Projections of daily mean temperature variability in the future: cross-validation tests with ENSEMBLES regional climate simulations

Jouni Räisänen and Olle Räty
Department of Physics, University of Helsinki, Finland

Submitted to Climate Dynamics, 14 June 2012
Returned for minor revision, 22 July 2012
Revised, 23 August 2012
Accepted, 27 August 2012

Corresponding author
Jouni Räisänen
Department of Physics, P.O. Box 48 (Erik Palménin aukio 1),
FI-00014 University of Helsinki, Finland
Phone +358-9-191 50872; Fax + 358-9-191 48802
Email: jouni.raisanen@helsinki.fi

Abstract

Because of model biases, projections of future climate need to combine model simulations of recent and future climate with information on observed climate. Here, ten methods for projecting the distribution of daily mean temperatures are compared, using six regional climate change simulations for Europe. Cross validation between the models is used to assess the potential performance of the methods in projecting future climate. Delta change and bias correction type methods show similar cross-validation performance, with methods based on the quantile mapping approach doing best in both groups due to their apparent ability to reduce the errors in the projected time mean temperature change. However, as no single method performs best under all circumstances, the optimal approach might be to use several well-behaving methods in parallel. When applying the various methods to real-world temperature projection for the late 21st century, the largest intermethod differences are found in the tails of the temperature distribution. Although the intermethod variation of the projections is generally smaller than their intermodel variation, it is not negligible. Therefore, it should be preferably included in uncertainty analysis of temperature projections, particularly in applications where the extremes of the distribution are important.

KEYWORDS: climate change, climate projection, temperature, daily variability, delta change, bias correction, cross validation, ENSEMBLES, Europe

1. Introduction

Despite decades of development, global and regional climate models (GCMs and RCMs) still show various kinds of biases in the simulation of the present-day climate (Randall et al. 2007, Christensen et al. 2007, van der Lindell and Mitchell 2009). Therefore, model-simulated future climate as such rarely provides a plausible projection of the actual future climate. To alleviate the impact of model biases, construction of climate projections also needs to extract information from the observed and simulated climates in the recent past.

As an example, three 30-year (930-day) time series of January daily mean temperature in Jyväskylä, central Finland, are shown in Fig. 1: the first from station observations in 1971-2000, the second from an RCM simulation during the same period, and the third from the same RCM in the end of this century (2069-2098). During the years 1971-2000, the RCM simulation exhibits both a cold bias and smaller than observed variability, and the distribution of the simulated temperatures shows less negative skewness than that observed. Considering these deficiencies, the simulation for 2069-2098 is unlikely to provide a good description of the climate in this period, even if the simulated changes in mean temperature and characteristics of variability turned out to be correct.

Two commonly used approaches to account for model biases are “delta change” and “bias correction” (Fig. 2). In the former, the projection for the future is obtained by perturbing an observed time series based on the difference between the simulated future and baseline climates. In the latter, the projection is built on the future simulation by the model, after correcting this based on the differences between the simulated and observed climate during the baseline period.

If only long-term climatic means are needed, the problem is technically simple. For example, a projection for the future mean temperature is easily derived as

$$\overline{p} = \overline{o} + (\overline{s} - \overline{c}) = \overline{s} - (\overline{c} - \overline{o}) \qquad (1)$$

where the overline indicates temporal averaging and the four letters stand for projection (p), baseline observations (o), scenario simulation for the future period of interest (*s*) and control simulation for the baseline period (*c*). In this case, the delta change and the bias correction approaches (the first and the second form in (1), respectively) give identical results. Note, however, that this result is neither unique nor necessarily optimal. As shown in recent studies (Buser et al. 2009; Boberg and Christensen 2012) and later in this paper, biases in simulated variability may also have implications for projections of the time mean climate.

If characteristics of variability are essential, the situation is more problematic. A constant delta change or constant bias correction would fail to account for either the changes or biases in the amplitude, shape of distribution, and temporal structure of the simulated variability. This can be improved by more sophisticated projection methods, but not without a potential trade-off. The more precisely a projection scheme attempts to correct for differences between simulated and observed climate or to incorporate simulated climate change, the more likely it is affected by features that are not statistically robust (e.g., random fluctuations in the tails of the distribution). The potential advantages of more sophisticated projection methods also need to be put in the context of the model- and scenario-related uncertainty in future climate change (Meehl et al. 2007, Christensen et al. 2007).

A large array of methods for projecting future climate variability on daily scales have been developed particularly for precipitation (see Maraun et al. 2010, for a recent review) but also for temperature (e.g., Engen-Skaugen 2007, Piani et al. 2010, Amengual et al 2012). The question thus arises, which of these different alternatives should be preferred? Although projection methods can be compared for their ability to reproduce present-day climate statistics (e.g. Themeßl et al. 2011, Dosiolo and Paruolo 2011), the crucial issue is their performance in future climate.

Although future climate is unknown, some inferences on the potential performance of different methods can be drawn from intermodel cross validation (Fig. 3). If simulations for both the recent past and the future are available for *N* models, any one of these can be left out for verification. In the same way as projections for the real future climate would be made by combining observations with the baseline and future simulations from different climate models, projections for the future climate in the verifying model can be derived by replacing observations with the baseline climate in this model. Unlike in the real world, this projection is verifiable against the actual future climate in the same model. Repeating this over all choices of the verifying model, statistics can be gathered that allow comparison between different methods of projection. Such a technique has already been used in studies focusing on projection of time mean climate (e.g. Räisänen and Ylhäisi 2011, Bracegirdle and Stephenson 2012, Maraun 2012), and it is also planned to serve as one of the main tools in the recently started European Concerted Research Action ES1102 VALUE (Validating and Integrating Downscaling Methods for Climate Change Research).

In the present study, which was in part inspired by VALUE, the focus is on the projection of daily mean temperatures. Ten different projection methods, broadly similar to those used in earlier studies, are applied to a subset of six RCM simulations for Europe from the ENSEMBLES (ENSEMBLE-based Predictions of Climate Changes and their Impacts) project (van der Linden and Mitchell 2009). Two main issues will be studied:

1. Which of the tested methods show the most promise for projection of future climate in light of their cross-validation performance? Can a single best method be identified, or would it be better to use several methods in parallel, to take into account the uncertainty in this choice (cf. Ho et al. 2012)?
2. How large is the uncertainty associated with the choice of the projection method compared with the variation of climate change between different models?

The model data and observations used are described in Section 2, and the projection methods are introduced in Section 3. The cross-validation results are presented in Section 4, whereas Section 5 studies the sensitivity of the projected future climate to the choice of the projection method. Finally, a synthesis of the main conclusions is presented in Section 6.

2. Data sets

Six RCM simulations from the ENSEMBLES project are used (Table 1), all using a different RCM and different driving GCM but the same (SRES A1B) emissions scenario. These data were retrieved from the ENSEMBLES Research Theme 3 web page (ensemblesrt3.dmi.dk/) in a regular 0.25° lon × 0.25° lat grid covering Europe and northernmost Africa. However, to reduce the computations, only 211 land grid boxes with 2.5° × 2.5° spacing were used in cross validation (Fig. 4a). Here we mainly use data for a 30-year baseline period (1971-2000) and a 30-year period in the end of this century (2069-2098), but some results for an earlier projection period (2001-2030) are also shown.

For testing real-world temperature projection, a total of 139 station time series were selected from the blended European Climate Assessment & Data archive available through the Royal Netherlands Meteorological Institute Climate Explorer (climexp.knmi.nl/), one per each 2.5° lat × 2.5° lon box where a station with near-complete time series (at least 99% of valid data in 1971-2000) was available (Fig. 4b). For each station, simulated time series were chosen from that one of the nearest nine 0.25° × 0.25° grid boxes which has the largest land fraction, using for consistency the same land-sea mask (from SMHI-BCM) in all cases. This procedure was adopted to avoid cases in which observations from a coastal but nevertheless land-based station would be combined with simulated time series from a sea-dominated grid box.

3. Projection methods

Ten different methods for constructing projections of future temperature variability are studied (Table 2). The delta change (M1-M5) and bias correction methods (M6-M10) are technically symmetric: the same computer subroutines can be used for both by only switching the order of the three input time series (observations and the baseline and future simulations) in the argument list. Therefore, only the delta change methods are described in the following.

M1 simply adds the time mean temperature change between the baseline and scenario simulations to each daily value in the observed time series. M2 also takes into account changes in the standard deviation, converting the values in the observed time series (o_i) to projected values (p_i) as

$$p_i = \bar{o} + (\bar{s} - \bar{c}) + (o_i - \bar{o})\frac{s_s}{s_c} \qquad (2)$$

where s_s and s_c are the simulated standard deviations during the scenario and baseline periods. In M3, changes in the sample third-moment skewness are also included, so that the skewness of the projected time series becomes

$$skew_p = skew_o + (skew_s - skew_c) \qquad (3)$$

where the subscripts p, o, s and c refer to the projected, observed, control simulation and scenario simulation time series, respectively. The condition (3) could be fulfilled by several different modifications to the data. Here, we follow the algorithm described by Ballester et al. (2010) in their electronic supplementary material.

M4 and M5 use the quantile mapping approach. Cumulative probability distributions of temperature are first estimated for both the control (F_c) and the scenario simulations (F_s), and each observed value o_i is then converted to

$$p_i = F_s^{-1}(F_c(o_i)) \qquad (4)$$

In implementing (4), two practical issues need to be solved. First, if the conversion $F_s^{-1}(F_c)$ is derived directly from an empirical quantile-quantile plot, it tends to become noisy near the tails of the distribution (see the crosses in Fig. 5a). To avoid this, some smoothing is needed. Second, if some of the observed values fall out of the range in the control simulation, the quantile-quantile relationship needs to be extrapolated beyond the simulated range.

