Clim Dyn (2010) 35:587–600
DOI 10.1007/s00382-009-0654-0

Regional climate change patterns identified by cluster analysis

Irina Mahlstein · Reto Knutti

Received: 22 December 2008 / Accepted: 17 August 2009 / Published online: 29 August 2009

Abstract Climate change caused by anthropogenic greenhouse emissions leads to impacts on a global and a regional scale. A quantitative picture of the projected changes on a regional scale can help to decide on appropriate mitigation and adaptation measures. In the past, regional climate change results have often been presented on rectangular areas. But climate is not bound to a rectangular shape and each climate variable shows a distinct pattern of change. Therefore, the regions over which the simulated climate change results are aggregated should be based on the variable(s) of interest, on current mean climate as well as on the projected future changes. A cluster analysis algorithm is used here to define regions encompassing a similar mean climate and similar projected changes. The number and the size of the regions depend on the variable(s) of interest, the local climate pattern and on the uncertainty introduced by model disagreement. The new regions defined by the cluster analysis algorithm include information about regional climatic features which can be of a rather small scale. Comparing the regions used so far for large scale regional climate change studies and the new regions it can be shown that the spacial uncertainty of the projected changes of different climate variables is reduced significantly, i.e. both the mean climate and the expected changes are more consistent within one region and therefore more representative for local impacts.

Keywords Regional classification · Regional climate change · Cluster analysis

I. Mahlstein (✉) · R. Knutti
Institute for Atmospheric and Climate Science, ETH Zurich, Universitätstrasse 16, 8092 Zurich, Switzerland
e-mail: irina.mahlstein@env.ethz.ch

1 Introduction

As a result of human induced changes in the atmospheric composition of trace gases, future climate is expected to change in different aspects, such as its mean state, interannual variability, and its extremes. Due to different feedbacks, climate will also change differently in the various regions in the world. Furthermore, the individual climate variables governing a climate regime show very distinctive patterns of change. Ideally, climate models should provide information on very small spatial scales, in particular for planning adaptation measures. Yet limitations in terms of processes and resolution prevent the interpretation of model results on single grid points. In order to simplify the communication of the results and to increase the robustness of the results, climate change patterns are therefore often aggregated over space. Regional climate change results have often been presented based on simple rectangular areas (Giorgi and Mearns 2003; Tebaldi et al. 2004, 2005; Giorgi and Bi 2005; Christensen et al. 2007) originally defined by Giorgi and Francisco (2000). The choice of these regions was rather pragmatic, based on a qualitatitive understanding of current climate zones and an expert assessment of the performance of climate models a decade ago. Here, a quantitive method is presented that attempts to address several shortcomings of the regions used so far. First, as computational capacity increases and models improve, the increasing complexity and resolution should provide information on smaller regional scales. Second, to facilitate communication of climate change results, regions should be based on similar expected future changes and not only on similar present day mean values. Third, climate is not bound to a rectangular shape. The shape of the region will rather depend on the variable of interest and the climate regime. Furthermore, local

topography variations can influence climate. Especially precipitation patterns can vary strongly due to regional differences in topography. These small scale variations should be taken into account when looking at climate changes and defining regions for aggregating climate change results.

Cluster analysis methods can be used to define regions where the climate change signal is similar in all grid cells encompassed by the region. The hypothesis is that, based on the variable(s) of interest, there is an optimal number of regions where models can provide robust information. Model agreement generally improves on larger scales (Räisänen 2007). Therefore, if a region is too small, the models may disagree in their signal. If the region is too large, the changes will be blurred and information is lost because different climate regimes are averaged together, e.g. averaging positive and negative precipitation changes will result in little net change. In addition, if the regions are very large, the information is no longer useful for local impacts, as the regional average is unlikely to be representative for local changes.

An algorithm is presented here which can contribute to answering the question of how and on what spatial scale regional climate change results from global climate models should be communicated. There is of course no perfect definition of such regions. Different questions require different answers which one single set of regions will not be able to provide. People interested in future temperature change will not make best use to work with regions based on precipitation since changes in these two variables differ. Therefore, the goal is to present one possible procedure to define regions which can be used to find answers to the question asked, i.e. to group climate data in such a way that regions encompass similar characteristics. The characteristics looked at depend on the question of interest.

2 Data and method

2.1 Data

This study uses up to 23 of the global coupled atmosphere ocean general circulation models (AOGCMs) (see Table 1) used for the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4) (IPCC 2007) which are available from the World Climate Research Programme (WCRP) Coupled Model Intercomparison Project Phase 3 (CMIP3) (Meehl et al. 2007). Detailed information of the participating models is available on the Program for Climate Model Diagnosis and Intercomparison (PCMDI) website: http://www-pcmdi.llnl.gov/ipcc/about_ipcc.ph.

All model data is regridded to a common T42 grid using a bilinear interpolation. One ensemble member of each model of the surface air temperature (TAS) and precipitation (PR) fields from the simulation of the twentieth century (20C3M) and the scenario A1B (SRES-A1B) simulations are used to construct an equally weighted multimodel mean (M).

Table 1 Climate models and their runs used in this study

	Temperature data	Precipitation data
BCCR-BCM2.0 run1	x	x
CCCMA-CGCM3.1 run1	x	x
CCCMA-CGCM3.1-T63 run1	x	x
CNRM-CM3 run1	x	x
CSIRO-MK3.0 run1	x	x
CSIRO-MK3.5 run1		x
GFDL-CM2.0 run1	x	x
GFDL-CM2.1 run1	x	x
GISS-AOM run1	x	x
GISS-EH run1	x	x
GISS-ER run1	x	
FGOALS-g1.0 run1	x	x
INGV-ECHAM4 run1		x
INM-CM3.0 run1	x	x
IPSL-CM4 run1	x	x
MIROC3.2(hires) run1	x	x
MIROC3.2(medres) run1	x	x
ECHO-g run1	x	x
ECHAM5/MPI-OM run1	x	x
MRI-CGCM2.3.2 run1	x	x
NCAR-CCSM3 run1	x	x
NCAR-PCM run1	x	x
UKMO-HadCM3 run1	x	x
UKMO-HadGEM1 run1	x	x

For the analysis, two 30-year monthly mean climatologies for the multimodel mean during the periods 1970–1999 and 2070–2099 were calculated. The simulated change in climate is simply the difference between these two time periods.

2.2 Method

2.2.1 Determination of the best cluster solution

The goal of a cluster analysis (CA) is to partition observations into groups (clusters) such that the pairwise dissimilarities between those observations assigned to the same cluster tend to be smaller than those in different clusters. Many different algorithms to obtain a classification exist in the literature (Jain et al. 1999). The one used for this analysis is the conventional *k*-means algorithm. It has the advantage of being simple and inexpensive. But on

the other hand, the number of clusters needs to be preassigned before running the algorithm. Furthermore, when applying k-means on datasets with a large sample size and a relatively smooth character (i.e. lack of obvious groups) multiple solutions may exist which cannot be improved any further by rearranging single objects of the different clusters. Although these solutions can be far away from the best solution (called the global optimum) the algorithm is trapped in a local optimum. Checking all possible combinations of the objects is computational infeasible due to the large sample size. Therefore, there is no way to determine whether a specific optimum is the global optimum (Philipp et al. 2007). One common way to apply k-means is to create seed partitions as first guesses, hence there is an arbitrary step in this procedure which leads to different realizations of solutions for the k-means algorithm applied on the same dataset. By running the algorithm multiple times the probability of converging to a local optimum of very low quality can be reduced. Other studies applied a similar technique in order to select the best cluster solution. Bonfils et al. (2004) repeated the clustering procedure several times with different drawings of initial centers. The best solution defined by the most distinct cluster was kept. Brewer et al. (2007a) ran the clustering procedure several times as well and checked visually that the clusters remained stable. Brewer et al. (2007b) decided to run the k-means algorithm 1,000 times and the standard deviation of the value attributed to the centroid was calculated after they already obtained a solution. A low value indicates that the centroids do not vary significantly and the found solution therefore is stable.

