PERFORMANCE OF CMIP3 AND CMIP5 GLOBAL CLIMATE MODELS OVER THE NORTH-EAST ATLANTIC REGION

Jorge Perez, Melisa Menendez, Fernando J. Mendez, Inigo J. Losada

Environmental Hydraulics Institute "IH Cantabria", Universidad de Cantabria.
Parque Científico y Tecnológico de Cantabria, 39011, Santander, Spain

Phone: +34-942-201616
Fax: +34-942-266361
e-mail: menendezm@unican.es

ABSTRACT

One of the main sources of uncertainty in estimating regional projections affected by global warming is the choice of the global climate model (GCM). The aim of this study is to evaluate the skill of GCMs from CMIP3 and CMIP5 databases in the north-east Atlantic Ocean region. It is well known that the seasonal and interannual variability of surface inland variables (e.g. precipitation and snow) and ocean variables (e.g. wave height and storm surge) are linked to the atmospheric circulation patterns. Thus, an automatic synoptic classification, based on weather types, has been used to assess whether GCMs are able to reproduce spatial patterns and climate variability. Three important factors have been analyzed: the skill of GCMs to reproduce the synoptic situations, the skill of GCMs to reproduce the historical inter-annual time-scale variability and the consistency of GCMs experiments during twenty-first century projections. The results of this analysis indicate that the most skilled GCMs in the study region are UKMO-HadGEM2, ECHAM5/MPI-OM and MIROC3.2(hires) for CMIP3 scenarios and ACCESS1.0, EC-EARTH, HadGEM2-CC, HadGEM2-ES and CMCC-CM for CMIP5 scenarios. These models are therefore recommended for the estimation of future regional multi-model projections of surface variables in the north-east Atlantic Ocean region.

Keywords: Downscaling, General circulation models, Projections, Skill, Weather types.

1 INTRODUCTION

Changes in the Earth's climate throughout the twenty-first century and their potential impacts have become a global concern during the last years. In this context, the World Meteorological Organization (WMO) and the United Nations Environment Programme (UNEP) established the Intergovernmental Panel on Climate Change (IPCC) in 1988. The IPCC has produced a series of reports which show abundant evidence of changes in the global climate system during the twenty-first century. Moreover, most of these changes are larger than those observed during the twentieth century (AR4, IPCC 2007).

The output of global climate models (GCMs) has been one of the most important sources of information since the first IPCC assessment in 1990. The outcomes from GCMs are extensively used in many studies to understand changes in climate dynamics and determine the affects of climate change on a range of impacts. Furthermore, GCMs are used as the basis for many dynamical and statistical downscaling experiments, providing refined information on variables that GCMs do not simulate directly, such as waves or storm surge (e.g. Marcos et al. 2011) or do not simulate at enough resolution (e.g. snow or precipitation). One of the main challenges associated with using GCMs, is model structural uncertainty. Notwithstanding the uncertainty of the forcings for the climate change scenarios, the skill of different GCMs is determined by the different methods used to solve the equations that describe atmospheric and oceanic dynamics. A systematic

evaluation of the performance of the models is, therefore, required to provide greater confidence in the use of GCMs.

One of the first opportunities for climate scientists to compare the skill of a large group of GCMs was phase 3 of the Coupled Model Intercomparison Project (CMIP3) (Meehl et al. 2007). The archived data, officially known as WCRP-CMIP3 multi-model dataset, has been widely studied. For example, analysis of temperature simulations in Australia based on probability density functions (Perkins et al. 2007; Maxino et al. 2008) or studies of precipitation over the Iberian Peninsula (Nieto and Rodriguez-Puebla 2006; Errasti et al. 2010). In these studies, different statistical measures (e.g. RMSE, KS-test, BIAS, correlation indices) are used for objective spatial and quantitative comparison. There are even some studies that aggregate several statistical measures to form a single metric (e.g. Gleckler et al. 2008). Similar studies based on later coordinated multi-model experiments have helped to the process of ongoing improvement of the models. For example, the analysis of the two generations of models used in ENSEMBLES project (van der Linden and Mitchell 2009) conducted by Brands et al. (2011). Recently, the efforts to reduce model uncertainty have led to a new generation of global climate models called Earth System Models as they incorporate the capability to explicitly represent biogeochemical processes that interact with the physical climate (Flato 2011). These models are the basis of the fifth phase of the Coupled Model Intercomparison Project (CMIP5, Taylor et al. 2012) constituting the most current set of coordinated climate model experiments. Several authors have analyzed subsets of CMIP5 models obtaining different rankings of models; e.g. Yin et al. (2012) studied the precipitation over South America, Brands et al. (2013) analyzed several variables in Europe and Africa and Su et al. (2012) studied precipitation and temperature over the Tibetan Plateau.

The main aim of this study is to define a methodology for evaluating the quality of GCMs in a region. The method can therefore assist GCM users in the choice of the most appropriate model to study changes in climate dynamics, to evaluate impacts or to downscale surface variables. A common procedure to evaluate the ability of GCMs is to compare outputs of model simulations against historical reconstructions (reanalysis) or observations. This can be achieved by analyzing differences between mean climatologies or even the whole probability density functions. Recent works have evaluated the skill of GCMs to reproduce synoptic climatology (e.g. Lorenzo et al. 2011; Belleflamme et al. 2012) by using classification methods. The circulation classification method has demonstrated to be a useful and computational efficient tool for the validation of GCMs (Huth 2000). The study of synoptic climatology from circulation patterns or weather types takes into account the natural climate variability and allows the evaluation of spatial relations between different locations.

In this work, we characterize the synoptic patterns from sea level pressure (SLP) fields. SLP provides information of surface climate conditions and it has been found to be a better predictor for downscaling purposes than other variables (e.g. von Storch et al. 1993; Busuioc et al. 2001; Frias et al. 2006).

Taking this into account, we have evaluated the performance of a range of GCMs within the north-east Atlantic Ocean region. The methodology, based on weather types and statistical metrics, analyzes not only the skill of the GCMs to reproduce mean climatologies but also the interannual

variability. Moreover, the consistency of future simulations is also evaluated. This method has been applied to 68 models from CMIP3 to CMIP5, providing useful information about the quality of the GCMs over the European region.

The rest of the paper is organized as follows. In section 2, the data from the model reanalysis databases used for comparison and the analyzed GCMs are presented. Section 3 explains the methodology that has been developed, describing the analyzed region, the weather type classification approach and the statistical analysis of the performance of the GCMs. The study is completed with the presentation of the results in section 4, and the conclusions in section 5.

2 DATA

2.1. Atmospheric reanalysis data

The evaluation of the performance of the GCMs requires the comparison against historical observations. Atmospheric reanalyses are long historical climate reconstructions that can be considered to be quasi-real data as they integrate multiple instrumental measurements and have been widely validated against independent observations. Nowadays, there are several global atmospheric reanalysis databases. In this work, we use 6-hourly SLP data obtained from the three global reanalysis covering the most extensive period of the 20th century: NCEP/NCAR Reanalysis I (NNR, Kalnay et al. 1996), ECMWF 40 Year Reanalysis (ERA-40, Uppala et al. 2005) and NOAA-CIRES 20th Century Reanalysis V2 (20CR, Compo et al. 2011).

NNR (1948-present), created by the National Centers for Environmental Prediction (NCEP) and National Center for Atmospheric Research (NCAR) has been widely used by the scientific community. This global reanalysis is generated by numerical simulation using models similar to those used for weather forecasting, and includes a data assimilation process. ERA-40 (1957-2002) was created by the European Centre for Medium-Range Weather Forecasts (ECMWF), with one version of the Integrated Forecasting System (IFS). 20CR (1871-2010) has been created by the NOAA ESRL/PSD (National Oceanic and Atmospheric Administration Earth System Research Laboratory/Physical Sciences Division). In this reanalysis, pressure observations have been combined with a short-term forecast ensemble of an NCEP numerical weather prediction model. In this study, NNR has been selected to characterize the synoptic patterns of atmospheric circulation because it has been widely validated by the scientific community, covers a large historical period and is an up to date database, nevertheless, ERA-40 and 20CR reanalyses have also been compared with the GCMs.

2.2. Global Climate Models

In this study, the available information on daily sea level pressure from 68 GCMs has been catalogued and subsequently stored. These models have been divided into two groups depending on which generation of scenarios have been simulated. One group includes 26 models from CMIP3 and ENSEMBLES projects and the other one includes 42 CMIP5 models. Tables 1 and 2 show the names of the models that have been used as well as the research centers and countries

that they belong to, the atmospheric resolution and the number of future simulations analyzed (runs). Data from 1961 to 1990 (reference period) have been used to characterize recent past conditions and projections from 2010 to 2100 have been taken to represent future conditions, as they are time periods available from most models.

The simulations analyzed in the CMIP3 and ENSEMBLES models are called 20C3M (Twentieth Century Climate in Coupled Models) for recent past conditions and SRES B1, SRES A1B and SRES A2 (Special Report on Emission Scenarios, Nakicenovic et al. 2000) for future scenarios. The three selected scenarios are generally taken to represent low, medium and high CO2 concentrations, respectively. A total of 44 20C3M simulations, 43 of A1B, 19 of A2 and 26 of B1 are studied. Eighteen models belong to CMIP3 and eight models (CNRM-CM33, ECHAM5C/MPI-OM, EGMAM, EGMAM2, IPSL-CM4v2, UKMO-HadCM3C and UKMO-HadGEM2) belong to the ENSEMBLES project. Data are obtained from the results of the models sent to the Program for Climate Model Diagnosis and Intercomparison (PCMDI) at the Lawrence Livermore National Laboratory in the USA (http://www-pcmdi.llnl.gov/ipcc/about_ipcc.php) and from the CERA database of the World Data Center for Climate (WDCC) in Hamburg (http://cera-www.dkrz.de/CERA/).

For the 42 CMIP5 models, the experiments analyzed are called historical for recent past conditions and RCP2.6, RCP4.5, RCP6.0 and RCP8.5 (Representative Concentration Pathways, Moss et al. 2010) for the future. The four selected RCPs included one mitigation scenario leading to a very low forcing level (RCP2.6), two medium stabilization scenarios (RCP4.5/RCP6.0) and one very high baseline emission scenario (RCP8.5) leading to high greenhouse concentration levels (van Vuuren 2011). This makes a total of 136 historical simulations, 48 of RCP2.6, 83 of RCP4.5, 31 of RCP6.0 and 63 of RCP8.5. CMIP5 data are available through the Earth System Grid - Center for Enabling Technologies (ESG-CET), on the page (http://pcmdi9.llnl.gov/).