M4 and M5 differ in how these practical issues are solved. In M4, the quantiles in the model simulations are smoothed using a running average, replacing the quantiles $F_c^{-1}(x), x \in [0,1]$ with

$$\tilde{F}_c^{-1}(x) = \int_{\max(x-D,0)}^{\min(x+D,1)} F_c^{-1}(x)dx \Big/ \int_{\max(x-D,0)}^{\min(x+D,1)} dx \qquad (5)$$

and similarly for $F_s^{-1}(x)$. Here we use D = 0.05, which was found to be close to optimal in terms of the cross-validation statistics. This smoothed quantile-quantile relationship is illustrated by the bold red line in Fig. 5a. A disadvantage of the smoothing is that it narrows the range of the data; for example $\tilde{F}_c^{-1}(0)$ equals the mean of the lowest 5% of the temperatures in the control simulation. We extrapolate towards low and high values assuming that the difference $\tilde{F}_s^{-1} - \tilde{F}_c^{-1}$ remains constant for $x < 0$ and $x > 1$ (dashed red lines in Fig. 5a).

In M5, simple linear regression is used to map F_c^{-1} on F_s^{-1} (blue line in Fig. 5a). This coarse-grained implementation of quantile mapping is used to study whether the more detailed treatment in M4 has additional value.

The projected scenario period (2069-2098) time series for the case introduced in Fig. 1 are shown in the left part of Fig. 6 (also note the statistics included in the figure panels). They illustrate the following key features:

1. In the delta change methods (M1-M5), the structure of the projection time series follows the observations, in the bias correction methods (M6-M10) the simulation for the scenario period.
2. The projected time mean temperature is the same for all of M1-M3 and M6-M8, but not for the quantile mapping methods M4-M5 and M9-M10. The causes of this difference will be discussed later in this section.
3. M2-M3 and M7-M8 all produce the same standard deviation of temperatures.
4. M3 and M8 additionally yield the same skewness of the distribution.
5. The skewness produced by M1-M2 and M5 (-0.9) is the same as in the observed time series in Fig. 1: none of these methods modifies the shape of the temperature distribution aside from its mean and standard deviation. Similarly, the projections based on M6-M7 and M10 have the same skewness (-0.5) as the RCM simulation for 2069-2098 in Fig. 1.
6. The extremes of the temperature distribution are particularly sensitive to the choice of the method. The maximum of the projected time series varies

from 5.5°C to 11.6°C, the minimum from -36.8°C to -23.2°C. For M8, the minimum is actually lower than that observed in 1971-2000 (-36.1°C).

As demonstrated by the quantile representation in the right part of Fig. 6, the selection of the method is not the only uncertainty in the projection. The choice of the RCM simulation also matters, although the size of the inter-RCM variation depends on the projection method used. We will study the relative roles of inter-method and inter-RCM uncertainty in more depth in Section 5.

M4-M5 give in this case a slightly lower mean temperature than the other delta change methods, which retain the change in mean temperature exactly as simulated by the model. This is explained by the cold bias in the simulation in 1971-2000 together with the decrease in variability occurring in 2069-2098 (Figs. 1 and 5a). Because the observed temperatures in 1971-2000 mostly fall in the upper part of the simulated distribution, where the quantile-quantile comparison indicates a smaller difference between 1971-2000 and 2069-2098, the mean temperature change as weighted by the distribution of observations becomes smaller than that directly simulated by the model. Conversely, M9-M10 indicate a larger increase in the mean temperature than the other methods. This is due to the underestimate in variability in the simulation for 1971-2000, which in bias correction type quantile mapping implies that a larger correction should be added to higher temperatures (Fig. 5b). As the simulated temperature distribution shifts upward from 1971-2000 to 2069-2098, the average correction in the latter period becomes larger than that in the former, thus amplifying the projected change in time mean temperature. Although the details of these results are case specific, the ability of quantile mapping to modify the time mean temperature change represents a generic difference from the other six methods.

In Figs. 5-6, we used (for simplicity of interpretation) only January data when estimating the changes (M1-M5) and biases (M6-M10) of the simulated January temperature distribution. This is not necessarily optimal, because the resulting relatively small sample size may introduce substantial noise. The noise can be reduced by using a wider time window in the estimation of climate changes and model biases, although potentially at the cost of some systematic error. In Section

4 below, we test three choices of the window length: one, two and three months. For the two-month window, for example, data from the second half of December to the first half of February are used in addition to January data when estimating the changes in January in M1-M5 and biases in M6-M10.

4. Cross validation

Should all of the ten methods be regarded as equally plausible, or are some of them more likely to give useful temperature projections than others? Here, we study this using cross validation between the six model simulations, as shown schematically in Fig. 3. One deterministic and two probabilistic statistics are computed, all based on a comparison between the quantiles of the projected and verifying temperature distributions ($T_{proj}(x)$ and $T_{ver}(x)$, $x = 0\ldots100\%$) in a given month and location. The mean square error is

$$MSE = A[(< T_{proj} > - T_{ver})^2] \tag{6}$$

where < > denotes the ensemble mean of the five (six minus verifying model) individual projections and A indicates averaging over the whole distribution from 0 to 100%, the 12 months, the 211 land grid boxes (weighted with the cosine of latitude), and the six choices of the verifying model. For calculating the continuous ranked probability score

$$CRPS = A[\int_{-\infty}^{T_{ver}} F(T_{proj})^2 dT_{proj} + \int_{T_{ver}}^{\infty} (1 - F(T_{proj}))^2 dT_{proj}] \tag{7}$$

we first form, separately for each quantile of temperature, a probabilistic forecast for T_{ver} from the discrete cumulative distribution $F(T_{proj})$ of the five T_{proj} values (cf. Räisänen and Palmer 2001). Our third score, *OutOfRange,* records the frequency of cases in which T_{ver} is below the lowest or above the highest of the five T_{proj} values. Unlike *MSE* and *CRPS, OutOfRange* is not a proper validation score in the sense that a lower value would always indicate a better forecast. By inflating the forecast distribution sufficiently, one could ensure *OutOfRange* = 0, while simultaneously making the forecast useless. A more useful interpretation is as follows. If T_{ver} and the five T_{proj} values are independent samples from the same statistical population (as they ideally should), then the probability that T_{proj} is the lowest or highest of these six values is 1/3. If *OutOfRange* exceeds this value, this indicates that the forecast obtained from the five T_{proj} values is underdispersive, thus underestimating the uncertainty in T_{ver}. Conversely, *OutOfRange* < 1/3 would indicate an overdispersive forecast.

A summary of the cross-validation statistics is given in Fig. 7. Focusing first on the statistics for the projection period 2069-2098 (using 1971-2000 as the baseline), we can note the following:

1. *MSE* and *CRPS* are relatively insensitive to the number of months used in estimating the changes (M1-M5) or the model biases (M6-M10). However, in nearly all cases, two-month aggregation of data performs better than the use of a single month, although some exceptions are found for individual verifying models particularly for the delta change methods (not shown). Differences between two and three months are unsystematic even when considering the six-model mean statistics shown in Fig. 7.
2. *MSE* and *CRPS* give a similar picture of the relative performance of the ten methods (although the differences in *MSE* are larger). The two simplest methods, M1 (constant change over the whole distribution) and M6 (constant bias correction for the whole distribution) perform less well than the others. The inclusion of the standard deviation in M2 and M7 gives a clear improvement, but there is little additional change in the statistics when also modifying the skewness (M3 and M8). Methods based on quantile mapping perform best within both the delta change group (M4-M5) and the bias correction group (M9-M10). M9 has both the lowest *MSE* and *CRPS* of the ten methods, although the difference from M10 is small.
3. *OutOfRange* is very close to the “desired” value of 1/3 (33.3%) for all of the bias correction methods. By contrast, the delta change methods tend to provide underdispersive projections, so that the verification falls more often out of the range of the five projections than it ideally should. This is understandable. Because the delta change projections from the five forecast models all modify the same underlying time series (the 1971-2000 time series in the verifying model), their differences do not fully

cover the effects of internal variability. This underdispersion becomes more pronounced for longer time windows in estimating the change.

The bottom row of Fig. 7 shows the corresponding verification statistics for the period 2001-2030, when climate changes from 1971-2000 are much smaller than in 2069-2098. The absolute intermodel differences in change are also smaller, making both *MSE* and *CRPS* lower during this period. Unlike in 2069-2098, M1 shows in 2001-2030 nearly identical performance with the other delta change methods. At this time, changes in the width and shape of the temperature distribution still have a very low signal-to-noise ratio. Their inclusion in the projection has, therefore, little impact on the cross-validation performance. Most of the bias correction methods are also close in performance to the delta change methods at this time, M9 being again the best in the whole group. However, M6 with its unrealistic assumption that model biases are constant throughout the distribution performs substantially worse than the other methods.

The *MSE* calculated for the whole temperature distribution can be written as the sum of two components: one arising from the error in the time mean temperature, and the other from errors in the differences between the individual quantiles and the mean value. The latter component (denoted as *debiased MSE* in Fig. 7) is typically much smaller than the former, particularly in 2069-2098. Therefore, most of the *MSE* and some of the intermethod differences in *MSE* actually reflect errors in the projected time mean temperature, rather than those in the width and shape of the distribution. In particular, the best performance of the quantile mapping methods (especially M9 and M10) in 2069-2098 results from the best projections of the time mean temperature. On the other hand, relatively large errors in the shape and the width of the distribution do distinguish the worst methods (M1 and M6 in 2069-2098 and M6 in 2001-2030) from the others.

In constructing Fig. 7, the same weight was given to all parts of the temperature distribution. Yet, for some applications the projection accuracy near the upper and/or lower tails of the distribution might be unproportionally important. To illustrate how the relative performance of the methods varies across the distribution, their *MSE* ranks in 2069-2098 are shown in Fig. 8 separately for all percentiles of temperature (similar analysis for *CRPS* gives very similar results). For this and all the later figures in this paper, the two-month time window is used.

