Vol. 44: 151–166, 2010
doi: 10.3354/cr00946

CLIMATE RESEARCH
Clim Res

Published December 10

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

Exploring metrics of extreme daily precipitation in a large ensemble of regional climate model simulations

Geert Lenderink*

PO Box 201, 3730 AE De Bilt, The Netherlands

ABSTRACT: The ability of a large ensemble of 15 state-of-the-art regional climate models (RCMs) to simulate precipitation extremes was investigated. The 99th, 99.9th and 99.99th percentiles of daily precipitation in the models were compared with those in the recently released E-OBS observational database for winter, spring, summer and autumn. The majority of the models overestimated the values of the precipitation extremes compared with E-OBS, on average by approximately 38%, but some models exceeded 50%. To measure model performance, a simple metric is proposed that averages a nonlinear function of the seasonal biases over the European area. The sensitivity of the metric to different assumptions in the construction and the quality of the observational data was explored. Generally, low sensitivities of the metric to spatial and seasonal averaging were found. However, large sensitivities to potential biases in the observational database were found. An alternative metric that measures the spatial pattern of the extremes (which is not sensitive to a potential constant offset in the observational data) was further explored. With this metric, the ranking between the models changed substantially. However, the 2 models with the worst score in the standard metric also displayed the worst scores with this alternative metric. Finally, the regional climate models displayed the largest biases compared with E-OBS in areas where the underlying station density used in E-OBS is low, thus suggesting that data quality is indeed an important issue. In summary, the results show that: (1) there is no metric that guarantees an objective and precise ranking or weighting of the models, (2) by exploring different metrics it nevertheless appears possible to indentify models that perform consistently worse than other models, and (3) the observational data quality should be considered when designing and interpreting metrics.

KEY WORDS: Precipitation · Extremes · Regional climate model · Model performance

1. INTRODUCTION

It is generally expected that precipitation extremes increase as the climate warms. This is of considerable societal interest because the impact of local or small-scale precipitation extremes on society through local flooding, erosion and water damage is large. There is widespread evidence from models and observations that precipitation extremes increase with higher temperatures (e.g. Frei et al. 2006, Fowler et al. 2007, Pall et al. 2007, Lenderink et al. 2007, Lenderink & van Meijgaard 2008, 2010, O'Gorman & Schneider 2009).

Despite the considerable societal interest in precipitation extremes, models that simulate them display substantial deficiencies. In particular, this applies to the simulation of the extremes in summer, which are mostly of convective origin. Convection is not resolved in the current generation of regional climate models (RCMs), and models use parameterizations to represent convection. It is known that these parameterizations have shortcomings, for example in the representation of the diurnal cycle (Guichard et al. 2004) or the sensitivity to soil conditions (Hohenegger et al. 2009).

*Email: lenderin@knmi.nl

In order to have confidence in model projections of the future, climate models should be based on sound representation of the physical processes, which should not only apply for the present-day climate but also be valid in a future climate. In a session of the General Assembly of the ENSEMBLES project (Hewitt & Griggs 2004, van der Linden & Mitchell 2009) on weighting, the vast majority of the approximately 80 participants — a mix of climate and impact modellers using both statistical and dynamics downscaling tools — agreed with this statement (Prague, November 2007, summary of session: http://www.knmi.nl/samenw/ensembles_rt5/rt5_files/ENSEMBLES_weightingworkshop_v03.pdf). The descriptive quality of the model physics is not readily quantified. A thorough evaluation of models on a process level is required to establish how well models represent the key physics of the climate system. This is a very demanding task, which relies on the understanding of the physical processes, the availability of observations at a process level and thorough insights in the technical details of the models. Presently, this can only be done on a case-by-case basis.

Another way of increasing our confidence in future climate projections is to evaluate the models' performance for the present-day climate over a wide range of different climate conditions and different variables. Although models can be (and are) tuned to optimize their performance for one particular area or variable, it is far more difficult to get a good and coherent performance for different climate regions and variables. Models that perform well overall are generally considered to be more realistic (that is, contain a better representation of the physics) than models that have considerable variation in performance between different areas and variables. To perform such an evaluation it is essential to have observational data covering large areas and long time periods.

In the present study we pursued the latter approach, and explored ways to test the models' performance with respect to precipitation extremes over Europe. This was done for a large ensemble of regional climate model simulations for the present-day climate as performed in the EU-funded ENSEMBLES project (Hewitt & Griggs 2004, van der Linden & Mitchell 2009).

Previous comparisons between regional climate models and observations often used model integrations that had been driven at their lateral boundaries by information derived from global climate model (GCM) integrations. Therefore, these regional model simulations are influenced by synoptic forcing from the GCM, which do not necessarily correspond (in a statistical sense) with the observations. This may affect the simulation of the extremes, in particular for the winter season where the influence of the information imported through the lateral boundaries is strong. One novelty in the ENSEMBLES project is the availability of a large number of regional climate models which are all forced by realistic boundaries from the ERA40 re-analysis project (Uppala et al. 2005). This also allows a more objective inter-comparison of the available regional climate models.

In this paper extremes of daily precipitation as simulated by the RCMs are compared to a new, high resolution gridded dataset of daily observations in Europe (Haylock et al. 2008). This E-OBS data set uses the same grid as the majority of the regional climate models, and the data is especially designed to represent area averages instead of local measurements. This allows an optimal comparison between model and observations. E-OBS is unique in its spatial and temporal extent covering the whole of Europe from 1950 to 2008, the high resolution of (approximately) 25 km, and the use of many observational stations (~2900). Yet, despite the fact that the E-OBS data set is the best available at the moment, it is known that the extremes could be underestimated due to the 'averaging' procedure from station data to area averages (Haylock et al. 2008, Hofstra et al. 2010).

Finally, we note that this study is part of a larger effort to establish model weighting systems based on exploring model performance (this CR Special). Besides the weights derived here, a number of other weights have been derived, e.g. inducing weights from the reproduction of patterns on the meso-scale, the seasonal cycle and synoptic pressure patterns. A synthesis of these weights is given in Christensen et al. (2010, this Special).

2. DATA AND METHODS

2.1. Observations and models

The RCMs were compared with the recent E-OBS observational data set (Haylock et al. 2008), which also has been developed in the ENSEMBLES project. E-OBS contains daily observations gridded onto 4 different grids: 2 regular latitude × longitude grids at 0.25 and 0.5° resolution, and 2 rotated grids at 0.22 and 0.44° resolution. For comparison with the model results, we used the E-OBS data set on the 0.22° degree rotated grid that is used by most RCMs. The E-OBS data set has been specially designed to represent grid box average values, instead of point values. This is essential to enable a direct comparison with the model data (see e.g. Chen & Knutson 2008). We used the second release of this data set (released in summer 2009).

The RCM integrations have been performed in contribution to the ENSEMBLES project. All RCMs have been driven at the lateral boundaries by the meteorological fields from the ERA40 re-analysis, thereby forc-

ing the atmospheric motions to be close to the observations (Sanchez-Gomez et al. 2009). Most models use an identical grid with a rotated pole at a 0.22° resolution. A few models use a Lambert conformal grid, with approximately the same resolution. In the observations, differences between a 0.25° regular and the 0.22° rotated grid are negligible for the measures considered here. Thus, the impact of the different grids in the RCM simulations is expected to be marginal. The different models are defined in Table 1.

2.2. Methods

The daily precipitation data are pooled in boxes of 2 × 2° longitude × latitude on a regular grid. With the used grid, ~40 to 60 grid points are within each box, except when a box contains a substantial sea fraction. The pooling of data allows the computation of the higher percentiles (rarer extremes) at the expense of degrading the horizontal resolution.

The 99th, 99.9th and 99.99th percentiles of the distribution were computed from the pooled data. These percentiles correspond to events with a frequency of occurrence of once every 100, 1000 and 10 000 d, respectively. This procedure was done for each season: winter (DJF), spring (MAM), summer (JJA) and autumn (SON). For each season, the above percentiles correspond roughly to a frequency of occurrence of once every year, once every 10 yr and once every 100 yr, respectively.

By pooling data from individual grid boxes, biases in the statistics of the extremes may result. If the precipitation distribution is not homogeneous in space, the highest percentiles could be biased towards those stations or grid boxes with the highest precipitation amounts within the 2 × 2° box. A negative bias in one part of the box could also be compensated for by a positive bias elsewhere, leading to good overall scores.

In addition, spatial correlations in precipitation could lead to an underestimation of the highest percentiles. For the Netherlands, however, Overeem et al. (2008, 2009) show that spatial correlations of daily precipitation extremes in rain radar data are very small for distances >50 km, and that the statistics of extremes with a return period <100 yr are barely affected by spatial correlation. Data from observations and models are treated in the same way, thus biases due to spatial correlation are expected to be similar in both observations and model results. However, if the spatial correlation in the observations is different from the correlation in the model result, this need not be the case. For instance, Leander & Buishand (2007) found larger spatial correlations in the output of an RCM integration than in the observational data set for sub-basins of the Meuse River. Also, the spatial correlation in the E-OBS data set could be affected by the low station density used in E-OBS in some areas (see Section 6). It is not trivial to quantify these effects, and we consider this outside the scope of this paper. In particular, in comparison with the substantial differences between model results and the observational data set, we think that these effects are likely to be small.

