Tellus (2007), 59A, 309–320
Printed in Singapore. All rights reserved

TELLUS

Probabilistic forecasts of near-term climate change: sensitivity to adjustment of simulated variability and choice of baseline period

By LEENA RUOKOLAINEN* and JOUNI RÄISÄNEN, *Department of Physical Sciences, Division of Atmospheric Sciences, P.O. Box 64, FIN-00014 University of Helsinki, Finland*

(Manuscript received 26 September 2006; in final form 26 January 2007)

ABSTRACT

The authors of this study recently proposed a resampling method for deriving probabilistic forecasts of near-term climate change and presented some results focusing on temperature and precipitation changes in southern Finland from 1971–2000 to 2011–2020. Here, the sensitivity of the resulting forecasts to two details of the methodology is studied. First, to account for differences between simulated and observed climate variability, a variance correction technique is devised. Second, the sensitivity of the forecasts to the choice of the baseline period is studied. In southern Finland, the variance correction technique generally widens the derived probability distributions of precipitation change, mirroring an underestimate of the observed precipitation variability in climate models. However, the impact on the derived probability distributions of temperature change is small. The choice of the baseline period is generally more important, but again the forecasts of temperature change are less sensitive to different options than those of precipitation change. Cross-verification suggests that the variance correction leads to a slight improvement in the potential quality of the probabilistic forecasts, especially for precipitation change. The optimal baseline length appears to be at least 30 yr, and the baseline should be as late as possible (e.g. 1971–2000 is preferable over 1961–1990).

1. Introduction

Projections of future climate change are affected by several sources of uncertainty (e.g. Cubasch et al., 2001). A natural way to deal with the uncertainties is to try to express the projections in probabilistic terms, analogously to what is commonly done in operational weather forecasting (e.g. Molteni et al., 1996). Consequently, a number of methods for deriving probabilistic projections of climate change have been proposed (e.g. New and Hulme, 2000; Räisänen and Palmer, 2001; Giorgi and Mearns, 2003; Räisänen and Alexandersson, 2003; Tebaldi et al., 2005).

The requirements for a good probabilistic climate change forecasting method depend, in part, on the period considered. Many of the studies in the field have focused on climate change in the relatively distant future, for example, the late–21st-century. On that timescale, uncertainties due to differences between climate models and emissions scenarios tend to dominate over natural climate variability. This makes issues such as weighting between different models (e.g. Giorgi and Mearns, 2003) and emissions scenarios (New and Hulme, 2000) important. For projections of climate change in the early–21st-century, however, differences between emission scenarios are unimportant, and uncertainty associated with climate models is also smaller than in longer-term forecasts (e.g. Räisänen, 2001). Conversely, the relative importance of forced and unforced natural climate variability becomes larger.

In a recent study, Räisänen and Ruokolainen (2006; hereafter RR06) derived probabilistic forecasts of near-term climate change, focusing on temperature and precipitation changes from the period 1971–2000 to the decade 2011–2020. To achieve a sufficient sample size, they used a resampling ensemble technique based on the assumption that the probability distribution of local climate changes is, at least to a first approximation, determined by the change in the global mean temperature averaged over a large number of climate models. By using cross-verification, they demonstrated that the enlarged sample of natural variability allowed by the resampling outweighs the errors that might arise due to the violation of this basic assumption. Nevertheless, there are some methodological issues in the technique of RR06 that deserve a closer investigation. In this paper we study the sensitivity of the resulting probabilistic forecasts to two choices in the methodology: a correction to take into account differences

*Corresponding author.
e-mail: leena.ruokolainen@helsinki.fi
DOI: 10.1111/j.1600-0870.2007.00233.x

between simulated and observed natural variability, and the selection of the baseline period from which the climate changes are calculated.

The need for a realistic estimate of natural climate variability in short-term climate change forecasts is obvious. For example, if the model simulations that are used as the basis of the probabilistic forecasts underestimate natural variability, this tends to make the derived probability distributions too narrow and thus give a misleading impression of certainty. Conversely, if models overestimate variability, the reverse happens. It is therefore important to compare simulated and observed variability with each other and, if there is evidence of differences, to try to adjust the derived probability distributions accordingly. In practice, this is not a straightforward task because natural variability on decadal and longer scales is difficult to estimate from observations. In this study, we use a technique based on a comparison of simulated and observed variability on interannual time scales. Although the basic assumption of this technique—namely, that variability on longer timescales is directly proportional to the interannual variability—may not be exactly valid, a cross-verification test indicates that the adjustments based on this technique are likely to be better than no adjustments at all.

The potential importance of the choice of the baseline period is best illustrated with an example. The winter climate in northern Europe in the late–20th-century exhibited strong interdecadal variability, with a pronounced warming from the 1960s to the 1990s that accompanied an increase in westerly flow from the Atlantic Ocean (e.g. Räisänen and Alexandersson, 2003; Scaife et al., 2005). For example, in a grid box in southern Finland (60°N, 25°E), the January mean temperatures for the decades 1951–1960, 1961–1970, 1971–1980, 1981–1990 and 1991–2000 were −6.3, −8.6, −6.2, −7.0 and −3.7°C, respectively, as inferred from the Climate Research Unit (CRU) TS 2.1 data set (Mitchell and Jones, 2005). The variability in multidecadal means was smaller but far from negligible (Table 1). For example, of the two overlapping 30-yr periods 1961–1990 and 1971–2000, the latter was 1.6°C warmer than the former. In addition to these two 30-yr periods, Table 1 also gives the mean temperatures for the 20-, 40- and 50-yr periods ending in 2000, all of which could be conceivably used as baselines when deriving climate scenarios. In particular, although 30 yr is the most commonly used baseline length in climatology, the Arctic Climate Impact Assessment (ACIA, 2005) and the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC, 2007) both chose a 20-yr baseline.

Table 1. An illustration of the sensitivity of climate scenarios to the choice of the baseline period. T_{obs} = observed January mean temperature in southern Finland (60°N, 25°E) for five alternative baseline periods; ΔT_{mean} = Change in January mean temperature between the various baselines and the decade 2011–2020, as averaged over simulations by 21 climate models; T_{scen} = 'best-estimate' temperature scenarios for 2011–2020, obtained as the sum of the observed baseline temperature and simulated temperature change.