For most of the distribution, the ranking of the methods is broadly consistent with Fig. 7. As an exception, the simple constant bias correction M6 actually has the lowest *MSE* at 16-34%, although it performs very poorly in the upper part of the distribution. Otherwise, M9 and M10 with the lowest overall *MSE*s dominate the top rank from the extreme lower tail up to 98%. However, in the extreme high end, the performance of M9 and particularly M10 deteriorates. Thus, these projection methods might not be optimal for applications that are particularly sensitive to extremely high temperatures. Considering how M9 and M10 function, this deterioration is not surprising. Both methods attempt to estimate the temperature dependence of model bias from comparison between the observed and simulated distributions during the baseline period (Fig. 5b). However, the highest temperatures simulated in the late 21st century by far exceed those in the baseline period. This results in substantial extrapolation uncertainty in the bias correction, which probably explains the deteriorating performance of M9 and M10 in the upper end of the distribution.

The intermethod variation of cross-validation statistics also depends on the location, month of the year, and the verifying model. We do not discuss the details of this variation here, but emphasize the general implication: a method that is best in an average sense will not be the best in all individual cases. This indicates that the choice between different projection methods represents a genuine uncertainty that cannot be fully eliminated by selecting a single best method.

Given the uncertainty in the projection methods, might it not be better to use several methods simultaneously instead of just one? To test this suggestion, four combinations of methods were chosen. B2, B4 and B8 include the best two (M9 and M10), four (M4-M5 and M9-M10) and eight methods (M2-M5 and M7-M10) in terms of the overall *MSE* and *CRPS* statistics for 2069-2098, while A10 includes all ten methods. In each case, the same weight was given to all of the methods included. Thus, for example, the best estimate projection from B8 becomes the mean of the multi-model means from all methods exluding M1 and

M6, while the corresponding probabilistic projection is formed by averaging the cumulative distribution functions from the same eight methods.

Figure 9 compares the cross-validation performance of these method combinations in 2069-2098 with that for the best four individual methods. When the number of methods combined increases, the range of projections included in the combination widens. Consequently, *OutOfRange* is already smaller for B2 than for the individual methods, and it decreases further when more methods are added (Fig. 9c). Still, as averaged over the whole temperature distribution, 15% of verification cases falling outside the predicted range remain even for A10.

As disussed above, decreases in *OutOfRange* do not necessarily imply a better projection. Indeed, *MSE* and *CRPS* (Figs. 9a,b) tell a partly different story. While *CRPS* averaged over the whole distribution is lower for all the four tested combinations than any individual method, it is at minimum for B4 that only includes the best four methods. Similarly, the B4 combination also has the lowest *MSE*. Its superiority over the other tested methods and combinations applies to most parts of the distribution (Figs. 9d,e). In particular, the probabilistic *CPRS* measure identifies B4 as the best approach with the only exception of the absolute extremes (0 and 100%). In the period 2001-2030 as well, *CRPS* and *MSE* averaged over the whole distribution are the lowest for B4 (not shown).

These findings suggest that temperature projections might be best derived by combining the information from the two delta change (M4-M5) and the two bias correction type quantile mapping methods (M9-M10). However, future research should test whether this conclusion remains valid for other model ensembles and other parts of the world.

5. Projections for the future

Here, we apply our methdology to real-world temperature projection for the set of 139 stations shown in Fig. 4b. Although the methodology provides projections for temperature in absolute units (cf. Fig. 6), we mostly show the results here as changes from the observed baseline distribution in 1971-2000.

5.1 Intermethod differences in projections

Intermethod differences in the projections are studied in Figs. 10 and 11. The first three rows of Fig. 10 summarize the projected six-RCM mean changes in the 1st, 10th, 50th, 90th and 99th percentiles of temperature, averaging the month- and station-specific values over the standard three-month seasons and over northern (48 stations north of 57.5°N), central (44 stations at 47.5°N-57.5°N) and southern Europe (47 stations south of 47.5°N). M1, which applies the same delta change in all parts of the distribution, provides a reference against which to compare the projections from the other methods. In line with earlier GCM and RCM studies (Räisänen et al. 2004, Kharin et al. 2007, Kjellström et al. 2007, Nikulin et al. 2011), the ENSEMBLES RCMs simulate seasonally varying changes in variability that are reflected in all of M2-M10. In winter and to some extent in autumn and spring, the simulated variability decreases particularly in northern Europe, resulting in larger changes in the lower than the upper end of the distribution. In summer, the reverse happens in central and southern Europe, with larger increases in the highest than in the lowest temperatures.

Differences also occur between the projections from M2-M10. For example, the contrast between the changes in the lowest and highest winter temperatures in northern Europe is less pronounced for M4 than the rest of M2-M10. This is at least partly due to the running averaging of the quantile-quantile relationship in M4, which contracts the range over which changes in variability can be taken into account (Fig. 5a). More strikingly, the apparent increase in the highest (lowest) summer temperatures in central and southern Europe is larger (smaller) for M6 than for the other methods. This is an artifact caused by the tendency of many of the models to overestimate present-day temperature variability in summer, a bias not corrected in M6.

The differences between the methods are smaller in the middle of the distribution than in the tails. However, M9 and M10 indicate a markedly smaller increase in median (50%) temperatures in central and southern Europe in summer than any other method. In southern Europe, M10 actually projects less warming than the other methods (excluding M6 near the lower tail) throughout the distribution. The explanation is analogous to the case shown in Figs. 5b and 6, but with the sign of the difference reversed. Because the simulated variability in southern and central

Europe in summer is too large, that is, the temperature bias increases with increasing temperature, a more negative bias correction is applied in M9 and M10 to the higher temperatures simulated in the future. This reduces the projected warming, just as recently shown for a similar bias correction method by Boberg and Christensen (2012). The results in Fig. 7 suggest that this feature may very well be an improvement: it was precisely the ability of M9 and M10 to modify the time mean temperature change that reduced the *MSE*s of these methods (and to some extent M4 and M5) in cross validation.

Another question of interest is how the choice of the method affects the intermodel variation of the projections (bottom row of Fig. 10). For M1, the intermodel standard deviation as calculated over all 139 stations is relatively small (1.1-1.2°C depending on season), being the same for all parts of the distribution. For the other methods, the standard deviation near the tails of the distribution is in most cases larger, particularly in the lower tail in winter and in the upper tail in summer. The intermodel variation tends to be the largest for M6, being amplified by uncorrected biases in variability. The standard deviation is in most cases smaller for the delta change than for the bias correction methods, because the former do not fully represent the uncertainty associated with internal variability (see the discussion of *OutOfRange* in Section 4). Note, however, the typically smaller standard deviations for M9-M10 than for the other bias correction methods.

To further compare the ten methods, pairwise intermethod root-mean-square (rms) differences in the six-RCM mean temperature projections are shown in Fig. 11 (see the caption for further details). Method pairs 2-3, 4-5, 7-8, and 9-10 all stand out as closely related, with very small differences in most of the distribution. In particular, the differences between M2 and M3 and between M7 and M8 are largely negligible, except for the extreme tails where changes and bias corrections of skewness have more substantial effects. Furthermore, while the M9-M10 differences are small in the lower and central parts of the distribution, they grow relatively large in the upper tail, reflecting the difficulty in the extrapolation of the bias correction beyond the range in the baseline period. As a whole, M6 and M1 are the methods furthest away from the others, but the tendency of M9 and M10 to differ relatively strongly from the other methods in the middle of the distribution also stands out.

5.2 Analysis of variance

In addition to the choice among the various projection methods, the projections also depend on the model simulation used. To assess the relative importance of these sources of uncertainty, fixed-effect analysis of variance was applied. The variance within each data set, consisting of all model- and method-specific projections for a given quantile of the temperature distribution at a given station and month, was decomposed as

$$V_{tot} = V_{mod} + V_{met} + V_{int} \tag{8}$$

where V_{mod} is the contribution of model differences (variation of multi-method mean projections across models), V_{met} that of method differences (variation of multi-model mean projections across methods), and V_{int} that of model-method interaction (method-dependence of intermodel differences, or equivalently model-dependence of intermethod differences). The computation of these terms is analogous to Eqs. (1)-(4) of Déqué et al. (2012). We stress that this decomposition does not aim to estimate the variances that would be observed within an infinite population of independent models and methods, but is rather used to diagnose the sources of variability within our specific set of (possibly non-independent) models and (certainly non-independent) methods. Furthermore, model simulations of climate always include unforced natural variability (Räisänen 2001, Yip et al. 2011). Some fraction of V_{mod} reflects this unforced variability rather than genuine intermodel differences in response to forcing, and to a lesser extent the unforced variability may also affect the other variance components.

As an illustration, the model- and method-dependence of projections for the 1st and 50th percentiles of January daily temperature in Jyväskylä, Finland in 2069-2098 is shown in Fig. 12. The projections for the 1st percentile vary widely between the models, but even more so between the methods. With all six models and all ten methods included, V_{tot} = 11.35 (°C)^2, of which 77% is attributed to method differences and only 6% to model differences, model-method interaction taking the remaining 17%. In particular, M1 gives systematically lower projections than the other methods, whereas the highest projections are generally obtained from M6. However, both of these methods are suspect due to their poor performance in cross validation. Indeed, the results in Fig. 9 suggest that it might be preferable to only retain the methods 4, 5, 9 and 10 included in the B4 combination. Doing this reduces the total variance by about 40%, but does not affect the relative shares of the different components in this particular case.

The projections for the 50th percentile are much less method-dependent (Fig. 12b). Almost 70% of the variance is attributed to model differences when all ten methods are included, and this increases to 88% when only the best four methods are retained. Conversely, the contribution of method differences is reduced from 15% in the fomer case to nearly zero in the latter.

