Developing a reduced-form ensemble of climate change scenarios for Europe and its application to selected impact indicators

**Martin Dubrovsky · Miroslav Trnka · Ian P. Holman ·
Eva Svobodova · Paula A. Harrison**

Abstract This paper presents a method for identifying a representative subset of global climate models (GCMs) for use in large-scale climate impact research. Based on objective criteria (GCM performance in reproducing the seasonal cycle of temperature and precipitation, and a subset ability to represent future inter-GCM variability), two candidate subsets are selected from a reference set of 16 GCMs. An additional subset based on subjective expert judgement is also analysed. The representativeness of the three subsets is validated (with respect to the reference set) and compared for future changes in temperature, precipitation and Palmer drought index Z (direct validation), and occurrence of the European corn borer and snow-cover characteristics implemented in the CLIMSAVE Integrated Assessment Platform (indirect validation).

The direct validation indicates that one of the objective-based subsets (ECHAM5/MPI-OM, CSIRO-Mk3.0, HadGEM1, GFDL-CM2.1 and IPSL-CM4 models) provides the best choice for the Europe-wide climate change impact study. Its performance is balanced between regions, seasons and validation statistics. However, the expert-judgement-based subset

M. Dubrovsky (✉)
Institute of Atmospheric Physics ASCR, Prague 14131, Czech Republic
e-mail: dub@ufa.cas.cz

M. Dubrovsky · M. Trnka · E. Svobodova
Institute for Agrosystems and Bioclimatology, Mendel University, Brno 61300, Czech Republic

I. P. Holman
Cranfield Water Science Institute, Cranfield University, Cranfield MK43 0AL, UK

M. Dubrovsky · M. Trnka · E. Svobodova
CzechGlobe - Global Change Research Centre AS CR, Brno 60300, Czech Republic

P. A. Harrison
Environmental Change Institute, Oxford University Centre for the Environment, Oxford OX1 3QY, UK

achieved slightly better results in the indirect validation. The differences between the subsets and the reference set are generally much lower for the impact indices compared to their mean (across all GCMs in the subset) changes due to projected climate change. The ranking of the candidate subsets differs between regions, climatic characteristics and seasons, demonstrating that the subset suitability for a specific impact study depends on the target region and the roles of individual seasons and/or climatic variables on the processes being studied.

1 Introduction

Climate change impact analysis often uses an ensemble of climate scenarios to represent some aspects of uncertainty (Holman et al. 2012). The scenario ensemble is usually based on a set of global climate models (GCMs)—commonly taken from the CMIP3 or CMIP5 datasets (Meehl et al. 2007; Taylor et al. 2012) related to the 4th and 5th IPCC's Assessment Reports; IPCC 2007, 2013—or regional climate models (RCMs; usually taken from datasets created within the frame of the PRUDENCE, ENSEMBLES, and CORDEX projects), which are run for one or more emissions scenarios. The choice of climate models involved in a given impact study is based on one of the following approaches: using (1) all available models or all those which provide the climate variables required for a given experiment (e.g. Semenov and Stratonovitch 2010), (2) a subset based on some quality criteria (e.g., Trnka et al. 2004), (3) ensembles of opportunity, when authors use a smaller-sized subset of models and often do not explicitly explain why the selected models were employed (Iglesias et al. 2010; Fronzek et al. 2012).

Having a large number of available climate models, the legitimate question arises as to how to reduce their number for a given study, so that the resultant ensemble of scenarios reasonably represents the inter-model uncertainties. This requirement was faced in the CLIMSAVE project (www.climsave.eu), whose main product is an interactive web-accessible "Integrated Assessment Platform" (IAP), which simulates climate change impacts, vulnerability and adaptation responses based on a series of linked metamodels which are run with climatic data representing present and/or future climates (Harrison et al. 2013, 2014; Holman et al. 2014). The requirement was motivated by the necessity to both reduce the number of GCMs on the user interface so as not to confuse users with excessive GCM choice, and to reduce the runtime for uncertainty analyses of the metamodels across the European target area.

This paper presents a methodology for identifying a representative subset of GCMs that adequately capture GCM uncertainty for use in large-domain climate impact assessments where it is impractical or impossible to include all GCMs. We apply the methodology in the context of creating climate change scenarios for the CLIMSAVE project (see section 'Construction of the climate change scenarios'). The candidate subsets are created by two approaches. The first (see section 'An expert judgement approach to identifying a representative GCM subset') is based on expert judgment, which was employed to create a representative subset in the first stage of the project. Later, a methodology based on the quality of a GCM and the ability of the subset to represent the inter-GCM variability was developed to create two objective-based subsets (see section 'An objective approach to identifying representative GCM subsets'). The representativeness of the candidate GCM subsets is validated in terms of selected climatic characteristics in section 'Direct validation of the GCM subsets' and impact indices implemented in the IAP in section 'Indirect validation of the GCM subsets'. The indirect validation tests focus on climate change impacts for pest occurrence and snow cover in the 2055.

2 Data

The objective methodology for choosing the representative GCM subsets started to be developed at the beginning of CLIMSAVE project (before the IAP was finalised) and it was also anticipated to be valuable in other projects. This gave rise to some differences between the direct validation, which was based on the data configuration used for the GCM selection, and the indirect validation, which was based on the configuration of the IAP that was finalised later during the CLIMSAVE project. Specifically, the target region for the direct validation includes only regions south of 60°N and west of 40°E (focusing on the main European agricultural regions), the climate characteristics relate to 2070–2099 vs 1961–1990 and SRES-A2 emissions, and the 0.5° grid corresponds to the CRU TS 2.1 dataset (Mitchell and Jones 2005) that was used as a reference in validating the GCMs in terms of the seasonal cycle of temperature and precipitation. On the other hand, the indirect validation of the GCM subsets was based on metamodels implemented in the IAP, which uses the CRU CL 2.0 dataset (10′×10′ grid; New et al. 2002) as a reference climate; the tests were performed for only those countries included in the IAP (EU27 plus Norway and Switzerland) and relate to climate change scenarios (for 2055 and SRES-A1B emissions) prepared specifically for the CLIMSAVE project. Despite these differences, the two validation experiments provide complementary information on the representativeness of the GCM subsets.

The analysis is based on GCM simulations made within the frame of the CMIP3 project (Meehl et al. 2007) and applied in the IPCC's Fourth Assessment Report (IPCC 2007). The initial GCM set (Table 1) includes 16 models which had been run for the SRES-A2 emission scenario and have available monthly time series of daily average (TAVG), minimum and maximum temperature, precipitation (PREC) and global solar radiation (SRAD) for the 1961–2099 period. The SRES-A2 simulations were preferred to SRES-A1B (for which more simulations are available), as they provide a larger signal-to-noise ratio allowing for more precise determination of the standardised scenarios. Where multiple runs were available from a given GCM, we used only the first run. The longitudinal and latitudinal resolutions of the GCMs ranges from 1.4 to 5° and 1.25 to 4°, respectively. For creating the subset using the objective algorithm, the GCM data were regridded (using bilinear interpolation) into the common 0.5°×0.5° grid of the reference CRU TS 2.1 data. The objective algorithm was then applied on grid cells found in a region bordered by the 11°W (western border) and 40°E meridians and the 34°N and 60°N parallels, which covers the majority of Europe and includes 3,799 "land" grid cells of the CRU TS 2.1 dataset. In addition to monthly mean climatologies of precipitation and daily maximum and minimum temperatures required by the metamodels, monthly means of global solar radiation that were not available in the CRU dataset were determined from sunshine duration (based on information in Rietveld 1978) and added to the CLIMSAVE datasets.