Baseline	T_{obs}	ΔT_{mean}	T_{scen}
1951–2000	−6.3	2.1	−4.2
1961–2000	−6.4	2.0	−4.4
1971–2000	−5.6	1.7	−3.9
1981–2000	−5.4	1.7	−3.7
1961–1990	−7.2	2.1	−5.1

Now, let us assume that we need a scenario for the mean January temperature in, for example, the decade 2011–2020. A usual way for deriving such a scenario is to take the observed mean temperature during the baseline period and add to this a model-simulated change from the baseline period to the forecast period. For the latter, Table 1 uses the mean change over the 21 model simulations specified in Section 2. The simulated change also depends on the baseline period, but because the averaging over several model simulations largely eliminates internal climate variability, this dependence is much smoother than the variation in the observed temperatures. The warming to 2011–2020 is larger from baseline periods centred earlier in time, but the differences are relatively modest. The resulting 'best-estimate' scenario for the January mean temperature in 2011–2020 varies from −5.1°C for the baseline 1961–1990 to −3.7°C for the baseline 1981–2000.

As illustrated by this example, projections of future climate are in some cases quite sensitive to the baseline period used in the calculations. The heart of this problem is internal climate variability, which may make the climate of any period either warmer or colder than on the average expected for the external conditions during this period. If internal climate variability made the selected baseline period warmer (colder) than on the average expected, this will introduce a warm (cold) bias in projections of future climate beginning from this baseline period. This conclusion is valid even when the projections are expressed in probabilistic terms, rather than as single 'best-guess' numbers as in Table 1. Although the resampling ensemble technique developed in RR06 takes into account the effects of internal climate variability in the baseline period as well as the forecast period, it is only able to make this in a statistical sense. In other words, the method accounts for the fact that the baseline period may have been 'too warm' or 'too cold', but it assumes that both of these alternatives were equally likely.

In this paper, we will study the impact of the chosen baseline period in more detail, focusing on climate changes in southern Finland (60°N, 25°E). Furthermore, we will use cross-verification to study the optimal choice of the baseline period. Here, there are two conflicting issues. On one hand, a longer baseline period tends to make the impact of internal variability smaller. On the other hand, a longer baseline period extends further into the past, which implies larger uncertainty in the externally forced climate change from the baseline period to the forecast period.

As in RR06, we assume that uncertainty in modelling climate response to anthropogenic forcing is sufficiently captured by differences between existing climate models. This is an unverified and debated assumption (e.g. Allen and Ingram, 2002). However, as illustrated by a sensitivity study in Section 6, the uncertainty in forecasts of near-term climate change is dominated by natural variability and our results should, therefore, be relatively insensitive to a possible misrepresentation of modelling uncertainty.

2. Data

As in RR06, we used coupled atmosphere ocean general circulation model simulations produced for the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR4). The 21 models are listed in RR06. For each model, a simulation covering the 20th century (20C3M) and forced by a mixture of anthropogenic and (in most models) natural forcing factors was combined with a scenario simulation of the 21st century climate, to obtain continuous time-series covering at least the years 1901–2098. For the 21st century, simulations based on the SRES A1B emissions scenario (Nakićenović and Swart, 2000) were used; note, however, that the choice of the scenario is unimportant for the present study that focuses on near-term climate changes. Although parallel runs started from slightly different initial conditions are available for some models, we used (as in RR06) only one combined 20C3M + A1B simulation for each model. Also in common with RR06, we gave in our probabilistic calculations the same weight for each of the 21 models.

We use in our analysis a common $2.5^\circ \times 2.5^\circ$ latitude–longitude grid, which is representative of the typical resolution of the models. To avoid interpolation-induced smoothing, the regridding to the analysis grid was made by using, for each model, the values of the nearest original grid point. This differs from RR06, in which bilinear interpolation was used. However, a comparison of the results labelled as 'Res' in Fig. 3 of this paper with fig. 4 of RR06 indicates that the effects of this difference are quite modest.

The variance correction technique discussed in Section 3.1 below requires data on the observed interannual variability of climate. For this purpose, the CRU TS 2.1 data set (Mitchell and Jones, 2005) was used.

3. Methods

The motivation and implementation of the resampling ensemble method were described in some detail in RR06. Briefly, the method assumes that the probability distribution of the climate changes that result from a combination of anthropogenic forcing and natural variability is determined by the simulated multimodel mean change in the global mean temperature. This allows one to select, for any transient climate change simulation, several different pairs of periods that can be used as a surrogate for estimating the climate change between the actual baseline (e.g. 1971–2000) and forecast periods (e.g. 2011–2020). As a result, a larger sample for probabilistic forecasts of climate change is achieved. For example, by subsampling the 10-yr forecast period with a 5-yr interval, we found from the time-series obtained from the 20C3M and A1B simulations (years 1901–2098) 20 pairs of periods which shared essentially the same 21-model mean global warming as simulated from 1971–2000 to 2011–2020. As we have 21 models, this gives a nominal sample size of 420, although the subsampling means that the effective sample size is smaller. By using cross-verification, RR06 showed that the increased sample size allowed by the resampling results in potentially better probabilistic forecasts of climate change.

In this study we examine the sensitivity of probabilistic climate change forecasts to two details in our resampling methodology: a simple correction to model-simulated variability (which was not used in RR06) and the choice of the baseline period used in estimating the changes. These methodological details are discussed in more depth in the following two subsections. In this paper, we focus entirely on climate change from the late–20th-century to the near future (2011–2020), counting in climate change both the gradually emerging anthropogenic climate change signal and the effects of natural variability.

3.1. Variance correction

The resampling ensemble method developed in RR06 attempts to derive a probability distribution for the combination of forced anthropogenic climate change and natural climate variability. However, this aim may be compromised if the amplitude of variability in the models is too low or too high. To alleviate this potential problem, a simple adjustment (hereafter: variance correction) for biases in simulated variability was devised.

Because our main interest here is on changes in decadal mean climate, the variance correction should be ideally based on comparison of simulated and observed variances on the interdecadal timescale. Unfortunately, small sample sizes make such a comparison practically meaningless. We therefore compared the simulated and observed variances on the interannual timescale and assumed that an overestimate (underestimate) of interannual variability is accompanied by an equally large relative overestimate (underestimate) of interdecadal variability. This assumption will only be exactly valid if the simulated and observed time-series share the same autocorrelation structure, which may not always be the case. To try to reduce this problem, we also tested using the variance of longer-term (2- to 5-yr) averages of temperature and precipitation, but cross-verification indicated that this had no advantage over the use of interannual variability (see Section 4.2).

Having calculated the observed and simulated variances (as detailed in the end of this subsection), we made the variance correction separately for each individual model. The resampling method gives, for each model i, n possible realizations of climate change. In doing the variance correction for temperature

changes, we first calculated for each model the mean m_i of these n realizations. Then each individual realization of change was adjusted using the formula

$$\Delta a_{ij} = m_i + \sqrt{\frac{v_o}{v_i}}\left(\Delta o_{ij} - m_i\right), \tag{1}$$

where v_o and v_i denote the observed and simulated variances, Δo_{ij} (where $1 \leq j \leq n$) the original realization of climate change and Δa_{ij} the adjusted realization of change. As a result of this adjustment, the variance of the original realizations Δo_{ij}, which was assumed to be wrong by the factor v_i/v_o, becomes multiplied by v_o/v_i, but the mean of the realizations remains unchanged.

