A Probability and Decision-Model Analysis of a Multimodel Ensemble of Climate Change Simulations

JOUNI RÄISÄNEN

Rossby Centre, Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

T. N. PALMER

European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

(Manuscript received 2 October 2000, in final form 31 January 2001)

ABSTRACT

Because of the inherent uncertainties in the computational representation of climate and because of unforced chaotic climate variability, it is argued that climate change projections should be expressed in probabilistic form. In this paper, 17 Coupled Model Intercomparison Project second-phase experiments sharing the same gradual increase in atmospheric CO_2 are treated as a probabilistic multimodel ensemble projection of future climate. Tools commonly used for evaluation of probabilistic weather and seasonal forecasts are applied to this climate change ensemble. The probabilities of some temperature- and precipitation-related events defined for 20-yr seasonal means of climate are first studied. A cross-verification exercise is then used to obtain an upper estimate of the quality of these probability forecasts in terms of Brier skill scores, reliability diagrams, and potential economic value. Skill and value estimates are consistently higher for temperature-related events (e.g., will the 20-yr period around the doubling of CO_2 be at least 1°C warmer than the present?) than for precipitation-related events (e.g., will the mean precipitation decrease by 10% or more?). For large enough CO_2 forcing, however, probabilistic projections of precipitation-related events also exhibit substantial potential economic value for a range of cost–loss ratios. The treatment of climate change information in a probabilistic rather than deterministic manner (e.g., using the ensemble consensus forecast) can greatly enhance its potential value.

1. Introduction

Probability forecasting using ensembles of deterministic integrations has become an established technique in short-, medium-, and extended-range weather forecasting [see Palmer (2000) for a recent review]. The flow-dependent growth in initial-condition uncertainty is represented by the dispersion of an ensemble, whose members start from not-quite-identical initial conditions. Well-calibrated probability forecasts are intrinsically more reliable than single deterministic forecasts. They also have a higher potential economic value in weather-sensitive risk management, for a range of possible forecast users. In this context, reliability is a precise and quantifiable quantity, forming a component of the well-known Brier score. ''Potential economic value'' is also a quantitative measure, based on a cost–loss ratio decision-making model already discussed by Anders Ångström around 1920 (see Liljas and Murphy 1994) and subsequently elaborated and used by several authors (e.g., Thompson 1952; Murphy 1977; Katz and Murphy 1997; Richardson 2000; Palmer et al. 2000).

The most traditional form of climate change prediction is not an initial value problem. Rather, the response of the invariant long-term statistical properties of climate to a change in external forcing such as a prescribed increase in CO_2 is sought. Even in this type of prediction, however, fundamental uncertainties are present. As discussed recently in Palmer (2001), the parameterization problem in climate modeling is not well posed and there may be fundamental and irreducible uncertainties in the representation of unresolved processes in deterministic models of climate. This suggests that climate change predictions could also benefit from an ensemble-based probabilistic approach, but where individual ensemble members differ, not necessarily in terms of initial state, but in terms of the computational representation of climate.

In practice, the traditional steady-state climate change experiment is of somewhat theoretical interest. In the real world, forcing conditions such as the concentration of CO_2 are changing with time. Thus, rather than asymptotic statistics, the response at a certain time in the future is sought. Predictions of such transient, time-

Corresponding author address: Jouni Räisänen, Rossby Centre, Swedish Meteorological and Hydrological Institute, S-60176, Norrköping, Sweden.
E-mail: jouni.raisanen@smhi.se

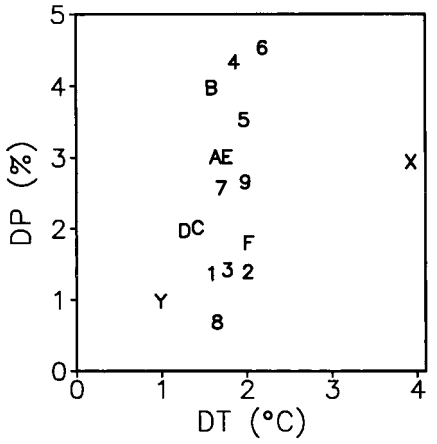

FIG. 1. Changes in global mean temperature (°C) and precipitation (%) at doubling of CO_2 in the 17 experiments. 1 = BMRC; 2 = CCC; 3 = CCSR/NIES; 4 = CERFACS; 5 = CSIRO; 6 = GFDL; 7 = GISS; 8 = IAP/LASG; 9 = LMD/IPSL; A = ECHAM3; B = MRI; C = NCAR-CSM; D = NCAR/DOE-PCM; E = HadCM2; F = HadCM3; X = NCAR–WM; Y = NRL.

dependent climate change are affected by uncertainty in the initial state as well as model uncertainty. However, for time horizons of several decades and longer, any detailed memory of the initial conditions has probably been virtually lost, so that the simulated climate statistics during a given period may be anywhere within the probability distribution determined by the external forcing and the model's internal variability. Last, predictions of real-world climate change are also affected by uncertainty in the external forcing, such as the rate of future CO_2 increase. Ensembles describing future climate change should ideally take into account all these sources of uncertainty (New and Hulme 2000).

A straightforward and pragmatic approach to the representation of uncertainty in the computational description of climate is through the multimodel ensemble. The dataset collected for the second phase of the Coupled Model Intercomparison Project (CMIP2; Meehl et al. 1997, 2000) provides a 17-member ensemble of model integrations with both present-day and increasing levels of CO_2 (however, the increase in CO_2 is the same in all the experiments, so that forcing-related uncertainty is not represented). Using this dataset, we address in this paper the question of whether a probabilistic representation and validation of climate change experiments, following the techniques used in short-, medium-, and extended-range weather prediction, is viable.

The question of validation of a probabilistic forecast of climate change is immediately problematic. First, a probability forecast cannot be verified from a single occurrence, unless the probabilities are 100% or 0%. So, if a climate change ensemble predicts that it will become drier over Stockholm 50 yr from now with 85% probability, and it turns out to be wetter, then it is impossible to say that this probability forecast was wrong. Second, even if the forecast probability was 100%, we would still have to wait 50 yr to see if the prediction was correct.

There are two approaches to these problems that are followed in this paper. First, we do not validate the forecasts over just one grid point. If we take all the grid points where it is predicted to be drier in 50 years' time, say with 85% probability, then, of these grid points, we would expect the actual precipitation to decrease in about 85% of cases. If the actual fraction of cases that are drier is significantly different from this, then the probability forecast should be considered wrong (more specifically ''unreliable'').

Second, it will be assumed that the individual components of the CMIP2 multimodel ensemble are realistic and equally likely representations of climate. With this assumption, it is enough to take at random, one member of the ensemble (which is then removed from the forecast ensemble) and treat it as if it were ''truth.'' Of course, in practice, the value of this procedure will be compromised if the ensemble of models are more similar to one another than they are to nature. It has been argued in Palmer (2001) that this may well be so from a set of GCMs formulated using deterministic equations. Moreover, as noted above, the CMIP2 experiments share the same idealized forcing scenario, which means that the uncertainty in future climate forcing is not reflected. As such, the present study could be thought of, not as giving definitive probabilistic predictions of climate change, but as setting out a basic template for the presentation and evaluation of such predictions.

One of the key results in this paper is that a probabilistic forecast of climate change can have significantly greater value than a deterministic forecast provided by the consensus or ensemble mean of available predictions. Yet the consensus forecast is commonly used to convey to potential users the results of climate change experiments, often in the mistaken belief that potential users would find probabilistic forecasts difficult to interpret.

In section 2, the basic CMIP2 dataset is described. In section 3, some examples of the resulting probability forecasts of climate change are shown. Brier score and reliability measures associated with these probability forecasts are discussed in section 4. In section 5, we describe the application of the cost–loss ratio decision analytic model to the climate change problem. The main findings are summarized and their implications are discussed in section 6.

2. The multimodel ensemble

CMIP2 (Meehl et al. 1997, 2000), compares the response of current atmosphere–ocean general circulation models to increasing CO_2. All the participating models share the same idealized forcing scenario, a 1% yr^{-1} compound increase in CO_2. This gives a doubling of CO_2 in 70 years. In the present study, the 17 models

available in September 2000 are included: Bureau of Meteorology Research Center (BMRC; Power et al. 1993); Canadian Centre for Climate Modelling and Analysis (CCC; Flato et al. 2000); Center for Climate System Research/National Institute of Environmental Studies (CCSR/NIES; Emori et al. 1999); Centre Europeen de Recherche et de Formation Avancée en Calcul Scientifique (CERFACS; Barthelet et al. 1998); Commonwealth Scientific and Industrial Research Organisation (CSIRO; Hirst et al. 2000); Geophysical Fluid Dynamics Laboratory (GFDL; Manabe et al. 1991); Goddard Institute for Space Studies (GISS; Russell and Rind 1999); Institute of Atmospheric Physics, Chinese Academy of Sciences (IAP/LASG; Zhang et al. 2000); Laboratoire de Metéorologie Dynamique, Institut Pierre Simon Laplace (LMD/IPSL; Braconnot et al. 1997); Max Planck Institute for Meteorology (MPI-ECHAM3; Voss et al. 1993); Meteorological Research Institute (MRI; Tokioka et al. 1995); National Center for Atmospheric Research Climate System Model (NCAR CSM; Boville and Gent 1998); NCAR Department of Energy (DOE) Parallel Climate Model (PCM; Washington et al. 2000); NCAR (Washington and Meehl 1996); Naval Research Laboratory (NRL; Li and Hogan 1999); the United Kingdom Met Office (UKMO-HadCM2; Johns et al. 1997); and UKMO-HadCM3 (Gordon et al. 2000).

