Vol. 44: 179–194, 2010
doi: 10.3354/cr00916

CLIMATE RESEARCH
Clim Res

Published December 10

Contribution to CR Special 23 'Regional Climate Model evaluation and weighting'

Weight assignment in regional climate models

Jens Hesselbjerg Christensen[1,*], Erik Kjellström[2], Filippo Giorgi[3], Geert Lenderink[4], Markku Rummukainen[2,5]

[1]Danish Meteorological Institute, Lyngbyvej 100, 2100 Copenhagen, Denmark
[2]Swedish Meteorological and Hydrological Institute, 601 76 Norrköping, Sweden
[3]The Abdus Salam International Centre for Theoretical Physics, PO Box 586, 34100 Trieste, Italy
[4]Royal Netherlands Meteorological Institute, 3730 AE de Bilt, The Netherlands
[5]Lund University, Sölvegatan 12, 223 62 Lund, Sweden

ABSTRACT: An important new development within the European ENSEMBLES project has been to explore performance-based weighting of regional climate models (RCMs). Until now, although no weighting has been applied in multi-RCM analyses, one could claim that an assumption of 'equal weight' was implicitly adopted. At the same time, different RCMs generate different results, e.g. for various types of extremes, and these results need to be combined when using the full RCM ensemble. The process of constructing, assigning and combining metrics of model performance is not straightforward. Rather, there is a considerable degree of subjectivity both in the choice of metrics and on how these may be combined into weights. We explore the applicability of combining a set of 6 specifically designed RCM performance metrics to produce one aggregated model weight with the purpose of combining climate change information from the range of RCMs used within ENSEMBLES. These metrics capture aspects of model performance in reproducing large-scale circulation patterns, meso-scale signals, daily temperature and precipitation distributions and extremes, trends and the annual cycle. We examine different aggregation procedures that generate different inter-model spreads of weights. The use of model weights is sensitive to the aggregation procedure and shows different sensitivities to the selected metrics. Generally, however, we do not find compelling evidence of an improved description of mean climate states using performance-based weights in comparison to the use of equal weights. We suggest that model weighting adds another level of uncertainty to the generation of ensemble-based climate projections, which should be suitably explored, although our results indicate that this uncertainty remains relatively small for the weighting procedures examined.

KEY WORDS: RCM · Ensemble forecast · Climate projections

Resale or republication not permitted without written consent of the publisher

1. INTRODUCTION

Global and regional climate models (GCMs and RCMs, respectively) share common computational aspects, as well as the need to provide quantitative projections of climate change by means of a better characterization of relevant uncertainties. This characterization requires the use of possibly large ensembles of model simulations to explore the different sources of uncertainty (Giorgi et al. 2008) and to design and calibrate procedures for use in constructing probabilistic regional climate scenarios (Déqué et al. 2010, this Special). This process includes the development of techniques for the generation of probabilistic predictions by statistical processing of ensemble integrations. In particular, weighting individual climate model members of an ensemble based on model performance has been suggested as a way to reduce the unwanted uncertainty in climate model projections (Giorgi & Mearns 2002, 2003, Murphy et al. 2004, Tebaldi et al. 2005, Tebaldi & Knutti 2007, Knutti et al. 2010). The underlying assumption is that uncertainties can be reduced if the results from the 'better performing' models are given a greater weight in the ensemble when used to produce probabilistic projections.

*Email: jhc@dmi.dk

© Inter-Research 2010 · www.int-res.com

In the European FP5 project PRUDENCE (Christensen et al. 2007, Christensen & Christensen 2007), it was concluded that the RCM formulation plays an almost equal role in determining uncertainty, compared with that related to the boundary conditions provided by the driving GCM, at least for summer conditions when the model interior is more decoupled from the large-scale boundary conditions (Déqué et al. 2005, 2007). This suggests that weighting of RCM output is a natural step to explore, as pursued in the European FP6 project ENSEMBLES. This project extended the approach in PRUDENCE by generating a matrix of experiments that would better span the uncertainty range that is due to both GCM and RCM formulation (van der Linden & Mitchell 2009). One of the objectives of ENSEMBLES was to explore performance-based RCM weights that could to be used in the generation of regional climate change probability distributions.

In such a pursuit, it becomes central to choose suitable metrics of model performance that are independent of the driving GCM and yet trace important characteristics of regional climate. Clearly, there is no objective universal approach to the choice of metrics, as an unmanageable number of different metrics could be utilized to cover all the degrees of freedom in a climate model. A subjective selection of a limited set of metrics with *a priori* largely unknown interdependencies is unavoidable. This adds an element of uncertainty that needs to be explored in the development of regional-scale climate projections based on ensemble information.

Knutti et al. (2010) and Weigel et al. (2010) demonstrated that multi-model-based climate change information with a weighting concept comes with special caveats. Equally weighted multi-model averages consistently outperform single models (e.g. Knutti et al. 2010). However, specific knowledge of the individual models is also required, such as aspects of the relative contribution of joint model errors and model noise, in order to avoid biased weights (Weigel et al. 2010). This is related to the fact that the internal variability of an RCM may be large, in Europe particularly during summer, and any performance-based weighting could thus be misleading in terms of apparent added value in comparison to no (i.e. equal) weighting. Furthermore, any estimate of model skill necessarily builds on model performance under current or past climate conditions, and thus needs to be somehow extrapolated to possible future conditions when applied to projections. Indeed, the assumption of stationarity under a changing climate can be an issue (Christensen et al. 2008, Buser et al. 2009).

In this study, which results from work completed as part of the ENSEMBLES project, we analyze the effect of different procedures for model weighting based on multiple performance metrics in an ensemble of RCM simulations for the European region. The set of RCM performance metrics adopted here were selected following 2 general guidelines. (1) They should measure RCM performance in climatic aspects that constitute an 'added value' compared with the driving global models; this involves, for example, the representation of sub-GCM-scale climate features simulated by RCMs and the simulation of extremes. (2) Model performance in reproducing observed large-scale climate characteristics should also be included among the metrics, as this is the primary driver of regional climate.

This resulted in the selection of 6 metrics and the development of corresponding weights for each of the ENSEMBLES RCMs. These are described in detail in other contributions to this Special. They are based on multi-decadal simulations using ERA-40 reanalysis fields as boundary conditions. Here we aggregate the information contained in the individual metrics and weights by exploring different methods for compounding these 6 weights into a single weight for each of the ENSEMBLES RCMs. The rationale for deriving a single weight based on multiple performance metrics is that a fundamental requirement for increasing the reliability (and thus the weight) of a model is that this model should perform well in a range of different metrics, to minimize possible compensation effects by different systematic model biases. In addition, such overall model weights can be more easily used in the generation of joint regional climate change projections. There may well be other aspects of uncertainty that will eventually need consideration in this context, such as the limited size of the ensemble, incomplete validation data sets used, etc. All such caveats of course need to be kept in mind in possible applications of the methods presented. In addition, we stress that the unavoidable element of subjectivity discussed above is still present in our method, so that our work should, at this stage, be considered mostly as exploratory of the relevance of model weighting. This is particularly the case in view of the fact that this work presents the first published attempt to develop and implement metrics and weights specifically designed for application to RCM ensembles.

2. EXPERIMENTAL SET-UP, EVALUATION METRICS AND WEIGHTS

2.1. RCMs and simulations

RCM data from 15 simulations at 13 institutes were used (Table 1). These RCMs cover most of Europe with a horizontal grid spacing of approximately 25 km. The experiments were set up so that each model covered the same minimum domain (Fig. 1). Apart from this, the exact extent of the model domain was

up to each institute to decide. As a result, a few models were run for larger domains covering a greater portion of the Atlantic (Models 1, 9–11, 15) than the other models. A majority of the models use an identical rotated latitude–longitude grid whereas the others use different Lambert-conformal projections for their respective grids (see Table 1). One of the institutes, the Hadley Centre, ran 3 members from a perturbed physics ensemble (Collins et al. 2010) in which the physical parameterizations in their RCM, HadRM3H, were different. These differences in the physics, also applied in their global model, lead to different climate sensitivity in the different model versions. The 3 model versions include a reference (denoted as Q0), one with high climate sensitivity (i.e. larger global temperature response to greenhouse gas forcing) (Q16), and one with low climate sensitivity (Q3). Two institutes, C4I and SMHI (see Table 1), ran the Rossby Centre RCA3.0 model (Kjellström et al. 2005). Differences between the 2 RCA3.0 simulations include a larger domain size and more vertical levels in the C4I simulations, compared with the SMHI simulations. Two institutes, DMI and Met. No, ran different versions of the HIRHAM model, with DMI using a newer system with modifications in the formulation of advection and some of the physics routines, as well as an entire rewrite of the code in Fortran90 (see references listed in Table 1).