In this study the best solution is selected by comparing the within cluster sum of deviations (WWS) (Philipp et al. 2007):

$$\mathrm{WWS} = \sum_{j=1}^{k} \sum_{i \in C_j} D(X_i, \overline{X}_j)^2, \tag{1}$$

where k is the number of clusters C, i the object number, $\overline{X}$ the centroid of the cluster, and D the Euclidean distance between the objects and its cluster centroids:

$$D(X_i, \overline{X}_j) = \left(\sum_{l=1}^{m} (X_{il} - \overline{X}_{jl})^2 \right)^{\frac{1}{2}}, \tag{2}$$

where m is the number of parameters describing the object. Thus, WSS needs to be minimized in order for the solution to be as close as possible to the global optimum.

One way of testing the convergence of the algorithm is to compare the few best solutions out of a large sample of solutions. If these are very similar, the resulting regions are robust and it is unlikely that a much better solution exists. This is tested here by calculating a correlation that measures the similarity of the ten best solutions (lowest WWS value) at each grid cell. For a given grid cell and for each of the ten best solutions, the cluster where the grid cell belongs to is determined, and all the grid cells belonging to the same cluster are labelled with one, whereas all other grid cells are set to zero. The correlation between two such maps indicates whether the regions (to which the particular grid cell is assigned to in the two solutions) are similar in shape and size. The average of all 100 pairwise correlations is determined at each grid point in that way. Figure 1a shows the correlation indices at each grid point. For comparison, Fig. 1b shows the best solution determined by the lowest WWS value. It shows a CA solution for 22 regions based on the multimodel mean M of the mean 30-year annual cycle of temperature data (details see Sect. 3). The high correlation leads to the conclusion that the solutions are of high quality since the different solutions do not differ greatly from each other. The largest differences are found at the edges of the clusters (compare with the regional classification shown in Fig. 1b), which is not surprising. The data used for the CA are relatively smooth fields, especially in case of temperature data. Precipitation data may be less smooth, but it is still rather difficult to define groups of data which are separated by a sharp border. Thus it is difficult for the algorithm to distinguish clearly between groups of data. In the following analysis, out of 60 realizations the best in terms of lowest WSS is chosen in order to guarantee a solution of high quality.

2.2.2 *Defining the number of clusters*

Several methods have been proposed to determine an appropriate number of clusters for a dataset. Kaufmann and Rousseeuw (1990) define a silhouette index which gives an idea of how well-separated the resulting clusters are. High

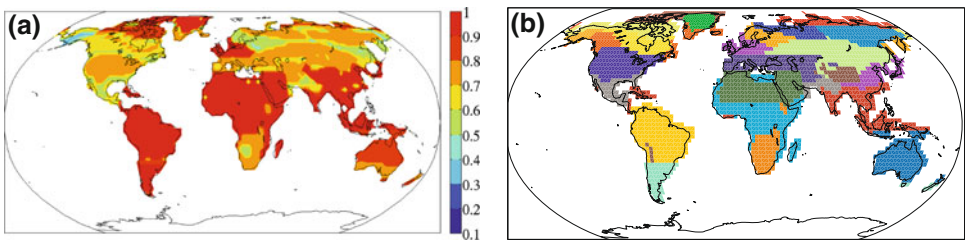

Fig. 1 Correlation of ten best k-means CA solutions for 22 clusters based on surface annual temperature data **a** and the best CA solution selected by the lowest WWS value **b**. In **b** grid points marked in the same color belong to the same region

values indicate that the objects are very distant from neighboring clusters. Wilks (1998) introduces a method based on the distances of the centroids. Furthermore, a stability index was tested which calculates the number of clusters leading to the most stable solution of all given clustering problems with a given data set. However, for the question addressed here, none of these methods were able to give a clear indication of how many clusters would be appropriate. Therefore, from a purely statistical side there seems to be no reliable procedure to determine the number of clusters. This finding agrees with the conclusion of Philipp et al. (2007). Defining an appropriate number of clusters seems to require specific solutions which depend on the problem looked at. Bonfils et al. (2004); Brewer et al. (2007a, b) for example developed a method which involves the inter-group and the intra-group variance to find the optimum number of clusters.

For this study one possibility to estimate the number of clusters is by trying to find a solution that minimizes the total uncertainty in local climate change given the information about the regional changes. From an impact perspective there are two contributions to the climate change uncertainty at a single grid point: First, models disagree about the changes in the region. This can be quantified by the spread of the different CMIP3 models relative to the change in each region, averaged over all regions:

$$\mathrm{UM}(k) = \frac{1}{k}\sum_{c=1}^{k}\left(\frac{\sigma(\overline{\triangle V_{M(c)}})}{\frac{1}{m}\sum_{i=1}^{m}(\overline{\triangle V_{M(c)}})_i}\right), \tag{3}$$

where UM(k) is the model uncertainty for k clusters, $\sigma(\overline{\triangle V_{M(c)}})$ the standard deviation across models of the mean expected change of parameter V in cluster c, m the number of models used in the analysis and $\frac{1}{i}\sum_{i=1}^{m}(\overline{\triangle V_{M(c)}})_i$ denotes the mean of the expected change of parameter V in cluster c across models.

Second, an uncertainty is introduced by the difference between the local climate signal (e.g. expected temperature rise in one single grid cell) which does not necessarily correspond to the mean expected climate in a region (e.g. mean expected temperature rise over all grid cell within one cluster). This uncertainty can be defined as the spread of the change at different grid points within one region:

$$\mathrm{UC}(k) = \frac{1}{k}\sum_{c=1}^{k}\left(\frac{\sigma(\triangle V_{G(c)})}{\frac{1}{g}\sum_{i=1}^{g}(\overline{\triangle V_{G(c)}})_i}\right), \tag{4}$$

where UC(k) is the uncertainty of the climatic pattern within one cluster for k clusters, $\sigma(\triangle V_{G(c)})$ the standard deviation of the expected change of parameter V across all the grid cells belonging to cluster c, g the number of gridcells within cluster c and $\frac{1}{g}\sum_{i=1}^{g}(\overline{\triangle V_{G(k)}})_g$ the mean expected change of parameter V in cluster c.

If the world is divided into many but very small regions, the models will often disagree on the projected changes. On the other hand, if a small number of large regions is chosen, the models will agree better but the changes aggregated over large regions may not be useful from an impacts point of view. The aim is therefore to minimize both types of uncertainties discussed above. Hence, adding up these two relative uncertainties and finding a minimum in this function (U_{tot}) is one way of defining an optimal number of clusters in terms of uncertainty. In order to reduce the noise in these functions the average of the best five solutions defined by the lowest WWS is taken. This method was applied on univariate datasets. In most studies performed before, as well as in the IPCC reports the changes in climate are communicated for only one variable at a time, such as temperature or precipitation. Therefore, the CA is executed for one variable only. More information can be found by performing a CA for precipitation and temperature combined. These results compare very well with the Köppen–Geiger climate classification (see Sect. 3).

3 Results

The best information about the expected changes can be provided by carrying out a CA for the different variables individually. Changes in precipitation are substantially different in where they are most pronounced as well as the direction of change and differ compared to patterns of temperature change. Therefore, the shape and the number of regions depend on the variable of interest.

3.1 Comparison to the old set of regions

In order to decide whether the new regions contain more information in terms of spatial and model spread of each variable, the same number of regions (22) as in the old set used by previous studies (Tebaldi et al. 2004; Giorgi and Bi 2005; Christensen et al. 2007; Furrer et al. 2007a) is generated. The regions are based on the monthly means of the current 30-year climatology (1970–1999) of the multimodel mean M of the variable of interest (temperature and precipitation in this case), the projected monthly mean changes, altitude, latitude and the labels of the continent are included for each grid cell over land areas only. The reason for working with land areas only is simply that these areas are mainly the ones leading to impacts concerning society. Since Europe and Asia are not separated by an ocean, these two continents denote one single continent termed Eurasia. The purpose of labeling the continents is to reduce the tendency of grouping grid cells in the same cluster which do not belong to the same continent. By

doing so we introduce categorical values in addition to the smooth character of the other parameters. This leads to difficulties in the standardization procedure. This is circumvented by adding each continent individually to the dataset, i.e. for each continent a binary vector (of length n where n is the number of land grid points) is defined where elements are set to one for grid points belonging to that particular continent and zero elsewhere, rather than using a single vector where a different number is given to each continent. Furthermore, Antarctica was excluded in this study because both spatial and model variations are large and because in terms of impact studies it is of interest to make accurate projections in those regions which have direct influence on society. Thus, a CA for 22 regions needs to be computed not taking into account Antarctica. Hence, the regions defined here are based on climatic feature as well as on geographical and political considerations. The choices made here on which parameters are included in the CA are partly subjective. Including continental labels simplifies communication because regions tend to be restricted to one continent only. The regions defined here are based on current climate as well as on the projected changes. An alternative approach would be to consider the changes only. However, from an impacts perspective it is of interest to look at existing climatic regimes as well as the expected changes (e.g. Tebaldi and Lobell 2008), as a drying of 10% for example will have more serious consequences in an area that is already dry and water limited compared to a wet area.

