Hadley Centre Technical Note No. 102:

Selection of CMIP5 members to augment a perturbed–parameter ensemble of global realisations of future climate for the UKCP18 scenarios

Carol McSweeney, James Murphy, David Sexton, John Rostron, Kuniko Yamazaki, Glen Harris

If printing double-sided you will need this blank page. If printing single sided, please delete this page

Contents

1. Executive Summary

- The next set of UK Climate Projections, UKCP18, will include a set of plausible realisations of 21st century climate formed directly from climate model output, in order to provide a flexible multivariate dataset with full spatial and temporal coherence for impacts analysis.
- This product, hereafter Land Strand 2 (LS2), will be based on a new perturbed parameter ensemble of variants of the HadGEM3 climate model (HadGEM3-PPE).
- By augmenting the core 15-member HadGEM3-PPE with subset of the CMIP5 multi-model ensemble (MME) we can include structural uncertainties that the PPE cannot capture, and potentially expand the range of projections offered in Land Strand 2 at both global and regional scale.
- The CMIP5 subset will also act as a 'benchmark' when evaluating the performance of the HadGEM3-PPE. CMIP5 models are known to vary considerably in their representation of regional climate processes we thereby consider the benchmark for good performance in the PPE to be at least as good as the 'good' CMIP5 members.
- 31 CMIP5 models with sufficient data availability have been screened for their performance against relevant regional and global performance criteria, and near neighbours in order to identify a subset of 10-20 members capable of providing credible additional realisations for use across a range of assessments of UK risks and impacts
- The regional criteria that the models were assessed against included the climatological circulation patterns of the N. Atlantic/European sector, distribution of daily storm track latitudes, mean temperature biases, frequencies of daily weather types, blocking frequency and the realism of the AMOC. Global criteria included global mean climate variability and drift, and SST errors.
- Three models were found to be too unrealistic in their representation of key characteristics of the regional climate to provide useful projections for the UK. These were IPSL-CM5B-LR (UK affected by a cool bias of 8-9°C and very unrealistic summer circulation), FGOALS-g2 (very unrealistic circulation patterns in both summer and winter) and MIROC5 (unrealistic summer circulation, lacking westerly flow direction over UK).
- A further six models were rejected due to very poor performance across multiple regional and/or global criteria.
- From remaining models, 13 were selected, sampling from groups of 'near-neighbours'.
- The subset of 13 models recommended for use in Strand 2 is: ACCESS1-3, bcc-csm1-1, CCSM4, CESM1-BGC, CanESM2, CMCC-CM, CNRM-CM5, EC-Earth, GFDL-ESM2G, HadGEM2-ES, IPSL-CM5A-MR, MPI-ESM-MR, MRI-CGCM3.
- The 13 members selected capture much of the full CMIP5 range of seasonal mean changes in temperature and precipitation, but the exclusion of poorest models does impact the range of regional projections for the UK by discounting the member with lowest warming in the region (INMCM4) and those with the largest temperature and precipitation increases (MIROC-ESM and MIROC-ESM-CHEM).

2. UKCP18 Strand 2 Overview

The next set of UK Climate Projections, UKCP18, will include a set of plausible global realisations of 21st century climate formed directly from climate model output, in order to provide a flexible multivariate dataset with full spatial and temporal coherence for impacts analysis. This product is referred to as Land Strand 2 (from hereon 'LS2') and will be provided alongside probabilistic projection information (Land Strand 1) and downscaled regional projections (Land Strand 3).

The LS2 simulations will consist of a core 15-member Perturbed Parameter Ensemble (PPE) based on HadGEM3-GC3.05 (from hereon, HadGEM3-PPE). A key requirement in the design of the LS2 projections is that they should be 'plausible' and 'diverse'. By 'plausible', we mean that the models should be sufficiently realistic in their representation of the key characteristics of the UK climate that they are useful for climate change impacts studies; by 'diverse' we mean that the projections should capture as wide a range of characteristics of plausible future climate changes as possible in order to reflect uncertainty and provide useful inputs to decision making.

In the design of the HadGEM3-PPE, the plausibility and diversity of model variants using different parameter combinations was assessed via several stages of modelling and assessment. This process is described in more detail in the UKCP18 Projections Report (Murphy et al, 2018), but briefly for context here, parameter combinations were 'ruled out' based on the following steps and criteria:

- **Quantitative screening**: Initially, several thousand 5-day 'Transpose AMIP'-style experiments were used to identify plausible parameter combinations based on global and regional-mean-square errors across a range of key atmospheric variables. Thresholds for acceptability were established relative to the performance of the model variants using standard unperturbed values for the model parameters. A similar quantitative assessment was applied to 495 5-year atmosphere-only (AMIP-Style) experiments.
- **Qualitative screening**: 49 members passed the quantitative screening based on AMIP-experiments. These were assessed qualitatively, via an assessment of key regional characteristics such as circulation patterns and significant local temperature biases to identify and reject members which demonstrated significantly poor regional performance
- **Sampling of diverse members based on idealised experiments**: Idealised experiments for the AMIP members allowed the quantification of aerosol forcing, CO_2 forcing, components of feedback and regional responses to SST warming patterns. From the remaining 49, 30 members were selected in order to best sample metrics representing aerosol forcing, components of feedback, and 9 quasi-independent Giorgi-Francisco regional responses in temperature and precipitation.
- **Screening of coupled simulations:** A final screening of the 25 coupled runs as they progressed allowed for any remaining members that were demonstrating implausible characteristics when run in coupled mode to be eliminated. While the intention was to provide a core PPE of 20-25 members, several issues resulted in a larger number of models being dropped at this stage than expected. Two members suffered frequent instabilities which made continuing the run unfeasible to complete in good time while three other members were found to exhibit a spin-down of the AMOC and

resulted in unrealistic extended Northern Hemisphere sea-ice extents. 5 further members were dropped at later stage following more extensive regional evaluation of temperature biases over Europe, AMOC strength and hemispheric cooling over the 20th Century (See Murphy et al, 2018 for further details).

The range of projections that could be offered in LS2 at both global and regional scale can potentially be expanded by including a subset of members of the WCRP CMIP5 coupled multi-model ensemble (from hereon, CMIP5-MME) (Taylor *et al.,* 2012, thus adding a range of structural model uncertainties to the parameter uncertainties captured by the HadGEM3-PPE. The CMIP5 ensemble of future projections under RCP8.5 consists of 42 models that vary significantly from one another in their performance in reproducing recent climate - particularly at the regional scale (e.g. McSweeney et al., 2015a). However, the ensemble also includes closely-related pairs or clusters of models that share significant parts of their code, and consequently share similar error and projection characteristics (e.g. Sanderson et al, 2015a&b). A subset of 10–20 models that are (a) reasonably independent of one another and (b) perform well against a set of criteria that are judged to be relevant is certainly easier to apply and interpret in impacts studies, and likely to be at least as skilful (if not more skilful) as using the full ensemble (Sanderson et al., 2015a).

A further consideration in augmenting the UKCP18 Strand 2 projections with CMIP5 is the consistency in the projection ranges across other products within the UKCP18 project. Land Strand 1 will consist of probabilistic projections that will reflect modelling uncertainties diagnosed from earlier PPEs that are based on the HadCM3 model, in combination with results from emissions-driven earth system model projections from CMIP5. These probabilistic projections will be provided for several emissions scenarios including RCP8.5, in which the ranges of response represent the combined effects of uncertainties in physical, aerosol and carbon cycle processes to a given emissions pathway. Including projections from concentration-driven CMIP5 models (many of which are the same, or closely related, to their emissions-driven earth system model counterparts) in the UKCP18 Land-Strand 2 projections is therefore likely to improve consistency between the Strand 1 and 2 products in UKCP18.

The HadGEM3 PPE simulations in Strand 2 are constructed from a physical ocean-atmosphere model lacking an interactive carbon cycle, and using prescribed CO_2 concentrations. However, in contrast to the CMIP5 simulations considered in this report, the HadGEM3 PPE members use *a range* of prescribed future CO_2 pathways designed to be approximately consistent with the range of CO_2 outcomes projected in Strand 1. This design choice was made to improve consistency between Strands 1 and 2, although it does mean that outcomes for future global temperature rise in the PPE reflect uncertainties in both physical and carbon cycle feedbacks, whereas the corresponding CMIP5 outcomes reflect only physical feedbacks.

3. Approach to Selection of CMIP5-MME subset

In order to be able to include any CMIP5 model in the UKCP18 Strand 2 ensemble, we require that daily data for core variables (precipitation, mean, daily min/max temperature) are available from both historical and rcp85 experiments. 31 models meet this requirement and therefore form the inital pool of models from which the subset is selected.

The selection of suitable parameter combinations for the HadGEM3-PPE follows the principles that ensemble members should be 'plausible' and 'diverse'. For the CMIP5-MME sub-selection, we apply a two-stage screening process.

- Firstly, all models are screened qualitatively for significant performance shortcomings at global and regional scale that are judged to significantly compromise the usefulness of those model projections for regional climate change studies relevant to the UK. The use of global information reflects our assumption that global performance provides evidence relating to the general plausibility of the physical assumptions built into a climate model, while the subsequent regional analysis indicates the extent to which a combination of remote and local drivers of error might compromise credibility relating specifically to UK applications.
- Secondly, remaining models are screened for 'near-neighbours' which share significant portions of code, and are known to generate simulations with similar error and projection characteristics. This stage of the screening is intended to reduce the number of models in the subset whilst maintaining the diversity of the projections.

Judging the 'plausibility' of projections is a significant challenge in climate change studies, and inevitably requires a level of subjectivity in terms of both the criteria/characteristics that are selected for assessment and the thresholds for plausibility. This study follows the approach of McSweeney et al. (2015a, 2015b) in using a qualitative performance assessment across multiple criteria to flag models with poorer performance. This approach lends some transparency to the subjective assessment. The following categories of performance 'flags' are used:

- **Red:** Model clearly not fit to provide useful regional projections for Europe because a feature is so poorly reproduced that relevant changes in regional climate can't be usefully diagnosed or applied in impacts assessments

- **Orange:** Significant bias/error, but its impact on the reliability of projections is unclear, or information not available for all/most models

- **Yellow:** Relatively poor/unrealistic compared with other models

- **Grey:** data/analysis not available.

The performance assessment draws on information available in published literature where possible, with some additional analysis of seasonal mean fields of key regional variables such as SST, near-surface circulation and precipitation patterns, and annual cycles of regional mean temperature and precipitation. In published studies, typically only a subset of CMIP5 models are considered, and so the 'grey' classification above becomes important where models are consistently excluded across multiple assessments (e.g. this may have occurred where models were later additions to the CMIP5 archive).

Where possible the criteria applied to the CMIP5-MME reflects equivalent criteria applied in the design of the HadGEM3-PPE. However, for several reasons relating to fundamental differences between MMEs and PPEs, the process and criteria must differ somewhat between the two ensembles. In particular:

- The CMIP5-MME displays a wider range of structural errors than those found in the HadGEM3-PPE. This is partly because the HadGEM3-PPE members are flux adjusted in order to suppress the development of systematic biases in sea surface temperature. The range of error characteristics exhibited by each ensemble may well not be sufficiently comparable to apply equivalent performance thresholds for inclusion to both ensembles. We see this in a comparison between SST errors between the CMIP5-MME and the initial set of 25 flux corrected coupled PPE members, whereby the lowest mean square error seen in CMIP5 is similar to the largest seen in the 25 coupled PPE simulations.

- We assess plausibility at different stages in the experiment in the two methods. For the CMIP5-MME, the assessment is post-hoc – this means that we have a data from full coupled historical and future simulation for all CMIP5-MME members to draw on when the assessing the models, as well as existing published literature to draw on. For the HadGEM3-PPE, the ensemble is designed by 'ruling-out' parts of parameter space or specific parameter combinations, using cheaper (i.e. shorter 5-day and 5-year) experiments that are run in advance of any coupled model experiments, in order to try to identify parameter combinations that we can expect to produce a set of plausible and diverse coupled runs. However, once the coupled PPE simulations are completed, it will be possible to compare their historical performance on a more like-for-like basis (subject to caveats associated with use of flux adjustments in the PPE), against the CMIP5 members selected in the present report. This will be done in future journal papers.
- Projection diversity was used as an *a priori* selection criterion for the HadGEM3-PPE for CMIP5 this is instead done by filtering for 'near-neighbours'. For the PPE, a relatively large number of members passed the performance criteria (50) which allowed remaining members to be filtered against further criteria which explicitly sampled a number of projection characteristics. For CMIP5, only 22 passed performance criteria, of which several were known to be very closely related. This smaller number of remaining members did not allow for a more explicit sampling of projection characteristics.

4. Evidence Relating to CMIP5 Model Performance

4.1. Global Characteristics

4.1.1. Global Variability and Drift

Figure 1 shows timeseries of annual global mean surface temperature from pre-industrial control runs that provide useful information about (a) drift and (b) internal variability.

Models which exhibit significant drift in the pre-industrial period are GFDL-CM3, IPSL-CM5B-LR, MIROC-ESM and MIROC-ESM-CHEM (orange).