For the analysis, the recent period 1971–2000 was used. We refer to this period as present-day climate, although we acknowledge that the climate may have already changed since then. For this period, the quality of the ERA40-derived boundaries for the regional models is high. The European (EU) domain that is analyzed consists of the land area extending from 10° W to 30° E and from 38 to 68° N.

Table 1. Overview of the different regional climate models used in this study. Model references and full names of the institutes can be found in Christensen et al. (2010, this Special) and at http://ensemblesrt3.dmi.dk/. LF grid: Lambert conformal grid; Reg. grid: a regular 0.25° latitude–longitude grid. The other models use the common rotated grid

Model no.	Institute	Model	Remarks
M1	C4i	RCA3.0	
M2	CHMI	ALADIN	LF grid
M3	CNRM	RM4.5	LF grid
M4	DMI	HIRHAM5	
M5	ETHZ	CLM	
M6	ICTP	RegCM3	LF grid
M7	KNMI	RACMO2	
M8	Met.No	HIRHAM	
M9	Meto-HC	HadRM3Q0	Normal climate sensitivity
M10	Meto-HC	HadRM3Q3	Low climate sensitivity
M11	Meto-HC	HadRM3Q16	High climate sensitivity
M12	MPI	REMO	
M13	OURANOS	MRCC4.2.3	Reg. grid
M14	SMHI	RCA3.0	
M15	UCLM	PROMES	LF grid

3. SIMULATION OF EXTREMES

3.1. Observations

The 99.9th percentile of daily precipitation (P99.9) in the observations is shown in Fig. 1 for winter, spring, summer and autumn. On average, P99.9 is 30 to 40 mm over most parts of Europe. The Alpine region is characterized by larger values of P99.9 throughout the year. Larger values of P99.9 are also

observed in the northern coastal areas of the Mediterranean Sea in autumn (and during parts of winter), along the coast of Norway and in the northwestern part of Spain and Portugal (mainly in autumn and winter).

3.2. Model results

The bias of P99.9 of daily precipitation in the different RCM simulations for winter and summer is shown in Figs. 2 & 3. Most models display a considerable positive bias for most regions. Exceptions to these rules are M3, M13 and M14 in winter, and M10 and M13 in summer. Large biases on the order +100% are observed over large areas for M4 and M8 in winter, and M5, M6 and M8 in summer. Most other models also have large areas with biases exceeding +50%.

The spatial (EU domain) bias averaged over the 4 seasons in the RCMs is +38%, and ranges from −5% (M13) to >+50% (M4, M5 and M6) and ~+70% (M8). The variations between the different seasons are considerable in the individual models, and differences between the seasons ranges from <10 to >40% (M10). However, averaged over all models, the seasonal difference is not large, and the mean bias ranges between +32% (in autumn) and +42% (in spring).

The ensemble mean precipitation amount and the ensemble mean bias are shown in Fig. 1 (middle and lower panels, respectively). Overall, the pattern of the bias is rather constant throughout the year. Relatively high values of the bias are observed in central parts of Spain, Eastern Europe and northern parts of Scandinavia. Biases are generally small in Ireland, England, the Netherlands, western Germany and southwestern Norway

To further condense the results, we employed the commonly used areas defined in the PRUDENCE project: Scandinavia, the British Isles, the Alps, Eastern Europe, mid Europe, France, the Iberian Peninsula and the Mediterranean (see Christensen & Christensen 2007 for the exact definition of these areas). The average P99.9 values are shown in Fig. 4 for both the models and the observational data set.

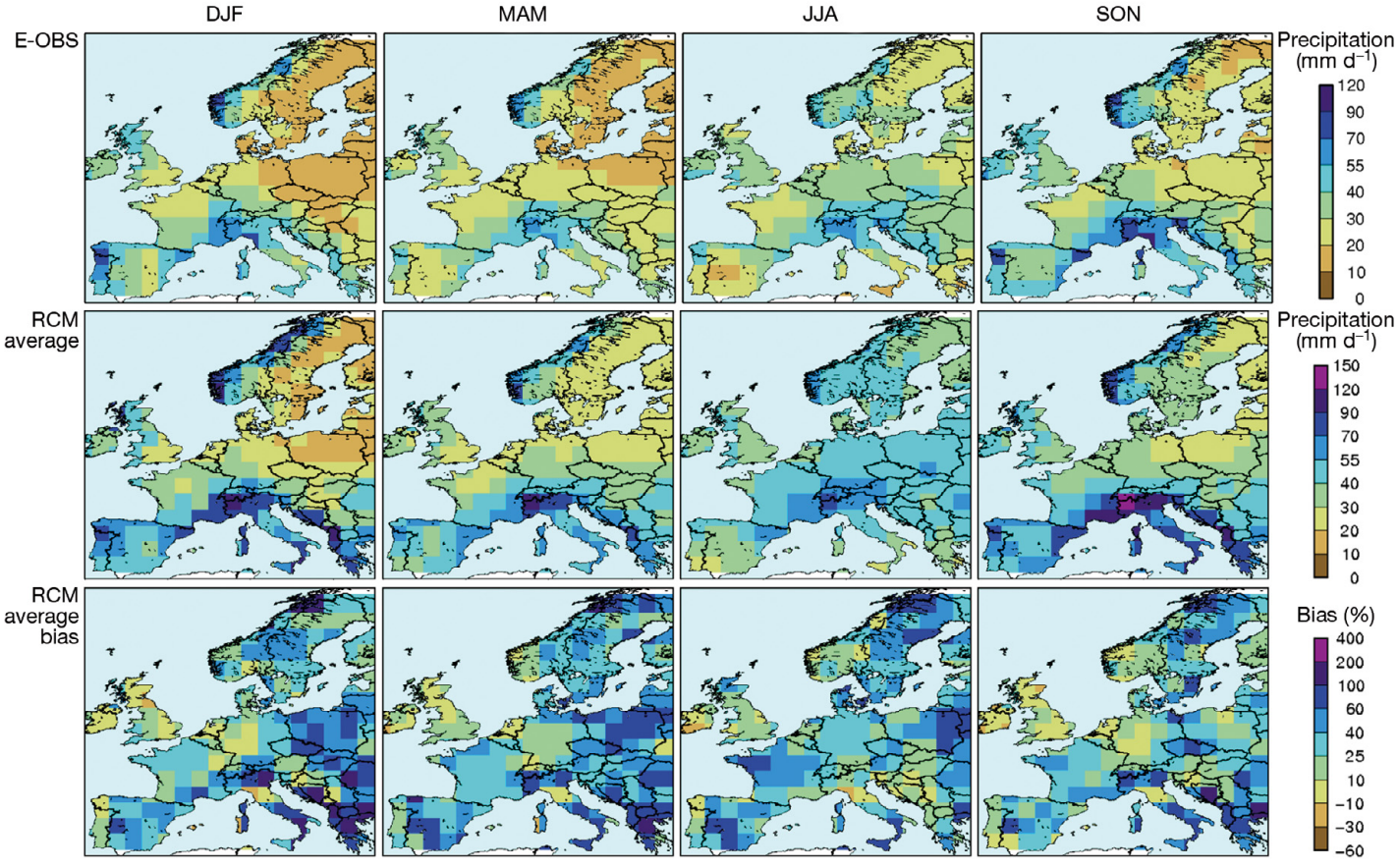

Fig. 1. The 99.9th percentile of daily precipitation (P99.9) for the observational data set (E-OBS; upper panels) and the average of the regional climate models (RCMs; middle panels). The lower panels show the bias of the model results averaged over all RCMs compared with the observations. The extremes are computed from the pooled data of 2 × 2° longitude × latitude boxes, and different seasons are plotted from left to right