In Fig. 2a and b the CA solutions based on temperature and precipitation data are shown. Compared to the regions proposed by Giorgi and Francisco (2000) (Fig. 2c), both the regions based on temperature as well as on precipitation include climatic characteristics (e.g. the influence of mountains) as well as features which are of a smaller scale than found in the old regions. These findings are consistent with the fact that the model resolution has improved over the past decades, although it is not clear what size and shape the regions derived with the same algorithm applied on the output of older models would have looked like. But since the spatial scale is smaller, the uncertainty stemming from large regions which blur different climate zones should be reduced.

Table 2 lists an area weighted mean spread over all regions and all months for the old set of regions and the regions defined by CA. This spatial spread is defined as the difference between the 90th and 10th quantile of all values (grid cells) within one region. It is calculated in each region and for each month for the 1970–2000 climatology and the projected changes. These values are then averaged over all months and over all regions, weighting by the size of the region. This provides an aggregated measure of how similar the present day climate and the projected changes are within one region, with smaller values indicating a more homogeneous pattern within each region, i.e. a better cluster solution. Thus, Table 2 provides information about the homogenity of the climate within one region. This number should not be confused with UC, which is introduced to determine the number of clusters.

As Table 2 indicates, the spatial spread in the current temperature within the regions is reduced by more than 3 K, or almost 25% compared to the old regions. For the expected warming the spatial spread is reduced by 0.4 K, more than 26%. For precipitation the improvement for current precipitation rates is 1.2 mm day^{-1}, for the expected change in precipitation the spread is similar in both cases. The spatial spread is therefore much smaller in the new regions, i.e. the regions capture the present day climate regimes and the expected changes better.

On the other hand, there is a danger that the models agree less in the regions defined by CA since the models tend to agree less on smaller scales (Räisänen 2001).

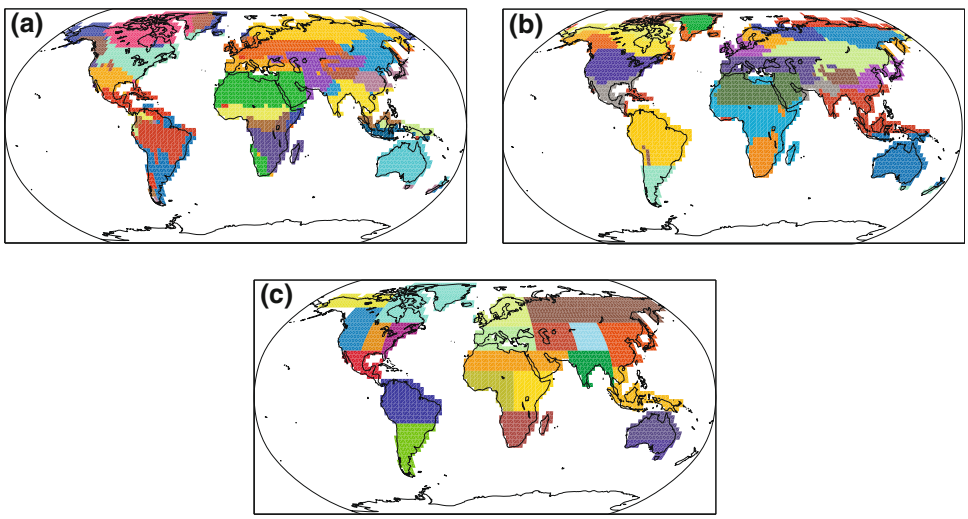

Fig. 2 Best solution of 22 regions generated using CA based on **a** precipitation data, **b** temperature data using the multimodel mean of the annual cycle and mean projected changes. For comparison the old set of regions is shown in **c**

Table 2 Spatial spread and improvement in the spatial spread of current mean temperature (*T*current), of the expected changes of temperature (ΔT), of current precipitation (P_{current}) and of the expected changes in precipitation (ΔP) in the two sets of regions

	T_{current} (*K*)	$\triangle T$ (*K*)	P_{current} (mm day^{-1})	$\triangle P$ (mm day^{-1})
Old regions	12.9	1.5	3.4	0.49
Clustered regions	9.7	1.1	2.2	0.36
Difference	−3.2 (−24.8%)	−0.4 (−26.7%)	−1.2 (−35.3%)	−0.13 (−26.5%)

Table 3 Model spread and changes in the model spread of current mean temperature (T_{current}), of the expected changes of temperature (ΔT), of todays precipitation (P_{current}) and of the expected changes in precipitation (ΔP) in the two sets of regions

	T_{current} (*K*)	$\triangle T$ (*K*)	P_{current} (mm day^{-1})	$\triangle P$ (mm day^{-1})
Old regions	5.7	4.3	1.2	0.50
Clustered regions	5.8	4.4	1.2	0.48
Difference	0.1 (1.8%)	0.1 (2.3%)	0.0 (0%)	−0.02 (−4%)

Indeed, analysing the model spread in an analogous way to above over the old set of regions and the ones defined by CA (see Table 3), the weighted mean over all regions and months indicates a slight decrease in model agreement, except for the projected changes in precipitation, which shows a slight improvement. The differences, however, are negligible small. These numbers should not be mistaken with UM, which was introduced to find the optimal number of clusters.

Overall, the reduction in the spatial spread is much larger than the increased model spread and should thereafter lead to more consistent signals within regions.

3.2 The effect of resolution on the number of regions

As mentioned before, increasing complexity and resolution of models should provide information on smaller regional scales. Therefore, one may argue that the number of regions should increase for better models. From a modeling perspective, one criteria could be to minimize the total uncertainty (see Sect. 6), i.e. minimizing the sum of the uncertainty in the spatial spread as well as in the model spread.

3.2.1 Temperature

Figure 3a shows the total (U_{tot}), spatial (UC) and model uncertainties (UM). As expected, the spatial uncertainty UC approaches zero as the number of clusters increases. This is expected since the fewer grid cells belong to one cluster, the smaller is the spread of the temperature pattern within this cluster. On the other hand, for the model spread the more clusters there are, the larger is the model spread UM. Adding up UM and UC leads to the function U_{tot} in which we seek a minimum in order to minimize the total uncertainty. Indeed, there is a minimum for about eight clusters but theses small variations are unlikely to be robust. For a large number of clusters the function converges to 0.35 (see Fig 3a). The results suggest that a range of values for the number of regions is possible without changing the uncertainty significantly. Having a large number of regions leads to the difficulty of how to communicate the findings in for example 50 different regions. By choosing a rather small number of regions we might loose information about regional climates. The number of regions is partly subjective but the conclusions will not strongly depend on the choices made. Figure 3b suggests a number of 35 to about 50 regions, because for more than 50 regions U_{tot} stays more or less constant, i.e. the gain of information is small for more than 50 regions and the changes in U_{tot} are rather large for less than 35 regions. But as mentioned before, due to communication issues and based on subjective considerations we argue that an upper limit of 35 clusters and a lower limit of 20 clusters is desirable. According to theses limits and the function of uncertainty 35 regions would be optimal.

In Fig. 3c and d the correlation index for different numbers of clusters is shown. This correlation index is derived by taking the mean of the correlations indices across all grid cells as shown in Sect. 3. For two clusters, the correlation index is very high as there are only few different plausible possibilities in clustering. The same is true for a high number of clusters. There is a minimum for about 80 clusters, apparently the solutions are most unstable for this number of clusters. Correlations are above 0.7 in all cases.