Models in the CMIP5 ensemble display a wide range of different characteristics in their global mean temperature variability on interannual-multi-decadal timescales (Figure 1). Direct comparisons of internal variability on decadal and longer timescales between observations and long control runs such as these are limited by the lack of availability of observations records of equivalent length, and difficulties in separating internal variability from forced responses given the significant uncertainties in forcings and responses (Sutton et al., 2015). We can, however, look at the characteristics on inter-annual variability, and given the wide ranging characteristic of the ensemble, identify some models which are clearly not demonstrating variability on interannual timescales which is broadly in order of that in the variability of observations. For example, several models in Figure 1 exhibit particularly low interannual variability – these include inmcm4 (orange), FGOALS-g2 (orange) and the GISS models (the latter are not part of our analysis here but are noted). This very low variability indicates a poor representation of the key modes of variability such as ENSO and the PDO, and/or poor representation of the atmosphere-ocean coupling.

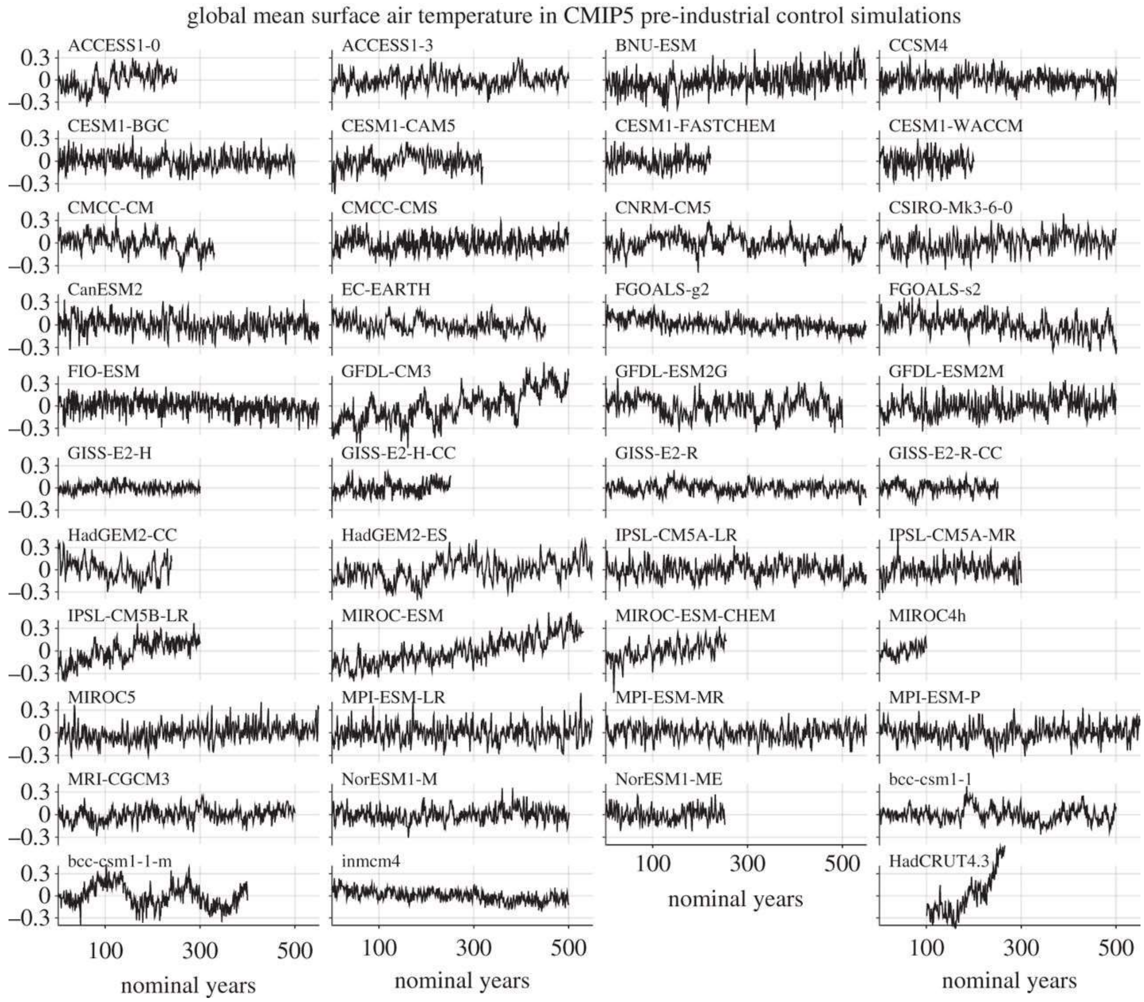

Figure 1: Internal variability and drift in annual mean global mean surface temperature anomalies for CMIP5 pre-industrial control experiments. HadCRUt4 observations are also shown for comparison. Taken from Sutton et al., 2017.

4.1.2. Remote SST errors

The Southern Ocean is the site of a climatically important interface between the ocean interior, the atmosphere and the cryosphere. The strong ocean–atmosphere coupling in the region means that the Southern Ocean makes a disproportionately large contribution to maintaining the global climate (Meijers, 2014). However, many models in the CMIP5 ensemble suffer from a pervasive warm bias in the Southern Ocean. We identify those models with the largest annual mean warm biases, as inmcm4, IPSL-CM5B-LR, MIROC5 and GFDL-ESM2M (orange), and to a lesser extent, CNRM-CM5, HadGEM2-ES, HadGEM2-CC, MRI-GCGCM3, NorESM1-M and GFDL-ESM2G (yellow).

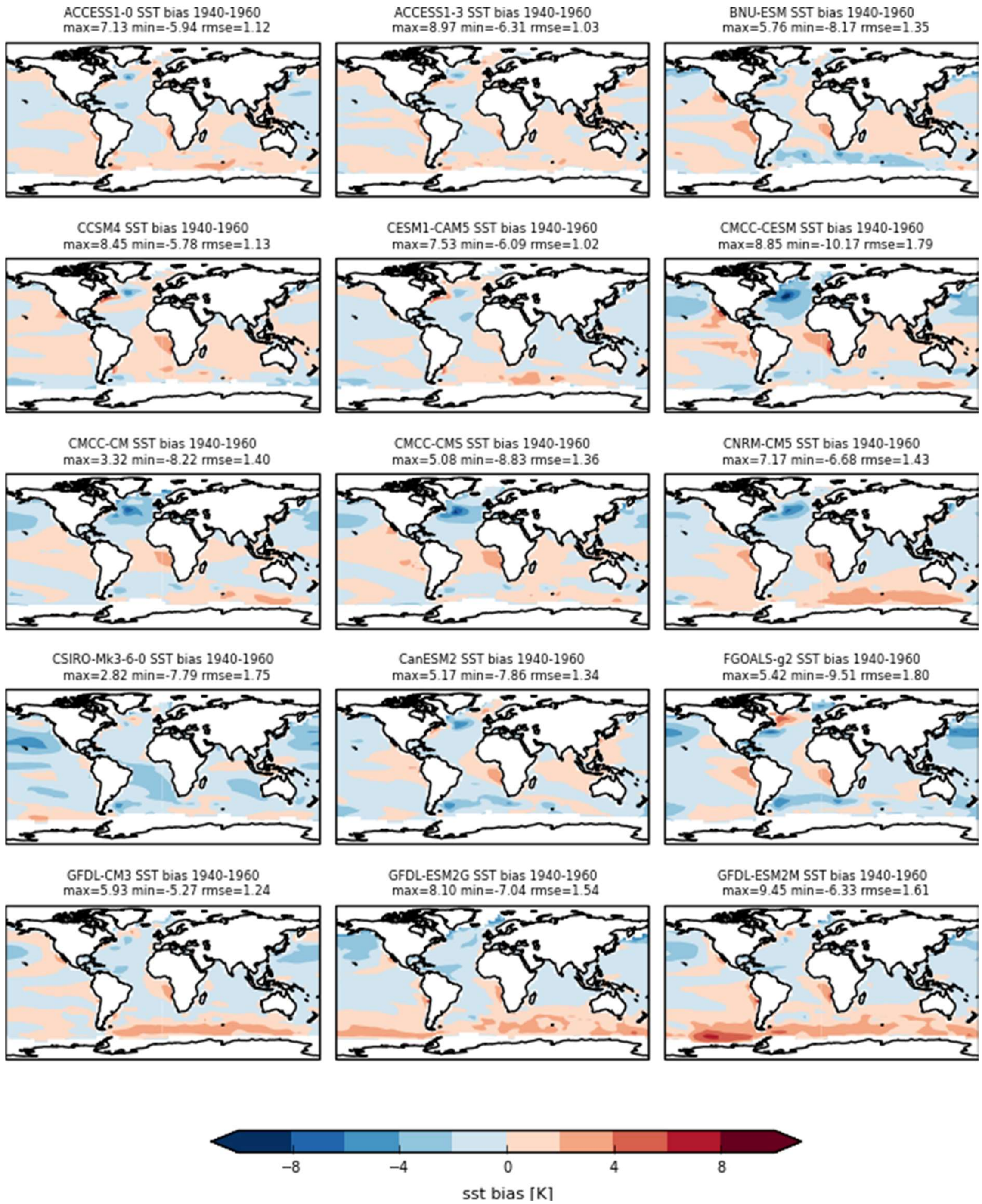

Figure 2(a): Annual mean SST biases in CMIP5 models relative to Had ISST for the period 1940-1960.

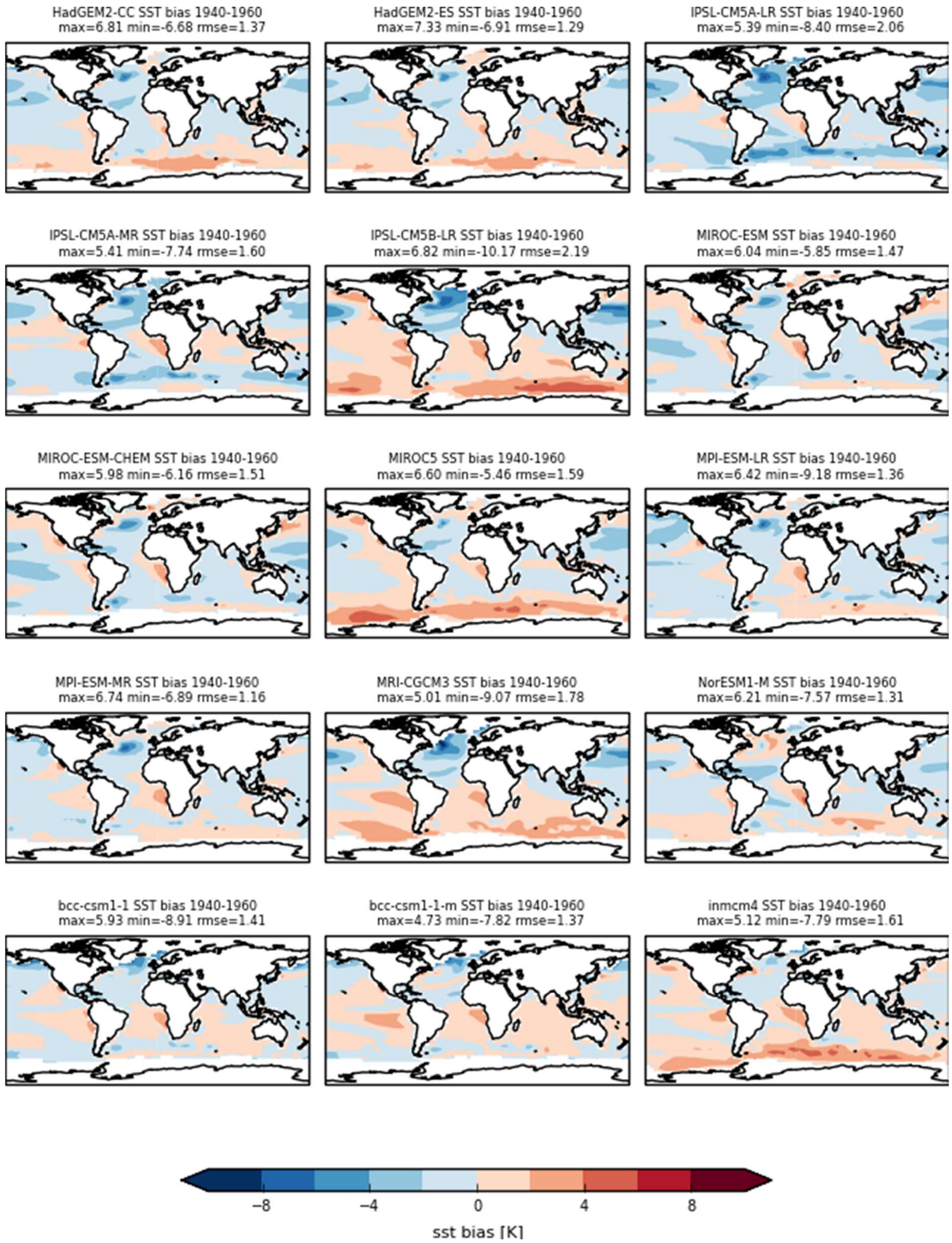

Figure 2 (b): Annual mean SST biases in CMIP5 models relative to Had ISST for the period 1940-1960.

Another region where pervasive SST errors persist in global coupled models is the equatorial pacific. An excessive and narrow SST cold tongue that extends too far west into the western Pacific in comparison to observations is a common bias in coupled GCMs. Whilst this feature is not evident in Figure 2, due to the relatively coarse scale of the plot, the cold tongue biases are clear in analysis McSweeney *et al*, 2015b , where CSIRO-mk3-6-0 and inmcm4 are identified as the most extreme examples of this error (orange).