Averaging over the 139 stations and 12 months confirms that intermodel differences strongly dominate the variance in the central parts of the temperature distribution (Fig. 13). Intermethod differences and model-method interaction both grow more important towards the tails of the distribution but do not become as dominant as in the case shown in Fig. 12a, especially not when only the best four methods are included. Averaging the variances over the whole distribution, V_{mod}, V_{int} and V_{met} contribute 73%, 13% and 14% in the 10-method case and 76%, 12% and 13% in the 4-method case, respectively. Therefore, the uncertainty associated with the choice of the projection method may be a secondary issue for many applications, although it clearly should not be neglected when and where the tails of the temperature distribution are particularly important. A similar conclusion – that uncertainty in bias correction is generally smaller than climate modeling uncertainty – was obtained by Chen et al. (2011), although their hydrological study addressed the bias correction uncertainty due to the choice of the baseline period rather than due to the choice of the method.

6. Conclusions

Projection of future climate cannot be generally based on model simulations alone but also requires information on the observed climate. This projection problem is often considered simple when only long-term climatic means are required, but it becomes more complicated when temporal variability is important. Here, we have focused on what is probably one of the easiest aspects of daily-scale variability for both climate models and in terms of its statistical properties, distributions of daily mean temperature in a changing climate. We first studied the relative strengths and weaknesses of ten projection methods using cross validation among six RCM simulations for Europe, all made with different RCMs and different driving GCMs. The main findings from these tests include the following:

1. Delta change and bias correction type methods showed similar overall performance in cross validation of late 21st century (2069-2098) temperature distributions. Within both groups, quantile mapping approaches performed best, due to their smallest errors in the projected time mean temperature. The simplest approaches assuming constant change or constant bias throughout the distribution were the worst, having larger errors in the distribution of temperature around the mean value than the other methods. In projections for early 21st century (2001-2030), the intermethod differences in verification statistics were smaller, except for the poor performance of the constant-bias bias correction method.
2. The performance of different projection methods may vary across the temperature distribution. In particular, quantile mapping type bias correction methods were found to be less reliable in the extreme upper tail than in the other parts of the distribution.
3. No single method performs best under all circumstances. Thus, to some extent, the choice of the projection method represents an uncertainty analogous to the choice of the climate model used for the projection. A natural way to take this uncertainty into account is to consider a few different but well-performing projection methods instead of just one. In our cross-validation exercise, the combination of the two delta change and two bias correction quantile mapping methods generally outperformed each individual method.

Second, we assessed the sensitivity of the resulting 21st century temperature projections to the choice of the method, to find that

1. The choice of the projection method has typically a larger impact in the tails of the temperature distribution than in the central parts. However, the latter

may also be affected. In particular, our quantile mapping type bias correction methods suggest a smaller warming in southern and central Europe in summer than would be inferred directly from the model simulations. This supports the recent findings of Boberg and Christensen (2012), in particular as these methods performed well in cross validation.

2. The uncertainty associated with the choice of the model simulation generally exceeds that due to the choice of the projection method. However, the relative importance of the method uncertainty increases towards the tails of the distribution, indicating that this uncertainty should also be considered at least in applications where extremely low or high temperatures are important.

Our study is based on only six RCM simulations and it only covers the European area. Its conclusions, particularly regarding the relative performance of different projection methods, should therefore be verified with other data sets. New opportunities for this will be provided by the CORDEX initiative (A COordinated Regional climate Downscaling EXperiment, cordex.dmi.dk/), and to some extent also by the GCM simulations conducted in the fifth phase of the Coupled Model Intercomparison Project (Taylor et al. 2011).

We also stress that the methods studied here were only designed for, and tested for their fidelity in, changing or correcting the local frequency distribution of daily mean temperatures. Issues that we have not addressed include the temporal (Haerter et al. 2011) and spatial autocorrelation structure (Huth 2002), as well as the correlation of temperature with other variables such as precipitation (Engen-Skaugen 2007).

Acknowledgments

The model simulations used in this work were funded by the EU FP6 Integrated Project ENSEMBLES (Contract number 505539). This work has been supported by the Academy of Finland RECAST project (decision 140801). The two reviewers are acknowledged for their constructive comments.

References

Amengual A, Homar V, Romero R, Alonso S, Ramis C (2012) A statistical adjustment of regional climate model outputs to local scales: application to Platja de Palma, Spain. J Climate 25: 939–957, doi: http://dx.doi.org/10.1175/JCLI-D-10-05024.1

Ballester, J., F. Giorgi, and X. Rodó (2010) Changes in European temperature extremes can be predicted from changes in PDF central statistics. Clim Change 98: 277-284, doi:10.1007/s10584-009-9758-0

Boberg F, Christensen JH (2012) Overestimation of Mediterranean summer temperature projections due to model deficiencies. Nature Climate Change 2: 433-436, doi: 10.1038/NCLIMATE1454

Bracegirdle TJ, Stephenson DB (2012) Higher precision estimates of regional polar warming by ensemble regression of climate model projections. Climate Dyn, doi: 10.1007/s00382-012-1330-3

Buser CM, Künsch HR, Lüthi D, Wild M, Schär C (2009) Bayesian multi-model projection of climate: bias assumptions and interannual variability. Climate Dyn 33: 849-868, doi: 10.1007/s00382-009-0588-6

Chen C, Haerter JO, Hagemann S, Piani C (2011) On the contribution of statistical bias correction to the uncertainty in the projected hydrological cycle. Geophys Res Lett 38: L20403, doi: 10.1029/2011GL049318

Christensen JH, Hewitson B, Busuioc A, Chen A, Gao X, Held I, Jones R, Kolli RK, Kwon W-T, Laprise R, Magaña Rueda V, Mearns L, Menéndez CG, Räisänen J, Rinke A, Sarr A, Whetton P (2007) Regional Climate Projections. Climate Change 2007: the Physical Science Basis, S Solomon et al., Eds., Cambridge University Press, pp 847-940

Déqué M, Somot S, Sanchez-Gomez E, Goodess CM, Jacob D, Lenderink G, Christensen OB (2012) The spread amongst ENSEMBLES regional scenarios: regional climate models, driving general circulation models and interannual variability. Climate Dyn 38: 951-964, doi: 10.1007/s00382-011-1053-x

Dosio A, Paruolo P (2011) Bias correction of the ENSEMBLES high-resolution climate change projections for use by impact models: evaluation on the present climate. J Geophys Res 116: D16106, doi: 10.1029/2011JD015934

Engen-Skaugen T (2007) Refinement of dynamically downscaled precipitation and temperature scenarios. Clim Change 84: 365-382, doi: 10.1007/s10584-007-9251-6

Haerter JO, Hagemann S, Moseley C, Piani C (2011) Climate model bias correction and the role of timescales. Hydrol Earth Syst Sci 15: 1065-1079, doi: 10.5194/hess-15-1065-2011

Ho CK, Stephenson DB, Collins M, Ferro CAT, Brown SJ (2012) Calibration strategies. A source of additional uncertainty in climate change projections. Bull Am Meteorolol Soc 93: 21-26, doi: 10.1175/2011BAMS3110.1

Huth R (2002) Statistical downscaling of daily temperature in Central Europe. J Climate 15: 1731–1742, doi: http://dx.doi.org/10.1175/1520-0442(2002)015<1731:SDODTI>2.0.CO;2

Kharin VV, Zwiers FW, Zhang X, Hegerl GC (2007) Changes in temperature and precipitation extremes in the IPCC ensemble of global coupled model simulations. Journal of Climate 20: 1419-1444, doi: 10.1175/JCLI4066.1

Kjellström E, Bärring L, Jacob D, Jones R, Lenderink G, Schär C (2007) Modelling daily temperature extremes: Recent climate and future changes over Europe. Clim Change 81: 249-265, DOI 10.1007/s10584-006-9220-5

Maraun D (2012) Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums. Geophys Res Lett 39: L06706, doi: 10.1029/2012GL051210

Maraun D, Wetterhall F, Ineson AM, Chandler ER, Kendon EJ, Widmann M, Brienen S, Rust HW, Sauter T, Themeßl, Cenema VKC, Chun KP, Goodess CM, Jones RG, Onof C, Vrac M, Thiele-Eich I (2010) Precipitation downscaling under climate change: recent developments to bridge the gap between dynamical models and the end user. Rev Geophys 48: RG3003, doi: 10.1029/2009RG000314

Meehl GA, Stocker TF, Collins W, Friedlingstein P, Gaye A, Gregory J, Kitoh A, Knutti R, Murphy J, Noda A, Raper S, Watterson I, Weaver A, Zhao Z-C (2007) Global climate projections. Climate Change 2007: the Physical Science Basis, S Solomon et al., Eds., Cambridge University Press, pp 747-845

Nikulin G, Kjellström E, Hansson U, Strandberg G, Ullerstig A (2011) Evaluation and future projections of temperature, precipitation and wind extremes over Europe in an ensemble of regional climate simulations. Tellus A 63: 41–55, doi: 10.1111/j.1600-0870.2010.00466.x

Piani C, Weedon GP, Best M, Gomes SM Viterbo P, Hagemann S, Haerter JO (2010) Statistical bias correction of global simulated daily precipitation and temperature for the application of hydrological models. J Hydrology 395: 199-215, doi: 10.1016/j.jhydrol.2010.10.024

Räisänen J (2001) CO_2-induced climate change in CMIP2 experiments. Quantification of agreement and role of internal variability. J Climate 14: 2088-2104, doi: http://dx.doi.org/10.1175/1520-0442(2001)014<2088:CICCIC>2.0.CO;2

Räisänen J, Palmer TN (2001) A probability and decision-model analysis of a multi-model ensemble of climate change simulations. J Climate 14: 3212-3226, doi: http://dx.doi.org/10.1175/1520-0442(2001)014<3212:APADMA>2.0.CO;2