For precipitation changes, the same idea was used with two modifications. First, because precipitation changes are expressed here in per cent units, the squared coefficient of variation was used in the comparison instead of the absolute variance. Second, the mean change m_i in eq. (1) was replaced by the quantity

$$m_{Pi} = 100\% \times \left(\frac{P_f}{P_c} - 1\right), \tag{2}$$

where P_f and P_c are the mean values of precipitation averaged over all the n forecast (f) and control (c) periods used in the resampling method. The replacement of m_i by m_{Pi} is motivated by the expectation that m_{Pi} should give a better estimate of the per cent precipitation change that would be obtained in the absence of internal variability. However, the difference mainly matters in arid areas with a very irregular precipitation climate and is unimportant for the results discussed in this paper.

The observed variances v_o were calculated from linearly detrended 100-yr (1901–2000) time-series of temperature and precipitation using the University of East Anglia CRU TS 2.1 data set (Mitchell and Jones, 2005). Similarly, 100-yr detrended time-series for the period 1901–2000 were used for calculating the variance in the models. Despite the detrending, the simulated and the observed variance estimates may both be slightly affected by forced anthropogenic climate change. However, the analogous treatment of the model simulations and the observed time-series should minimize the effect that this contamination may have on the comparison of their variances.

The CRU data set has a much higher resolution ($0.5^\circ \times 0.5^\circ$) than the model simulations. Consequently, we calculated the variances v_o in two different ways. In the first method, the CRU temperature and precipitation time-series were first averaged to the same grid boxes ($2.5^\circ \times 2.5^\circ$) that were used for analysing the model data, after which the variances of these time-series were calculated. In the second method, the variances were calculated directly in the $0.5^\circ \times 0.5^\circ$ grid and then averaged over the $2.5^\circ \times 2.5^\circ$ grid boxes. The results shown in Sections 4 to 5 below use the first method, which provides a more fair comparison with the model simulations. On the other hand, the second variance estimate comes closer to truly local climate variability, which is relevant in many climate impact studies. The sensitivity of our probabilistic forecasts to the use of this second estimate is discussed briefly in Section 6.

3.2. Choice of the baseline period

In RR06, climate changes were calculated against the 30-yr baseline period 1971–2000. As illustrated in Section 1, however, scenarios of future climate may be affected by the choice of the baseline. We therefore repeated our probabilistic climate change calculations using three other baseline periods extending to the year 2000 (1951–2000, 1961–2000, 1981–2000), and also the baseline 1961–1990. The period 1961–1990 was included because it is still used as the baseline of present-day climate at some occasions, for example in climate impact research (e.g. Mearns et al., 2001).

To make the forecasts obtained with the various baselines comparable with each other, all results are expressed as temperature and precipitation differences relative to our reference baseline 1971–2000. To do this, we combine the model-simulated climate changes with the observed differences (from the CRU data set) in climate between 1971 and 2000 and the other baselines. The adjusted temperature changes ΔT are then defined as

$$\Delta T = \Delta T_s - \Delta T_o, \tag{3}$$

where ΔT_s is the simulated change from the selected baseline (e.g. 1951–2000) to the decade 2011–2020 and ΔT_o is the observed temperature difference between the reference baseline and the alternative baseline, for example $\Delta T_o = T(1971 - 2000) - T(1951 - 2000)$.

For changes in precipitation, the modification is different because we examined relative changes. The adjusted precipitation change becomes

$$\Delta P = 100\% \times \frac{\Delta P_s - \Delta P_o}{\Delta P_o + 100\%}, \tag{4}$$

where ΔP_s is the simulated per cent change representing the difference between the forecast period (e.g. 2011–2020) and the alternative baseline (e.g. 1951–2000), and ΔP_o is the observed per cent difference between this baseline and the reference baseline (1971–2000).

3.3. Cross-verification

For obvious reasons, forecasts of future climate change cannot be verified directly. However, under the assumption that the differences between the model simulations give a meaningful measure of uncertainty, the potential quality of the forecasts can be estimated by using cross-verification (RR06). In cross-verification the climate changes in one model are treated as a pseudo-truth and the probabilistic forecasts obtained by using the other 20 models are verified against it. This is repeated for all individual models, and the verification statistics are averaged over all cases and over the global area. When we calculated changes with

the variance correction in cross-verification mode the variance in the verifying model was assumed to be correct and was thus substituted for v_o in eq. (1).

The cross-verification results are expressed by using the continuous ranked probability score CRPS (Stanski et al., 1989; Hersbach, 2000; Candille and Talagrand, 2005). A lower CRPS indicates a better forecast. In this study, the cross-verification results are given separately for annual, seasonal (mean of CRPS in the four standard 3-month seasons) and monthly (mean of CRPS in the 12 calendar months) temperature and precipitation changes. By comparing the CRPS scores for the resampling method without and with the variance correction, and between different baseline periods, we wished to test which choices in the forecast methodology are likely to produce the best forecasts.

Obviously, cross-verification is a relative rather than absolute measure of forecast quality. It gives advice on how well different forecast strategies use the information available in the multimodel ensemble. However, if the ensemble as a whole would turn out to be substantially biased when compared with the real climate system, then the actual quality of the forecasts would be lower than the cross-verification suggests. The extent to which multimodel ensembles like the one used in this study capture the actual uncertainty in anthropogenic climate change is still actively debated (e.g. Allen and Ingram, 2002). We therefore include in Section 6 a test which illustrates the sensitivity of our probabilistic forecasts to an artificial amplification of intermodel differences.

4. Impact of variance correction

We first consider differences between the simulated and observed variability of temperature and precipitation. Figure 1a illustrates the ratio between the simulated and the observed standard deviation of annual mean temperature in Europe. The standard deviation in the models is represented by $Sim = \sqrt{v_s}$, where v_s is the 21-model mean interannual variance. In most of Europe, excluding the southernmost parts of the area, the simulated temperature variability exceeds the observed variability. For precipitation (Fig. 1b), the situation is quite different. In most of northern and Western Europe, the simulated interannual coefficient of variation of precipitation is notably smaller than the observed coefficient of variation. The reverse only happens in the Mediterranean area.