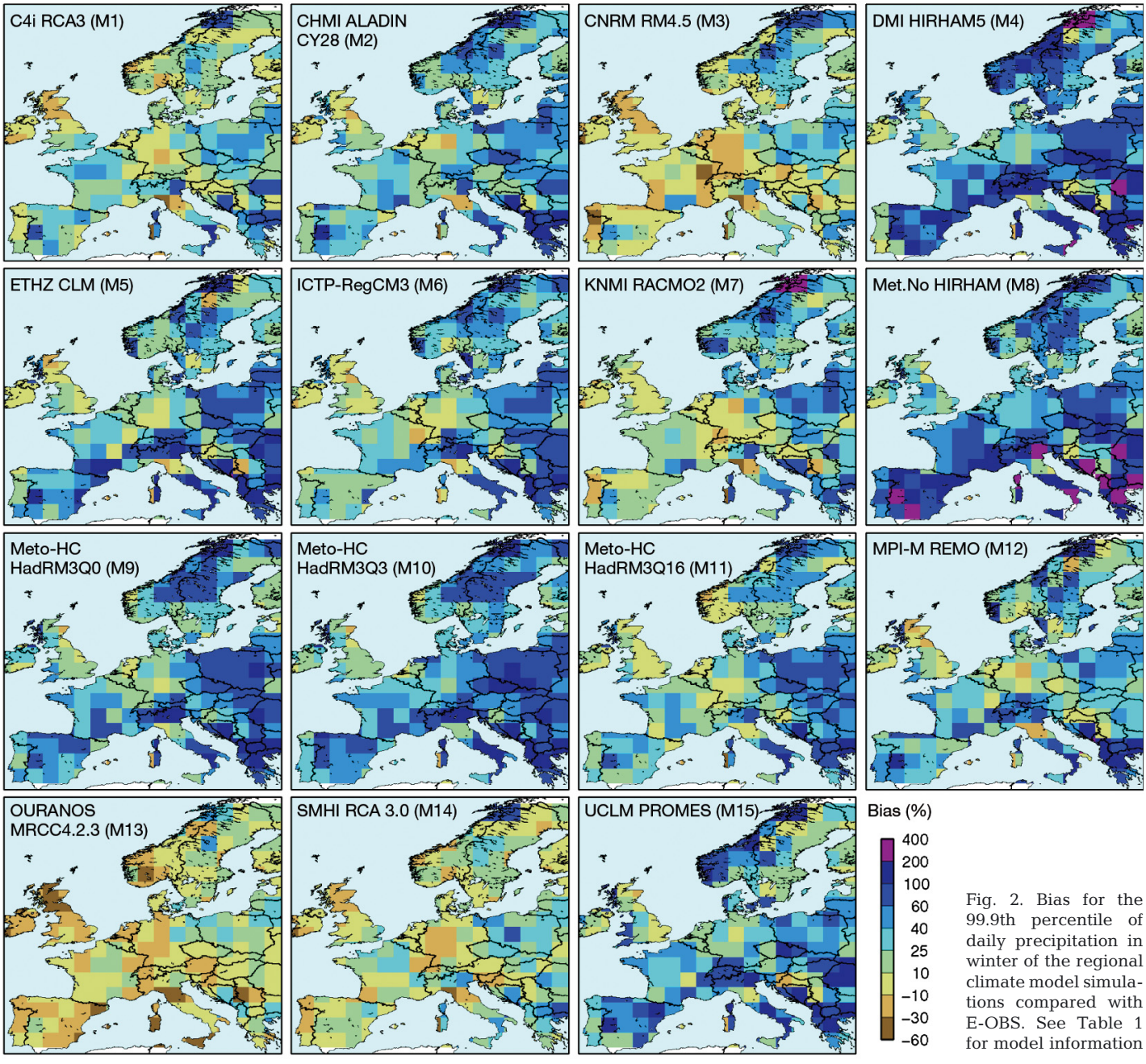

Fig. 2. Bias for the 99.9th percentile of daily precipitation in winter of the regional climate model simulations compared with E-OBS. See Table 1 for model information

In autumn and winter simulated precipitation extremes are very high in southern Europe; in many models, P99.9 > 80 mm d^{-1} in the Iberian Peninsula and the Mediterranean, whereas observed amounts are ~50 mm d^{-1}. Models also produce high amounts in the alpine region. Extremes are (on average) relatively low in Scandinavia, mid Europe and Eastern Europe. We note that in Scandinavia, the high extremes along the coast in Norway are not reflected in the means.

The biases of the area mean extremes in the RCMs compared with the observations are shown in Fig. 5. The vast majority of the models overestimate P99.9 for all seasons and all areas considered. In particular, models M8 and, to a lesser extent, M4, M5 and M6 have large values of the bias for all seasons and the majority of the areas. One model, M13, however, tends to slightly underestimate the extremes for most seasons and areas. The different versions of HadRM3 have different versions of the physics parameteriza-

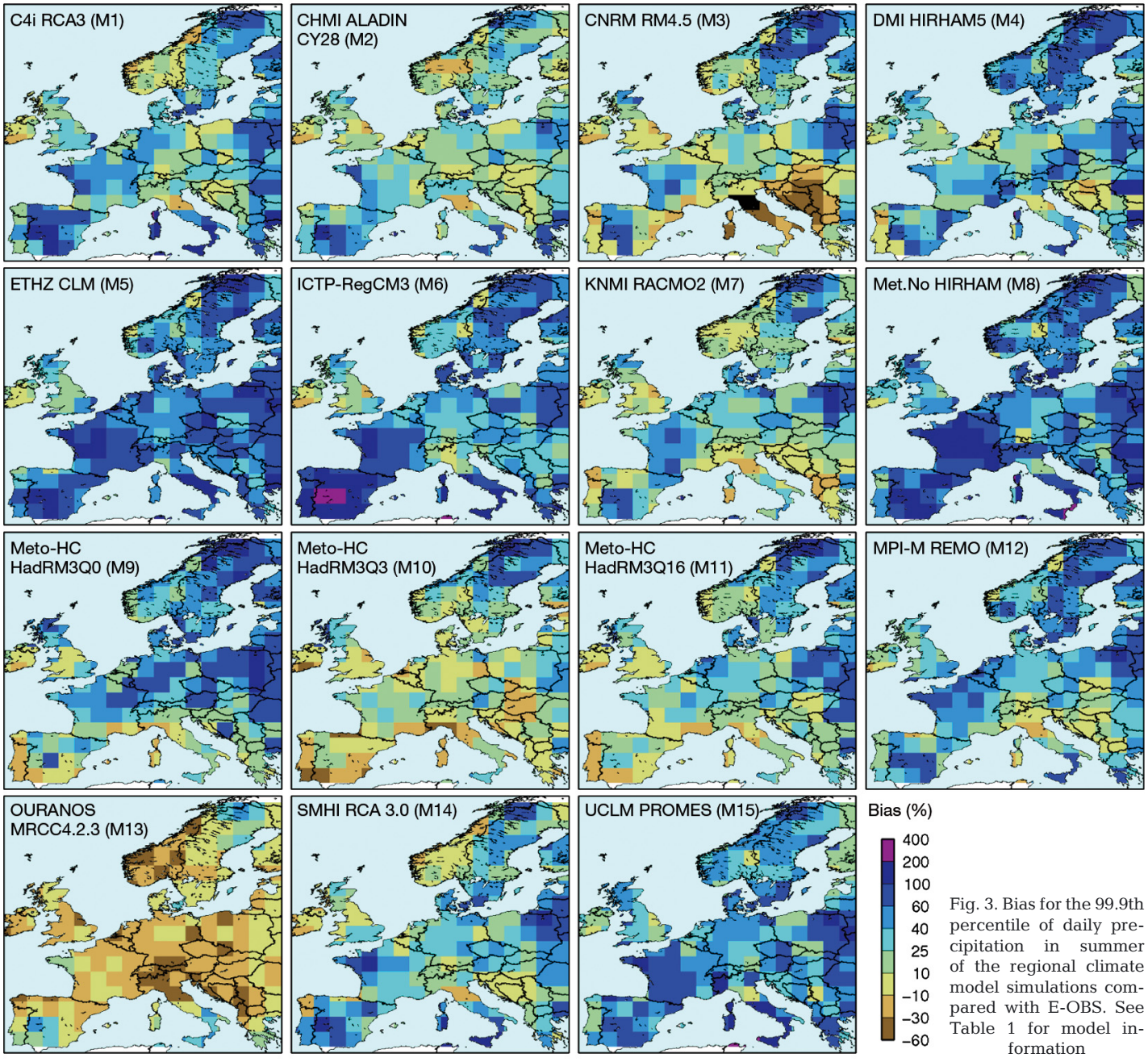

Fig. 3. Bias for the 99.9th percentile of daily precipitation in summer of the regional climate model simulations compared with E-OBS. See Table 1 for model information

tions that lead to different climate sensitivities, which is the response of the global mean temperature to a doubling of greenhouse gases. The low sensitivity version of HadRM3, M10, performs relatively well for the summer period, but has large positive biases for the other seasons. The differences between the 3 versions of HadRM3 are considerable, in particular for the summer season. This shows that the changes in the physics that are responsible for the different global climate sensitivity also directly impact the hydrological cycle. This likely relates to changes in the deep convection scheme compared with the reference model. The mean model biases vary slightly with the season, with the largest biases in winter and spring and the smallest biases in summer and particularly autumn. Overall, the largest biases occur in Scandinavia, France, the Iberian Peninsula and the Mediterranean.