Generally, 80 yr of data are available both for the transient greenhouse runs with increasing CO_2 concentration, and the control runs with constant (same as in the beginning of the greenhouse runs) CO_2. For NRL, however, only a 3-yr control run (but a full 80-yr greenhouse run) is available. We retain NRL in the present study because this is the CMIP2 model with the smallest global warming. Excluding such outliers could make the results unduly optimistic.

The control climates are here defined, when possible, as means over the whole 80-yr control simulations. The NRL control climate is formed by combining the 3-yr control run with the first 7 yr of the greenhouse run, when the CO_2 forcing is still small. The climate in the transient greenhouse runs is defined using a 20-yr averaging window. Thus, for calculating climate changes at the doubling of CO_2, the control run means are subtracted from greenhouse run means over the years 61–80. The global annual-mean warming at this time is smallest, 1.0°C, in the NRL model and largest, 3.9°C, in the oldest NCAR (Washington and Meehl 1996) model. The remaining 15 models are within 1.3°–2.2°C. Changes in global precipitation vary more evenly between 0.7% and 4.5% (see Fig. 1). To compare these figures with the possible impact of control run climate drift, the largest global temperature difference between the years 61–80 in the control run and the whole 80-yr mean is 0.25°C, and the difference in precipitation 0.65% (both in GISS). Much larger local control run trends occur in some models in some areas, most notably in temperature over the high-latitude Southern Ocean.

TABLE 1. Globally averaged probabilities of the events DT $>$ 0, DT $>$ 1°C, DP $<$ $-$10%, DP $>$ 0, and DP $>$ 10%, all for 20-yr JJA means. Results are shown for the control simulations (CTRL) and for four 20-yr periods in the transient greenhouse runs (GHG).

	DT $>$ 0	DT $>$ 1°C	DP $<$ $-$10%	DP $>$ 0	DP $>$ 10%
CTRL	50.5%	1.2%	10.2%	49.2%	10.4%
GHG 1–20	73.1%	1.7%	10.7%	48.5%	9.7%
GHG 21–40	95.5%	14.0%	13.6%	51.2%	11.3%
GHG 41–60	98.8%	54.4%	16.7%	54.5%	17.0%
GHG 61–80	98.9%	82.2%	20.3%	56.2%	24.1%

Such drift may both contribute to intermodel differences in climate change and amplify the estimated natural variability, but not to the extent that this would affect the conclusions of the present study.

The CMIP2 multimodel ensemble is here treated as a probabilistic climate change forecast. The probabilities for different events related to climate change (e.g., that the warming at the doubling of CO_2 will exceed 1°C) are inferred by counting the number of models in which the event occurs. For comparison, the probability of similar anomalies in the control simulations is computed. This is made by selecting 13 partly overlapping[1] (years 1–20, 6–25, . . . , 61–80) 20-yr periods from the control runs and counting the number of periods when the difference from the mean for the remainder of the control run is above or below the threshold considered. These probabilities are then averaged over 16 models, excluding NRL (the control run probabilities for individual models are not used individually because they are subject to large sampling variability). The 16-model mean control run ''climatological'' probabilities estimated in this way are generally in good agreement with the climate change probabilities calculated for the first 20-yr periods of the greenhouse runs, when the impact of the CO_2 forcing on the probability distributions is still small.

Differences in climate change between the CMIP2 experiments result, disregarding the effects of climate drift, from two factors: differences between the models themselves, and internal variability manifested as a sensitivity of the simulated climate change to the control and greenhouse run initial conditions. Räisänen (2001) estimated the relative importance of these factors for the interexperiment variance in climate change. For the 20-yr period 61–80 centered at the doubling of CO_2, neither factor is negligible but model differences appear the more important, at least regarding temperature change [even when the NRL and NCAR (Washington and Meehl 1996) (NCAR–WM) models with the smallest and largest global warming are excluded]. Earlier

[1] The results in this paper would remain very similar even without the subsampling.

during the experiments when the CO_2 forcing is weaker, internal variability is, in relative terms, more important.

3. Climate change probabilities

As a somewhat arbitrary example of the CMIP2 results, the 17-model ensemble means of temperature and precipitation change in the northern summer [June–July–August (JJA)] around the doubling of CO_2 (years 61–80) are shown in Figs. 2a,b. The ensemble mean relative precipitation change is calculated as the relative difference between the greenhouse and control run ensemble means, which avoids problems in those areas where the control run precipitation in some individual model(s) is very near zero. Dark shading marks areas where the ensemble mean temperature change (DT) exceeds 1°C or the ensemble mean precipitation change (DP) is below −10%. The former is true almost everywhere, excluding the Arctic Ocean[2] and a few locations in the North Atlantic and the Southern Ocean. Areas with ensemble mean DP < -10 % are much less common. The widest of these covers the region surrounding the Mediterranean, and a part of southwestern Asia. In most of the world, however, the ensemble mean indicates an increase in precipitation.

The probability of the events DT > 1°C and DP < -10%, estimated by the fraction of the 17 experiments in which the events occur, is shown in Figs. 2c,d. As expected, the occurrence of these events in the ensemble mean does not generally imply a probability of 100%. Conversely, the events commonly occur in some models even where this is not the case with the ensemble mean. There are some areas with DT > 1°C in all models (as there are only 17 experiments, probabilities exceeding the highest contour of 98% are all 100%) but these are quite few, partly because the warming in NRL is below this threshold in about 70% of the world. Likewise, there are some areas where none of the models indicates JJA precipitation to decrease by 10% or more, mainly over the Southern Ocean. Otherwise, the inferred probabilities are above 0% but below 100%. As expected from the ensemble means, DT > 1°C seems in most areas a much more probable event than DP < -10%.

The average frequency of the two events in the control simulations is shown in Figs. 2e,f. The 20-yr mean JJA control run temperature anomalies exceeding 1°C are almost nonexistent, except in the high-latitude Southern Ocean, where, however, the estimated probabilities may have been amplified by climate drift in some models. Negative precipitation anomalies exceeding 10% are more frequent in the control simulations, although even in this case the probability is geographically variable. It ranges from nearly zero over the Southern Ocean to 25%–50% in arid subtropical areas in both hemispheres. Obviously, the contrast between the greenhouse run and control run probabilities is much larger for DT > 1°C than for DP < -10%.

Globally averaged probabilities of these two events, and three other events related to the JJA climate, are given in Table 1 for the control runs and different 20-yr periods in the transient greenhouse runs. In the first 20-yr period, the average greenhouse run probabilities are close to those in the control runs, except for DT > 0. Although the CO_2 forcing at this time is still weak, it is already sufficient to make warm anomalies substantially more commonplace than cold anomalies. Later during the greenhouse runs, the gradually increasing CO_2 makes DT > 0 a virtually ''certain'' event in most of the world, and DT > 1°C also becomes dramatically more probable. There is likewise a slight increase in the frequency of positive precipitation anomalies (DP > 0).

The diametrically opposite events DP < -10% and DP > 10% both also grow more common in the course of time, roughly by a factor of 2 by the doubling of CO_2. While the magnitude of precipitation variability at any given point might change with increasing CO_2, this alone is unlikely to explain these increased probabilities. More likely, they are mainly due to the fact the change in the ''noise-free'' mean precipitation (around which internal variability operates) differs in sign between different parts of the world. A simple numeric example is illustrated in Fig. 3. Assume that 20-yr JJA precipitation anomalies in a control run are normally distributed with a standard deviation of 8%. In that case, DP < -10% and DP > 10% both have a probability of 10.6%. Further assume that the increase in CO_2 forces an 8% increase in the noise-free mean precipitation in one-half of the world (GHG+) and an 8% decrease in the other half (GHG−), with no change in the magnitude of internal variability. In GHG+, 10% wet anomalies relative to the control run mean are now much more common (40.1%). The 10% dry anomalies become quite rare (1.2%), but the absolute decrease in their frequency is smaller than the increase in the frequency of the 10% wet anomalies. Similarly, in GHG−, 10% dry anomalies increase more in frequency than 10% wet anomalies decrease. As averaged over both areas, the two events both become more common by almost a factor of 2 in this example. In practice, the forced precipitation change at a given point also frequently differs in sign between different models, so that, when taken over all CMIP2 experiments, there are areas in which both DP < -10% and DP > 10% become more common.

4. Brier skill scores and reliability diagrams

A basic verification statistic for probability forecasts is the Brier score (Brier 1950; Wilks 1995). As evaluated over N forecast cases,

[2] The muted warming over the Arctic Ocean is specific to summer, when melting ice buffers the simulated temperatures near the freezing point. In the annual mean, a maximum in warming occurs in this area (e.g., Räisänen 2001).

