Vol. 44: 195–209, 2010
doi: 10.3354/cr00866

CLIMATE RESEARCH
Clim Res

Published December 10

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

OPEN ACCESS

Weighted frequency distributions express modelling uncertainties in the ENSEMBLES regional climate experiments

M. Déqué*, S. Somot

Météo-France, Centre National de Recherches Météorologiques, CNRS/GAME, 42 Avenue Coriolis, 31057 Toulouse Cedex 01, France

ABSTRACT: Fourteen regional climate models (RCMs) were driven by general circulation models (GCMs) in FP6-ENSEMBLES to provide 17 fine-scale (25 km) climate change scenarios for the period 2021–2050. In a preliminary exercise, these RCMs were driven by gridded observations (ERA40 reanalysis) to simulate as accurately as possible the 1961–2000 period. The quality of this reproduction was used to calculate a weight for each model. Each individual model climate had an uncertainty due to the finite sampling (30 yr). These spreads were combined by those weights to produce an ensemble uncertainty. We illustrate here the daily and climatological frequency distributions for winter and summer temperature and precipitation in 3 European cities (Budapest, Dublin and Lisbon). The distribution obtained by ENSEMBLES weights was compared with a distribution using equal weights, distributions using random weights and distributions based on a single model. As far as the reproduction of the observed distribution (1961–1990) is concerned, there is no evidence that the ENSEMBLES weight system provides results closer to observation than equal weights or weights drawn at random. A single model taken at random yields a quality score not better than ENSEMBLES in the case of precipitation, and worse than ENSEMBLES in the case of temperature. As far as climate change for 2021–2050 is concerned, the use of ENSEMBLES weights instead of equal weights also leads to a similar response at daily as well as 30 yr mean time scales.

KEY WORDS: Temperature · Precipitation · Frequency distribution · Ensemble · Scenario

1. INTRODUCTION

As climate has changed in the last century, and as the cause of this change has been clearly identified (IPCC 2007), robust indices of these changes have been developed, both for observations and for expectations from numerical model scenarios. The easiest to understand and, therefore, the most widespread is the time average. However, human society and natural ecosystems are more sensitive to changes to certain extremes, which act as limiting factors. For example, Sun et al. (2007) showed that an increase in greenhouse gas (GHG) concentrations would lead to an intensification of extreme precipitation in many parts of the globe.

A probability density function (PDF) is a function that describes the probability of a continuous random variable, such as precipitation, occurring at a given point in time. The term 'PDF' encompasses 2 different concepts—(1) a statistical description of a variable: a climate change in the tail of the distribution may be different from a change in the mean; (2) the uncertainty about the climate change itself: different climate models produce different expectations of the future climate, and there is no way to decide which model tells the truth. Frei et al. (2003) evaluated the capacity of a few European models to simulate the PDF of precipitation, according to the former meaning of this term. More recently, the PRUDENCE project (Christensen et al. 2002) offered a database of regional projections over Europe which was exploited by Boberg et al. (2009) to derive climate projections for the PDF of precipitation. However, the PDF described in this study did not introduce

*Email: deque@meteo.fr

the uncertainty about the choice of a model within a population of available experiments. Buser et al. (2009) used a Bayesian approach to derive a multi-model PDF from PRUDENCE data, according to Concept (2), above. In the domain of seasonal prediction, it is usual to build a PDF from several model results (e.g. Doblas-Reyes et al. 2000). In the present study, our use of 'PDF' is based on the assumptions in Concept (2) (above), because the predictand is most often a deterministic quantity, i.e. the observed seasonal mean of a particular year. In addition, we have the possibility to calibrate and validate a probability approach in seasonal forecasting, because we have about 50 independent past cases for verification, which is not the case for a probability approach in climate change.

One of the new features of the ENSEMBLES project (Hewitt & Griggs 2004) with respect to the PRUDENCE project is the ability to construct probability distributions based on a non-uniform weighting of the models, using the quality of their reproduction of the present climate. To this purpose, a simulation of 1961–2000 was added to the traditional reference/scenario pair. This simulation is not exploited here, but is used in the other papers of this Special to rate the different regional climate models (RCMs). In the present study we used only the weights from this validation experiment. The weighting approach offers an alternative approach to the traditional Bayesian one (e.g. Tebaldi et al. 2005), and has been introduced in climate change studies by Giorgi & Mearns (2002). The aim of the present study was not to compare the 2 approaches, but to focus on the sensitivity of the combined distributions to the choice of the weights.

Specifically, we calculated a distribution for daily temperature and precipitation of the reference simulation and compared it to the observed one. This comparison was extended to the case of distributions based on a single model. We then compared the distributions of daily values in the reference and scenario simulations, and calculated a bivariate frequency distribution for the temperature/precipitation climate change.

2. AVAILABLE DATA

We used 3 kinds of data from the ENSEMBLES project: the weights from the ERA40-driven RCM simulations, the gridded observations (Haylock et al. 2008) and the general circulation model (GCM)-driven RCM simulations. We used 14 RCMs referred to by the name of their institutes:

- C4I (Jones et al. 2004), Community Climate Consortium for Ireland
- CNRM (Radu et al. 2008), Météo-France Research Center
- DMI (Christensen et al. 1996), Danish Meteorological Institute
- ETHZ (Böhm et al. 2006), Swiss Federal Institute of Technology in Zurich
- HadC (Collins et al. 2006), Met Office Hadley Centre, UK
- ICTP (Giorgi & Mearns 1999), International Center for Theoretical Physics in Trieste, Italy
- KNMI (Lenderink et al. 2003), Dutch Meteorological Service
- METN (Haugen & Haakensatd 2006), Norwegian Meteorological Service
- MPI (Jacob 2001), Max Planck Institute for Meteorology in Hamburg, Germany
- OURA (Plummer et al. 2006), Ouranos Consortium in Montreal, Canada
- SMHI (Kjellström et al. 2005), Swedish Meteorological and Hydrological Institute
- UCLM (Sanchez et al. 2004), University of Toledo, Spain

The total number of RCMs is not 12 but 14, because HadC corresponds to 3 versions: HC-lo, HC-med and HC-hi. These 3 models share the same dynamics and a very similar description of the sub-grid processes. They have been produced by arbitrary perturbations of a large set of model parameters which lead to very different responses to GHG concentration in their GCM version (Murphy et al. 2007).

The RCMs all have approximately the same integration domain widely covering Europe (from Iceland to Cyprus, and more to avoid the lateral boundary undesirable numerical effects) and the same horizontal resolution of 25 km. The ERA40-driven RCMs have been run from 1961 to 2000 (some of them longer). The GCM-driven RCMs have been run from 1950 to 2050 (some of them to 2100). The driving GCMs, referred to by the name of their institutes, are:

- BCM (Furevik et al. 2003), University of Bergen, Norway
- CGCM (Scinocca et al. 2008), Canadian Meteorological Service
- CNRM (Gibelin & Déqué 2003), the global version of CNRM RCM with variable resolution; this atmospheric GCM takes its sea surface temperature data from CNRM-CM3 (Salas-Mélia 2002)
- HadC (Gordon et al. 2000), the global version of HadC RCM; 3 driving runs are available (HC-lo, HC-med and HC-hi)
- MPI (Roeckner et al. 2003), the global version of MPI RCM

As for the RCMs, we considered the 3 versions of the HadC model (HC-lo, HC-med and HC-hi) as 3 different GCMs because their climate responses are very different by construction. Table 1 indicates for each

Table 1. Regional climate model (RCM) × general circulation model (GCM) matrix. An X indicates that the corresponding RCM × GCM pair was available in ENSEMBLES at the time of the study. The last column indicates the weight attributed to each RCM and the last row the weight attributed to each GCM. See Section 2 for RCM and GCM details

RCM	GCM BCM	CGCM	CNRM	HC-lo	HC-med	HC-hi	MPI	Weight
C4I						X		0.069
CNRM			X					0.062
DMI			X				X	0.063
ETHZ					X			0.067
HC-lo				X				0.045
HC-med					X			0.062
HC-hi						X		0.069
ICTP							X	0.076
KNMI							X	0.093
METN	X							0.061
MPI							X	0.072
OURA		X						0.060
SMHI	X			X			X	0.079
UCLM					X			0.059
Weight	0.069	0.093	0.069	0.045	0.063	0.081	0.081	

RCM which GCM has been used. There are a total of 17 GCM-driven simulations.