3 METHODS

The methodology developed to study the skill of the GCMs is summarized in a diagram in Figure 1. Data from reanalysis and GCMs are collected first. The study area is then defined and SLP fields are preprocessed to the spatial domain in the selected region (chart upper level). In order to get the estimated indicators of the performance of the GCMs, a weather type (WT) classification from the reanalysis data is carried out. The occurrence rate of each synoptic situation group is assessed from both the reanalysis data and the GCMs for several time periods (chart middle level). Finally, different statistical indices are computed to compare the occurrence rates (chart bottom level). The comparison between the observed and simulated historical WT frequency indicates the skill of the GCMs to simulate the recent past climate. The results of this comparison are used to analyze the similarity of the synoptic situations and the ability of the GCMs to reproduce the interannual variability. On the other hand, the comparison between the historical and future WT frequency from GCMs determine the simulated rates of change. These rates of change are used to analyze the consistency of future projections.

3.1. Study area

The domain of interest in this work is the North Atlantic. This region is dominated by the North Atlantic Oscillation (NAO), which is one of the most prominent climate fluctuation patterns in the Northern Hemisphere (Hurrell et al. 2003). NAO is usually described with an index based on the pressure difference between Iceland and the Azores and it has important influence on climate from the United States to Siberia, and from the subtropical Atlantic to the Arctic. We have therefore selected an area in the north-east Atlantic from 25ºN to 65º N and from 52.5º W to 15º E. In this region, many surface variables are highly correlated with pressure fields, such as wind waves (Izaguirre et al. 2012), precipitation (Rodriguez-Puebla and Nieto 2010), snow (Seager et al. 2010) and cereal production (Rodriguez-Puebla et al. 2007). Given the fact that data from GCMs are provided in different spatial resolution grids, in order to make a coherent comparison, all SLP data have been interpolated by means of bilinear interpolation to a grid of 2.5º latitude by 2.5º longitude, identical to the mesh of the NNR results. The analyzed spatial domain and resolution is shown in Figure 2.

3.2. Classification of weather types

Non-initialized simulations by GCMs aim to simulate long-term statistics of observed weather rather than day-to-day chronology. For this reason, mean climatologies from GCMs are usually compared against reanalysis to evaluate the ability of the GCMs. However, mean climatology comparison ignores the climate variability of the atmospheric circulation, which causes a wide variety of meteorological situations, even severe storm conditions. The evaluation of GCMs throughout a classification of weather types reduces this problem, since classification aims to group similar meteorological situations minimizing the variability within each group. Therefore, each group is more or less homogeneous and distinct from other groups. Many authors are aware of the importance of the models to reproduce climate variability over a region and have used atmospheric circulation type classifications; e.g. (Belleflamme et al. 2012; Lee and Sheridan 2011; Pastor and Casado 2012). Here, the circulation type classification is developed by applying the non-hierarchical clustering technique K-means (McQueen 1967) over the SLP fields in the study region. To do this, 3-daily averaged SLP fields, *SLP(x,t)*, from the NNR are analyzed. The three days time scale is chosen to be able to capture mid- latitude cyclogenesis situations.

First, we process each 3-day averaged SLP field anomaly, $SLPA(x,t) = SLP(x,t) - \overline{SLP(t)}$, where t represents each 3-days interval and $\overline{SLP(t)}$ is the mean SLP in the 3-days interval in the spatial domain. So, two situations with similar patterns but slightly different mean SLP can be grouped together. Then, we apply principal components analysis (PCA) to the processed 3-daily SLP fields of NNR from 1950 to 1999. PCA helps the clustering technique reduce dimensions whilst conserving the maximum data variance. That is, the covariance of the SLP anomalies in the study region is used to obtain uncorrelated principal components. In this case, eleven components have explained more than 95% of variance. In order to get a set of synoptic climatologies (weather types), the K-means algorithm has been applied over these modes. The K-means technique divides

the data space into N classes, which are represented by their centroids. Each class represents a group of atmospheric states of similar characteristics. We force the K-means algorithm to start with dissimilarity-based compound selection (Snarey et al. 1997) and the number of classes has been set to N=100. The selection of a hundred classes is made based on the compromise between the best possible characterization of synoptic climatologies, represented by the largest number of clusters and including an average number of 40 data per group. A proximity criterion is applied over the N=100 obtained WTs, and the centroids are visualized in a 10x10 lattice (Figure 3). The proximity criterion is based on minimizing the sum of Euclidean distances between each centroid and its neighbors. This organization helps to interpret results since weather types of similar characteristics appear near to one another. For example, the dominant winter pattern is characterized by a low pressure center over the Azores Islands, while a high pressure center dominates the summer synoptic situation. The weather types located in the right side of figure 3 are characterized by low pressures in Iceland and high pressures in the Azores Islands, which is usually associated with a positive phase of NAO.

3.3 Evaluation of the performance of GCMs

3.3.1 Similarity of synoptic situations

Here, the climate information obtained from the synoptic classification of NNR has been used to evaluate the skill of GCMs. First, the relative frequency of each of the one hundred weather types has been calculated for NNR, as the reference pattern (Figure 4). The relative frequencies are estimated from the number of 3-day atmospheric states that can be attributed to each WT, characterized by its centroid, during the reference period of 30 years (from 1st January 1961 to the 31st of December 1990). The Euclidean distance in the reduced EOF-space has been used to assess which centroid is the closest. Then, the same methodology has been applied to compute the relative frequencies from ERA40, 20CR and GCMs.

Objective indexes to measure the differences between frequencies of the reference pattern and those for the GCMs during the same period in the historical/20C3M simulations have been applied. The scatter index and a metric based on the relative entropy have been used for this purpose. The scatter index (*SI*) is the root mean square error normalized by the mean frequency:

$$SI = \sqrt{\frac{\sum_{i=1}^{N}(p_i - p_i^{'})^2}{N}} \Bigg/ \frac{\sum_{i=1}^{N}(p_i)}{N} \qquad (1)$$

being p_i the relative frequency of the i^{th} weather type from the reanalysis for the reference period, p'_i the relative frequency of the i^{th} weather type from a GCM simulation for the reference period and N the number of weather types. This index has been used to compare the relative frequencies of each simulation of each GCM with the ones of the reanalysis during the reference period. The metric based on the relative entropy (*RE*) is defined here as:

$$RE = \sum_{i=1}^{N} p_i \left| \log \frac{p_i}{p_i^{'}} \right|, \quad (2)$$

Lower values of *SI* and *RE* therefore indicate a high degree of similarity and hence a better performing GCM. This index has been used to analyze the skill of the different GCMs to simulate weather types of low probability of occurrence. The analysis of these situations, which could be associated to extreme events, requires a relative index, such as *RE* since the scatter index analysis gives more importance to commonly occurring situations.

This analysis has been done both for annual time-scale as well as seasonal time-scales, considering the following distribution: winter (December, January and February), spring (March, April and May), summer (June, July and August) and fall (September, October and November). An example of the application of these indexes is shown in figure 4. The reference pattern represents the relative frequency of each characterized synoptic situation (weather type) for the recent past conditions. NNR has been used to derive this pattern although ERA-40 and 20CR show similar characteristics.

The frequencies obtained from ECHAM5 (CMIP3) and ACCESS1.0 (CMIP5), provide low *SI* and *RE* since the most common and unusual situations are well reproduced. These models show only small variations between occurrence of neighboring weather types which represent near synoptic situations and probability of occurrence. Alternatively, CNRM-CM3 (CMIP3) and FGOALS-g2 (CMIP5) show less similarity with the reanalysis reference pattern and consequently larger *SI* and *RE*. These models tend to overestimate the frequency of particular WT´s associated to synoptic situations with weaker gradients between low and high pressure centers. Note that here and henceforth, *SI* and *RE* are interpreted in relative values (i.e. lowest values versus highest values across the ensemble).

3.3.2 Interannual variability

The skill of a model to represent the climate state is the most important test to evaluate its quality. It is for this reason that mean climatologies over several decades are often used to compare GCMs with observations. It is however, important to note that the variance (i.e. interannual variability) is also a requirement for good model performance. We have analyzed the skill of GCMs to represent interannual climate variability because it is an indicator of their ability to respond to changing conditions. The magnitude of the interannual variability has been measured for each WT by assessing the standard deviation of the 30 annual values of relative frequency over the reference period (1961-1990). The comparison of the variability values of the reanalysis with those that correspond to each GCM is conducted by the scatter index of the standard deviations of the N weather types (*stdSI*).

$$stdSI = \sqrt{\frac{\sum_{i=1}^{N} (std(p_i) - std(p_i^{'}))^2}{N}} \Bigg/ \frac{\sum_{i=1}^{N} (std(p_i))}{N} \quad (3)$$

The lower the *stdSI* the better the performance of the GCM to simulate the interannual climate variability.

3.3.3 Consistency of future projections

We have evaluated the skill of GCMs to reproduce historical climate and its variability. However, good model performance evaluated from the present climate does not necessarily guarantee reliable predictions of future climate (Reichler and Kim 2008). This is mainly due to projections consider future greenhouse gas forcings outside the used range in the historical period of validation. Consequently, the skill of GCMs to reproduce future climate projections cannot be directly evaluated. However, multi-model ensembles are often used to analyze future projections. In order to provide information about uncertainty on the ensembles, we have evaluated the consistency between GCMs during future projections.

To assess the consistency between future projections of GCMs, we have divided the twenty-first century in three different periods: short term (2010-2039), mid-term (2040-2069) and long-term (2070-2099), while evaluating which models predict inconsistent variations in each of these periods, i.e. magnitudes of change much larger or much lower than those of most models. We assume the stationary hypothesis over climate dynamics, that is, the WT classification remains valid throughout the twenty-first century. For every analyzed simulation and future time period, we have calculated two metrics of the magnitude of change towards the reference period. The magnitude of change in the frequency of synoptic situations has been evaluated through SI and the magnitude of change in the interannual variability has been analyzed through *stdSI*. The mean magnitude of change has been used in case of several simulations of the same model. For each scenario, future period and metric, we have computed the quartiles of the magnitudes of change. The interquartile range (IQR) is the difference between the upper quartile (Q3, 75 percentile) and the lower quartile (Q1, 25 percentile). IQR is a robust statistic to measure the dispersion of a set of data. In this study, models with magnitudes of change lower than Q1 - 1.5(IQR) or higher than Q3 + 1.5(IQR) are considered outliers, i.e. GCMs of a very different behavior compared with the rest of GCMs.