3 Selection of representative GCM subsets

3.1 Construction of the climate change scenarios

To account for multiple uncertainties, as well as to provide the GCM-based climate change scenarios for those GCM-emissions combinations not available in the source GCM database, the climate change scenarios for the CLIMSAVE project were determined using a pattern scaling approach (Santer et al. 1990; Mitchell 2003), in which the scenario is defined as the product of a standardised scenario and a change in global mean temperature, ΔT_G. The

Table 1 The statistics used to choose the "*best*" GCM (columns *2–4*), the "*central*" GCM (columns *5–7*), and the *most diverse* GCM triplet (columns *8–12*) for Europe. The GCMs are ordered according to their frequency of selection among three best GCMs, the *underlined parts* of the model acronyms (column *1*) are used in the figures. The numbers in columns *2–7* indicate the percentage of 3,799 European grid cells where the GCM is selected among 1–3 best GCMs and 1–3 most central GCMs. Columns *8–11* present the four most frequently selected triplets with numbers indicating percentage of grid cells where the triplets were selected, and column *12* gives the percentages of grid cells where the GCM is in the most diverse GCM triplet

1		2	3	4	5	6	7	8	9	10	11	12
		A. Choosing the "best" GCM			B. Choosing the "central" GCM			C. Choosing the most diverse GCM triplet				
GCM	Modelling centre	Best	Top 2	Top 3	Central	2 central	3 central	1st = OBJ5a	2nd	3rd	4th = OBJ5b	Σ
ECHAM5/MPI-OM	Max-Planck-Institut for Meteorology (Germany)	**25.4**	**41.2**	**51.3**	**17.9**	**34.6**	**49.6**	6.9				1.1
HadCM3	UK Met. Office (UK)	**8.9**	**24.3**	**40.0**	11.2	18.8	27.3					9.4
CGCM3.1 (T47)	Canadian Center for Climate Modelling and Analysis	**13.9**	**23.6**	**31.8**	1.9	8.8	18.6					14.8
HadGEM1	UK Met. Office (UK)	14.3	21.6	26.1	**18.1**	29.5	34.6	6.9			5.4	29.8
GFDL-CM2.0	Geophysical Fluid Dynamics Laboratory (USA)	6.2	15.1	24.1	4.3	12.0	21.9					5.0
MRI-CGCM2.3.2	Meteorological Research Institute (Japan)	5.7	14.9	23.5	0.8	2.4	4.8				5.4	22.7
CSIRO-Mk3.0	CSIRO (Australia)	4.1	11.9	20.7	**22.5**	**38.9**	**49.0**					8.8
GFDL-CM2.1	Geophysical Fluid Dynamics Laboratory (USA)	4.3	10.5	19.9	0.0	0.1	0.3	6.9	6.5	6.0		49.1
NCAR-PCM	National Centre for Atmospheric Research (USA)	4.5	8.0	15.0	2.7	9.4	19.8					10.7
MIROC3.2(medres)	National Institute for Environmental Studies (Japan)	3.3	8.1	13.0	2.0	7.7	15.9			6.0		8.0
BCM2.0	Bjerknes Centre for Climate Research (Norway)	2.7	5.9	8.9	0.1	0.7	1.5		6.5	6.0	5.4	51.4
INM-CM3.0	Institute for Numerical Mathematics (Russia)	2.1	4.4	7.5	2.8	5.3	9.7					10.7
NCAR-CCSM3	National Centre for Atmospheric Research (USA)	1.0	2.6	5.7	3.5	8.5	15.2					12.7
ECHO-G	Met.Inst.Univ Bonn (Germany)+Met. Res.Inst. (Korea)+Model and Data Group at MPI-M (Germany)	1.5	3.2	5.2	9.9	18.2	24.8					17.0
CNRM-CM3	Centre National de Recherches Meteorologiques (France)	1.1	2.4	4.1	2.1	4.8	6.7					17.8
IPSL-CM4	Institute Pierre Simon Laplace (France)	1.0	2.2	3.4	0.1	0.2	0.4	6.9	6.5			30.9

Bold entries shows the three GCMs which are most frequently selected among the three best and three most central GCMs

standardised scenarios, which are given in terms of the changes in climatic characteristics per 1 °C global warming, are derived from 1961 to 2099 GCM-simulated monthly series using a linear regression approach (Dubrovsky et al. 2005), in which the GCM-simulated change in global mean temperature is an independent variable, and the simulated grid cell-specific change in a given climatic variable is a dependent variable. The standardised temperature and precipitation changes combined across the five GCMs selected for CLIMSAVE are presented later (see section 'Direct validation of the GMC subsets'). The change in global mean temperature is determined by the MAGICC (version 5.3) model (Harvey et al. 1997; Hulme et al. 2000) run for a given emission scenario and climate sensitivity. The final set of scenarios for CLIMSAVE is obtained by combining multiple ΔT_G values (changes with respect to 1975 are assumed here) with a selected GCM subset. The applied set of ΔT_G values—Table 1 of the electronic supplementary material (ESM)—is based on combining three climate sensitivities (low = 1.5 °C, medium = 3.0 °C, high = 4.5 °C), four emission scenarios (SRES-A1B, SRES-A2, SRES-B1, SRES-B2), and two target periods (2025 and 2055).

The climate change scenarios within the IAP consist of percentage changes in daily sums of precipitation and solar radiation, and additive changes in daily averages, minima and maxima of temperature. The changes are defined for 12 months and four seasons of the year, and for 23,871 grid cells taken from the reference CRU CL 2.0 dataset ($10' \times 10'$ grid). In downscaling the coarse-resolution GCM-based scenarios to the 10′ grid, we used a simple linear interpolation assuming that the changes in individual climate characteristics have much smaller spatial variability than that of their absolute values. When defining the future-climate values of the relevant climatic characteristics for individual grid cells, the reference-climate (based on CRU CL 2.0 data for 1961–1990) temperature characteristics are modified additively, and the precipitation and solar radiation characteristics are modified multiplicatively. In applying the multiplication modification, we used an approach (illustrated in Figure 1 of the ESM), in which the scaling of the standardised change towards lower values (when $\Delta_S X < 0$) is exponential and the scaling towards higher values (when $\Delta_S X > 0$) is exponential until a 1 °C rise after which the scaling is linear with the slope equalling the derivative of the exponential rise at $\Delta T_G = 1$ °C. This approach helps to avoid excessive decreases (if $\Delta_S X$ is negative) or increases (if $\Delta_S X$ is positive) for higher ΔT_G values.

To prevent overloading the IAP users with excessive GCM choices, we reduced the number of GCMs included in the IAP to five, which we subjectively considered to be a reasonable number to represent the inter-GCM uncertainty. Two approaches to selecting a GCM subset based on the standardised scenarios derived from the SRES-A2 simulations by 16 GCMs are described in the following sections.