All models were run with lateral boundary conditions and sea-surface temperature (SST) taken from the European Centre for Medium-range Weather Forecasts reanalysis product ERA40 (Uppala et al. 2005), although a few of the participating groups did not use the full horizontal and vertical resolution offered by ERA40. Sensitivity experiments (with Model 4, data not shown) with different resolutions of ERA40 boundary conditions indicated that this has only a marginal effect on the overall model performance. The ERA40 period covers

Table 1. Regional climate models (RCMs) from which data have been analyzed. Forcing conditions are either set constant (C) or variable (V). In the case of variable forcing, changes are included either as explicit changes in CO_2 or other greenhouse gases (GHGs) and aerosol concentrations, or as changes in CO_2 equivalents (CO_{2eq}). Aerosols are treated explicitly in these RCMs only in the Met Office Hadley Centre HadRM3 model, which has an explicit formulation of the sulphur cycle (S). *z*: number of vertical levels; Rel. zone: sponge or relaxation zone used to drive the model

Model no.	Acronyms	Institute	Model (long × lat × *z*)	Domain size (points)	Rel. zone	GHG and aerosol forcing	Source
1[a]	C4I:	The Community Climate Change Consortium for Ireland, Met Eireann	RCA3.0	206 × 206 × 31	8	V (CO_{2eq})	Kjellström et al. (2005)
2	CHMI:	Czech Hydrometeorological Institute	ALADIN	183 × 205 × 31	8	V (CO_{2eq})	Farda et al. (2010)
3	CNRM:	Centre National de Recherches Meteorologique, Meteo France	RM4.5	229 × 229 × 31	8	V (GHG) + V(sulfate aerosols)	Radu et al. (2008)
4[a]	DMI:	Danish Meteorological Institute	HIRHAM5	194 × 210 × 19	10	C	Christensen et al. (2006)
5[a]	ETH:	Swiss Federal Institue of Technology	CLM	193 × 201 × 32	8	V (CO_{2eq})	Jaeger et al.(2008)
6	ICTP:	The Abdus Salam Intl. Centre for Theoretical Physics	RegCM3	190 × 206 × 18	12	V(GHG)	Pal et al. (2007)
7[a]	KNMI:	The Royal Netherlands Meteorological Institute	RACMO2	206 × 224 × 40	8	V (GHG) C (aerosols)	van Meijgaard et al. (2008)
8[a]	Met.No:	The Norwegian Meteorological Institute	HIRHAM	198 × 213 × 31	5	V (CO_2) C (aerosols) (2006)	Haugen & Haakenstad
9[a]	MetoHC:	UK Met Office, Hadley Centre for Climate Prediction and Research	HadRM3Q0	214 × 220 × 19	8	V (GHG, S)	Collins et al. (2010)
10[a]	MetoHC:	UK Met Office, Hadley Centre for Climate Prediction and Research	HadRM3Q3	214 × 220 × 19	8	V (GHG, S)	Collins et al. (2010)
11[a]	MetoHC:	UK Met Office, Hadley Centre for Climate Prediction and Research	HadRM3Q16	214 × 220 × 19	8	V (GHG, S)	Collins et al. (2010)
12[a]	MPI-M:	Max-Planck-Institute for Meteorology	REMO	193 × 217 × 27	8	V (GHG) C (aerosols)	Jacob (2001)
13	OURANOS:	Consortium on Regional Climatology and Adaptation to Climate Change	MRCC4.2.3	209 × 209 × 29	9	V (GHG, aerosols)	Music & Caya (2007)
14[a]	SMHI:	Swedish Meteorological and Hydrological Institute	RCA3.0	186 × 206 × 19	8	V (CO_{2eq})	Kjellström et al. (2005)
15	UCLM:	Universidad de Castilla La Mancha	PROMES	223 × 209 × 28	10	V (GHG) Aerosols not considered	Sanchez et al. (2004)

[a]Models that operate on the same rotated longitude/latitude grid in the common minimum domain (170 × 190 grid points)

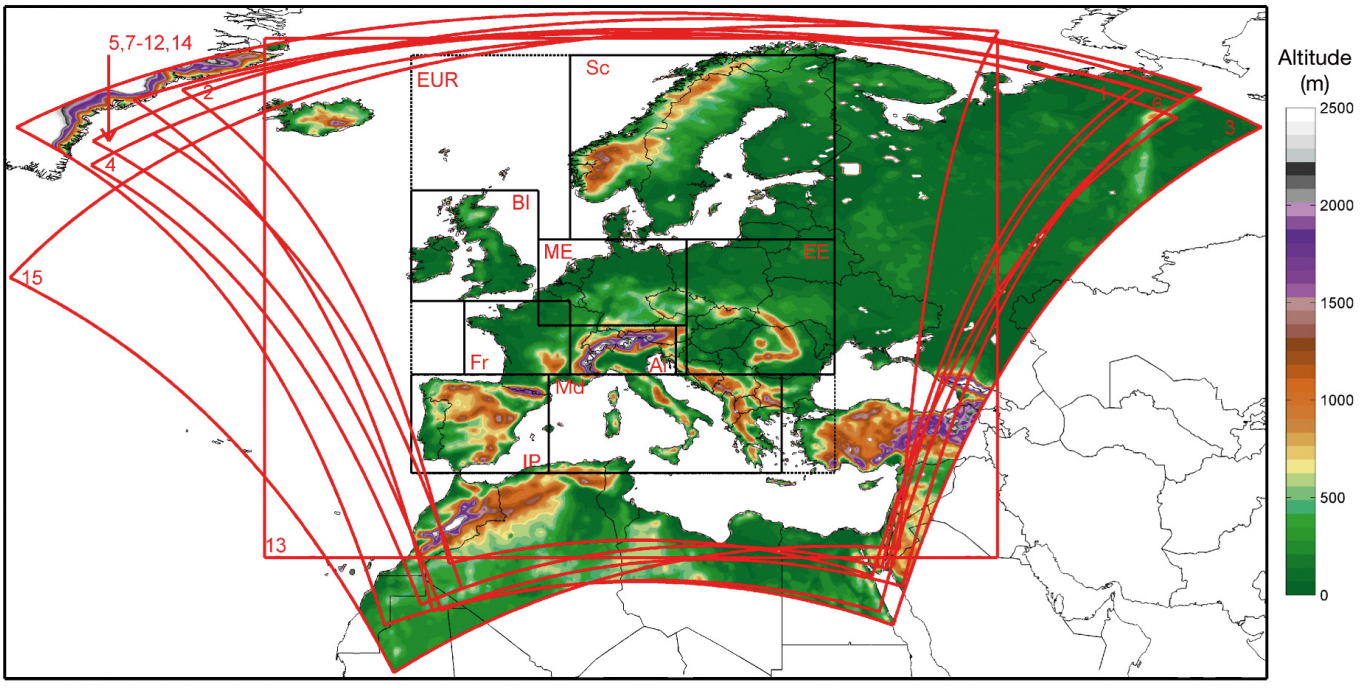

Fig. 1. The domain of the regional climate models (RCMs) running on the RCM domains using a rotated latitude longitude grid. The colours depict the altitude in one of the models (UCLM-PROMES). Eight subdomains are shown (BI: British Isles; IP: Iberian Peninsula; Fr: France; ME: mid-Europe; Sc: Scandinavia; Md: Mediterranean region; Al: Alps; EE: Eastern Europe), as well as the larger European domain (EUR), for which most metrics have been calculated. Numbers indicated in the corners of each domain identify the models (see Table 1)

1958–2002, a few model simulations used a slightly later initial time, but all included the 1960s. Therefore, here we base the model evaluation on data for the common period 1961–2000, unless otherwise noted. We also point out that, apart from differences in model formulations between different RCMs, physiographic characteristics—such as topography, land/sea and land/lake contrasts, vegetation, surface albedo, soil type and other fields related to such quantities—also vary across the models. In addition, the treatment of greenhouse gases (GHGs) and aerosols differs across the RCMs. Most, but not all, models prescribe increasing CO_2 concentrations following observations, but the treatment of other GHGs differs across the RCMs. Furthermore, HadRM3H is the only model that explicitly takes into account changing sulphate aerosol concentrations (Table 1).

2.2. Individual metrics used for RCM validation

A set of 6 model performance metrics was established in the procedure to generate RCM weights. As mentioned above, although not fully comprehensive, the proposed metrics cover a wide range of the so-called added value for dynamical downscaling; such as sub-GCM-scale and meso-scale information, fine-scale processes to better capture higher order statistics and extreme events (Rummukainen 2010). In addition, the metrics take into account some of the basic requirements for a model to be assessed as credible in terms of representing essential features of both the observed climate and possible future climate conditions. These metrics include the annual cycle of precipitation and temperature, large-scale agreement with the driving model to capture variations in the large-scale atmospheric flow, and observed temperature trends. Table 2 summarizes the data sets used for deriving the final individual weights f_1, f_2, …, f_6. In the following we give a short description of the 6 metrics. For a more comprehensive description, see references in Table 2.

2.2.1. f_1: large-scale circulation based on a weather regime classification

This metric tests whether the RCMs are able to reproduce observed weather regimes (Sanchez-Gomez et al. 2008). Large-scale circulation is an important constraint of regional climate and its variability (van Ulden et al. 2007) and a satisfactory representation of large-scale regimes is a prerequisite for acceptable RCM performance. Sanchez-Gomez et al. (2008) evaluated several related metrics: (1) mean behaviour

Table 2. Metrics used to infer the final weights. Z_{500hPa}: geopotential height; P: precipitation; T: mean temperature; EUR: European domain. See Fig. 1 for areas considered

Metric	Variables	Period	Reference data set	Data type	Seasons	Area	Source
f_1	Z_{500hPa}	1961–2000	ERA40	Daily	DJFM, JJAS	Minimum domain	Sanchez-Gomez et al. (2008)
f_2	P, T	1961–2000	CRU TS1.2	Monthly	DJF, MAM, JJA, SON	EUR	Coppola et al. (2010)
f_3	P, T_{min}, T_{max}	1961–1990	EOBS2.0	Daily	DJF, MAM, JJA, SON	EUR	Kjellström et al. (2010)
f_4	P, T_{min}, T_{max}	1971–2000	EOBS2.0	Daily	DJF, MAM, JJA, SON	EUR	Lenderink (2010), Buonomo (unpubl.)
f_5	T	1961–2000	EOBS2.0	Monthly	DJF, MAM, JJA, SON, ANN	Average of 8 subdomains	Lorenz & Jacob (2010)
f_6	P, T	1961–2000	EOBS2.0	Monthly		EUR	Halenka et al. (unpubl.)

for each season in terms of mean frequency of occurrence of weather regimes, mean persistence and spatial structure of the composite; (2) interannual variability of the frequency of occurrence of weather regimes; and (3) daily chronology of weather regimes. To calculate the first metric (f_1), the work of Sanchez-Gomez et al. (2008) was repeated for 13 of the available RCM simulations at 25 km. Models 9–11 are run with the same model dynamics, the same lateral boundary forcing technique and, for the most part, the same physics. They are therefore considered as equivalent in terms of large-scale dynamical regimes and thus only 1 model configuration analyzed was considered representative for all 3 model versions.