If we consider that the modelling uncertainties of the RCMs are fairly represented by 12 models coming from most European countries, this is not the case with GCMs. Five GCMs (of which BCM and CNRM share the same atmospheric component) is much less than the 23 GCMs used in the IPCC Fourth Assessment Report (AR4); however, there are more degrees of freedom in a GCM than in an RCM. In addition, the anthropogenic radiative forcing is based on IPCC Special Report on Emissions Scenarios A1B concentrations for all models, neglecting the carbon cycle feedback. We must keep this point in mind when building a probabilistic framework below.

In the present study we have limited our analyses to winter (DJF) and summer (JJA) screen level temperature and precipitation. In the case of daily distribution of precipitation, it is important to avoid spatial averages. It is impossible to display or analyze the approximately 50 000 grid points of each model. Therefore, we have restricted the analysis to 3 locations—Dublin, Lisbon and Budapest—on the basis of the nearest grid point. There is indeed in Europe a north–south gradient in precipitation and JJA temperature and a west–east gradient in DSF temperature in the typical response to anthropogenic climate change (e.g. Jacob et al. 2007 for PRUDENCE). By choosing these locations, which also have rather different climates (oceanic, Mediterranean and continental, respectively) we hope to show a panel of European situations. In practice, we have taken the nearest grid point for each RCM and for the 0.25° observation grid. In Déqué (2009), the ENSEMBLES-based local climate changes are displayed for 34 European capitals and 4 seasons.

3. AN ENSEMBLES-BASED PROBABILITY DISTRIBUTION FOR THE PRESENT CLIMATE

If we keep in mind that the GCM sampling includes only part of the uncertainties about modelling and that the weights have not been calculated in a rigorous probabilistic framework as proposed by Rougier (2007), we can build a simple probabilistic model which represents the spread of the results in pooled ENSEMBLES simulations.

For each of the 17 models and each season (DJF and JJA), a PDF based on bin counting was calculated for temperature (resolution = 0.5°C) and precipitation (resolution = 0.25 mm d^{-1}) using the daily values of the reference period 1961–1990 in the GCM-driven experiment. For precipitation, we calculated the probability of a dry day (precipitation < 1 mm d^{-1}) and the conditional probability of heavy rain (precipitation > 20 mm d^{-1}). This separation avoids giving too much importance to the weak precipitation values (which represent more than 90% of the data set in JJA in Lisbon). The temperature and precipitation PDFs represent the probability law of this quantity for a day drawn at random in the series produced by a given model. If we consider now the ENSEMBLES multi-model by pooling the different results, the probability of obtaining a value x for temperature or precipitation is given by the density:

$$\mathrm{PDF}(x) = \sum_{i=1}^{17} \mathrm{PDF}_i(x) p_i \quad (1)$$

where $\mathrm{PDF}_i(x)$ is the PDF of the variable x for model i, and p_i is the probability of this model. The simplest probability is to consider that the ENSEMBLES models are equiprobable, then $p_i = 1/17$. A more complex system consists of using the weights calculated with ERA40-driven RCM data (see the other papers in this Special). These weights are given in last column of Table 1. Although the ENSEMBLES project did not produce an official weighting system for the GCMs, it seems necessary to introduce one here in our sensitivity study. Indeed, SMHI and DMI RCMs appear several times with different GCMs, whereas the other RCMs appear only once in our multi-model. This means that a comparison with equal weights would cause bias. The choice of the GCM weights here was dictated by the

wish to favor the GCMs which propose the best large-scale patterns over the RCM common domain. The GCM weights do not include any information on model performance derived from outside the study region (e.g. the skill of their simulations of the El Niño Southern Oscillation or the Indian monsoon). The method is the same as used for the evaluation of RCMs in ENSEMBLES ERA40-driven simulations by Sanchez-Gomez et al. (2009). For each calendar season, a cluster analysis is applied to ERA40 500 hPa height daily data over the Europe-Atlantic domain. Four patterns (weather regimes) are identified. Then, each day of each simulation was attributed to a regime by a spatial distance criterion. For each calendar season of each model we have thus 4 frequencies which are compared with the corresponding 4 ERA40 frequencies. A score based on the mean square differences (Brier score) is calculated. Finally, the scores are transformed into weights by a linear decreasing function, so that we get the same range of weights for the GCMs as for the RCMs (last line of Table 1). Hereafter, ENSEMBLES weights will refer to the product of the GCM weights by the RCM weights.

Fig. 1 shows the temperature multi-model PDF (thick dashed line) for Budapest, Dublin and Lisbon in DJF and JJA, as well as the 17 individual models (thin lines). To improve readability, the pooled PDF curves have been smoothed by Gaussian kernels (see Déqué 2009). The standard deviation of the kernels is 0.5°C for temperature and 1 mm d^{-1} for precipitation. This choice is an empirical compromise between graphics where all individual models produce local density maxima in the pooled PDF and graphics which appear too bell-shaped.

In Budapest, the multi-model PDF has more spread than the observation PDF, particularly in JJA. This behavior is expected, because the multi-model ensemble introduces different biases, some models being too cold, some being too warm. The multi-model ensemble is warmer than the observation. In Dublin, the multi-model fits the observed PDF rather well. In JJA, the multi-model underestimates the spread: most individual models underestimate variability and have a small bias, contrary to Budapest. In Lisbon, the multi-model again overestimates the spread, as in Budapest.

The PDF obtained with equal weights (data not shown) is very close to the dashed curve of Fig. 1. In addition to the ENSEMBLES and equal weight systems, one can imagine many other systems by drawing the 17 weights at random with a positive constraint. In the present study, the weights are generated by independent uniform random numbers in the 0 to 1 range. The weights are then normalized to yield the 17 probabilities p_i. The shaded area in Fig. 1 corresponds to a 95% envelope of the multi-model PDFs. It is constructed by taking, in each bin, the 25th and the 975th temperature probabilities out of 1000 weightings. The same kernel filter as above is applied.