3.2 An expert judgement approach to identifying a representative GCM subset

A subset of five GCMs was identified based on expert judgement, through an iterative process that focused on GCM-based climate change scenarios. Three experts from the CLIMSAVE project team independently assessed maps of annual and seasonal standardised changes in temperature and precipitation from the available GCMs (changes related to summer and winter are shown in Figure 2 of the ESM), subjectively choosing a subset which each believed to best capture the range of variability (magnitude and spatial patterns) across these two variables. No quantitative criterions were employed in this procedure. After this first stage, all three experts agreed on three GCMs (HadGEM1, NCAR-CCSM3 and GFDL-CM2.1). After discussion, it was agreed to add MIROC3.2(medres), which got two votes in the first round, due to it being a relatively wet and hot model with a strong northeast to southwest gradient. Finally, in order to

complete the set, the ECHAM5/MPI-OM model (no votes in the first round) was selected, as a less extreme central model from a leading European modelling centre. The final selection, consisting of HadGEM1, NCAR-CCSM3, GFDL-CM2.1, MIROC3.2(medres) and ECHAM5/MPI-OM (hereafter referred to as EXP5 subset) combined broad representation of model variability and leading modelling centres from different parts of the world.

3.3 An objective approach to identifying representative GCM subsets

In the first stage of the screening procedure, a subset of five GCMs was identified for each European 0.5×0.5° grid cell in three steps, consisting of: (1) the "best" GCM at reproducing the observed seasonal cycle of temperature and precipitation, (2) the "central" GCM which is closest to the multi-GCM mean climate change scenario, and (3) the most diverse GCM triplet which, together with the "best" and "central" GCMs, would best represent the inter-GCM climate change scenario variability. The three steps were motivated by the most common requirements coming from climate change impact researchers, who require scenarios based on the "best" model and/or an "average" scenario and/or a reasonably-sized set of scenarios.

The "best" GCM For the purpose of the CLIMSAVE project, where input climatological data consist of surface weather characteristics, the GCM quality metric is based on the ability of the GCM to reproduce the reference (1961–90) seasonal cycles of temperature and precipitation derived from the CRU TS 2.1 dataset. The metric is defined by a Q score, which is based on the RV scores (reduction of variance) determined separately for temperature and precipitation:

$$Q(\text{GCM}) = \sum_{X\in\{T,P\}} \frac{[\text{RV}_{\text{GCM}}(X)- <\text{RV}(X)>] \times |\text{RV}_{\text{GCM}}(X)- <\text{RV}(X)>|}{\text{var}[\text{RV}(X)]}$$

where $\text{RV}_{\text{GCM}}(X) = 1-\sum_{m=1\ldots12}(X^*_{\text{GCM},m}-X_{\text{CRU},m})^2/\sum_{m=1\ldots12}(<X_{\text{CRU},m}>-X_{\text{CRU},m})^2)$, X relates to temperature and precipitation, $X^*_{\text{GCM},m}$ are the debiased GCM simulated monthly mean temperature and monthly precipitation sum, $X_{\text{CRU},m}$ are monthly means derived from the CRU data, and $<\bullet>$ and $\text{var}[\bullet]$ are the average and variance over all GCMs. The highest Q-score indicates the "best" GCM in each grid cell. Figure 1 shows a map of the "best" grid cell-specific GCMs, whilst Table 1 (3 columns in section A) shows the frequencies of individual GCMs being selected as among the best models.

The "central" GCM To find the GCM which is closest to the multi-GCM mean climate change scenario, each GCM is represented by an eight-dimensional position vector R_{GCM}, whose co-ordinates are normalised (using grid cell-specific average and standard deviation over all GCMs) GCM-simulated standardised changes in seasonal means of temperature and precipitation. Euclidean distance is then calculated between two GCMs: $d(\text{GCM}_1, \text{GCM}_2) = \sum_{i=1,\ldots,8}(r_{1i}-r_{2i})^2$, where r_{1i} and r_{2i} are components of the GCMs position vectors. The "central" GCM is selected as a GCM which has the smallest distance from the centroid of all GCMs ($R_{\text{centroid}} = \sum_{\{\text{all GCMs}\}} (R_{\text{GCM}})/n$, where $n=16$ is a number of all GCMs). Figure 1 shows a map of the grid cell-specific "central" GCMs, and Table 1 (3 columns in section B) shows the number of grid cells in which each individual GCM is among the one (2, 3) GCMs closest to the centroid.

The most diverse GCM triplet consists of the three GCMs which maximise the sum of distances between the three GCMs [$\text{SUM}_{ijk} = d(\text{GCM}_i, \text{GCM}_j) + d(\text{GCM}_i, \text{GCM}_k) + d(\text{GCM}_j,\text{GCM}_k]$ over all possible triplets. Figure 3 of the ESM shows the grid cell-specific
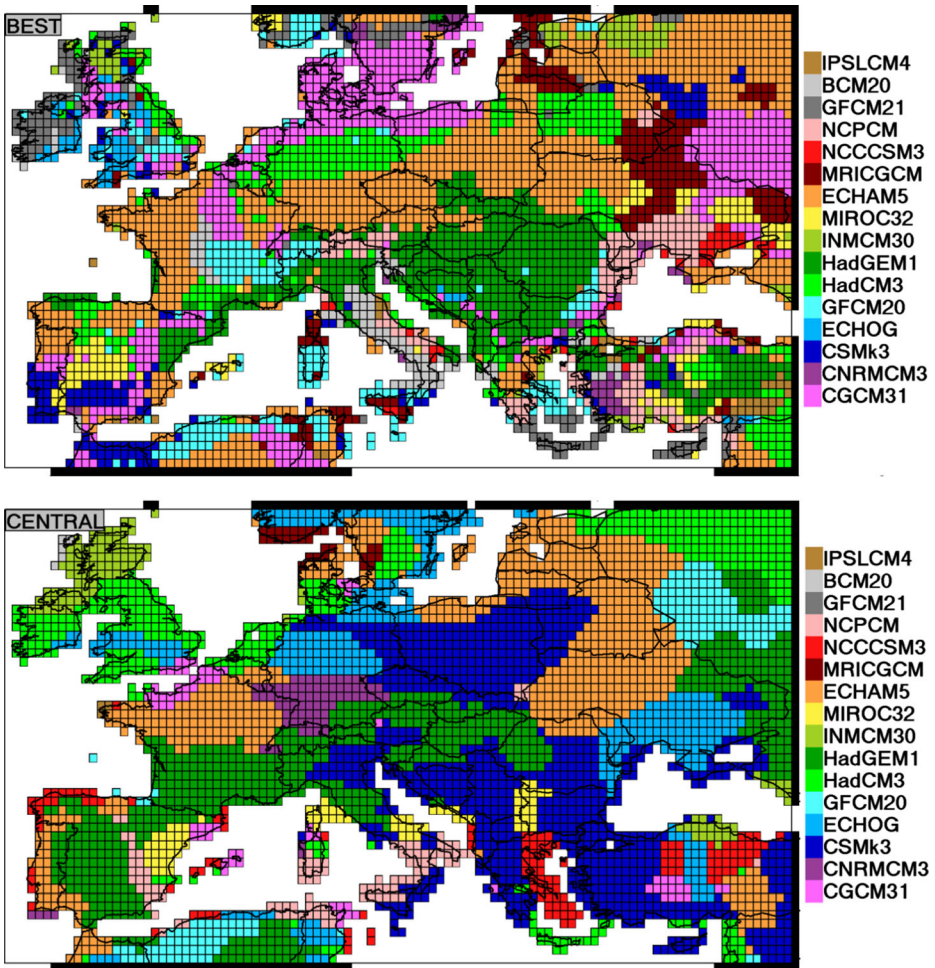
Fig. 1 Maps of the grid cell-specific "*best*" GCM [model with the highest Q-score (Eq. 1); *top*] and "*central*" GCM (model with the lowest distance from the all-GCM centroid; *bottom*)

triplets for the whole of Europe, and Table 1 (columns in section C) shows the frequency (number of grid cells) of the four most frequently chosen GCM triplets in Europe. As a consequence of its definition, the most diverse triplet (if applied as a representative GCM subset on its own) would seemingly overestimate inter-GCM variability, but this overestimation is suppressed by adding the "best" and "central" GCMs into the overall representative GCM subset.