The spatial distribution of the extremes over Europe is visualized in Fig. 6. Here, the anomaly of the area mean to the (un-weighted) mean over all

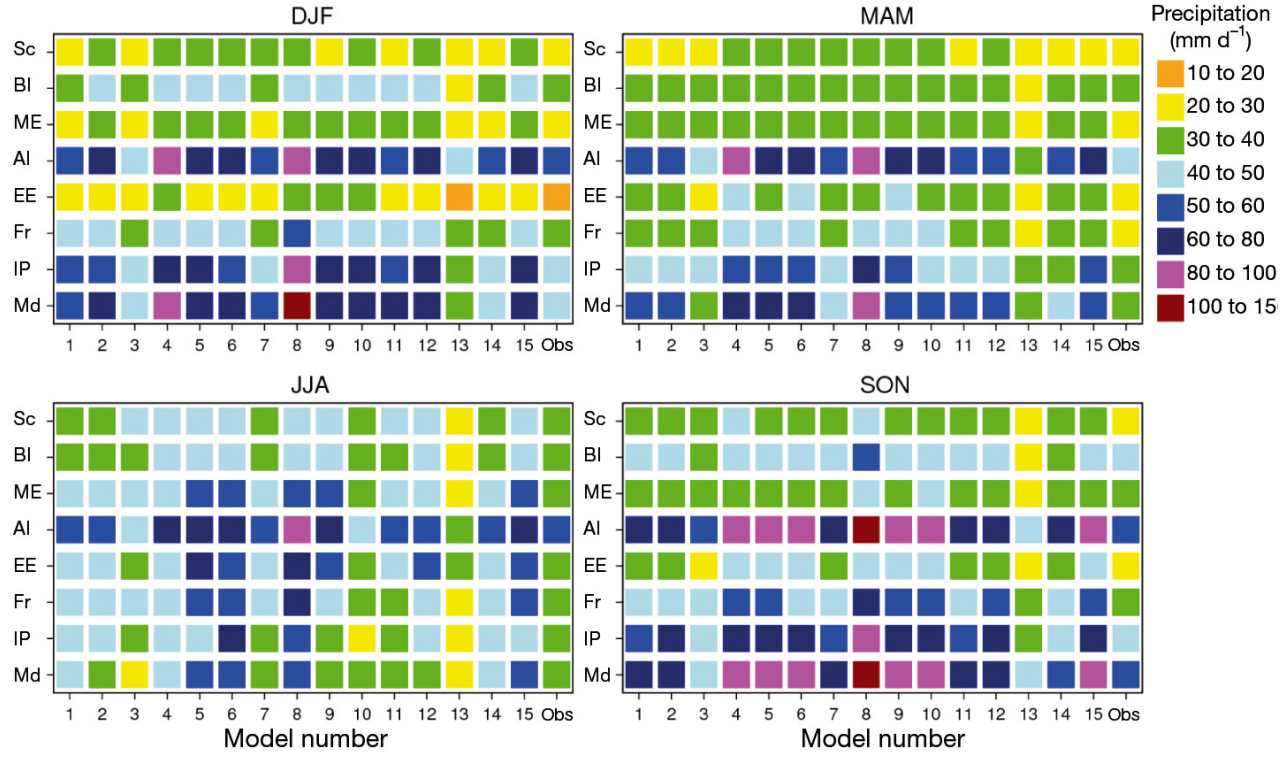

Fig. 4. Area averages of the 99.9th percentile of daily precipitation for the 15 different regional climate models (see Table 1 for model descriptions) and the observational data set (E-OBS) in different seasons. AI: Alps; BI: British Isles; EE: Eastern Europe; Fr: France; IP: Iberian Peninsula; Md: Mediterranean; ME: mid Europe; Sc: Scandinavia

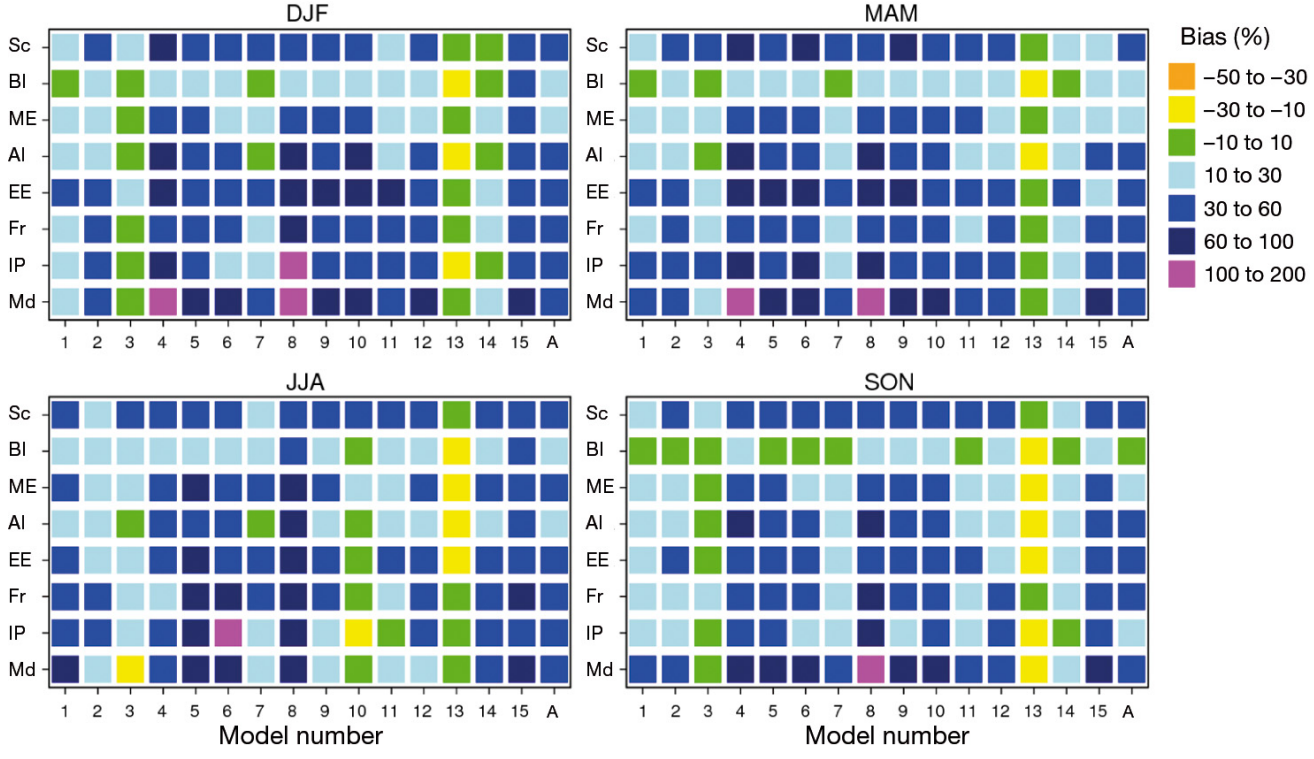

Fig. 5. Bias of the 99.9th percentile of daily precipitation for the 15 different regional climate models (see Table 1 for model descriptions) compared with the observations for the different seasons. A: mean bias averaged over all models. See Fig. 4 for area abbreviations

areas is plotted for models and observations. In general, all models simulate the differences between the different areas reasonably well for all seasons. Exceptions are the relative overestimation of the extremes in the Mediterranean (except for summer) and the relative underestimation of the extremes for the British Isles.

To illustrate the typical model scatter in simulating the differences between the different areas, we show the results of models M4 and M13 separately. The absolute bias in these models is very different, with large biases in M4 and small biases in M13. However, despite this large difference in model performance, the relative differences between these areas is simulated almost equally well for both models when averaged over all seasons. In fact, model M4 scores very well for the summer period (with the smallest squared difference with the observations of all models when averaged over all areas). For the other seasons, M13 scores better than M4.

4. A METRIC OF MODEL PERFORMANCE

Below, a simple metric is proposed to measure the model performance with respect to the precipitation extremes. In the following, we often use the term weight. This refers to the value of the metric. The intended use of these weights is to quantitatively weight model results according to their performance, and thus obtain more precise climate change predictions. How this can be accomplished is not a topic of this paper, but is discussed by Christensen et al. (2010) elsewhere in this CR Special issue.

To compute the RCM weights, we first computed the bias B (%) and then converted the bias to a weight (W) using a simple transformation:

$$B = 100\left[\frac{P_{\mathrm{RCM}}}{P_{\mathrm{obs}}} - 1\right]$$
$$W = 1 + B/100, \qquad B < 0$$
$$= \frac{1}{1 + B/100}, \qquad B > 0 \tag{1}$$