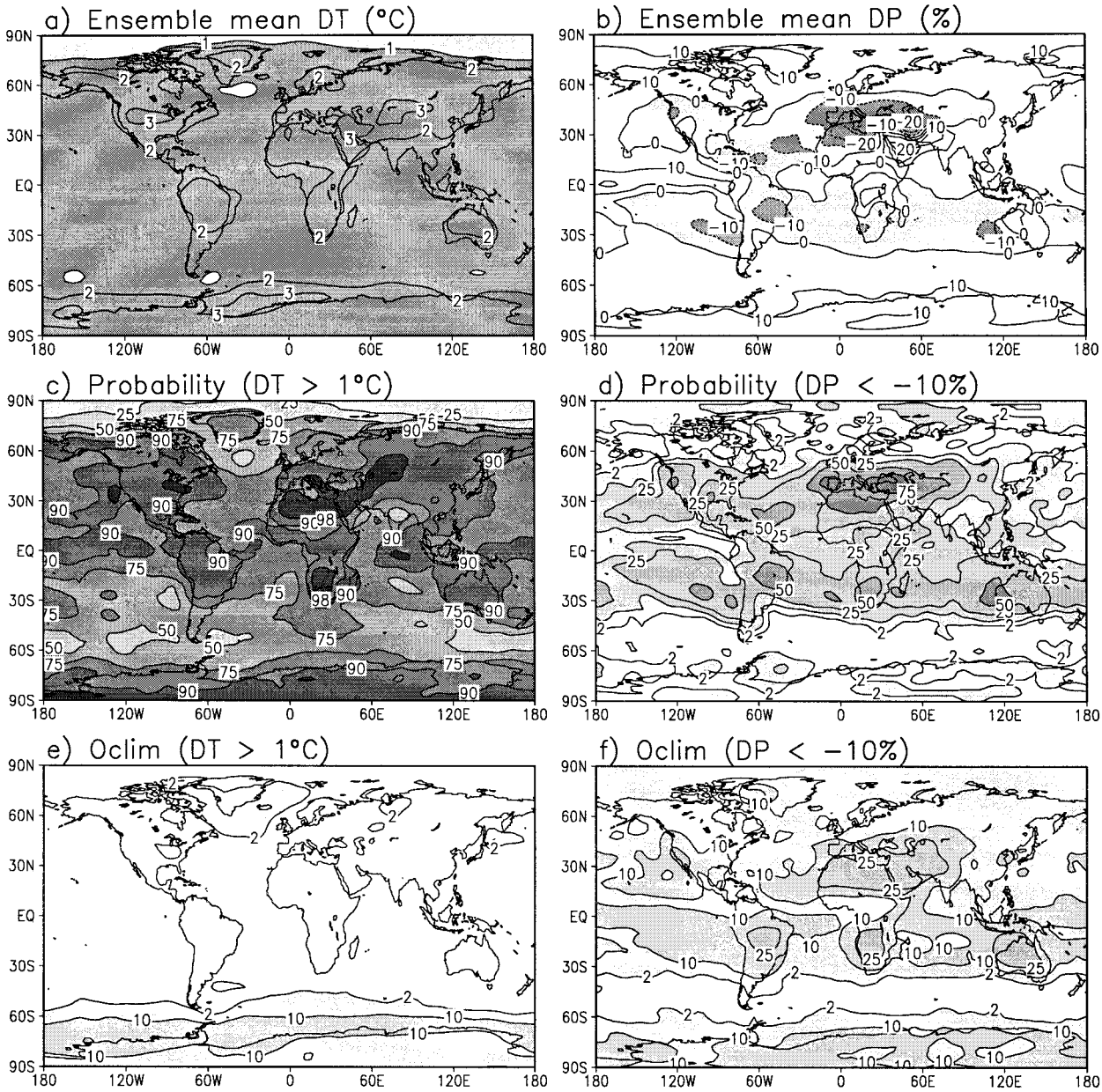

FIG. 2. The 17-model mean changes in Jun–Aug mean (a) temperature DT and (b) precipitation DP around the doubling of CO_2 (years 61–80). Areas where the average DT exceeds 1°C or the average DP is below −10% are shaded in dark. Areas with $0 < \mathrm{DT} < 1$°C or $-10\% < \mathrm{DP} < 0$ are shaded in light. The fraction of models (%) with (c) DT > 1°C and (d) DP < −10%. The probability of obtaining (e) DT > 1°C and (f) DP < −10% as a result of internal variability, as estimated from the control simulations. A slight smoothing is applied for legibility.

$$b = \frac{1}{N}\sum_{i=1}^{N}(p_i - v_i)^2; \qquad 0 \le p_i \le 1, \qquad v_i \in \{0, 1\}, \tag{1}$$

where p_i is the forecast probability of event E in case i and v_i is the actual verified occurrence of the event. Here $v_i = 1$ implies that the event occurs and $v_i = 0$ that it does not. For assessing if the forecast has any nontrivial skill, b is commonly converted to the Brier skill score

$$B = 1 - \frac{b}{b_{\mathrm{cli}}}, \tag{2}$$

where b_{cli} is the climatological Brier score, that is, the Brier score for a forecast system which always predicts E to occur with its average observed probability $\overline{o}_{\mathrm{cli}}$.

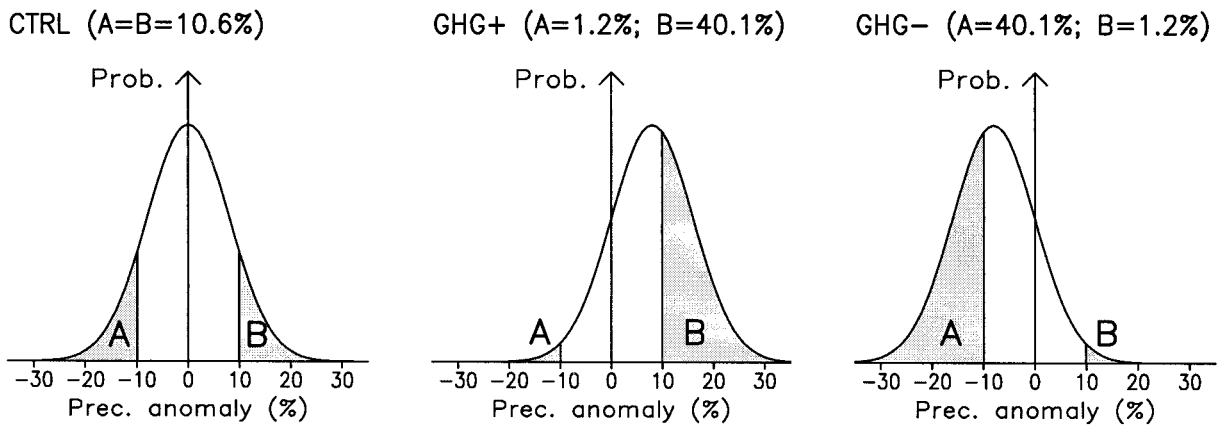

FIG. 3. Schematic example of how the probability distribution of precipitation anomalies is affected by climate change. (left) Control run with a Gaussian distribution and a standard deviation of 8%. (middle) Greenhouse run with an 8% forced increase but no change in internal variability. (right) Greenhouse run with an 8% forced decrease. Here A (B) denotes the probability of over 10% dry (wet) anomalies with respect to the noise-free control run mean.

These statistics are affected by both the physical fidelity of the forecast system and the ensemble size. For an extreme example, assume that E occurs in 50% of cases in both reality and the forecasts, but the forecast ensemble members are uncorrelated with each other and the actual outcome. Then, an infinite forecast ensemble would always indicate $p_i = 0.5$, which leads to $b = 0.25$. Likewise, if $\overline{\sigma}_{\rm cli}$ is derived from a large sample, this is close to 0.5 and $b_{\rm cli}$ approaches 0.25. By contrast, if the forecast ensemble consists of only one member, the forecast probability is always either 0 or 1 and the expected b is 0.5 (because 50% of the forecasts are wrong). Thus, in the usual case that the forecast ensemble is smaller than the dataset used to derive $\overline{\sigma}_{\rm cli}$, pure chance will result in negative skill scores.

As there are no observations of future climate, these statistics cannot be evaluated directly for the probabilistic CMIP2 forecast. Rather, the cross-validation approach outlined in the introduction is used. Seventeen realizations of the Brier score b are calculated, verifying at each time a 16-model probability forecast against the one remaining model. Global averages are obtained by weighting, in the summing indicated by (1), individual grid points with the cosine of latitude. Similarly, $b_{\rm cli}$ is calculated separately for each of the 17 verifying models using the 16-model mean control run climatological probabilities discussed above. As mentioned, this cross verification is blind to errors shared by all 17 models in their response to increased CO_2 and it gives therefore an upper estimate of the actual prognostic skill of the ensemble.

TABLE 2. Global CMIP2 cross-verification Brier skill scores for events related to JJA climate. The skill scores on the first four rows are calculated from the 17-case means of b and $b_{\rm cli}$, evaluated for 20-yr JJA mean anomalies of temperature and precipitation. For the last 20-yr period, the lowest (61–80, min) and highest (61–80, max) single realizations are also given, as well as the average skill score for yearly (rather than 20-yr mean) JJA anomalies (61–80, seasonal).

