturbation in each input parameter from a second set of random simulations). (c) The RMSE in the prediction of the all fields climatology for the verification set, scaled by natural variability. (d) The RMS response in all fields climatology to a small parameter perturbation [as in (b)].

examples of "proper" network behavior. In this case, the training set is provided by 60% of the available models in the ensemble, totaling about 4000 examples. Using a larger number of models in the training set did not noticeably improve accuracy.

The ideal number of neurons to be used in the hidden layer should ensure accuracy while avoiding overfitting. Figure 3a shows the mean fitting error of the network in predicting an unseen "verification" set of models as a function of the number of neurons. This plot suggests little increase in accuracy for more than six neurons.

Figure 3b shows the effects of overfitting on the input data. Here we take a sample of random parameter combinations and perturb them slightly, examining the impact on the predicted sensitivity (thereby estimating the steepness of the response). For less than eight neurons, this results in a slight mean perturbation to the predicted sensitivities. However, the tests conducted with 9 and 10 neurons show large discrepancies between the original and perturbed simulations, indicating an overfitted network, with large gradients in response. Thus, a conservative six neurons are used in the hidden layer.

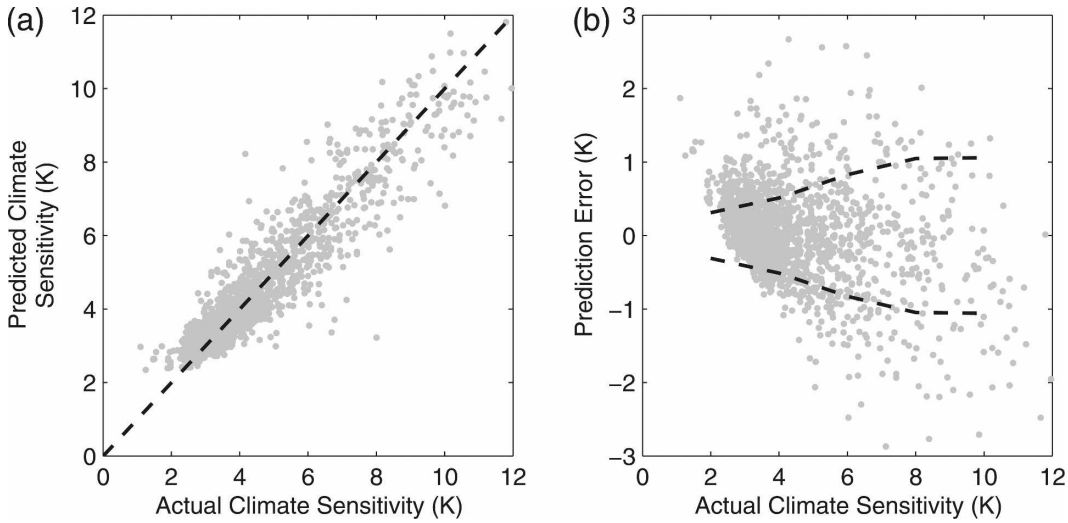

FIG. 4. (a) A plot showing the predicted sensitivities of the verification set of climate*prediction*.net models as a function of their actual sensitivities. (b) A plot of the prediction error in S as a function of S. Each point represents a member of the verification set in the climate*prediction*.net ensemble and the width between lines represents the standard error in prediction at a given climate sensitivity.

A similar process is conducted for the prediction of w_{ik}: to measure the prediction ability of the network, we predict the K principal components for each member of the verification set. The prediction error is root-mean-square difference between the neural network estimation and the actual value in the verification set. We measure the smoothness of the response surface by taking the root-mean-square response to a small parameter perturbation, as before. Again, six neurons are appropriate for predicting w_{ik}.

The cost function used in the iterating training procedure measures network performance as a combination of the mean squared prediction error (85%) and the mean squared weight and bias values (15%). This prevents any single neuron from being weighted too highly, which was found to further help prevent the network from overfitting.

Once the network has been trained and verified, we perform a Monte Carlo parameter perturbation experiment, emulating an ensemble many orders of magnitude greater than the original climate*prediction*.net dataset. The ensemble densely samples the emulated parameter space, allowing a search for the best-performing models in different (0.1 K) bins of climate sensitivity, as judged by various observational constraints.

The underlying function of minimized model–observation error as a function of sensitivity $E(S)$ is thus discretized into 0.1-K bins of S. The Monte Carlo ensemble is sufficiently densely populated so that the following statements are true:

- $E(S)$ is a smooth, continuous function.
- $E(S)$ does not alter if the sampling density is further increased.

Note that the issues of prior sampling of climate sensitivity raised in Frame et al. (2005) are not relevant here, because we do not attempt to assign probabilities to different values of S. The sampling of S is simply used to outline the shape of the underlying function $E(S)$.

3. Results

a. Verification

We first show a demonstration of the ability of the neural network to predict an unseen verification set within the ensemble itself. Figure 4a illustrates the network's ability to predict S. Figure 4b shows that the standard error in prediction increases with increasing sensitivity, an effect also noted both in Piani et al. (2005) and Knutti et al. (2006). This is simply explained by considering that observables tend to scale with λ, the inverse of S. Although in practice, a direct prediction of S with the neural network is considerably more accurate than a linear prediction of λ for large values of S.

The network must be able to predict model climatology for previously unseen parameter combinations. Figure 5 uses the verification set to demonstrate the network's ability to predict the total RMSE from observations for each of the different observation types.