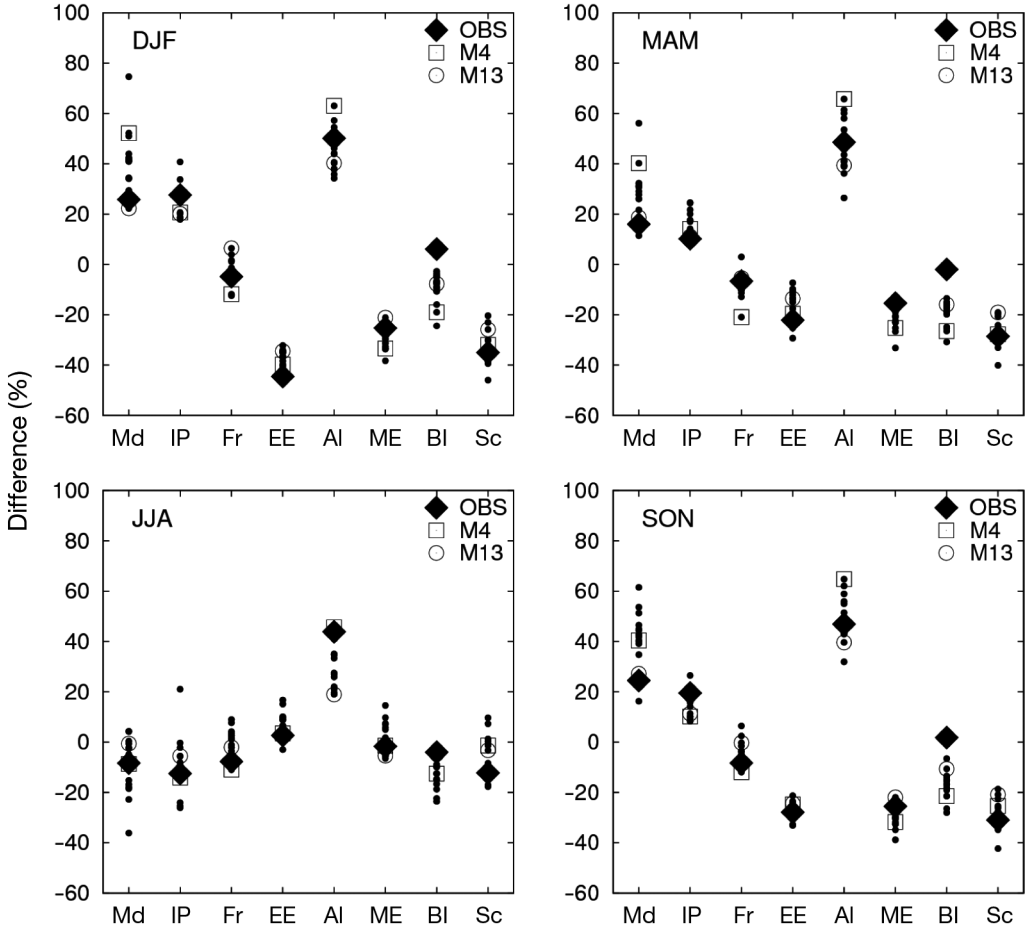

Fig. 6. Difference in the 99.9th percentile of daily precipitation relative to the mean of all areas (see Fig. 4 for area abbreviations), showing the spatial difference across Europe for all models (black dots) and the E-OBS observations (black diamond). Results of model M4 and M13, as discussed in Section 3.2, are indicated separately

where P_{RCM} and P_{obs} are RCM-predicted and observed precipitation, respectively. The metric W gives the same penalty to models that produce half of the observed rainfall amount as models that produce twice the observed amount. This asymmetry is needed to prevent negative W for $B > +100\,\%$. In particular, when observed amounts are small, large biases $>100\,\%$ could (and do) occur. For small values of B, the metric is, however, symmetric around $B = 0$. For example, the relative difference between the weights assigned to a positive and negative bias of 20 % is only 4 % (0.8333 versus 0.8, respectively). Of course, we acknowledge that the definition of the metric is, to a considerable extent, subjective. However, the definition has 2 important properties: it is symmetrical for small values of the bias, and it produces a well-constrained range between 0 and 1 for all possible values of the bias, with $W = 1$ only if the bias is zero.

For each model and each season, the weights were computed following the procedure outlined below. We first computed W for each season, each percentile and each $2 \times 2°$ box. Then, for each season, we averaged first over the weights belonging to the different percentiles for each box, and then over the different boxes in the European area. These seasonal weights were finally averaged to yield final weights for the different models. We note that the order of averaging is only relevant when a percentile could not be computed, in which case it was neglected. For example, the highest percentile could sometimes not be computed when only a few grid points were contained in a box. In that case, the average for that box only consists of the average of the lower percentiles. We note, however, that the impact of the order of averaging on the final weights is very small (<0.01 in the final weights).

Table 2 gives the seasonal and final weights for the different RCMs. Typically, the value of the weights are 0.75 with a standard deviation amongst the RCMs of 0.06. The models with high scores are M13, M14 and M3, whereas M8 has the lowest score. Most models score rather evenly over the seasons, with models M6, M10 and M14 showing the largest interseasonal variations. There is no season in which the models perform significantly better or worse.

We also computed the metric from the bias in P99.9 only. Differences from the standard metric, computed from the biases in P99, P99.9 and P99.99, turned out to be very small; typically <0.01 in the yearly average and the seasonal means (not shown).

5. SENSITIVITY TESTS

To highlight the explorative nature of this research, we performed several sensitivity tests to establish how much the results depend on several subjective choices in the construction of the metric. In addition, we briefly investigate how sensitive our metric is to potential errors in the observational database. Finally, given the large biases in most models but the comparatively good simulation of the spatial differences within Europe, we also explore the possibility of designing a metric based on spatial patterns and compare this metric with the standard metric.

5.1. Sensitivity to averaging of the seasonal weights

First, we investigated how the method of temporal averaging of the seasonal values of W influences the yearly value. In the case of the reference metric, we simply performed an arithmetic mean of weights of the different seasons. However, within ENSEMBLES there has been some debate whether weights should be obtained by performing multiplication of the separate weights. The philosophy behind the approach of multiplying weights is that in order to receive a high weight a model needs to perform well in all metrics considered, so as to avoid a possible counterbalancing effect of different systematic biases.

The geometric mean of n values is defined as the nth root of the product of these n values. Thus, the multiplication of the weights of the 4 seasons is identical to the

Table 2. Weights obtained with the standard metric for the different regional climate models (RCMs). Shown are seasonal (DJF, MAM, JJA and SON) and yearly means (AVE). In addition to the RCMs (M1 to M15), the first version of E-OBS is also shown

Model	DJF	MAM	JJA	SON	AVE
M1	0.82	0.76	0.75	0.82	0.79
M2	0.74	0.77	0.80	0.76	0.77
M3	0.84	0.79	0.77	0.81	0.80
M4	0.64	0.64	0.73	0.70	0.68
M5	0.72	0.68	0.65	0.72	0.69
M6	0.72	0.64	0.65	0.74	0.69
M7	0.77	0.77	0.79	0.80	0.78
M8	0.64	0.66	0.64	0.65	0.65
M9	0.71	0.66	0.72	0.73	0.70
M10	0.68	0.70	0.80	0.72	0.72
M11	0.76	0.73	0.78	0.78	0.76
M12	0.74	0.75	0.71	0.75	0.74
M13	0.85	0.86	0.82	0.83	0.84
M14	0.85	0.79	0.75	0.83	0.81
M15	0.73	0.77	0.68	0.75	0.73
Mean ± SD	0.75 ± 0.07	0.73 ± 0.07	0.74 ± 0.06	0.76 ± 0.05	0.74 ± 0.06
E-OBS v1	0.93	0.93	0.94	0.94	0.93

fourth power of the geometric mean of the seasonal weights. Fig. 7a shows that there is no practical difference between the geometric and the (normal) arithmetic mean, a rather trivial result given the relatively small spread in the seasonal weights between the models.

Thus, multiplying weights instead of averaging does not change the ranking (between good and bad models) of the models as long as the spread in sub-weights is relatively small. It does, however, change the final spread in the weights, but a similar spread could also be obtained by the arithmetic mean of the nth power of the sub-weights (where n is the number of sub-weights).

5.2. Sensitivity to spatial averaging of weights

Second, we investigated how sensitive the results are to the method of spatial averaging. An alternative weight was computed from the model biases of the 8 European areas. The same formula as in the standard method was used to convert the mean bias of the 99.9th percentile for each area into a weight, and then averaged over the 8 European areas (without taking into account the difference in size between the different areas). Thus, this method uses a different spatial averaging (less sensitive to small-scale spatial structures) and only the 99.9th percentile instead of the 3 percentiles used in the standard method. As noted earlier, the influence of the latter difference is small, thus differences are mainly due to the spatial averaging method.