Note that the GCMs selected in the preceding three categories relate to the metrics used to define the GCM quality and the distance between GCMs, both of which may be defined differently by other authors.

As a result of the screening procedure, the selected grid cell-specific subset consists of five GCMs in most grid cells. Occasionally, when the "best" GCM is the same as the "central" GCM or one of the most diverse triplet, it consists only of four GCMs. In the second stage of the procedure, we analysed the frequency of occurrence (across all 3,799 land grid cells) of the

individual GCMs selected in the "best", "central" and "the most diverse triplet" categories (Table 1) and defined the set of five GCMs that were most representative for the whole of Europe:

1. ECHAM5/MPI-OM is selected as the "best" GCM because it is found to be the "best" GCM in the largest number of grid cells (it is among the 3 "best" GCMs in about half of all grid cells).
2   CSIRO-Mk3.0 is selected as the "central" GCM because it is found to be the "central" GCM in the largest number of grid cells (and is among the three most central GCMs in about half of all grid cells). Interestingly, our "best" model (ECHAM5/MPI-OM) is also among the three most central GCMs. This fact corresponds well to the earlier finding (Gleckler et al. 2008) that the "mean model" consistently outperforms other models.
3. Two versions of the most diverse triplet were selected for further tests. The first one, consisting of HadGEM1, GFDL-CM2.1 and IPCM4, is the most frequently selected triplet, and is based strictly on maximising the diversity among GCMs. The second triplet (HadGEM1, MRI-CGCM2.3.2, BCM2.0) also accounts for the quality of the GCMs. It consists of GCMs which together exhibit the best quality among the ten most frequently selected triplets (in terms of the sum of their frequencies of being selected as the "best" grid cell-specific GCM).

    As a result of the aforementioned procedure, two candidate subsets emerge: OBJ5a = (ECHAM5/MPI-OM, CSIRO-Mk3.0, HadGEM1, GFDL-CM2.1, IPCM4) and OBJ5b = (ECHAM5/MPI-OM, CSIRO-Mk3.0, HadGEM1, MRI-CGCM2.3.2, BCM2.0). Note the relatively small differences between individual subsets: OBJ5a differs from both OBJ5b and EXP5 only by two GCMs.

4 Direct validation of the GCM subsets

The "direct" validation of the candidate subsets examined their ability to reproduce the multi-model means and standard deviations (using all 16 GCMs as the reference set, which will be referred to as ALL16) of projected changes (2070–2099 vs. 1961–1990) in temperature (ΔTAVG), precipitation (ΔPREC) and drought in individual grid cells as well as for Europe as a whole. Drought is represented by the relative Z-index (rZ), which is an intermediate product of the relative Palmer Drought Severity Index (PDSI) model obtained by minor modification (Dubrovsky et al. 2009) of the self-calibrated PDSI (Wells et al. 2004). To characterise the change in drought conditions, the underlying PDSI model is first calibrated using the reference climate data, and then applied to future climate data. As a result, the rZ value characterises the future (2070–2099) soil moisture content anomalies with respect to the reference (1961–1990) normal conditions. The values above/below zero indicate wetter/drier conditions compared to the reference normal conditions, the <−2.7, 2.7> interval encompasses about 96 % of rZ values. The rZ index was chosen as it accounts for both temperature and precipitation changes and allows validation of the subsets for more complex climatic characteristic which are dependent on multiple climatic variables.

    The validation tests focus on the multi-GCM means and standard deviations of ΔTAVG, ΔPREC and rZ related to individual seasons. Reproduction of the multi-GCM means is quantified by the mean subset bias (defined as a difference between the means of the subset

vs. ALL16; $B = A_{subset} - A_{ALL16}$), and the percentage of grid cells where B is higher (lower) than B_T ($-B_T$) threshold, where $B_T = 0.5 \times S_{ALL16}$ and S_{ALL16} is the multi-GCM standard deviation of the climatic characteristic based on all 16 GCMs. Similarly, validation of the inter-GCM variability is based on comparing the multi-GCM standard deviations derived from the subset, S_{subset}, with S_{ALL16}: we count the percentage of grid cells where the ratio of the subset-to-ALL16 standard deviations, $R = S_{subset}/S_{ALL16}$, is higher (lower) than 3/2 (2/3). The preceding thresholds subjectively represent a significant degree of mismatch that simultaneously provides sufficient frequency of exceedance to enable an efficient comparison of performances of individual subsets.

The quality of the GCM subsets in terms of their ability to reproduce the means and variability of ΔPREC, ΔTAVG and rZ is shown in the top half of Table 2 and Table 2 of the ESM and Fig. 2 and Figure 4 of the ESM. The spatial patterns of standardised changes in PREC and TAVG for summer and winter based on the OBJ5a subset (the subset, which was finally selected for use in CLIMSAVE) and the biases with respect to the ALL16 set are shown in Fig. 3.

The results of the direct validation (see section 2 of the ESM for details) show that the subsets determined using the objective criteria perform better than those based on expert judgement. The OBJ5a subset is the best when we require a Europe-wide performance that is balanced between regions, seasons, climate change characteristics and statistics (multi-GCM means and standard deviation). When summed over ΔTAVG, ΔPREC and rZ, OBJ5a shows the lowest percentage of seasonal exceedances across Europe for both the subset mean and inter-GCM variability. The spatial pattern of its performance is similar over the whole of Europe, albeit with possible weaknesses in regions to the north of the Black Sea and the Iberian Peninsula. However, the results also show that the optimal subset may be different if the stress is put on a specific climate characteristic, region, and/or season. The performance of the EXP5 subset is most significantly affected by underestimating temperature variability in summer and autumn. However, its performance is almost as good as the best (OBJ5a) of the objective-based candidate subsets, due to the two subsets having three common GCMs. Furthermore, the EXP5 subset was found to be the best in terms of some partial criteria, as well as the best subset for the winter season.

5 Indirect validation of the GCM subsets

The indirect validation shows how imperfections in the subsets representation of the inter-GCM variability of ΔPREC and ΔTAVG affect the ability of the subsets to capture the ALL16 central value and variability of changes in the impact indices in each 10′ grid cell. Four indices from two climatically driven metamodels implemented in the IAP were selected: (1) ONEI (ecoclimatic index) describes the overall suitability of climate conditions for the establishment and long-term presence of a pest population (the European corn borer, *O. nubilalis*) in a given grid cell. ONEI ranges from 0 to 100, where ONEI ≥25 indicates very favourable climate conditions for species occurrence, 10≤ONEI <25 as favourable and ONEI <10 as limiting for species survival (Hoddle 2003). (2) ONNG is the number of generations of the corn borer that can be completed within a year. (3–4) SNOW3 and SKI10 are the average number of days in a year with more than 3 and 10 cm of snow (corresponding to 3 and 10 mm, respectively, of snow water equivalent), which would provide frost protection for crops and allow skiing activities, respectively. While the response of the pest metamodel is driven primarily by temperature changes, the snow cover model is driven by a combination of changes in precipitation and temperature in the cold part of the year. More details on the two metamodels