Years	DT > 0	DT > 1°C	DP < −10%	DP > 0	DP > 10%
1–20	0.21	−0.03	−0.06	−0.05	−0.05
21–40	0.84	0.22	−0.02	0.00	−0.04
41–60	0.95	0.59	0.05	0.09	0.03
61–80	0.96	0.83	0.12	0.14	0.15
61–80, min	0.78	−1.11	−0.65	−0.27	−0.06
61–80, max	0.99	0.94	0.27	0.23	0.25
61–80, seasonal	0.87	0.73	−0.03	−0.02	−0.02

Brier skill scores for some events related to 20-yr means of JJA climate in the greenhouse runs are given in Table 2. The (weighted) average skill scores in the first four rows were obtained by substituting into (2) the 17-case means of b and $b_{\rm cli}$. As expected, the two temperature-related events (DT $>$ 0 and DT $>$ 1°C) yield much higher skill scores than the three precipitation-related events (DP $<$ −10%, DP $>$ 0 and DP $>$ 10%). In addition, reflecting the increase in greenhouse gas forcing with time in the experiments, the average skill scores also increase with time. During the first 20-yr period, the score is still slightly negative for four of the five events, basically as a result of larger sampling variability in the 16-model ensemble climate change probabilities than in the climatological probabilities averaged over 16 models and the whole 80-yr control runs. The only event with a positive average skill score in the years 1–20 is DT $>$ 0. Even during the later parts of the experiments, B is higher for this event than for DT $>$ 1°C. The models agree very well that there will be some warming at almost all points around the world, but somewhat less well on whether the warming during a given period will be larger or smaller than any positive threshold.

The discussed average skill scores are relatively insensitive to individual outlying models. For example, while the exclusion of the NRL and NCAR–WM models with the smallest and largest global warming yielded an increase in B for the DT $>$ 1°C event, this increase was

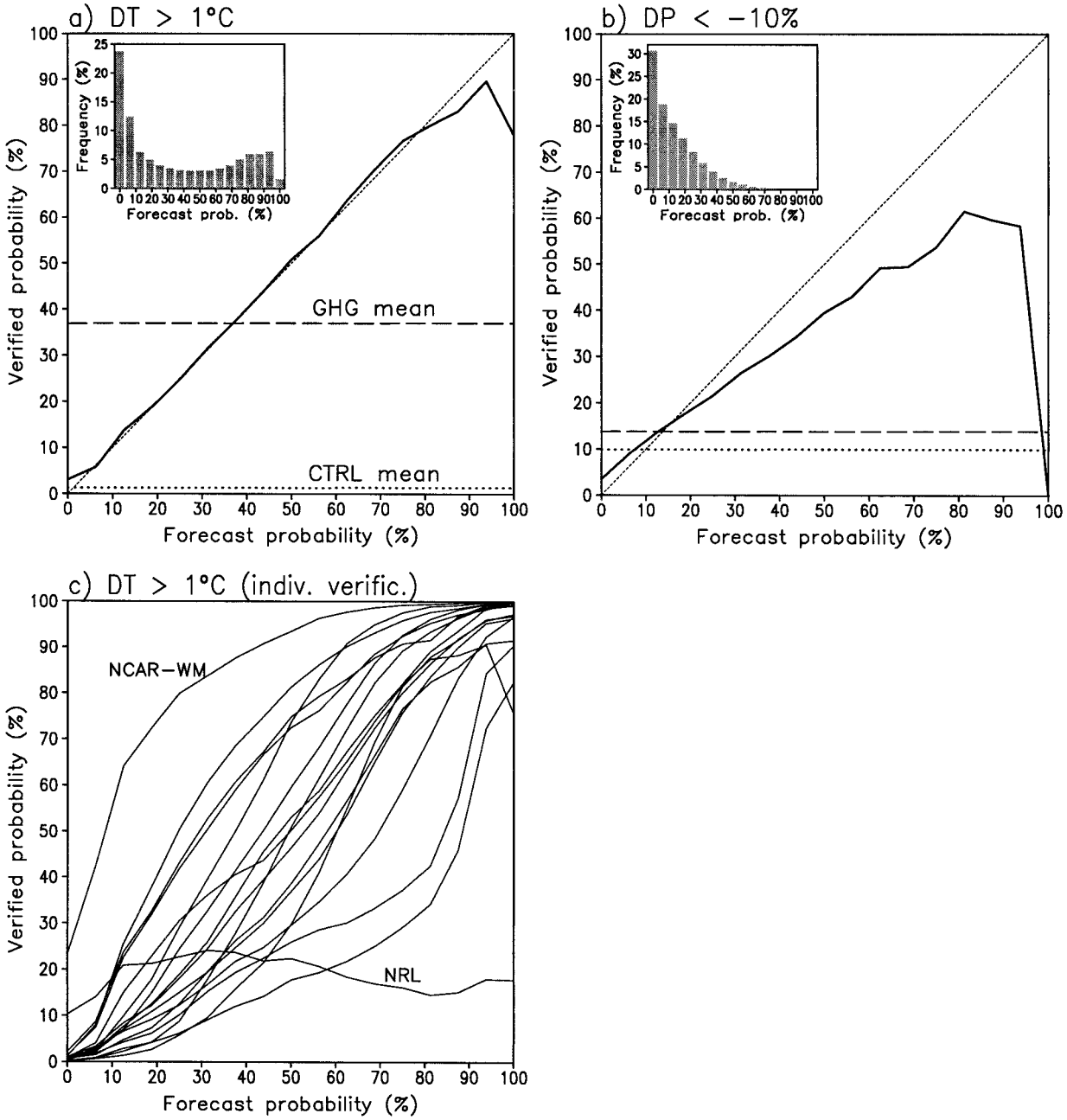

FIG. 4. Reliability diagrams for the events (a) DT $>$ 1°C and (b) DP $<$ −10%. The verified probability is averaged over the 17 choices of the 16-model forecast ensemble and the one verifying model. (c) As in (a), but the individual 17 verifications are shown separately. The horizontal dashed and dotted lines indicate the average frequencies of the events in the 80-yr greenhouse runs and in the control runs, respectively. Frequency histograms of the forecast probabilities are given in the inset panels.

not larger than 0.03 in any 20-yr period. On the other hand, averaging over several realizations of climate change is only possible with model data: in the real world only one realization of future climate will occur. For this reason it is important to note that the calculated skill scores do vary substantially between the individual verifying models. The fifth and sixth rows in Table 2 give the lowest and highest individual scores (out of 17) at the doubling of CO_2. Excluding DT $>$ 0, there is always at least one outlier in the dataset whose climate changes are predicted worse in terms of the Brier score by the remaining 16 models than by the control run climatology. The very low minimum skill score for DT $>$ 1°C is for verification with the NRL model. In this model, the warming is still below 1°C in about 70% of the world, whereas it generally exceeds this threshold

in all other models. The climatological forecast that indicates DT $>$ 1°C to be very improbable is therefore in this case closer to the verifying truth than the high probabilities indicated by the remaining 16-model ensemble. Conversely, if some of the models with close-to-average climate changes turns out to be right, then the observed skill score will be higher than is indicated by the mean over the 17 realizations.

These Brier skill scores for the temperature-related events are much higher than the scores obtained from multimodel seasonal-timescale ensemble integrations. For example, Palmer et al. (2000) report, for an ensemble of atmospheric GCM simulations forced by observed sea surface temperatures in 14 different years, a Brier skill score of 0.12 for the event DT(850 hPa) $<$ 0. Even when a perfect model assumption was used (i.e., basing verification on one member of the ensemble), the skill score at most increased to 0.29. This indicates that beyond year 20, the forcing from enhanced CO_2 in the CMIP2 experiments is stronger relative to unforced bidecadal temperature variability in atmosphere–ocean GCMs, than the forcing from anomalous sea surface temperatures relative to unforced interannual atmospheric variability.

The unforced variability of 20-yr seasonal means is naturally much smaller than the variability of individual seasonal means. For a more fair comparison with the seasonal forecasts, the average CMIP2 cross-verification skill scores are also given for temperature and precipitation anomalies in individual JJA seasons during the years 61–80 (last row of Table 2). For DT $>$ 0 and DT $>$ 1°C, these scores are only moderately lower than the corresponding scores for 20-yr JJA means. This indicates that, around the doubling of CO_2, the simulated temperature changes are very substantial even when compared with interannual variability. By contrast, no positive skill scores are obtained for the precipitation-related events on this timescale.

A good probabilistic forecast system should be *reliable.* This means that, when one takes all cases in which the forecast probability of event E is p%, then E should actually occur in about p% of these cases. In addition, the system should have *resolution.* A system that always predicts the same probability will be perfectly reliable if this probability equals the average frequency of E in the verification data, but the forecasts are not very useful. A probabilistic system is said to have good resolution if the forecast probability varies from case to case, and the average verified frequencies for the low- and high-probability forecasts are substantially different. A perfect deterministic system would have both perfect reliability and perfect resolution: it would only forecast probabilities of 0% and 100%, and the verified frequencies for these two groups of forecasts would be 0% and 100%. For the mathematical definitions of reliability and resolution, see Murphy (1973) or Palmer (2000).