4.2. Regional characteristics

4.2.1. Atlantic/European climatological circulation

The UK climate is characterised by the transit of large-scale weather systems across the North Atlantic, and capturing the climatological prevailing westerly/south-westerly surface wind direction can therefore be considered to be fundamental to representing UK climate. Here we look at the climatological mean (1981-2000) 850hpa circulation compared with ERA40 (Figure 3).

In winter (DJF), a southern bias in the latitude of the region of strongest south-westerly flow, and a tendency for flow to be too zonal are common to many of the CMIP5 members. One member is flagged for particularly unrealistic performance - FGOALS-g2 (Red), in which the region of strongest flow is so far south that it misses all but the very southern edge of the UK.

Other members with orange or yellow flags are:

- MIROC5 (orange), which contrary to the common error toward a too zonal pattern has a strongly ‘omega shape’ pattern such that flow over the UK is too weak and north westerly.
- CSIRO-Mk3-6-0 (orange) – too far south and turning southerly over Northern UK/Scotland
- MIROC-ESM-CHEM/MIROC-ESM/BNU-ESM (orange) – flow is particularly too far south and too zonal, so flow weak over northern UK and Scotland.
- CCSM4, bcc-csm1-1, bcc-csm1-1-m, MRI-CGCM3 (yellow) excessively strong flow.

These members can be compared directly to examples of model variants which were rejected in the design of the HadGEM3-PPE (Figure 4). Rejected members in the PPE display similar characteristics and magnitudes of errors as models with ‘red’ and ‘orange’ flags in the CMIP5-MME assessment. Similar errors in both ensembles are instances of the exaggerated ‘omega’-shape in MIROC5 and PPE-member 1102900, and the overly zonal characteristic of the CSIRO-Mk-3-6-0 and FGOALS-g2 models and PPE-member 1102823.

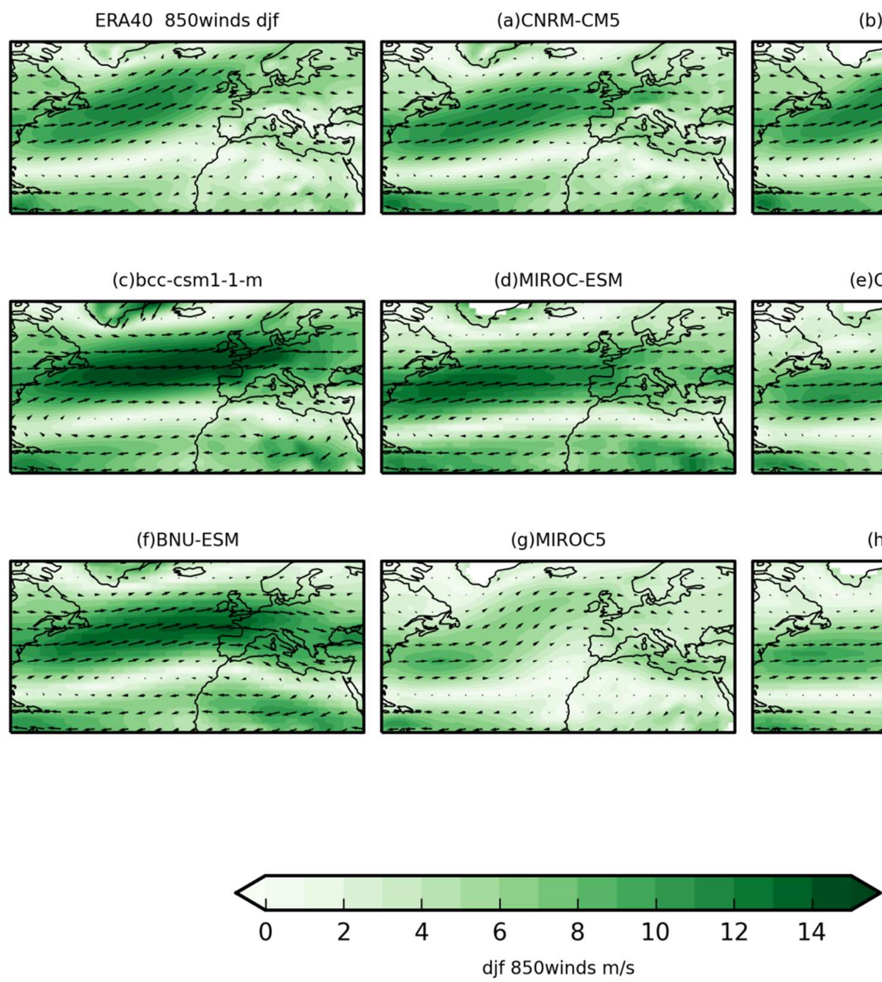

Figure 3: 850hpa circulation in December-January-February (DJF) 1981-2000 for ERA40 observations and in selected example models (a-h) from CMIP5. (a)-(b) indicate acceptable models, (c) is rated 'yellow', (d)-(f) are 'orange' and (h) 'red'. For a full set of plots see Appendix.

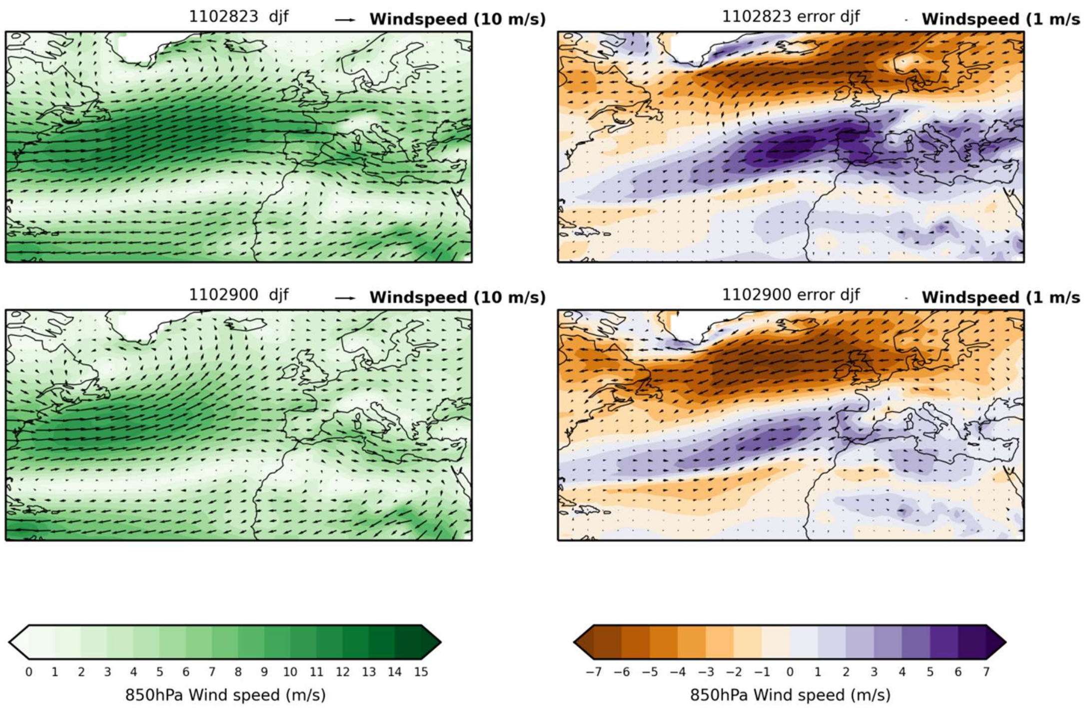

Figure 4: Examples of 'unacceptable members' DJF circulation patterns in 5-year AMIP experiments used in the design of the HadGEM3-PPE. Plots show DJF mean for years 2004-2009, errors relative to ERA40.

In summer (JJA) the prevailing westerly in the mean circulation is less strong but remains a significant driver of weather systems across the Atlantic. Some models fail to capture flow extending across the UK (Figure 5), either because the latitude of peak flow is too far south (e.g. FGOALS-g2 - red) or is too weak and dissipates too far west or because flow is too weak or both (MIROC5, IPSL-CM5B-LR – red).

Other flagged models are:

- MIROC-ESM, MIROC-ESM-CHEM, CMCC-CESM, flow too far south, thus very weak over northern UK/Scotland (yellow)
- bcc-csm1-1-m flow consistently too strong throughout region (yellow).
- CNRM-CM5 – flow too far south and southerly over UK (yellow).

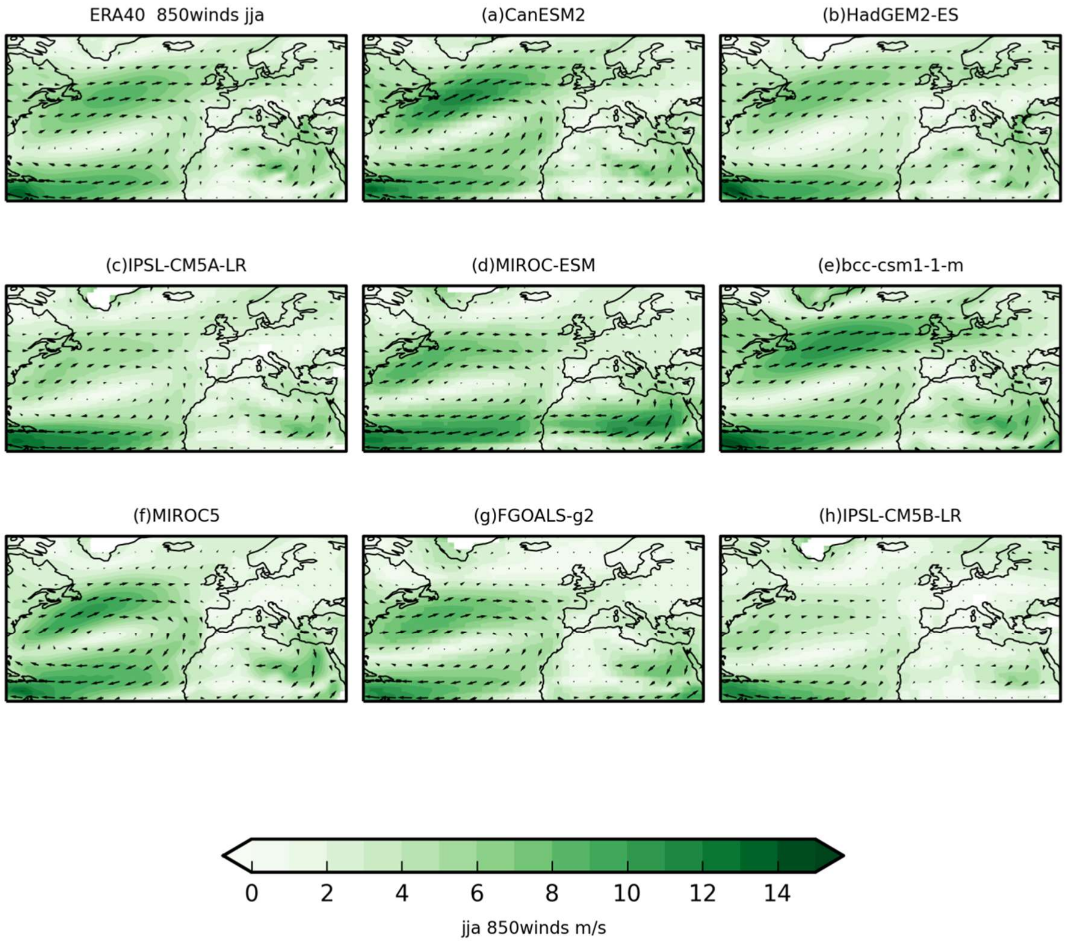

Figure 5: 850hpa circulation in June-July-August (JJA) 1981-2000 for ERA40 observations and in selected example models (a-h) from CMIP5. (a)-(b) indicate acceptable models, (c)-(e) are rated 'yellow', (f)-(h) 'red'. For a full set of plots see Appendix.

4.2.2. North Atlantic Storm track

The north Atlantic storm track in the storm seasons of autumn and winter is characterised by a tri-modal latitudinal distribution, with peaks at around 38-42, 55-60 and 70-75°N. While this will in part be determined by the realistic representation of the climatological circulation already assessed in Section 4.2.1, this additional analysis tells us about the variability within the season.

While most CMIP5 models broadly capture the three modal peaks in the distribution in some form, (Figure 6), almost all over estimate the frequency of tracks occurring at 40-50°N. FGOALS-g2 (red) is notable for its absence of either of the more northern peaks, reflecting the exceptionally too-far –south mean circulation in Figure 3. Four other models are distinct from the ensemble in failing to represent the two distinct modal peaks at 35-40 and 55-60, with too many tracks occurring throughout the 40-60°N latitudes – bcc-csm1-1, BNU-ESM, MIROC-ESM and MIROC-ESM-CHEM (orange), while bcc-csm1-1-m fails to capture the northern-most peak (orange).