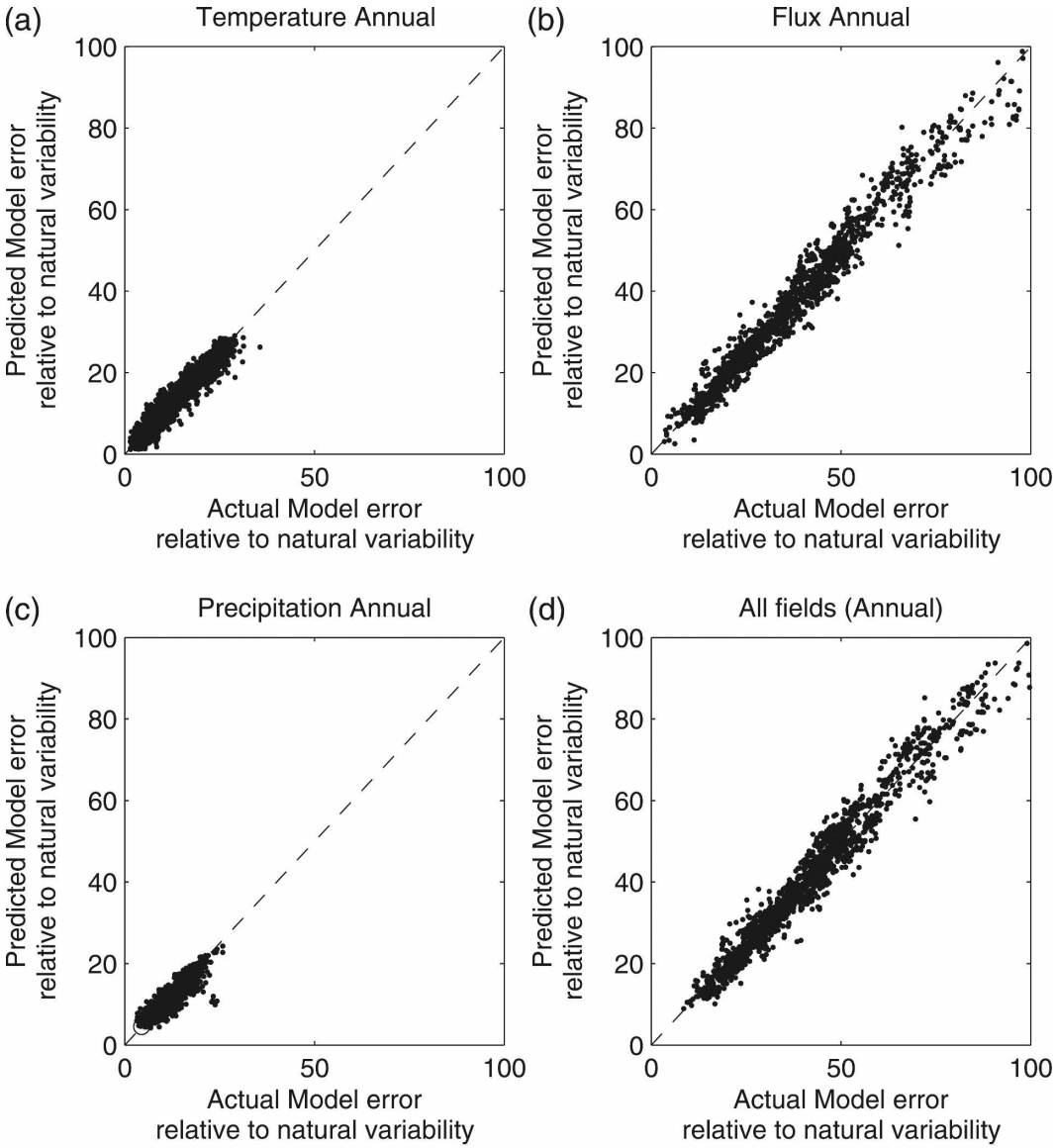

FIG. 5. (a)–(d) The vertical axes show the ANN predicted combined model errors as compared to annual mean observations taken from observations for the verification set of climate*prediction*.net models for each observation type. The horizontal axes show the actual RMSE, before EOF truncation. The model errors are normalized by the variance of the projection of the first EOF onto a 500-yr control simulation.

This is not a test of the network's ability to interpolate between discrete parameter values.

PARAMETER INTERPOLATION

The climate*prediction*.net ensemble uses a parameter sampling strategy that chooses one of a small number of possible values for each parameter. However, once trained, the ANN emulator may be used to interpolate between these values and map out the parameter space more completely. Given that we do not know the true behavior of models in this unsampled parameter space, the ANN is designed such that there is a smooth transition between the model responses at known, discrete parameter values.

This process is demonstrated by perturbing each in turn of P individual parameter settings within the limits of the sampled climate*prediction*.net range, while keeping the other $(P - 1)$ parameters at the standard HadAM3 value. Thus we can observe the emulator's ability to interpolate climatology and greenhouse gas response between the known discrete parameter settings. Section 2c described how the choice of network