Table 2 Results of the direct (*top part*) and indirect (*bottom*) validation tests. The *direct validation* gives averages—over 4 seasons (DJF, MAM, JJA, SON)—of the validation statistics for comparing multi-GCM seasonal climate characteristics—future changes in seasonal precipitation (ΔPREC) and temperature (ΔTAVG), and seasonal means of future-climate rZ—based on the GCM subset vs. ALL16 set. $B=A_{subset}−A_{ALL16}$ is a multi-GCM mean bias between the grid cell-specific climate characteristic calculated for a subset minus ALL16 set, $R=S_{subset}/S_{ALL16}$ is a ratio of subset to ALL16 standard deviations of the climatic characteristics, $<B>$ and $S(B)$ are averages and standard deviations (over all 3,799 European grid boxes) of the bias; columns 3–5 relate to percentages of the grid cells where B is lower than zero[a], lower than $−B_T$, and larger than B_T ($B_T = 0.5×S_{ALL16}$), respectively; similarly, columns 6–8 present percentages of grid cells where R is below 1[a], below 2/3 and above 3/2, respectively. The numbers in the "*optimal value*" row give the value of the statistic which would indicate a perfect fit between the subset and ALL16 set. The *AVG* rows show averages of the significant exceedances of ΔPREC, ΔTAVG and rZ. The *indirect validation* gives values of the same validation statistics (comparing the subset-based vs. ALL16-based characteristics), except for the changes in four impact indices

Validation statistic →	1 $<B>$	2 $S(B)$	3 $P^*(B<0)$[a] [%]	4 $P(B<−B_T)$ [%]	5 $P(B>B_T)$ [%]	6 $P(R<1)$[a] [%]	7 $P(R<2/3)$ [%]	8 $P(R>3/2)$ [%]
Optimal value →	0	0	50	0	0	50	0	0
Direct validation								
OBJ5a								
ΔPREC [%]	−2.90	5.34	71.3	26.4	5.2	**44.4**	**13.6**	**2.8**
ΔTAVG [°C]	**0.13**	**0.23**	**32.5**	**2.1**	**16.8**	29.4	**4.2**	**2.2**
rZ	**−0.10**	**0.29**	**64.1**	19.9	5.8	**45.1**	**12.5**	**5.6**
AVG				**16.1**	**9.2**		**10.1**	**3.6**
OBJ5b								
ΔPREC [%]	**0.52**	4.40	**44.1**	**8.0**	**12.8**	59.6	19.1	1.6
ΔTAVG [°C]	−0.34	0.31	80.8	41.6	0.5	**56.2**	18.5	0.5
rZ	0.11	0.26	34.3	**5.3**	**18.2**	58.8	23.3	3.5
AVG				18.3	10.5		20.3	1.8
EXP5								
ΔPREC [cuo]	−2.66	5.02	70.2	26.3	5.0	44.3	13.3	4.3
ΔTAVG [°C]	0.35	0.26	10.0	0.0	34.3	70.7	32.4	1.4
rZ	−0.14	**0.23**	70.8	20.6	3.0	60.5	23.3	2.6
AVG				15.6	14.1		23.0	2.7
Indirect validation								
OBJ5a								
ONEI	−0.07	2.16	40.8	9.4	10.7	36.5	10.5	5.4
ONNG	0.00	0.19	**49.7**	2.2	2.9	**46.7**	**9.9**	**3.0**
SNOW3 [days]	1.02	2.30	33.5	1.4	14.2	**43.8**	7.5	6.2

Table 2 (continued)

Validation statistic →	1 $\langle B \rangle$	2 $S(B)$	3 $P^*(-B < 0)^a$ [%]	4 $P(-B < -B_\tau)$ [%]	5 $P(-B > B_\tau)$ [%]	6 $P(-R < -1)^a$ [%]	7 $P(-R < 2/3)$ [%]	8 $P(-R > 3/2)$ [%]
SKII10 [days]	1.43	**2.66**	32.6	**1.3**	**13.4**	**41.1**	7.2	6.0
AVG				3.6	10.3		8.8	5.2
OBJ5b ONEI	**−0.06**	**0.95**	**54.5**	7.9	6.7	83.0	38.0	0.1
ONNG	−0.02	0.19	51.3	3.9	1.2	58.6	19.5	0.8
SNOW3 [days]	**−0.43**	3.47	**61.4**	21.0	4.7	41.8	14.1	2.3
SKII10 [days]	0.02	3.33	**59.0**	20.1	3.9	39.7	13.4	2.0
AVG				13.2	4.1		21.2	1.3
EXP5a ONEI	−0.38	1.57	61.2	**10.1**	**2.6**	**58.4**	**10.0**	**4.7**
ONNG	**−0.01**	**0.14**	50.4	**2.1**	**1.4**	**46.5**	9.1	5.1
SNOW3 [days]	0.63	**2.53**	36.4	4.1	9.9	37.6	**5.3**	**3.8**
SKII10 [days]	0.88	3.02	35.0	4.8	11.1	33.9	**3.7**	**5.1**
AVG				5.3	6.2		7.0	4.7

[a] When calculating the percentage of grid cells with below-zero and below-one values of B and R, we add the half of the grid cells, where the subset bias B is zero, or R ratio is 1, respectively: $P^*(-B < 0)$ is defined here as $P(B < 0) + 0.5\,P(B = 0)$, and $P^*(-R < 1) = P(R < 1) + 0.5\,P(R = 1)$. That is, the value in column 3 or 6 would be 50 if all values of B are zero, or all values of R are 1, respectively