4 RESULTS

4.1 Skill of GCMs to perform climatologies

The ability of the GCMs to represent the relative frequency of synoptic situations in the reference period can be assessed by direct comparison with the reference pattern. Figure 5 summarizes the bias of the GCMs for the 20C3M simulations (CMIP3 and ENSEMBLES) and the historical simulations (CMIP5). Dots in the WTs indicate agreement on the sign of the bias for more than 80% of the models. Small bias has been estimated on GCMs over all WTs, indicating a good ability of the models to reproduce common synoptic situations, i.e. mean climatologies. CMIP5 simulations show a general better agreement than CMIP3. Some discrepancies, however, are found on unusual events associated to deep low pressures centered over different areas of the North Atlantic (right hand side of the figure) and relatively stable atmospheric states (WTs at the bottom of the figure). The former are over-estimated, whilst the latter tend to be slightly underestimated. Note that the overestimated WTs might be associated to extreme storm events during intense

Northern Annular Mode (NAM). This overestimation is in agreement with previous studies. For instance, Gerber et al. (2008) found that climate models vaguely capture the NAM variability, over-estimating persistence on sub-seasonal and seasonal timescales.

The performance of individual GCMs has been measured using the *SI* and *RE* indices. The results are summarized in Figure 6 for 20C3M simulations and in Figure 7 for historical simulations. In both figures the models have been sorted according to their *SI* and the number of simulations analyzed for each model is shown between brackets. The *SI* score of the models with only one simulation is represented by the small vertical black lines. When several simulations are available these vertical black lines represent the mean value of the *SI* while the horizontal ones represent the range between the minimum and the maximum *SI*. The mean *RE* is represented by a black dot. The *SI* and *RE* scores have also been obtained for the reanalyses ERA-40 (*SI=0.16*, *RE=0.10*) and 20CR (*SI=0.26*, *RE=0.14*) during the reference period. 20CR has also been analyzed in 1901-1930 (*SI=0.30*, *RE=0.18*) and in 1931-1960 (*SI=0.30*, *RE=0.19*). The similar scores for different periods of the twentieth century support the use of the same synoptic classification in the twenty-first century. These values provide an indicator of SI and RE values which better represent the performance of GCMs. The *SI* scores of the reanalyses have been represented in the figures by vertical dotted lines. It can be observed that ERA-40 is very similar to NNR whereas 20CR present larger differences. This was expected since 20CR only assimilates surface pressure data.

The models that best reproduce the occurrence rate of synoptic climatology for 20C3M simulations with *SI* lower than 0.5 and *RE* lower than 0.3, are: UKMO-HadGEM2 (*SI=0.37*, *RE=0.22*), ECHAM5/MPI-OM (*SI=0.46*, *RE=0.26*) and MIROC32HIRES (*SI=0.49*, *RE=0.28*). Alternatively, the five models which have *SI* larger than 1 and, therefore, have a lower simulation performance with regard to the frequency of the different synoptic situations, are: CCSM3, GISS-ER, FGOALS-g1.0, CNRM-CM3 and CNRM-CM33. For CMIP5 models, there are nine models with *SI* lower than 0.5. Three of them: ACCESS1.0 (*SI=0.33*, *RE=0.19*), EC-EARTH (*SI=0.36*, *RE=0.21*) and HadGEM2-CC (*SI=0.37*, *RE=0.21*) have both *SI* and *RE* lower than the best model for 20C3M simulations. The other six: HadGEM2-ES, MPI-ESM-P, CMCC-CM, GFDL-CM3, MPI-ESM-LR and CMCC-CMS have *SI* slightly larger but *RE* is still lower than 0.3. Note that, only two CMIP5 models: IPSL-CM5B-LR (*SI*=1.03, *RE*=0.57) and FGOALS-g2 (*SI*=1.17, *RE*=0.60) show *SI* larger than one.

The differences between runs of a single model are one order of magnitude lower than the differences between models. This shows that the internal variability is well taken into account by using a 30-year period. Moreover, results are qualitatively similar for the two indicators (*RE* and *SI*) that have been used to analyze the representation of the synoptic situations, indicating that the model performance is consistent across the two performance measures. Both indexes reveal an improvement in CMIP5 models with respect to the analyzed set of models from CMIP3 and ENSEMBLES. In addition, the values of *RE* are smaller for CMIP5 models than for CMIP3 models with similar values of *SI*, indicating that CMIP5 models have improved their capacity to detect synoptic situations with low relative frequency.

4.2 Skill of GCMs to perform climate variability

The results of the diagnosis in each season are shown in Figures 8 and 9, with the models and simulations analyzed as in Figures 6 and 7, respectively. The *SI* scores for ERA40 and 20CR are very similar in fall (0.34 vs. 0.35, respectively) and winter (0.34 vs. 0.39), being the differences slightly larger in spring (0.30 vs. 0.40). The largest differences can be found in summer (0.31 vs. 0.59). The *RE* scores cannot be included because several WTs have zero occurrences in some seasons.

For CMIP3 and ENSEMBLES models (Figure 8) the diagnosis in spring and fall is analogous to the annual one except for minor differences. In both seasons, most models show very similar performance with *SI* between 0.5 and 1. Only three models in spring and seven models in fall show noticeably larger *SI*. On the contrary, in winter and summer the differences are larger. In winter some ENSEMBLES models: EGMAM (*SI=0.76*), EGMAM2 (*SI=0.71*) and UKMO-HADCM3C (*SI=0.80*) perform as well as the best models. FGOALS-g1.0 shows results of lower quality (*SI=3.70*) in summer and hence performs poorly on the annual scale. On the other hand, CCSM3 and PCM only show low *SI* in summer, and perform with lower quality in the rest of the seasons. A similar observation occurs with models from Commonwealth Scientific and Industrial Research Organisation (CSIRO), with the *SI* of CSIROmk35 and CSIROmk30, the first and third lowest on this season. For CMIP5 models (Figure 9) the seasons that show larger discrepancies with respect to the global evaluation shown in Figure 7 are also winter and summer, with the diagnosis in spring and fall similar to the global evaluation. Interestingly, the CMIP5 models which provide the worst diagnostic in winter (*SI* larger than 1.4), namely CCSM4, CESM1(BGC), CESM1(FASTCHEM), BNU-ESM and BCC-CSM1.1(m) are some of the best models in summer. Note that the *SI* in summer of CCSM4 is 0.62, only slightly larger than the one of 20CR. On the contrary IPSL-CM5B-LR and FGOALS-g2 are the poorest performing models at the annual scale and during summer season but they perform well in winter. Curiously, the model with the third largest *SI* in summer INM-CM4 is one of the best models in the other seasons. The seasonal analysis show that the performance of the models depends on the season, especially in summer and winter, indicating that, in some cases, the most adequate models depend on the purposes.

The interannual variability analysis has been based on the *stdSI* score described in section 3.2. As shown in figures 10 and 11, in which the order of GCMs of the previous figures has been kept, the *stdSI* scores for ERA-40 (*stdSI=0.17*) and 20CR (*stdSI=0.21*) are more similar than their *SI*. The results for 20C3M simulations (Figure 10), show that UKMO-HadGEM2 (*stdSI=0.24*) and ECHAM5/MPI-OM (*stdSI=0.27*) provide the highest quality results, with *stdSI* lower than 0.3, while CNCM33 and GISS-ER are the ones that provide results of lower quality with *stdSI* larger than 0.6. For the historical simulations of the CMIP5 models (Figure 11) the values of *stdSI* are slightly better than the ones for 20C3M simulations. Five models ACCESS1.0, MPI-ESM-P, EC-EARTH, HadGEM2-CC and HadGEM2-ES have *stdSI* lower than 0.3. Furthermore, there are no models with *stdSI* larger than 0.6 and only two models: IPSL-CM5B-LR and FGOALS-g2 exceed 0.5. Results obtained for interannual variability confirm those obtained from the similarity of synoptic situations, with the models with the highest and lowest performance the same for both analyses.

4.3 Consistency of future projections

Analysis of future projections is made in a different way to the analysis of past climate. Historical simulations can be compared with reanalysis data, but the future projections can only be compared to each other. The analysis of future projections can be used to detect models with anomalous behavior but not to determine which models are best. The results of the consistency of future projections have been synthesized in Figure 12 for the three SRES scenarios considered (B1, A1B and A2) and figure 13 for the four RCP (RCP2.6, RCP4.5, RCP6.0 and RCP8.5). For each scenario, the magnitudes of change of the frequency of the synoptic situations and the magnitudes of change in the interannual variability are shown for three future time periods. On each box, the central mark is the median, the edges of the box are the lower and upper quartiles and the whiskers extend to the most extreme magnitudes of change within the range defined by *Q1 - 1.5(IQR)* and *Q3 + 1.5(IQR).* The numbered red dots represent models with magnitudes of change outside this range.

For SRES scenarios (Figure 12) only the mid-term and long-term periods are shown because few simulations cover the short term period. For these scenarios, INM-CM3 (19), GISS-ER (23) and CNRM-CM3 (25) show magnitudes of change notably high in particular combinations of scenario, indicator and time-period. For CMIP5 (Figure 13) short-term, mid-term and long-term can be shown because information for the full twenty-first century is available. In this case there are two different groups of models with anomalous magnitudes of change. HadGEM2-AO (03), GFDL-CM3 (08), IPSL-CM5A-MR (28), IPSL-CM5A-LR (35), MIROC-ESM-CHEM (38), FGOALS-s2 (39) and FGOALS-g2 (42), show in several cases high magnitudes of change whereas MPI-ESM-MR (15), INM-CM4 (20), MRI-CGCM3 (31) and BCC-CSM1.1(m) (40) show in some cases low magnitudes of change. Results indicate that the magnitudes of change and their spread are larger in the long-term period than in the short-term period and for high-emissions scenarios, e.g., A2 and RCP8.5, than for low-emission scenarios. It is interesting to note the connection between the ability of models to reproduce the present climate (the higher the number, the worse the performance) and the consistency of their future simulations. The models with anomalous magnitudes of change mostly belong to the group of models with low skill in the reference period. However, some of the models with anomalous magnitudes of change perform reasonably well in the recent past. It may indicate that these models are unable simulate the climate variability associated to larger changes in the forcings during the twenty-first century.