2.2.2. f_2: meso-scale signal based on seasonal mean temperature and precipitation

RCMs are expected to provide added information on the meso-scale in comparison to coarse resolution GCMs. As a measure of the meso-scale signal, Coppola et al. (2010) use a spatial filter that removes the large-scale component (>200–250 km) in both observations and RCMs. For temperature and precipitation, they calculate separately the spatial correlation coefficients between observations and RCMs and the interannual variability for both temperature and precipitation. They also evaluate the correlation between temperature and precipitation at the meso-scale. The resulting 5 sub-weights are multiplied to obtain the weight function (f_2).

2.2.3. f_3: probability density distributions of daily and monthly temperature and precipitation

The statistical properties of daily and monthly temperature and precipitation are important for many application purposes, but they also provide a summary measure of the overall model performance in capturing means and higher order moments. The chosen metric (f_3) considers empirical probability density functions (PDFs) for precipitation and maximum and minimum temperature (Kjellström et al. 2010). Whereas daily data are seen as largely representing regional information and, hence, form a relevant RCM metric, the monthly precipitation field is much more strongly controlled by the driving GCM. Therefore, the metric developed for the monthly precipitation statistics is not used directly, but is given a lower weight in the final metric. This is achieved by taking the square root of the resulting sub-weights derived from the monthly means when combining the sub-weights into the final metric.

2.2.4. f_4: extremes in terms of re-occurrence periods for temperature and precipitation

Consideration of the far tails of distributions (99th, 99.9th and 99.99th percentiles) provides additional information to the full PDF information as captured in f_3. We evaluate such extremes in 2 ways: by considering daily precipitation extremes taken directly from the empirically deduced PDFs (Lenderink 2010) and by using generalized extreme-value theory (Buonomo unpubl.) for daily precipitation and daily maximum/minimum temperatures. Weighting factors from each of the extremes are combined after being individually averaged over the seasons into annual numbers and then multiplied by each other to obtain the f_4 weight function.

2.2.5. f_5: long-term trends in temperature

RCMs should be capable of capturing forced trends when these are present in the driving boundary condi-

tions. However, the omission or misrepresentation in RCMs of possible local to regional forcings, such as from aerosols or land-use changes, may inhibit this behaviour. Because changes in local/regional forcings are potentially important for the future regional climate, this component is assessed by comparing model trends with corresponding observations (Lorenz & Jacob 2010). Linear trends are calculated for Europe and evaluated to yield f_5. Only temperature is considered, because moisture in ERA40 is strongly influenced by changes in the observational system during the reanalysis period. This metric considers skill scores for how well the observed trends are matched by the RCMs in all seasons as well as for annual averages. To calculate f_5, ½ of the annual skill score and ⅛ of each seasonal skill score are added.

2.2.6. f_6: annual cycle in temperature and precipitation

The annual cycle of temperature and precipitation is a basic measure of model performance. The model's ability to capture the annual cycle provides insight into the model's response to altered radiative forcing, such as that arising from increased levels of GHGs. Halenka et al. (unpubl.) investigate this and obtain f_6.

2.3. Combining individual metrics into one weight per RCM

Each of the 6 metrics chosen above describes essential features of the European climate. To receive a high weight, a model needs to perform well in all metrics considered, so as to minimize possible counterbalancing effects of different systematic biases. This consideration needs to be accounted for in combining the different performance metrics into an overall weight. The simplest possible combination of the relative scores or weights from the individual metrics (normalised to have a sum equal to 1) would be to either add or multiply them. The former case can be seen as a form of averaging the importance of each metric, whereas the latter implicitly assumes that the weights are essentially independent of each other and that a model needs to perform well for all metrics to obtain a high score.

The multiplicative approach was chosen as our baseline, following Giorgi & Mearns (2002), but we also consider alternative approaches as described in the next section. As part of the ENSEMBLES project, these metrics were defined for different seasons and both for European sub-regions (indicated in Fig. 1; the same as those defined within the PRUDENCE project; see e.g. Christensen & Christensen 2007) and for Europe as a whole. Information on more traditional model evaluation is also available on the ENSEMBLES web site (http://ensembles-eu.metoffice.com/) and in van der Linden & Mitchell (2009). Here we focus on the whole European scale and on the annually averaged quantities. We stress that, by making this choice, many detailed aspects of model skill are washed out and the combined score may not reflect the specific model performance for a particular region or season.

As mentioned above, the performance of a particular RCM as assessed by the metrics above can be combined into a single weight for each RCM by, for example, a multiplication of the weights f_1, f_2, …, f_6:

$$W_{\mathrm{PROD}} = \prod_{i=1}^{6} f_i^{n_i} \quad (1)$$

where all the individual weights are first normalized to yield a value between 0 and 1 (see accompanying papers in this issue, and references in Table 2) before entering Eq. (1). The final weight (W_{PROD}) for each model is also normalized across the models in order to facilitate application to the model ensemble. The simple multiplication can be refined by allowing for the exponent n_i in Eq. (1) to be chosen as any positive number. Assuming $n_i = 1$ implies weighting the various metrics equally, whereas choosing any positive value different from 1 shifts the emphasis across the individual metrics (a value of 0 would imply equal weighting of the RCMs). This latter approach would be warranted if some metrics were considered to be more fundamental than others, for example when applying the method to a specific impact sector or if some of the metrics were not independent from each other. Other methods could be introduced based on more sophisticated approaches, for example paying attention to how the different metrics are correlated or formulated.

Given the subjective nature of the metrics aggregation into weights, in order to explore the sensitivity to different aggregation approaches we also introduce 2 additional ways to combine the 6 individual metrics. The first considers a varying value of the exponents n_i. This second total weight (W_{REDU}) is a variant of the baseline in that the spread for all the subweights is 'normalized' such that the ratio between the highest and lowest assigned individual weight is 1.2. This choice implies a maximum overall ratio across model weights of 3 and is simply chosen to illustrate an intermediate case between equal weights ($n = 0$) and the weights according to Eq. (1) with $n = 1$. Formally, this is obtained by using n_i different from 1 in Eq. (1) (see the explanation in Table 3), defining the re-normalized sub-weights $\tilde{f}_{i,j}$ as:

$$\tilde{f}_{i,j} = f_{i,j}^{n_i} \quad (2)$$

for each model j, and each sub-weight i, and choosing n_i such that:

$$\frac{\max_j\left(\tilde{f}_{i,j}\right)}{\min_j\left(\tilde{f}_{i,j}\right)} = 1.2 \quad (3)$$

for each sub-weight i.

Finally, we examine a third way of obtaining the total weight (W_{RANK}) by means of first ranking all models according to their order of performance in terms of each of the metrics. We then sum these 6 ranks and transform this rank sum into a model weight by dividing the sum of the ranks by the rank sum of each model and then normalize it so that the total sum of the weights is equal to 1.

3. RESULTS

3.1. Final weights

The final weight for each model and each of the 3 approaches (W_{PROD}, W_{REDU} and W_{RANK}) is shown in Fig. 2 and Table 3 along with the corresponding 6 individual metrics. The total weight picks out a 'winner' in the sense that one model (KNMI-RACMO2) has a significantly higher score than the others. The weight obtained by KNMI-RACMO2 is almost 17 times as high as that for the model receiving the lowest score when using our baseline approach (W_{PROD}). A closer inspection of the individual metrics reveals that some are more discriminating than others (Table 3). In particular, f_2 and f_4 contribute most substantially to the spread. This is not surprising, as these metrics are both calculated as products of several different sub-metrics. By contrast, most of the other metrics are constructed by averaging only one metric over the different seasons.