To adjust the derived probability distributions of temperature and precipitation change for the biases in the simulated variability, we applied the variance correction described in Section 3.1. The effect of this correction on the width (taken here as the difference between the 95th and 5th percentiles) of the probability distributions of annual mean temperature and precipitation change from 1971–2000 to 2011–2020 is illustrated in Fig. 2. In northern and Eastern Europe, the variance correction produces slightly narrower distributions of temperature change than the basic resampling method without the correction, whereas in southernmost Europe the distribution grows wider (Fig. 2a). For changes in precipitation (Fig. 2b), the variance correction makes the distribution wider in northern and Western Europe and in some areas of Eastern Europe. These results are consistent with the differences between the observed and simulated variability shown in Fig. 1.

4.1. *Probability forecasts of climate change from 1971–2000 to 2011–2020 in southern Finland*

In the following we introduce probability forecasts for temperature and precipitation changes from 1971–2000 to 2011–2020 for a grid box in southern Finland (60°N, 25°E). We used both the basic resampling method of RR06 and the resampling method with the variance correction for monthly, seasonal

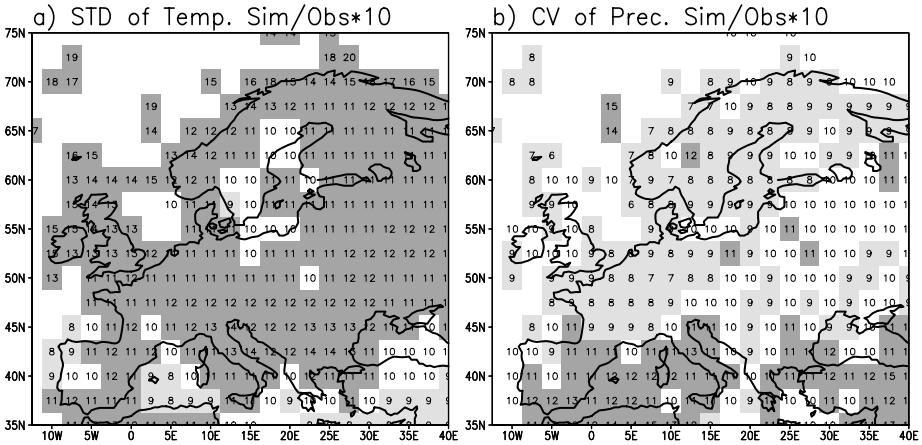

Fig. 1. (a) Ratio (×10) between the simulated and the observed interannual standard deviations (STDs) of annual mean temperatures in Europe. Dark (light) shading indicates areas where the simulated standard deviation is at least 5% larger (smaller) than the observed standard deviation. (b) as (a) but for the coefficient of variation of annual precipitation. See text for further details.

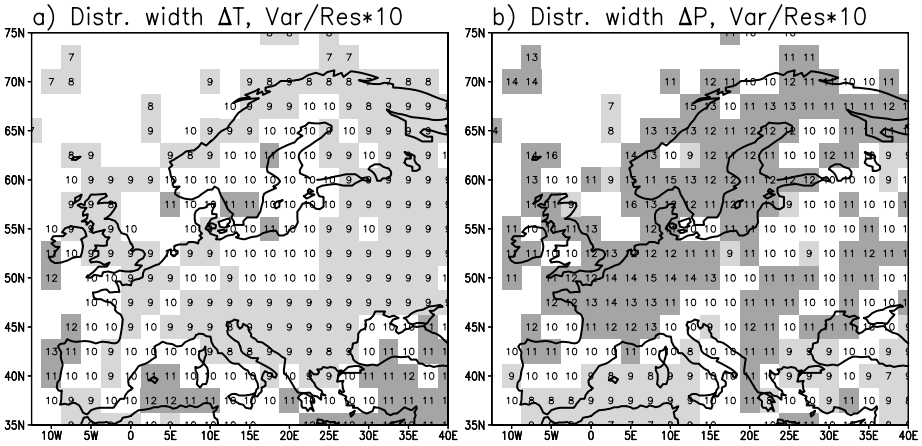

Fig. 2. Effect of the variance correction on the distribution width (difference between the 95th and 5th percentiles) of forecasts of annual mean (a) temperature and (b) precipitation change from 1971–2000 to 2011–2020. The numeric values give the ratio (×10) between the widths of the variance-corrected (Var) and uncorrected (Res) forecast distributions. Shading as in Fig. 1.

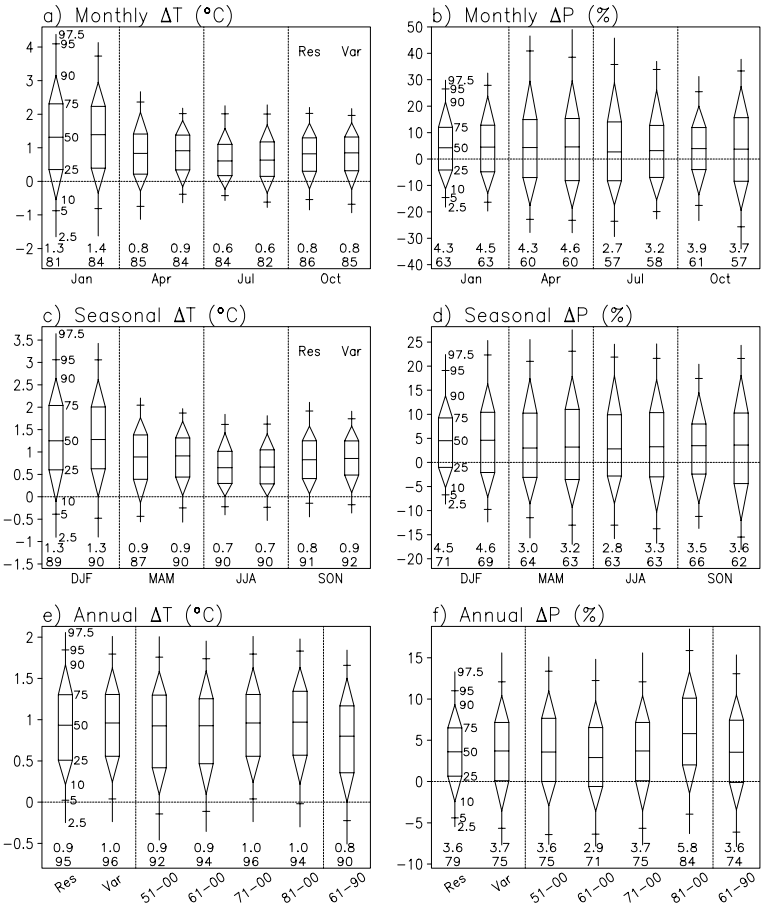

Fig. 3. Probabilistic forecasts for monthly (a–b), seasonal (c–d) and (e–f) annual mean temperature (left) and precipitation changes (right) from 1971–2000 to 2011–2020, in southern Finland. The whiskers show the 2.5st, 5th, 10th, 25th, 50th, 75th, 90th, 95th and 97.5th percentiles of the distributions, as indicated for the leftmost whiskers in each panel. The two rows of numbers in the bottom of the panels denote medians of the estimated probability distributions (in°C for temperature and in % for precipitation change) and the probabilities of increase (%). The whiskers in (a–d) and the two leftmost whiskers in (e–f) compare the distributions obtained without (Res) and with (Var) the variance correction, using 1971–2000 as the baseline. The five rightmost whiskers in (e–f) compare the distributions of annual mean temperature and precipitation change obtained by using different baseline periods in the calculations (with the variance correction included in all cases). Note that the vertical scales in the three rows differ.