The reliability and resolution characteristics of a probabilistic forecast system can be illustrated with a reliability diagram (Wilks 1995), in which the observed average frequency $o(p)$ is plotted against the forecast probability p. Such diagrams based on the CMIP2 ensemble are shown in Fig. 4. In calculating $o(p)$, all cases in which the predicted probability was p were selected, using data for all four seasons and for 13 partly overlapping 20-yr periods (1–20, 6–25, . . . , 61–80) in the 80-yr greenhouse runs. This was first made individually for each 17 choices of the verifying model and the 16-model forecast ensemble, and for Figs. 4a and 4b further averaging over the 17 alternatives was done. The diagrams thus reflect the interexperiment agreement on how the climate changes vary with the strength of the CO_2 forcing and the season, as well as with the geographical location.

For both DT $>$ 1°C and DP $<$ −10%, the 17-case mean $o(p)$ exhibits a marked general increase with the forecast probability, suggesting that the forecasts have useful resolution (Figs. 4a,b). In both cases but much more markedly for DP $<$ −10%, however, the slope of the curve is below 45°, meaning that the 16-model ensembles tend to give too certain predictions of the behavior of the remaining verifying simulation. This substantial systematic unreliability indicates that, for probabilistic predictions of the DP $<$ −10% event, larger ensembles would be preferable. Note, on the other hand, that the detailed behavior of the DP $<$ −10% reliability curve at the highest forecast probabilities deserves relatively little attention. As indicated by the insets, cases with a high forecast probability for this event are very rare, which makes the results sensitive to sampling variability. The striking fall in the verified probability down to zero when the forecast probability of DP $<$ −10% is 100% is entirely attributed to this factor.

Figure 4a may give the impression that, for DT $>$ 1°C, probabilistic climate change forecasts are highly reliable (i.e., the verified probability is very close to the forecast probability for a wide range of the latter). This impression is, however, a by-product of averaging $o(p)$ over the 17 individual verifications. Reliability curves plotted for the 17 cases separately (Fig. 4c) show that, for any forecast probability p, the verified probability $o(p)$ varies quite strongly with the verifying model. Excluding p = 0% and p = 100%, the individual $o(p)$ curves are scattered on both sides of p, so that the 17-case mean verified probability is necessarily closer to p than the individual verifications are on the average [by contrast, the averaging causes no artifacts in the inferred resolution characteristics, that is, differences in $o(p)$ between high and low p]. The verified probabilities are generally smallest and largest for verification against the NRL and the oldest NCAR (NCAR–WM) models with the smallest and largest global mean warming. In the NRL case, in addition, the verified probability is actually highest for relatively low forecast probabilities. This is probably due to the fact that the largest warming in the NRL experiment occurs near the margin of the

TABLE 3. Frequencies associated with the precautionary action and the detrimental binary event. The economic expense in each case is given in parentheses.

		Occurs	
		No	Yes
Action	No	α (0)	β (L)
	Yes	γ (C)	δ (C)

Antarctic continent, where many of the other 16 experiments indicate a minimum in warming.

5. A decision-model analysis

Assume that the CMIP2 experiments in fact give the probability distribution of future climate changes: one of the models is right but we do not know which one. Are these probability distributions practically useful? To approach the question quantitatively, we consider a simple decision model (Thompson 1952; Murphy 1977; Katz and Murphy 1997) whose inputs are probabilistic forecast information and whose output is potential economic value. A specific possible application of this model (reservoir construction) is discussed below, and some of its limitations in the climate change context are noted in the end of this section.

The decisionmaker would like to know if a detrimental climate event E will occur. If E occurs and no precautionary action is taken, this will cause a loss L. The loss can be avoided with precautionary action, but this itself involves an expense C. Thus, when averaged over a large number of cases, the expected economic expense normalized by L will be

$$M = \frac{C}{L}(\gamma + \delta) + \beta, \tag{3}$$

where β is the relative loss frequency (no action is taken but the event occurs), and γ and δ are the corresponding frequencies of ''false alarms'' and correct precautionary action (see Table 3).

For a perfect deterministic forecast, $\beta = \gamma = 0$ and $\delta = \overline{o}$, where $\overline{o}$ is the relative frequency of cases in which E occurs, and the normalized expense is thus

$$M_{\text{per}} = \frac{C}{L}\overline{o}. \tag{4}$$

A decisionmaker expecting no change in climate would assume that the probability of the event in the future will equal its relative frequency in the present climate ($\overline{o}_{\text{cli}}$). With this assumption, the precautionary action would appear useful if and only if $\overline{o}_{\text{cli}} > C/L$. The expected cost of using the present climate information is thus

$$M_{\text{cli}} = \begin{cases} C/L, & \overline{o}_{\text{cli}} > C/L \\ \overline{o}, & \overline{o}_{\text{cli}} < C/L, \end{cases} \tag{5}$$

where, if climate does change, $\overline{o}$ may be substantially different from $\overline{o}_{\text{cli}}$.

Alternatively, a decision to take or not take precautionary action could be based on the probabilistic climate change forecast. The action is taken if the forecast probability exceeds a preselected threshold p_t. In this case, the expense will depend on how p_t is chosen [cf. (3)]. If p_t is set high, the action is taken seldom, which keeps the precautionary cost low [γ and δ in (3) are small] but increases the risk of damage (β may be large). Conversely, small p_t keeps the risk of damage small but the precautionary cost is high. The optimal p_t is the one that minimizes the expected expense. This optimal value depends on C/L: the smaller the precautionary cost, the less eventual false alarms matter, and the smaller the optimal p_t. The relative value of the probabilistic forecast with threshold p_t is defined as

$$V = \frac{M_{\text{cli}} - M(p_t)}{M_{\text{per}} - M_{\text{cli}}} \tag{6}$$

and the optimal value as

$$V_{\text{opt}} = \max_{p_t} V(p_t). \tag{7}$$

It is easy to show that the optimal p_t for a perfectly reliable probabilistic forecast system equals C/L. When the probabilistic system is not perfectly reliable, however, the optimal p_t needs to be chosen empirically, as indicated by (7). In other words, the user needs to calculate the expected expense for different choices of p_t in advance, based on past verification statistics of the system or (in the climate change application where such statistics are not available) on the cross verification. If the selection of p_t is based on a too small sample (e.g., a few individual grid points), this procedure involves a risk of getting too high estimates of V_{opt}. This is because, if the sample is small, the calculated expense might be low at some p_t just by chance. For the large-area value estimates presented in this section, however, this caveat appears unimportant. The value estimates obtained by simply using p_t = C/L were only slightly lower than, or in some cases almost identical with, the estimates based on the empirically optimized p_t.

The potential value of the CMIP2 climate change ensemble is estimated using the cross-verification approach applied above. The expenses associated with the probabilistic climate change forecast (using different choices of p_t) and the climatological forecast (based on the average control run probabilities) are both first calculated for all 17 choices of the verifying model and then averaged over the 17 cases. Last, the optimal p_t is selected. In Fig. 5, the two events studied in Fig. 2 are considered: DT(JJA) > 1°C, and DP(JJA) $< -10\%$. The potential economic value may be seen just as another diagnostic quantifying the agreement between the different model experiments, but the results might also have practical relevance.