Räisänen J, Ylhäisi J (2011) How much should climate model output be smoothed in space? J Climate 24: 867-880, doi: http://dx.doi.org/10.1175/2010JCLI3872.1

Räisänen J, Hansson U, Ullerstig A, Döscher R, Graham LP, Jones C, Meier HEM, Samuelsson P, Willén U (2004) European climate in the late 21st century: regional simulations with two driving global models and two forcing scenarios. Climate Dynamics 22: 13-31, doi: 10.1007/s00382-003-0365-x

Randall DA, Wood RA, Bony S, Colman R, Fichefet T, Fyfe J, Kattsov V, Pitman A, Shukla J, Srinivasan J, Stouffer RJ, Sumi A, Taylor KE (2007) Climate models and their evaluation. Climate Change 2007: the Physical Science Basis, S Solomon et al., Eds., Cambridge University Press, pp 589-662

Taylor KE, Stouffer RJ, Meehl GA (2011) A summary of the CMIP5 experiment design. Available at http://cmip-pcmdi.llnl.gov/cmip5/docs/Taylor_CMIP5_design.pdf

Themeßl MJ, Gobiet A, Leuprecht A (2011) Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. Int J Climatology 31: 1530-1544, doi: 10.1002/joc.2168

van der Linden P, Mitchell JFB, Eds. (2009) ENSEMBLES: Climate Change and its Impacts: Summary of research and results from the ENSEMBLES project. Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK, 160 pp. Available at http://ensembles-eu.metoffice.com/docs/Ensembles_final_report_Nov09.pdf

Yip S, Ferro CAT, Stephenson D (2011) A simple, coherent framework for partitioning uncertainty in climate change predictions. J Climate 24: 4634-4643, doi: 10.1175/2011JCLI4085.1

Tables

Table 1 *The RCM simulations used in this study*

Driving GCM	RCM	Institution	Shorthand
ARPEGE	ALADIN	CNRM	CNRM-A
HadCM3Q0	CLM	ETHZ	ETHZ-H0
HadCM3Q3	HadRM3Q3	Met Office	METO-H3
HadCM3Q16	HadRM3Q16	Met Office	METO-H16
ECHAM5-r3	REMO	MPI	MPI-E5
BCM	RCA3	SMHI	SMHI-BCM

The first column indicates the driving global climate model, the second the regional climate model and the third the institution that conducted the simulations, using model and institution acronyms that follow the ENSEMBLES Research Theme 3 web page (http://ensemblesrt3.dmi.dk/). The last column gives the shorthand notations used in this article

Table 2. *The projection methods used in this study*

M1	Delta change: mean
M2	Delta change: mean + standard deviation
M3	Delta change: mean + standard deviation + skewness
M4	Delta change: quantile mapping using smoothing
M5	Delta change: quantile mapping using linear regression
M6	Bias correction: mean
M7	Bias correction: mean + standard deviation
M8	Bias correction: mean + standard deviation + skewness
M9	Bias correction: quantile mapping using smoothing
M10	Bias correction: quantile mapping using linear regression

Figures

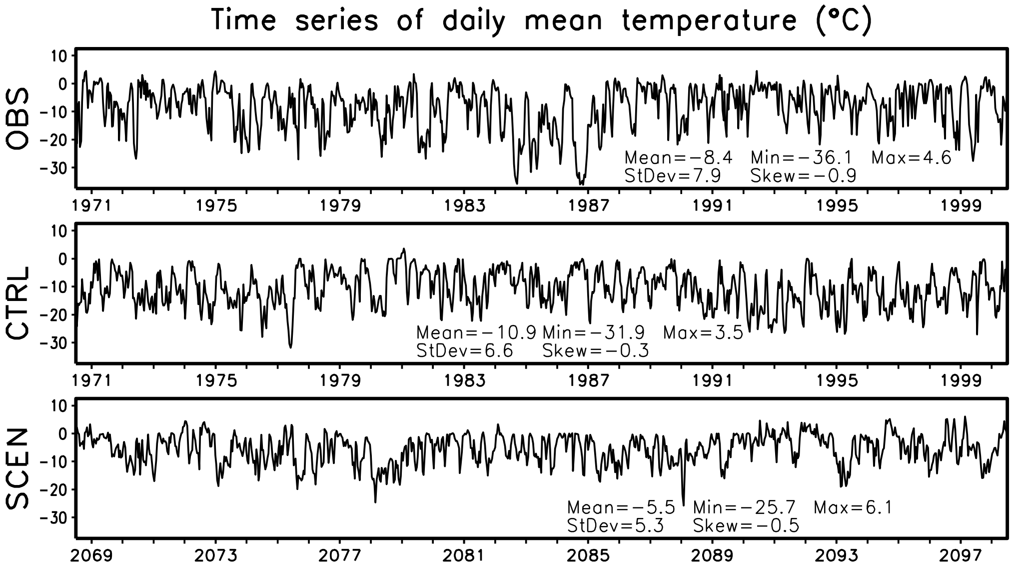

Fig. 1 *30-year time series of January daily mean temperature in Jyväskylä, Finland (62.4°N, 25.7°E), as observed in 1971-2000 (top), in the ETHZ-H0 (see Table 1) RCM simulation during the same period (middle), and in the same RCM simulation in 2069-2098 (bottom)*