5 CONCLUSIONS

A methodology to analyze the performance of GCMs based on weather types (WTs) and statistical metrics has been developed. The method analyzes the ability of the models to reproduce three characteristics: the historical synoptic climatologies, the interannual variations and the consistency of future projections. The use of statistic metrics based on the scatter index and the relative entropy allow a quantitative estimation of the GCMs performance.

The method has been applied to the Northeast Atlantic region. The three models that best simulate the recent past climate conditions from the CMIP3 and ENSEMBLES datasets are: UKMO-

HadGEM2, ECHAM5/MPI-OM and MIROC3.2 (hires). Furthermore, these models are consistent during the twenty-first century for the SRES simulations analyzed. For CMIP5, seven models perform above the rest during the twentieth century: ACCESS1.0, EC-EARTH, HadGEM2-AO, HadGEM2-CC, HadGEM2-ES, MPI-ESM-P and CMCC-CM. During the twenty-first century five of them are consistent but two of them are not. HadGEM-AO overestimates the changes for RCP45 in the short term and there are no future simulations for MPI-ESM-P.

These results are consistent with other studies of SLP in the Northern Hemisphere. For example, Walsh et al. (2008) evaluated 15 GCMs of CMIP3 over the Northern extratropical domains focusing in Greenland and Alaska. They found that ECHAM5/MPI-OM is one of the top-performing models. Errasti et al. (2011) found ECHAM5/MPI-OM and MIROC3.2 (hires) as the best CMIP3 model in the Iberian peninsula. Brands et al. (2011) found similar results within ENSEMBLES models in the Northeast Atlantic region for the two best models (UKMO-HadGEM2 and ECHAM5/MPI-OM) and they also concluded that the two worst performing models are CNRM-CM3 and CNRM-CM33. Brands et al. (2013) also obtained HadGEM2-ES outperforming the remaining models in a group of seven CMIP5 models.

It is important to highlight that an evaluation of the quality of the GCMs depends on the study area and the considered predictor, showing different results to those obtained for other variables or regions. Note that the performance of the GCMs also varies depending on the analyzed season. Therefore, the choice of the most adequate models depends on the specific purposes (e.g. studies focus on extreme wave heights during winter or ice melting during summer). On the contrary, from the analysis carried out the importance of the atmospheric resolution is not clear. The models with the highest performance are not always performing the best.

The small differences in the skill indexes among runs of the same model indicate that the methodology is robust because it is not considerably affected by the natural variability of climate. In spite of this, notable differences can be observed in future simulations, even among the best rated models. Therefore, the use of ensembles or multi-model groups is recommended since it diminishes the effects of individual simulations allowing us to have greater confidence on the results.

ACKNOWLEDGEMENTS

The work was partly funded by the projects iMar21 (CTM2010-15009) and GRACCIE (CSD2007-00067) from the Spanish government and the FP7 European project CoCoNet (287844). The authors would like to acknowledge the climate modeling groups which have generated the data used in this study, as well as the PCMDI and WDCC for facilitating access to it. We would also like to acknowledge the valuable suggestions made by B. P. Gouldby.

REFERENCES

Belleflamme A, Fettweis X, Lang C, Erpicum M (2012) Current and future atmospheric circulation at 500 hPa over Greenland simulated by the CMIP3 and CMIP5 global models. Clim Dyn. doi: 10.1007/s00382-012-1538-2

Blazquez J, Nuñez MN (2012) Analysis of uncertainties in future climate projections for South America: comparison of WCRP-CMIP3 and WCRP-CMIP5 models. Clim Dyn. doi: 10.1007/s00382-012-1489-7

Brands S, Herrera S, San-Martín D, Gutiérrez JM (2011) Validation of the ENSEMBLES global climate models over southwestern Europe using probability density functions, from a downscaling perspective. Clim Res 48:145–161

Brands S, Herrera S, Fernández J, Gutiérrez JM (2013) How well do CMIP5 Earth System Models simulate present climate conditions in Europe and Africa?. Clim Dyn. doi:10.1007/s00382-013-1742-8

Busuioc A, Chen D, Hellström C (2001) Performance of statistical downscaling models in GCM validation and regional climate change estimates: application for Swedish precipitation. Int J Climatol, 21(5), 557-578

Camus P, Mendez FJ, Medina R, Cofiño AS (2011) Analysis of clustering and selection algorithms for the study of multivariate wave climate. Coast Eng. doi: 10.1016/j.coastaleng.2011.02.003

Compo GP, Whitaker JS, Sardeshmukh PD, Matsui N, Allan RJ, Yin X, Gleason BE et al. (2011) The Twentieth Century Reanalysis Project. Q J Roy Meteor Soc 137(654), 1–28. doi:10.1002/qj.776

Errasti I, Ezcurra A, Sáenz J, Ibarra-Berastegi G (2010) Validation of IPCC AR4 models over the Iberian Peninsula. Theor Appl Climatol 103(1-2), 61-79. doi:10.1007/s00704-010-0282-y

Flato GM (2011) Earth system models: an overview. Wiley Interdiscip Rev Clim Change, 2(6), 783-800

Frias MD, Zorita E, Fernández J, Rodríguez-Puebla C (2006) Testing statistical downscaling methods in simulated climates. Geophys Res Lett, 33(19)

Gerber EP, Polvani LM, Ancukiewicz D (2008) Annular mode time scales in the Intergovernmental Panel on Climate Change Fourth Assessment Report models. Geophys Res Lett, 35(22), L22707. doi:10.1029/2008GL035712

Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. J Geophys Res, 113(D6), D06104

Hurrell JW, Kushnir Y, Ottersen G, Visbeck M (2003) An overview of the North Atlantic oscillation. Geophysical Monograph-American Geophysical Union, 134, 1-36.

Huth R (2000) A circulation classification scheme applicable in GCM studies. Theoret Appl Climatol, 67(1-2), 1-18

IPCC (2007) Climate Change 2007: The Physical Science Basis. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.

Izaguirre C, Menéndez M, Camus P, Méndez FJ, Mínguez R, Losada IJ (2012) Exploring the interannual variability of extreme wave climate in the Northeast Atlantic Ocean, Ocean Model.

Kalnay EM, Kanamitsu R, Kistler W, Collins D, Deaven L, Gandin M, Iredell S, Saha G, White J, Woollen Y, Zhu M, Chelliah W, Ebisuzaki W, Higgins J, Janowiak KC, Mo C, Ropelewski J, Wang A, Leetmaa R, Reynolds R, Jenne R, Joseph D (1996) The NCEP/NCAR 40-year reanalysis project. Bull Am Meteorol Soc 77, 437–470.

Lee CC, Sheridan SC (2012) A six-step approach to developing future synoptic classifications based on GCM output. Int J Climatol, 32(12), 1792-1802

MacQueen J (1967, June). Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 281-297, p. 14)

Lorenzo MN, Ramos AM, Taboada JJ, Gimeno L (2011) Changes in present and future circulation types frequency in northwest Iberian Peninsula. PLoS One, 6(1), e16201.

Marcos M, Jordà J, Gomis D, Pérez B (2011) Changes in storm surges in southern Europe from a regional model under climate change scenarios. Global Planet Change, Volume 77, Issues 3–4, Pages 116-128

Maxino CC, McAvaney BJ, Pitman AJ, Perkins SE (2008) Ranking the AR4 climate models over the Murray Darling Basin using simulated maximum temperature, minimum temperature and precipitation. Int J Climatol 28:1097–1112.

Meehl GA, Covey C, Taylor KE, Delworth T, Stouffer RJ, Latif M, McAvaney B et al. (2007) The WCRP CMIP3 Multimodel Dataset: A New Era in Climate Change Research. Bull Am Meteorol Soc, 88(9), 1383–1394. doi:10.1175/BAMS-88-9-1383

Nakicenovic N et al. (2000): IPCC Special Report on Emissions Scenarios. Cambridge University Press, 599 pp.

Nieto S, Rodríguez-Puebla C (2006) Comparison of precipitation from observed data and general circulation models over the Iberian Peninsula. J Clim, (1992), 4254–4275.

Pastor MA, Casado MJ (2012) Use of circulation types classifications to evaluate AR4 climate models over the Euro-Atlantic region. Clim Dyn, 39(7-8), 2059-2077

Perkins SE, Pitman AJ, Holbrook NJ, McAneney J (2007) Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature and precipitation over Australia using probability density functions. J Clim, 20, 4356–4376

Reichler T, Kim J (2008) How well do coupled models simulate today's climate. Bull Am Meteorol Soc, 89(3), 303.

Rodríguez-Puebla C, Ayuso SM, Frias MD, Garcia-Casado LA (2007) Effects of climate variation on winter cereal production in Spain. Climate Res, 34(3), 223

Rodríguez-Puebla C, Nieto S (2010) Trends of precipitation over the Iberian Peninsula and the North Atlantic Oscillation under climate change conditions. Int J Climatol, 30(12), 1807-1815

Seager R, Kushnir Y, Nakamura J, Ting M, Naik N (2010) Northern Hemisphere winter snow anomalies: ENSO, NAO and the winter of 2009/10. Geophys Res Lett, 37(14).

Snarey M, Terrett NK, Willett P, Wilton DJ (1997) Comparison of algorithms for dissimilarity-based compound selection. J Mol Graphics Modell, 15(6), 372-385

Su F, Duan X, Chen D, Hao Z, Cuo L (2012) Evaluation of the Global Climate Models in the CMIP5 over the Tibetan Plateau. J Clim.