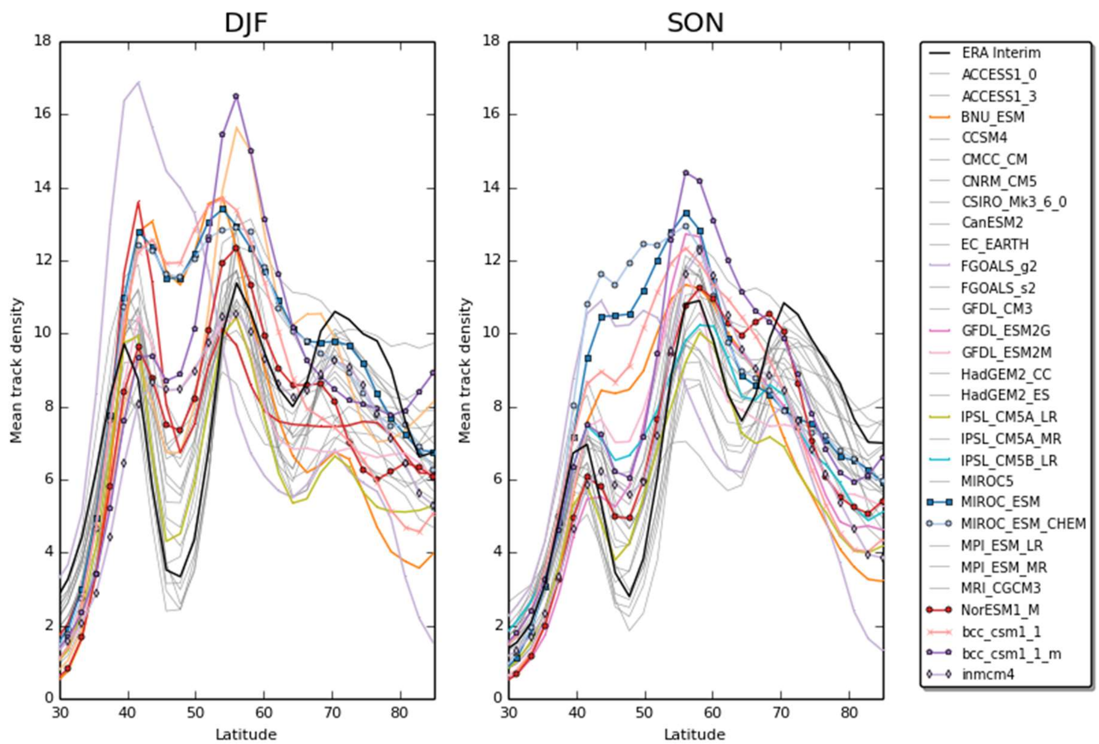

Figure 6: Zonal mean (5-25°) storm track density in CMIP5 historical simulations versus ERA-Interim. 'least realistic' models (those with highest RMS and lowest correlation between observations and cmip5 member) are highlighted in colour/markers. Figure originally appeared in McSweeney et al., 2015a).

4.2.3. Regional mean temperature biases

CMIP5 models span a range of biases in mean air temperature of around +- 2.5°C (Figure 7A notable exception to this is IPSL-CM5B-LR (red) which displays a more significant cold bias of -8-9°C in winter (Figure 7). This extreme cold bias takes mean winter temperatures well below zero. Figure 8 shows the geographical extent of the cool bias; surface air temperature is cooler than observed throughout the North Atlantic and European region but the most excessive cold biases of more than 12 degrees particularly affect northern Europe, including Northern England and Scotland.

Looking similarly at SST biases in the north Atlantic sector (Figure 2), there is a tendency for several models to demonstrate significant cool biases. The larger biases are negative, notably IPSL-CM5B-LR demonstrating extreme cool temperatures, but also 8 further models in which there are significant cool regions in the north Atlantic of 4-6°– CMCC-CM, CMCC-CMS, CMCC-CESM, IPSL-CM5A-LR, IPSL-CM5A-MR, MRI-CGCM3 (orange), and to a lesser extent, bcc-csm1-1-m and CNRM-CM5 (yellow).

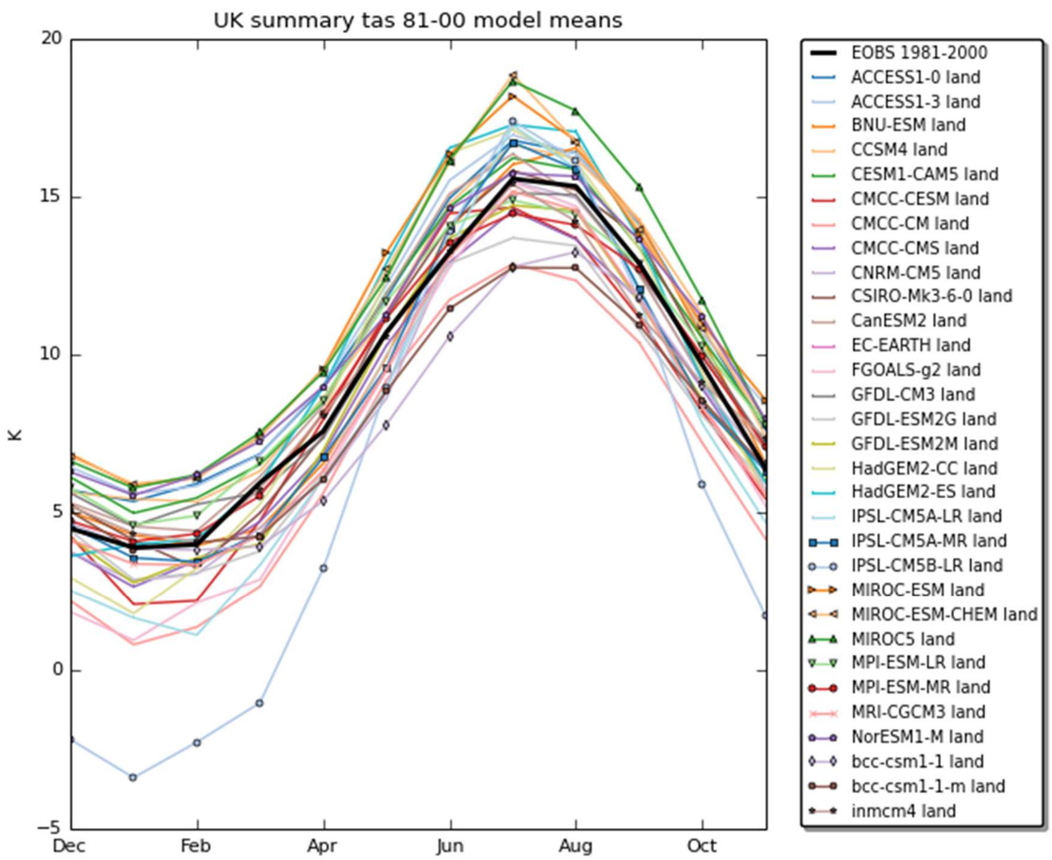

Figure 7: Mean near-surface air temperature 1981-2000 over UK land grid points between 10W-5E, 50N-63N in CMIP5 models and E-OBS observations (Hofstra et al, 2009).

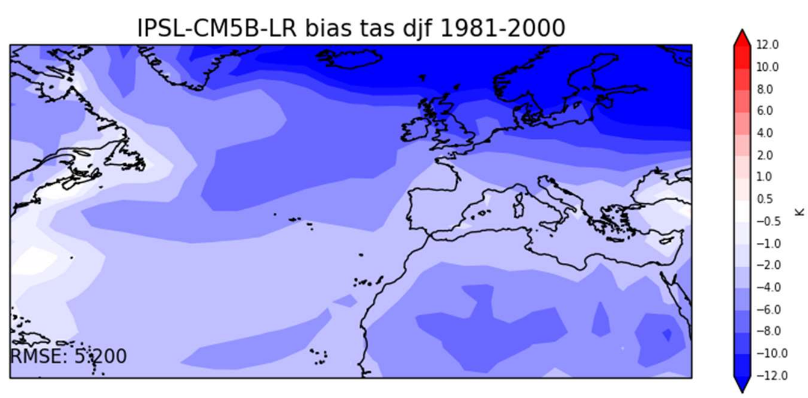

Figure 8: DJF mean temperature bias versus E-OBS observations (Hofstra et al, 2009). 1981-2000 in IPSL-CM5B-LR.

4.2.4. Weather type frequencies

Weather typing provides the opportunity to assess a models ability to reproduce the variability in synoptic states in the North Atlantic region which cause the wide range of meteorological conditions in the UK. Perez *et al*. (2014) use a classification system based on 100 weather types applied to daily mean sea level pressure from members of CMIP5 and ERA40 data sets to assess the relative realism of each member. Perez *et al*. (2014) calculate a performance 'scatter index, normalised by frequency,' that accounts for the overall frequency of weather types and also the interannual variability. A relative entropy score is also provided, which representing the models ability to capture low-probability weather types.

Given that these indices represent an aggregated view of performance over many individual weather patterns, we use the information in do not attempt to make absolute (red) judgements on credibility, limiting ourselves to relative judgements (orange, yellow). FGOALS- g2 and IPSL-CM5B-LR (orange) performed significantly worse than the other models in the frequency of the weather types (Figure 9), which is perhaps not surprising given the errors in mean climatological circulation shown in section 4.2.1. Other models which appeared in the bottom 10 models by this index are bcc-csm1-1-m, BNU-ESM, IPSL-CM5A-LR, MIROC5, MIROC-ESM, MIROC-ESM-CHEM (yellow). Inmcm4 demonstrated particularly poor scores in summer (not shown), offset by good performance in other seasons, so this is also flagged (yellow).

4.2.5. Blocking Frequency

Blocking events can be responsible for extended periods of settled UK weather, which often lead to damaging impacts such as dry, cold snaps in winter and heatwaves in summer. Blocking frequency is, however, typically underrepresented by GCMs largely as a result of biases in the mean state (Kennedy *et al*., 2016; Scaife *et al*, 2010), such as errors in the latitude of the tropospheric jet (Masato *et al.,* 2013).

Annual mean blocking frequency is summarised in Figure 10 from the IPCC AR5 WGI report. While a number of the models underestimate the magnitude of blocking frequency, several mis-represent the longitudinal position of the key blocking region in the Atlantic sector (-60-60); MIROC-ESM-CHEM and FGOALs-g2, and bcc-csm1-1-1m (orange) demonstrate significantly underestimated blocking frequency in both the Atlantic and Pacific sectors. BNU-ESM (red) shows almost no blocking at the longitudes of the UK (-10 to 5).

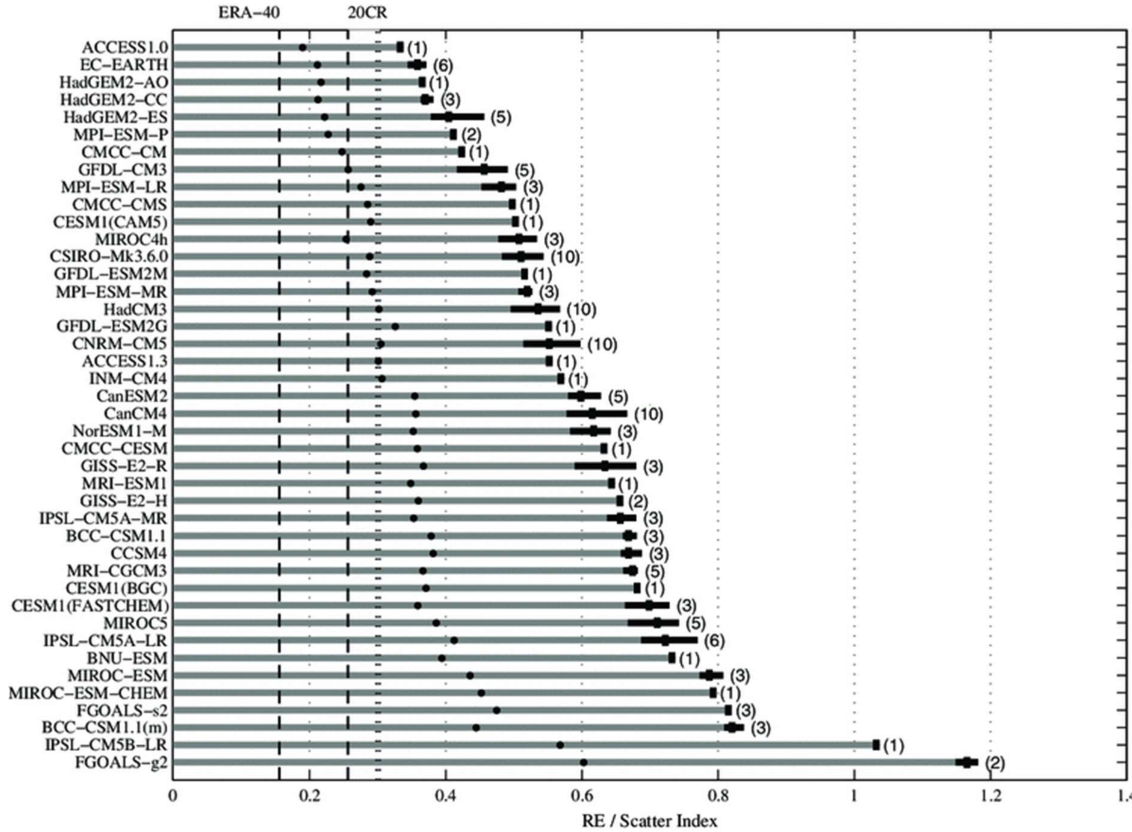

Figure 9: CMIP5 models sorted by models performance (higher performance gives a lower 'SI').The mean Relative Entropy is marked with a black dot. The vertical dashed and dotted lines show RE and SI scores obtained by comparing alternative reanalyses of observations. Figure reproduced from Perez et al., 2014.