For example, the analysis for DP(JJA) $< -10\%$ can

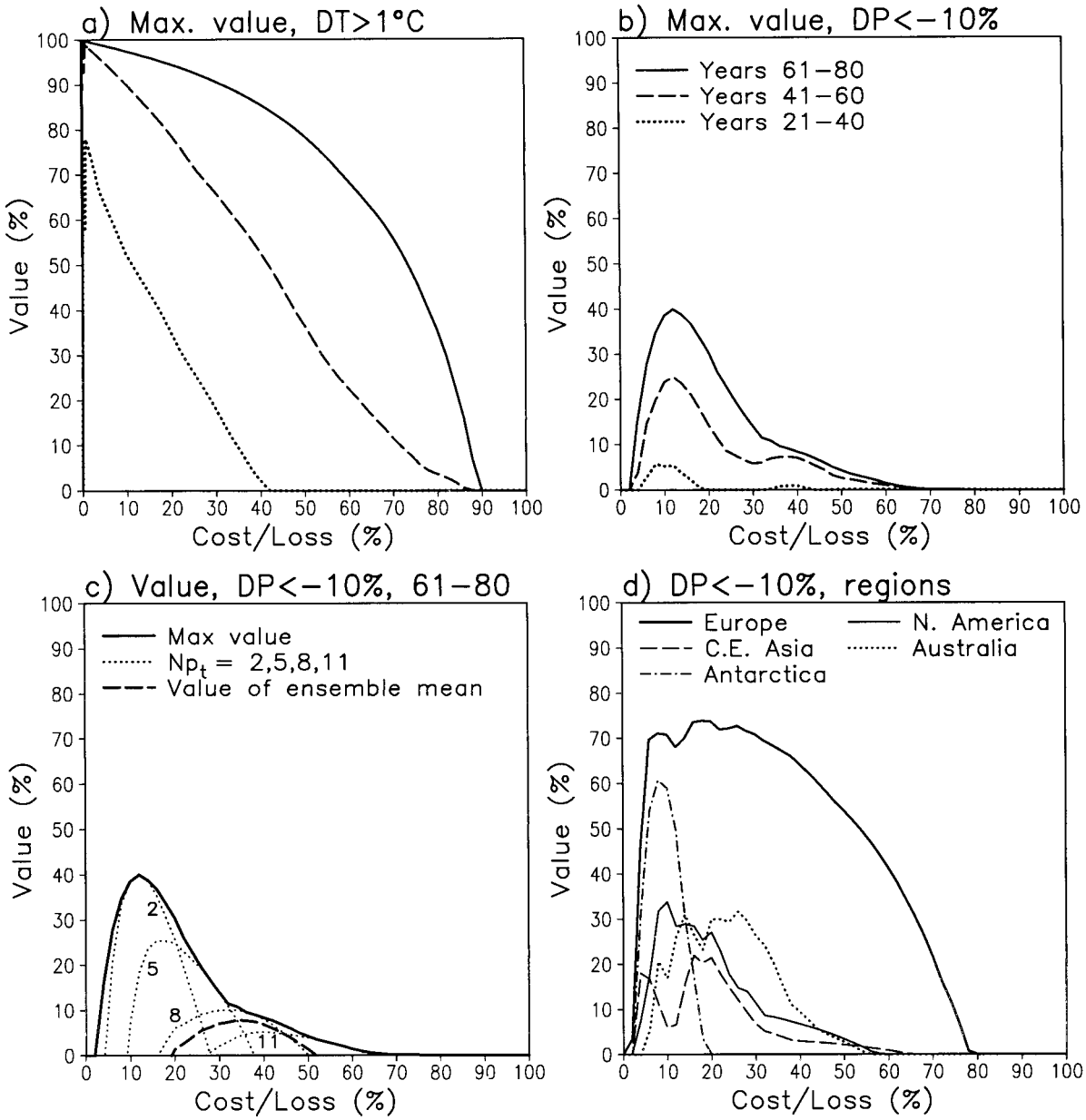

FIG. 5. The relative global economic value of the optimal probabilistic forecasts for (a) DT $>$ 1°C and (b) DP $<$ −10% in the years 61–80 (solid lines), 41–60 (dashed), and 21–40 (dotted). (c) The economic value for DP $<$ −10% in the years 61–80 for the optimal probabilistic forecast (solid), probabilistic forecasts with some individual Np_t (dotted), and for using the ensemble mean as a deterministic forecast (dashed). Here Np_t denotes the minimum number of models with DP $<$ −10% (out of the 16 in the forecast ensemble) required for taking the action. (d) The optimal economic value for probabilistic forecasts of DP $<$ −10% in the years 61–80 in some geographic areas.

be motivated by the following scenario. Imagine that water authorities in a number of different regions have to decide whether to build new reservoirs. The cost C of construction is considerable, but if water rationing had to be imposed, this would involve an expense L to industry in particular, and to society in general. Clearly, L will depend on how frequently water rationing is needed (e.g., once in 20 yr or every summer), and it will therefore be a complex nonlinear function of climatic conditions. However, for the sake of argument, let us suppose that L can be quantified. Let us also suppose, for the sake of argument, that the frequency of drought conditions is related in a simple step function manner to the 20-yr mean precipitation anomaly DP: it increases from negligible values to substantial values when DP varies around −10%. This is clearly a gross simplification but one that suffices for present illustrative purposes.

Hence, in this idealized scenario, the water authorities would base their decision on whether to invest in new

reservoirs on the probability that 20-yr mean JJA rainfall anomalies are below −10% in future decades. If it were assumed that climate is stationary, that is, that the probability of such long-term rainfall deficit in the future is the same as it is now ($\overline{\sigma}_{\text{cli}}$), then it would appear beneficial to build a new reservoir if C/L $< \overline{\sigma}_{\text{cli}}$. Conversely, if C/L $> \overline{\sigma}_{\text{cli}}$, the cost of building the reservoir would be judged excessive in relation to the likelihood of water shortage. On the other hand, the water authorities could use information from the CMIP2 ensemble. For some of them, this would lead to a different decision (either building a reservoir when the first strategy did not suggest this is useful, or vice versa). Will the use of the CMIP2 ensemble lead to smaller net expenses (when averaged over all water authorities)? If yes, then the ensemble of climate change integrations can be said to have economic value.

For a third alternative, the water authorities might use the climate change information in a deterministic manner, trusting that the ensemble mean (''consensus forecast'') precipitation change is right. Thus, reservoirs are built if and only if the ensemble mean precipitation decreases by at least 10%. Is this strategy as good or worse than using the climate change information in a probabilistic manner?

Figures 5a,b give the optimal global economic value (7) for the two events in three 20-yr periods of the greenhouse runs: years 21–40, 41–60, and 61–80. Like the Brier skill scores in Table 2, the value increases with time, reflecting the gradual increase in CO_2 forcing. Thus, the longer the planning horizon, the larger the potential advantage of using climate change information in the decision process. During the first 20 yr of the greenhouse runs, the value is still negligible for both events (not shown).

As expected, the value for DT > 1°C is much higher than that for DP $< -10\%$. Positive value for DT > 1°C extends, during the latter two periods, for all but the very highest C/L. However, the value is particularly high, almost 100% in the last two periods and over 70% in the years 21–40, for low C/L. This is because, in a large part of the world, $\overline{\sigma}_{\text{cli}}$ for this event is zero. A decisionmaker relying on this information alone would never take the precautionary action, even when the cost of precaution is negligible compared with the eventual loss. This strategy works poorly in a changed climate in which the event is quite common. The probabilistic climate change forecast shows that the event in fact may happen, and precaution is therefore judged useful. This leads to greatly reduced expenses at low C/L, in spite of occasional false alarms.

For DP $< -10\%$, the maximum value reaches 0.40 in the years 61–80, 0.25 in the years 41–60, and only 0.06 in the years 21–40. The range of C/L for which positive value exists is narrower than that for DT $>$ 1°C, but in the last period it still extends from about 0.02 to 0.7. Last, the maximum value is obtained at higher C/L (in the years 61–80, at C/L $= 0.12$) than for DT > 1°C. $\overline{\sigma}_{\text{cli}}$ is, in most of the world, higher for DP $< -10\%$ than for DT > 1°C. Precautionary action against DP $< -10\%$ at very low C/L therefore generally appears useful even with no climate change information.

TABLE 4. The matrix elements α, β, γ, and δ, and the extra costs over a perfect deterministic forecast (losses, false alarms, total) for the event DP(JJA, years 61–80) $< -10\%$, C/L $= 0.12$. The first column is for the climatological forecast, the second is for the optimal probabilistic forecast, and the third is for the ensemble mean precipitation change. The relative economic value of the forecasts is given in the last row.

	Climatology	Probabilistic forecast	Ensemble mean
α (correct rejection)	0.631	0.365	0.761
β (losses)	0.111	0.022	0.164
γ (false alarms)	0.167	0.433	0.037
δ (correct precaution)	0.091	0.181	0.039
Cost (losses)	0.098	0.019	0.144
Cost (false alarms)	0.020	0.052	0.004
Cost (total)	0.118	0.071	0.148
Relative value	0	40%	−25%

Figure 5c shows, for DP $< -10\%$ in the years 61–80, the value curves for four individual choices of p_t (Np_t in the figure indicates how many models out of 16 are required to predict the event to occur). With low p_t, the value is relatively high for low C/L but it also drops relatively rapidly to zero when C/L increases. With larger p_t the top of the value curve moves toward higher C/L. Thus, as expected, decision makers with a low cost–loss ratio should use a lower threshold probability than those whose cost–loss ratio is high.

Relying on the ensemble consensus climate change (i.e., the action is taken if the ensemble mean precipitation decreases by at least 10%) is, for DP $< -10\%$, a rather poor strategy (see the dashed line in Fig. 5c). The ensemble mean does have some modest value for C/L $\approx$ 0.2–0.5 (as the mean is generally near the median, this is similar to the value of using $Np_t = 8$–9), but it has no value for C/L ≈ 0.12 where the value of the probabilistic forecast is highest. For decision makers with a low C/L ratio, it is important to identify potentially disastrous events that may happen even if their probability is well below 50%. A probabilistic interpretation of climate change results provides a potentially useful method for such risk assessment, unlike the simpler approach of treating the consensus of the model results as truth.

As stated above, the high potential value for the DT > 1°C event is associated with greatly reduced losses associated with failure to take precautionary action. The same is qualitatively true for DP $< -10\%$, basically because even this event is somewhat more common in the greenhouse than in the control runs (Table 1). As an example, the components of the action–occurrence matrix for C/L $= 0.12$ are given in Table 4. As averaged over the 17 verifying models, JJA precipitation decreases by over 10% in 20.3% of the globe ($\beta + \delta$). In most

TABLE 5. The highest relative potential economic value (in %) reached at any C/L between 0.02 and 0.98 for probabilistic forecasts of some events, and the corresponding highest value for using the ensemble mean climate change (in parentheses). All results are for the global domain.