Taking into account the two contrary effects of either prefering a large number of clusters to minimize uncertainty, or rather fewer to guarantee stable cluster solutions, we conclude that the number of clusters for temperature

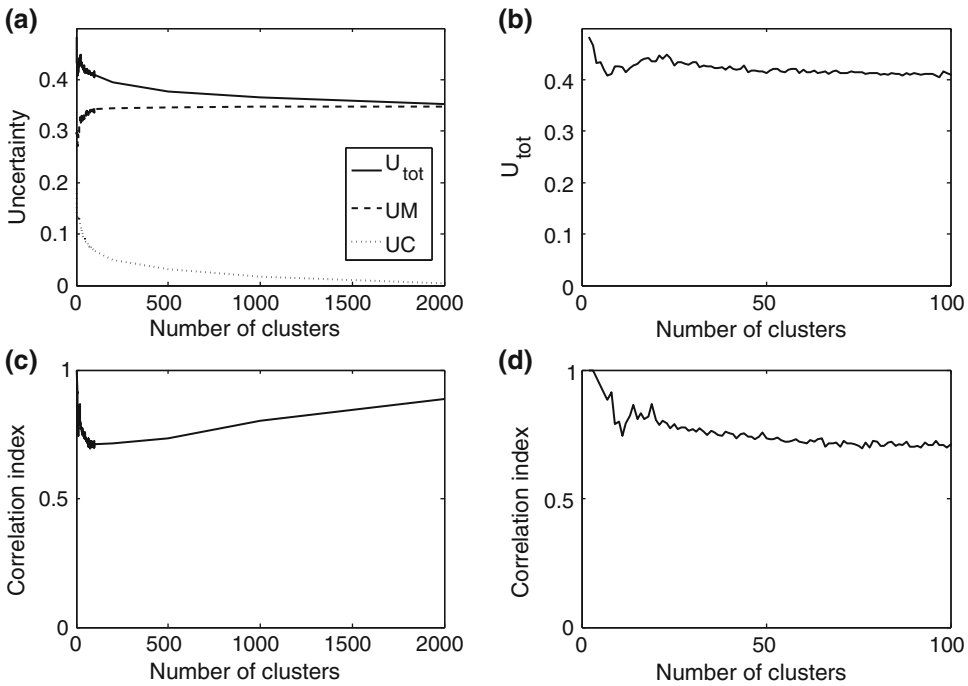

Fig. 3 **a** Total relative error (U_{tot}), model uncertainty (UM) and the uncertainty of the climatic pattern (UC) for the CA solutions for different numbers of regions of the annual cycle of temperature, **b** detailed view of U_{tot}, **c** correlation for different number of clusters for the annual cycle of temperature and **d** its detailed view

should be between 30 and 35. The quantitative conclusions drawn here do not depend on the number of clusters, except for extreme choices. For illustration Fig. 4a shows the CA solution for 33 regions.

By introducing more but smaller regions the question whether the model disagreement is too large needs to be addressed. Looking at the distribution of all possible simulated changes (i.e. every single grid cell of a cluster in every model) and by doing the same for the old set of regions we can compare whether the increase in the number of regions leads to a loss of information due to model disagreement. Note that this assumes that each model is an equally plausible representation of the real world, and that the models are approximately covering the range of uncertainty. These assumptions are strictly not correct but hard to overcome (Tebaldi and Knutti 2007). For illustration the different distributions for South America are shown in Fig. 5. Comparing Fig. 5a and b it becomes obvious that there are different effects in different regions. The northernmost region has about the same distribution in both cases, the new regions provide no improvement in this case. The southern region has a distribution which is narrower in the case of the CA solution, the new regions show the advantage of having less uncertainty concerning the expected warming. But on the other hand the mountain region only found in the CA solution shows a rather broad distribution, but still being of the same range or even less compared to the two regions in Fig. 5b.

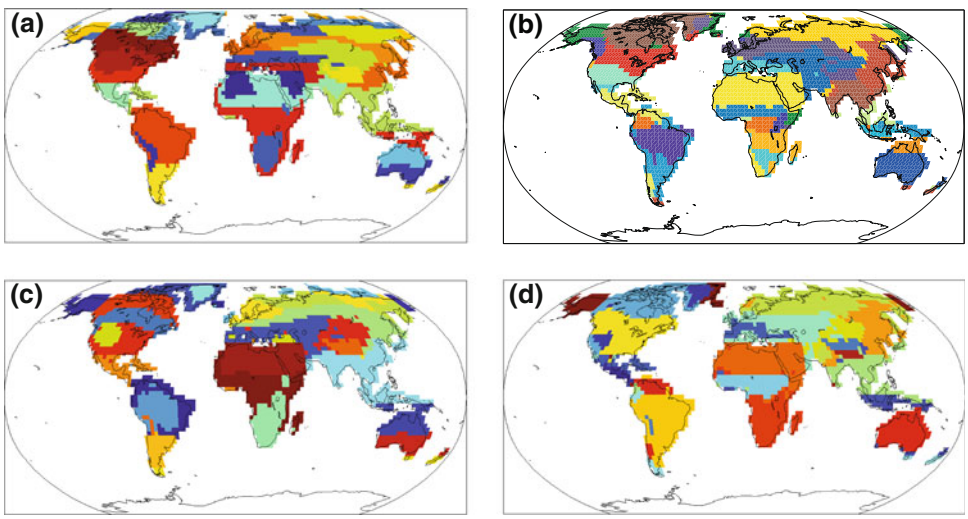

Fig. 4 CA solution for **a** temperature based on the current mean and the projected changes in the annual cycle (33 regions), **b** precipitation based on the current mean and the projected changes in the annual cycle (24 regions), **c** current mean and projected changes of JJA temperatures (26 regions), and **d** current mean and projected changes of JJA precipitation (22 regions)