(DJF = December–January–February, MAM = March–April–May, JJA = June–July–August, SON = September–October–November) and annual climate changes. As already shown in RR06, the probability distribution of temperature change in winter is much wider than the distributions for the other seasons or for the annual mean change. This is a consequence of larger temperature variability in winter than in the other seasons. In the case of precipitation change, the distributions in individual months are much wider than the seasonal and annual distributions (note that Figs. 3a–b, 3c–d and 3e–f have different scales).

For each of the four individual months studied (Figs. 3a–b), for the four 3-month seasons (Figs. 3c–d) and for the annual mean (left of Figs. 3e–f), Fig. 3 shows two probability whiskers: the one on the left for the basic resampling method (Res) and the one on the right for the resampling with the variance correction (Var). A comparison of these two adjacent whiskers indicates that the variance correction has in most cases only modest effects of the derived probability distributions of temperature change. However, the effect is mostly towards narrowing the distributions, particularly so in spring (MAM) and April and to a slightly lesser extent in winter (DJF) and January. On the other hand, the distributions of precipitation change grow in most cases wider by using the variance correction, which indicates increased uncertainty. This also means in general that the estimated probabilities of precipitation increase become smaller but the differences are only a few percentage units. For example, the probability of annual precipitation increase diminishes from 79% to 75%; seasonal and monthly differences are equal or slightly smaller. The probability of annual mean warming is very high both without (95%) and with (96%) the variance correction.

4.2. *Cross-verification*

Is it justified to assume that the probability distributions including the variance correction are more realistic than those derived with the basic resampling method without the correction? To obtain at least a tentative answer to this question, cross-verification was conducted as detailed in Section 3.3 and in RR06. CRPS scores were calculated both without and with the variance correction, and they were averaged over the global area and over all 21 choices of the verifying model. The leftmost column in Table 2 gives the resulting ratios of CRPS between the two methods, with values below one indicating that the variance correction leads to a potential improvement of the probabilistic forecasts.

The results suggest that the variance correction generally improves the forecasts, at least in the cross-verification framework. However, the improvement in the globally averaged CRPS scores is modest particularly for temperature changes. For changes in precipitation, the improvement is larger, but still only a half of the improvement obtained when replacing the straightforward method of only using one realization of climate change per one model with the basic resampling method (RR06).

For changes in precipitation, the variance correction leads to a decrease in the cross-verification CRPS scores in about 80% of the global area (not shown). For changes in temperature, however, there are wide areas particularly in low latitudes where the variance correction worsens the cross-verification performance (as illustrated for seasonal data in Fig. 4). It is probably not a pure coincidence that the largest deterioration in CRPS is seen over the Tropical Pacific. A basic assumption in our variance correction method is that the frequency spectrum of variability has similar shape for all models (and between models and observations), so that the correction factors needed on the interdecadal timescale can be derived directly from a comparison of interannual variances. This is probably not a good assumption in the Tropical Pacific where temperature variability is strongly dominated by the El Niño phenomenon and the timescale of El Niño varies markedly between different models and between many models and observations (AchutaRao and Sperber, 2006).

We also tested the variance correction method by replacing the interannual variances by the variances of longer-term (2- to 5-yr) averages of temperature and precipitation. However, cross-verification indicated that this had no advantage over the use of interannual variances (not shown). Although the use of longer-term averages is expected to reduce the errors associated with the non-universal frequency spectrum of variability, this advantage appears to be more than compensated by the disadvantage of reduced sample size.

Despite its limitations, the variance correction method appears in the light of the cross-verification results generally preferable over the basic resampling method. Consequently, we include the variance correction when studying the impact of the choice of the baseline period in the next section.

Table 2. Cross-verification CRPS ratios between climate change forecasts with the variance correction and without it (Var/Res), and between calculations based on five different baselines and the reference baseline period 1971–2000 (last five columns). The first three rows give the ratios of globally averaged CRPS scores for annual, seasonal and monthly temperature changes, and the last three the same for precipitation changes. Ratios below one indicate improvement. The forecast period is 2011–2020 and emission scenario is A1B in all calculations. See text for further details.

		Var/Res	1951–2000/ 1971–2000	1961–2000/ 1971–2000	1981–2000/ 1971–2000	1991–2000/ 1971–2000	1961–1990/ 1971–2000
ΔT	Annual	0.998	1.053	1.020	0.986	1.011	1.094
	Seasonal	0.994	1.035	1.012	1.001	1.059	1.080
	Monthly	0.991	1.019	1.003	1.014	1.097	1.064
ΔP	Annual	0.979	0.992	0.984	1.046	1.171	1.019
	Seasonal	0.978	0.976	0.974	1.040	1.172	1.010
	Monthly	0.974	0.967	0.967	1.037	1.175	1.004

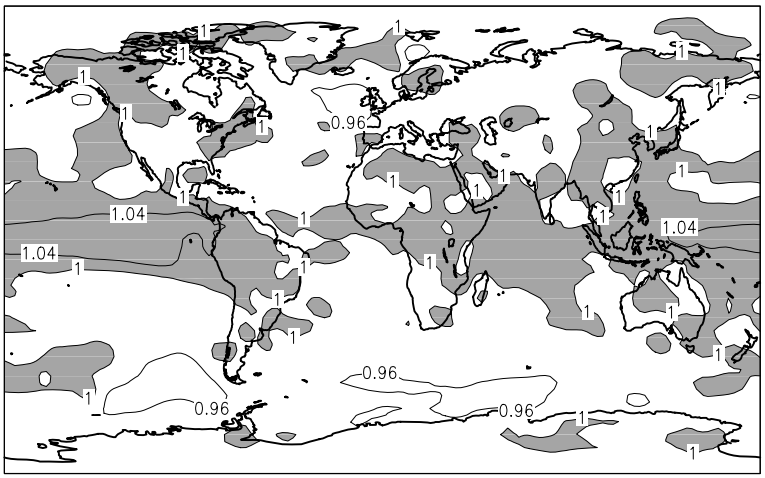

Fig. 4. The ratio of cross-verification CRPS scores of seasonal mean temperature change between the resampling method with the variance correction (Var) and without it (Res). Where the ratio is below one (unshaded areas) the variance correction improves the cross-verification performance. See text for further details.