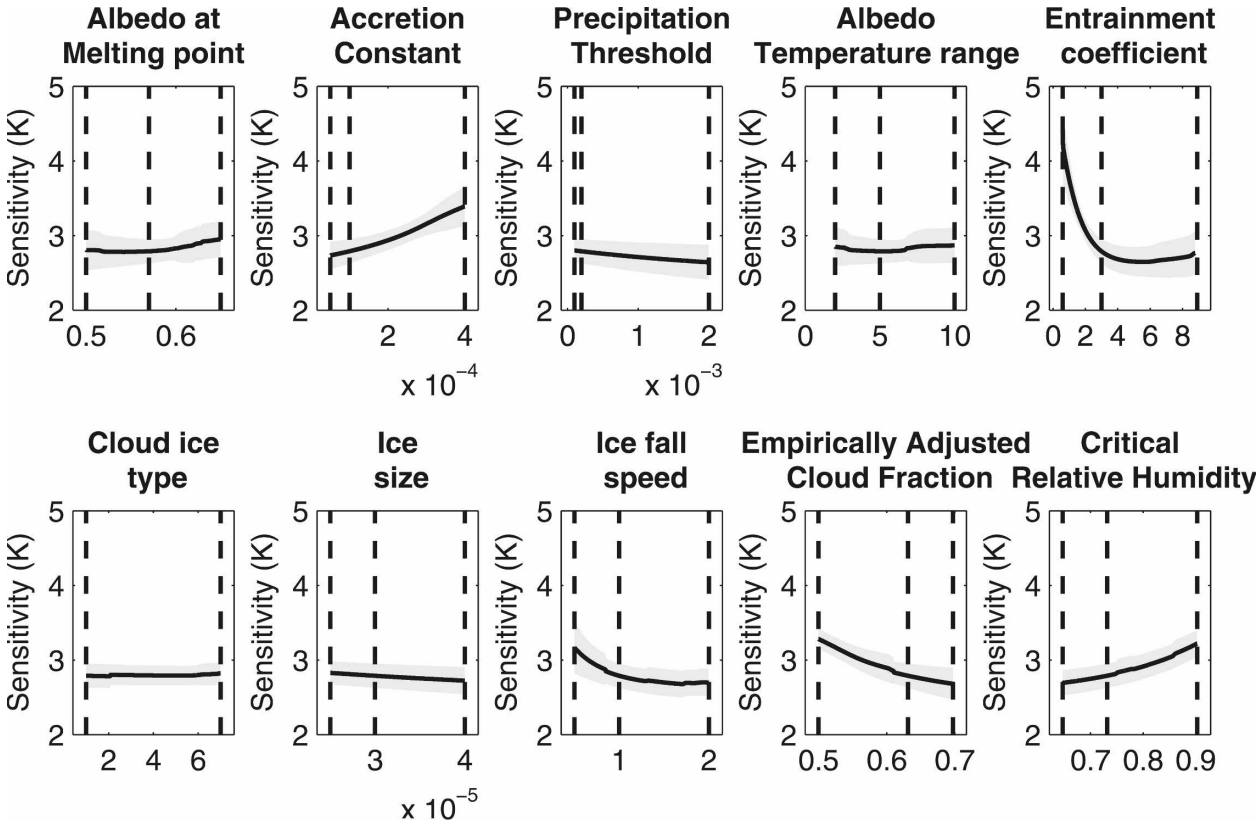

FIG. 6. A plot showing the results of a set of experiments to calculate the dependence of S on individual perturbed parameters. Each parameter value is incremented within the bounds of the climate*prediction*.net limits, while keeping all other parameters at their default value. The predicted S is calculated and plotted as a function of each parameter. The shaded area represents the 1 sigma spread in an ensemble of 10 attempts to fit the data (each using a different training set within the GCM ensemble).

design was chosen to minimize overfitting to training data, without sacrificing accuracy. The shapes of the response functions shown in Fig. 6 show a cross section of the fitted surface in each of the 10 parameter dimensions—in each case with the other 9 parameters held at the default HadAM3 value.

Previous findings (Stainforth et al. 2005) have shown that the single perturbations with the most dominant influence on S are those of the entrainment coefficient (or entcoef), critical relative humidity (RHCrit), and the ice fall speed (or VF1).

b. Monte Carlo simulation

We emulate a much larger ensemble using a Monte Carlo sampling scheme in which the value of each parameter is randomly selected between the maximum and minimum values in the GCM ensemble. Because the trained neural network is computationally inexpensive, we are able to emulate an ensemble many orders of magnitude larger than the original GCM ensemble.

We emulate a large (one million member) Monte Carlo–style ensemble in which each parameter value is ascribed a random value within the limits of the discrete parameter settings used in the climate*prediction*.net experiment. The parameters are generated randomly using an exponential probability distribution that ensures that model parameters are equally likely to be above or below the default value for HadAM3.

These models make up the emulated ANN ensemble. For each model we use the trained neural network to predict its climate sensitivity and the amplitudes of the truncated EOF set used to represent control climatology. Once we obtain an estimate for the truncated EOF amplitudes for each emulated model, we can use Eq. (1) to calculate a prediction of the model error for that simulation as compared to the observations.

As described in section 2c, we then divide the ensemble into 0.1-K bins of S and determine the best-performing models in each bin (i.e., those with the lowest E_{is}). Figure 7 shows the best models in each bin of sensitivity as simulated by the original GCM ensemble, plus the best models emulated in the ANN Monte