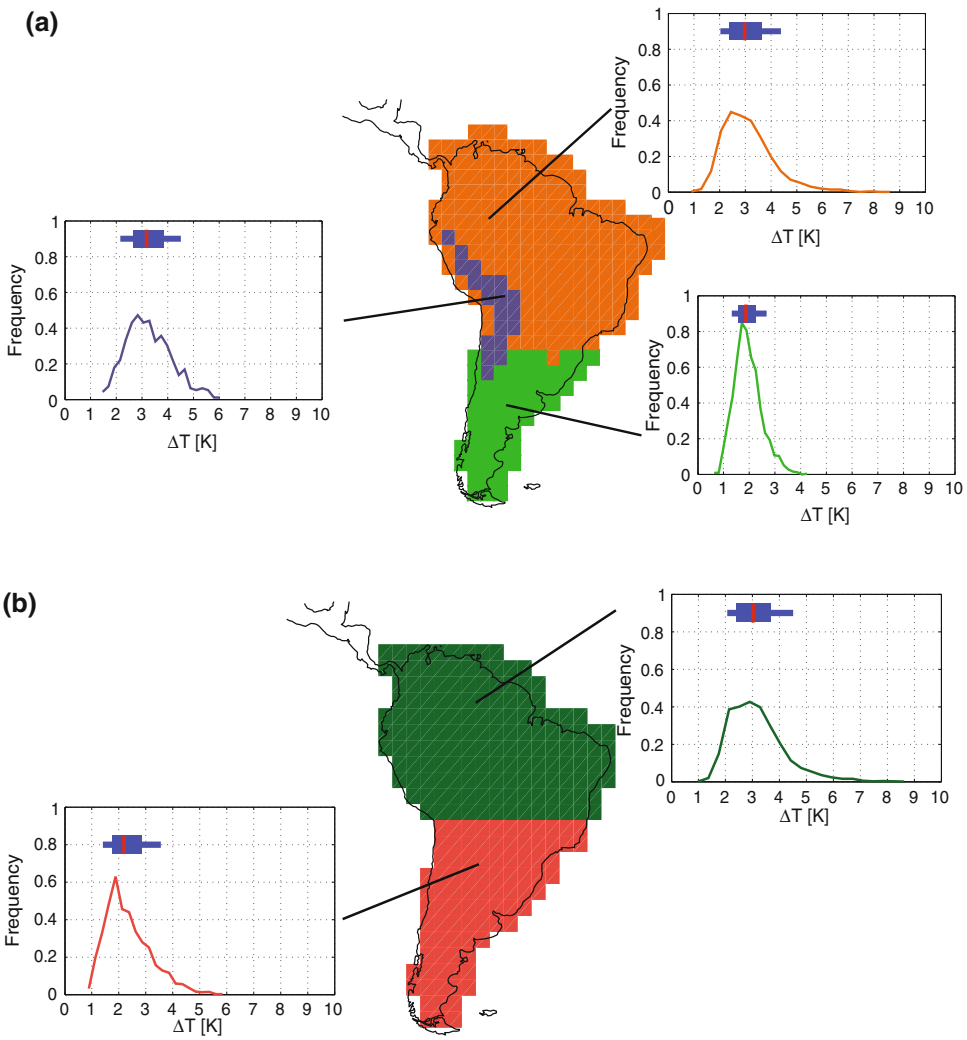

Fig. 5 **a** CA solutions for 33 regions based on the annual cycle of temperature (shown here is only South America). **b** shows for the same area the old set of regions. For each region the distribution of the expected warming for each grid cell in the region and across all the models is shown. The *boxplot* indicates the median (*red*), the 25 and 75% quantile (*box*) and the 10 and 90% quantile (*blue line*)

3.2.2 *Precipitation*

The case for precipitation it is not as clear as for temperature. As illustrated in Fig. 6 the uncertainties are much larger for precipitation. Note that due to very small mean precipitation changes in some areas, the denominator of Eqs. 3 and 4 leads to high uncertainties. Therefore, these two terms are limited here to a maximum value of 100%. We find a minimum in U_{tot} for about 200 clusters. But as mentioned above, due to communication problems the same lower and upper limit of 20 and 35 regions can be used for precipitation. The two curves of the uncertainty and correlation suggest that in case of precipitation it is desirable to have a rather low number of clusters. We believe that restricting the number of clusters to 20 to 25 makes best use of the information available. Both, Fig. 6b and d suggest a low number of clusters since the uncertainty increases with an increasing number of clusters and the correlation decreases with increasing number of clusters. For more than 25 regions the function U_{tot} reaches its maximal value and the decline thereafter is rather slow. Furthermore, in case of precipitation it is rather difficult to find a clear signal of change. As shown in Zhang et al. (2007) a signal in twentieth-century precipitation trends is found only by averaging the data in latitudinal bands. Thus, using small regions the signal of climate change may be lessend. Therefore, it is more suitable to work with larger and hence fewer regions in case of precipitation to improve the signal to noise ratio. In Fig. 4b the 24 regions for precipitation are shown.

Again the question needs to be addressed whether introducing a higher number of regions leads to a decrease in model agreement which offsets the increased spatial detail. As for temperature data, the distribution of the projected changes in precipitation is derived by looking at the distribution of all the grid cells in one cluster for all models. In the case of precipitation the results are shown for North America. Note that due to the rather complicated regional mask, some regions shown in Fig. 7 are not limited to North America, i.e. there are grid points on other

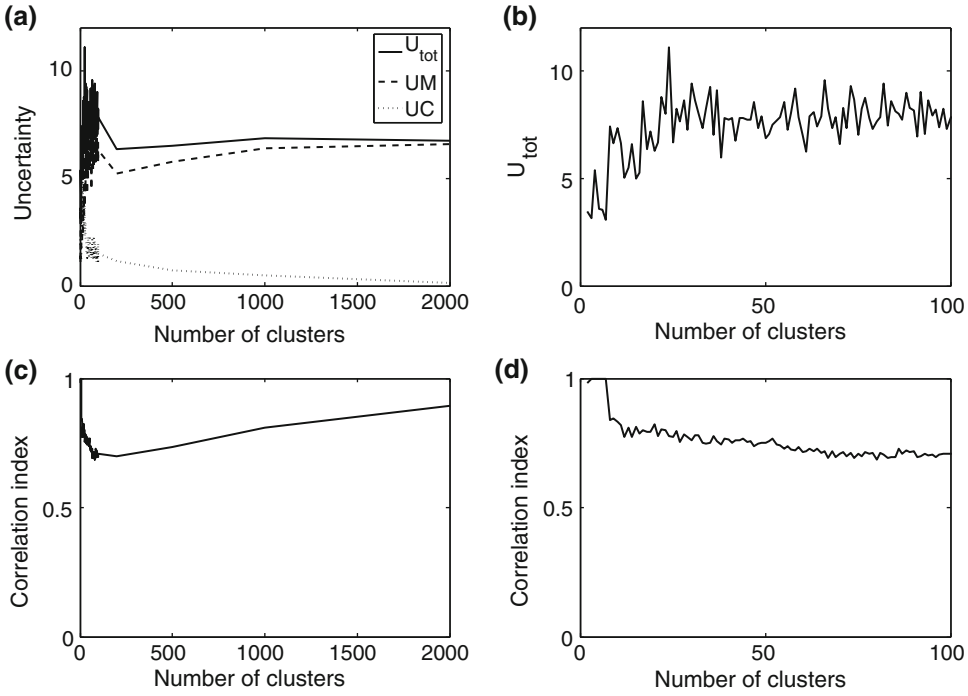

Fig. 6 **a** Total relative error (U_{tot}), model uncertainty (UM) and the uncertainty of the climatic pattern (UC) for the CA solutions for different numbers of regions for the annual cycle of precipitation, **b** detailed view of U_{tot}, **c** correlation for different number of clusters for the annual cycle of precipitation and **d** its detailed view

continents belonging to the same region. For the analysis all grid cells belonging to the same region are taken into account, not only the ones shown in the frames. Figure 7 highlights two results. First, the spread (width of box plots) is slightly reduced by defining the regions using CA. Note again that in Fig. 7 the spatial spread plus the spread of the models are included. Furthermore, by defining the regions using CA the signal for changes in precipitation can be enhanced. There are two cases in the CA-derived solution (the two northern most regions) for which both the 10 and 90% quantiles are positive. For the old set of regions this is only true for one region.

3.2.3 Monthly versus annual data input

For some studies it is not desirable to look at regions based on the whole annual cycle but based on a specific season. Figure 4c and d shows the CA solution for temperature or precipitation based on monthly means of the current climate and projected changes in June, July and August (JJA). The two solutions derived by using the annual cycle or only JJA show great similarities, but still there are some regional distinctions. Whether the appropriate number of clusters is the same for JJA as for the whole annual cycle is unclear. As the analysis shows, for the correlation index the findings are similar to the ones in Fig. 3 (compare with Fig. 8b). On the other hand, the results of U_{tot} is different (Fig. 8a) in that the function stays approximately constant for 25 to 30 clusters. Hence, in the case of a CA based on JJA temperatures a lower number of regions is favored, which in turn leads to more stable cluster solutions as well. In the case of precipitation the two curves for U_{tot} and the correlation index look similar to the ones in Fig. 6, although there are some differences. For JJA precipitation it is even clearer that fewer regions leads to more stable cluster solutions, and fewer regions are also favored by the function of U_{tot}. Therefore, about 20 to 23 regions is recommended in this case. Figure 4c and d shows the suggested regions for temperature and precipitation based on JJA values.