5. Role of the baseline period

5.1. Probability forecasts of climate change to 2011–2020 as estimated by using different baselines

The impact of the chosen baseline period on the probability distributions of annual mean temperature and precipitation change in southern Finland (60°N, 25°E) is illustrated with whisker plots in Figs. 3e–f. In addition, Tables 3 and 4 give, for each of the five baselines studied, the median estimates of change and the estimated probabilities of increase from 1971–2000 to 2011–2020 for monthly, seasonal and annual means of temperature and precipitation. To facilitate the discussion of these results, the observed (CRU) temperature and precipitation differences between the other baselines and 1971–2000 are given in Table 5.

Whereas the impact of the variance correction is largely limited to the width of the probability distributions, the choice of the baseline period affects both the location and the width of the distributions but generally more the former than the latter. In the case of the annual mean temperature change, however, the sensitivity of the forecast to the choice of the baseline is relatively modest (Fig. 3e). The median estimate of the warming from 1971–2000 to 2011–2020 varies from 0.8°C for the baseline 1961–1990 to 1.0°C for the baselines 1971–2000 and 1981–2000, and the probability of temperature increase from 90% to 96% (for the baselines 1961–1990 and 1971–2000, respectively).

Forecasted changes in annual mean precipitation (Fig. 3f) are more sensitive to the choice of the baseline period than the forecasts of temperature change. The estimated probability of precipitation increase from 1971–2000 to 2011–2020 varies from 71% for the 1961–2000 baseline to 84% for the 1981–2000 baseline. These differences reflect the observed variations of precipitation given in the bottom row of Table 5; in particular the period 1981–2000 was about 3.5% wetter than the 1961–2000 as a whole. In addition, the shortest (20-yr) baseline results in notably wider probability distribution of precipitation change than the longer ones, which is caused by the increasing random variation of precipitation amount with decreasing averaging period.

Table 3. Probabilistic forecasts of temperature change in southern Finland from five different baselines to the period 2011–2020. The changes are expressed as temperature differences from the reference period 1971–2000. The medians (°C) of the derived probability distributions and the probabilities of warming (%; in parenthesis) are given for four months, four seasons and the annual means. The variance correction is included in all cases (see text).

ΔT	Baseline period					
Med (Prob)	1951–2000	1961–2000	1971–2000	1981–2000	1961–1990	Ranges
Jan	1.0 (79)	0.9 (75)	1.4 (84)	1.4 (84)	0.1 (54)	0.1–1.4 (54–84)
Apr	0.7 (81)	0.8 (83)	0.9 (84)	1.1 (91)	0.7 (81)	0.7–1.1 (81–91)
Jul	0.5 (71)	0.6 (75)	0.6 (82)	0.6 (77)	0.5 (75)	0.5–0.6 (71–82)
Oct	1.1 (87)	1.1 (89)	0.8 (85)	1.2 (91)	1.3 (91)	0.8–1.3 (85–91)
DJF	1.1 (85)	1.0 (83)	1.3 (90)	1.3 (88)	0.5 (68)	0.5–1.3 (68–90)
MAM	0.6 (77)	0.8 (86)	0.9 (90)	1.0 (93)	0.7 (84)	0.6–1.0 (77–93)
JJA	0.7 (90)	0.7 (92)	0.7 (90)	0.5 (84)	0.7 (89)	0.5–0.7 (84–92)
SON	1.1 (95)	1.0 (94)	0.9 (92)	0.9 (92)	1.1 (96)	0.9–1.1 (92–96)
ANN	0.9 (92)	0.9 (94)	1.0 (96)	1.0 (94)	0.8 (90)	0.8–1.0 (90–96)

Table 4. As Table 3 but for precipitation changes. The median changes are given in per cent of the mean precipitation in 1971–2000.

ΔP	Baseline period					
Med (Prob)	1951–2000	1961–2000	1971–2000	1981–2000	1961–1990	Ranges
Jan	6 (67)	2 (56)	5 (63)	17 (85)	−3 (42)	−3 to 17 (42–85)
Apr	7 (66)	8 (66)	5 (60)	−1 (47)	7 (65)	− 1 to 8 (47–66)
Jul	7 (70)	4 (59)	3 (58)	5 (60)	7 (71)	3–7 (58–71)
Oct	1 (55)	4 (58)	4 (57)	9 (67)	0 (49)	0–9 (49–67)
DJF	7 (76)	3 (63)	5 (69)	11 (87)	−1 (47)	−1 to 11 (47–87)
MAM	3 (63)	4 (67)	3 (63)	5 (69)	3 (63)	3–5 (63–69)
JJA	4 (67)	2 (57)	3 (63)	6 (70)	4 (65)	2–6 (57–70)
SON	2 (59)	3 (64)	4 (62)	3 (58)	6 (71)	2–6 (58–71)
ANN	4 (75)	3 (71)	4 (75)	6 (84)	4 (74)	3–6 (71–84)

Table 5. Observed (CRU) mean temperature and precipitation differences between four different baselines and the reference period 1971–2000, in southern Finland. For temperature the absolute (°C) and for precipitation the relative (%) baseline minus 1971–2000 differences are given.

Baseline period	1951–2000	1961–2000	1981–2000	1961–1990	1951–2000	1961–2000	1981–2000	1961–1990
ΔOBS	ΔT (°C)	ΔT	ΔT	ΔT	ΔP (%)	ΔP	ΔP	ΔP
Jan	−0.7	−0.7	0.3	−1.6	2.3	−2.3	12.4	−7.8
Apr	−0.3	−0.1	0.3	−0.3	2.2	2.6	−4.8	1.6
Jul	−0.2	−0.2	0.0	−0.3	3.3	1.1	1.6	4.3
Oct	0.1	0.2	0.5	0.2	−2.0	−0.6	5.3	−4.2
DJF	−0.6	−0.5	0.2	−1.1	1.3	−1.8	7.0	−5.6
MAM	−0.5	−0.2	0.3	−0.4	−0.4	0.9	2.5	−0.9
JJA	−0.1	−0.1	0.0	−0.1	−1.3	−2.0	3.1	−0.3
SON	0.1	0.1	0.2	0.1	−1.4	0.2	−0.4	1.6
ANN	−0.3	−0.2	0.2	−0.4	−0.6	−0.8	2.7	−0.9

Turning to the changes in seasonal and monthly mean climate, the numbers in Table 3 show that, in southern Finland, the forecasts of temperature change are relatively insensitive to the choice of the baseline period in spring, summer and autumn. In winter and particularly in January, however, the sensitivity is much larger. The median estimate of January mean temperature change from 1971–2000 to 2011–2020 varies from only 0.1°C (with the baseline 1961–1990) to 1.4°C (with the baselines 1971–2000 and 1981–2000), and the probability of warming from 54 to 84%. For the changes in winter (DJF) mean temperature, the corresponding ranges of median and probability of warming are 0.5°–1.3°C and 68–90%, respectively. The large differences between the forecasts obtained with the baseline 1961–1990 and the baselines 1971–2000 and 1981–2000 reflect the strong observed warming of winters in the late–20th-century (see Table 5 and Table 1), which accompanied an increase in westerly flow from the Atlantic Ocean and greatly exceeded the warming typically simulated by climate models between these two periods (e.g. Räisänen and Alexandersson, 2003).