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc, 93(4), 485-498

Uppala SM, KÅllberg PW, Simmons AJ, Andrae U, Bechtold VDC, Fiorino M, Gibson JK, et al. (2005) The ERA-40 re-analysis. Q J Roy Meteor Soc, 131(612), 2961–3012. doi:10.1256/qj.04.176

van der Linden P, and JFB Mitchell (eds) (2009). ENSEMBLES: Climate Change and its Impacts: Summary of research and results from the ENSEMBLES project. Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK. 160pp.

van Vuuren DP, Edmonds J, Kainuma MLT, Riahi K, Thomson A, Matsui T, Hurtt G, Lamarque J-F, Meinshausen M, Smith S, Grainer C, Rose S, Hibbard KA, Nakicenovic N, Krey V, Kram T (2011) Representative concentration pathways: An overview. Climatic Change, 1–27. 10.1007/s10584-011-0148-z

von Storch H, Zorita E, Cubasch U (1993) Downscaling of global climate change estimates to regional scales: an application to Iberian rainfall in wintertime. J Climate, 6(6), 1161-1171

Yin L, Fu R, Shevliakova E, Dickinson RE (2012) How well can CMIP5 simulate precipitation and its controlling processes over tropical South America?. Clim Dyn. doi:10.1007/s00382-012-1582-y

Figure

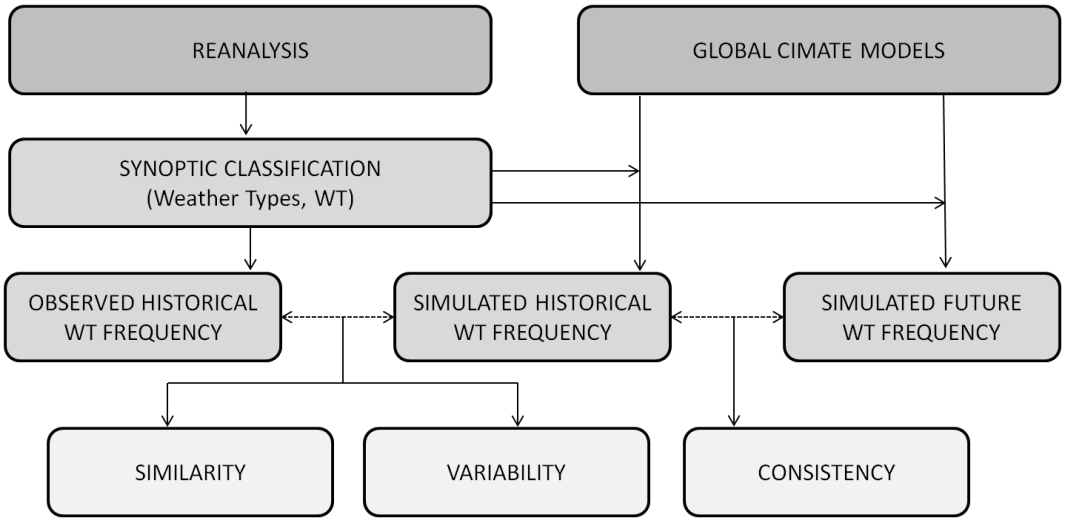

Fig. 1 Flowchart representing the methodology.

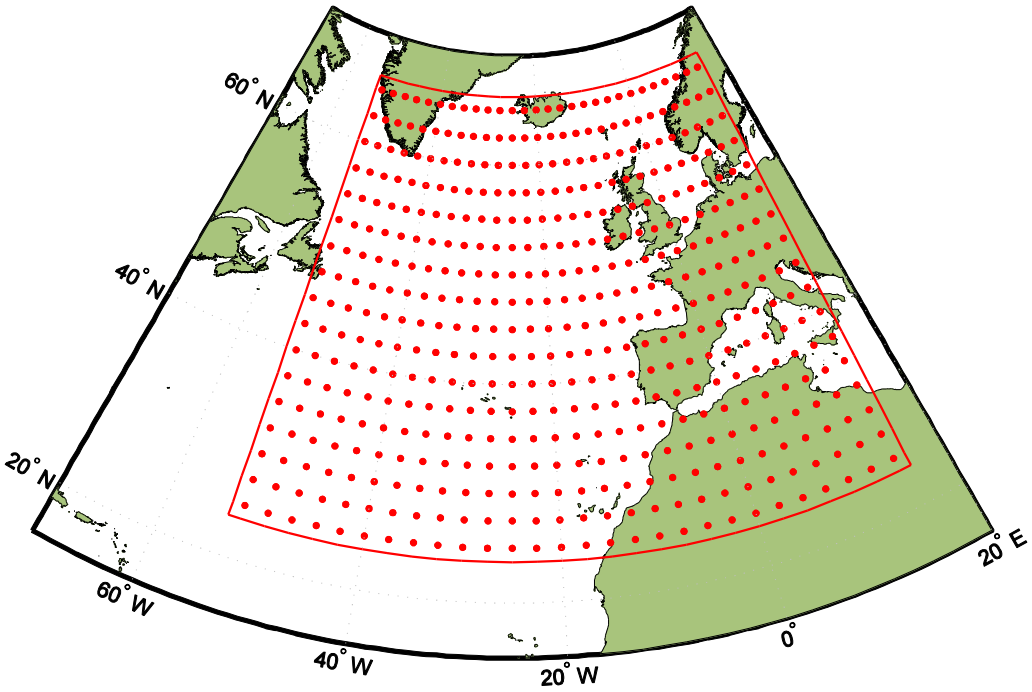

Fig. 2 Spatial domain of the study area

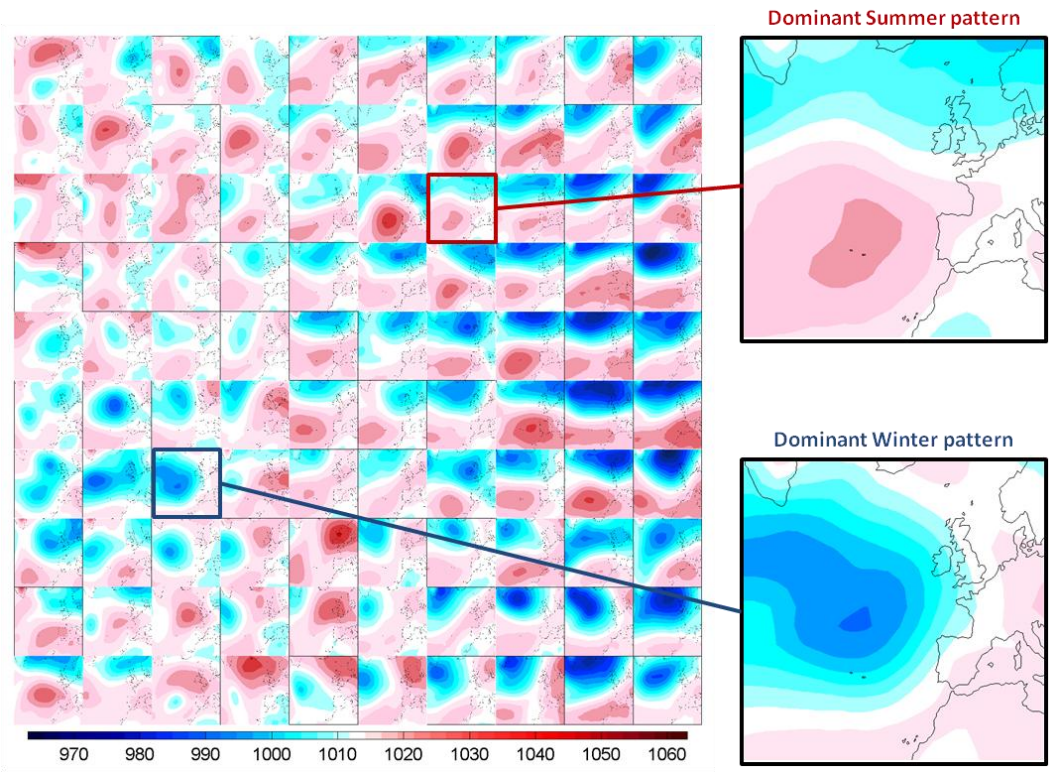

Fig. 3 The 100 weather types represented by the SLP fields (mbar). Right panels show the most frequently occurring weather types in winter and summer.

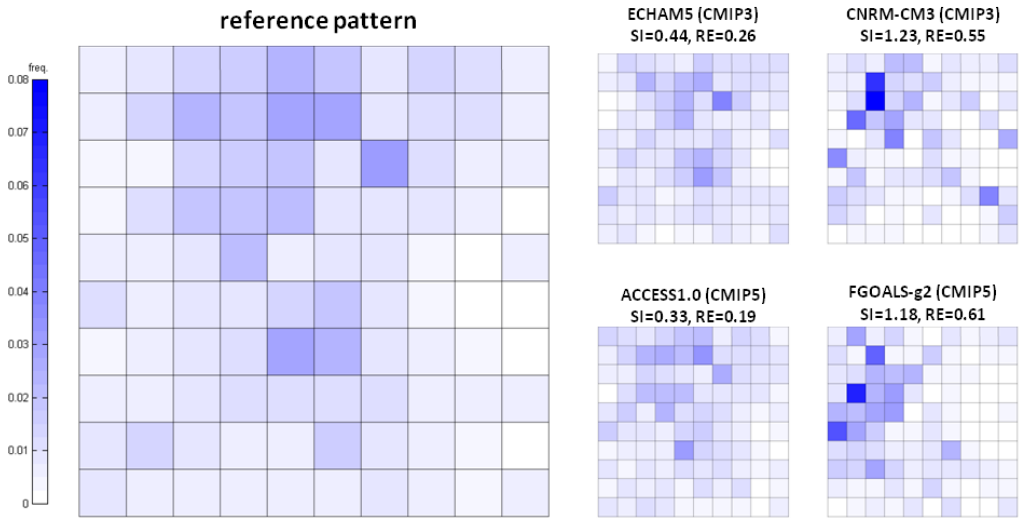

Fig. 4 Relative frequency of the 100 weather types in the reference period for NCEP-NCAR reanalysis (quasi-observations) and four GCMs. The darker blue colors being weather types with high frequency and the lighter blue the less frequent weather types.

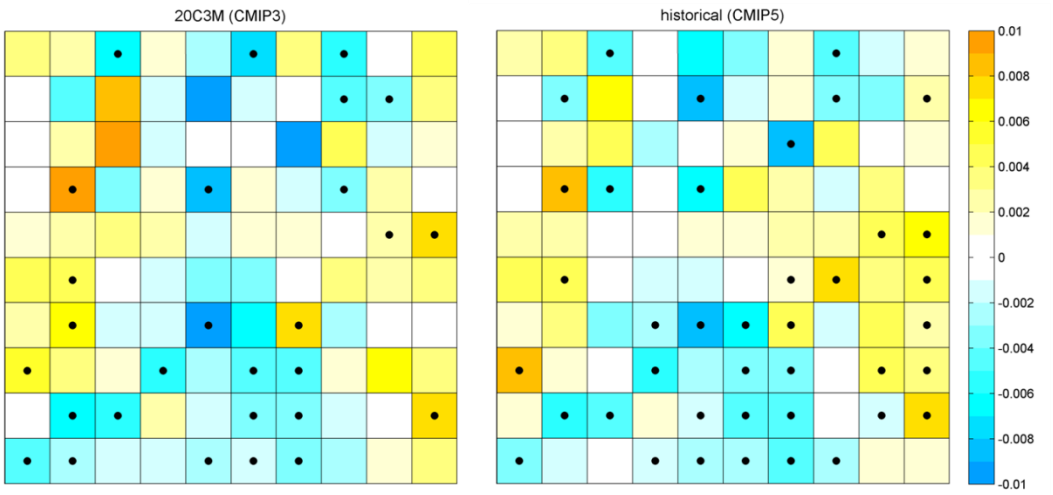

Fig. 5 Bias of 20C3M (left) and historical (right) ensembles. The small dots indicate agreement on the sign for more than 80% of the models.