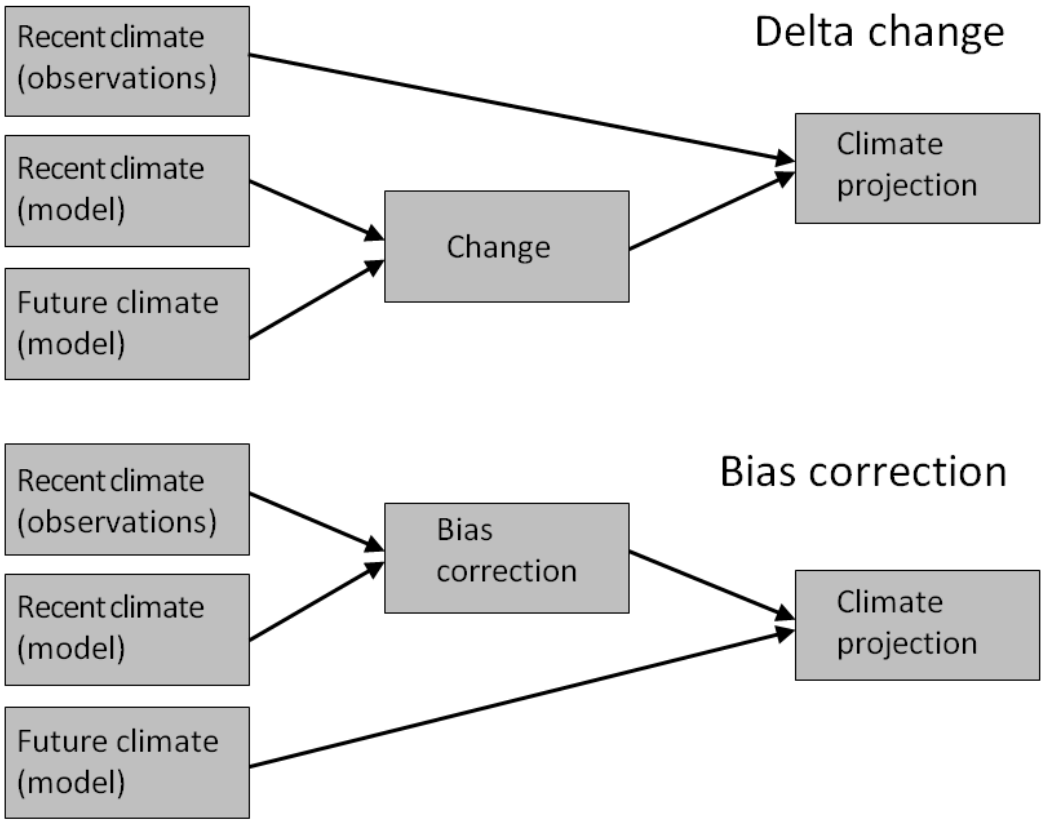

Fig. 2 *A schematic illustration of delta change and bias correction methods*

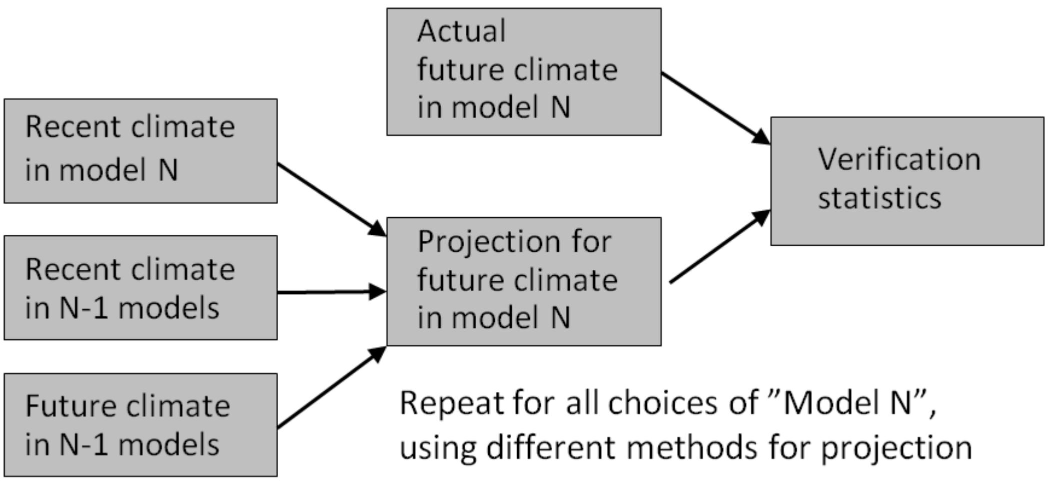

Fig. 3 *Principle of cross validation, as used in this study*

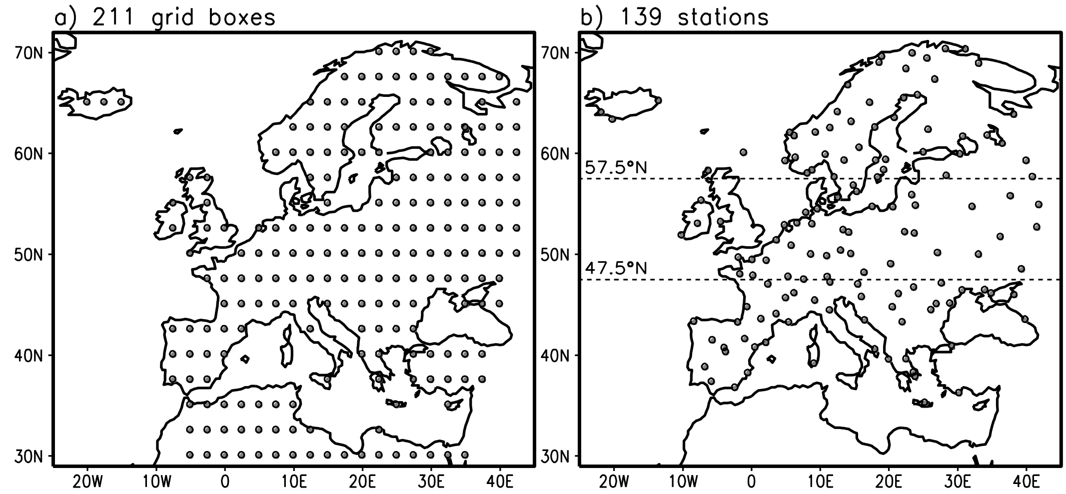

Fig. 4 *Locations of (a) the 211 grid boxes used in cross validation in Section 4, and (b) the 139 weather stations used in real-world temperature projection in Section 5. The latitudes 47.5°N and 57.5°N used for the division in Fig. 10 are also indicated in (b)*

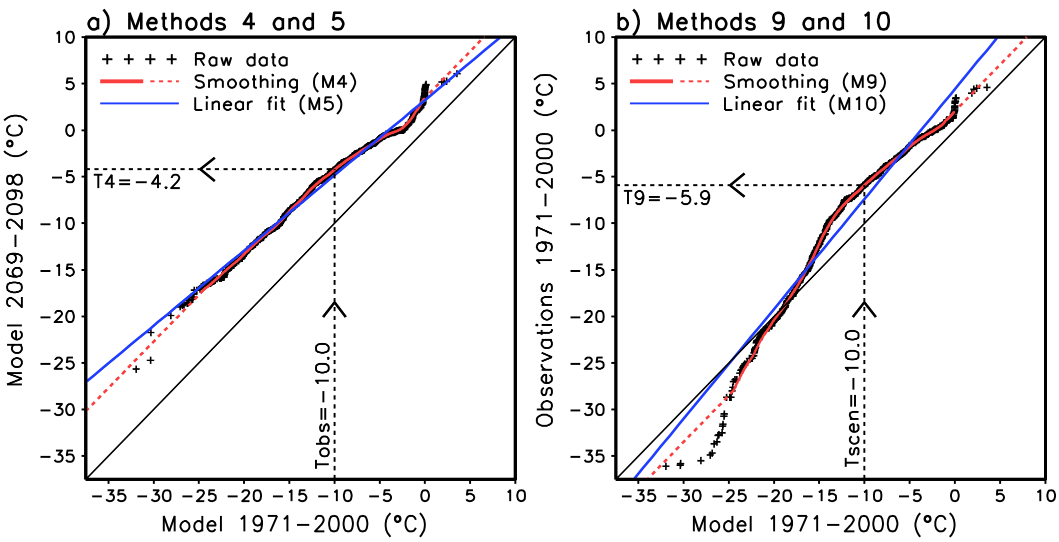

Fig. 5 *An illustration of quantile mapping methods for the case depicted in Figs. 1 and 6. In (a), the crosses show the quantile-quantile plot obtained directly from the simulations for 1971-2000 and 2069-2098, the red line gives the smoothed curve used in M4 and its extrapolation, and the blue line depicts the linear regression used in M5. Using M4, an observed temperature of -10.0°C would be converted to -4.2°C in the projection for 2069-2098. (b) is the same for the comparison of the simulation and observations in 1971-2000, as used in M9 and M10. In M9, a temperature of -10.0°C in the scenario period simulation would be converted to -5.9°C in the projection*

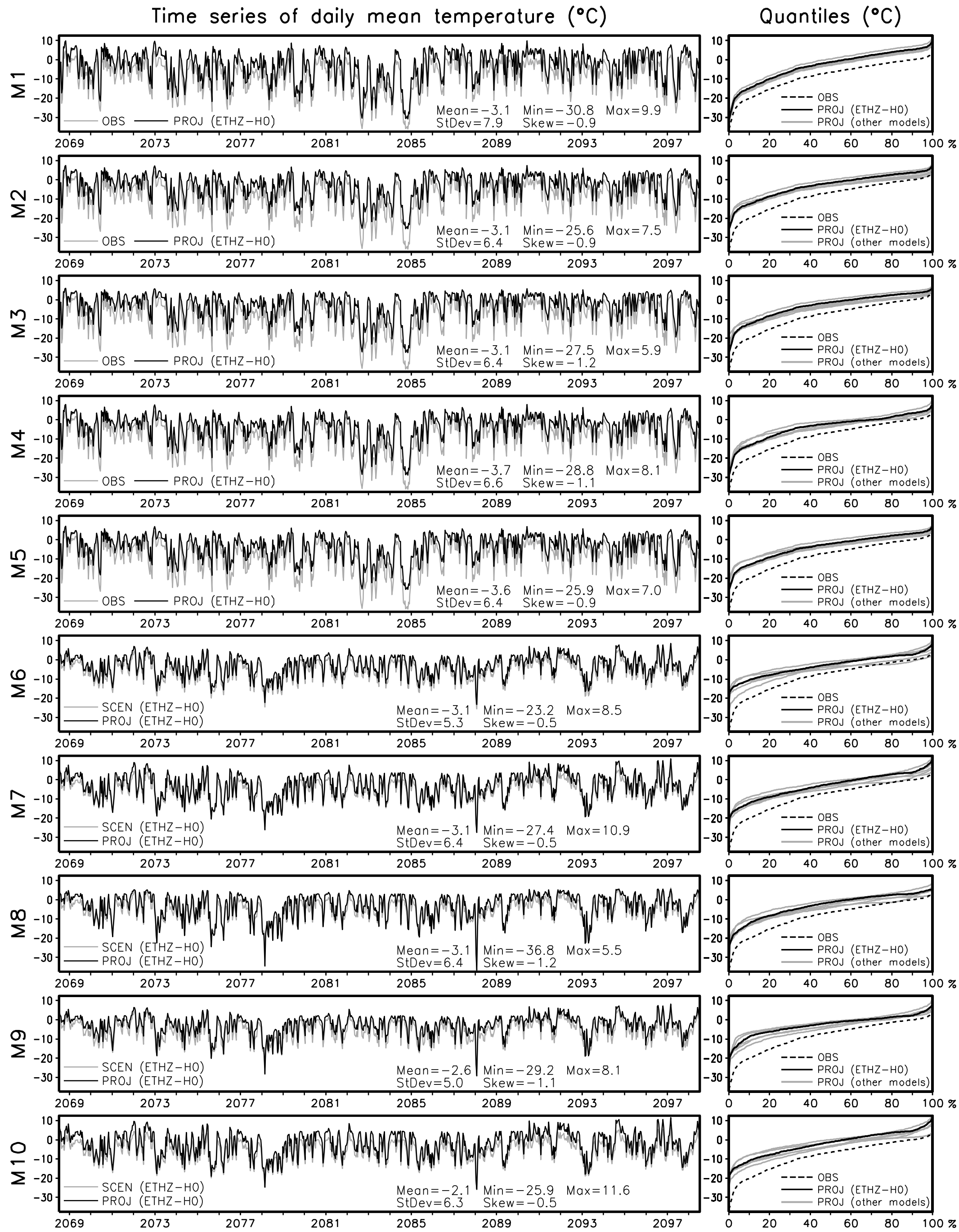

Fig. 6 *(left) Projected 30-year (2069-2098) time series of January daily mean temperature in Jyväskylä, Finland, using data from the ETHZ-H0 simulation, in methods M1-M10 (black lines). The grey lines in the top five (bottom five) panels show observations for 1971-2000 (the ETHZ-H0 scenario simulation for 2069-2098). The numeric values in each panel give the mean, minimum and maximum, standard deviation and the skewness of the distribution within the projected time series. (right) Quantiles of the observed distribution in 1971-2000 (dashed, same in all panels), and of the projected distributions obtained using data from ETHZ-H0 (black solid line) and the other five RCMs (grey lines)*