The resulting weights obtained with the alternative method of spatial averaging are similar to the weights from the standard method. There is a clear correlation between the performance of the models for the 2 metrics (Fig. 7b). The influence of the spatial averaging

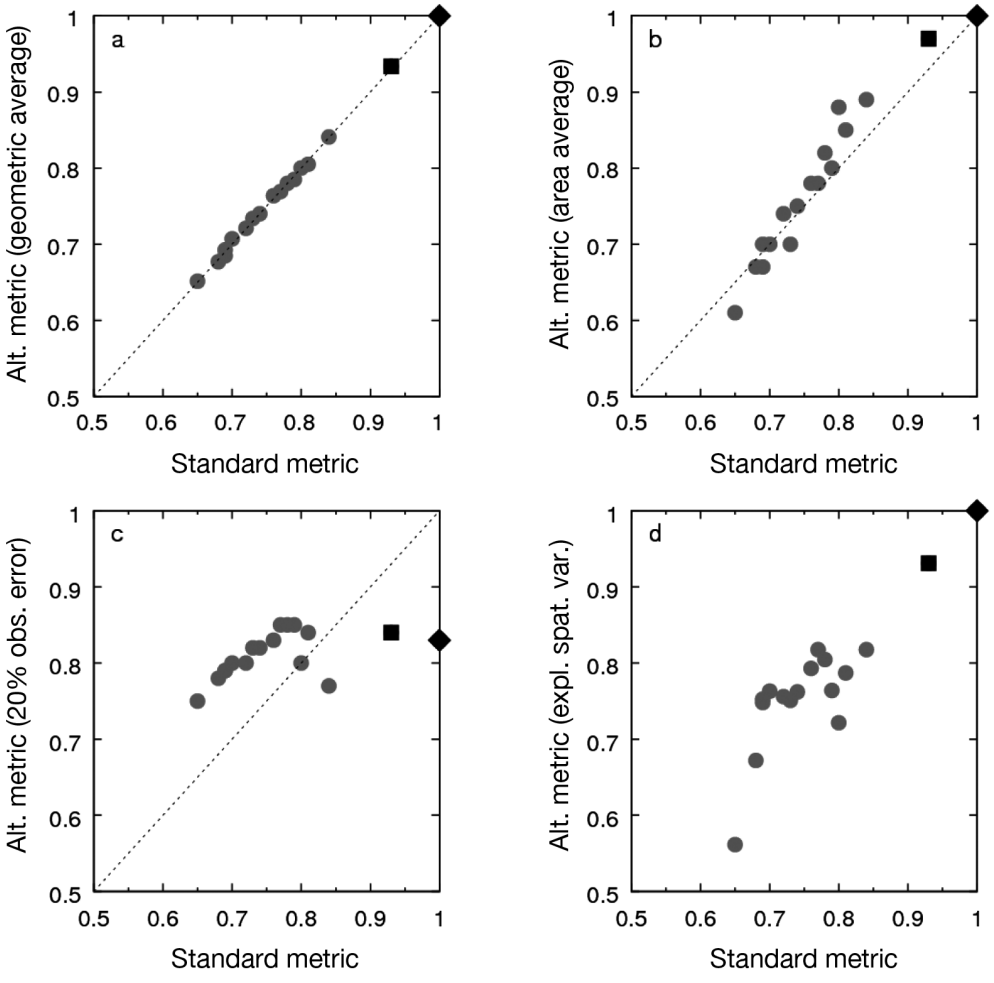

Fig. 7. Comparison of 4 alternative (Alt.) metrics (on the y-axis) to the standard metric (on the x-axis). Shown are weights from (a) a metric based on geometric (instead of arithmetic) average of the seasonal weights, (b) a metric based on averages over the 8 European PRUDENCE areas, (c) the standard metric, but using an artificially constructed observational data set (E-OBS amplified by 20%), and (d) a metric based on the explained spatial variance (see section 5 for details on these alternative metrics). Grey dots: model results; black diamonds: E-OBS data; black squares: version 1 of E-OBS

method is, however, larger than the influence of geometric versus arithmetic averaging. For example, small changes in the ranking of the models result from the alternative spatial averaging method.

5.3. Sensitivity to observation data

Third, we investigated how potential biases in the observational database could affect the model scores. The black squares in Fig. 7 show weights of the first release of the data set when compared with the second release. Treating the first release as a model, it gets a weight of 0.93 with the standard metric, compared with the average weight of 0.75 of the models. Computing weights from the area averages (as outlined in Section 5.2), the first release of the database gets an even higher score. This is consistent with the fact that most of the changes between the 2 versions of the database are on a small scale. Thus, the previous version of the observational database can be distinguished from the model results by means of the proposed metrics.

The uncertainty estimates of the observational data set are, however, larger than is reflected in the difference between the 2 releases. For each grid box and each day, E-OBS contains an estimate of the interpolation error in the E-OBS database (see Haylock et al. 2008). This error mainly describes the error of the interpolation of the station observations to a very high resolution (0.1° base grid and the subsequent aggregation onto the 0.22° grid that is used here). In particular, the error related to spatial aggregation is difficult to assess because it is strongly dependent on estimates of the spatial correlation (shared variance), which is difficult to assess due to the low station density (see Hofstra et al. 2009, 2010). It is also not entirely clear whether this standard error should be interpreted as a random error or whether (in part) this error could be systematic. As a simplistic approach, we interpreted the error as a systematic bias, and computed the percentiles from the distribution of E-OBS plus the standard error. The bias in the 99.9th percentile compared with E-OBS from this crude approach is shown in Fig. 8. We note that this figure mainly reveals the spatial differences in the error estimates, and that the absolute magnitude of the error in the extreme should be considered with caution.

As a measure of the uncertainty in E-OBS, we also compared the RCMs with an alternative data set for the Rhine catchment area issued by the International Rhine Commission (CHR) (Krahe et al. 2010). The data consists of precipitation measurements in 134 sub-catchments. Here, we used only the 119 sub-catchments between Lobith (at the German–Dutch border) and Rheinfelden (at the Swiss–German border). The average size of these catchments is ~1000 km^2, which corresponds to a grid size of 33 km, which is somewhat larger than the resolution of the RCM and E-OBS. Selecting only catchments with sizes between 400 and 900 km^2 (between 20 and 30 km resolution) showed no substantial differences; thus the difference in spatial resolution appears to be a minor issue. For the summer period, the difference between the 2 data sets amounts to approximately +20% for the extremes, with the higher amounts occurring in the CHR database (Fig. 9). We note that +20% is well within the uncertainty estimates from the E-OBS database, when interpreted as a systematic error. For the winter period, the differences were approximately +15%.

If we assume (for the sake of this experiment) that the E-OBS data has a uniform bias of +20% over Europe, how would this impact on the weights? Fig. 7c shows the resulting weights obtained with this artificial data set compared with the weights obtained with E-OBS. Clearly, this changes the rating of many models, and models which are close to E-OBS (or with a small negative bias) now get a much lower score. The model with the highest score, M13, ends up with one of the lowest scores against this artificial data set. Most models, however, get a higher score, which is expected because most models had a positive bias. E-OBS itself gets a score of 0.83 (1 divided by 1.2) compared with

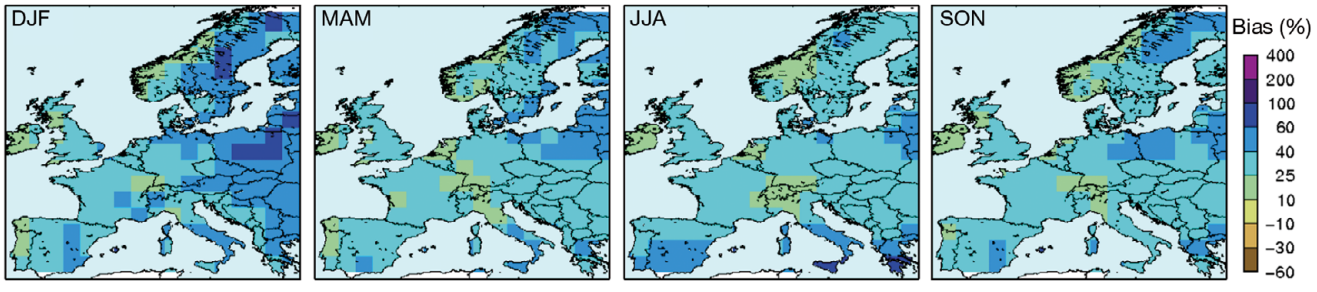

Fig. 8. Relative difference (bias) in P99.9 computed from a data set, constructed by adding the standard error estimates to E-OBS on a daily basis, compared to E-OBS itself. We note that this is a crude approach to estimate the errors in P99.9 in the observations, and assume that the error estimates are (primarily) systematic

this artificially constructed data set, which is now close to the best models. Thus, it is clear that a potential offset in the observational database highly influences the scores of the models.

5.4. An alternative metric based on spatial patterns

Finally, we investigated an alternative metric that focuses entirely on the spatial patterns and is not sensitive to an offset in the data. This is also motivated by the fact that most models reproduce the spatial patterns over Europe reasonably well (see Fig. 6). To this end, we used the relative precipitation intensity of the 99.9th percentile of each 2 × 2° box (S99.9). S99.9 is defined by the value of the 99.9th percentile at that grid box divided by the European mean for that percentile. Thus, for each model (and the observations), S99.9 measures the relative difference for each box to the European mean for that model, similar to Fig. 6 for the European sub-regions.

We employed the explained spatial variance (E) that is defined by van Ulden & van Oldenborgh (2006) as:

$$E = 1 - \frac{\sigma^2_{\mathrm{diff}}}{\sigma^2_{\mathrm{obs}}} \quad (2)$$

where σ^2_{diff} is the variance of the difference in S99.9 between the model and the observations, and σ^2_{obs} is the variance of S99.9 in the observations. This explained spatial variance measures not only the spatial correlation between observed and modelled S99.9, but also the amplitude of the spatial variations. A perfectly correlated field, but with an amplitude reduced by a factor of 2, would yield an E of 0.75. For S99.9, E is negative when the model deviates more from the observations than the observations deviate from 1 over all Europe. The spatial variance is first computed for each season separately, and then averaged to yield the alternative metric.