How the 3 alternative aggregation methods modify the 'classical' case of no weighting is illustrated in

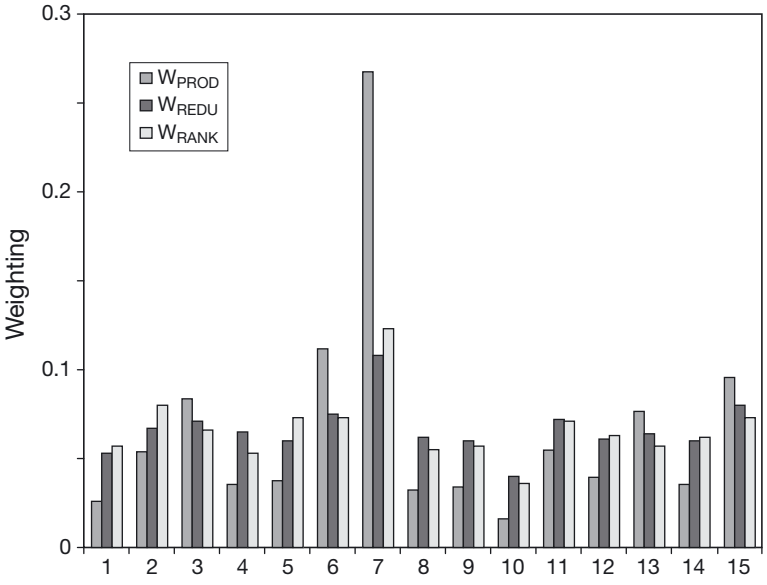

Fig. 2. W_{PROD}, W_{REDU} and W_{RANK} for each of the regional climate models (RCMs). See Table 1 for model numbers and Table 3 for definition of weights

Table 3. Revised individual regional climate model (RCM) metrics and their 3 different combinations into RCM weights. W_{PROD} is calculated based on f_1 to f_6 according to Eq. (1) with n_1 to $n_6 = 1$. W_{REDU} is calculated in a similar way, but with n_1 to n_6 calculated to maintain the ratio between the highest and lowest assigned individual weight equal to 1.2 ($f_1^{\wedge 0.69}$, $f_2^{\wedge 0.145}$, $f_3^{\wedge 3.1}$, $f_4^{\wedge 0.125}$, $f_5^{\wedge 0.905}$, $f_6^{\wedge 1.33}$). W_{RANK} is calculated according to the rank of the RCMs. The numbers are rounded to retain 3 decimals. The RCM acquiring the **highest** and the lowest respective value for each metric and weight is indicated

Model	f_1	f_2	f_3	f_4	f_5	f_6	W_{prod}	W_{redu}	W_{rank}
1	0.058	0.050	0.067	0.044	0.066	0.069	0.026	0.058	0.057
2	0.071	0.058	0.067	0.070	0.060	0.069	0.054	0.066	0.08
3	0.069	0.059	0.067	0.113	0.066	0.061	0.084	0.064	0.066
4	0.068	0.039	0.066	0.062	0.070	0.068	0.035	0.066	0.053
5	**0.075**	0.073	0.067	0.036	0.059	**0.069**	0.038	0.067	0.073
6	0.073	0.112	0.065	0.066	0.069	0.067	0.112	0.075	0.073
7	0.070	**0.137**	**0.069**	**0.132**	0.066	0.068	**0.268**	**0.094**	**0.123**
8	0.070	0.041	0.067	0.057	0.065	0.067	0.032	0.064	0.055
9	0.061	0.048	0.067	0.054	0.071	0.066	0.034	0.063	0.057
10	0.061	0.049	0.066	0.030	0.064	0.062	0.016	0.047	0.036
11	0.061	0.051	0.067	0.080	**0.073**	0.066	0.055	0.069	0.071
12	0.068	0.072	0.066	0.038	0.069	0.069	0.039	0.068	0.063
13	0.072	0.089	0.065	0.063	0.065	0.066	0.077	0.065	0.057
14	0.057	0.053	0.067	0.054	0.067	0.069	0.035	0.063	0.062
15	0.067	0.068	0.067	0.099	0.070	0.065	0.096	0.074	0.073

Fig. 2 (with no weighting or equal weighting given by a value of 0.066 for each model). The baseline case (W_{PROD}) exhibits the largest differentiating effect. The other methods result in a reduced variation across the RCM ensemble, but retain the same overall performance ranking of the RCMs. The latter result is not entirely trivial because the spread in W_{PROD} is determined primarily by 2 sub-weights (f_2 and f_4).

These weights are only meaningful vis-à-vis the entire set of RCMs used in their construction. (1) The weights for the models are relative to one another. If instead some subset of the RCMs were to be used, the values for the weights would need to be recalibrated with respect to each other. (2) The evident sensitivity of the combined weights to specific assumptions on certain metrics emphasizes the subjective nature of the weighting procedure and the exploratory nature of this work. Due consideration of the underlying assumptions concerning the weights, as well as the overall subjective component of the choice of metrics and aggregation process, is necessary. (3) Even though weights based on extremes seem to differentiate models, this is conditional to the underlying observational data used to calculate the metrics. For example, the E-OBS gridded daily data set used for deriving f_4 underestimates extremes (Hofstra et al. 2009). This could mean that a good match between a particular model and the data does not necessarily identify the best model. The 'renormalization' procedure picks out weights f_2 and f_4 to be the most discriminating ones, as the values of n_i are close to 0.1 in both cases. As already mentioned, the proposed normalization and ranking procedures reduce the sensitivity to any individual metric, but still retain the overall separation into higher and lower weights. The actual choice of 1.2, viz. normalization, in the example shown here is an ad hoc choice for illustrative purposes.

3.2. Calculating a weighted ensemble mean

Here we illustrate how the weights can be used to calculate weighted ensemble averages for seasonal mean temperature and precipitation using W_{PROD} as defined above. Biases in the ensemble means are calculated with respect to the E-OBS2.0 data set (Haylock et al. 2008), both for weighted means and for the corresponding un-weighted means. By comparing the biases we investigate whether the weighting procedure improves the ensemble mean.

The differences between the 2 ensemble means are relatively small for all seasons, both for temperature (Figs. 3 & 4) and precipitation (Figs. 5 & 6). The weighted mean does not always outperform the un-weighted mean. In fact, sometimes the weighting leads to a lower quality ensemble mean. This is because the weights are not based on performance metrics related to climate means. A summary over the entire European area is given in Table 4. The differences in seasonal mean precipitation are quite small, whereas temperatures are improved for summer (JJA). However, weighting also leads to overall worse agreement for temperature during winter (DJF) both in terms of mean absolute error and fractional area improved by applying the weighting.

The weights in Table 3 are for the whole Europe domain and mostly for annual averages. The RCM performance against observations, however, varies, for example, between winter and summer. This is illustrated for seasonal mean temperature and precipitation in Figs. 3–6 by showing biases in the ERA40-forced RCM simulations with respect to the E-OBS gridded data for the period 1961–1990. Figs. 3 & 5 show that the 'best' performing model (KNMI-RACMO2) has a warm and dry bias in Eastern Europe in summer. As this model is given a high total weight for the all-Europe metrics, it contributes to the deterioration of the weighted ensemble mean in this region compared with other RCMs (e.g. SMHI-RCA3.0). For winter, Fig. 4 reveals an example of the opposite, in which Meto-HC-HadRM3Q3 performs relatively well over Western Europe compared with many other RCMs. In this case, the downgrading of this model owing to its low weight contributes to the poorer agreement of the weighted ensemble mean compared to the un-weighted ensemble mean in some areas (e.g. parts of France).

Fig. 4 shows a large cold bias in OURANOS-MRCC4.2.3. However, the final weight for this model (Table 3) is among the better ones, ranking as number 4. This shows that the weighting system does not strongly penalize this poor performance in wintertime

Table 4. Weighted / unweighted means over all land grid points in the European domain. MAE: mean absolute error; RMSE: root mean square error; areal fraction: area where the weighting leads to smaller MAE; **bold:** measures that improved when weighting was applied

Variable	Period	MAE	RMSE	Areal fraction
Precipitation	JJA	0.292 / 0.286	0.565 / 0.543	**0.55**
	DJF	**0.372** / 0.377	**0.455** / 0.465	**0.53**
Temperature at 2 m	JJA	**0.740** / 0.824	**0.928** / 1.018	**0.74**
	DJF	1.049 / 0.985	1.452 / 1.407	0.34
Total cloud fraction	JJA	**10.60** / 11.10	**12.87** / 13.33	**0.63**
	DJF	7.34 / 6.08	7.97 / 7.70	0.45