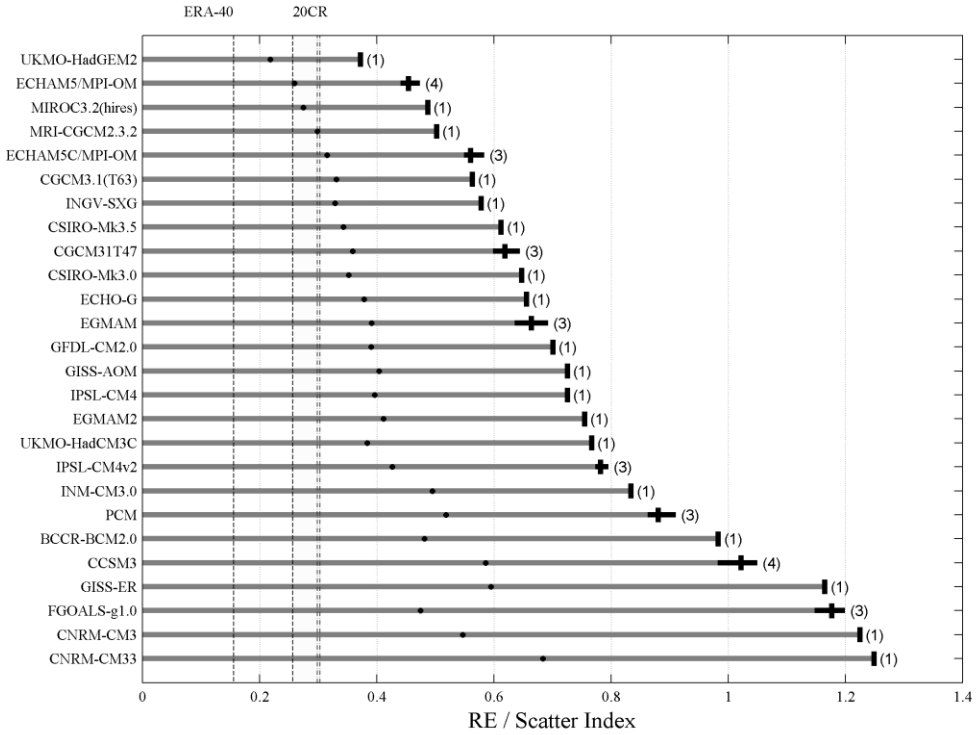

Fig. 6 GCMs of CMIP3 and ENSEMBLES sorted out by performance to model synoptic situations (the higher performance, the lower SI)

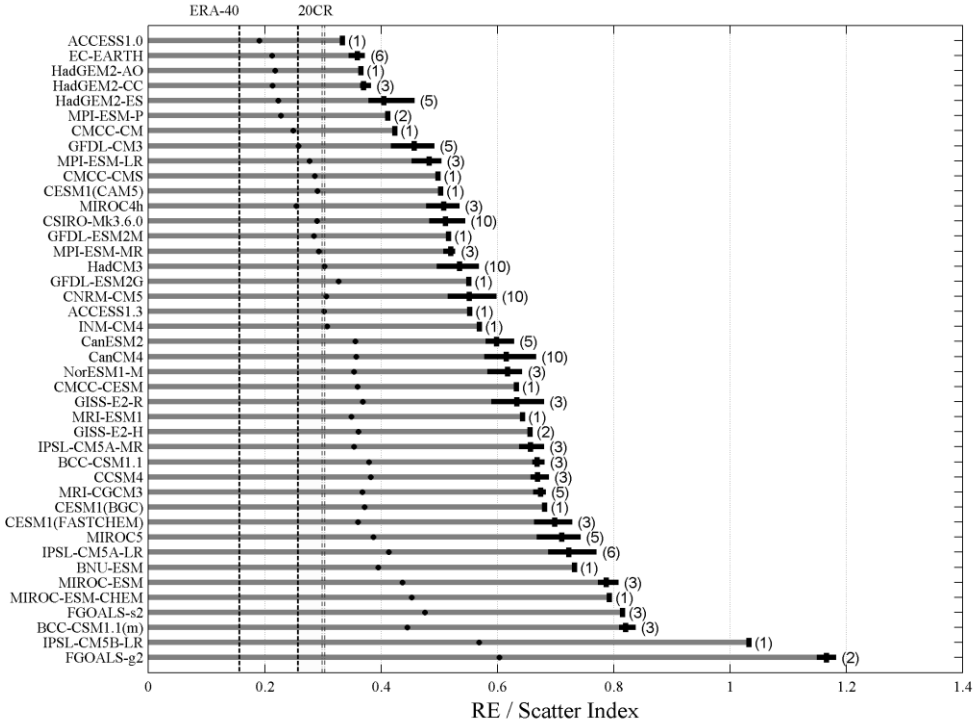

Fig. 7 GCMs of CMIP5 sorted out by performance to model synoptic situations (the higher performance, the lower SI)

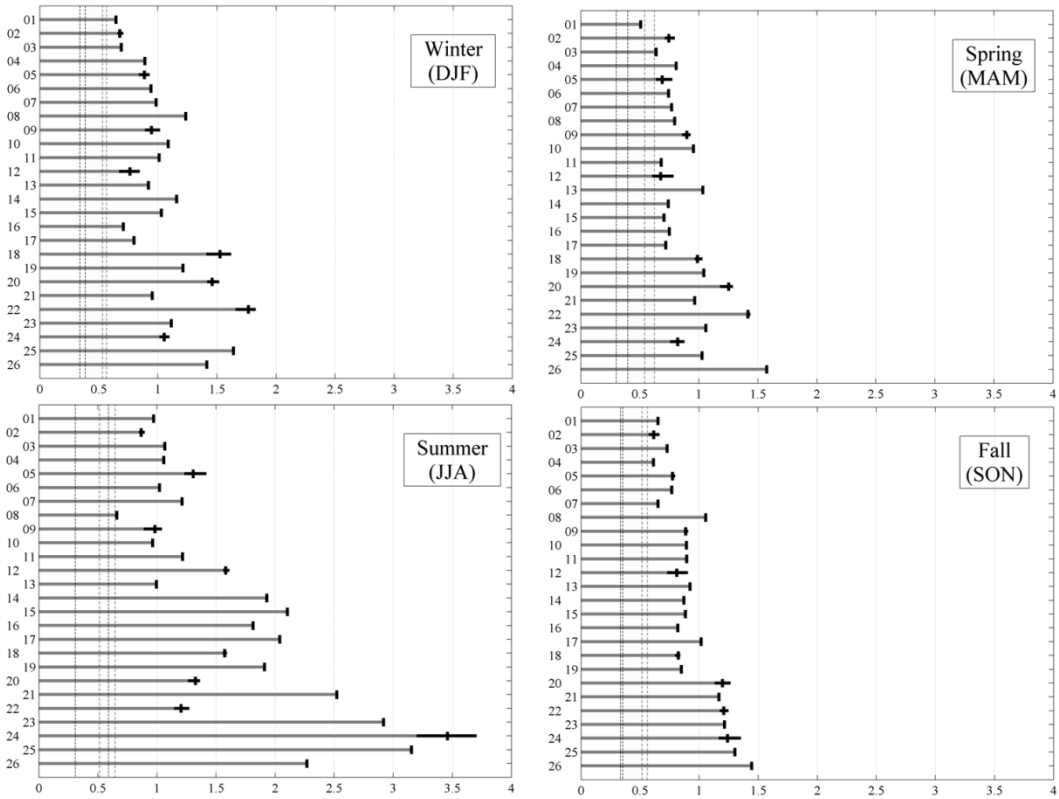

Fig. 8 GCMs of CMIP3 and ENSEMBLES performance to model synoptic situations on each season 1) UKMO-HadGEM2; 2) ECHAM5/MPI-OM; 3) MIROC3.2(hires); 4) MRI-CGCM2.3.2; 5) ECHAM5C/MPI-OM; 6) CGCM3.1(T63); 7) INGV-SXG; 8) CSIRO-Mk3.5; 9) CGCM31T47; 10) CSIRO-Mk3.0; 11) ECHO-G; 12) EGMAM; 13) GFDL-CM2.0; 14) GISS-AOM; 15) IPSL-CM4; 16) EGMAM2; 17) UKMO-HadCM3C; 18) IPSL-CM4v2; 19) INM-CM3.0; 20) PCM; 21) BCCR-BCM2.0; 22) CCSM3; 23) GISS-ER; 24) FGOALS-g1.0; 25) CNRM-CM3; 26) CNRM-CM33