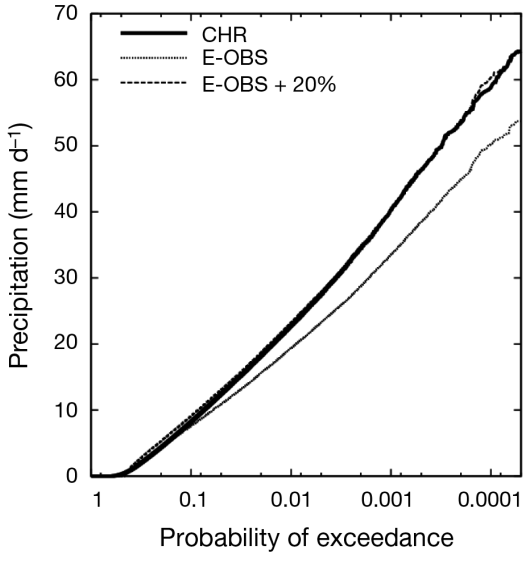

Fig. 9. Extreme statistics of daily precipitation for summer (JJA) using the International Rhine Commission (CHR) database, E-OBS and E-OBS increased by 20%. A probability of exceedance of 0.001 corresponds to the 99.9th percentile

Fig. 7d compares the metric based on explained spatial variance E with the standard metric. Because the alternative metric measures a rather different property of the 99.9th percentile field, the difference between both metrics is not surprising. However, it is reassuring that the 2 models with the lowest score in W also attain the lowest scores in E. The scores of the 2 metrics, however, are almost uncorrelated for the other models.

The values of E for the different models and seasons are shown in Fig. 10. Whereas with the standard metric, comparable scores are obtained for the different seasons, E is found to depend rather strongly on the season. High scores (on average between 0.8 and 0.9) are obtained for the winter season, but low scores (on average between 0.2 and 0.5) are obtained for the summer season. Scores for autumn and spring are slightly lower than for the winter season.

E computed from the relative spatial pattern (S99.9) is, by definition, identical to E from the absolute spatial pattern of the bias-corrected 99.9th percentile field (biases corrected with respect to the EU mean bias of P99.9 for the model compared with the observations). How well do the spatial patterns without bias correction match? Except those models with a low mean bias (M3, M13 and M14), which display equal or slightly lower scores, all other model get a much lower score, with 8 models having a score close to zero or negative. Models apparently simulate the relative differences within Europe better than the absolute differences.

6. QUALITY OF THE OBSERVATIONS AND STATION DENSITY

The E-OBS data set is, at the moment, the best long-term observational set available for Europe covering the period 1950–2008 at a daily time scale with high spatial resolution. As such, it is a very useful data set to evaluate the output of RCMs. Nevertheless, it is known that, in particular, the extremes in E-OBS are affected by station density (Haylock et al. 2008, Hofstra et al. 2009, 2010). For the construction of the E-OBS data set, approximately 2900 stations are used, whereas the number of grid cells in the 0.22° rotated model grid is approximately 16 000. This implies that many grid boxes do not contain any observations. Hofstra et al. (2010) conclude that extremes are significantly affected when station densities are low and that extremes in

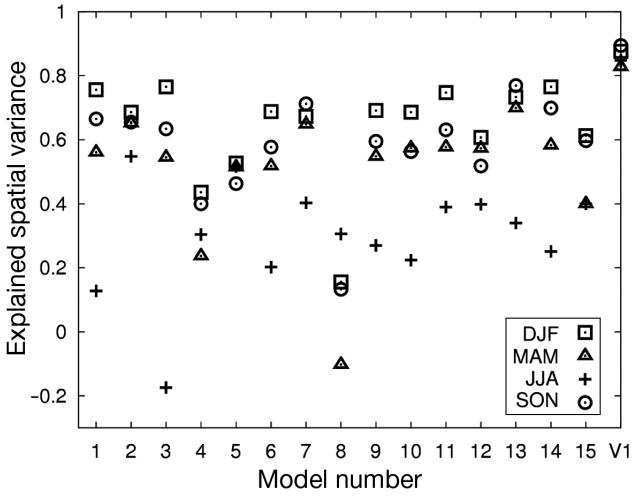

Fig. 10. Explained spatial variance of the 15 different regional climate models (see Table 1 for model information) and the first version of the E-OBS database (V1) for the different seasons

precipitation could be considerably underestimated (by 15 to 51% in the grid boxes considered). Hofstra et al. (2009) also found substantial differences between E-OBS and 3 other high-resolution data sets, in particular over topographically complex terrain such as the Alps. In agreement, we found differences between E-OBS and an alternative data set for the Rhine catchment area of approximately +20%. However, it is hard to generalize these statements because the distribution of stations is very inhomogeneous, and biases are therefore likely to be variable in space.

We compared the station density (number of stations used in the E-OBS data set per 2 × 2° box) and the mean bias averaged over all models and all seasons (Fig. 11). These 2 fields are clearly correlated. In areas with low station densities, the average bias in the model ensemble tends to be large (e.g. in France, Spain and Sweden), whereas biases in areas with high station densities tend to be lower (e.g. in The Netherlands and Ireland). An exception is the area south of the Alps, where both station density and bias are large.

Fig. 11 does not prove that the average of the models is more trustworthy than the observations, and the color coding has been chosen such as to emphasize the correspondence in spatial structure. However, it certainly emphasizes that observational errors might be substantial in areas with low station densities, and this needs to be considered in deriving weights. Therefore, structural biases seen in the RCMs may actually not be model errors, but could also be (in part) errors in the observational database.

7. DISCUSSION

7.1. Choice of the metric

The results above show that the model scores and ranking strongly depend on the metric chosen, even when based on the same index (here the 99.9th percentile). This rather trivial result is worrying as the choice of the metric is arguably rather subjective. However, it is reassuring that there are models that consistently score low with all different metrics. In addition, the first version of the observational database has consistently higher scores when evaluated against the second version for all models. Nevertheless, an important question is whether we can choose the metric on a more objective basis.

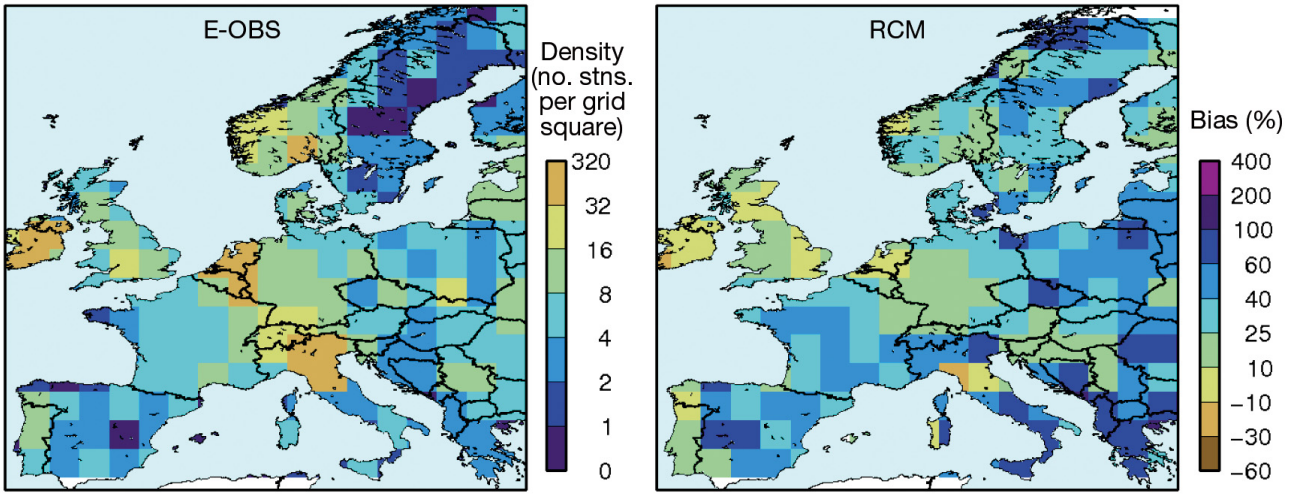

Fig. 11. Station density used in the construction of the E-OBS database in number of stations per 2 × 2° degree boxes (left) versus average model bias compared with E-OBS, averaged over all regional climate models (RCMs) and all seasons (right)

The ultimate purpose of metrics is to help reduce uncertainty in future climate predictions. It should facilitate the quantification of our confidence in the model simulations of the future climate. Therefore, a metric should relate directly or indirectly to the climate change signal. An indirect way to assure this to some extent is to consider the model performance over a range of different climate conditions. If a model performs well for different climate conditions in the present-day climate, we are more confident that it will also perform well for the future climate. This may imply that a model with a considerable but constant bias for different climate zones could in fact be more trustworthy than a model with, on average, no bias but with a misrepresentation of the differences between different climate zones. As such, the metric based on the explained spatial variance may be more trustworthy than the standard metric based on the bias.