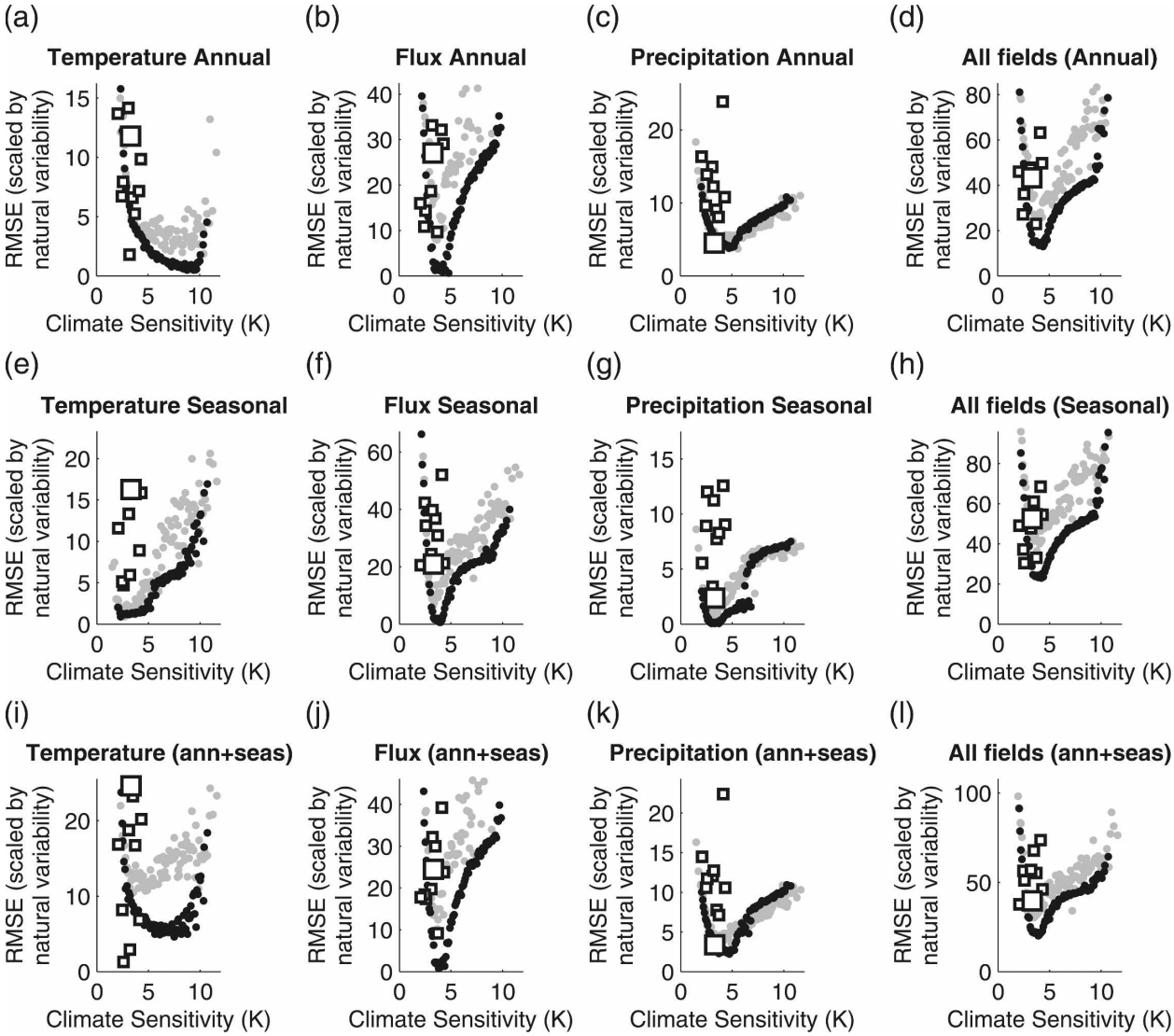

FIG. 7. (a)–(l) The error distributions $E(S)$ as produced by different observational fields. (top) Annual mean data only, (middle) seasonal differences alone, and (bottom) information on both summer and winter absolute values. The vertical axes represent the RMSE between models and observations in the respective field, in units of natural variability. Models shown by gray dots are the best-performing simulations from the original climate*prediction*.net ensemble, while black dots are the best models from the million-member Monte Carlo neural network emulated ensemble. Small squares represent the performance of models from the AMIP ensemble of atmospheric models (with values for S taken from the corresponding CMIP model). The larger square shows the performance of the unperturbed HadAM3 model. The right-hand column shows the combined error taking into account all three observation types simultaneously.

Carlo ensemble. By using different observation types, we may compare the ability of the different observations to constrain the value of S within the ANN ensemble.

We measure model performance by a selection of different criteria: first using annual regional mean surface temperatures and then again using the JJA − DJF seasonal differences. This process is then repeated for total precipitation and top of atmosphere (TOA) radiative flux balance (an expanded vector with elements for shortwave and longwave, clear-sky and cloudy-sky fluxes).

Each EOF must be scaled by an estimation of its natural variability. The control climates in the ensemble are means of a 15-yr period; hence we estimate natural variability by projecting each EOF onto 33 separate 15-yr periods in a 500-yr HadCM3 control simulation and taking the standard deviation of the pro-

jection coefficients. The principal components of this EOF in the perturbed ensemble may then be scaled using this value.

For each observation type, we also include a selection of models from the Atmospheric Model Intercomparison Project (AMIP), which is best suited for comparison with the atmospheric models with observed ocean temperatures used in the ensemble. Each model is treated identically to ensemble members; the anomaly is taken from the climate*prediction*.net mean, onto which the regional EOFs are projected for each observation type. AMIP models are not processed by the neural network and are shown for comparison only. Double CO_2 equilibrium experiments were not conducted for the AMIP ensemble, so the corresponding Coupled Model Intercomparison Project (CMIP) sensitivities are shown for each model. Sensitivities are thus provided for comparison only, though the results of Meehl et al. (2003) suggest that in several AOGCMs, the atmospheric model was dominant in determining equilibrium response.

The results show significantly different error distributions imposed by the different observations. In general, the neural network emulated ensemble tends to produce slightly smaller minimum model/observational error than the original climate*prediction*.net ensemble because of the large number of emulated models. We consider each observational constraint in turn:

- Regional temperature fields alone show no clearly defined minimum in error as a function of S; emulated models with S of up to 10 K may have control surface temperatures that are consistent with observations. The lower bound, however, is well defined—perturbed models with S less than 4 K are predicted to significantly differ from observations of annual mean surface temperature.

