

Selecting CMIP5 GCMs for downscaling over multiple regions

Article

Published Version

Creative Commons: Attribution 4.0 (CC-BY)

McSweeney, C. F., Jones, R. G., Lee, R. W. ORCID: <https://orcid.org/0000-0002-1946-5559> and Rowell, D. P. (2015) Selecting CMIP5 GCMs for downscaling over multiple regions. *Climate Dynamics*, 44 (11). pp. 3237-3260. ISSN 0930-7575 doi: 10.1007/s00382-014-2418-8 Available at <https://centaur.reading.ac.uk/68059/>

It is advisable to refer to the publisher's version if you intend to cite from the work. See [Guidance on citing](#).

To link to this article DOI: <http://dx.doi.org/10.1007/s00382-014-2418-8>

Publisher: Springer

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the [End User Agreement](#).

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Selecting CMIP5 GCMs for downscaling over multiple regions

C. F. McSweeney · R. G. Jones · R. W. Lee · D. P. Rowell

Received: 12 February 2014 / Accepted: 13 November 2014 / Published online: 16 December 2014
© The Author(s) 2014. This article is published with open access at Springerlink.com

Abstract The unprecedented availability of 6-hourly data from a multi-model GCM ensemble in the CMIP5 data archive presents the new opportunity to dynamically downscale multiple GCMs to develop high-resolution climate projections relevant to detailed assessment of climate vulnerability and climate change impacts. This enables the development of high resolution projections derived from the same set of models that are used to characterise the range of future climate changes at the global and large-scale, and as assessed in the IPCC AR5. However, the technical and human resource required to dynamically-downscale the full CMIP5 ensemble are significant and not necessary if the aim is to develop scenarios covering a representative range of future climate conditions relevant to a climate change risk assessment. This paper illustrates a methodology for selecting from the available CMIP5 models in order to identify a set of 8–10 GCMs for use in regional climate change assessments. The selection focuses on their suitability across multiple regions—Southeast Asia, Europe and Africa. The selection (a) avoids the inclusion of the least realistic models for each region and (b) simultaneously

captures the maximum possible range of changes in surface temperature and precipitation for three continental-scale regions. We find that, of the CMIP5 GCMs with 6-hourly fields available, three simulate the key regional aspects of climate sufficiently poorly that we consider the projections from those models ‘implausible’ (*MIROC-ESM*, *MIROC-ESM-CHEM*, and *IPSL-CM5B-LR*). From the remaining models, we demonstrate a selection methodology which avoids the poorest models by including them in the set only if their exclusion would significantly reduce the range of projections sampled. The result of this process is a set of models suitable for using to generate downscaled climate change information for a consistent multi-regional assessment of climate change impacts and adaptation.

Keywords RCM · Ensemble design · Uncertainty · CMIP5

1 Introduction

Modelling centres participating in the fifth Coupled Model Inter-comparison Project (CMIP5) experiment (Taylor et al. 2012) agreed to make available the 6-hourly instantaneous fields of prognostic variables from GCMs for use as lateral boundary conditions (LBCs) for driving regional climate models (RCMs). This provides the opportunity for those interested in higher-resolution baseline and future climates derived by downscaling with multiple combinations of global and regional climate models or statistical downscaling techniques, allowing exploration of a wide range of high-resolution projections for one or more regions of the world consistent with the latest GCM-based climate projections assessed in the Intergovernmental Panel on Climate Change (IPCC) Assessment Report 5 (AR5). However,

Electronic supplementary material The online version of this article (doi:10.1007/s00382-014-2418-8) contains supplementary material, which is available to authorized users.