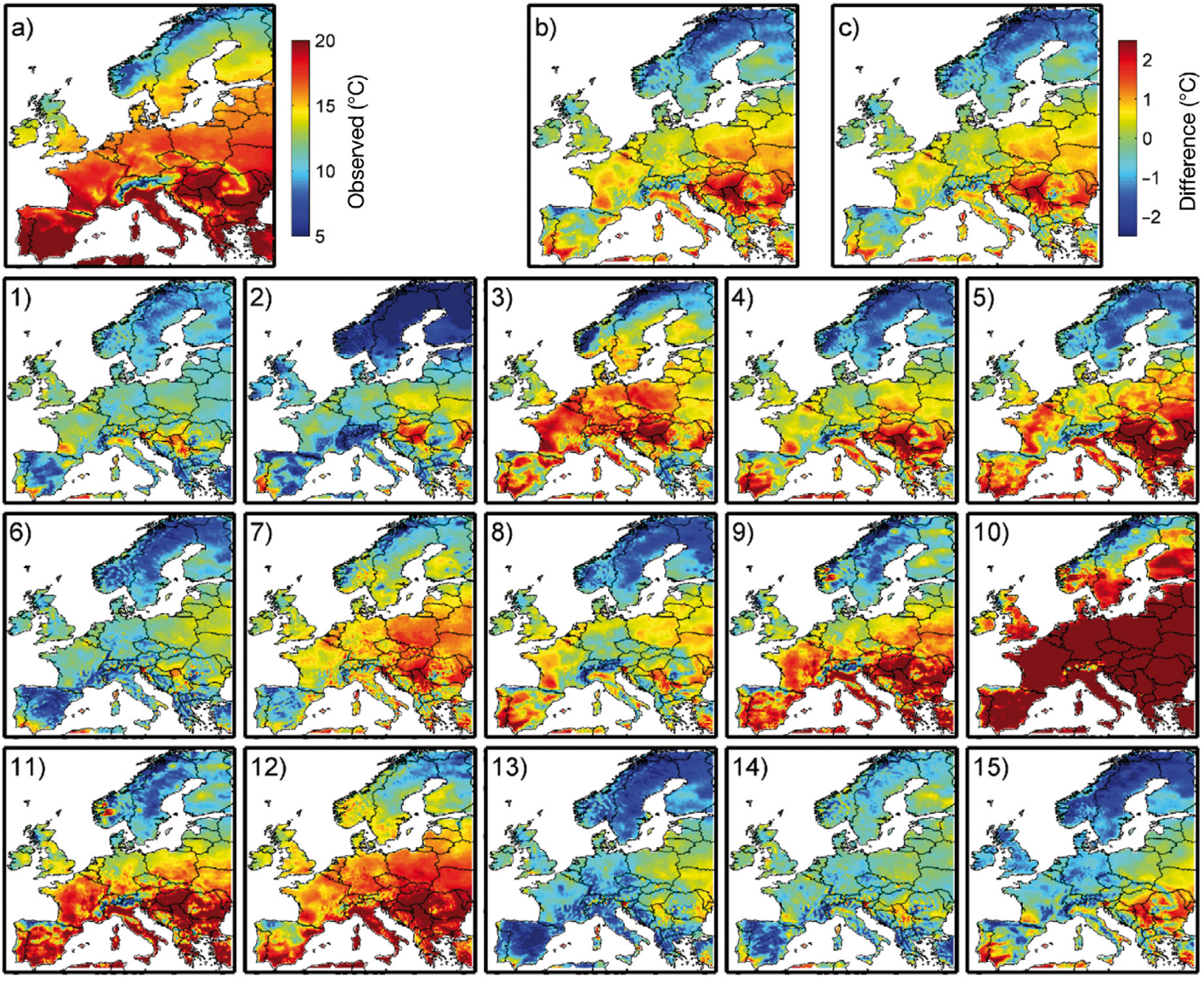

Fig. 3. Summer (JJA) temperature at 2 m (T_{2m}; °C). (a) E-OBS; (b) difference between unweighted ensemble mean and E-OBS; and (c) difference between weighted ensemble mean and E-OBS. Panels 1–15 show the difference between model and E-OBS for each individual regional climate model (see Table 1 for model numbers). The left-most color scale applies to panel (a) only; the right-most color scale applies to all other panels

temperature compared with other models. In terms of the individual weights, it can be seen that OURANOS-MRCC4.2.3 has the lowest score for f_3, which holds information about the entire probability distribution of temperatures. However, as f_3 does not contribute strongly to the spread across RCMs in the final weights (W_{PROD}), OURANOS-MRCC4.2.3 does not show a low final weight compared to the other RCMs. The ranking procedure shows that all models but 2 have a poor ranking, i.e. they are among the 3 worst models for at least 1 of the 6 metrics considered. Vice versa, all models rank among the top 3 for at least 1 of the 6 metrics.

4. DISCUSSION AND CONCLUSIONS

We applied performance indices and weighting schemes to a large ensemble of RCM simulations for present-day climate using realistic boundary conditions from the ERA40 reanalysis completed as part of the ENSEMBLES project. Details about these separate indices can be found in the accompanying papers in this issue. Here, we concentrate on ways of combining the information from the 6 performance indices into a weighting scheme.

The different weighting schemes yield a wide range of inter-model spread of weights, although the general

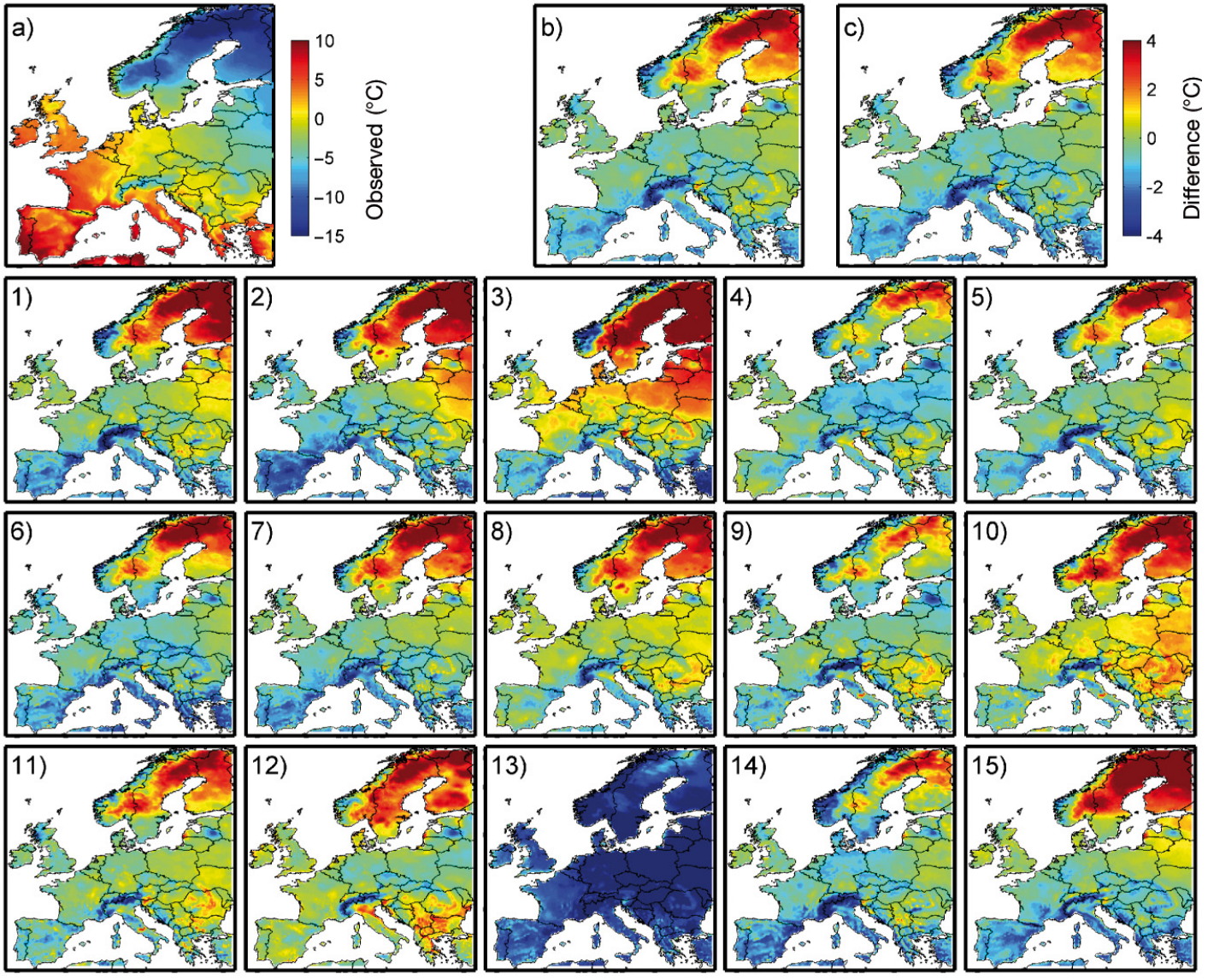

Fig. 4. Winter (DJF) temperature at 2 m (T_{2m}; °C). (a) E-OBS; (b) difference between unweighted ensemble mean and E-OBS; and (c) difference between weighted ensemble mean and E-OBS. Panels 1–15 show the difference between model and E-OBS for each individual regional climate model (see Table 1 for model numbers). The left-most color scale applies to panel (a) only; the right-most color scale applies to all other panels

ranking of model performance is maintained for all schemes. Our baseline weighting, in which the 6 individual indices are multiplied in order to give a stringent performance test, provide the largest inter-model spread, with one model emerging as best performer. In this case the weighting scheme has a significant effect on the ensemble performance.

However, the different weighting procedures do not appear to provide a strong and consistent superiority in simulating ensemble means when compared with other and simpler methods to combine multiple model information (such as simple unweighted averaging). One reason for this is that mean biases do not enter the set of performance indices utilized. This finding generally confirms results from previous work (e.g. Wilby & Harris 2006, Fowler & Ekström 2009, Knutti et al. 2010).