Bold entries indicates the best fit (of the three subsets) for a given climate characteristic and statistic; in this case, we summed the above-threshold and below-threshold significant exceedances (columns 4 and 5 for the averages and columns 7 and 8 for the standard deviations) and compared the subsets based on these sums
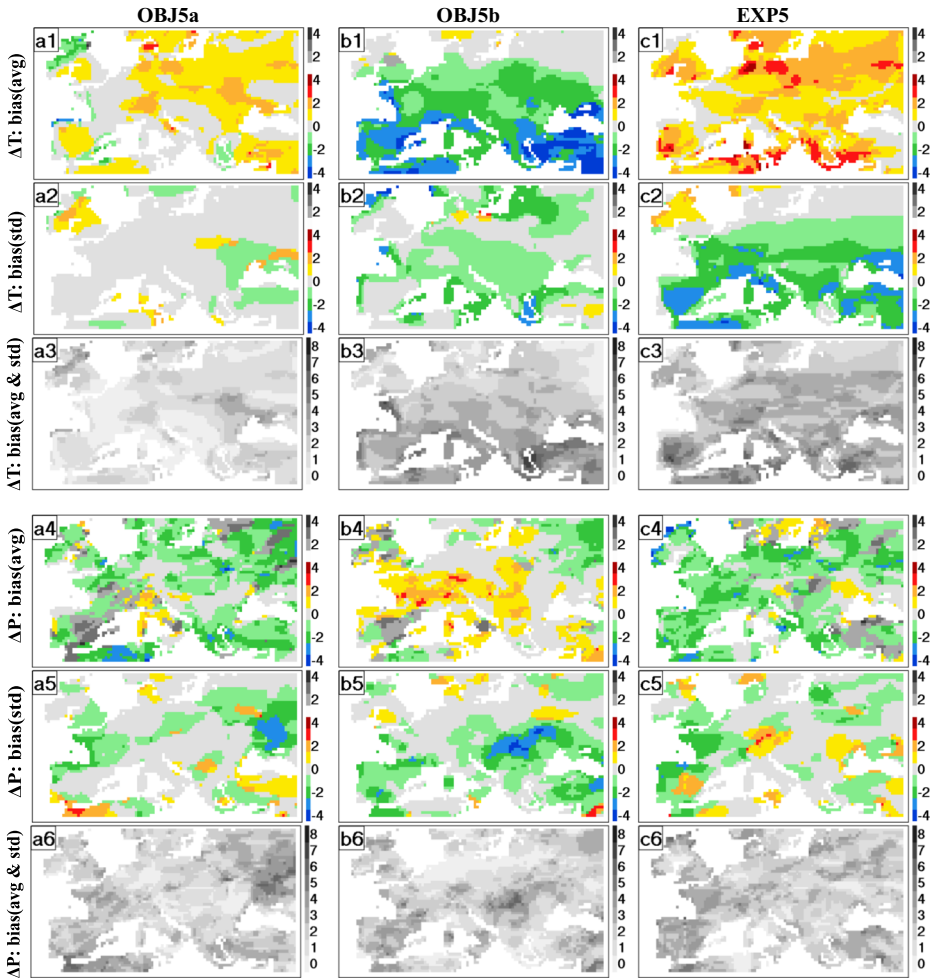
Fig. 2 Comparison of the three subsets vs. ALL16 set in terms of the multi-GCM mean and standard deviation of future temperature (*top three rows*) and precipitation (*bottom three rows*) changes. *First rows* in the *top* and *bottom* blocks: number of seasonal exceedances of the subset mean outside $<A_{ALL16} \pm 0.5 \times S_{ALL16}>$ interval; *middle rows in each block*: number of seasonal exceedances of the subset standard deviation outside $<2/3 \times S_{ALL16}, 3/2 \times S_{ALL16}>$ interval; *bottom rows in each block*: sum of the *first* and *second rows*. Colours in *the first and second rows* show the number of exceedances are of the same sign (positive: *yellow* to *red*; negative: *green* to *blue*), whilst the *grey scale* indicates that the given number of exceedances are not all of the same sign

are given in SM1. The metamodels were run for the baseline and 2055 target period assuming SRES-A1B emissions and 3.0 °C climate sensitivity.

The statistics (analogous to those used in the direct validation) quantifying the reproduction of the grid cell-specific multi-GCM means and standard deviations of changes in the four impact indices are summarised in the lower half of Table 2, the spatial patterns of the impact indices for reference and future (ALL16 mean) climates and subset vs. ALL16 biases are shown in Fig. 4 and Figure 5 of the ESM, whilst the significant subset vs. ALL16 deviations (aggregated over changes in all four indices) are shown in Fig. 5.
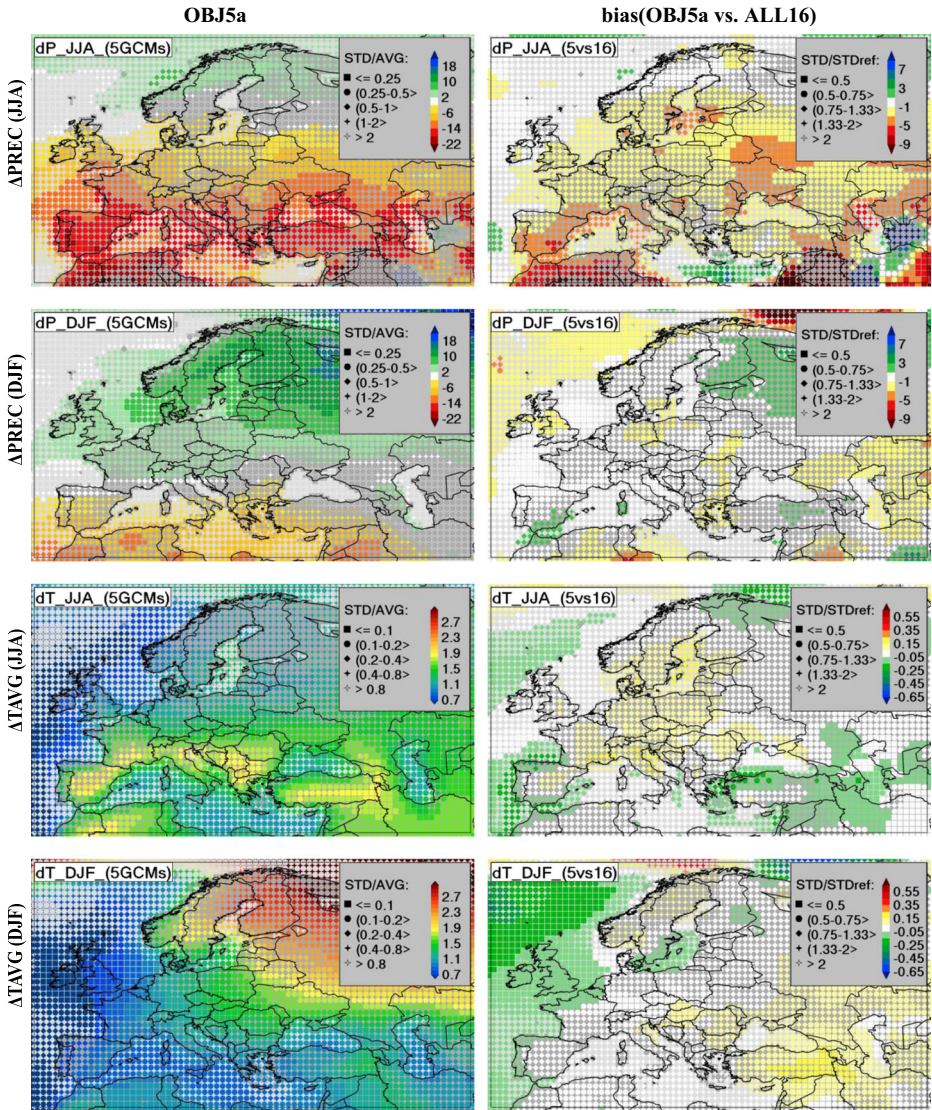
Fig. 3 The standardised climate change scenario (regridded into 1 × 1° resolution) based on OBJ5a subset (*left*) and the subset bias with respect to the ALL16 set (*right*). The maps relate (from *top to bottom*) to the changes in summer precipitation sum, winter precipitation sum, summer mean temperature and winter mean temperature. In the *left hand column*, the shape of the grid cell-specific symbols relate to the inter-GCM variability (represented by *STD/AVG* ratio, where AVG and STD are average and standard deviation across the five GCMs of OBJ5a) of the displayed characteristic. In the *right hand column*, the shape of the symbols represents the ratio of the multi-GCM standard deviations of the displayed characteristic based on the subset vs. ALL16 set