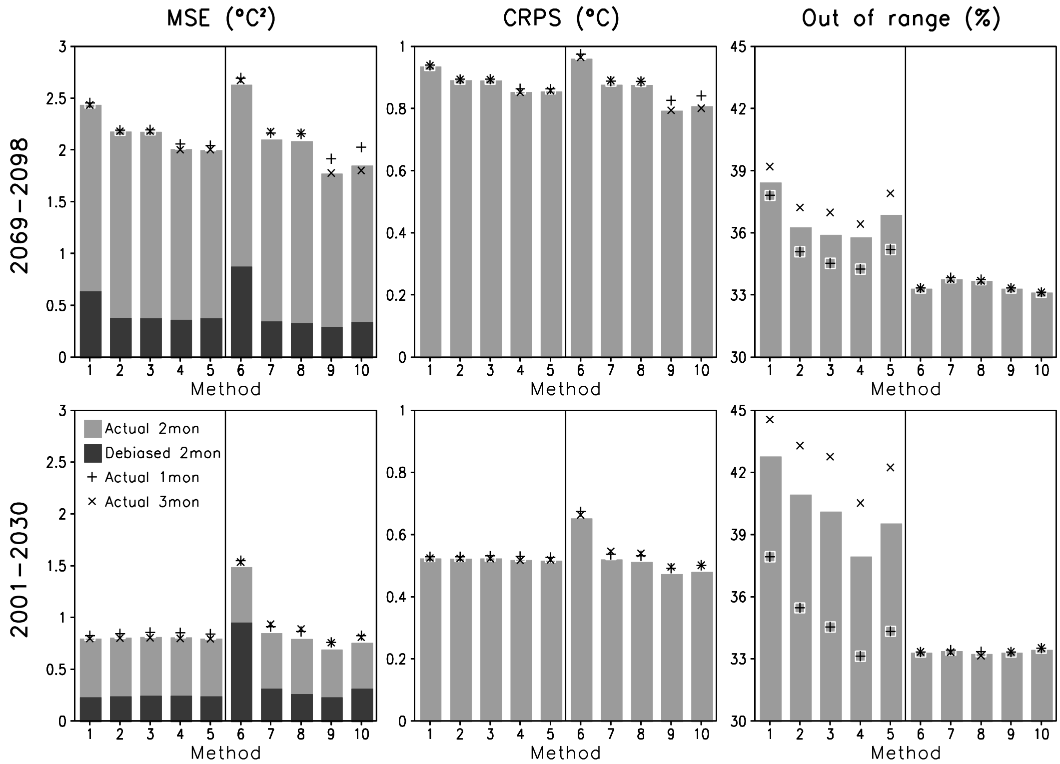

Fig. 7 *Cross-validated MSE, CRPS and OutOfRange for temperature distributions in the years 2069-2098 (top) and 2001-2030 (bottom). The MSE that would have been reached if always predicting the correct 30-year monthly mean temperature is also shown (dark part of the bars in the left column). The bars give statistics based on two-month sampling of climate changes (M1-M5) and biases (M6-M10); the plus signs (+) and crosses (×) show the corresponding values for 1-month and 3-month sampling*

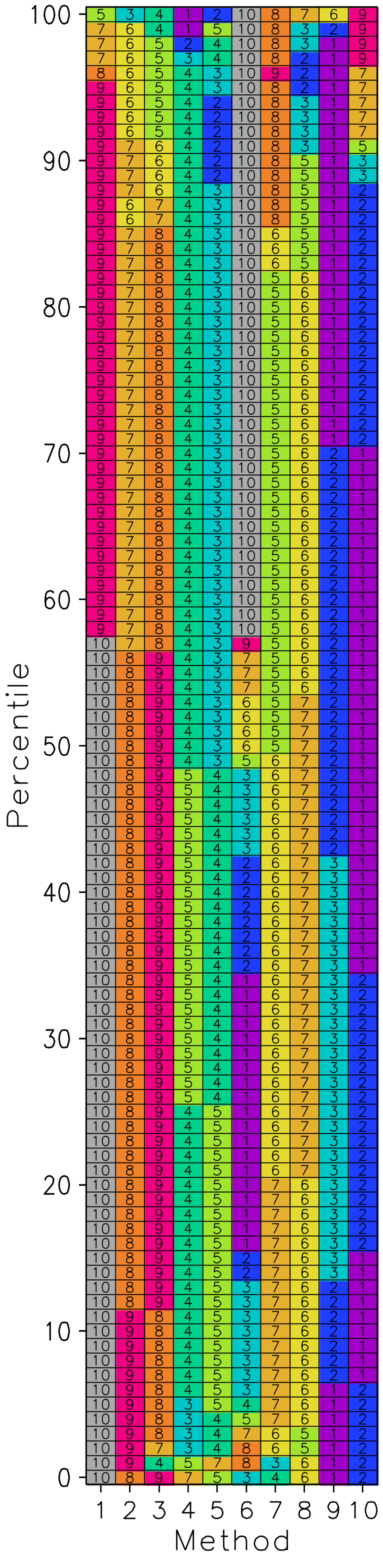

***Fig. 8** Ranking of the 10 methods (1 best, 10 worst) for cross-validated MSE of different percentiles of the temperature distribution in 2069-2098 (0% = absolute monthly minima, 100% = absolute monthly maxima)*

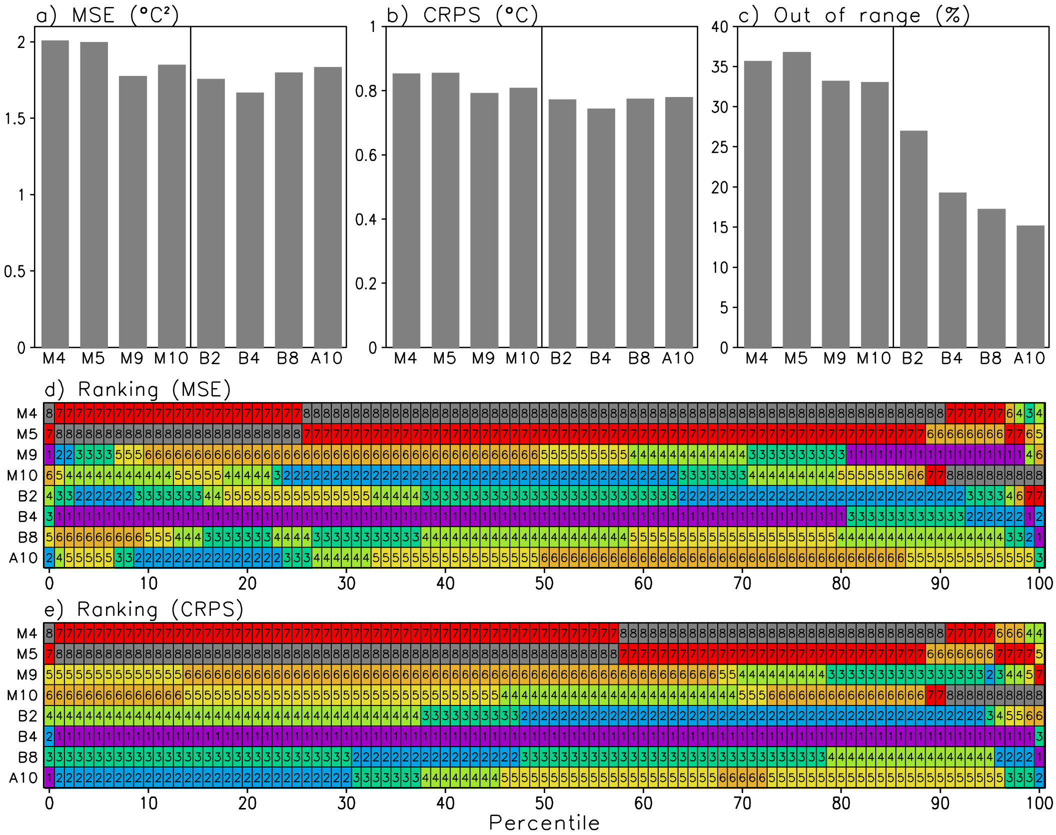

Fig. 9 *Cross-validation statistics for temperature in the years 2069-2098. The top row shows MSE, CRPS and OutOfRange separately for four individual methods (M4, M5, M9 and M10) and for the combinations B2, B4, B8 and A10 defined in the text. The last two panels indicate the ranking (1 = best, 8 = worst) of the MSE and CRPS values within this sample of methods in different parts of the temperature distribution*

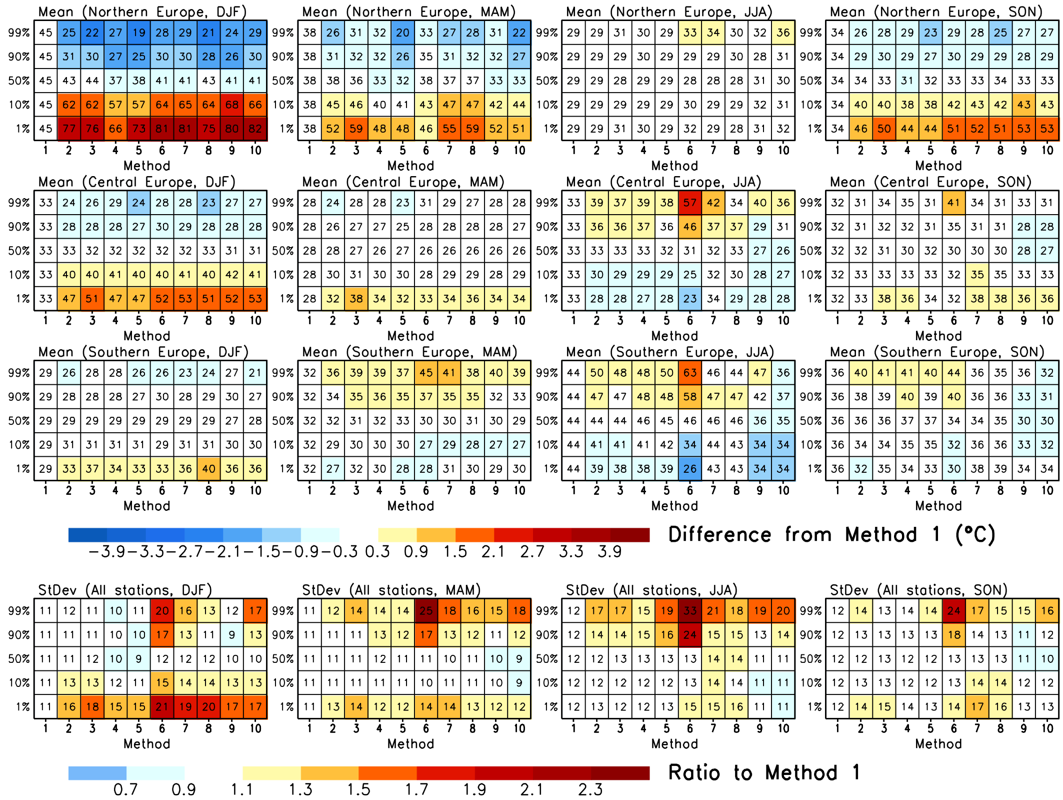

Fig. 10 *Summary of temperature projections for the years 2069-2098. The first three rows show the six-model mean changes in five quantiles of the temperature distribution (1% to 99%), as averaged over northern, central and southern Europe (48 stations north of 57.5°N, 44 stations at 47.5°N-57.5°N and 47 stations south of 47.5°N, respectively). The quantiles were first calculated for each month and then averaged over the three-month seasons identified in the figure headers. The numeric values indicate the absolute difference from the observed value in 1971-2000 (unit: 0.1°C), and the shading gives the difference from the value for M1. The numeric values in the last row show the intermodel standard deviation of the projections (unit: 0.1°C), as calculated from the variance averaged over all 139 stations and the three months in each season. The shading indicates the ratio to the standard deviation for M1*