The results furthermore illustrate that not all aspects of model quality are captured by the present weighting, which is based on circulation, temperature and precipitation. For some applications it might be relevant to focus more on other climate aspects, such as snow, wind or soil moisture deficit, which may be more directly tied to the specific problem at hand. Indeed, from a practical point of view, metrics that are more tied to the needs of some specific impact sectors could be selected. To illustrate this, we have made an additional analysis by comparing

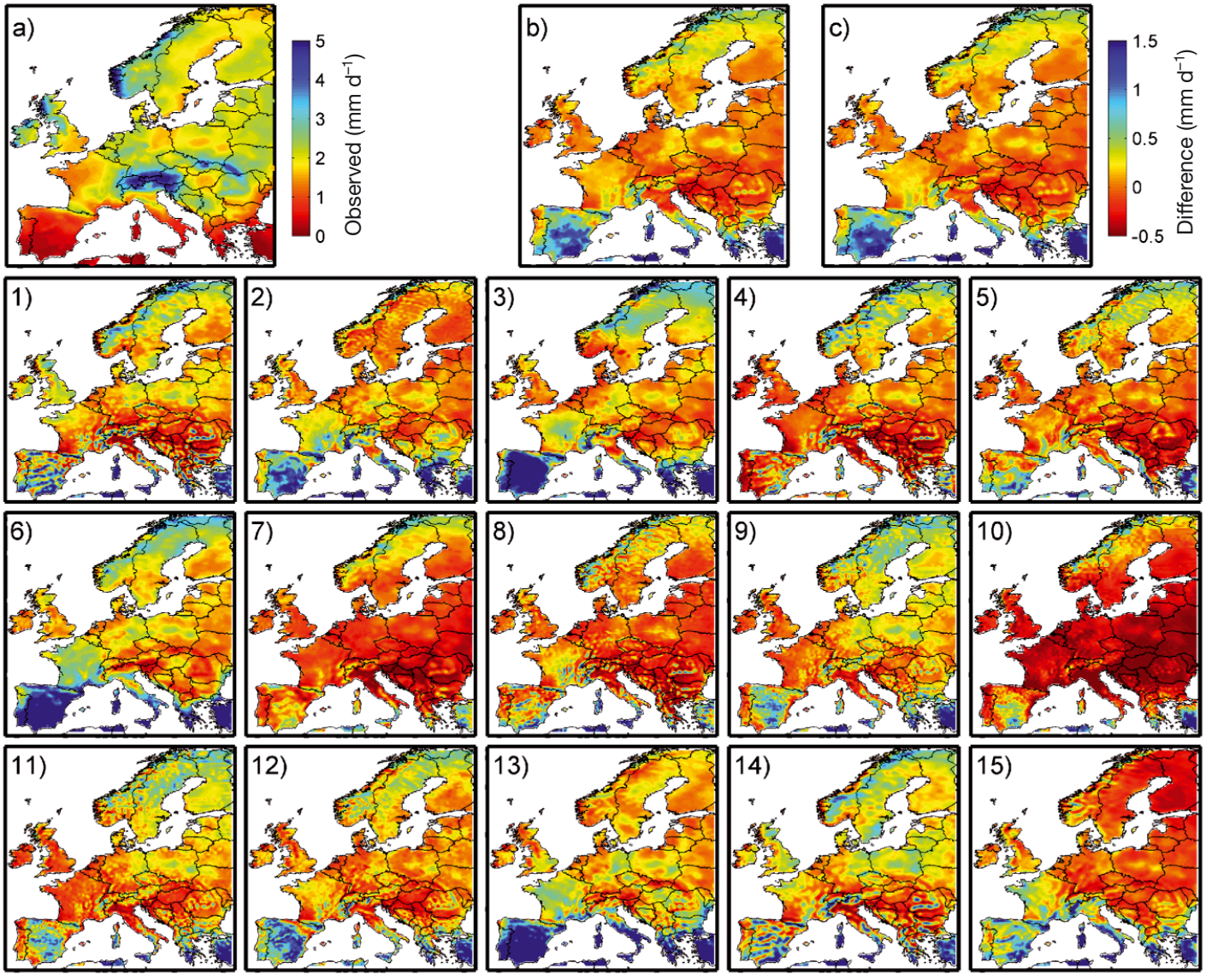

Fig. 5. Summer (JJA) precipitation (mm d^{-1}). (a) E-OBS; (b) difference between unweighted ensemble mean and E-OBS; and (c) difference between weighted ensemble mean and E-OBS. Panels 1–15 show the difference between model and E-OBS for each individual regional climate model (see Table 1 for model numbers). The left-most color scale applies to panel (a) only; the right-most color scale applies to all other panels

total seasonal mean cloud fraction from the RCMs with the ISCCP-D2 data set (Rossow et al. 1996). Note that this comparison utilizes a data set independent from those used to construct the weights. Data for the period 1983–2000 were used for this comparison and the results show a large spread among the RCMs (Figs. 7 & 8). A common feature in many, but not all, RCMs is that they appear to overestimate the north–south gradient in cloud cover, with excessive cloudiness in Scandinavia and too few clouds in southeast Europe. Applying the weights to calculate a weighted ensemble mean leads to a small improvement in summer, but only a marginal difference or no improvement in winter (Table 4).

An intrinsic limitation of the weighting procedure lies in the quality of the observational data used to calculate the weights. In the case of the E-OBS data set, a number of problems have been identified and corrected in the second version used here. It is nevertheless possible that important errors remain. An indication of such problems can be given by studying the ensemble mean biases in seasonal mean temperature for winter (Fig. 4). In Latvia there is a large positive bias along the Baltic Sea coast and a relatively strong negative bias in the eastern part of the country. These biases are local in nature and much larger than those for the rest of Europe. This may be an indication of

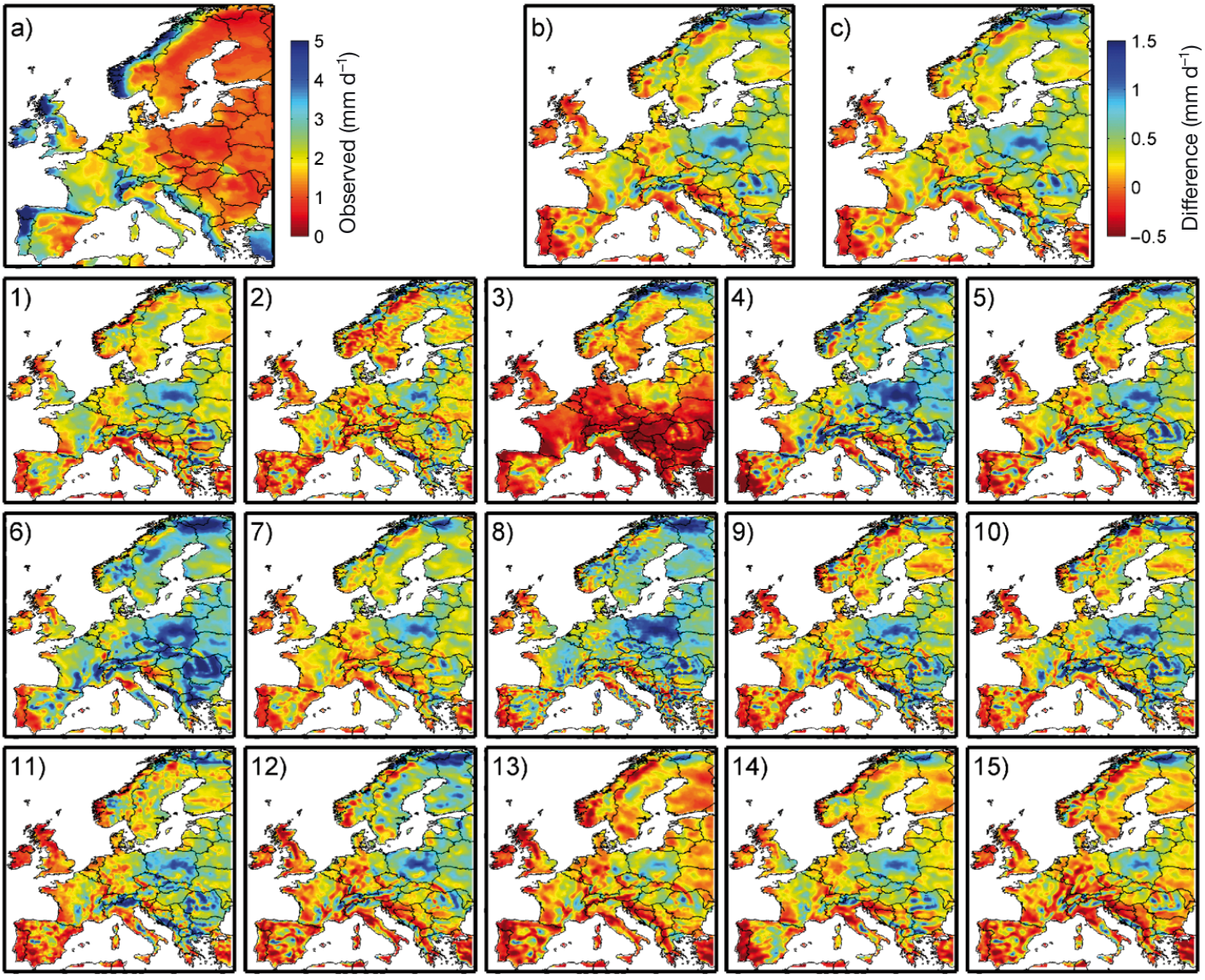

Fig. 6. Winter (DJF) precipitation (mm d^{-1}). (a) E-OBS; (b) difference between unweighted ensemble mean and E-OBS; and (c) difference between weighted ensemble mean and E-OBS. Panels 1–15 show the difference between model and E-OBS for each individual regional climate model (see Table 1 for model numbers). The left-most color scale applies to panel (a) only; the right-most color scale applies to all other panels

problems with the observational data set, as there is no obvious *a priori* reason why all RCMs would have localized problems in this low-altitude area with no complex terrain. Another example of when an ensemble of RCMs may aid in identifying suspect values in observational data sets can be seen for wintertime precipitation in large parts of Poland and northern Finland (Fig. 6). In both areas, the observations have much lower precipitation than found in all simulations. Although E-OBS is accompanied by a measure of interpolation uncertainty for each day and each grid box, it is not obvious how this uncertainty measure translates into the metrics considered. Yet it is clear that observational data quality issues should be included in future analysis. This underlines the overall difficulty in assessing the quality of a model when observations are still likely to be affected by significant uncertainties.