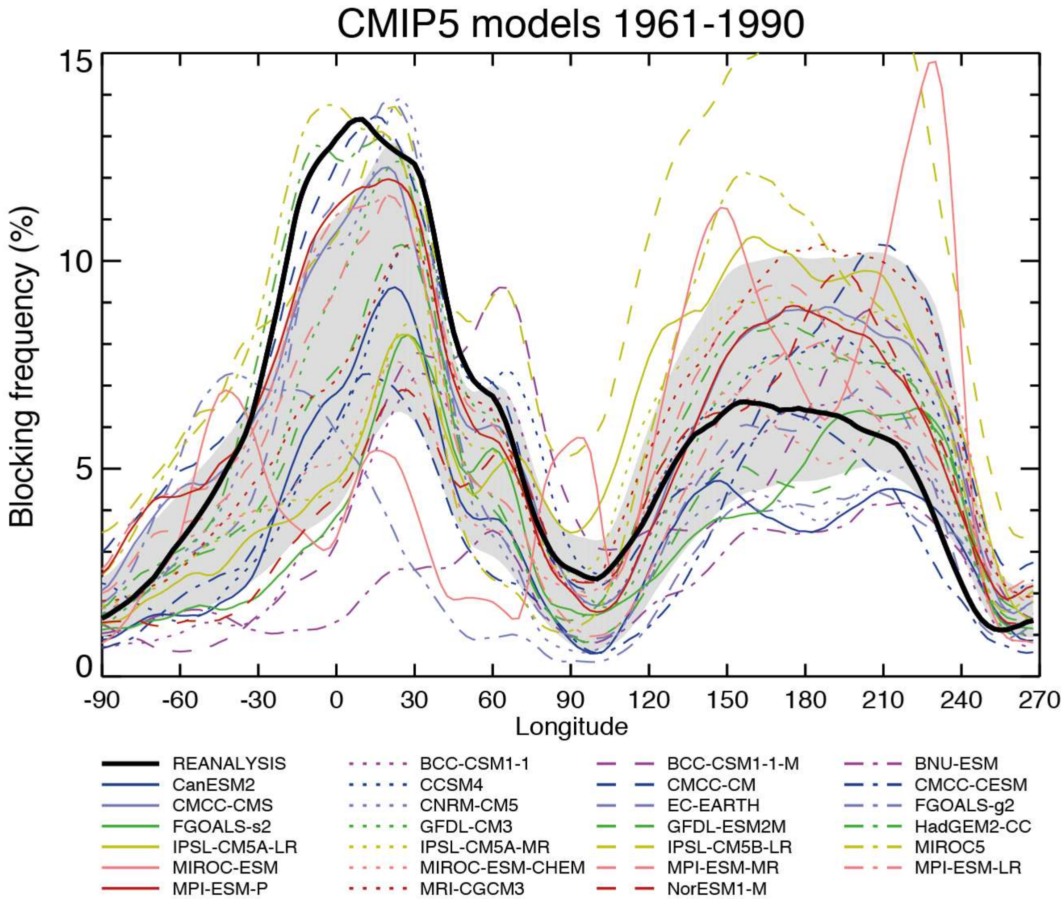

Figure 10: Annual mean blocking frequency in the NH (expressed in % of time, that is, 1% means about 4 days per year) as simulated by a set of CMIP5 models (colour lines) for the 1961–1990 period of one run of the historical simulation. Grey shading shows the mean model result plus/minus one standard deviation. Black thick line indicates the observed blocking frequency derived from the National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalysis. Only CMIP5 models with available 500 hPa geopotential height daily data at http://pcmdi3.llnl.gov/esgcet/home.htm have been used. Blocking is defined as in Barriopedro et al. (2006), which uses a modified version of the (Tibaldi and Molteni, 1990) index. Daily data was interpolated to a common regular 2.5° × 2.5° longitude–latitude grid before detecting blocking. Figure reproduced from Christensen et al. 2015.

4.2.6. Atlantic Meridional Overturning Circulation (AMOC)

The AMOC plays an important role in transporting heat northwards in the Atlantic Ocean, exerting significant influence on the regional mean temperature, but also influence dynamical aspects of the regional climate e.g. via its influence on the storm track (e.g. Brayshaw et al., 2009).

Global coupled models vary in their representation of the AMOC in terms of both the magnitude of the circulation and its variability. Time mean streamflow functions shown for some CMIP5 models, indicate significant diversity in both magnitude and structure across the ensemble (Menary and Wood, *In Press*) (Figure 11). Some models exhibit a stronger streamflow than others –those models with stronger flow tend also to be those in which the flow also extends to deeper levels (e.g. NorESM1-M, FGOALS-g2). Available observations of the AMOC for evaluation of these features are limited, but are available for a short recent period of 2004-2015 at a 26.5°N transect via the RAPID project (Cunningham *et al*., 2009). These are shown as timeseries' from the models and this observational dataset in Figure 12 indicating that, aside from the underestimation of the interannual variability by almost all models, the magnitude of the AMOC is also significantly too strong in the Norwegian models (NorESM1-M), and FGOALS-g2 (orange). Apart from these outliers, while the models exert a wide range of AMOC magnitudes, they are broadly in line with the range of values during the short observation period recorded via the RAPID project. One further characteristic is the 'noisy' nature of the streamflow in the inmcm4 model, which although difficult to verify due to a shortage of observations, is unlikely to be realistic (Matt Menary, *pers. comm*.).

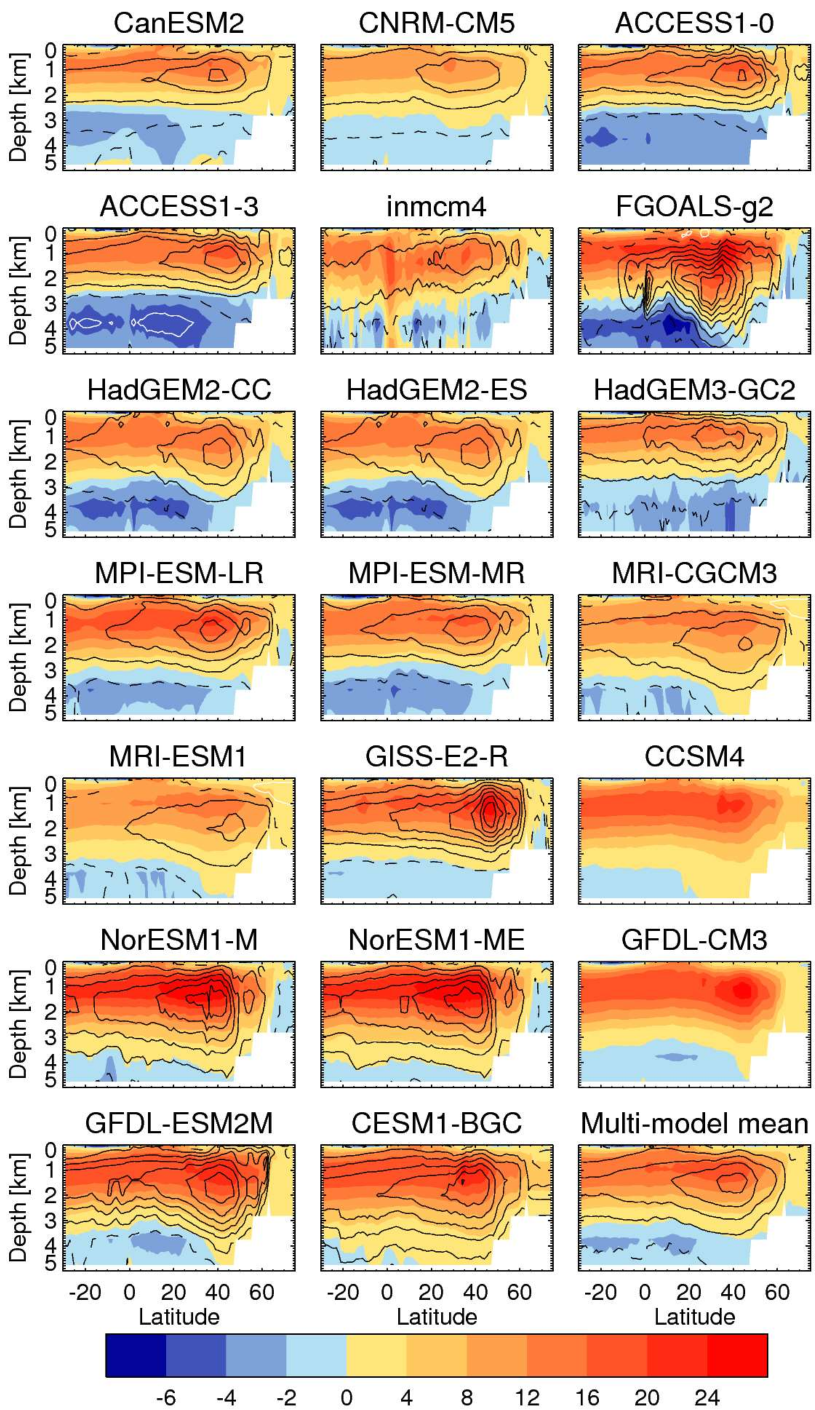

Figure 11: Time Mean Atlantic Overturning stream functions for models for which data were available from the CMIP5 archive (Menary and Wood, 2018). Black contours show the projected change under warming scenarios (not referred to in this report).

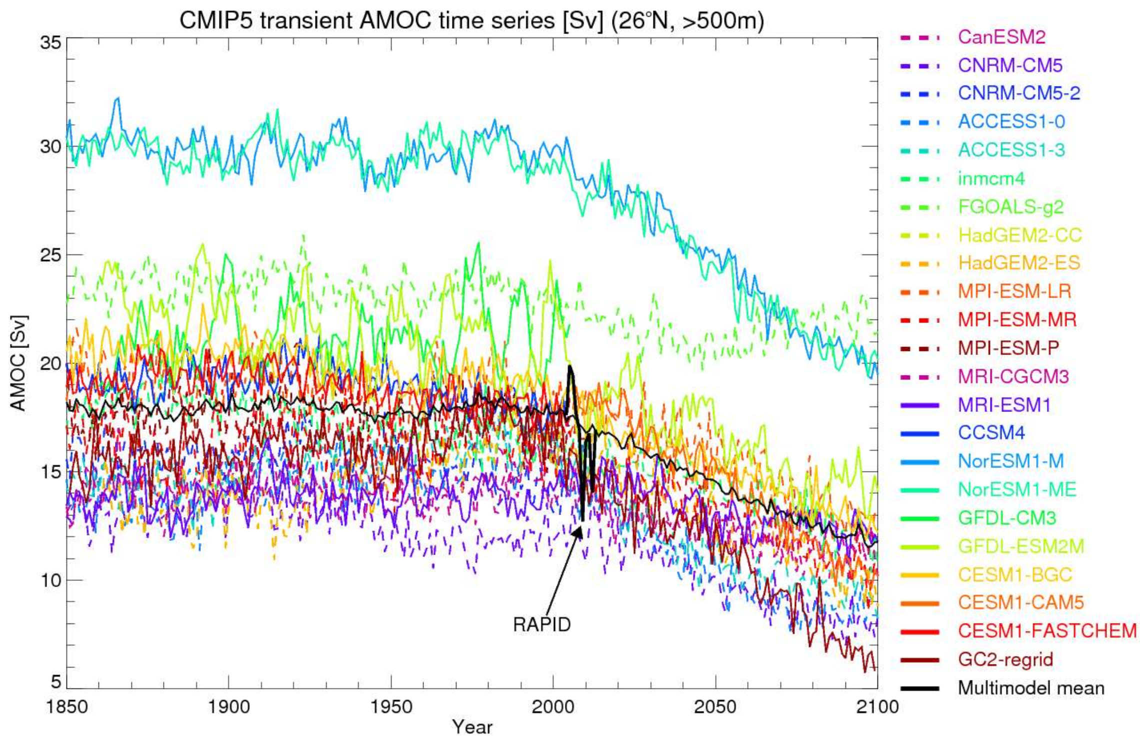

Figure 12: Timeseries of AMOC as volume transport at 26°Nand below 500m in CMIP5 simulations in both historical and projection simulations (NB only the period up until present day is considered here). Observations from the RAPID project dataset between 2004 and 2015.

4.3. Performance Summary

The qualitative performance flags red/orange/yellow allocated for regional and global performance criteria are summarised in Table 1.

Models which are clear cases for elimination based on this assessment are FGOALS-g2 and IPSL-CM5B-LR based on significant misrepresentation of European large-scale behaviours which drive key aspects of the regional weather and climate. For IPSL-CM5B-LR, the very large cool bias of around 8-9° affecting the European region means that it is hard to justify using projections from this model in climate impacts studies – while smaller biases might justifiably be mitigated for the purpose of applying these projections to impacts assessments, it would clearly not be appropriate to attempt to treat such a large bias in this way.

FGOALS-g2 and MIROC5 exhibit very unrealistic structure in the climatological circulation, and in the case of FGOALS-g2, the distribution of storm track positions. BNU-ESM exhibits almost no blocking events affecting the UK region. The unrealistic representation of these key features of these key drivers of regional weather and climate indicate fundamental issues in the representation of the UK climate, which again make it difficult to justify using these projections to diagnose or explore plausible future climate changes for the region.

For several other models, indications across multiple evaluation criteria indicate they exhibit significant shortcoming across several of the assessed criteria. MIROC-ESM, MIROC-ESM-

CHEM, CSIRO-mk3-6-0, inmcm4, and bcc-csm-1-m have all been flagged with an ‘orange’ tag under at least two criteria in Table 1. While it may not be clear what impact these significant errors individually might have on the plausibility of the projections for the UK, overall confidence in the credibility of these projections is reduced as a result of these multiple significant shortcomings. These 6 models are therefore rejected for inclusion in the CMIP5 subset. This leaves 22 models for further consideration in the following section.