The results of the indirect validation (see section 5 of the ESM for details) show EXP5 to be the best-performing subset (based on the statistics in the lower half of Table 2). The OBJ5a subset exhibits a slightly poorer performance, and OBJ5b shows the largest portion of grid cells in which the subset-based results significantly differ from the reference ALL16 set. Of the
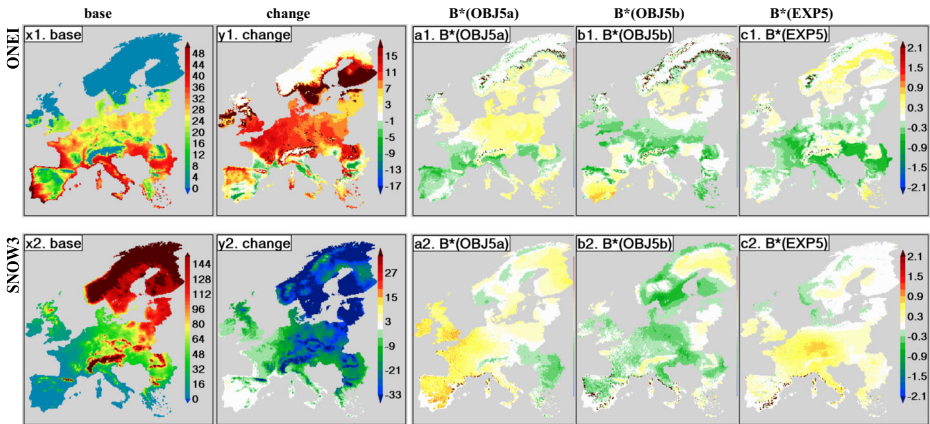
Fig. 4 The spatial pattern of the two impact indices. The *top row* relates to *ONEI* [dimensionless index] and the *bottom row* to SNOW3 [days]. From *left* to *right*, the maps show: "base" = the reference climate values; "change" = changes due to climate change (2055 vs 1961–1990, SRES-A2, middle climate sensitivity, mean over 16 GCMs); $B*$(OBJ5a), $B*$(OBJ5b) and $B*$(EXP5) = the significance of a subset's climate change impact bias defined as a B/S_{ALL16} ratio, where B is the subset mean bias (the difference between mean changes in impact indices averaged over the subset vs. ALL16 set) and S_{ALL16} is the standard deviation of the future-climate impact index based on all 16 GCMs in the reference set; the values outside ±0.55 approximately indicate statistically insignificant (0.05 level) biases

two better subsets, both EXP5 and OBJ5a show some regions of weakness: EXP5 underestimates the reduction in snow cover in Central Europe and underestimates changes in pest indices in the rest of the continental Europe, OBJ5a shows more pronounced (compared to ALL16) decreases in snow cover in the Alps and Balkan Peninsula along with major increases in the UK. Considering the superiority of OBJ5a in the direct validation, the superiority of EXP5 in the indirect validation may be related to the following reasons. Firstly, different regions were involved in the direct and indirect validation: e.g. Turkey and parts of the former Soviet Union, where EXP5 performed worse in the direct validation, were not included in the indirect validation (as they are outside of the CLIMSAVE study area). On the other hand, the direct validation tests did not include Europe north of 60°N latitude, where EXP5 outperforms OBJ5a. Secondly, OBJ5a's performance is partly reduced by the grid cells in Scandinavia, where the ONEI values are identified as significantly different as a result of one GCM outlier (with respect to the other members of the subset).

6 Summary and conclusions

This paper presented two methodologies for identifying a representative subset of GCMs for use in climate change impact studies. One candidate subset (EXP5) was selected using expert judgment, whilst two other subsets were selected based on quantitative criteria: GCMs performance in reproducing the reference seasonal cycle of temperature and precipitation, and an ability of the subset to represent future inter-GCM variability. The three candidate subsets were validated and mutually compared in terms of future changes in climatic characteristics (direct validation) and impact indices (indirect validation). As a part of the indirect validation, climate change impacts on the European corn borer and snow cover have been briefly assessed. The main results of the present experiments may be summarised in the following points.
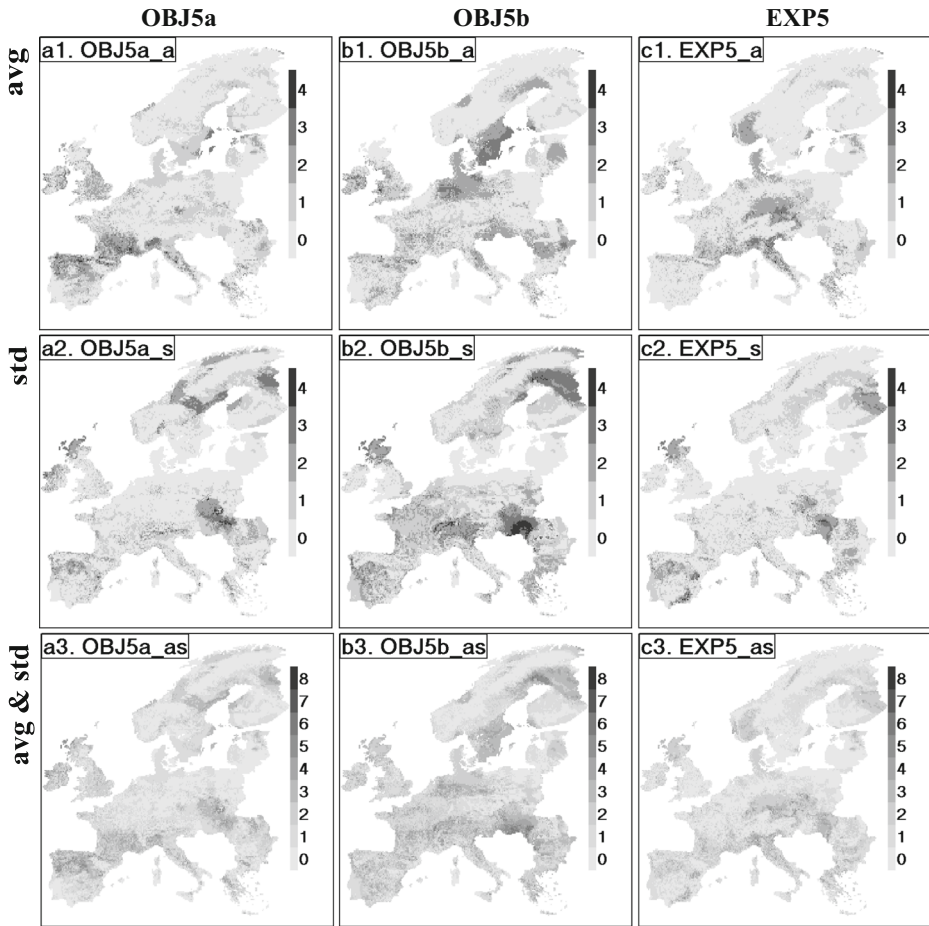
Fig. 5 Comparison of the three subsets vs ALL16 set in terms of the multi-GCM mean and standard deviation of changes in the four impact indices. *Top row*: number of exceedances of the subset mean outside $<A_{ALL16} \pm 0.5 \times S_{ALL16}>$ interval; *middle row*: number of exceedances of the subset-based standard deviation outside $<2/3 \times S_{ALL16}, 3/2 \times S_{ALL16}>$ interval; *bottom row*: total number of exceedances (= sum of the *top* and *middle rows*)

The results of the direct validation tests, which focused on the ability of the subsets to reproduce the multi-GCM variability of changes in seasonal means of precipitation, temperature, and drought conditions, indicated the objective criteria-based OBJ5a subset to be the best choice. The performance of this subset, which produced slightly better results than the other two subsets, is balanced between regions, seasons, climate variables (ΔTAVG and ΔPREC) and statistics (multi-GCM means and standard deviation). However, the results also show that the ranking of the three candidate subsets differs between climatic characteristics, seasons and region, so that the suitability of a given subset for a specific impact study would depend on the choice of the target region and the roles of individual seasons and/or climatic variables on the processes being studied.