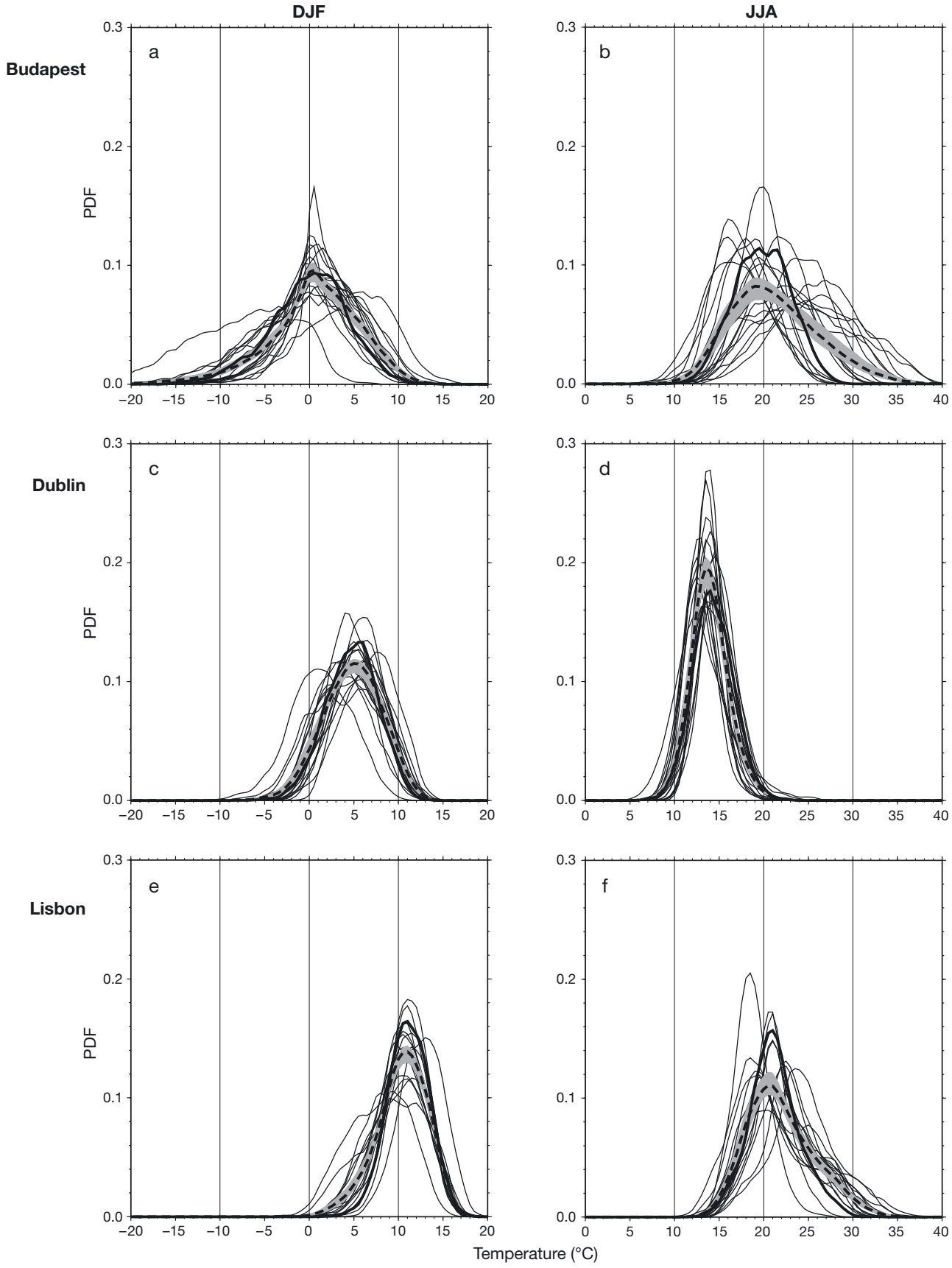
Table 2. Daily probability of receiving <1 mm precipitation (i.e. probability of dry day) for the reference 1961–1990 period based on: ENSEMBLES weights (ENS), equal weights (equal), 2.5% quantile of random weights (R2.5%), 97.5% quantile of random weights (R97.5%) and observation (obs). DJF: December–February; JJA: June–August

City and season	Probability of dry day				
	ENS	Equal	R2.5%	R97.5%	Obs
Budapest					
DJF	0.657	0.664	0.645	0.684	0.754
JJA	0.725	0.726	0.697	0.754	0.708
Dublin					
DJF	0.506	0.518	0.500	0.537	0.526
JJA	0.498	0.506	0.473	0.540	0.591
Lisbon					
DJF	0.622	0.632	0.604	0.661	0.616
JJA	0.943	0.942	0.932	0.951	0.945

As far as precipitation is concerned, Table 2 gives the probability of dry days (i.e. precipitation < 1 mm). The ENSEMBLES and equal weight systems are almost equivalent. In Budapest, they invert the annual cycle with respect to observation (more dry days in JJA than in DJF with the multi-model). In Dublin, the multi-model is too wet in JJA. In Lisbon it fits the observation well. Except in Budapest in DJF and Dublin in JJA, the observed probability is inside the 95% interval of random weights. This indicates that in those 2 cases all models have a bias of the same sign, and thus no positive combination can correct it.

Fig. 2 shows the PDF of precipitation for rainy days. Here again, the 2 weight systems are almost equivalent and, therefore, equal weights are not shown. The model agrees with observations in Budapest and Dublin, but produces too much low precipitation (1 to 3 mm) and too little moderate precipitation (4 to 12 mm) in Lisbon. This figure is not suitable for evaluating extreme precipitations events, but Boberg et al. (in press) showed that ENSEMBLES models are satisfactory in this respect.

Fig. 1. Probability density functions (PDFs) of daily temperature for the 3 cities in winter (DJF) and summer (JJA) based on ENSEMBLES weights (dashed line) and observation (thick solid line) in the reference period 1961–1990. Grey area: 95% range for the PDF when combining the models with random weights. Thin solid lines: the 17 individual models


DJF
JJA
Budapest
Dublin
Lisbon
a
b
c
d
e
f
PDF
Temperature (°C)

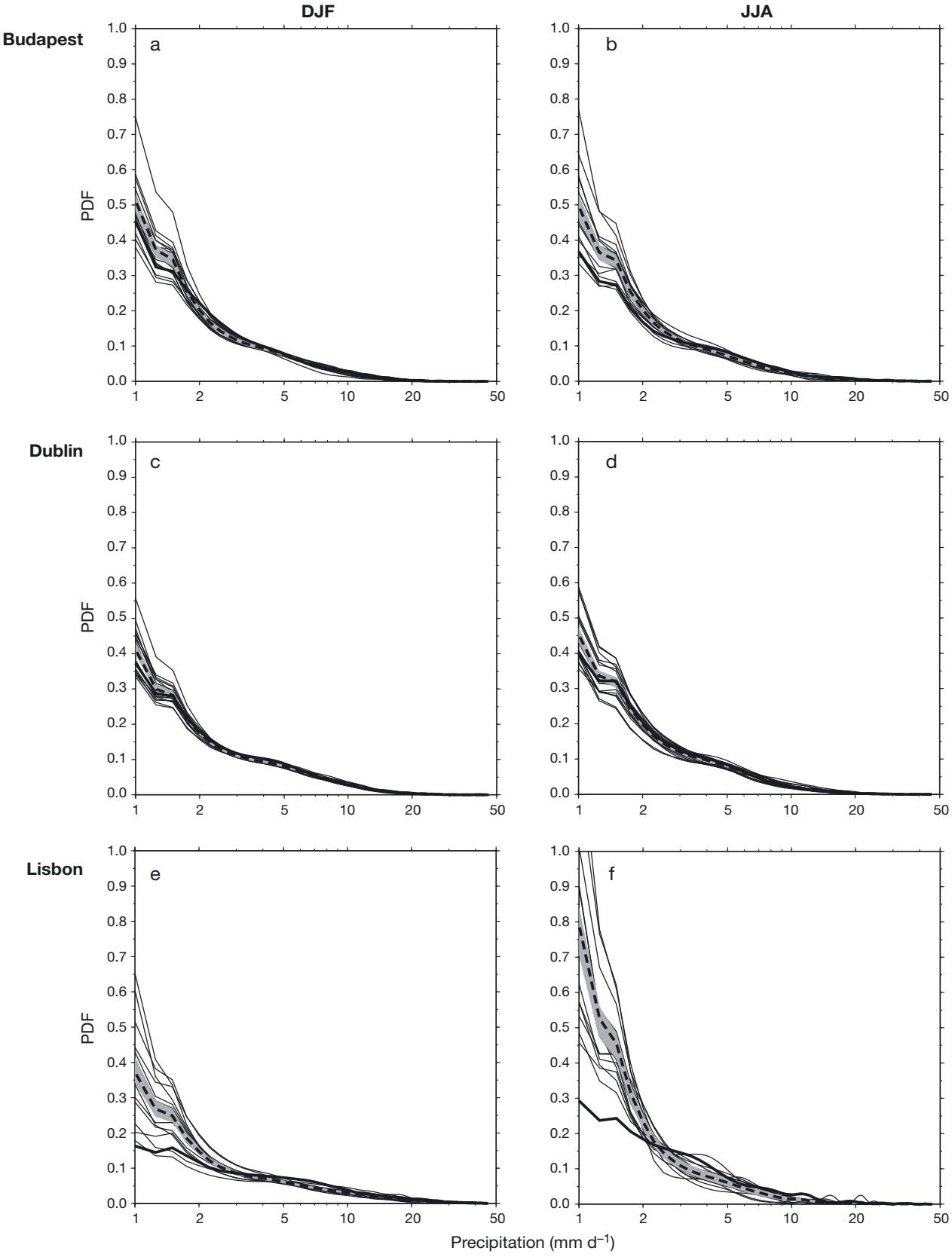


Fig. 2. PDFs of daily precipitation for the 3 cities in winter (DJF) and summer (JJA). Other details as in Fig. 1