3.3 Clustering temperature and precipitation

Various ecological risks are associated with the prospect of a changing climate. Novel temperature regimes as well as changes in precipitation lead to novel and disappearing climates by the end of the twentyfirst century, which in turn may lead to novel species associations and other unexpected ecological responses (Williams et al. 2007). Therefore, in order to quantify the impact of climate change it may be best to look at changes in temperature and precipitation at the same time. By clustering current means of temperature and precipitation, as well as the projected changes in both of these variables, it is possible to identify regions where the current climate as well as both temperature and precipitation changes are similar. Similar patterns of current temperature and precipitation (but not trends) are the basis of the Köppen–Geiger climate classification, which represent the different vegetation groups, as plants are indicators for many climatic elements (Kottek et al. 2006). Figure 9 shows that

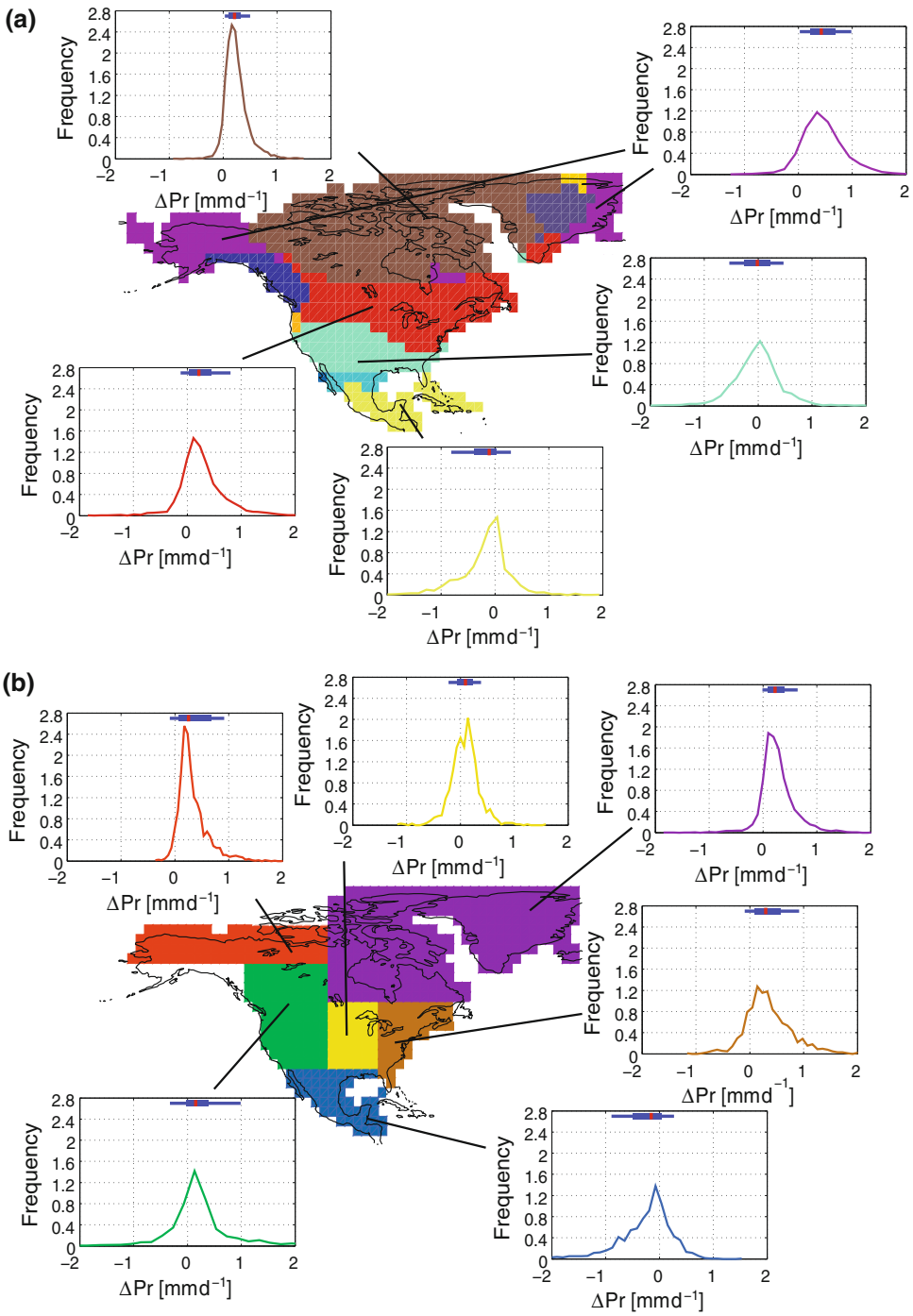

Fig. 7 **a** CA solutions for 24 regions based on current mean and projected changes in the annual cycle of precipitation data (shown here is only North America). **b** Shows for the same area the old set of regions. For each region the distribution of the expected changes in precipitation for each grid cell in the region and across all the models is shown. The *boxplot* indicates the median (*red*), the 25 and 75% quantile (*box*) and the 10 and 90% quantile (*blue line*)

the cluster solution for 31 clusters (the number of Köppen–Geiger climate classifications) is very similar to the Köppen classification as given by Kottek et al. (2006). Note that a Köppen–Geiger climate classification can be derived based on observations, but here we use model data to calculate the classification. The advantage of the CA solution is that the projected changes are already incorporated in the regional partitioning. Hence, the above defined regions are better suited to study climate change and climate change impacts.

Whether having 31 regions instead of 23 introduces more uncertainty again needs to be checked for temperature and precipitation. As before, the old set of regions serves as reference. The distributions are derived the same way as for Figs. 5 and 7. Figure 10 shows that there is improvement in the temperature signal even for regions that are not only based on temperature data but on precipitation data as well. On average the distributions are narrower for the new regions. But again, as for precipitation, some of the regions

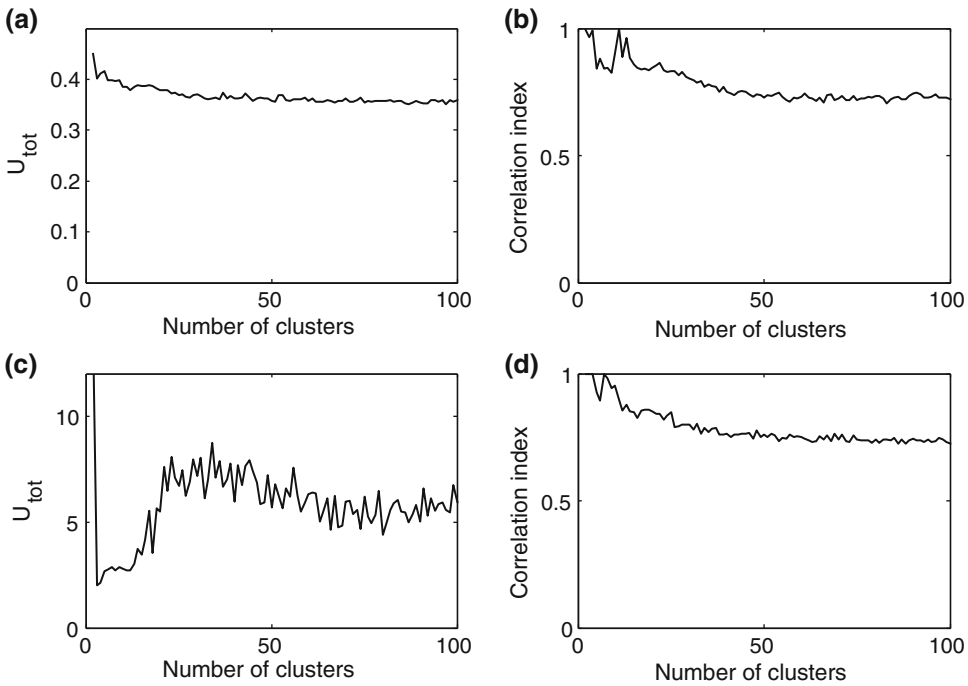

Fig. 8 Total uncertainty (U_{tot}) **a** and correlation index **b** of CA solutions for different numbers of regions based on current and projected changes in JJA temperature data and U_{tot} **c** and correlation index **d** for current and projected JJA precipitation data

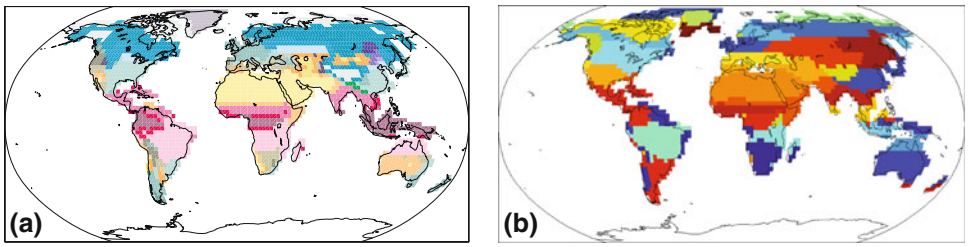

Fig. 9 **a** Köppen–Geiger climate classification derived from multimodel mean data and **b** CA solution for 31 regions based on the current mean and projected changes in the annual cycle of temperature and precipitation

shown are not limited to North America but for the analysis all the grid cells belonging to this region are used.