  The use of seasonal cycle data alone shows the most likely models occurring between 3 and 4 K. This is broadly consistent with the findings of Knutti et al. (2006), who found the seasonal cycle in temperature-constrained S to lie between 1.5 and 6.5 K at the 5%–95% confidence intervals.

  The inclusion of both annual and seasonal data in the input vector still produces a very poor constraint on sensitivity; models with S between 4 and 10 perform comparably well. Thus absolute values of the control surface temperatures of a model are a very poor predictor of response to greenhouse gas forcing. The reason for this may lie in the nature of the flux-corrected model, where control simulation ocean temperatures are adjusted to observations by corrective fluxes. We infer that this allows control mean surface temperatures to remain close to observations, even in high-sensitivity models.
- Results using radiative fluxes show a tight constraint on climate sensitivity, irrespective of the use of seasonal or annual data. The neural network is able to predict models that lie closer to observations than both the original climate*prediction*.net ensemble and the AMIP ensemble. The only models fully consistent with observations (within natural variability) are predicted within S values of 3.8–4.5 K for the annual data, and 3.9–4.2 K for the seasonal inputs.

  At high sensitivities the original ensemble produces a small number of models that score better than the emulated ANN ensemble. We attribute this to the imposed smoothness in the neural network response, which may eliminate some of the opportunity for outliers. In addition, the original climate*prediction*.net measurements of S are subject to some degree of sampling noise, especially at higher sensitivities (Stainforth et al. 2005).
- Annual mean precipitation provides a weaker constraint than the radiative fluxes, with the GCM, emulated, and AMIP ensembles all failing to reproduce annual values of precipitation. The best-performing models with S of 3.5–5.5 K have comparable errors. In contrast, many of the ANN emulated models are able to reproduce observed seasonal cycles in precipitation in an S range of 2.5–5 K.

An examination of the minimized observational discrepancies scaled by natural variability shows that the radiative flux constraint is the strongest of the three, irrespective of the use of annual mean or seasonal data. In the cases of surface temperature and precipitation, some models are able to match observed seasonal cycles, but annual mean values are not reproduced within the ensembles. These individual constraints using only seasonal data consistently show the best-performing models to lie in the range 3–5 K.

Combining all observations together, weighting each of the observations equally produces the "all fields" plots. The plot using annual mean data is only comparable with Fig. 2c in Stainforth et al. 2005, and replicates the result that shows climate*prediction*.net models with an S of greater than 10 K showing comparable RMSE to some members of the AMIP ensemble. However, the emulated ANN ensemble shows a clear minimum in model error between 4 and 5 K, an attribute that is poorly defined in the original climate*prediction*.net ensemble. Clearly, the CMIP sensitivities shown represent the most likely values for S as evaluated by a number of modeling groups, hence this ensemble should not

be expected to cover the full range of possible values for S.

It is notable that the inclusion of additional observations actually decreases the error of high sensitivity simulations relative to the most likely simulations. Even in the experiment using only seasonal data, where the separate constraints on S are consistent for the three observation types; the combined all fields plot shows an increased systematic error (the irreducible error of the best-performing model in the ensemble, at the minimum of the error curve, is significantly increased when the observations are combined into a single metric).

Although the ensemble contains models that are individually able to match the different observation types, this is achieved at the expense of making other fields less well simulated. Hence there is no parameter combination that allows all observations to be matched *simultaneously*. This more challenging requirement produces an irreducible error—the minimum error using the best-tuned model when summed over all the observations (Rougier 2007).

Thus, as the number of observational fields is increased, the error of the models with the *most likely* value of S increases from negligible to some finite irreducible value ε. However, for *less likely* values of S where a single observation produces an irreducible error, increasing the number of observational fields is unlikely to produce the same large relative increase in error. Hence, as the number of observational fields is increased, the apparent score of the best-performing models is worsened.

The methodology employed here to provide constraints on S is significantly different from that of Piani et al. (2005) or Knutti et al. (2006). While each of these papers searched for predictors of S using all members of the ensemble, we have instead used information from only the most likely possible model for each value of S. Therefore, a relation between some observable quantity and S may be stronger or weaker when all simulations are considered, compared to the method used here when only the best simulations for each value of S are used. In addition, in Knutti et al. (2006), regions where observations lay outside the entire ensemble were ignored. In contrast, in the methodology presented in this paper, such regions will influence the model "score."

Finally, we find that an increase in the number of observations used tends to increase the systematic error associated with the best-performing ensemble member, implying that (unsurprisingly) a perfect model may be impossible to achieve using only perturbations of parameters. Hence, any prediction trained within the ensemble and applied to the "perfect" observations may be to some extent an extrapolation. Piani et al. (2005) approached this issue by taking the unperturbed base model error as a crude estimate of the systematic error in the prediction, but the treatment of such errors in the prediction of S from imperfect ensembles remains an unresolved issue. However, we propose that the method illustrated here provides a systematic means of finding the irreducible component of model–observation discrepancy giving an upper limit for the systematic error, which must be included when applying ensemble-trained predictors of an unknown quantity such as S.

4. Parameter dependence

Using all observations simultaneously (the all fields case), the most likely models for each sensitivity "bin" are shown in Fig. 7. By looking at the input parameters for these models, we can examine the parameter changes necessary (if the ANN interpolation is correct) to achieve the best models at different climate sensitivities. The results (shown in Fig. 8) predict the optimal parameter settings required to produce a model of a given sensitivity, while making each model as close to observations as possible.