This expectation was confirmed in the indirect validation tests, in which the performance of the subsets was assessed in terms of the changes in impact indices derived from the outputs of metamodels fed by the GCM-based climate scenarios: the EXP5 subset achieved slightly

better ranking compared to the OBJ5a subset. However, the OBJ5a's score was partly reduced by the rather random significant differences between the subset vs ALL16 values of the pest indices in Scandinavia, and by the different domains involved in the direct and indirect validation tests. Nevertheless, the differences between the subsets and the reference ALL16 set were generally much lower compared to the mean (across all GCMs in the subset) changes in impact indices due to projected climate change.

Though the results of the direct validation favoured the subset based on an objective methodology, the performance of the expert-judgment-based subset was found to be similar to OBJ5a in several aspects. The main problem with the EXP5 subset related to its underestimation of inter-GCM variability in ΔTAVG, especially in autumn. Despite the results of the direct validation, the EXP5 exhibited the best results if summarised across the whole of Europe and the four impact indices. Though based on given quantitative criteria, the objective methodology provided outputs (in terms of the selected subsets and their biases) which may differ according to the strategies employed to select the GCMs based on their frequencies of being among the best, central, and/or most diverse models. This is demonstrated by the differences between OBJ5a and OBJ5b, which imply mutually different spatial patterns of the subset biases.

It should be born in mind that our objective methodology involves several features which could be specified differently and imply different GCM subsets. These features include, firstly, the number of GCMs in a subset, where we chose five to partly cover the most common demands of users for having scenarios which are based on the best model, mean model and the representation of modelling uncertainty. Secondly, our choice to quantify the GCM quality by its ability to represent seasonal means of precipitation and temperature (which are the most common climatic characteristics used as an input to impact models), whereas other approaches to quantify the GCM performance may be used. For example, Evans et al. (2014) ranked GCMs using the fractional demerit score, which was based on 11 validation statistics used in experiments made by other authors; obviously, this score, or any single validation statistic (see also Gleckler et al. 2008) could serve to identify the "best" GCM. Thirdly, the metrics used to quantify the between-GCM distance, which we based on future changes in temperature and precipitation, could be defined differently using other climatic characteristics such as Bishop and Abramowitz (2012) measure based on the covariance in model errors to select the subset of the most independent models. Fourthly, a different strategy might be used instead of our procedure based on choosing the "best", the "central" and "the most diverse GCM triplet" in three successive steps; for example, the worst performing models might be excluded in the first step (e.g. Evans et al. 2014).

To summarise, there is a need to represent GCM uncertainty in impact studies, but the computational costs of applying impact models at detailed spatial resolutions across large domains such as Europe means that it is often not possible to run all climate models; hence, there is a need to identify representative subsets. We have developed, applied and tested a novel objective approach to solving this problem, which we have implemented in the CLIMSAVE IAP (www.climsave.eu/iap). Our objective approach has been compared with the results of a subjective, expert-judgement driven approach. No method is uniformly superior, but our objective method has the strong advantage of being transparent and repeatable, and will be applicable to new impact studies using updated GCM/RCM ensembles, possibly with some modifications related to the settings of the specific impact study such as the number of GCMs in a subset, metrics for model quality and inter-model distance, and the strategy of selecting the subset.

References

Bishop CH, Abramowitz G (2012) Climate model dependence and the replicate Earth paradigm. Climate Dynam 41:885–900

Dubrovsky M, Nemesova I, Kalvova J (2005) Uncertainties in climate change scenarios for the Czech Republic. Clim Res 29:139–156

Dubrovsky M, Svoboda MD, Trnka M, Hayes MJ, Wilhite DA, Zalud Z, Hlavinka P (2009) Application of relative drought indices in assessing climate-change impacts on drought conditions in Czechia. Theor Appl Climatol 96:155–171

Evans JP, Ji F, Lee C, Smith P, Argüeso D, Fita L (2014) Design of a regional climate modelling projection ensemble experiment – NARCliM. Geosci Model Dev 7:621–629

Fronzek S, Carter TR, Jylhä K (2012) Representing two centuries of past and future climate for assessing risks to biodiversity in Europe. Glob Ecol Biogeogr 21:19–35

Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. J Geophys Res 113(D6): D06104. doi:10.1029/2007JD008972

Harrison PA, Holman IP, Cojocaru G, Kok K, Kontogianni A, Metzger MJ, Gramberger M (2013) Combining qualitative and quantitative understanding for exploring cross-sectoral climate change impacts, adaptation and vulnerability in Europe. Reg Environ Chang 13:761–780

Harrison PA, Holman IP, Berry PM (2014) Assessing cross-sectoral climate change impacts, vulnerability and adaptation: an introduction to the CLIMSAVE project. Clim Chang (this issue)

Harvey LDD, Gregory J, Hoffert M, Jain A et al (1997) An introduction to simple climate models used in the IPCC Second Assessment Report. IPCC Tech paper 2, Intergovernmental Panel on Climate Change, Geneva

Hoddle MS (2003) The potential adventive geographic range of glassy-winged sharpshooter, Homalodisca coagulata and the grape pathogen Xylella fastidiosa: implications for California and other grape growing regions of the world. Crop Prot 23:691–699

Holman IP, Allen DM, Cuthbert MO, Goderniaux P (2012) Towards best practice for assessing the impacts of climate change on groundwater. Hydrogeol J 20:1–4

Holman IP, Harrison PA, Metzger MJ (2014) Cross-sectoral impacts of climate and socio-economic change in Scotland: implications for adaptation policy. Reg Environ Chang. doi:10.1007/s10113-014-0679-8

Hulme M, Wigley TML, Barrow EM, Raper SCB, Centella A, Smith S, Chipanshi AC (2000) Using a climate scenario generator for vulnerability and adaptation assessments: MAGICC and SCENGEN Version 2.4 Workbook. Climatic Research Unit, Norwich, UK

Iglesias A, Quiroga S, Schlickenrieder J (2010) Climate change and agricultural adaptation: assessing management uncertainty for four crop types in Spain. Clim Res 44:83–94

IPCC (2007) The Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Available at http://www.ipcc.ch/report/ar4/. Accessed November 2014

IPCC (2013) The Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Available at http://www.ipcc.ch/report/ar5/wg1/. Accessed November 2014

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

Mitchell TD (2003) Pattern scaling: an examination of the accuracy of the technique for describing future climates. Clim Chang 60:217–242

Mitchell TD, Jones PD (2005) An improved method of constructing a database of monthly climate observations and associated high-resolution grids. Int J Climatol 25:693–712

New M, Lister D, Hulme M, Makin I (2002) A high-resolution data set of surface climate over global land areas. Clim Res 21:1–25

Rietveld MR (1978) A new method for estimating the regression coefficients in the formula relating solar radiation to sunshine. Agric Meteorol 19:243–252

Santer BD, Wigley TML, Schlesinger ME, Mitchell JFB (1990) Developing climate scenarios from equilibrium GCM results. Report no. 47, Max Planck Institute für Meteorologie, Hamburg, Germany
Semenov M, Stratonovitch P (2010) Use of multi-model ensembles from global climate models for assessment of climate change impacts. Clim Res 41:1–14
Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93:485–498
Trnka M, Dubrovský M, Žalud Z (2004) Climate change impacts and adaptation strategies in spring barley production in the Czech Republic. Clim Chang 64:227–255
Wells N, Goddard S, Hayes MJ (2004) A self-calibrating Palmer Drought Severity Index. J Clim 17:2335–2351