Table 3. Ranked probability score (RPS) between model and observation for daily temperature in the 1961–1990 reference period with various weight systems: ENSEMBLES (ENS), equal and random (R). With random weights, the table proposes the quantiles R2.5%, R50% and R97.5% as well as the absolute minimum (R_{min}). When a single model is considered (last 3 columns), the best, median and worst correspond to the minimum, 9th ranked and maximum RPS amongst the 17 climate change scenarios. DJF: December–February; JJA: June–August

City and season	Ranked probability score								
	ENS	Equal	R2.5%	R50%	R97.5%	R_{min}	Best	Median	Worst
Budapest									
DJF	0.0034	0.0023	0.0008	0.0024	0.0065	0.0001	0.0001	0.0095	0.1272
JJA	0.0102	0.0128	0.0051	0.0128	0.0276	0.0002	0.0034	0.0477	0.1745
Dublin									
DJF	0.0008	0.0008	0.0006	0.0019	0.0022	0.0001	0.0011	0.0054	0.0868
JJA	0.0036	0.0043	0.0019	0.0045	0.0084	0.0001	0.0005	0.0094	0.0510
Lisbon									
DJF	0.0048	0.0052	0.0024	0.0052	0.0095	0.0001	0.0004	0.0053	0.0866
JJA	0.0050	0.0046	0.0026	0.0042	0.0092	0.0001	0.0004	0.0098	0.0893

It is possible to go further in the evaluation of the PDF by introducing a quantitative criterion of similarity to the observation PDF. The most natural one would be the quadratic difference (or Euclidean distance) between the PDFs. However, this index has 2 drawbacks: (1) if the distance is calculated on many intervals, the PDF is poorly sampled and exhibits many peaks which artificially increase the distance, unless using a filter as in Figs. 1 & 2; (2) if the 2 PDFs have a small dispersion and their means are different, the distance is almost independent of the difference between the means, which is not desirable.

For these reasons, it is a common practice in long-range probabilistic forecasting to use the ranked probability score (RPS; Epstein 1969), which is the quadratic distance between the cumulative density functions (CDFs):

$$\mathrm{RPS} = \frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{CDF_m}(t_i) - \mathrm{CDF_o}(t_i)\right]^2 \qquad (2)$$

where m is model, t_i are given thresholds, generally quantiles of the observed distribution, o is observation, and:

$$\mathrm{CDF_m}(x) = \mathrm{Prob}(\mathrm{model} < x);$$
$$\mathrm{CDF_o}(x) = \mathrm{Prob}(\mathrm{observation} < x) \qquad (3)$$

where $\bar{x}$ is a dummy variable.

The continuous form of Eq. (2) can be written as an integral, but with n = 100 and t_i the centiles of observed distribution, the RPS is close to its asymptote.

Table 3 shows the RPS for the temperature probability distribution. The values are squared differences between probabilities. They are of order 10^{-2} with large variations. The calculation is done without Gaussian kernel filtering because the CDF has—contrary to the PDF—smooth variations due to the constraint of being an increasing function bounded by 0 and 1. The results are in agreement with the visual analysis of Fig. 1 with excellent scores for Dublin in DJF.

The 2.5 and 97.5% quantiles of the corresponding RPS distribution provide a 95% interval for random weights. These quantiles are estimated with the same 1000 Monte-Carlo drawings of the weights as for the quantiles used in Figs. 1 & 2 (shaded area). As expected, the equal weights provide an RPS inside this 95% interval. This is also the case with the ENSEMBLES weights, which shows that the ENSEMBLES weighting system is not significantly better than a random choice of weights. In addition, the median (50% quantile) of the RPS distribution is provided. Except for Dublin in DJF, this median is close to the RPS with equal weights.

Since the RPS is a Euclidean distance, it is also possible to compute (linear regression with constraints) the weights which minimize the RPS in each case. The 7th column of Table 3 indicates that this minimum RPS (R_{min}) is very small, in fact much smaller than one can expect from a standard random drawing of the weights. This shows that a simple linear combination of model data with positive numbers (the solution is trivial with positive and negative numbers) can reproduce the statistical properties (mean, spread) of an observed series. Of course, if all models were too cold or too warm somewhere, it would not be possible to correct this bias by any linear combination.

Table 4 shows the various RPSs of the precipitation distribution, according to the various weight systems. The score is particularly good in Dublin and in Budapest (DJF only), but is poor in Lisbon in JJA. However, the ENSEMBLES weights do not perform better than a random weight system. This suggests that using more models is more efficient, in an RPS sense, than discriminating the best models. However, the minimum RPS amongst all possible random weights is small in Budapest and Dublin. In Lisbon in JJA, no linear combination of the models is able to produce a PDF very close to the observed one, as all models underestimate high precipitation.

Table 4. RPS between model and observation for daily precipitation in the 1961–1990 reference period. Other details as in Table 3

City and season	Ranked probability score								
	ENS	Equal	R2.5%	R50%	R97.5%	R_{min}	Best	Median	Worst
Budapest									
DJF	0.0003	0.0003	0.0003	0.0004	0.0005	0.0001	0.0001	0.0013	0.0156
JJA	0.0041	0.0039	0.0024	0.0040	0.0058	0.0001	0.0004	0.0053	0.0221
Dublin									
DJF	0.0002	0.0002	0.0001	0.0002	0.0003	0.0001	0.0001	0.0006	0.0039
JJA	0.0001	0.0001	0.0001	0.0002	0.0003	0.0001	0.0001	0.0011	0.0082
Lisbon									
DJF	0.0050	0.0052	0.0032	0.0053	0.0079	0.0003	0.0014	0.0058	0.0321
JJA	0.0277	0.0261	0.0203	0.0275	0.0329	0.0043	0.0044	0.0249	0.1003

4. ENTROPY OF THE WEIGHTING

In the previous section we compared the ENSEMBLES weights with equal weights, which is a natural way to combine the models. The fact that a weighting is even can be measured by the weighted average of the logarithm of the weights, with opposite sign:

$$S = -\sum_{i=1}^{n} p_i \ln(p_i) \quad (4)$$

This quantity, named entropy, is maximum with equal weights (the maximum is ln[n], here 2.83) and minimum (0) when one weight is 1 and the others are 0. The ENSEMBLES distribution has a high entropy, 2.80, close to the maximum. In the random choice of weights examined above, the 1000 weightings have an entropy ranging between 2.18 and 2.83. The weights which minimize the RPS have a lower entropy, ranging between 0.30 and 1.59.

The fact that our random weightings are rather even leads us to consider the most uneven weightings, which are those based on a single model. To avoid multiplying tables and to facilitate the comparisons, we have added 3 columns to Tables 3 & 4. The first of these corresponds to the best model, i.e. the model for which the RPS of the daily PDF of a given variable, at a given location and for a given season, is the minimum of the 17 RPSs. The second corresponds to the median model, i.e. the 9th model when sorting the 17 RPSs. The third corresponds to the worst model, i.e. the maximum RPS. The best and worst models give a range for the RPS obtained when a single model is considered to build the PDF. The median model is more meaningful than the averaged RPS. Indeed, the spread of the 17 RPSs is large, and the average would be dominated by the 1 or 2 largest RPSs. This median model gives a reasonable hint of what we could get if we take a single model at random.

Table 3 shows that taking the best model is better than a random weighting, except in Dublin in DJF. In Budapest in DJF, the best model has an RPS close to the absolute minimum. The worst model is far beyond the interval of the random weights. The median model is also beyond this interval, except in Lisbon: in DJF at this location, the median model has the same RPS as equal weights, but in JJA it is worse. So, considering a single model is generally worse than using an ensemble with weights, as far as the reproduction of the observed PDF is concerned.