A possible way to move forward is to first establish whether there are predictors or observables in the present-day climate that relate to the climate change signal. An understanding of the essential physical processes may help reveal these relationships, which could be explored making use of the ensemble of simulations of the present-day and future climate, available in the ENSEMBLES database. We note that in this sense it is unfortunate that there is no common climate change simulation in ENSEMBLES using the same GCM boundaries for all RCMs. If such a relationship between present-day climate observables and the climate change signal is established, the observables could be employed to construct a metric. Examples of such approaches are now emerging in the literature (e.g. Piani et al. 2005, Bony & Dufresne 2005). Another potential example of this approach is to explore the relationships between precipitation extremes, temperature and atmospheric moisture content for the present-day climate and for the climate change signal (e.g. Lenderink & van Meijgaard 2008, 2010).

Considering metrics for RCMs, another argument is often used. As RCMs are used as downscaling tools, spatial detail is an important issue. With the application of the RCMs in mind, measures of model performance therefore tend to be focused on small-scale structures. This reasoning guarantees, to some extent, that the model results can be used directly, or with relatively small corrections, in impact models. Related to this is the argument that metrics should measure the 'added value' of RCMs compared with coarse-resolution GCMs.

We clearly do not have definite answers to the question of how metrics should be chosen, but we aim to stimulate discussion with this paper. We would also like to draw attention to a recent paper by Knutti (2010) that describes many of these issues.

7.2. Weighting using metrics of model performance

In the final ENSEMBLES weighting system (Christensen et al. 2010), the metric is used to compute weighted average results. For this purpose, the final weights of each model are normalized such that the sum of weights of the RCMs is 1. Thus, the relative weights of all of the models enter the final weighting system. Our results display a ratio between the best and the worst model of 1.3 (Table 2).

The ratio between the weights of the best and worst model is subjective. By using different conversions from bias to weight, this ratio can be modified easily. For example, taking the second power of W, thereby giving larger deviations from the observations a larger penalty, gives a ratio of 1.7. The ratio can be used to express how well the metric can distinguish between good and bad models. Given the uncertainty in the observational data set, a small ratio of 1.3 seems appropriate.

8. CONCLUSIONS

We evaluated the simulation of extreme daily precipitation in an ensemble of 15 RCMs performed in the ENSEMBLES project (Hewitt & Griggs 2004, van der Linden & Mitchell 2009) against the recently developed E-OBS database (Haylock et al. 2008). The E-OBS database has been developed specially to represent daily spatially averaged observations, and covers the whole of Europe for the period 1950–2008 on a grid of ~25 km. The model simulations have been forced by 'perfect boundaries' from the ERA40 reanalysis project, and the time period 1971–2000 is used for evaluation. The observational and model data were first pooled in grid boxes of $2 \times 2°$, and then, from the pooled data, different percentiles of the distribution were computed for the 4 different seasons. The vast majority of the models considerably overestimated the extremes compared with E-OBS. For P99.9, the European mean bias is on average +38%, but ranges from –10 to +70% in the different models.

We proposed a simple metric of extreme precipitation to measure the model performance. A nonlinear function of the bias in the 99th to 99.99th percentiles was used to compute model scores (referred to as weights). Weights were computed for each member of the ensemble of model simulations. The final weights differed by a factor of 1.3 between the best and worst model. These weights have been used in the final ENSEMBLES weighting scheme, which is discussed in Christensen et al. (2010).

To highlight the explorative nature of this research, different sensitivity tests were performed. A sensitivity

test showed relatively low impacts of the spatial averaging and the combination of the seasonal weights to the annual weights. An alternative weight based on the spatial pattern of the extremes, however, resulted in large differences from the standard metric. Despite this, it is reassuring that the same 2 models yielded the lowest score for both metrics, thus implying that models that perform consistently worse than the other models can be identified.

The quality of the observational data is an important issue. Although E-OBS is a considerable advance in the availability of observational data for Europe, it is known that the extremes in E-OBS could be biased. For example, Hofstra et al. (2010) showed substantial biases in E-OBS for areas where the underlying station density is low. In agreement, our results showed a remarkable correspondence between the patterns of the mean bias in the model ensemble and the station density, with, on average, low biases where the station density is high and high biases where the station density is low. A sensitivity test showed that a potential bias in the observations of +20% could turn the model with the highest score into a model with one of the lowest scores.

Considering the potential biases in the observational data and the high degree of subjectivity in constructing metrics, at the moment it does not seem appropriate to discriminate too much based on the proposed metric(s). Nevertheless, comparing the RCMs with the observations by means of these metrics revealed important insights in the behavior and performance of models and the quality of the observations.

Acknowledgements. Financial support by the EU FP6 Integrated Project ENSEMBLES (contract no. 505539) and the Dutch Climate change and Spatial Planning program (CcSP) and Knowledge for Climate (KfC) is gratefully acknowledged. J. Attema and E. van Meijgaard are thanked for careful proofreading, and 3 reviewers for their comments on an earlier version of this paper.

LITERATURE CITED

Bony S, Dufresne JL (2005) Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models. Geophys Res Lett 32:L20806 doi:10.1029/2005GL023851

Chen CT, Knutson T (2008) On the verification and comparison of extreme rainfall indices from climate models. J Clim 21: 1605–1621

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

Christensen JH, Kjellström E, Giorgi F, Lenderink G, Rummukainen M (2010) Weight assignment in regional climate models. Clim Res 44:179–194

Fowler HJ, Ekström M, Blenkinsop S, Smith AP (2007) Estimating change in extreme European precipitation using a multimodel ensemble. J Geophys Res 112:D18104 doi:10.1029/2007JD008619

Frei C, Schöll R, Fukutome S, Schmidli J, Vidale P (2006) Future change of precipitation extremes in Europe: intercomparison of scenarios from regional climate models. J Geophys Res 111:D06105 doi:10.1029/2005JD005965

Guichard F, Petch JC, Redelsperger JL, Bechtold P and others (2004) Modelling the diurnal cycle of deep precipitating convection over land with cloud-resolving models and single-column models. Q J R Met Soc 130:3139–3172

Haylock MR, Hofstra N, Klein Tank AMG, Klok EJ, Jones PD, New MA (2008) European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. J Geophys Res 113:D20119 doi:10.1029/2008JD010201

Hewitt C, Griggs D (2004) Ensembles-based predictions of climate changes and their impacts (ENSEMBLES). EOS 85:566 doi:10.1029/2005EO520005

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

Hofstra N, New M, McSweeney C (2010) The influence of interpolation and station network density on the distribution and extreme trends of climate variables in gridded data. Clim Dyn 35:841–858

Hohenegger C, Brockhaus P, Bretherton CS, Schär C (2009) The soil moisture-precipitation feedback in simulations with explicit and parameterized convection. J Clim 22: 5003–5020

Knutti R (2010) The end of model democracy? An editorial comment. Clim Change 102:395–404

Krahe P, Eberle M, Carambia M, Buitefeld H, Wilke K (2010) A hydrometeorological reference data set for the river Rhine (CHR_OBS). BfG-Berichte. Federal Institute of Hydrology, Koblenz

Leander R, Buishand TA (2007) Resampling of regional climate model output for the simulation of extreme river flows. J Hydrol 332:487–496

Lenderink G, van Meijgaard E (2008) Increase in hourly precipitation extremes beyond expectations from temperature changes. Nat Geosci 1:511–514

Lenderink G, van Meijgaard E (2010) Linking increases in hourly precipitation extremes to atmospheric temperature and moisture changes. Environ Res Lett 5:025208

Lenderink G, van Ulden A, van den Hurk B, Keller F (2007) A study on combining global and regional climate model results for generating climate scenarios of temperature and precipitation for the Netherlands. Clim Dyn 29:157–176

O'Gorman PA, Schneider T (2009) The physical basis for increases in precipitation extremes in simulations of 21st-century climate change. Proc Natl Acad Sci USA 106: 14773–14777

Overeem A, Buishand TA, Holleman I (2008) Rainfall depth-duration-frequency curves and their uncertainties. J Hydrol (Amst) 348:124–134

Overeem A, Buishand TA, Holleman I (2009) Extreme rainfall analysis and estimation of depth-duration-frequency curves using weather radar. Water Resour Res 45:W10424 doi:10.1029/2009WR007869

Pall P, Allen M, Stone D (2007) Testing the Clausius–Capeyron constraint on changes in extreme precipitation under CO_2 warming. Clim Dyn 28:351–363

Piani C, Frame DJ, Stainforth DA, Allen MR (2005) Constraints on climate change from a multi-thousand member ensemble of simulations. Geophys Res Lett 32:L23825 doi: 10.1029/2005GL024452

Sanchez-Gomez E, Somot S, Déqué M (2009) 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

Uppala S, Kallberg P, Simmons A, 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, FitzRoy Road, Exeter

van Ulden AP, van Oldenborgh G (2006) Large-scale atmospheric circulation biases in global climate model simulations and their importance for climate change in Central Europe. Atmos Chem Phys 6:863–881

Submitted: October 29, 2009; Accepted: October 6, 2010

Proofs received from author(s): November 30, 2010