Also shown in Fig. 8 is the spread of each parameter setting seen in the best 100 simulations (out of a typical 10 000) in each 0.1-K bin of S. Hence, parameters showing only a small amount of spread show a unique optimal configuration for minimized error at a given value of S. Those also showing a large variation over the range of S, while remaining well constrained at any given value of S, are deemed the most important parameters for determining model response (as emulated by the ANN).

Optimum values of two parameters—the "entrainment coefficient" and the "ice fall speed"—show first order sensitivity to S. Other investigations (Stainforth et al. 2005; Sanderson and Piani 2007; Knight et al. 2007) have suggested that the entrainment coefficient is dominant in establishing different relative humidity profiles that lead to strongly different responses to greenhouse gas forcing.

Entcoef fixes the rate at which environmental air is mixed into an ensemble of simulated rising convective plumes. A high value of entcoef results in a moist midtroposphere, with weak convective activity. A low value of entcoef increases the depth of convection, transporting moisture to higher levels in the tropics (Gregory and Rowntree 1990).

A close examination of Fig. 8 shows that at any given

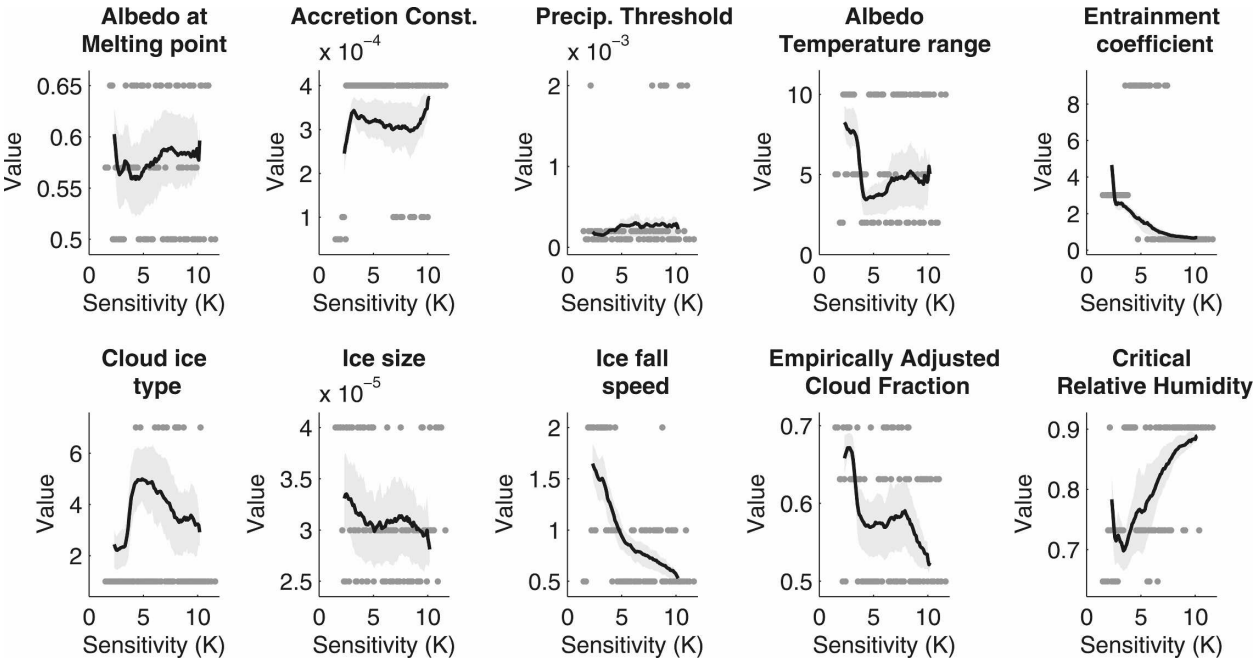

FIG. 8. A plot showing the parameter settings for the best-performing model emulated in each sensitivity bin in Fig. 7l. The black lines show the mean parameter value for the 100 best-performing models in each 0.1-K bin of climate sensitivity. The gray background shows the 10th–90th range of the 100 best-performing models. The dots show the parameter value for the best-performing model in the bin (if one exists) for the original climate*prediction*.net ensemble.

value of S, the value of entcoef is well constrained, showing no significant spread among the best-performing 100 simulations. As S rises, the value of entcoef falls monotonically from its default down to its lower limit for high values of S. The majority of the variation, however, occurs at values of S less than 6 K—indicating that other parameters are responsible for further increases in S.

Sanderson and Piani (2007) found that the reduction of entcoef caused an increase in clear-sky absorption of longwave radiation, as midtropospheric humidity was increased by strengthened convection, especially in the tropics.

The ice fall speed also shows little spread at any given value of S, and likewise is observed to reduce monotonically throughout the range of simulated S. A large value of this parameter allows the fast fallout of cloud ice. Smaller values of similar parameters in radiative–convective equilibrium models lead to increasingly moist, warm, convectively unstable atmospheric profiles (Wu 2001; Grabowski 2000).

Sanderson and Piani (2007) found that a reducing ice fall speed increased longwave clear-sky and cloudy forcing by allowing the air to remain moister. A reducing VF1 was found to increase low-level layer clouds, and this increased their positive longwave cloud feedback upon warming. The results here are consistent with those findings.

Extremes of S are achieved with additional secondary parameters:

- Low sensitivities ($S < 3$ K)—Fig. 7 makes it clear that at very low sensitivities, even the best simulated atmospheres move rapidly away from the observations. An examination of Fig. 8 shows that two parameters in particular have large variation in this region: the empirically adjusted cloud fraction (EACF) and the albedo temperature range.