In case of precipitation the results in Fig. 11 indicate some improvement as well. Over all, the distributions are narrower with one exception. The region to the very south shows a rather broad distribution. But on the other hand, for this region a clearer signal towards less precipitation can be found compared to matchable region in the old set of regions.

4 Conclusions

Climate change is one of the most serious problems that our society and economy is facing. In order to quantify the impact of climate change on a regional scale and to decide on adaptation and mitigation measures a quantitative picture of the magnitude of change of the different variables is necessary. So far, regional climate change results have often been presented on simple rectangular areas defined in a rather ad hoc way instead of being based on climatic features. A procedure is presented here which offers the opportunity to define regions in which certain variables of interest, e.g. the current climate, or the projected changes, have similar values. The number as well as the shape of the regions depends on the variable(s) and the time scale of interest. It is shown that by using a cluster analysis for the regional classification focused on one variable (e.g. temperature or precipitation) the spatial spread can be reduced significantly without introducing too much uncertainty in the model disagreement compared to the old set of regions used in previous studies (Tebaldi et al. 2004) and IPCC reports (Christensen et al. 2007). Furthermore, by using the uncertainty of the climate change pattern and the model disagreement in case of temperature we can conclude that the best information of the models can be obtained by increasing the number of regions compared to the old set of regions. Therefore, regions with climatic features of a smaller scale are found by clustering the data. On the other hand, due to large uncertainties, especially in the model agreement (Wang 2005) in the precipitation data, we recommend to work with rather large regions. Although the number of regions is still larger for all the variables shown in this study than the number of regions used so far. This

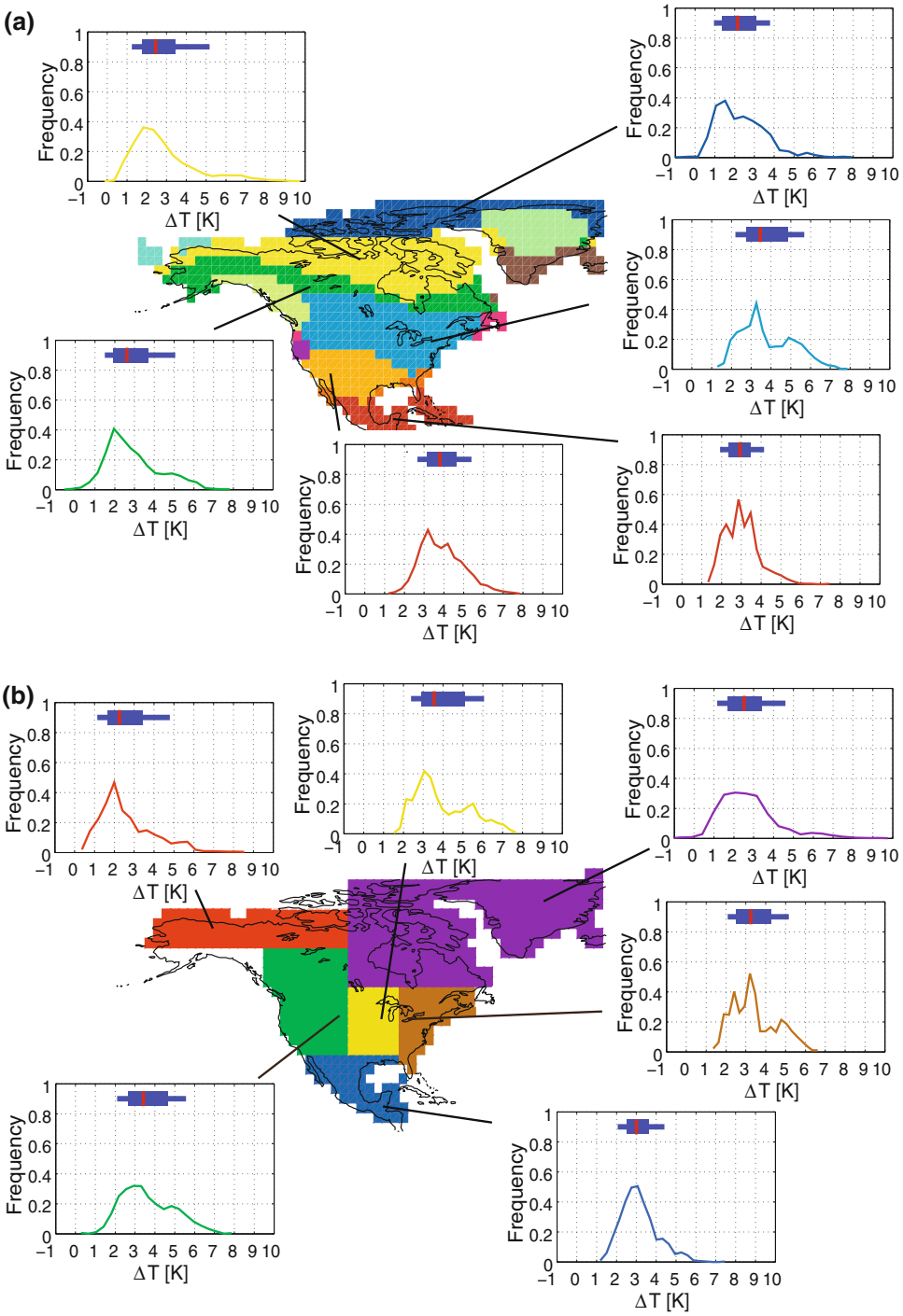

Fig. 10 **a** CA solution for 31 regions based on the current and projected annual cycle of temperature and precipitation data (shown here is only North America). **b** Shows for the same area the old set of regions. For each region the distribution of the projected warming for each grid cell in the region and across all the models is shown. The *boxplot* indicates the median (*red*), the 25 and 75% quantile (*box*) and the 10 and 90% quantile (*blue line*)

leads to regions encompassing climatic features of a rather small scale which in turn could introduce uncertainties due to model disagreement. But as could be shown in this study the total uncertainty, defined as spatial uncertainty and model disagreement, is on average still smaller than in the old set of regions.

One caveat with the presented algorithm is that there is the subjective component for the estimation of the optimal number of regions. There is small range concerning the number of regions which can be chosen from, the conclusions made here do not depend on the choices made. By introducing a lower and an upper limit of the number of regions we ensure a solution of high quality concerning the robustness of the cluster analysis solution as well as the total uncertainty including model disagreement and the local climate pattern. Furthermore, to derive a stable cluster analysis solution the multimodel mean had to be used with each model having an equal weight. This assumption of