For precipitation (Table 4), the best model is always close to the absolute minimum RPS, except in Lisbon in DJF. The median model stays within the 95% interval of random weightings in Budapest (JJA) and Lisbon (DJF and JJA). In the other cases it is beyond this range, but remains relatively small. So, if we exclude the few bad models, considering a single model for precipitation is less detrimental to the reproduction of observed daily PDF than it is for temperature.

In order to further examine the relation between entropy and RPS, we calculated the correlation between entropy and the RPS. We expanded the random sample to 50 000 weightings to increase our confidence. The correlation is always negative (Table 5), albeit weak, a result confirmed in the other seasons (data not shown). This indicates that an uneven distribution of weights has less chance to fit the observed PDF. This is a classical result in statistics: when one ignores the exact solution, the mean square error is minimized by predicting the mean than to predict one particular case selected at random.

Table 5. Correlation between the entropy of the weights and the ranked probability score of daily probability distribution functions versus observation in the case of random weightings (10 000 samples)

City and season	Temperature	Precipitation
Budapest		
DJF	−0.08	−0.25
JJA	−0.06	−0.02
Dublin		
DJF	−0.23	−0.13
JJA	−0.03	−0.20
Lisbon		
DJF	−0.04	−0.03
JJA	−0.08	−0.01

5. DAILY PROBABILITY DISTRIBUTION FOR REFERENCE AND SCENARIO

The main difference between the 2021–2050 and 1961–1990 periods is an increase in GHG concentrations, resulting in an increase in temperature. Fig. 3 shows the PDF of daily temperature in the 3 locations with ENSEMBLES weights (thick lines). At first glance, the shape of the PDF is not largely deformed and the climate change manifests itself as a shift of the PDF towards warm values. If we use equal weights (data not shown) instead of ENSEMBLES weights, the resulting figure is identical to Fig. 3. A slight flattening of the PDFs is seen in this scenario, except in DJF in Dublin and Lisbon. If we consider the median model introduced in Section 4 (thin lines), the same conclusions hold, except in Budapest in JJA where the scenario PDF becomes skewed. The single model PDF also has less dispersion than the weighted PDF, except in Budapest. Note that this model is not the same for the different seasons and locations, and the features might have been different with other choices of the model. The aim of the thin lines is to illustrate, not to analyze.

In the case of precipitation, plotting the PDFs, as in Fig. 2, does not allow us to capture the climate change, since the curves are too close to each other. Table 6 provides numerical values that are easier to compare (some data from Table 2 is repeated for comparison). In Budapest and Dublin, the probability of dry days decreases in DJF and increases in JJA. In Lisbon, this probability increases in both seasons. The 2 weight systems are in full agreement on this result. In the case of heavy rain (>20 mm d^{-1}) the probability increases in all cases, except in Lisbon in JJA where both weightings of the models lead to a decrease in probability. Table 6 also includes the probability for the median model. As far as dry days are concerned, there is an agreement with the multi-model, except in Lisbon in DJF. As far as heavy rains are concerned, there is also an agreement except in Lisbon.

Table 6. Probability of a dry day (precipitation < 1 mm) and heavy rain (precipitation > 20 mm) in winter (DJF) and summer (JJA) for reference (R, 1961–1990) and scenario (S, 2021–2050) periods with ENSEMBLES weights (ENS), equal weights (equal) and a single model (median). DJF: December–February; JJA: June–August

City, season and period	Dry day			Heavy rain		
	ENS	Equal	Median	ENS	Equal	Median
Budapest						
DJF						
R	0.657	0.664	0.745	0.015	0.015	0.016
S	0.657	0.663	0.734	0.018	0.019	0.033
JJA						
R	0.725	0.726	0.890	0.029	0.029	0.010
S	0.738	0.742	0.894	0.036	0.036	0.024
Dublin						
DJF						
R	0.506	0.518	0.507	0.020	0.021	0.026
S	0.502	0.512	0.504	0.024	0.024	0.031
JJA						
R	0.498	0.506	0.532	0.019	0.020	0.019
S	0.520	0.529	0.575	0.024	0.024	0.027
Lisbon						
DJF						
R	0.622	0.632	0.775	0.102	0.101	0.124
S	0.639	0.648	0.755	0.111	0.112	0.103
JJA						
R	0.943	0.942	0.918	0.016	0.016	0.003
S	0.954	0.954	0.945	0.014	0.013	0.007

6. WEIGHTED RESPONSE TO CLIMATE CHANGE

In the previous section we examined the reference and scenario simulations separately. Here we will address directly the differences between the scenario and reference simulations, also known as the climate response. At the daily scale, looking at the distribution of the differences between data in the scenario run and data in the control run is not relevant because the day-to-day variability is much larger than the pure climate change. At the seasonal scale, relevance is increased, but the year-to-year variability still dominates. We will work here at the climate scale, using the traditional definition of climate: an average over 30 consecutive years. To calculate the frequency distribution of the difference between scenario and reference climates, we use the same formula as in Eq. (1). Here $PDF_i(x)$ is replaced by the distribution for model *i*. Ideally, this distribution should be calculated with an ensemble of at least 10 simulations with the same model and boundary conditions coming from the same coupled GCM run 10 times with different initial conditions. Another less expensive method consists of considering that the climate response is an average of 30 differences, one per year. They are not strictly independent, but if one plots the 30 differences for a given city and season, one can see that this assumption is largely valid. Thanks to Gauss's central limit theorem, we can consider that the 30 yr-based difference follows a Gaussian law with the same mean as the 1 yr-based differences. The variance is that of the 1 yr-based differences divided by 30. Note that here the seasonal mean precipitation is the actual mean, not the mean of the rainy days as was used in the previous sections for the daily values.

Rather than plotting one curve for temperature and one for precipitation as we did in the last sections, we will display 2D graphics. Indeed, there is a correlation between temperature and precipitation responses due