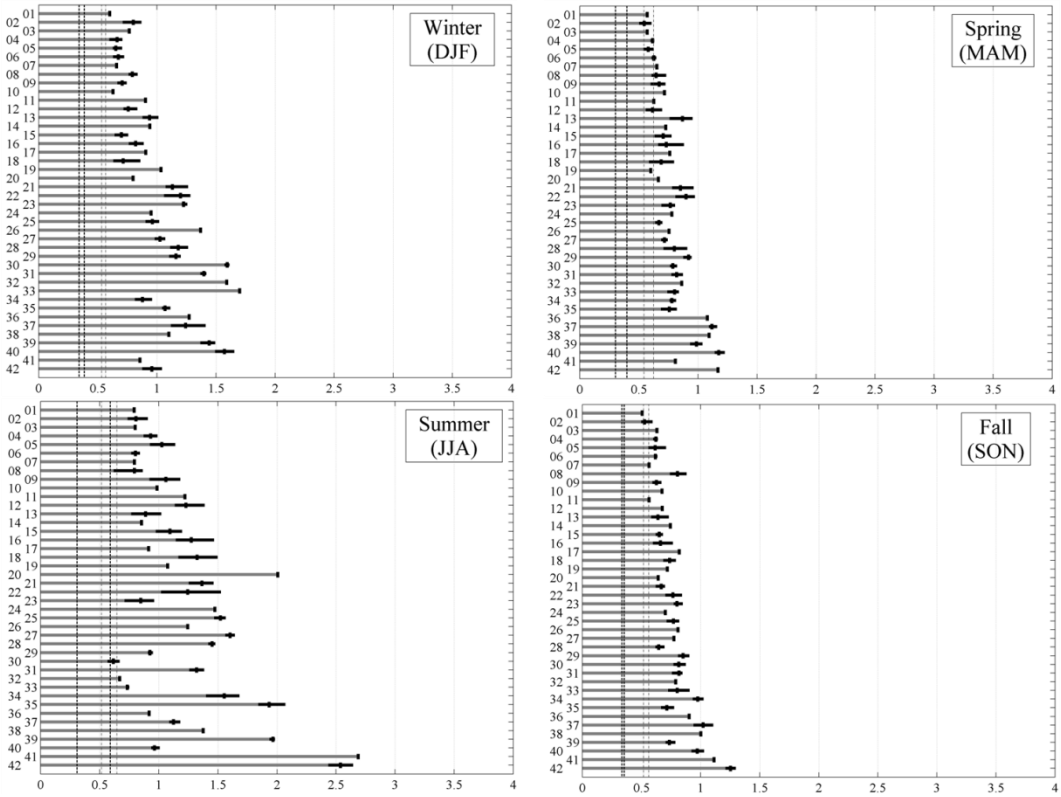

Fig. 9 GCMs of CMIP5 performance to model synoptic situations on each season 1) ACCESS1.0; 2) EC-EARTH; 3) HadGEM2-AO; 4) HadGEM2-CC; 5) HadGEM2-ES; 6) MPI-ESM-P; 7) CMCC-CM; 8) GFDL-CM3; 9) MPI-ESM-LR; 10) CMCC-CMS; 11) CESM1(CAM5); 12) MIROC4h; 13) CSIRO-Mk3.6.0; 14) GFDL-ESM2M; 15) MPI-ESM-MR; 16) HadCM3; 17) GFDL-ESM2G; 18) CNRM-CM5; 19) ACCESS1.3; 20) INM-CM4; 21) CanESM2; 22) CanCM4; 23) NorESM1-M; 24) CMCC-CESM; 25) GISS-E2-R; 26) MRI-ESM1; 27) GISS-E2-H; 28) IPSL-CM5A-MR; 29) BCC-CSM1.1; 30) CCSM4; 31) MRI-CGCM3; 32) CESM1(BGC); 33) CESM1(FASTCHEM); 34) MIROC5; 35) IPSL-CM5A-LR; 36) BNU-ESM; 37) MIROC-ESM; 38) MIROC-ESM-CHEM; 39) FGOALS-s2; 40) BCC-CSM1.1(m); 41) IPSL-CM5B-LR; 42) FGOALS-g2

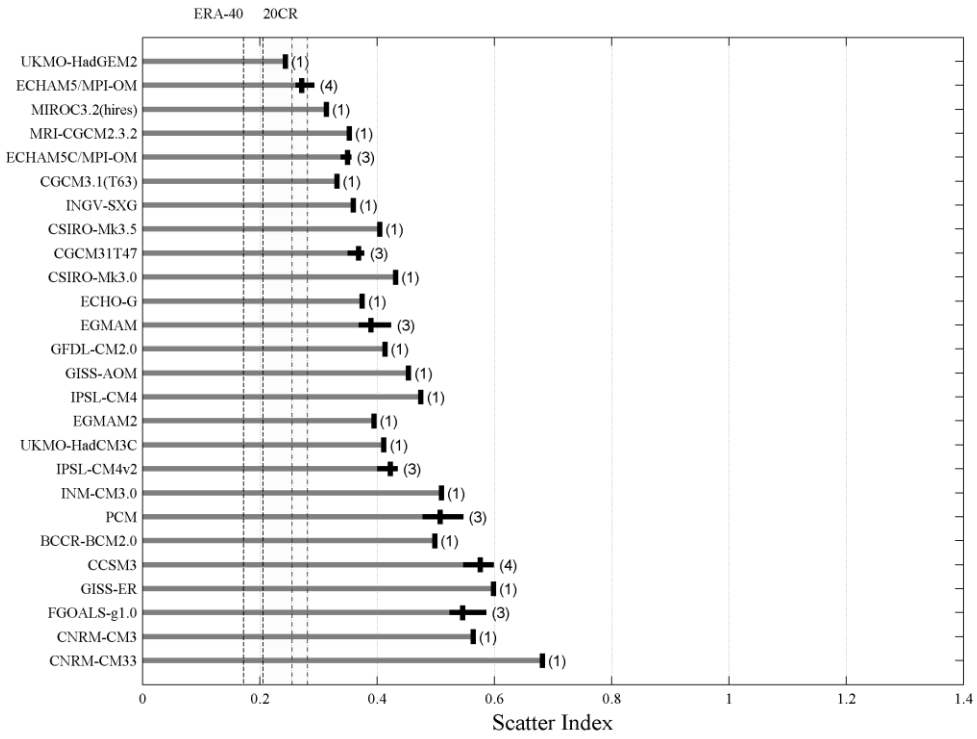

Fig. 10 GCMs of CMIP3 and ENSEMBLES performance to simulate interannual variability (the higher performance, the lower SI)

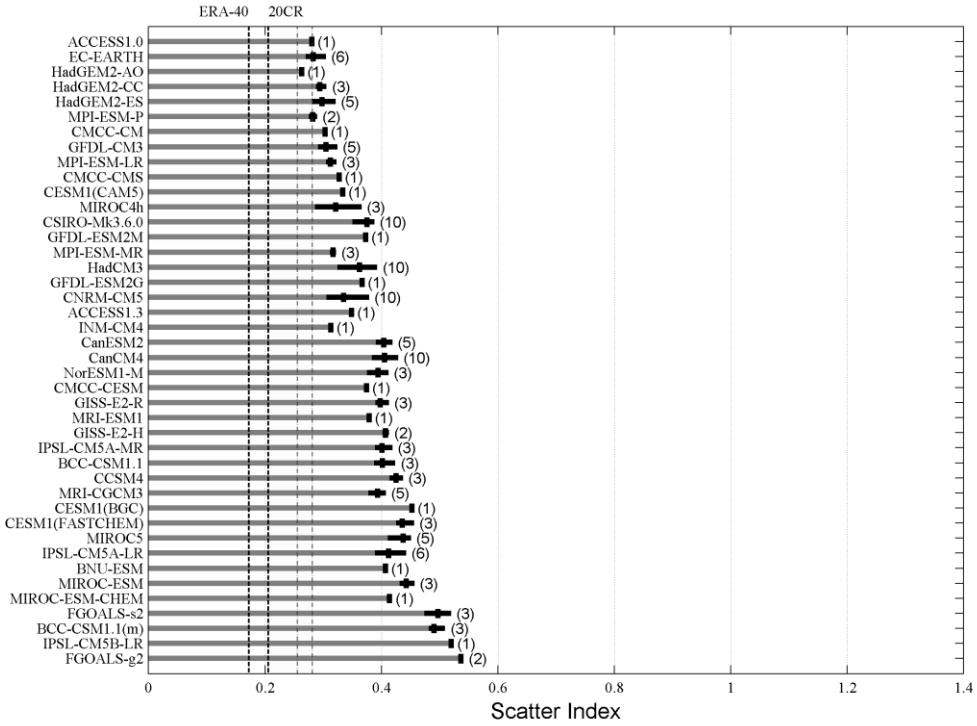

Fig. 11 GCMs of CMIP5 performance to simulate interannual variability (the higher performance, the lower SI)

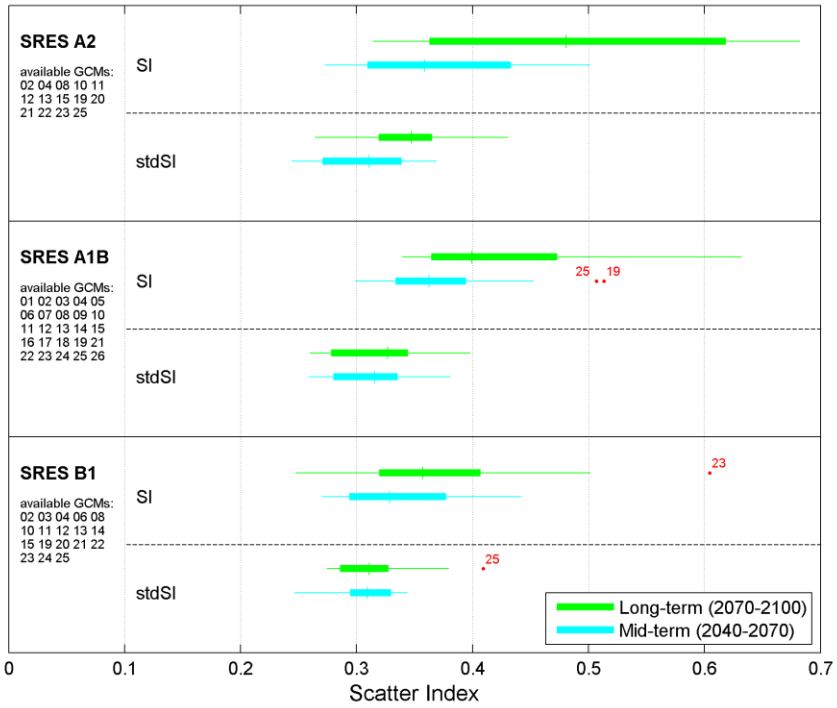

Fig. 12 Box plots of the two indicators of consistency for scenarios B1, A1B and A2. Numbering in accordance with figure 8.