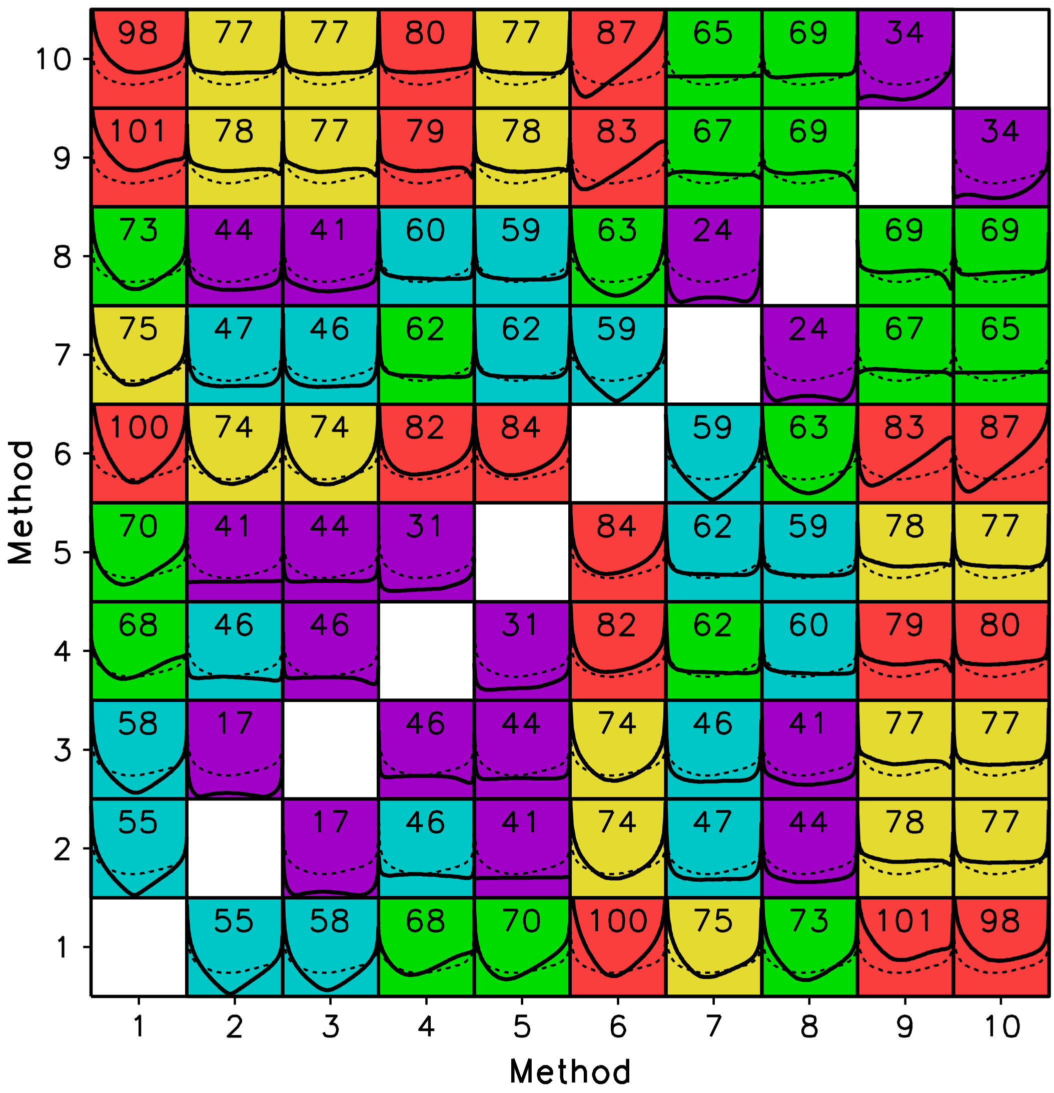

Fig. 11 *Intermethod rms differences of six-model mean temperature projections for the years 2069-2098, using data for all 12 months and the 139 stations. In each cell, the thick solid line shows the rms difference between the two methods compared (scale from 0 to 2°C) for quantiles ranging from 0 to 100% (left to right), while the dashed line gives the rms difference averaged over all method pairs. The cells are coloured according to the rms difference calculated for the whole distribution (violet: lowest 20% of cases … red: highest 20% of cases), which is plotted in each cell in units of 0.01°C*

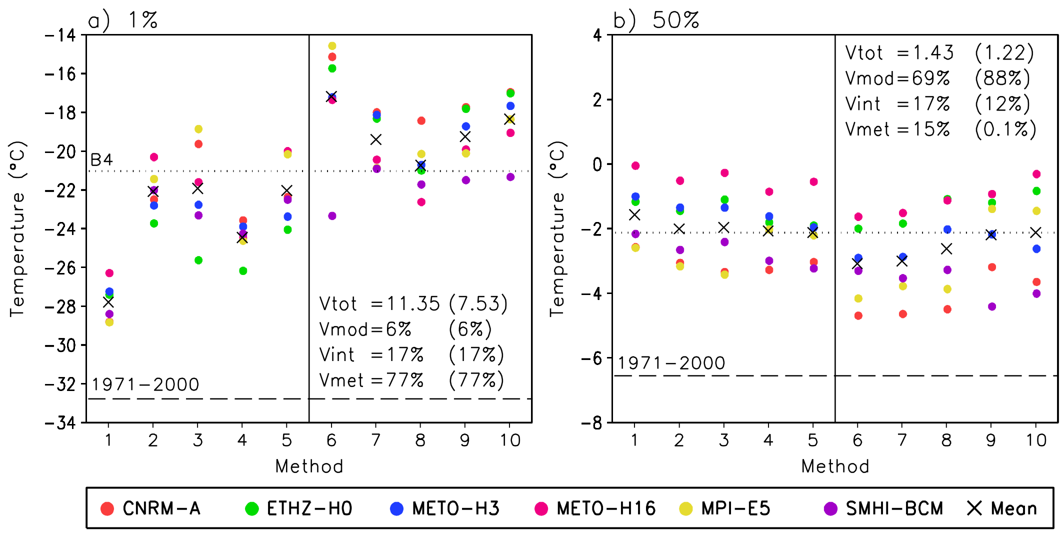

Fig. 12 *Projections for (a) the 1st and (b) the 50th percentile of daily mean temperature in January in Jyväskylä, Finland in 2069-2098, based on the different methods and RCM simulations. The dashed lines show the observed values in 1971-2000, and the dotted line the six-model means averaged over the methods 4, 5, 9 and 10 The total variance in (°C)^2 and the relative contributions of model differences, model-method interaction and method differences are also shown, both when including all 10 methods (first numbers) and when only including methods 4, 5, 9 and 10 (numbers in parentheses)*

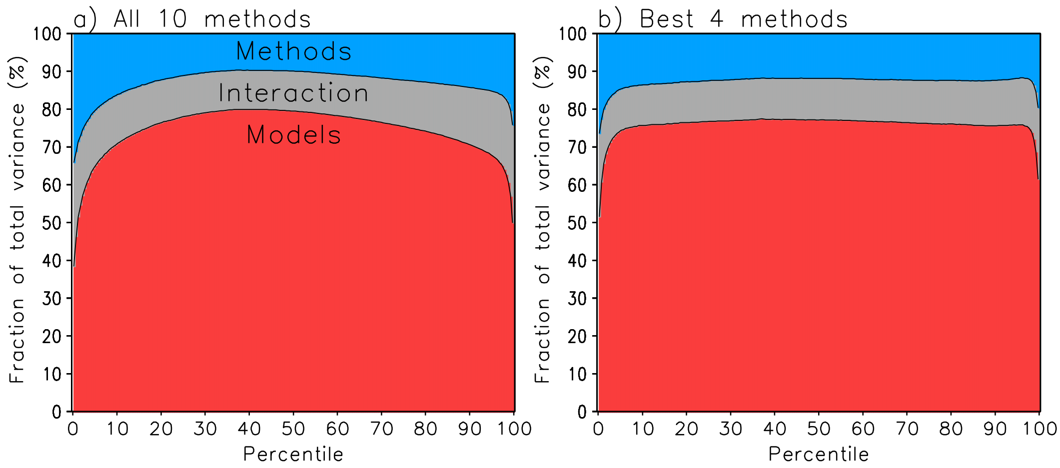

Fig. 13 *The relative contributions of model differences, model-method interaction and method differences to the variance of the projected temperatures in 2069-2098 as a function of the percentile of the distribution. The variances are averaged over the 12 months and the 139 stations. In (a), all 10 methods are included in the analysis, in (b) only methods 4, 5, 9 and 10*