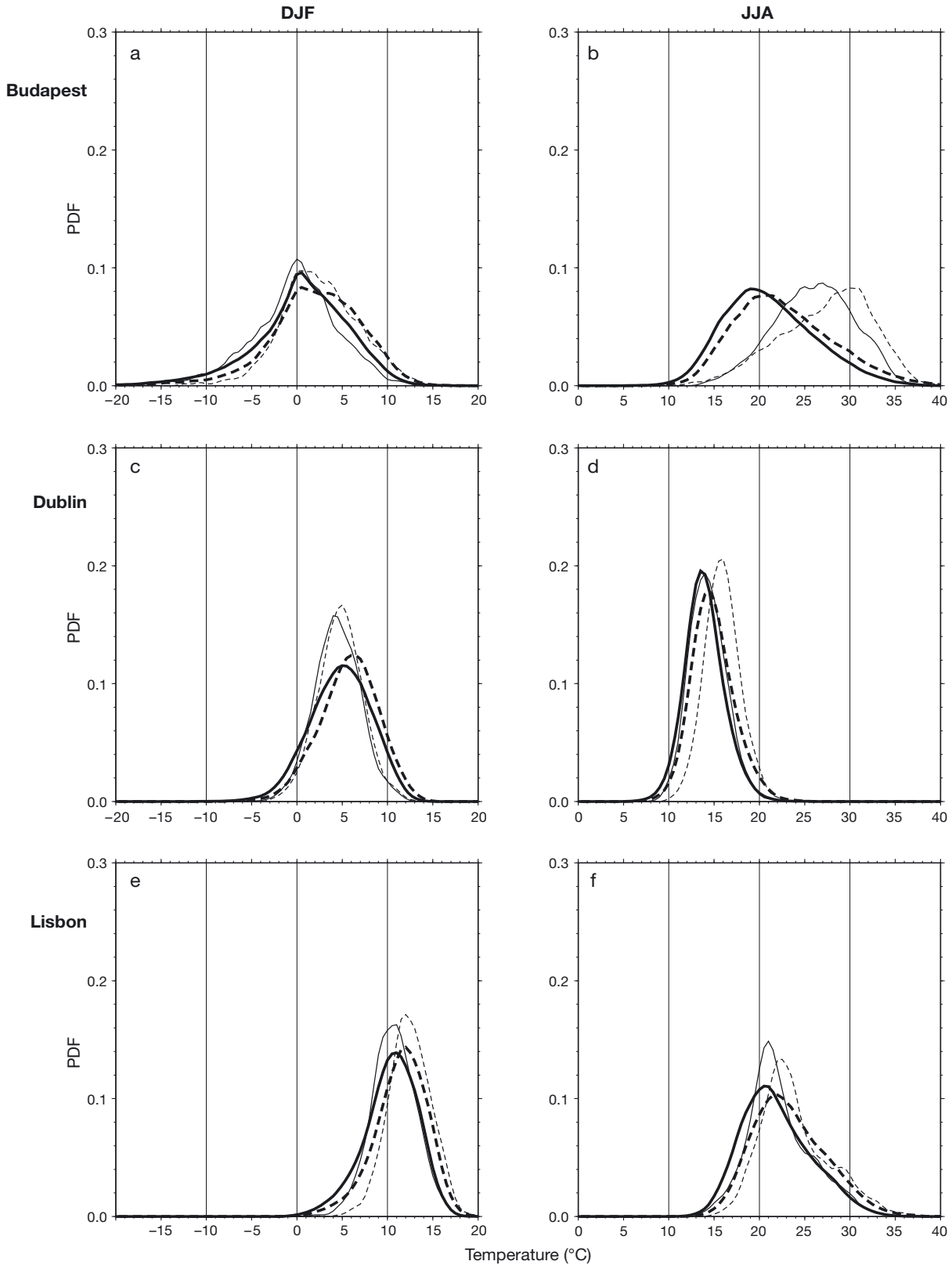


Fig. 3. Probability density functions (PDFs) of daily temperature for the 3 cities in winter (DJF) and summer (JJA) for the 1961–1990 (solid lines) and 2021–2050 (dashed lines) periods based on ENSEMBLES weights (thick lines) and for a single model based on median RPS (thin lines)

to 2 factors: (1) for a given model, warm summers are often dry summers in Europe, and (2) models which have a stronger response in temperature may also have a stronger response in precipitation.

The Gauss theorem applies also to 2D random variables like the temperature–precipitation pair, the correlation being the same as the correlation of the 30 yr-based differences. Fig. 4 shows the bivariate PDF calculated with the ENSEMBLES weights. As each model has a narrow distribution for its response (it might be a Dirac function if we neglect the natural variability), we need to apply a Gaussian kernel filter to the bivariate distribution for an easier interpretation.

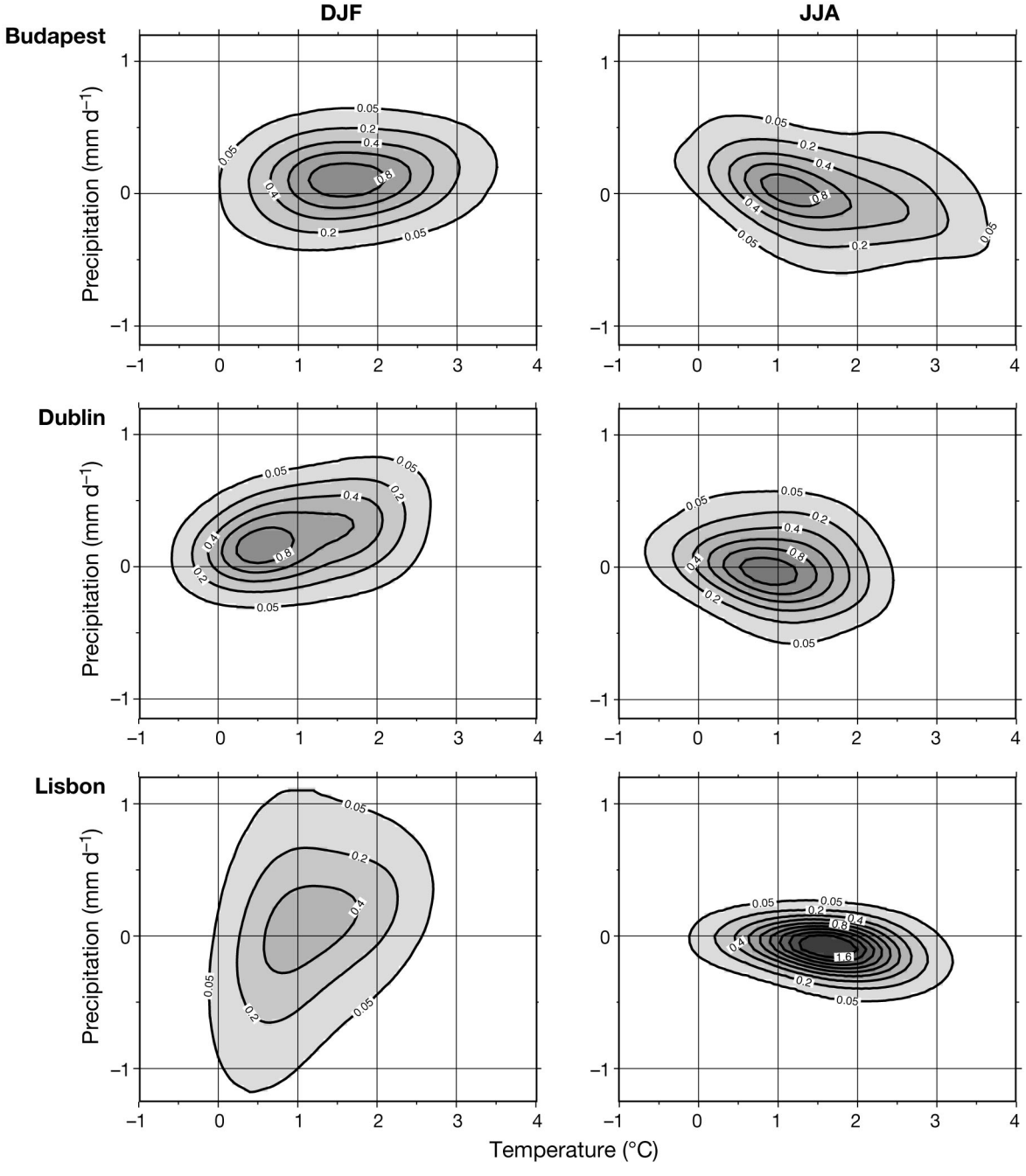


Fig. 4. Bivariate probability density functions (PDF) of temperature (°C) and precipitation (mm d^{-1}) 30 yr mean climate change responses for the 3 cities in winter (DJF) and summer (JJA). The density contour interval is 0.2°C^{-1} mm^{-1} and the 0.05 contour is added

From a technical point of view the generalization from 1D to 2D kernels is straightforward. We assume a correlation r between temperature (T) and precipitation (P) in the kernels. The filtered PDF (f_2) reads:

$$f_2(T,P) \propto \sum_{T',P'} f_1(T',P') \exp\left[-\frac{1}{2(1-r^2)}\left(\left(\frac{T-T'}{\sigma_T}\right)^2+\left(\frac{P-P'}{\sigma_P}\right)^2-2r\left(\frac{T-T'}{\sigma_T}\right)\left(\frac{P-P'}{\sigma_P}\right)\right)\right] \quad (5)$$

The bin sizes ΔT and ΔP are smaller than in Section 3: 0.04°C and 0.04 mm d^{-1}. Dummy variables T' and P' are summed up over (–2°C; 6°C) and (–2 mm d^{-1}; 6 mm d^{-1}), respectively. The values of σ_T and σ_P have been empirically tuned to remove the effect of individual models, without affecting—as far as is possible—their envelope (0.4°C and 0.1 mm d^{-1}). Some anisotropy in the filter is also introduced by the parameter r which is the average of the 17 correlations of the individual distributions. The difference with the same plots calculated with equal weights (data not shown) is negligible. Fig. 4 shows different interesting features. In DJF, the correlation is positive (lower left to upper right shape), which means that the warmest scenarios are associated with the wettest ones. As expected, the opposite is seen in JJA (upper left to lower right shape). Except in Dublin in DJF, the PDF is centered along the *x*-axis, which means that a positive or a negative precipitation response are almost equiprobable. The case of Lisbon is different from the other 2 cities: in DJF the precipitation distribution is very flat, whereas in JJA it is very sharp.