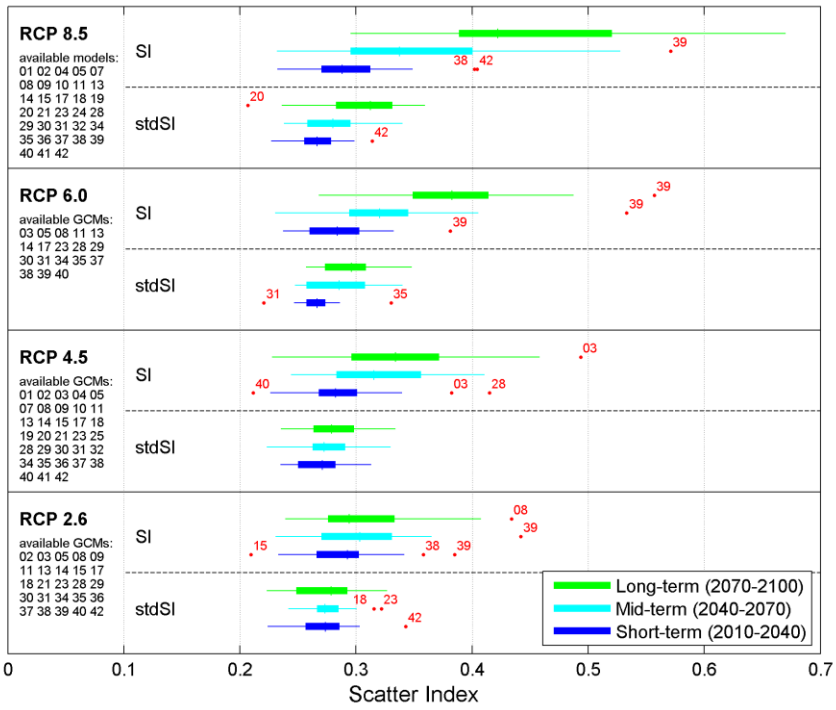

Fig. 13 Box plots of the two indicators of consistency for scenarios RCP2.6, RCP4.5, RCP6.0 and RCP8.5. Numbering in accordance with figure 9.

Table

Model	Institution	Country	Atmospheric resolution (lat x lon, number of layers)	Runs B1-A1B-A2
BCCR-BCM2.0	Bjerknes Centre for Climate Research	Norway	1.9º x 1.9º, L31	1-1-1
CCSM3	National Center for Atmospheric Research	USA	1.4º x 1.4º, L26	2-2-2
CGCM3.1(T47)	Canadian Centre for Climate Modelling and Analysis	Canada	2.8º x 2.8º, L31	0-3-0
CGCM3.1(T63)	Canadian Centre for Climate Modelling and Analysis	Canada	1.9º x 1.9º, L31	1-1-0
CNRM-CM3	Centre National de Recherches Météorologiques	France	2.8º x 2.8º, L45	1-1-1
CNRM-CM33	Centre National de Recherches Météorologiques	France	1.9º x 1.9º, L19	0-1-0
CSIRO-MK3.0	CSIRO Atmospheric Research	Australia	1.9º x 1.9º, L18	1-1-1
CSIRO-MK3.5	CSIRO Atmospheric Research	Australia	1.9º x 1.9º, L18	1-1-1
ECHAM5/MPI-OM	Max-Planck-Institute for Meteorology	Germany	1.9º x 1.9º, L31	3-4-3
ECHAM5C/MPI-OM	Max-Planck-Institute for Meteorology	Germany	3.75º x 3.75º, L19	0-3-0
ECHO-G	University of Bonn	Germany	3.9º x 3.9º, L19	1-1-1
EGMAM	Freie Universitaet Berlin, Institute for Meteorology	Germany	3.75º x 3.75º, L39	3-3-3
EGMAM2	Freie Universitaet Berlin, Institute for Meteorology	Germany	3.75º x 3.75º, L39	0-1-0
FGOALS-g1.0	Institute of Atmospheric Physics	China	2.8º x 2.8º, L26	3-3-0
GFDL-CM2.0	Geophysical Fluid Dynamics Laboratory	USA	2º x 2.5º, L24	1-1-1
GISS-AOM	Goddard Institute for Space Studies	USA	3º x 4º, L12	1-1-0
GISS-ER	Goddard Institute for Space Studies	USA	4º x 5º, L20	1-1-1
INGV-SXG	Istituto Nazionale di Geofisica e Vulcanologia	Italy	1.12º x 1.12º, L19	0-1-0
INM-CM3.0	Institute of Numerical Mathematics	Russia	4º x 5º, L21	1-1-1
IPSL-CM4	Institut Pierre Simon Laplace	France	2.5º x 3.75º, L19	1-1-1
IPSL-CM4v2	Institut Pierre Simon Laplace	France	2.5º x 3.75º, L19	0-3-0
MIROC3.2 (hires)	Center for Climate System Research, NIES and RCGC	Japan	1.12º x 1.12º, L56	1-1-0
MRI-CGCM2.3.2	Meteorological Research Institute	Japan	2.8º x 2.8º, L30	1-1-1
PCM	National Center for Atmospheric Research	USA	2.8º x 2.8º, L18	2-0-1
UKMO-HadCM3C	Met Office Hadley Centre	UK	2.5º x 3.75º, L38	0-2-0
UKMO-HadGEM2	Met Office Hadley Centre	UK	1.25º x 1.9º, L38	0-3-0

Table 1 caption. Analyzed CMIP3 and ENSEMBLES GCMs names, institutions, countries, atmospheric resolutions and runs.

Model	Institution	Country	Atmospheric resolution (lat x lon, number of layers)	Runs RCP2.6-RCP4.5 - RCP6.0-RCP8.5-
ACCESS1.0	CSIRO-BOM	Australia	1.25º x 1.9º, L38	0-1-0-1
ACCESS1.3	CSIRO-BOM	Australia	1.25º x 1.9º, L38	0-1-0-1
BCC-CSM1.1	Beijing Climate Center	China	2.8º x 2.8º, L26	1-1-1-1
BCC-CSM1.1(m)	Beijing Climate Center	China	1.12º x 1.12º, L26	1-1-1-1
BNU-ESM	College of Global Change and Earth System Science	China	2.8º x 2.8º, L26	1-1-0-1
CanCM4	Canadian Centre for Climate Modelling and Analysis	Canada	2.8º x 2.8º, L35	0-10-0-0
CanESM2	Canadian Centre for Climate Modelling and Analysis	Canada	2.8º x 2.8º, L35	5-5-0-5
CCSM4	National Center for Atmospheric Research	USA	0.94º x 1.25º, L26	3-3-3-3
CESM1(BGC)	Community Earth System Model Contributors	USA	0.94º x 1.25º, L26	0-1-0-1
CESM1(CAM5)	Community Earth System Model Contributors	USA	0.94º x 1.25º, L26	1-1-1-1
CESM1(FASTCHEM)	Community Earth System Model Contributors	USA	0.94º x 1.25º, L26	0-0-0-0
CMCC-CESM	Centro Euro-Mediterraneo per I Cambiamenti Climatici	Italy	3.71º x 3.75º, L39	0-0-0-1
CMCC-CM	Centro Euro-Mediterraneo per I Cambiamenti Climatici	Italy	0.75º x 0.75º, L31	0-1-0-1
CMCC-CMS	Centro Euro-Mediterraneo per I Cambiamenti Climatici	Italy	1.9º x 1.9º, L95	0-1-0-1
CNRM-CM5	Centre National de Recherches Météorologiques	France	1.4º x 1.4º, L31	1-1-1-1
CSIRO-Mk3.6.0	CSIRO-QCCCE	Australia	1.9º x 1.9º, L18	10-10-10-10
EC-EARTH	EC-EARTH consortium	Various	1.1º x 1.1º, L62	1-5-0-5
FGOALS-g2	LASG-CESS	China	2.8º x 2.8º, L26	1-1-0-1
FGOALS-s2	LASG-CESS	China	1.7º x 2.8º, L26	1-0-1-3
GFDL-CM3	NOAA Geophysical Fluid Dynamics Laboratory	USA	2º x 2.5º, L48	1-0-1-1
GFDL-ESM2G	NOAA Geophysical Fluid Dynamics Laboratory	USA	2º x 2.5º, L48	1-1-1-1
GFDL-ESM2M	NOAA Geophysical Fluid Dynamics Laboratory	USA	2º x 2.5º, L48	1-1-1-1
GISS-E2-H	NASA Goddard Institute for Space Studies	USA	2º x 2.5º, L40	0-0-0-0
GISS-E2-R	NASA Goddard Institute for Space Studies	USA	2º x 2.5º, L40	0-2-0-0
HadCM3	Met Office Hadley Centre	UK	2.5º x 3.75º, L19	0-10-0-0
HadGEM2-AO	Met Office Hadley Centre	UK	1.25º x 1.9º, L38	1-1-1-0
HadGEM2-CC	Met Office Hadley Centre	UK	1.25º x 1.9º, L60	0-1-0-3
HadGEM2-ES	Met Office Hadley Centre	UK	1.25º x 1.9º, L38	4-4-4-3
INM-CM4	Institute for Numerical Mathematics	Russia	1.5º x 2º, L21	0-1-0-1
IPSL-CM5A-LR	Institut Pierre-Simon Laplace	France	1.9º x 3.75º, L39	4-4-1-4
IPSL-CM5A-MR	Institut Pierre-Simon Laplace	France	1.25º x 2.5º, L39	1-1-1-1
IPSL-CM5B-LR	Institut Pierre-Simon Laplace	France	1.9º x 3.75º, L39	0-1-0-1
MIROC-ESM	MIROC	Japan	2.8º x 2.8º, L80	1-1-1-1
MIROC-ESM-CHEM	MIROC	Japan	2.8º x 2.8º, L80	1-1-1-1
MIROC4h	MIROC	Japan	0.56º x 0.56º, L56	0-3-0-0
MIROC5	MIROC	Japan	1.4º x 1.4º, L40	3-3-1-3
MPI-ESM-LR	Max-Planck-Institut für Meteorologie	Germany	1.9º x 1.9º, L47	3-3-0-3
MPI-ESM-MR	Max-Planck-Institut für Meteorologie	Germany	1.9º x 1.9º, L95	1-1-0-0
MPI-ESM-P	Max-Planck-Institut für Meteorologie	Germany	1.9º x 1.9º, L47	0-0-0-0
MRI-CGCM3	Meteorological Research Institute	Japan	1.1º x 1.1º, L48	0-0-0-0
MRI-ESM1	Meteorological Research Institute	Japan	1.1º x 1.1º, L48	0-0-0-0
NorESM1-M	Norwegian Climate Centre	Norway	1.9º x 2.5º, L26	0-0-0-0

Table 2 Analyzed CMIP5 GCMs names, institutions, countries, atmospheric resolutions and runs.