Here we have explored different methodologies for producing model weights based on aggregated information of different metrics of model performance. The rationale behind the choice of a single aggregated weight is that a model should perform well in all metrics so as to minimize the possible effects of counterbalancing systematic errors. Two caveats should, however, be emphasized. (1) In our baseline approach we produced our overall model weight from the product of

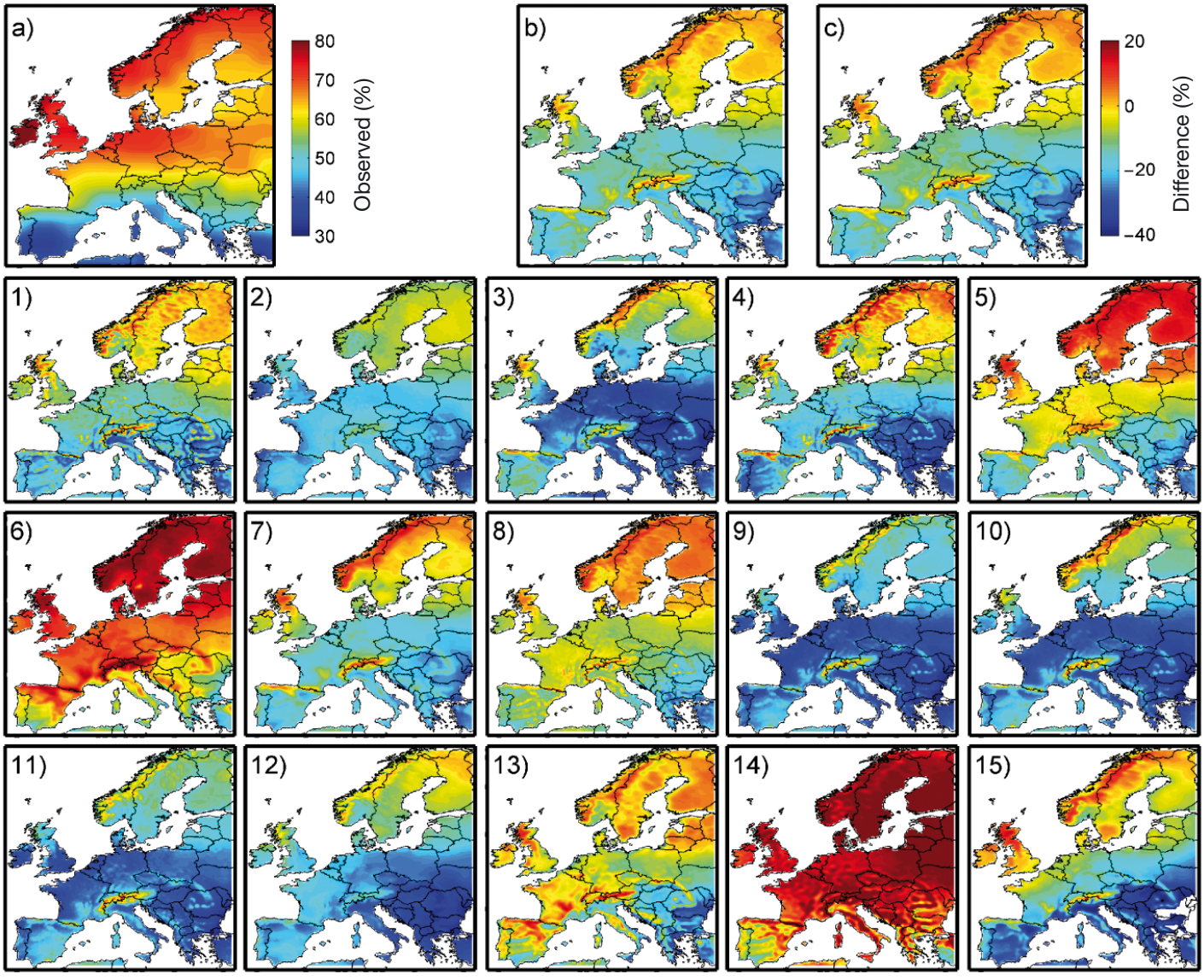

Fig. 7. Summer (JJA) cloud fraction (%). (a) ISCCP-D2; (b) difference between unweighted ensemble mean and ISCCP-D2; and (c) difference between weighted ensemble mean and ISCCP-D2. Panels 1–15 show the difference between model and ISCCP-D2 for each individual regional climate model (see Table 1 for model numbers). The left-most color scale applies to panel (a) only; the right-most color scale applies to all other panels

the weights associated with the individual performance metrics. This implies a very stringent test, as a good model is expected to have relatively high weights in all metrics. However, each weight can be considered differently depending on the specific application, as allowed by our general weighting framework. In fact, as shown in Table 3, some weights exhibit a much higher inter-model spread than others. This may be associated with the way the weight is produced, can profoundly affect the overall weight of the model. (2) We stress in some of our approaches that our weights are normalized using the full ensemble, i.e. they are relative in that they measure not the absolute performance of a model but the relative performance compared with the other models in the ensemble. The specific weights generated are thus likely sensitive to the exact mix of RCMs employed. Further work to address this aspect could involve resampling subsets of the available RCMs.

The choice of the metrics and how the individual weights are combined has an obvious and unavoidable subjective component. Given the extremely large number of degrees of freedom in a climate model, it is impossible to devise an all-encompassing objective weighting. We tried to use a wide range of metrics important for estimating the added value and the per-

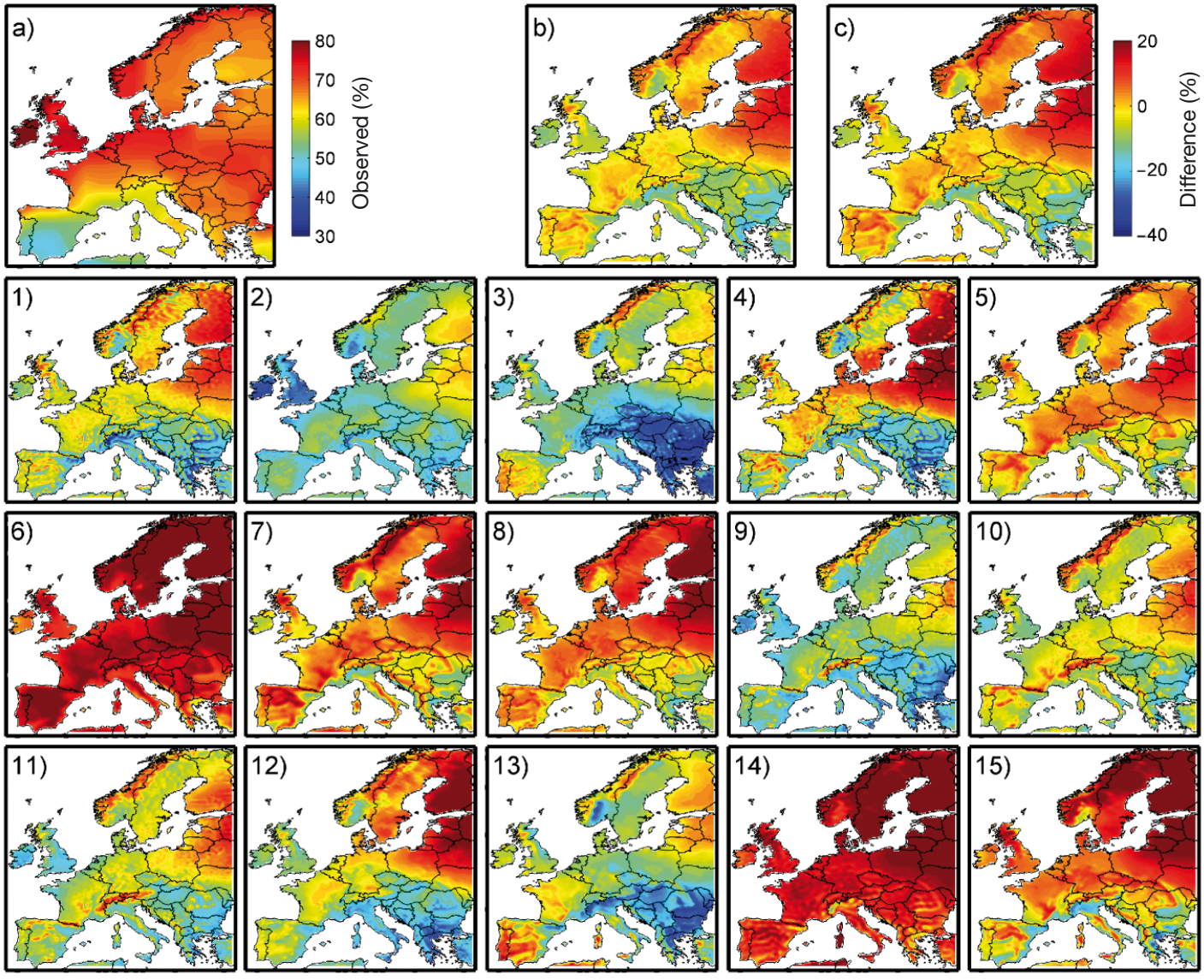

Fig. 8. Winter (DJF) cloud fraction (%). (a) ISCCP-D2; (b) difference between unweighted ensemble mean and ISCCP-D2; and (c) difference between weighted ensemble mean and ISCCP-D2. Panels 1–15 show the the difference between model and ISCCP-D2 for each individual regional climate model (see Table 1 for model numbers). The left-most color scale applies to panel (a) only; the right-most color scale applies to all other panels

formance of RCMs and combined them in stringent ways. The inevitable subjectivity of this approach, however, makes it necessary to evaluate the sensitivity of the overall weights to the criteria used to derive them as well as their interdependencies. Given the strong interlinkages between atmospheric flow, temperature and precipitation, the chosen metrics may in fact be correlated. A preliminary result from a simple test of the dependencies between the different metrics suggests that, particularly, metrics f_1, f_3 and f_6 are correlated. However, as these metrics do not discriminate strongly between the models, this may not be a very robust result.