Table 1: Summary of qualitative performance scores for CMIP5 models

	Regional Criteria								Global Criteria		
	NE Atlantic Weather Types (Perez et al, 2014)	Annual mean blocking frequency : IPCC	Mean DJF circulation	Mean JJA circulation	Mean temp bias - regional	Atlantic SST bias	AMOC (Menary and Wood, In press	Storm Track (McSweeney et al, 2015a)	Global temperature variability and/or drift in piControl (Sutton et al, 2015)	Remote SST biases – southern ocean and Tropical pacific	High latitude temperature variability (Jones et al, 2013)
bcc-csm1-1											
bcc-csm1-1-m											
BNU-ESM											
CanESM2											
CESM1-BGC											
CMCC-CESM											
CMCC-CM											
CMCC-CMS											
CNRM-CM5											
ACCESS1-0											
ACCESS1-3											
CSIRO-Mk3-6-0										(CT)	
EC-EARTH											
inmcm4										(CT &SO)	
IPSL-CM5A-LR											
IPSL-CM5A-MR											
IPSL-CM5B-LR											
FGOALS-g2											
MIROC5											
MIROC-ESM											
MIROC-ESM-CHEM											
HadGEM2-CC											
HadGEM2-ES											
MPI-ESM-LR											
MPI-ESM-MR											
MRI-CGCM3											
CCSM4											
Nor-ESM1-M											
GFDL-CM3											
GFDL-ESM2G											
GFDL-ESM2M											
CESM1-CAM5											

5. Screening for near-neighbours

Some pairs /groups of models within the CMIP5 ensemble are known to share significant parts of their code, and therefore also share error and projection characteristics (Sanderson et al., 2015a; Sanderson et al ., 2015b; Knutti et al., 2013). Sanderson et al. (2015a and 2015b) use a multivariate metric of present day climatology to quantify similarity between pairs of models from the CMIP5 (and CMIP3) ensemble (see Figure 13), showing clear relationships between some groups of models. These groups of 'near neighbours' include sets of models from particular centres, but also highlighting some groups of models that come from different centres but still share significant portions of code, which may not be so easy to identify without this type of analysis. Sanderson *et al* 2015b extends this analysis to future projections, demonstrating that the degree of dependence between models in their present-day climatology applies similarly to projection characteristics.

In order to minimise double-counting of structural modelling assumptions, we therefore sample a subset of models from the remaining 22 by excluding some of those 'near neighbours'. Groups of models with calculated distances in the order of <0.6 are shown in Table 1. The three CMCC models (CMCC-CESM, CMCC-CM and CMCC-CMS) are not shown in Figure 13, but appear in an equivalent figure in Sanderson et al, 2015a. EC–Earth similarly does not appear in either Sanderson analysis, thus in the absence of this information, is treated as an independent ensemble member.

We can now use the additional performance information from Table 1 to select preferred members from each group, i.e. those which have fewest poor performance flags or those which have been included in most studies. We identify 2 members from groups of 4 or more, and one from each other group (shown in green in Table 2). This process results in a 13 member subset: ACCESS1-3, bcc-csm1-1, CCSM4, CESM1-BGC, CanESM2, CMCC-CM, CNRM-CM5, EC-Earth, GFDL-ESM2G, HadGEM2-ES, IPSL-CM5A-MR, MPI-ESM-MR, MRI-CGCM3.

Figure 14 gives an indication of what impact the 2-stage sub-selection has on the range of projected changes in mean temperature and rainfall indicated by the ensemble. The sub-selection process does narrow the range of projections of both of these quantities somewhat, and this is mainly due to the exclusion of poor performers which have projections at the extremes of the range. The large increases in summer precipitation are excluded due to the poor performance of MIROC-ESM and MIROC-ESM-CHEM, and similarly, the coolest end of the range of temperature projections is excluded with inmcm4. The reduction in range due to the filtering of near-neighbours has a smaller impact on the range of projections. While we have filtered for near-neighbours based on existing quantitative information about near neighbours, other similar studies have explicitly sampled models to span a range of projections for specific variables, seasons and regions (e.g. McSweeney et al, 2015a, 2015b; Whetton et al., 2012). Those sampling approaches may maximise the spread captured for the very specific regions, seasons and variables on which the selection is based but here we have used an approach that draws on information about the relationships between models across a wider range of variables, regions and seasons and may reasonably be expected to be less sensitive to the specific choice of region, variable or season.

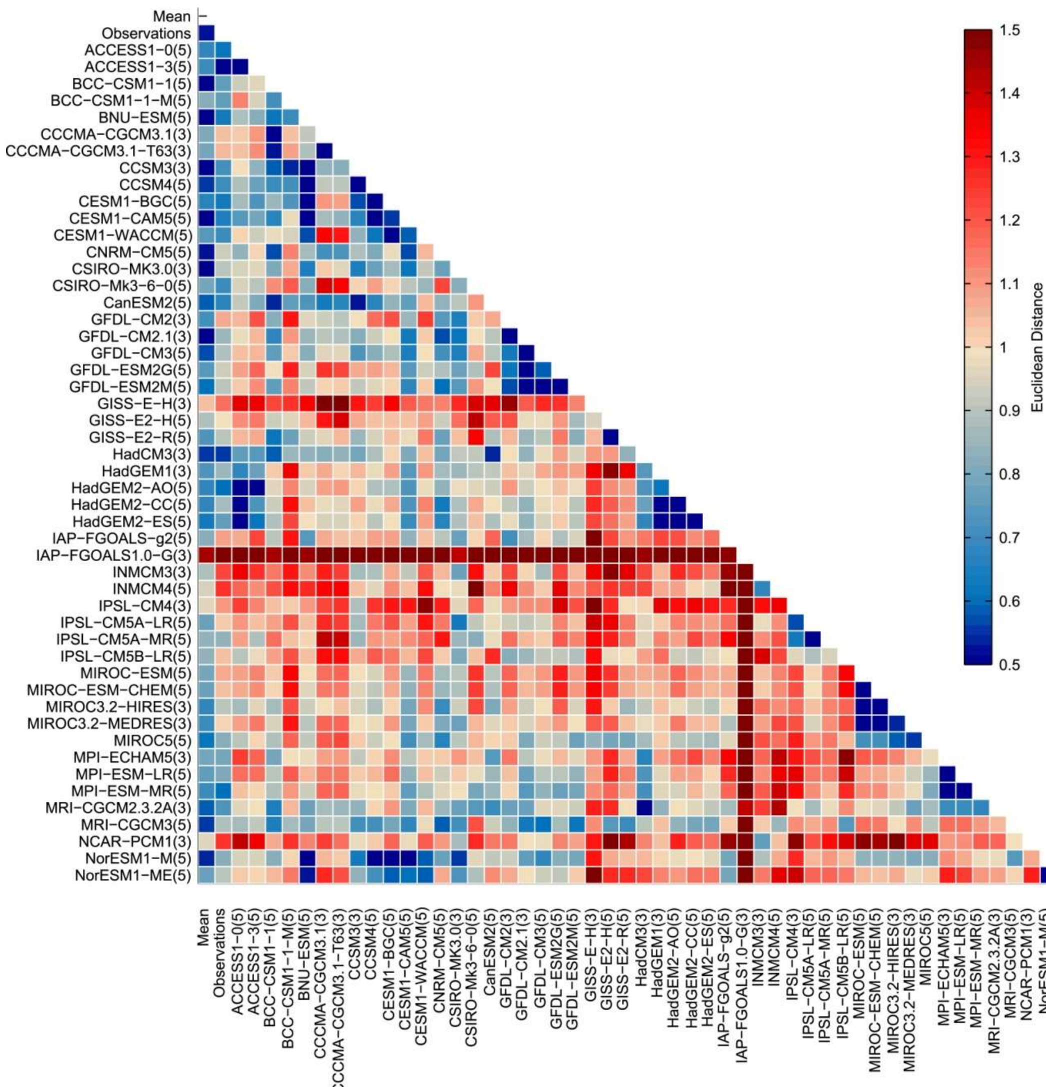

Figure 13: A graphical representation of the intermodal distance matrix for CMIP3, CMIP5, the multi-model mean, and a set of observed values. Each row and column represents a single climate model (or observation). Each box represents a pairwise combination, where warm colours indicate a greater distance. Distances are measured as a fraction of the mean model bias in the combined CMIP3 and CMIP5 ensembles (taken from Sanderson et al., 2015b)

Table 2: CMIP5 models clustered according to near-neighbours identified from Sanderson et al 2015a&b. EC-EARTH and CNRM-CM5 are grouped based on a shared atmosphere model (Knutti et al., 2013). Models in red had already been excluded based on evaluation against key performance criteria outlined in Section 4.3. Green models are the preferred remaining models for each cluster.

bcc-csm1-1 bcc-csm1-1-m	BNU-ESM CCSM4 CESM1-BGC CESM1-CAM5 Nor-ESM1-M	CanESM2	CMCC-CESM CMCC-CM CMCC-CMS	CNRM-CM5
IPSL-CM5A-LR IPSL-CM5A-MR IPSL-CM5B-LR	MIROC5 MIROC-ESM MIROC-ESM-CHEM	FGOALS-g2	MPI-ESM-LR MPI-ESM-MR	MRI-CGCM3
EC-EARTH	inmcm4	CSIRO-Mk3-6-0	GFDL-CM3 GFDL-ESM2G GFDL-ESM2M	ACCESS1-0 ACCESS1-3 HadGEM2-CC HadGEM2-ES

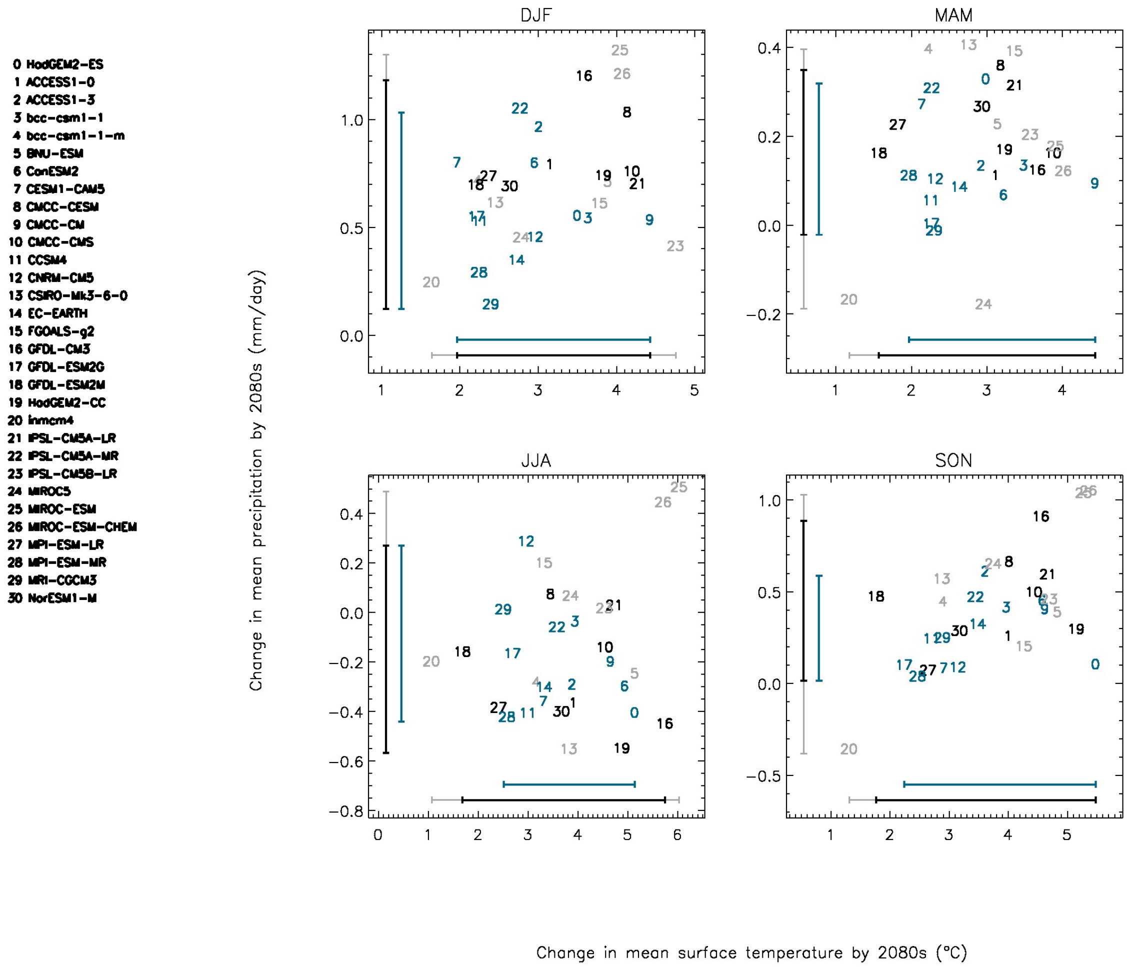

Figure 14: Regional mean changes in mean temperature and precipitation by 2070-2100 relative to 1961-90 under RCP8.5 in CMIP5 GCMs. Models included in the selected subset shown in blue, models which were rejected on the grounds of performance are shown in grey.