C. F. McSweeney (✉) · R. G. Jones · D. P. Rowell
Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK
e-mail: carol.mcsweeney@metoffice.gov.uk

R. G. Jones
Oxford University School of Geography and Environment, Dyson Perrins Building, South Parks Road, Oxford OX1 3QY, UK

R. W. Lee
Department of Meteorology, University of Reading, Earley Gate,
PO box 243, Reading RG6 6BB, UK

the human and computational resource-intensive nature of high-resolution downscaling places a restriction on the size of ensembles generated and downscaling the full ensemble may not be desirable or necessary to generate a representative range of future climate conditions relevant to assessing risks associated with future climate change. This implies the need to develop strategies to sample from the available General Circulation Models (GCMs) and Representative Concentration Pathway (RCP) scenarios in order to generate projections that are policy relevant and manageable to develop, analyse and disseminate.

This paper explores the selection of GCMs from the CMIP5 ensemble in order to identify a subset that is representative of the range of future climate outcomes indicated by the full ensemble. Such an approach could be adopted by large model intercomparison projects such as the Coordinated Regional Downscaling Intercomparison Project (CORDEX) (Giorgi et al. 2009).

The selection process also provides the challenge and opportunity to discount any models which we find unsatisfactory in their representation of key processes or features of climate. The down-weighting or exclusion of GCMs has been explored in a number of studies (e.g. Tebaldi et al. 2005; Greene et al. 2006; Tebaldi and Sanso 2009; Watterson and Whetton 2011; Sexton et al. 2012). However, this is a challenging problem with respect to both the practicalities of identifying unsatisfactory models, and the more philosophical considerations of how we relate apparently poor performance to the plausibility of future projections (see Knutti 2010; Knutti et al. 2010 for discussion). Critically, the elimination of some GCMs may narrow the range of uncertainty represented by the remaining models (e.g. Overland et al. 2011). While this is often considered desirable given the policy challenges in responding to projections with large uncertainty ranges, provision of a falsely narrow range of projections may lead to over confidence and mal-adaptation. The IPCC is one key institution that has avoided attempting to weight or eliminate individual models, adopting a ‘one-model-one-vote’ interpretation of CMIP3 and CMIP5 projections in its fourth and fifth assessment reports, respectively (IPCC 2007, 2013).

Previous examples of weighting or selection on the basis of realism have argued that it may be more justifiable for specific regions or applications where the key aspects of model behaviour can be identified and assessed (e.g. Overland et al. 2011; McSweeney et al. 2012). We explore whether it is either practical or justifiable to extend a regional approach in order to identify a subset of models which are suitable for use across multiple regions. Such an approach would yield practical benefits by reducing the overhead of interfacing between GCMs and RCMs for modelling groups involved in providing downscaled information for more than one region, as well as reducing the

need for multiple selection studies. Further, projects with a global scope such as ISIMIP (Intersectoral-Impacts model inter-comparison project) (Warszawski et al. 2013) have expressed interests in the use of a consistent set of GCMs for downscaling in order to generate datasets with consistently generated uncertainty ranges globally.

We present an approach to selection that considers the large-scale performance of the models in the regions of interest, with a view to excluding those that are considered very unrealistic, while also considering the effect on the spread of models in the final subset of eliminating models that perform poorly. We demonstrate the application of this approach to the selection of 8–10 CMIP5 models that could be used in generating climate information for three climatologically diverse continental-scale regions of the world: Southeast Asia, Europe and Africa.

Section 2 describes the rationale underlying the proposed methodology. Section 3 describes the CMIP5 model data used throughout. In Sect. 4 we explain the evaluation criteria used, assess for each of the three regions how well each of the GCMs generates key features of the large-scale climate, before applying these results to make a multi-regional decision on elimination in Sect. 5. Section 6 explores the subsequent selection of an ‘optimal’ subset of 8–10 of the remaining models in order to span most fully the range of future outcomes. In Sect. 7 we discuss the benefits and limitations of the approach proposed.