  The models with the lowest S show a very large value for the empirically adjusted cloud fraction (EACF). EACF is a modification to the cloud scheme of the model that adjusts the fractional cloud coverage to observations in relation to total and condensed water; a higher value produces a greater overall cloud fraction (Wood and Field 2000). By setting this parameter to its maximum, the model cloud fraction is maximized. Meanwhile, the lowest sensitivity models also exhibit a high value for the temperature range of sea ice–albedo variation. This has the effect of increasing the effective albedo in ice-covered regions.

  Hence, it seems that the lowest sensitivities are achieved by maximally increasing albedo,

maximizing shortwave negative feedbacks upon warming. However, Fig. 7 suggests that this approach rapidly leads to unrealistic atmospheres in all three observation types.

- High sensitivities ($S > 5$ K)—the simulated models with the highest sensitivities all show entcoef and the ice fall speed to be set to low values. However, the best-performing models with high S show two additional parameter perturbations: critical relative humidity and again EACF.

  The critical relative humidity is the relative humidity at which cloud vapor will start to form (Smith 1990). It is the dominant parameter in determining the sensitivity of the simulated models with S greater than 5 K.

  In the low entcoef simulations, Sanderson and Piani (2007) found that the strong positive longwave feedback produced by the increased humidity is partly offset by a negative feedback caused by increased albedo due to high-level cirrus clouds that condense in the moist upper troposphere. The amplitude of this negative feedback is modulated by the value of RHCrit, a high value making cloud formation more difficult, thus reducing the negative albedo feedback.

  At S values of 8–9 K, RHCrit nears the upper limit defined in the GCM ensemble, and a further reduction in the negative feedback is achieved by a decrease in EACF, which is reduced to its minimum value to achieve the highest values of S in the ensemble.

  Hence, it is by suppressing cloud formation that the simulated ensemble achieves very high values of S. Without a negative shortwave response, longwave clear-sky feedbacks enhanced by high-level water vapor are left to dominate the response to warming. However, a comparison with Fig. 7 shows that this quickly causes very large discrepancies from observations of top of atmosphere radiative fluxes in the mean control state.

5. Conclusions

A two-layer feed-forward neural network was trained to accurately emulate and interpolate model output from a multithousand-member climate model ensemble. Having trained a network with data from the climate*prediction*.net dataset, we were able to predict both equilibrium temperature response and the amplitudes of leading EOFs of climatology for various different model outputs in an unseen verification set of models.

The network was successfully used to examine the equilibrium response to individual parameter changes and to smoothly interpolate between known discrete parameter settings. A much larger, neural network emulated ensemble was designed, which employed a Monte Carlo sampling scheme replacing the original climate*prediction*.net discrete sampling. The neural network was used to simulate model output from a very large ensemble in order to fully sample the parameter space within the discrete sampling of the original climate*prediction*.net experiment. The model output was divided into bins of climate sensitivity, such that in each bin a model most consistent with observations could be found.

Various different observational fields were employed, giving dramatically different constraints on climate sensitivity. The strongest constraints were found to result from observations of top of atmosphere radiative fluxes. A clear minimum in model error was apparent for models with climate sensitivities between 3 and 5 K. The simulated ensemble predicted some models to be closer to observations than all members of the climate*prediction*.net or AMIP ensembles. Seasonal data in radiative fluxes produced a similar constraint. The use of these diagnostics as tuning parameters may thus help to explain the clustering of simulated values of S in ensembles such as CMIP and AMIP.

Using only observations of surface temperature to constrain the models resulted in no upper bound constraint on S. The lower bound suggested that only models with S less than 3 K could produce reasonable annual means in surface temperature. However, observations of the seasonal cycle in temperature produced a constraint on S, with some models between 2 and 5 K in agreement with observations.

Observations of precipitation showed that all models in the climate*prediction*.net, ANN simulated, and AMIP ensembles could not reproduce annual mean data within the bounds of natural variability, hence the constraint is weaker than for the radiative case. However, the best-performing models were able to reproduce seasonality in rainfall where they could not reproduce absolute values, and models with S between 3 and 8 K could reproduce seasonal rainfall differences.

Requiring models to match all observations simultaneously proved a more difficult task for all of the ensembles. The ANN simulated ensemble suggested that model parameters could at best be tuned to a compromise configuration with a finite error from the observations. This "best model discrepancy" was found to increase with the inclusion of increasing numbers of separate observations, and was not itself a strong function of S.

Hence although models can be found to independently reproduce seasonal differences in the three observation types, there is no single model that can reproduce all three simultaneously. The relative errors of best models at different sensitivities will decrease and the irreducible error of the best-performing model increases dramatically as more observations are added. Thus the "all fields" approach yields no models that are fully consistent with observations, although it shows a minimum in error at $S = 4$ K.

Such an effect is a natural by-product of tuning an imperfect model to match observations: it is easy to tune parameters to match a single observation, but impossible to match all simultaneously. Such an effect must be considered in predictions of sensitivity such as Knutti et al. (2006) and Piani et al. (2005), where trends determined through analysis of an imperfect ensemble were applied directly to observations. We have found that the perfect model state may be unattainable through parameter perturbations alone, hence an estimation of irreducible error should be included when using ensemble-trained predictors of S.

The neural network was also used to show the parameter settings for the best-performing models over a wide range of S. We propose this as a convenient tool for the intelligent sampling of parameter space in future ensembles. For example, using the parameters suggested in Fig. 8 would provide a small, efficient ensemble containing only the most relevant models necessary for wide distribution of S. Simulation of these runs is beyond the scope of this study as the slab-ocean experiment has now ended, but this method of parameter sampling is under consideration for the next generation of climate*prediction*.net models using the Hadley Centre Global Environmental Model (HadGEM).