6. Summary

In this exercise, a set of 13 members of the CMIP5-MME has been identified that will be used to augment a 15-member PPE to provide a set of plausible and diverse projections for UK climate impacts assessments. The 13 member subset has been selected using a 2 stage approach in which the models were initially screened qualitatively for global and regional performance to exclude poorly performing models, and then to reduce the size of the remaining ensemble by removing some models which are near-neighbours. The regional criteria that the models were assessed against included the climatological circulation patterns of the N. Atlantic/European sector, distribution of daily storm track latitudes, mean temperature biases, frequencies of daily weather types, blocking frequency and the realism of the AMOC. Global criteria included global mean climate variability and drift, and SST errors.

Three models were found to be too unrealistic in their representation of one or more key characteristics of the regional climate to provide useful projections for the UK. These were IPSL-CM5B-LR (UK affected by a cool bias of 8-9°C and very unrealistic summer circulation), FGOALS-g2 (very unrealistic circulation patterns in both summer and winter) and MIROC5 (unrealistic summer circulation, lacking westerly flow direction over UK). Six further models were found to perform very poorly against several of the regional and/or global criteria. From the remaining 22 models, 12 were selected by sampling from groups of 'near-neighbours', in order to reduce the reliance on shared model components in the final chosen subset. The 13 members selected capture much of the full CMIP5 range of seasonal mean changes in temperature and precipitation, but the exclusion of poorest models does impact the range of regional projections for the UK by discounting the member with lowest warming in the region (INMCM4) and those with the largest temperature and precipitation increases (MIROC-ESM and MIROC-ESM-CHEM).

Ideally, the criteria for including MME members would be equivalent to those applied to members in the design of the PPE. However, in practice, differences in the experimental design, structural differences between the two ensembles and differences in availability of performance information mean that it is not practical to apply the same absolute thresholds for inclusion to each ensemble. While some qualitative criteria applied to both ensembles (e.g. the European seasonal mean circulation patterns in Section 4.2.1), similar thresholds for acceptable/unacceptable performance could be applied. However, in other cases, such as for global SSTs, this was not practical because the RMS errors were systematically larger in the CMIP5-MME compared with the coupled PPE members, due to use of flux adjustments in the latter. A model sub-selection activity is inevitably a balance between eliminating some models in which we have reduced confidence and while retaining enough models to provide projections which are both plausible *and* diverse. In order to achieve this balance, the assessment of performance of members within a given climate model ensemble (as carried out here) inevitably becomes an assessment of relative performance. We note that UKCP18 users will use our CMIP5 subset in combination with PPE members to evaluate impacts and risks, so it will be important to extend the assessment in subsequent work using a common set of metrics applicable to members of both ensembles once the PPE simulations have completed. In these evaluation activities, the CMIP5 subset will also act as a 'benchmark' for gauging the performance of the HadGEM3-PPE. The CMIP5 models are currently considered to represent the 'state-of-the art' in climate modelling, but are also known to vary considerably in their representation of regional climate processes.

We therefore consider that a reasonable benchmark for acceptable performance in the PPE to be that which is as good as these ‘good’ CMIP5 members or better across a range of evaluation metrics that will appear in UKCP18 documentation (e.g. Murphy et al, 2018).

Selecting a subset of models based on performance criteria is challenging for a number of reasons. One of these is that the sub-selection process requires subjective judgements at a number of stages, in (a) the choice of evaluation criteria (b) allocating flags to particular models and (c) deciding on inclusion/exclusion of models based on the flags across multiple assessment criteria. However, while differences in these subjective choices could influence the outcome, it is clear that several models clearly exhibit shortcomings across multiple criteria, which suggests some robustness in the judgements that these are weaker models we should have less confidence in. The criteria used here were chosen specifically to identify models sufficiently realistic to be useful in impacts studies for the UK – other sub-selection activities may make different subjective decisions where the purpose differs. The selected 13 CMIP5 models will be augment the 15-member HadGEM3-PPE, resulting in a 28-member ensemble of global realisations in UKCP18 Land Strand 2 which can be used to explore potential impacts of these future climate scenarios. A more detailed account of the design of the HadGEM3-PPE, and evaluation of the historical performance and future projections in the resultant 28-member ensemble can be found in the UKCP18 Science Report (Murphy *et al*, 2018).

7. References

Brayshaw, D. J., T. Woollings, and M. Vellinga, 2009: Tropical and Extratropical Responses of the North Atlantic Atmospheric Circulation to a Sustained Weakening of the MOC. Journal of Climate, 22, 3146-3155.

Christensen, J.H., K. Krishna Kumar, E. Aldrian, S.-I. An, I.F.A. Cavalcanti, M. de Castro, W. Dong, P. Goswami, A. Hall, J.K. Kanyanga, A. Kitoh, J. Kossin, N.-C. Lau, J. Renwick, D.B. Stephenson, S.-P. Xie and T. Zhou, 2013: Climate Phenomena and their Relevance for Future Regional Climate Change. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA

Cunningham, S. A., and Coauthors, 2007: Temporal variability of the Atlantic meridional overturning circulation at 26.5 degrees N. Science, 317, 935-938.

Hofstra, N, Haylock, M, New, M and Jones, PD (2009) Testing E-OBS European high-resolution gridded data set of daily precipitation and surface temperature. Journal of Geophysical Research D: Atmospheres, 114 (21).

Jones, G. S., P. A. Stott, and N. Christidis, 2013: Attribution of observed historical near-surface temperature variations to anthropogenic and natural causes using CMIP5 simulations. Journal of Geophysical Research-Atmospheres, 118, 4001-4024.

Knutti, R., D. Masson, and A. Gettelman, 2013: Climate model genealogy: Generation CMIP5 and how we got there. Geophysical Research Letters, 40, 1194-1199.

McSweeney, CF, Jones, RG, Lee, RW and Rowell, DP (2015a) *Sub-selecting CMIP5 GCMs for downscaling over multiple regions*. Climate Dynamics, DOI 10.1007/s00382-014-2418-8

McSweeney, C, R. Rahmat, G. Redmond, C.Marzin, J. Murphy., R. Jones., W.-K. Cheong, S.-Y. Lim, X. Sun. (2015b) *Chapter 3: Sub-Selection of CMIP5 GCMs for Downscaling over Singapore*. In: Singapore 2nd National Climate Change Study – Phase 1. Available at: http://ccrs.weather.gov.sg/publications-second-national-climate-change-study-science-reports/

Meijers, A. J. S., 2014: The Southern Ocean in the Coupled Model Intercomparison Project phase 5. Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences, 372.

Menary M. and Wood R (2018) An anatomy of the projected [1]North Atlantic warming hole in CMIP5 models. Climate Dynamics, Volume 50, Issue 7–8, pp 3063–3080

Murphy JM, Harris GR, Sexton DMH, Kendon EJ, Bett PE, Brown SJ, Clark RT, Eagle K, Fosser G, Fung F, Lowe JA, McDonald RE, McInnes RN, McSweeney CF, Mitchell JFB, Rostron J, Thornton HE, Tucker S and Yamazaki K (2018) UKCP18 Land Projections: Science Report.

Perez, J., M. Menendez, F. J. Mendez, and I. J. Losada, 2014: Evaluating the performance of CMIP3 and CMIP5 global climate models over the north-east Atlantic region. Climate Dynamics, 43, 2663-2680.

Sanderson, B. M., R. Knutti, and P. Caldwell, 2015a: A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble. Journal of Climate, 28, 5171-5194.

——, 2015b: Addressing Interdependency in a Multimodel Ensemble by Interpolation of Model Properties. Journal of Climate, 28, 5150-5170.

Sutton, R., E. Suckling, and E. Hawkins, 2015: What does global mean temperature tell us about local climate? Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences, 373.

Whetton, P., K. Hennessy, J. Clarke, K. McInnes, and D. Kent, 2012: Use of Representative Climate Futures in impact and adaptation assessment. Climatic Change, 115, 433-442.

8. Acknowledgements

This work was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101).

We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modelling groups (listed in Table XX of the appendix) for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. We are also grateful to Jamie Kettleborough, Ian Edmond and Emma Hibling at the Met Office Hadley Centre for developing tools which have allowed us to download and access this data easily.

Thanks to Robert Lee (University of Reading) for providing the North Atlantic Storm Track densities shown in Figure 6. Thanks to Matt Menary (Met office) for providing Figure 12.

9. Appendix

Table A1: Models and modelling groups

Modelling Group	Group Acronym	Model Designation
Commonwealth Scientific and Industrial Research Organization (CSIRO) and Bureau of Meteorology (BOM), Australia	CSIRO-BOM	ACCESS1-0
Commonwealth Scientific and Industrial Research Organization (CSIRO) and Bureau of Meteorology (BOM), Australia	CSIRO-BOM	ACCESS1-3
Beijing Climate Center, China Meteorological Administration	BCC	bcc-csm1-1
Beijing Climate Center, China Meteorological Administration	BCC	bcc-csm1-1-m
College of Global Change and Earth System Science, Beijing Normal University	GCESS	BNU-ESM
Canadian Centre for Climate Modelling and Analysis	CCCMA	CanESM2
National Center for Atmospheric Research	NCAR	CCSM4
Community Earth System Model Contributors	NSF-DOE-NCAR	CESM1-CAM5
Centro Euro-Mediterraneo per I Cambiamenti Climatici	CMCC	CMCC-CESM
Centro Euro-Mediterraneo per I Cambiamenti Climatici	CMCC	CMCC-CM
Centro Euro-Mediterraneo per I Cambiamenti Climatici	CMCC	CMCC-CMS
Centre National de Recherches Météorologiques / Centre Européen de Recherche et Formation Avancée en Calcul Scientifique	CNRM-CERFACS	CNRM-CM5
Commonwealth Scientific and Industrial Research Organization in collaboration with Queensland Climate Change Centre of Excellence	CSIRO-QCCCE	CSIRO-Mk3-6-0
EC-EARTH consortium	ICHEM	EC-EARTH
LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences and CESS,Tsinghua University	LASG-CESS	FGOALS-g2
NOAA Geophysical Fluid Dynamics Laboratory	NOAA GFDL	GFDL-CM3
NOAA Geophysical Fluid Dynamics Laboratory	NOAA GFDL	GFDL-ESM2G
NOAA Geophysical Fluid Dynamics Laboratory	NOAA GFDL	GFDL-ESM2M
Met Office Hadley Centre	MOHC	HadGEM2-CC
Met Office Hadley Centre	MOHC	HadGEM2-ES
Institute for Numerical Mathematics	INM	inmcm4
Institut Pierre-Simon Laplace	IPSL	IPSL-CM5A-LR
Institut Pierre-Simon Laplace	IPSL	IPSL-CM5A-MR
Institut Pierre-Simon Laplace	IPSL	IPSL-CM5B-LR
Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies	MIROC	MIROC5
Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies	MIROC	MIROC-ESM
Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies	MIROC	MIROC-ESM-CHEM
Max-Planck-Institut für Meteorologie (Max Planck Institute for Meteorology)	MPI-M	MPI-ESM-LR
Max-Planck-Institut für Meteorologie (Max Planck Institute for Meteorology)	MPI-M	MPI-ESM-MR
Meteorological Research Institute	MRI	MRI-CGCM3
Norwegian Climate Centre	NCC	Nor-ESM1-M

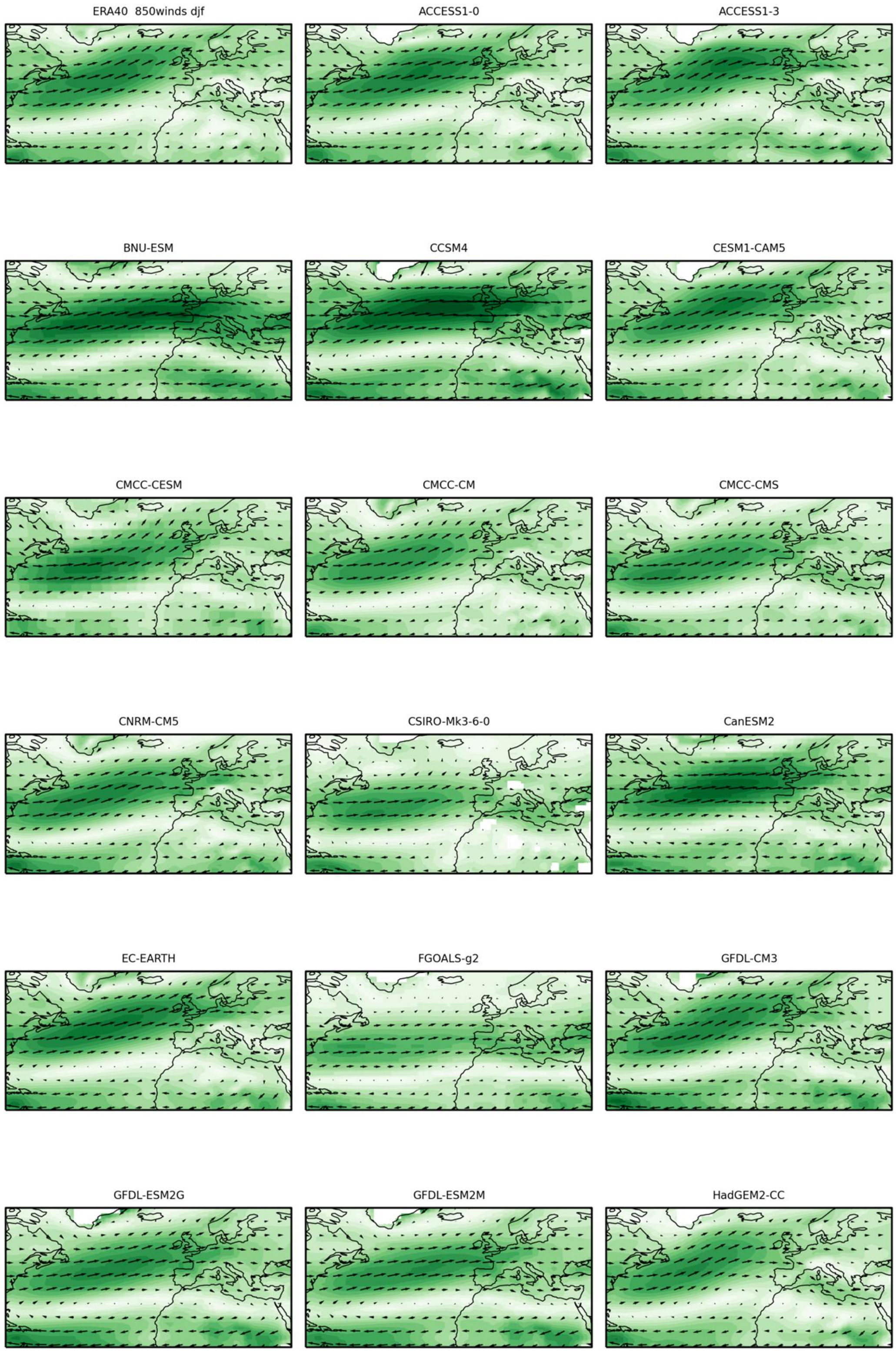

Figure A1(a): 850hpa circulation in December-January-February (DJF) 1981-2000 for ERA40 observations and CMIP5 models.