	DT > 0°C		DT > 1°C		DP < −10%		DP > 0		DP > 10%	
Years	DJF	JJA	DJF	JJA	DJF	JJA	DJF	JJA	DJF	JJA
1–20	47	54	8	4	1	0	3	4	1	0
(ens. mean)	(45)	(52)	(0)	(0)	(0)	(0)	(2)	(1)	(0)	(0)
21–40	93	94	56	74	2	6	17	15	8	1
(ens. mean)	(93)	(94)	(30)	(7)	(0)	(1)	(16)	(14)	(0)	(1)
41–60	98	98	98	98	19	25	28	27	40	27
(ens. mean)	(98)	(98)	(62)	(67)	(1)	(6)	(27)	(27)	(10)	(1)
61–80	98	99	99	99	29	40	33	34	59	52
(ens. mean)	(98)	(99)	(94)	(96)	(7)	(9)	(33)	(34)	(22)	(10)

of this area ($\beta = 11.1\%$), the climatological probability is below 0.12, precaution is judged useless, and loss occurs. The optimal probabilistic forecast (the best choice at this C/L is to take the action if at least 2 out of the 16 models predict the event to occur) reduces the loss frequency to only 2.2%. Therefore, despite a much larger false alarm frequency γ for the probabilistic (43.3%) than for the climatological (16.7%) forecast, the total expected expenses are 40% smaller for the former. Using (3), (4), and the definition $\beta + \delta = \overline{o}$, the expense due to an imperfect forecasting system can be written as

$$M - M_{\text{per}} = M_L + M_{\text{FA}} = \beta\left(1 - \frac{C}{L}\right) + \gamma\frac{C}{L}, \quad (8)$$

where M_L denotes the loss due to damage (no precautionary action when the event occurs) and M_{FA} the expense due to false alarms (precaution when this is not needed). According to this division, over 80% of the unnecessary expenses associated with the climatological forecast are due to losses, whereas over 70% of the expenses associated with the optimal probabilistic forecast are due to false alarms.

Last, a user relying solely on the ensemble mean precipitation change would judge the precautionary action useless in over 92% of the world (right column of Table 4). This would lead to a loss rate β of 16.4%, well in excess of the 11.1% loss rate associated with the climatological forecast. Despite a substantial decrease in the frequency of false alarms, this makes the total unnecessary expense at this C/L 25% larger for the ensemble mean than for the climatological forecast (note that the negative value is not visible in Fig. 5c where the scale starts from zero).

The value estimates in Figs. 5a–c use all grid points in the world. This is an efficient and objective way to characterize the overall performance of the ensemble, but, in practice, decisionmakers operate in subglobal domains. Figure 5d therefore shows V_{opt} for DP < −10% in the years 61–80 for some smaller albeit still relatively large areas. The five regions considered are Europe (land grid boxes at 35°–70°N, 10°W–45°E), North America (25°–70°N, 150°–60°W), central-east Asia (25°–60°N, 80°–145°E), Australia (12°–40°S, 115°–155°E), and Antarctica (60°–90°S). Clearly, the value estimates are quite area dependent. Much of this variability can be understood by comparing the 17-model climate change probabilities in Fig. 2d with the corresponding control run probabilities in Fig. 2f. The larger the contrast between the two, the more value is available.

Of the regions considered, the least value is generally found in central East Asia. The simulated precipitation changes in this area vary on both sides of zero, and the inferred probability of DP < −10% is relatively close to (although mostly somewhat above) that in the control runs. Somewhat more potential value is diagnosed for North America and Australia. However, the highest value for any of the five regions is obtained in Europe, where V_{opt} peaks at 74% at C/L = 0.2 and remains substantially positive for C/L up to over 0.7. This is due to the fact that a majority of the models simulate an over 10% decrease in precipitation in southern Europe (for the European area north of 50°N, V_{opt} only reaches 50% at C/L ≈ 0.06 and drops below 10% at C/L > 0.2). Unlike in the global domain, even the ensemble mean precipitation change has very substantial potential value in Europe, at best 67% at C/L = 0.3.

The value curve for Antarctica differs substantially in shape from those for the other regions, reaching 60% at C/L = 0.08 but dropping to zero for C/L > 0.2. In this area, DP < −10% is less common in the greenhouse runs than in the control runs (see Figs. 2d and 2f). Therefore, on Antarctica, hypothetical water authorities with a low C/L ratio would benefit from the probabilistic forecast by not building new reservoirs where the present climate statistics suggest this to be useful.

The impacts of climate change are not restricted to the two events studied in Fig. 5 nor only to the northern summer. Table 5 therefore lists, for a slightly larger sample of events in both solstitial seasons and in different 20-yr periods, the global maximum value for the probabilistic climate change forecast (highest V_{opt} at any C/L) and the corresponding maximum value for using

the ensemble mean climate change. Several features are worth noting.

1) Unlike the other events, DT $>$ 0 shows substantial maximum value already in the first 20-yr period (in accord with the Brier skill scores in Table 2).
2) The maximum value for a 10% increase in precipitation is higher than that for a 10% decrease, in particular in DJF. This is at least partly because the frequency of +10% anomalies increases more in the models than the frequency of −10% anomalies, as expected from increasing global mean precipitation. Again, the global value estimates hide substantial geographical variations. For all of Europe, North America, and central East Asia, the maximum value of DP(DJF) $>$ 10% approaches 90% in the years 61–80 and it already exceeds 40% in the period 21–40 (not shown).
3) In terms of the maximum value, there is little difference between the probabilistic forecast and the ensemble mean forecast for DT $>$ 0 and DP $>$ 0. For these events with control run climatological probabilities close to 50%, the maximum value is obtained at C/L around 0.5. At such midrange C/L, the optimal probabilistic forecast and the ensemble mean climate change have generally very similar value (at low and high C/L, however, the probabilistic forecast beats the ensemble mean even in these cases). During the last 20-yr period, the difference in maximum value between the probabilistic and ensemble mean forecasts is also small for DT $>$ 1°C, simply because the warming in most areas exceeds 1°C in a majority of the models.
4) For the events DP $<$ −10% and DP $>$ 10%, and for DT $>$ 1°C before the last 20-yr period, the maximum value is substantially higher for the probabilistic forecast than for using the ensemble mean. Thus, the potential advantage of using the model results in a probabilistic way is largest for those events that are relatively rare in the control simulations and for which some but no unanimous interexperiment agreement on the future development exists.

Reliance on the ensemble mean is of course not the only deterministic approach to using the model results. An even simpler method is to select just one model simulation and take it as truth. The average (over different choices of the forecast model and the verifying model) value of this strategy was also evaluated for the DP(JJA, years 61–80) $<$ −10% event and was found to be, for most C/L ratios, even worse than that of using the ensemble mean (not shown). At low C/L, however, the reverse was true. In a narrow C/L range between 0.15 and 0.2, individual ensemble members were found to have slightly positive, but the ensemble mean negative, globally averaged value. This can be understood from the fact that ensemble averaging tends to damp the amplitude of both positive and negative precipitation anomalies, making large anomalies of either sign look more rare than they are in the individual simulations. The resulting increase in loss frequency makes the ensemble mean a particularly poor choice for users with a low C/L ratio.

Obviously, the actual situation faced by decision-makers in front of climate change is more complex than indicated by this simple cost–loss model. The model, originally developed for the context of short-term weather forecasts, assumes a detrimental event that either does not happen or happens just once during the forecast period. In the long-term climate change context, the relevant eventual losses are rather related to how frequently the event occurs during the whole period when the precautionary measure is effective. In the case of reservoir construction, this period is certainly longer than just 20 yr. In addition, in quantifying the cost–loss ratio in the case that the costs occur much earlier in time than the eventual losses, discounting of future expenses may be in order. Last, it is expected that advances in research will gradually lead to more accurate projections of future climate. Thus, in the case that the model results point toward a need of precaution after several decades but not in the immediate future, the best strategy might be to wait with the decision until more information is available.

6. Summary with discussion

Climate change will have crucial implications on socioeconomic planning in coming decades. Projections of future climate are, however, affected by several sources of uncertainty, ranging from external forcing conditions (e.g., future CO_2 and aerosol concentrations) to model deficiencies and internally generated climate variability. This raises the fundamental question of how climate change information should be presented to users: as a deterministic consensus view, such as an ensemble mean over all available model results, or in probabilistic terms, making full use of the spectrum of possibilities? The purpose of this paper has been to show that a probabilistic interpretation of climate change simulations is feasible and has potential advantages over the more conventional deterministic interpretation.