Because of the subjective component of the weighting approach and the many uncertainties associated with it, we suggest that the weighting itself is a source of uncertainty in the generation of climate change scenarios and that this uncertainty is suitably explored by considering multiple metrics and aggregation procedures. Ensemble model weighting is still a highly controversial issue and, as our results confirm, the value of weighting has not been clearly demonstrated. Our work should thus be considered exploratory, in particular because it provides the first attempt to apply targeted model weighting to a large ensemble of RCMs.

Finally, when RCMs are forced with boundary conditions from GCMs, the systematic biases in the latter affect the results and should be considered. When using specific reanalyses as boundary conditions, this is less troublesome. When forcing is acquired from GCMs, bias patterns (not shown) differ, the implication being a wider spread across the ensemble of RCMs (e.g. Jacob et al. 2007).

In conclusion, we still know rather little about how to construct a credible and robust weighting procedure for multi-model regional climate change projections. The use of intercomparative quantities as performance indicators is one possible alternative.

Acknowledgements. This work was carried out under a contract with the European Union through the ENSEMBLES project (contract no. GOCE-CT-2003-505539). We are grateful to the GCM and RCM modellers who have worked hard to produce the valuable model results, which can now be accessed from the ENSEMBLES regional climate model data base at http://ensembles-rt3.dmi.dk. We are particularly in debt to Dr. Ole Bøssing Christensen (DMI) for maintaining the archive and for his support in accessing the data, to Dr. Fredrik Boberg (DMI) for making the figures and to Peter Thejll (DMI) for enlightening us on possible dependencies between the weights.

LITERATURE CITED

Buser CM, Künch HR, Lüthi D, Wild M, Schär C (2009) Bayesian multi-model projections of climate: bias assumptions and interannual variability. Clim Dyn 33:849–868

Christensen JH, Christensen OB (2007) A summary of the PRUDENCE model projections of changes in European climate by the end of this century. Clim Change 81 (Suppl 1):7–30

Christensen OB, Drews M, Christensen JH, Dethloff K, Ketelsen K, Hebestadt I, Rinke A (2006) The HIRHAM Regional Climate Model Version 5 (β). Tech Rep 06-17. ISSN 1399-1388. DMI, Copenhagen

Christensen JH, Carter TR, Rummukainen M, Amanatidis G (2007) Evaluating the performance and utility of regional climate models: the PRUDENCE project. Clim Change 81(Suppl 1):1–6

Christensen JH, Boberg F, Christensen OB, Lucas-Picher P (2008) On the need for bias correction of regional climate change projections of temperature and precipitation. Geophys Res Lett 35:L20709 doi:10.1029/2008GL035694

Collins M, Booth BBB, Bhaskaran B, Harris GR, Murphy JM, Sexton DMH, Webb MJ (2010) Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model ensembles. Clim Dyn doi:10.1007/s00382-010-0808-0

Coppola E, Giorgi F, Rauscher SA, Piani C (2010) Model weighting based on mesoscale structures in precipitation and temperature in an ensemble of regional climate models. Clim Res 44:121–134

Déqué M, Somot S (2010) Weighted frequency distributions express modelling uncertainties in the ENSEMBLES regional climate experiments. Clim Res 44:195–209

Déqué M, Jones RG, Wild M, Giorgi F and others (2005) Global high resolution vs. regional climate model climate change scenarios over Europe: quantifying confidence level from PRUDENCE results. Clim Dyn 25:653–670

Déqué M, Rowell DP, Lüthi D, Giorgi F and others (2007) An intercomparison of regional climate simulations for Europe: assessing uncertainties in model projections. Clim Change 81(Suppl 1):53–70

Farda A, Déqué M, Somot S, Horányi A, Spiridonov V, Tóth H (2010) Model ALADIN as regional climate model for central and eastern Europe. Stud Geophys Geod 54: 313–332

Fowler HJ, Ekström M (2009) Multi-model ensemble estimates of climate change impacts on UK seasonal precipitation extremes. Int J Climatol 29:385–416

Giorgi F and others (2008) The regional climate change Hyermatrix framework. EOS 89:445–446

Giorgi F, Mearns LO (2002) Calculation of average, uncertainty range and reliability of regional climate changes from AOGCM simulations via the 'reliability ensemble averaging' (REA) method. J Clim 15:1141–1158

Giorgi F, Mearns LO (2003) Probability of regional climate change based on the reliability ensemble averaging (REA) method. Geophys Res Lett 30:1629

Haugen JE, Haakenstad H (2006) Validation of HIRHAM version 2 with 50 and 25 km resolution. RegClim general Tech Rep no. 9. Norwegian Meteorological Institute, Oslo, p 159–173

Haylock MR, Hofstra N, Klein Tank AMG, Klok EJ, Jones PD, New M (2008) A European daily high-resolution gridded dataset of surface temperature and precipitation. J Geophys Res 113:D20119 doi:10.1029/2008JD010201

Hofstra N, Haylock M, New M, Jones PD (2009) Testing E-OBS European high-resolution gridded dataset of daily precipitation and surface temperature. J Geophys Res 114: D21101 doi:10.1029/2009JD011799

Jacob D (2001) A note to the simulation of the annual and inter-annual variability of the water budget over the Baltic Sea drainage basin. Meteorol Atmos Phys 77:61–73

Jacob C, Bärring, L, Christensen OB, Christensen JH and others (2007) An inter-comparison of regional climate models for Europe: design of the experiments and model performance. Clim Change 81 (Suppl 1):31–52

Jaeger EB, Anders I, Lüthi D, Rockel B, Schär C, Seneviratne S (2008) Analysis of ERA40-driven CLM simulations for Europe. Meteorol Z 17:1–19

Kjellström E, Bärring L, Gollvik S, Hansson U and others (2005) A 140-year simulation of European climate with the new version of the Rossby Centre regional atmospheric climate model (RCA3). Rep Meteorol Climatol 108. SMHI, Norrköping, Sweden

Kjellström E, Boberg F, Castro M, Christensen JH, Nikulin G, Sanchez E (2010) Daily and monthly temperature and precipitation statistics as performance indicators for regional climate models. Clim Res 44:135–150

Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA (2010) Challenges in combining projections from multiple climate models. J Clim 23:2739–2758

Lenderink G (2010) Exploring metrics of extreme daily precipitation in a large ensemble of regional climate model simulations. Clim Res 44:151–166

Lorenz P, Jacob D (2010) Validation of temperature trends in the ENSEMBLES regional climate model runs driven by ERA40. Clim Res 44:167–177

Murphy JM and others (2004) Quantification of modeling uncertainties in a large ensemble of climate change simulations. Nature 429:768–772

Music B, Caya D (2007) Evaluation of the hydrological cycle over the Mississippi River basin as simulated by the Cana-

dian regional climate model (CRCM). J Hydrometeorol 8:969–988

Pal JS and others (2007) The ICTP RegCM3 and RegCNET: regional climate modeling for the developing world. Bull Am Meteorol Soc 88:1395–1409

Radu R, Déqué M, Somot S (2008) Spectral nudging in a spectral regional climate model. Tellus Ser A Dyn Meterol Oceanogr 60:898–910

Rossow WB, Walker AW, Beuschel DE, Roiter MD (1996) International Satellite Cloud Climatology Project (ISCCP) documentation of new cloud datasets. WMO/TD-No. 737, World Meteorological Organization, Geneva

Rummukainen M (2010) State-of-the-art with regional climate models. Clim Change 1:82–96

Sanchez E, Gallardo C, Gaertner MA, Arribas A, Castro M (2004) Future climate extreme events in the Mediterranean simulated by a regional climate model: a first approach. Global Planet Change 44:163–180

Sanchez-Gomez E, Somot S, Déqué M (2008) Ability of an ensemble of regional climate models to reproduce weather regimes over Europe-Atlantic during the period 1961–2000. Clim Dyn 33:723–736

Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. Philos Trans R Soc Lond A 365:2053–2075

Tebaldi C, Smith R, Nychka D, Mearns LO (2005) Quantifying uncertainties in projections of regional climate change: A Bayesian approach to the analysis of multi-model ensembles. J Clim 18:1524–1540

Uppala SM, Kållberg PW, Simmons AJ, Andrae U and others (2005) The ERA-40 Re-analysis. Q J R Meteorol Soc 131: 2961–3012

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, Exeter

van Meijgaard E, van Ulft LH, van de Berg WJ, Bosveld FC, van den Hurk BJJM, Lenderink G, Siebesma AP (2008) The KNMI regional atmospheric climate model RACMO, version 2.1 KNMI-publication TR-302. KNMI, De Bilt

van Ulden A, Lenderink G, van den Hurk B, Meijgaard E (2007) Circulation statistics and climate change in Central Europe: PRUDENCE simulations and observations. Clim Change 81:179–192

Weigel AP, Liniger MA, Appenzeller C, Knutti R (2010) Risks of model weighting in multi-model climate projections. J Clim 23:4175–4191

Wilby R, Harris I (2006) A framework for assessing uncertainties in climate change impacts: low flow scenarios for the River Thames, UK. Water Resour Res 42:W02419

Submitted: May 28, 2010; Accepted: July 30, 2010

Proofs received from author(s): Decemmber 2, 2010