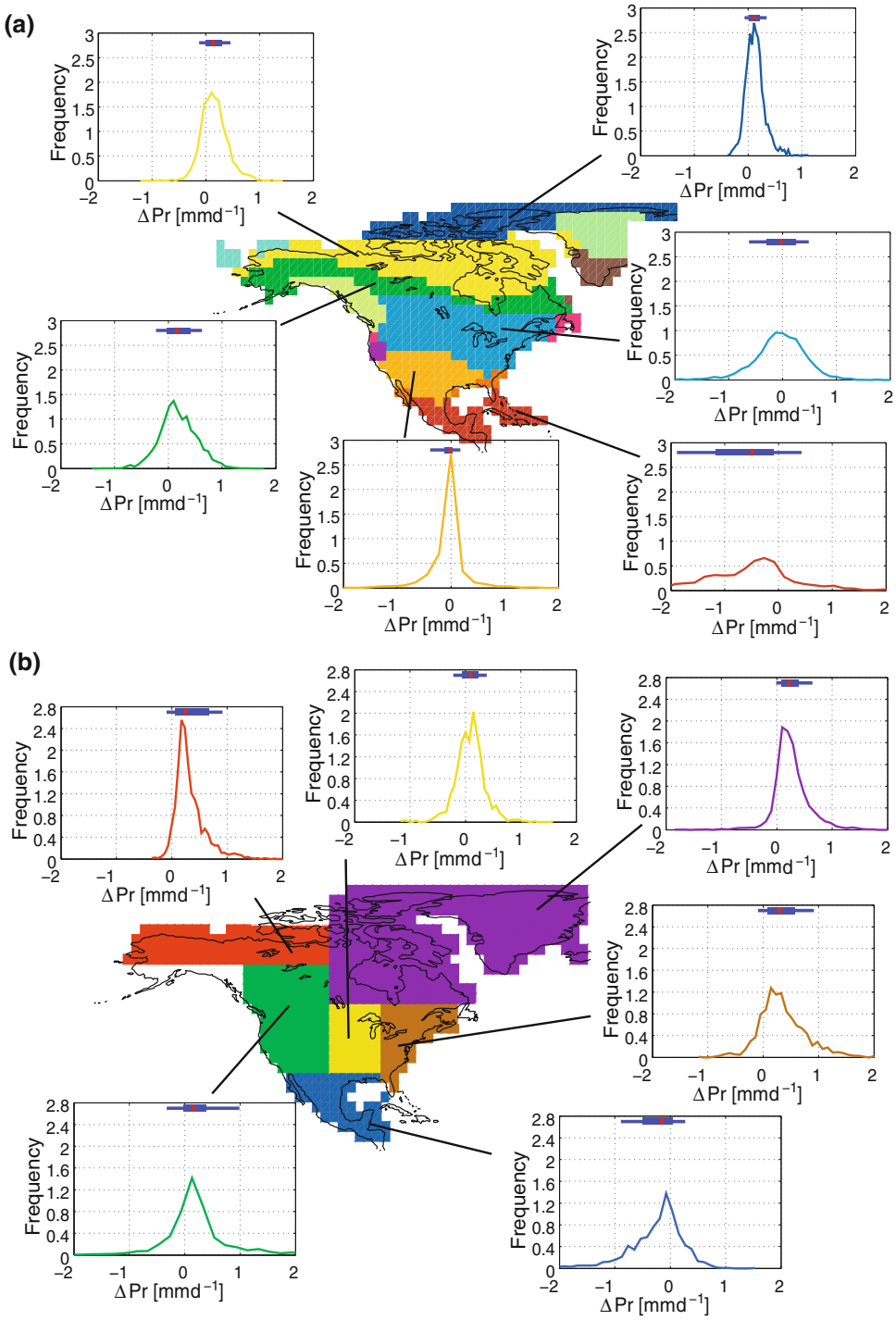

Fig. 11 Same as Fig. 10 except that for each region the distribution of the projected changes in precipitation for each grid cell in the region and across all the models is shown

each model giving a plausible representation of reality is not necessarily true (Tebaldi and Knutti 2007).

Cluster analysis also offers the possibility to combine different aspects of a climate such as temperature and precipitation, two characteristics which are important for impact studies because of their relevance in plant phenology and therefore in ecosystems. Comparing the Köppen classification with the regions defined by cluster analysis we find great similarities. Hence, these regions offer the opportunity to study climate change from an impacts perspective. The proposed regions can be seen as a basis for discussions on the issue whether the old set of regions is still appropriate considering to the improvements that have been made in climate modeling, and whether it is justified to calculate regional climate change projections of different variables with the same set of regions if the pattern of

different variables looks quite differently. Furthermore, it should be noted that the presented algorithm is not limited to the models and data used here. Instead of working with the CMIP3 models it is also possible to apply the same algorithm on regional model output or on data with a different resolution than T42.

Acknowledgments We thank Christof Appenzeller, Jonas Bhend and Martin Jaggi for stimulating discussions. We also acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multi-model dataset. Support of this dataset is provided by the Office of Science, US Department of Energy.

References

Bonfils C, de Noblet-Ducoudre N, Guiot J, Bartlein P (2004) Some mechanisms of mid-holocene climate change in Europe, inferred from comparing pmip models to data. Clim Dyn 23(1):79–98 doi:10.1007/s00382-004-0425-x

Brewer S, Alleaume S, Guiot J, Nicault A (2007a) Historical droughts in mediterranean regions during the last 500 years: a data/model approach. Clim Past 3(2):355–366

Brewer S, Guiot J, Torre F (2007b) Mid-holocene climate change in Europe: a data-model comparison. Clim Past 3(3):499–512

Christensen JH, Hewitson B, Busuioc A, Chen A, Gao X, Held I, Jones R, Kolli RK, Kwon WT, Laprise R, Magaña Rueda V, Mearns L, Menéndez CG, Räisänen J, Rinke A, Sarr A, Whetton P (2007) Regional climate projections. Climate change 2007. In: Solomon S et al (eds) The physical science basis contribution of working group I to the fourth assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, pp 847–845

Furrer R, Knutti R, Sain SR, Nychka DW, Meehl GA (2007a) Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis. Geophys Res Lett 34. doi: 10.1029/2006GL027754

Giorgi F, Bi X (2005) Updated regional precipitation and temperature changes for the 21st century from ensembles of recent AOGCM simulations. Geophys Res Lett 32:L21,715 doi:10.1029/2005GL024288

Giorgi F, Francisco R (2000) Uncertainties in the prediction of regional climate change. Global change and protected areas, pp 127–139

Giorgi F, Mearns LO (2003) Probability of regional climate change based on the reliability ensemble averaging (REA) method. Geophys Res Lett 30:1629. doi:10.1029/2003GL017130

IPCC (2007) Climate change 2007. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) The physical science basis. Contribution of working group I to the fourth assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, p 996

Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comp Surv 31(3):264–323

Kaufmann L, Rousseeuw P (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York

Kottek M, Grieser J, Beck C, Rudolf B, Rubel F (2006) World map of the Köppen–Geiger climate classification updated. Meteorologische Zeitschrift 15(3):259–263 doi:10.1127/0941-2948/2006/0130

Meehl GA, Covey C, Delworth T, Latif M, McAvaney B, Mitchell JFB, Stouffer RJ, Taylor KE (2007) The WCRP CMIP3 multimodel dataset—a new era in climate change research. Bull Am Meteorol Soc 88:1383−1394 doi:10.1175/BAMS-88-9-1383

Philipp A, Della-Marta PM, Jacobeit J, Fereday DR, Jones PD, Moberg A, Wanner H (2007) Long-term variability of daily north Atlantic-European pressure patterns since 1850 classified by simulated annealing clustering. J Clim 20(16):4065–4095 doi: 10.1175/JCLI4175.1

Räisänen J (2001) CO_2-induced climate change in CMIP2 experiments: quantification of agreement and role of internal variability. J Clim 14(9):2088–2104

Räisänen J (2007) How reliable are climate models? Tellus Ser A Dyn Meteorol Oceanogr 59(1):2–29 doi:10.1111/j.1600-0870.2006.00211.x

Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. Phil Trans R Soc A 365:2053-2075 doi:10.1098/rsta.2007.2076

Tebaldi C, Lobell DB (2008) Towards probabilistic projections of climate change impacts on global crop yields. Geophys Res Lett 35(8). doi:10.1029/2008GL033423

Tebaldi C, Mearns LO, Nychka D, Smith RL (2004) Regional probabilities of precipitation change: a Bayesian analysis of multimodel simulations. Geophys Res Lett 31:L24,213 doi: 10.1029/2004GL021276

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

Wang GL (2005) Agricultural drought in a future climate: results from 15 global climate models participating in the IPCC 4th assessment. Clim Dyn 25(7–8):739–753 doi:10.1007/s00382-005-0057-9

Wilks DS (1998) Statistical methods in the atmospheric sciences. Elsevier, New York

Williams JW, Jackson ST, Kutzbach JE (2007) Projected distributions of novel and disappearing climates by 2100 AD. Proc Natl Acad Sci USA 104:5738–5742 doi:10.1073/pnas.0606292104

Zhang XB, Zwiers FW, Hegerl GC, Lambert FH, Gillett NP, Solomon S, Stott PA, Nozawa T (2007) Detection of human influence on twentieth-century precipitation trends. Nature 448(7152):461–464 doi:10.1038/nature06025