2 Selection rationale

The default ‘one-model-one vote’ approach to interpreting ensemble projections can be considered to be precautionary—that is, generally speaking, we cannot confidently link the observed shortcomings in the realism of baseline simulations directly to the plausibility of that model’s projections. Therefore when considering projections of future change for planning and decision-making purposes, all projections should be considered to have a non-negligible likelihood of occurring. However, in the context of generating higher resolution projections for an increasingly large ensemble of available GCMs, this is no longer a case of ‘should we select?’, but now a question of ‘*how* should we select?’, downscaling all the available projections so that all may be considered is simply not an option for most experiments, except perhaps those undertaken at the largest climate centres or collaboratively. Given that selection is desirable, the equivalent precautionary approach would be to select based on a requirement to span the range of outcomes most effectively, giving no consideration to each model’s relative realism or the plausibility of its projections. Let us consider a hypothetical ensemble in which 1 out of the 20 available ensemble members (‘Model X’)

Table 1 Decision making matrix for potential elimination of ensemble members

Model performance	Model projections	
	Outlier	Other models predict similar outcomes too.
Model suffers sufficient shortcoming that it significantly reduce our confidence in its projections ('Implausible')	<i>Exclude: we should carefully document justification for this, however, as exclusion will affect the range of outcomes</i>	<i>Exclude: We can avoid using these models without much affecting the range of projected outcomes</i>
Model suffers significant shortcomings which we cannot clearly link to confidence in its projections ('Biases/Significant Biases')	Include: we do not have strong enough evidence to exclude these outcomes from the projections	<i>Exclude: We can avoid using these models without much affecting the range of projected outcomes</i>
Model performance is satisfactory ('Satisfactory')	Include	Include

displays a clear and distinct shortcoming compared with the other 19 models, which leads us to have a considerably reduced confidence in the plausibility of the projections from that model. In the situation that we can consider all 20 models (i.e. we do not need to downscale) and we choose to follow the accepted precautionary approach then Model X contributes 1/20 of the results of the ensemble. If we were to select 5 of those models with no reference to any performance metrics, and happened to include Model X, then Model X now contributes 1/5 of the results of the ensemble. The inherent up-weighting of selected models means that we must re-consider the precautionary approach that results from our lack of confidence in discounting projections from models which perform less well in validation.

In reality, making decisions about elimination is difficult and often subjective. An earlier paper by McSweeney et al. (2012) describes a two-stage approach whereby initially all models were assessed to ensure that they realistically reproduced key aspects of the regional climate, and secondly, a subset of all remaining 'plausible' projections was selected to span the range of outcomes in surface variables. However, we suggest that these decisions could be restricted to only a few cases by combining our knowledge about performance with some information about the future projections to identify models which present key decisions. Returning to our hypothetical ensemble of 20 models and the poorly-performing Model X, we can also consider the model's position in the ensemble of projections. If Model X sits well within the range of future projections compared with other models, then we could easily avoid including this model in favour of others which give similar projections, but in which we have more confidence, avoiding the difficult question of whether the projection should be considered implausible. A more significant decision arises if the projections from Model X lie outside the range of the rest of the ensemble; in this case we must make this key decision based on our best knowledge. However, by employing this approach we minimise the burden and impact of this decision-making process.

This approach to the decision-making is summarised as a matrix in Table 1. Here the key decisions occur in allocating a model its position on the y-axis between 'implausible' and 'significantly biased' if it is classed as an 'outlier'. Our criterion in this situation is that if it is clear that a model fails to simulate a large-scale process that is a significant driver of the climate of a region, for example extra-tropical storm tracks or monsoonal circulations, then this model is unlikely to correctly capture how global climate change will manifest itself over the region. It may be unlikely, for example, to represent realistically future changes in the transport of heat or moisture resulting from climate change into or out of the region. Where we find evidence of very significant shortcomings of this nature in a model then we feel it reasonable to class it as 'implausible' and eliminate it. However, only these clearly justified process based assessments are used to eliminate models. Other aspects of performance, such as the realism of surface variables, which may *indicate* shortcomings in key processes, but may also reflect less significant errors such as coarse resolution, are not eliminated outright, but classed as 'biased' or 'significantly biased'.