The medians of precipitation change and probabilities of precipitation increase from 1971–2000 to 2011–2020 obtained by using the five different baselines are given in Table 4. The two longest baselines, 1951–2000 and 1961–2000, give results quite similar to those obtained with the baseline 1971–2000. For all the four single months considered, for all 3-month seasons and for the annual mean, the medians of change are positive (by 1–8%) for these three baselines. The same is in most cases true for the baseline 1961–1990, but in January and winter (DJF) this baseline gives median changes of −3% and −1% with probabilities of increase of only 42% and 47%. The forecasts obtained with the shortest baseline (1981–2000) are in some cases (e.g. in January and winter) markedly different from those for the reference baseline 1971–2000.

As a whole, the forecasts of precipitation change appear to be more sensitive to the chosen baseline than the forecasts of temperature change. However, regardless of the baseline used, the probability of annual precipitation increase exceeds 70%.

5.2. *Cross-verification*

Given that probabilistic forecasts of climate change are in some cases quite sensitive to the choice of the baseline period, which baseline period is likely to give the best results? To shed some light on this issue, we used cross-verification in the same way as described in Section 4.2. The resulting globally averaged CRPS values, normalized by the values obtained by using the baseline period 1971–2000, are shown in the five last columns of

Table 2. For illustration, we also include in this table one very short baseline (1991–2000), which was not considered in the previous subsection.

1. For changes in temperature, the lowest CRPS scores are obtained by using the baseline 1971–2000 (changes in individual calendar months) or 1981–2000 (changes in annual mean temperature). Both longer (1951–2000 and 1961–2000) and shorter (1991–2000) baselines work less well.
2. For changes in precipitation, the best baseline period appears to be 1951–2000 (monthly changes) or 1961–2000 (annual mean changes). The shortest baseline, 1991–2000, results in by far the worst cross-verification CRPS scores.
3. For both variables but particularly for temperature, the baseline 1961–1990 is worse than 1971–2000.

These results can be understood by considering the following facts. First, if internal climate variability were the only uncertainty in climate change forecasts, the baseline period should be as long as possible to reduce the impact of internal variability. Second, however, the uncertainty in forced (mainly greenhouse-gas induced) anthropogenic climate change increases with the difference in forcing conditions between the baseline period and the forecast period. In broad terms: the further in the past the baseline period is centred, the larger the uncertainty in the forced climate change grows. This effect is most clearly illustrated by the differences between 1961–1990 and 1971–2000. Third, the relative impact of the uncertainty in the forced climate change as compared with internal variability is larger for temperature than for precipitation, the changes of which have a lower signal-to-noise ratio (e.g. Räisänen, 2001). This tends to make the optimal baseline period longer for precipitation than temperature changes. The same argument explains why the optimal baseline period appears to be longer for changes in monthly than in annual mean climate.

Considering the results for both temperature and precipitation, 1971–2000 and 1961–2000 appear to be the best of the tested baselines. 1961–1990, although still used at some occasions because lack of suitable data for 1971–2000, is not recommendable from the cross-verification perspective.

6. Further sensitivity tests

As noted in Section 3.1, two estimates of interannual variability were derived from the CRU data set. For the first one, which was used in Sections 4 and 5 above, the CRU temperature and precipitation fields were averaged over the 2.5° × 2.5° grid boxes before calculating their interannual variance. For the second one, the order of variance calculation and horizontal averaging was reversed, to obtain estimates of variance that are more representative of climate variability on small horizontal scales.

As expected, the second method yields higher estimates of variance and, when substituted to the variance correction procedure, it results in wider probability distributions of climate change. In practice, however, this difference only matters for precipitation, and even for precipitation the effect is relatively modest. As an illustration, the 5–95% uncertainty ranges of DJF, JJA and annual mean temperature and precipitation change in the southern Finland (25°E, 60°N) grid box are compared between the two variance calculation options in the first two columns of Table 6. In the case of precipitation change, the second method gives up to 10% wider uncertainty ranges. For changes in temperature, the difference is at most 1%.

Another caveat in our derived probability distributions concerns the ability of the multimodel ensemble to capture the uncertainty in the noise-free anthropogenic climate change signal, that is, in the changes that would occur in the absence of natural variability. To study the potential importance of this issue, a sensitivity test was conducted in which *intermodel* differences in climate change were artificially amplified by the factor 1.5. Denoting the overall mean of the derived probability distribution as M, the variance-corrected climate changes Δa_{ij} were replaced, for each model i and realization j, with

$$\Delta b_{ij} = M + 1.5(m_i - M) + (\Delta a_{ij} - m_i), \quad (5)$$

where m_i is the mean of Δa_{ij} for the n realizations from model i. Thus, this adjustment amplified the differences between the model-specific mean changes but left the differences between the individual realizations from each model unchanged.

Table 6. 5-to-95% uncertainty ranges of winter (DJF), summer (JJA) and annual (ANN) mean temperature and precipitation changes from 1971–2000 to 2011–2020 in southern Finland. Var = standard variance correction method. Var2 = variance correction method with the alternative CRU variability estimate (see text). Fact1.5 = as Var, but with an artificial 50% increase in intermodel differences according to Eq. (5). For Var2 and Fact1.5, the numbers in parentheses give the relative increase in the width of the 5–95% range as compared with Var.