As in Section 4, we can compare this frequency distribution with the distribution obtained from a single model. We have ranked the 17 RPSs for each city and season. Here the ranking is different from the ranking in previous sections. Indeed, we are considering here bivariate Gaussian laws for 30 yr seasonal means in both reference and observed data sets. Fig. 5 shows that the bivariate response of a single model is sharper than with the weighted combination. If we exclude Lisbon in DJF—for which precipitation spread is as large as the multi-model—the sign of the precipitation response is significant: positive in Budapest, positive (DJF) and negative (JJA) in Dublin, and negative in JJA in Lisbon. Also, the slope of the ellipses is the same as in Fig. 4.

From the bivariate multi-model distributions, it is easy to compute a marginal distribution for temperature and precipitation (data not shown). The distribution of precipitation has a rather Gaussian shape, but the distribution of temperature is skewed toward warm values (see Déqué 2009). One can then compute a cumulative distribution and a confidence interval based on quantiles. Table 7 shows the quantiles 0.5 and 99.5% which provide a 99% interval with the 2 weight systems. As usual, the results with ENSEMBLES weights and with equal weights are very similar. Note that the sign of the precipitation response is uncertain at this 99% level. At the 90% level (data not shown), this sign becomes significant for Dublin (DJF increase) and Lisbon (JJA decrease). As far as temperature is concerned, the warming is significant in Lisbon in both seasons, and in Budapest in DJF only. When the climate response is expressed only as an averaged value, a part of the ENSEMBLES message about the climate of the next 50 yr is missed. In Table 7, contrary to Fig. 4, we did not apply a Gaussian kernel filtering, because the tails of the distributions depend on its standard deviation. In addition, we use here the CDF and not the PDF (see Section 3 discussion about the RPS). The stability of the results with respect to the weighting method indicates that smoothing is not necessary for this application. If we apply kernel filtering anyway, the 99% intervals of Table 7 are broader: this is an indication that we do not have enough models in ENSEMBLES to sample the full uncertainty. As the choice of σ_T and σ_P is empirical and as its impact on the results is far from negligible (e.g. temperature interval

Table 7. Lower (L) and upper (U) boundaries of a 99% confidence interval for temperature (°C) and precipitation (mm d^{-1}). The probability law of the climate change is a weighted average of 17 individual Gaussian laws with ENSEMBLES (ENS) or equal weights (Equal), or the Gaussian law of the model which has a ranked probability score of 9 with respect to observation (Median). DJF: December–February; JJA: June–August

City, season and boundary	Temperature ENS	Equal	Median	Precipitation ENS	Equal	Median
Budapest						
DJF						
L	0.30	0.32	1.62	−0.34	−0.34	−0.14
U	3.36	3.42	3.78	0.64	0.62	0.62
JJA						
L	−0.10	−0.10	0.82	−0.56	−0.58	−0.02
U	3.50	3.54	3.50	0.62	0.64	0.54
Dublin						
DJF						
L	−0.06	−0.04	1.18	−0.22	−0.22	−0.02
U	2.30	2.34	2.42	0.80	0.82	0.90
JJA						
L	−0.16	−0.14	1.06	0.50	−0.52	−0.42
U	2.00	2.00	1.86	0.58	0.58	0.14
Lisbon						
DJF						
L	0.24	0.24	0.34	−1.32	−1.30	−1.26
U	2.45	2.42	1.32	1.50	1.50	1.50
JJA						
L	0.40	0.40	1.46	−0.38	−0.38	−0.30
U	2.82	2.84	2.48	0.14	0.14	0.10

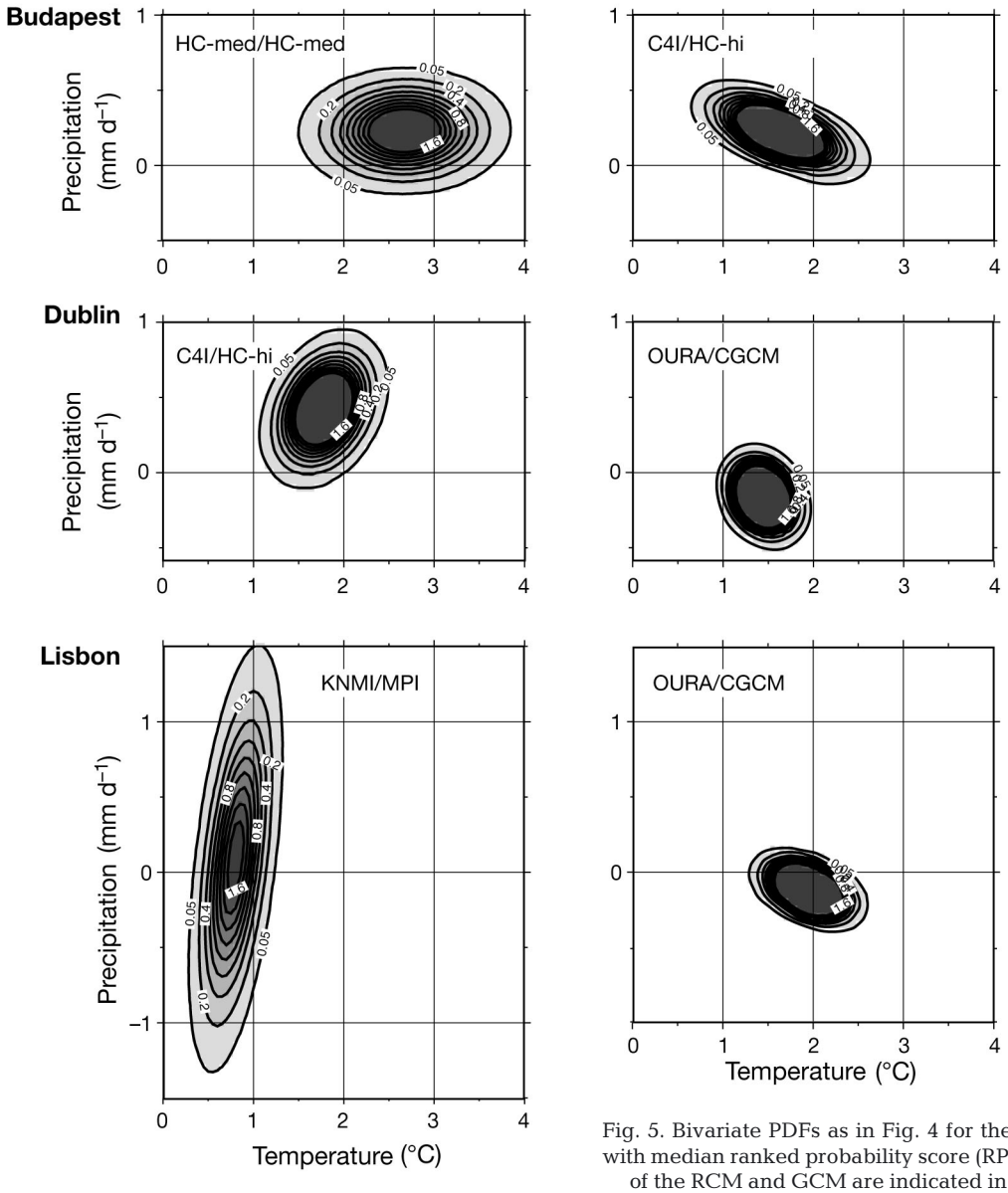


Fig. 5. Bivariate PDFs as in Fig. 4 for the single model with median ranked probability score (RPS). The names of the RCM and GCM are indicated in each panel

in DJF in Lisbon goes from (0.2; 2.4) to (–0.3; 2.8) with kernel filtering), we prefer to stay close to the strict definition of spread in our 17 member ensemble.