The probabilistic interpretation was applied to an ensemble of 17 CMIP2 experiments, which share the same 1% yr^{-1} increase in CO_2 but differ in terms of model characteristics and internal variability. Evaluation tools commonly used for probabilistic forecasts from daily to seasonal time ranges, that is, Brier skill scores, reliability diagrams, and estimates of potential economic value, were applied to this multimodel climate-change ensemble. Given that no observations of future climate exist, a cross-verification approach was used, in which each of the 17 simulations was taken, in turn, as the verifying truth for the ensemble formed by the other 16 simulations. Such a cross verification will of course only give unbiased results if the different experiments span

the actual probability distribution of future climate changes. This condition is unlikely to be fulfilled, both because the models might share common errors in their response to increased CO_2 and because the CMIP2 experiments share the same idealized forcing scenario. The quantitative results presented in this paper should therefore be taken with care, but we find it justified to believe that most of the qualitative conclusions are reasonably robust.

The main findings from the cross-verification are summarized below.

1) In terms of the Brier skill score and potential economic value, probabilistic climate change forecasts improve with increasing time horizon, as the impact of increasing CO_2 becomes better discernible from internal variability. In practice, this particular result may be to some extent compromised by forcing-related uncertainty (due to, e.g., unknown future CO_2 emissions), which grows increasingly important with time (e.g., Houghton et al. 1996) but is not presented by the CMIP2 ensemble.
2) Probabilistic forecasts for many temperature-related events (e.g., will the 20-yr mean summer temperature around the doubling of CO_2 be at least 1°C above its present level) have, with the caveats associated with the cross verification, very high quality and large potential economic value.
3) Probability forecasts for precipitation-related events (e.g., will summer precipitation decrease by 10% or more) are likely to be less skillful than similar forecasts for temperature-related events. Despite this, probabilistic forecasts for precipitation-related events may also have substantial economic value for decision makers with planning horizons of several decades and longer.
4) In terms of the potential economic value, a probabilistic interpretation of climate change simulations has a distinct advantage over the deterministic interpretation of treating the ensemble mean of model results as the truth. This is most clearly the case when the scatter between different model results is substantial, as is generally the case with precipitation changes. An ensemble mean only shows the events that on the average happen in the models. In climate-related risk assessment, however, information is also needed on those high-cost events that may happen even if their probability is judged to be well below 50%.

The last point is of particular importance. Because of the inherent uncertainties that exist in climate prediction, the notion of providing users only with deterministic forecasts is a misguided strategy. More informed decisions can be made given a reliable probability forecast, as compared with a deterministic forecast of uncertain accuracy. It is therefore hoped that the framework for probabilistic climate change predictions developed in this paper will become an established part of the analysis of climate change ensemble integrations. This framework can and should be applied both to ensembles including different forcing scenarios and to more complicated but practically important climate events such as the occurrence of various types of extremes.

Acknowledgments. All CMIP2 modeling groups are acknowledged for conducting and making available the simulations requested by the CMIP Panel. CMIP is supported and the model data are distributed by the Program for Climate Model Diagnosis and Intercomparison (PCMDI) at the Lawrence Livermore National Laboratory (LLNL). The Rossby Centre is part of the Swedish SWECLIM programme financed by MISTRA and by SMHI. This paper benefited from helpful comments by two anonymous reviewers.

REFERENCES

Barthelet, P., L. Terray, and S. Valcke, 1998: Transient CO_2 experiment using the ARPEGE/OPAICE non flux corrected coupled model. *Geophys. Res. Lett.,* **25,** 2277–2280.

Boville, B. A., and P. R. Gent, 1998: The NCAR Climate System Model, version one. *J. Climate,* **11,** 1115–1130.

Braconnot, P., O. Marti, and S. Joussaume, 1997: Adjustment and feedbacks in a global coupled ocean–atmosphere model. *Climate Dyn.,* **13,** 507–519.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probabilities. *Mon. Wea. Rev.,* **78,** 1–3.

Emori, S., T. Nozawa, A. Abe-Ouchi, A. Numaguti, M. Kimoto, and T. Nakajima, 1999: Coupled ocean–atmosphere model experiments of future climate change with an explicit representation of sulfate aerosol scattering. *J. Meteor. Soc. Japan,* **77,** 1299–1307.

Flato, G. M., G. J. Boer, W. G. Lee, N. A. McFarlane, D. Ramsden, M. C. Reader, and A. J. Weaver, 2000: The Canadian Centre for Climate Modelling and Analysis global coupled model and its climate. *Climate Dyn.,* **16,** 451–467.

Gordon, C., C. Cooper, C. A. Senior, H. Banks, J. M. Gregory, T. C. Johns, J. F. B. Mitchell, and R. A. Wood, 2000: The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Climate Dyn.,* **16,** 147–168.

Hirst, A., S. P. O'Farrell, and H. B. Gordon, 2000: Comparison of a coupled ocean–atmosphere model with and without oceanic eddy-induced advection. Part I: Ocean spinup and control integrations. *J. Climate,* **13,** 139–163.

Houghton, J. T., L. G. Meira Filho, B. A. Callander, N. Harris, A. Kattenberg, and K. Maskell, Eds., 1996: *Climate Change 1995: The Science of Climate Change*. Cambridge University Press, 572 pp.

Johns, T. C., R. E. Carnell, J. F. Crossley, J. M. Gregory, J. F. B. Mitchell, C. A. Senior, S. F. B. Tett, and R. A. Wood, 1997: The second Hadley Centre Coupled ocean–atmosphere GCM: Model description, spinup and validation. *Climate Dyn.,* **13,** 103–134.

Katz, R. W., and A. H. Murphy, 1997: Forecast value: Prototype decision-making models. *Economic Value of Weather and Climate Forecasts,* R. W. Katz and A. H. Murphy, Eds., Cambridge University Press, 183–217.

Li, T. F., and T. F. Hogan, 1999: The role of the annual mean climate on seasonal and interannual variability of the tropical Pacific in a coupled GCM. *J. Climate,* **12,** 780–792.

Liljas, E., and A. H. Murphy, 1994: Anders Ångström and his early papers on probability forecasting and the use/value of weather forecasts. *Bull. Amer. Meteor. Soc.,* **75,** 1227–1236.

Manabe, S., R. J. Stouffer, M. J. Spelman, and K. Bryan, 1991: Transient responses of a coupled ocean–atmosphere model to gradual changes of atmospheric CO_2. Part I: Annual mean response. *J. Climate*, **4,** 785–818.

Meehl, G. A., G. J. Boer, C. Covey, M. Latif, and R. J. Stouffer, 1997: Intercomparison makes for a better climate model. *Eos, Trans. Amer. Geophys. Union*, **78,** 445–446, 451.

——, ——, ——, ——, and ——, 2000: The Coupled Model Intercomparison Project (CMIP). *Bull. Amer. Meteor. Soc.*, **81,** 313–318.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12,** 595–600.

——, 1977: The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Mon. Wea. Rev.*, **105,** 803–816.

New, M., and M. Hulme, 2000: Representing uncertainty in climate change scenarios: A Monte-Carlo approach. *Integr. Assess.*, **1,** 203–213.

Palmer, T. N., 2000: Predicting uncertainty in forecasts of weather and climate. *Prog. Rep. Phys.*, **63,** 71–117.

——, 2001: A nonlinear dynamical perspective on model error: A proposal for nonlocal stochastic-dynamic parametrisation in weather and climate prediction models. *Quart. J. Roy. Meteor. Soc.*, **127,** 273–304.

——, C. Brankovic, and D. S. Richardson, 2000: A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Quart. J. Roy. Meteor. Soc.*, **126,** 2013–2033.

Power, S. B., R. A. Colman, B. J. McAvaney, R. R. Dahni, A. M. Moore, and N. R. Smith, 1993: The BMRC coupled atmosphere/ocean/sea-ice model. BMRC Research Rep. 37, Bureau of Meteorology Research Centre, Melbourne, Australia, 58 pp.

Räisänen, J., 2001: CO_2-induced climate change in CMIP2 experiments: Quantification of agreement and role of internal variability. *J. Climate*, **14,** 2088–2104.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126,** 649–667.

Russell, G. L., and D. Rind, 1999: Response to CO_2 transient increase in the GISS coupled model: Regional coolings in a warmer climate. *J. Climate*, **12,** 531–539.

Thompson, J. C., 1952: On the operational deficiencies in categorical weather forecasts. *Bull. Amer. Meteor. Soc.*, **33,** 223–226.

Tokioka, T., A. Noda, A. Kitoh, Y. Nikaidou, S. Nakagawa, T. Motoi, S. Yukimoto, and K. Takata, 1995: A transient CO_2 experiment with the MRI CGCM. Quick Report. *J. Meteor. Soc. Japan*, **73,** 817–826.

Voss, R., R. Sausen, and U. Cubasch, 1998: Peridiocally synchronously coupled integrations with the atmosphere–ocean general circulation model ECHAM3/LSG. *Climate Dyn.*, **14,** 249–266.

Washington, W. M., and G. A. Meehl, 1996: High-latitude climate change in a global coupled ocean–atmosphere–sea ice model with increased atmospheric CO_2. *J. Geophys. Res.*, **101,** 12 795–12 801.

——, and Coauthors, 2000: Parallel climate model (PCM) control and transient simulations. *Climate Dyn.*, **16,** 755–774.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.

Zhang, X., G. Shi, H. Liu, and Y. Yu, Eds., 2000: *IAP Global Ocean–Atmosphere–Land System Model.* Science Press, 249 pp.