The classification of model performance into the four categories of 'Implausible', 'Significantly biased', 'Biased' and 'Satisfactory' allows us to assess models against a range of criteria, including both quantitative and qualitative assessments and with reference to results of our own analyses as well as those which appear in the literature. This classification scheme is designed to allow necessarily subjective decisions to be made in a transparent way. We use 4 classifications for model performance in this matrix, allowing for 2 degrees of 'biased' and 'significantly biased' to reflect a range in performance, but also reflecting the relative importance of some aspects over others. We apply these classifications based on the framework described in Table 2 (the criteria relevant to each region are discussed further in Sect. 4). These criteria are based on the underpinning principles for selection proposed in McSweeney et al. (2012) based on guidance in Knutti (2010) as follows:

Table 2 Criteria for performance categories

What might cause a model's performance to be judged as...

Implausible	Significant biases	Biases
Very unrealistic representation of a key large-scale circulation feature e.g. where lower tropospheric flow in a key region is reversed compared with observations	<p>Very unrealistic representation of a surface climate feature (e.g. the annual cycle of rainfall or temperature). The poor representation of surface features may indicate a failure in the large-scale circulation in the model, but might also be a result of the restricted resolution or parameterisation errors in the GCM and therefore could be improved by downscaling</p> <p>Very poor representation of a key circulation feature, but where other models do not offer much improvement</p>	<p>Substantial systematic biases—i.e. where the position and timing of features are realistic but magnitude is inaccurate</p> <p>Where particularly poor performance is indicated in external literature but the assessment does not include all models that we are selecting from</p> <p>The model performs relatively less well than almost all other models in the ensemble</p> <p>Some necessary variables/fields that we have assessed for other models are missing from the CMIP5 archive</p> <p>Key aspects cannot be assessed due to missing data</p>

1. Metrics and criteria for evaluation should be demonstrated to relate to projection.
2. It may be less controversial to downweight or eliminate specific projections that are clearly unable to mimic important processes than to agree on the best model.
3. Process understanding must complement 'broad brush metrics'.

After the decision making framework is applied, and models eliminated, a further selection process is required in order to identify statistically the subset of n models which best span the range of the remaining models. Figure 1 shows how the 3 stages of the selection process that we describe compare with the simpler 2-stage process of McSweeney et al. (2012).

3 CMIP5 model data

The coupled models analysed are listed in Table S1, where the models for which 6-hourly atmospheric fields required as lateral boundary conditions (LBCs) for dynamical downscaling are available are highlighted (29 of 43 analysed, from hereon referred to as 'LBC-Avail'). The experimental design of the CMIP5 experiment is described in Taylor et al. (2012).

Our model validation assesses the historical simulations of the period 1961–1990, while future changes are based on the period 2070–2100 relative to 1961–1990 in the RCP8.5 experiment in order to use projections with the greatest potential signal compared to internal variability. In all cases we use a single realisation. All fields are re-gridded to a common $2.5^\circ \times 3.75^\circ$ grid for all analyses except those of the storm tracks assessment in section referred to in Sect. 4.2 (see Hodges et al. 2011 for a full description of the storm tracking methodology employed) and the

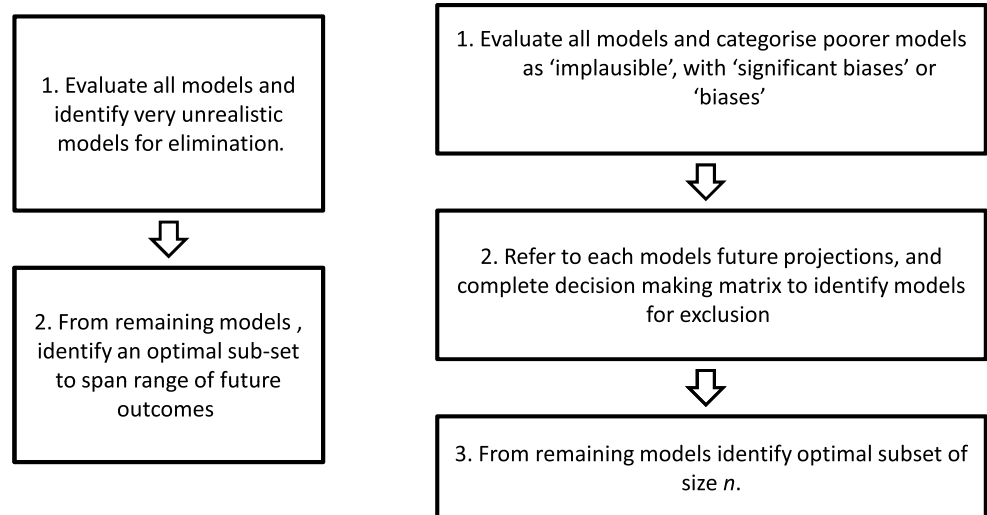
teleconnections for Africa in Sect. 4.3 (See Rowell 2013, for documentation of this analysis).