Furthermore, by highlighting regions of interest in the parameter space (e.g., those with steep gradients in the response function for S), efforts could be made to conduct additional simulations in those regions, further improving the fit in regions where the response is ambiguous.

We propose that a possible extension of this work with the advent of future coupled ensembles providing more comprehensive data for each model would be to evaluate the model climatology with EOFs of fully gridded data, rather than regional means. The added information in such an analysis would allow more a comprehensive metric for model verification.

Finally, the approach illustrated here is not restricted to an investigation of climate sensitivity. The method could be equally well applied to provide improved sampling and constraints for any climate model output diagnostic of interest, with potential for a multivariate predictor such as the joint probability of regional change in temperature and precipitation.

Acknowledgments. We thank all participants in the "climate*prediction*.net" experiment, and those who have worked toward its continuing success. We also thank the CMIP II modeling groups, the ERBE team, and the NCEP reanalysis team for the use of their data.

REFERENCES

Annan, J. D., J. C. Hargreaves, N. R. Edwards, and R. Marsh, 2005a: Parameter estimation in an intermediate complexity earth system model using an ensemble kalman filter. *Ocean Modell.,* **8,** 135–154.

——, ——, R. Ohgaito, A. Abe-Ouchi, and S. Emori, 2005b: Efficiently constraining climate sensitivity with ensembles of paleoclimate simulations. *Sci. Online Lett. Atmos.,* **1,** 181–184.

Collins, M., B. B. B. Booth, G. R. Harris, J. M. Murphy, D. M. H. Sexton, and M. J. Webb, 2006: Towards quantifying uncertainty in transient climate change. *Climate Dyn.,* **27,** 127–147.

Frame, D. J., B. B. B. Booth, J. A. Kettleborough, D. A. Stainforth, J. M. Gregory, M. Collins, and M. R. Allen, 2005: Constraining climate forecasts: The role of prior assumptions. *Geophys. Res. Lett.,* **32,** L09702, doi:10.1029/2004GL022241.

Giorgi, F., and R. Francisco, 2000: Uncertainties in regional climate change predictions. A regional analysis of ensemble simulations with the HadCM2 GCM. *Climate Dyn.,* **16,** 169–182.

Grabowski, W. W., 2000: Cloud microphysics and the tropical climate: Cloud-resolving model perspective. *J. Climate,* **13,** 2306–2322.

Gregory, D., and P. Rowntree, 1990: A mass flux convection scheme with representation of cloud ensemble characteristics and stability dependent closure. *Mon. Wea. Rev.,* **118,** 1483–1506.

Hagan, M. T., and M. Menhaj, 1994: Training feedforward networks with the marquart algorithm. *IEEE Trans. Neural Net.,* **5,** 989–993.

——, H. B. Demuth, and M. H. Beale, 1996: *Neural Network Design*. PWS Publishing.

Hansen, J., M. Allen, D. Stainforth, A. Heaps, and P. Stott, 2001: Casino-21: Climate simulation of the 21st century. *World Resour. Rev.,* **13,** 187–198.

Hsieh, W. W., and B. Tang, 1998: Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bull. Amer. Meteor. Soc.,* **79,** 1855–1870.

Johns, T. C., and Coauthors, 2003: Anthropogenic climate change for 1860 to 2100 simulated with the HadCM3 model under updated emissions scenarios. *Climate Dyn.,* **20,** 583–612.

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

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

Meehl, G. A., W. M. Washington, and J. M. Arblaster, 2003: Factors affecting climate sensitivity in global coupled climate models. Preprints, *14th Symp. on Global Change and Climate Variations,* Long Beach, CA, Amer. Meteor. Soc., 2.1.

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

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

Pope, V. D., M. L. Gallani, P. R. Rowntree, and R. A. Stratton, 2000: The impact of new physical parameterizations in the Hadley Centre climate model, HadAM3. *Climate Dyn.,* **16,** 123–146.

Rodwell, M. J., and T. N. Palmer, 2007: Using numerical weather prediction to assess climate models. *Quart. J. Roy. Meteor. Soc.,* **133** (622A), 129–146.

Rougier, J., 2007: Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climatic Change,* **81,** 247–264.

Sanderson, B. M., and C. Piani, 2007: Towards constraining climate sensitivity by linear analysis of feedback patterns in thousands of perturbed-physics gcm simulations. *Climate Dyn.,* **30** (2–3), 175–190.

Smith, R. N. B., 1990: A scheme for predicting layer clouds and their water content in a general circulation model. *Quart. J. Roy. Meteor. Soc.,* **116** (492), 435–460.

Stainforth, D., J. Kettleborough, M. Allen, M. Collins, A. Heaps, and J. Murphy, 2002: Distributed computing for public-interest climate modeling research. *Comput. Sci. Eng.,* **4,** 82–89.

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

Stott, P., and J. R. Kettleborough, 2002: Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature,* **416,** 723–725.

Vogl, T., J. Mangis, J. Rigler, W. Zink, and D. Alkon, 1988: Accelerating the convergence of the backpropagation method. *Biol. Cybern.,* **59,** 257–263.

Wood, R., and P. R. Field, 2000: Relationships between total water, condensed water, and cloud fraction in stratiform clouds examined using aircraft data. *J. Atmos. Sci.,* **57,** 1888–1905.

Wu, X., 2001: Effects of ice microphysics on tropical radiative–convective–oceanic quasi-equilibrium states. *J. Atmos. Sci.,* **59,** 1885–1897.