		Var	Var2	Fact1.5
ΔT	DJF	−0.5 to 3.1	−0.5 to 3.1 (+1%)	−0.5 to 3.1 (+3%)
	JJA	−0.2 to 1.6	−0.2 to 1.6 (+1%)	−0.3 to 1.8 (+12%)
	ANN	0.0 to 1.8	0.0 to 1.8 (0%)	−0.1 to 1.9 (+14%)
ΔP	DJF	−9.7 to 22.3	−10.9 to 23.6 (+8%)	−9.8 to 22.5 (+1%)
	JJA	−13.8 to 21.6	−15.1 to 22.7 (+7%)	−14.7 to 23.9 (+9%)
	ANN	−5.7 to 12.1	−6.5 to 12.9 (+10%)	−5.9 to 12.4 (+3%)

At the limit where intermodel differences dominated over the uncertainty associated with internal variability, so that the differences ($m_i - M$) were much larger than the differences ($\Delta a_{ij} - m_i$), the adjustment (eq. 5) would widen the derived probability distributions by almost 50%. In practice, the effect was much smaller (compare the first and third column in Table 6). The derived 5–95% uncertainty range in annual mean temperature change became 14% wider, and the relative increase in the corresponding seasonal ranges was even smaller. For the changes in precipitation, the effect was systematically smaller than for temperature changes. Thus, because of the dominant role of internal variability, probabilistic forecasts of near-term regional climate change appear to be relatively insensitive to the representation of modelling uncertainty.

7. Summary and discussion

In an earlier study, RR06 derived probabilistic forecasts of near-term climate change in southern Finland by using a resampling ensemble method. Their results included, for example, a 95% probability of annual mean warming from 1971–2000 to 2011–2020 and an 80% probability of increasing annual precipitation. In this paper, we studied the sensitivity of these probabilistic forecasts to two details in the methodology: a variance correction attempting to correct biases in the amplitude of model-simulated natural variability, and the choice of the baseline period. Following RR06, we also used cross-verification to study which choices in the methodology are likely to give the best probabilistic forecasts. Our main findings are listed below.

(1) **Sensitivity of forecasts to the variance correction.** The variance correction affects the width of the derived probability distributions. In southern Finland, the variance correction generally widens the distributions of annual, seasonal and monthly precipitation change so that the probability of precipitation increase becomes somewhat smaller. By contrast, the variance correction has practically no systematic effect on probabilistic forecasts of temperature change.

(2) **Sensitivity of forecasts to the choice of baseline period.** The selection of the baseline period affects both the location and width of the derived probability distributions. The probability distributions of precipitation change vary in some cases quite substantially, but the forecasts to temperature change are sensitive to the chosen baseline only in winter. On the whole, the forecasts of temperature change are less sensitive to the details of the methodology than forecasts of precipitation change.

(3) **Cross-verification.** Our cross-verification tests suggest that, in general, the variance correction should improve probabilistic forecasts of near-term climate change, especially in the case of precipitation change. Secondly, the optimal length of the baseline period for forecasts of temperature change appears to be close to 30 yr, and that for forecasts of precipitation change at least 30 yr. Furthermore, at least from the cross-verification perspective, 1971–2000 is a better baseline period than 1961–1990 for both temperature and precipitation.

Our probabilistic forecasts are based on the assumption that the uncertainty in the response of climate to changes in atmospheric composition is represented adequately by differences between existing climate models and that the forcing scenario used in the calculations is realistic. An eventual violation of these assumptions might to some extent compromise the accuracy of our results, although (as implicated by the results shown in Section 6) this caveat should be smaller for forecasts of near-term than long-term climate change. Regardless of this, our findings suggest that the amplitude of model-simulated variability and the choice of the baseline period are issues that may need careful consideration also when deriving probabilistic forecasts of near-term climate change for other parts of the world.

8. Acknowledgments

We acknowledge the international modelling groups for providing their data for analysis, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) for collecting and archiving the model data, the JSC/CLIVAR Working Group on Coupled Modelling (WGCM) and their Coupled Model Intercomparison Project (CMIP) and Climate Simulation Panel for organizing the model data analysis activity, and the IPCC WG1 TSU for technical support. The IPCC Data Archive at Lawrence Livermore National Laboratory is supported by the Office of Science, U.S. Department of Energy. We also thank two anonymous reviewers for their helpful comments. This research has been supported by the Academy of Finland (decision 106979).

References

AchutaRao, K. and Sperber, K. R. 2006. ENSO simulation in coupled ocean-atmosphere models: are the current models better? *Climate Dyn.* **27**, 1–15.

ACIA 2005. *Impacts of a Warming Arctic: Arctic Climate Impact Assessment*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1042.

Allen, M. R. and Ingram, W. J. 2002. Constraints on future changes in climate and the hydrologic cycle. *Nature* **419**, 224–232.

Candille, G. and Talagrand, O. 2005. Evaluation of probabilistic prediction systems for a scalar variable. *Q. J. R. Meteorol. Soc.* **131**, 2131–2150.

Cubasch, U., Meehl, G. A., Boer, G. J., Stouffer, R. J., Dix, M. and co-authors. 2001. Projections of future climate change. In: *Climate change 2001. The Scientific Basis* (eds J. T. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell, and C. A. Johnson). Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 525–582.

Giorgi, F. and Merns, L. O. 2003. Probability of regional climate change based on Reliability Ensemble Averaging (REA) method. *Geophys. Res. Lett.* **30**, 1629. (doi:10.1029./2003.GL0171.30).

Hersbach, H. 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15**, 559–570.

IPCC 2007. *Climate Change 2007: The Physical Science Basis*. (To be updated when full bibliographic information available).

Mearns, L. O., Hulme, M., Carter, T. R., Leemans, R., Lal, M. and co-authors. 2001. Climate scenario development. In: *Climate Change 2001. The Scientific Basis* (eds J. T. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell, and C. A. Johnson). Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 739–768.

Mitchell, T. D. and Jones, P. D. 2005. An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *International Journal of Climatology* **25**, 693–712.

Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T., 1996. The new ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–119.

Nakićenović, N. and Swart, R. (eds.) 2000. *Emissions Scenarios. A Special Report of Working Group III of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, 599.

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

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

Räisänen, J. and Palmer, T. N. 2001. A probability and decision model analysis of a multimodel ensemble of climate change simulations. *J. Climate* **14**, 3212–3226.

Räisänen, J. and Alexandersson, H. 2003. A probabilistic view on resent and near future climate change in Sweden. *Tellus* **55A**, 113–125.

Räisänen, J. and Ruokolainen, L. 2006. Probabilistic forecasts of near-term climate change based on a resampling ensemble technique. *Tellus* **58A**, 461–472.

Scaife, A., Knight, J. R., Vallis, G. K. and Folland, C. K. 2005. A stratospheric influence on the winter NAO and North Atlantic surface climate. *Geophys. Res. Lett*., **32**, L18715.

Stanski, H. R., Wilson, L. J. and Burrows, W. R. 1989. Survey of common verification methods in meteorology. Research report 89–5, Atmospheric Environment Service Forecast Research Division, Canada.

Tebaldi, C., Smith, R., Nychka, D. and Mearns, L. O. 2005. Quantifying uncertainty in projections of regional climate change: a Bayesian Approach. *J. Climate* **18**, 1524–1540.