Coming back to single model uncertainties, Table 7 yields the intervals for the median model. This is, for each season and location, the same model as in Fig. 5. We did not sort the model RPSs separately for temperature and precipitation, in order to maintain consistency with Fig. 5. As expected, the confidence interval is narrower, but in some instances (e.g. DJF temperature in Budapest) the single model interval extends outside the multi-model interval. So when restricting to a single model, one can be at the same time more confident (smaller spread) and more pessimistic (larger mean).

7. CONCLUSIONS

Assigning weights to models which generate a huge bundle of diagnostics, and which are good for certain aspects and poor for others, is a difficult exercise. The present study is a fair evaluation of these weights, because we did not use any of the material used to construct them (ERA40-driven RCM data), but it does not evaluate all aspects of a weighting system, just its ability to simulate the observed temperature and precipitation distribution and evaluate the spread of their response at a few locations. These 2 aspects are not necessarily linked. Better simulation of the observed climate is not proof that the climate response is more reliable.

On the basis of the reconstruction of the reference climate PDF, the ENSEMBLES method is not better or worse than the use of equal or random weights. The reconstructed reference distribution is not always close to the observed distribution, but here the RCMs are driven by GCMs with their own errors. As far as the response to climate change is concerned, the results obtained with ENSEMBLES weights are rather similar to those obtained with equal weights.

These 2 results should not be interpreted in a pessimistic way, as far as the ENSEMBLES outcome is concerned. These ENSEMBLES weights are based on a physical and multi-purpose approach, so they are more reliable than equal weights, even if, in the end, the results are similar. In addition, these ENSEMBLES weights are designed to be used with any combination of model results, not only temperature and precipitation frequencies.

As a certain number of impact studies are based on a single model, we have compared our ensemble-based spread with a single model-based spread. It is tempting to consider the model which has the PDFs closest to observation in the reference period as the 'best model'. However, with 3 locations, 2 seasons and 2 variables, the best model will change from one case to another, and includes a component of chance. Instead, we prefer to use the concept of a 'median model', which is better than 50% of the other models and represents the status of an ENSEMBLES regional model in a more stable way. As far as the reconstruction of the observed PDF is concerned, a single model is worse in an RPS sense than a weighted combination, in particular for temperature. If we wanted to calculate a probability for climate change, using a single model PDF would rely upon the wrong hypothesis that this model is perfect. Using several models introduces the concept of imperfection amongst the models. In PRUDENCE, most RCMs were driven by the same GCM. Here in ENSEMBLES, we used 5 GCMs. We must acknowledge that, compared to IPCC AR4, this GCM sampling is too poor to deliver a definite response to the difficult question of a probabilistic prediction of the next 50 yr, although the RCM sampling is the best we can do. So our results must be interpreted as a sensitivity study to the weighting system in a multi-model context, not as a tool for decision makers.

Acknowledgements. This work was supported by the European Commission Programme FP6 under contract GOCE-CT-2003-505539 (ENSEMBLES). Thanks are due to the global and regional modellers of the project who produced the simulations, and to O. B. Christensen who made them available through the DMI web server.

LITERATURE CITED

Boberg F, Berg P, Thejll P, Gutowski WJ, Christensen JH (2009) Improved confidence in climate change projections of precipitation evaluated using daily statistics from the PRUDENCE ensemble. Clim Dyn 32:1097–1106

Boberg F, Berg P, Thejll P, Gutowski WJ, Christensen JH (in press) Improved confidence in climate change projections of precipitation further evaluated using daily statistics from the ENSEMBLES models. Clim Dyn doi:10.1007/s00382-009-0683-8

Böhm U, Kücken M, Ahrens W, Block A, Hauffe D, Keuler K, Rockel B, Will A (2006) CLM—the climate version of LM: brief description and long-term applications. COSMO Newsl 6:225–235

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

Christensen JH, Christensen OB, Lopez P, van Meijgaard E, Botzet M (1996) The HIRHAM4 regional atmospheric climate model. Scientific Report 96-4, Danish Meteorological Institute, Copenhagen

Christensen JH, Carter TR, Giorgi F (2002) PRUDENCE employs new methods to assess European climate change. EOS Trans Am Geophys Union 83:147

Collins M, Booth BBB, Harris GR, Murphy JM, Sexton DMH, Webb MJ (2006) Towards quantifying uncertainty in transient climate change. Clim Dyn 27:127–147

Déqué M (2009) Temperature and precipitation probability density functions in ENSEMBLES regional scenarios. ENSEMBLES Technical Report 5, Météo-France/CNRM, Toulouse

Doblas-Reyes FJ, Déqué M, Piedelievre JP (2000) Model and multimodel spread in the PROVOST seasonal forecasts: application to probabilistic forecasts. Q J R Meteorol Soc 126:2069–2088

Epstein ES (1969) A scoring system for probability forecasts of ranked categories. J Appl Meteorol 8:985–987

Frei C, Christensen JH, Déqué M, Jacob D, Jones RG, Vidale PL (2003) Daily precipitation statistics in regional climate models: evaluation and intercomparison for the European Alps. J Geophys Res 108:4124 doi:10.1029/2002JD002287

Furevik T, Bentsen M, Drange H, Kindem IKT, Kvamstø NG, Sorteberg A (2003) Description and evaluation of the Bergen climate model: ARPEGE coupled with MICOM. Clim Dyn 21:27–51

Gibelin AL, Déqué M (2003) Anthropogenic climate change over the Mediterranean region simulated by a global variable resolution model. Clim Dyn 20:327–339

Giorgi F, Mearns LO (1999) Introduction to special section: regional climate modeling revisited. J Geophys Res 104:6335–6352

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

Gordon C, Cooper C, Senior CA, Banks HT and others (2000) The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. Clim Dyn 16: 147–168

Haugen JE, Haakensatd H (2006) Validation of HIRHAM version 2 with 50 and 25 km resolution. RegClim General Technical Report 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 data set of surface temperature and precipitation for 1950–2006. J Geophys Res 113:D20119 doi 10.1029/2008JD010201

Hewitt CD, Griggs DJ (2004) Ensembles-based predictions of climate changes and their impacts. EOS Trans Am Geophys Union 85:566

IPCC (2007) Climate change 2007: the physical science basis. In: Solomon S, Qin D, Manning M, Chen Z and others (eds) Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge

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 D, 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 81:31–52

Jones CG, Willen U, Ullesrtig A, Hanssson U (2004) The Rossby Centre regional atmospheric climate model. I. Model climatology and performance for the present climate over Europe. Ambio 33:199–210

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). Reports Meteorology and Climatology 108, Swedish Meteorological and Hydrological Institute, Norrkoping

Lenderink G, van den Hurk B, van Meijgaard E, van Ulden A, Cuijpers H (2003) Simulation of present day climate in RACMO2: first results and model developments. Technical Report 252, Dutch Meteorological Service, De Built

Murphy JM, Booth BBB, Collins M, Harris GR, Sexton D, Webb M (2007) A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. Philos Trans R Soc A 365:1993–2028

Plummer D, Caya D, Coté H, Frigon A and others (2006) Climate and climate change over North America as simulated by the Canadian regional climate model. J Clim 19: 3112–3132

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

Roeckner E, Bäuml G, Bonaventura L, Brokopf R and others (2003) The atmospheric general circulation model ECHAM5. I. Model description. Report 349, Max Planck Institute for Meteorology, Hamburg

Rougier JC (2007) Probabilistic inference for future climate using an ensemble of climate model evaluations. Clim Change 81:247–264

Salas-Mélia D (2002) A global coupled sea ice-ocean model. Ocean Model 4:137–172

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 (2009) Ability of an ensemble of regional climate models to reproduce the weather regimes during the period 1961–2000. Clim Dyn 33:723–736

Scinocca JF, McFarlane NA, Lazare M, Li J, Plummer D (2008) The CCCma third generation AGCM and its extension into the middle atmosphere. Atmos Chem Phys Discuss 8:7883–7930

Sun Y, Solomon S, Dai A, Portman RW (2007) How often will it rain? J Clim 20:4801–4818

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

Submitted: September 25, 2009; Accepted: April 21, 2010

Proofs received from author(s): June 22, 2010