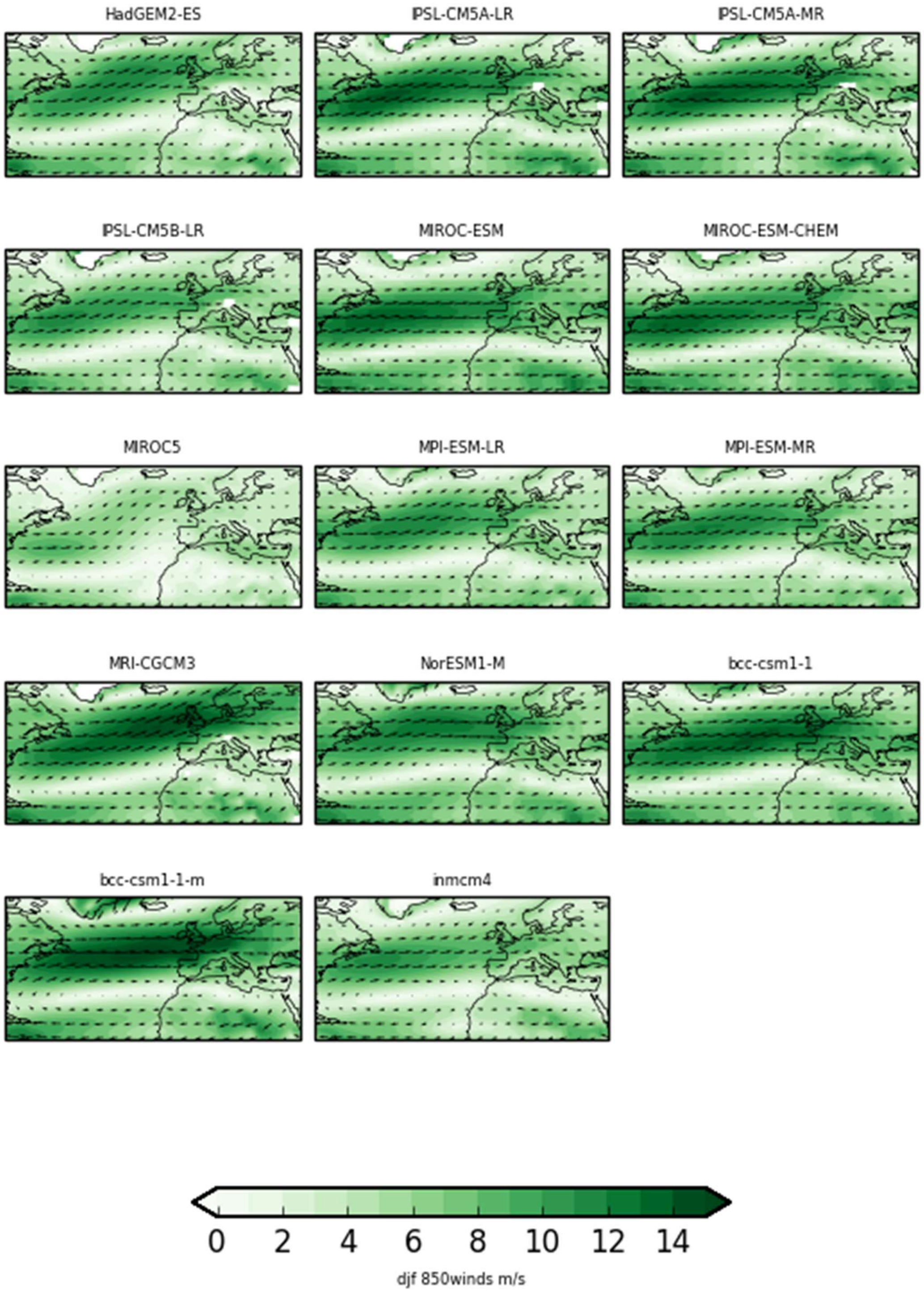

Figure A1(b): 850hpa circulation in December-January-February (DJF) 1981-2000 for ERA40 observations and CMIP5 models.

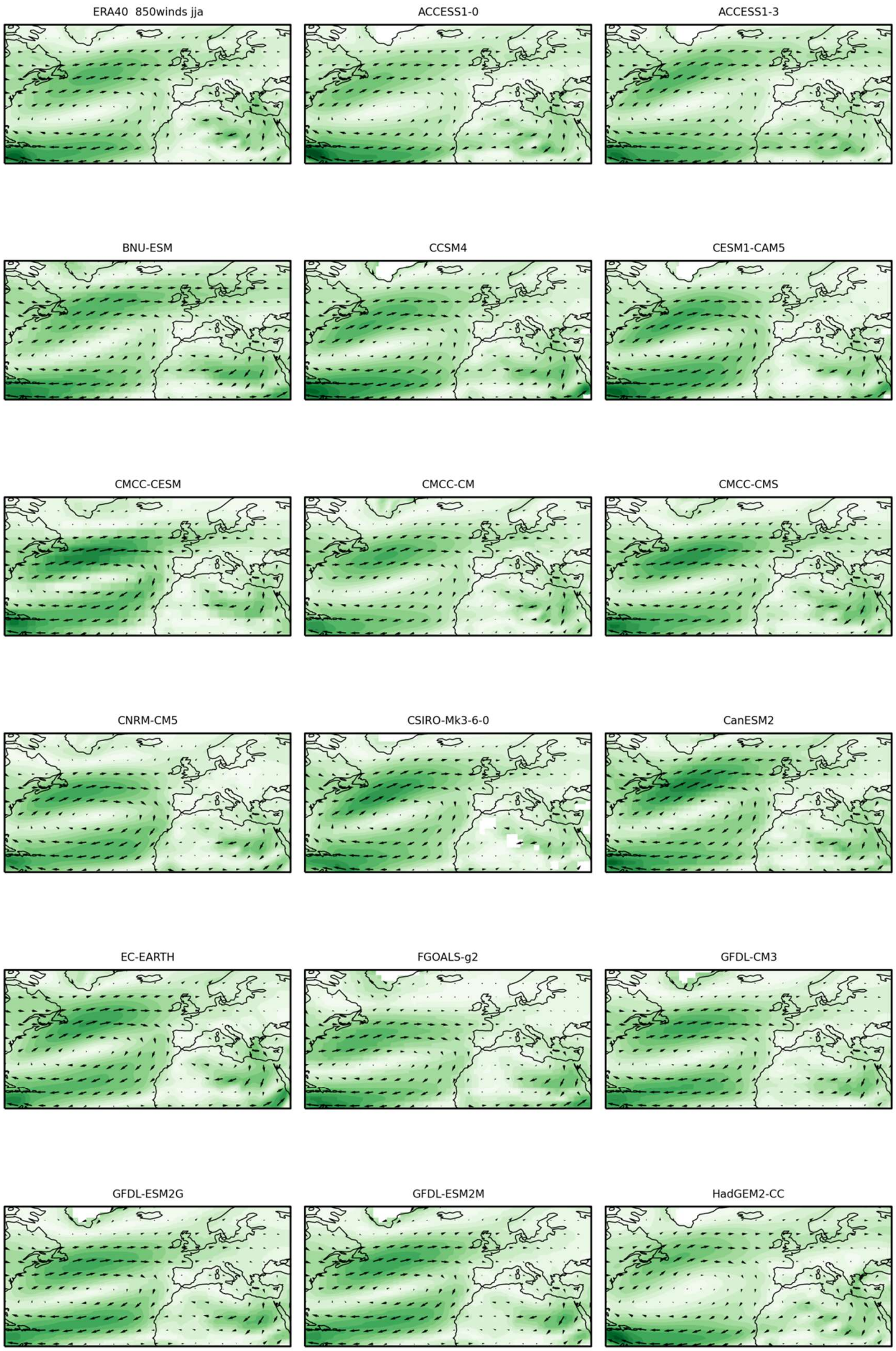

Figure A2(a): 850hpa circulation in December-January-February (DJF) 1981-2000 for ERA40 observations and CMIP5 models.

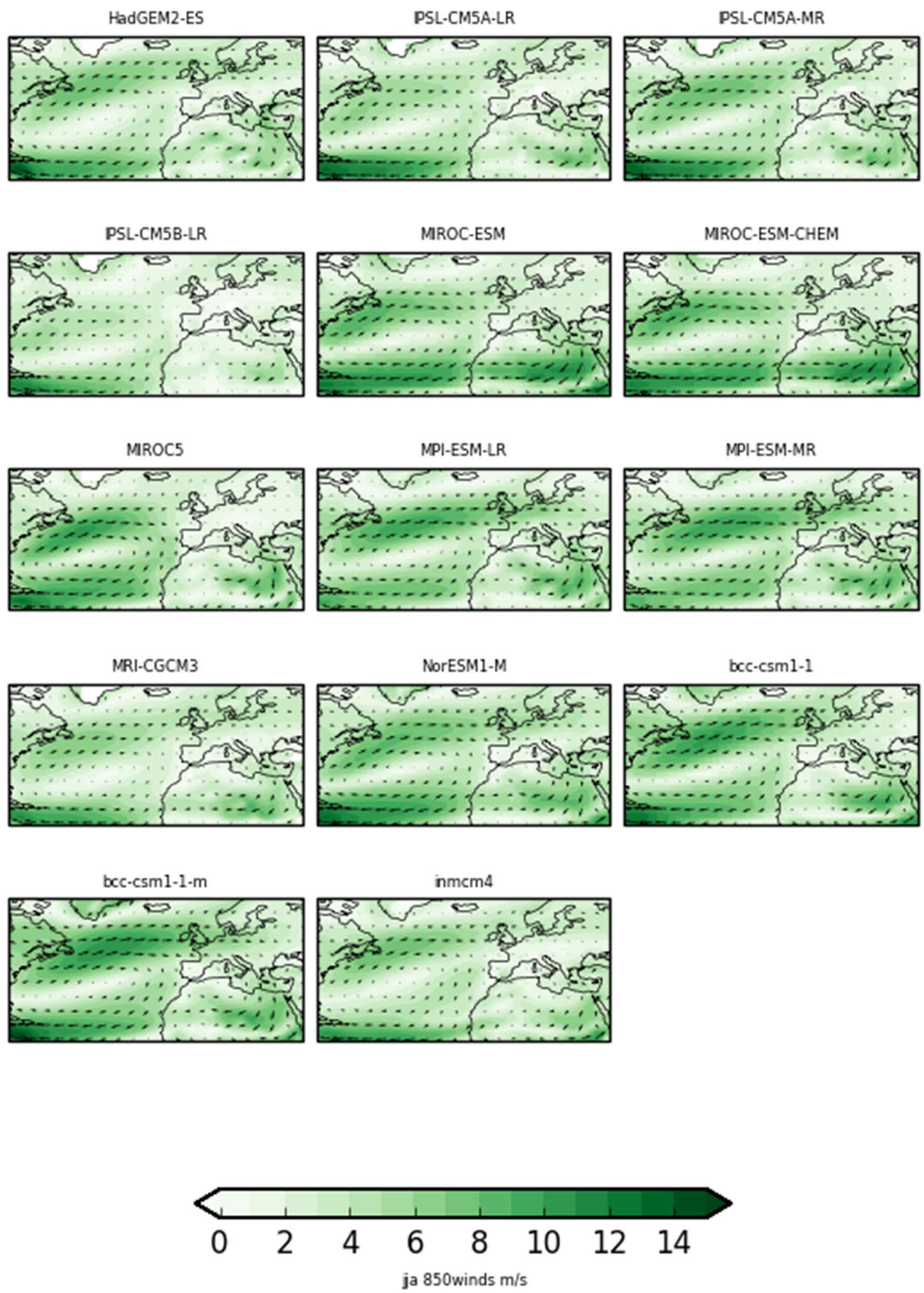

Figure A2(b): 850hpa circulation in December-January-February (DJF) 1981-2000 for ERA40 observations and CMIP5 models.