We analyse all models for which the relevant historical and RCP8.5 variables are available, including those which we do not have the option to select for downscaling. We do this for several reasons. (1) The behaviour of all available models provides a useful benchmark for those models which we can select. (2) It is possible that modelling centres have provided LBC fields only for their 'better' models, and the comparison might give us some impression of this. (3) For the purposes of interpreting the scenarios generated by downscaling, it is useful to know whether the set of models downscaled is representative of the full range of CMIP5 projections, rather than just those for which LBC data are available. A full set of evaluation figures available in the supplementary information in order to all readers to consult this information in their own selection studies.

4 Evaluation criteria and results

Our selection of evaluation criteria includes key aspects of the large-scale climate of the regions (i.e. relevant to driving the high resolution downscaling) and also takes advantage of specific CMIP5 regional assessments documented in the existing literature. However, we note that for many of these existing studies the analysis is limited to only a subset of the LBC-avail models due to the lag in data availability in the CMIP5 archive. An example of this is the thorough evaluation of the Asian summer monsoon of Sperber et al. (2013) in 25 of the CMIP5 GCMs. This limits our scope to make use of this potential very valuable source of information. We therefore note the outcomes of other studies where relevant, and use them as supporting evidence for our model assessment, but we cannot rely on this evidence alone to influence the selection of a model if the assessment

Fig. 1 The GCM selection process demonstrated in McSweeney et al. (2012) (left) compared with the approach proposed in this paper (right)



does not extend to all models. As the body of literature on the assessment of CMIP5 models becomes more complete over coming years and an increasing volume of well-documented processed-based evaluation results becomes available, in future it might be possible to undertake a thorough selection exercise based solely on information available in existing literature.

Our assessments combine the use of metrics with more qualitative assessments (e.g. the visual inspection of climatological fields). In order to combine this information into a summary which can be used to make elimination decisions, we assess all models on a 4-point qualitative scale for each criterion, corresponding with our rationale described in Sect. 2. Any aspects of the models behaviour identified as 'biased' (B), 'significantly biased' (SB) or 'implausible' (IP) are indicated in text where the relevant evidence is described.

For each region, we assess the realism of the annual cycles of rainfall against GPCP (Adler et al. 2003) and CMAP (Xie and Arkin 1997) gridded datasets and annual cycles of temperature against CRU (Mitchell and Jones 2005) observations. For this analysis we use 2 metrics for each of several sub-regions in each region described in Table 3 to describe (a) the pattern correlation of the monthly average values with those in observations and (b) the root mean square error (RMSE). We use these metrics to identify models for visual inspection by highlighting the 5 lowest scores for each metric. It is not clear how such metrics could be used solely to assess models due to the complexities in identifying appropriate thresholds, or indeed ensuring that these are useful metrics in every case, but they allow us to reduce the amount of data required for visual inspection. Judging the implications of the realism of these surface variables for the credibility of the model is more problematic—for example, given good representation

of synoptic-scale dynamics in a global model, downscaling with a regional RCM may improve the representation of poorly-resolved local surface characteristics. However, in some cases, poor representation of the surface variables may be an indicator of large scale deficiencies which would be inherited by any RCM. We are also cautious of excluding models based on metrics where the underlying data used to generate these observed climatologies is relatively sparse or of poor quality and thus may not provide a good estimate of the models' performance. We therefore only apply 'biased' or 'significantly biased' ratings to models for characteristics of surface variables, reserving the 'implausible' category for clear deficiencies in large scale features.

The following sections summarise the key aspects of climate analysed for each region, and the outcomes are summarised in Sect. 4.4.

4.1 Southeast Asia

Following the methodology of McSweeney et al. (2012), we look at the monsoon circulation as the main driver of seasonal rainfall in the region, as well as extending this analysis to include the north-east (winter) monsoon (Fig. 2) as well as the south-west (summer) monsoon (Fig. 3).

A key detail of the north-east monsoon is the north-easterly near-surface flow (850 hpa winds) over the South China Sea directing near-surface flow towards the Malaysian peninsula. In many of the CMIP5 models, the flow in this region has too strong an easterly component, such that flow is directed more towards the coast of Vietnam rather than further south towards the Malaysian peninsula—this is particularly true of the models *inmcm4*, *MIROC-ESM*, *MIROC-ESM-CHEM*, *NorESM-1-M* and *NorESM-ME* (SB), and, to a lesser extent, *CCSM4*, *CNRM-CM5*, *HadGEM2-ES* and *HadGEM2-CC* (B). The 'significant biases' and 'biases'

Table 3 Definitions of regions and sub-region

		North	South	East	West
<i>Europe</i>		72.5	35	25	−10
UK	UK	60	50	2.5	−10
Scandinavia	SCA	72.5	55	25	5
Mediterranean	MED	45	35	25	−10
W Europe	WEU	55	35	17.5	−10
E Europe	EEU	55	35	25	17.5
<i>Africa</i>		36	−35	53	−20
N Africa	NAF	36	20	40	−20
W Sahel	WSH	20	10	20	−20
E Sahel	ESH	20	10	40	20
W Tropical Africa	WTA	10	−10	27.5	−20
Horn of Africa	HA	15	−15	53	27.5
Southern Africa	SA	−10	−35	42	10
<i>Southeast Asia (SEA)</i>		25	−11	127	93
Continental SEA	CSEA	25	7.5	110	93
Malaysia peninsula and Sumatra	MPS	7.5	−6	107	95
Java	JV	−5	−11	127.5	105
Borneo	BN	7.5	−5	125	107.5
New Guinea	NG	0	−10	150	130
Philippines	PL	20	5	127.5	120

categories are used here because this is a relatively subtle characteristic of the flow which may be corrected to some degree in the higher-resolution RCM simulations.

Most models capture the observed broad-scale characteristics of the south-west (SW) monsoon, i.e. that the occurrence of strongest flow in the core of the Somali Jet is clear, and flow is largely westerly across peninsular India before diverting to a south-westerly flow across the Bay of Bengal, westerly across continental southeast Asia and finally turning directly southerly before reaching the Philippines. While most models exhibit some variations on these key features, *MIROC-ESM-CHEM* and *MIROC-ESM* both have a monsoon flow which diverts to a southerly flow before reaching continental southeast Asia, representing a substantial deviation from the patterns observed. The implications of this unrealistic representation of the large-scale characteristics of the SW monsoon in *MIROC-ESM* and *MIROC-ESM-CHEM* is that representation of the characteristics of flow over southeast Asia are particularly poor—notably, the resulting flow over the South China Sea is predominantly north-westwards instead of north-eastwards as seen in observations. We argue that this significant shortcoming suggests strongly that these models will be unable to represent the potential implications of changes in the SW monsoonal circulation in southeast Asia (Implausible—IP).

The model *inmcm4* has an 850 hpa flow which is significantly weaker than observations throughout the region,

although the features are otherwise reasonably realistic (Significant Biases—SB), while *IPSL-CM5B-LR* and *MRI-CGCM3* (SB) both have a very weak Somali jet combined with flow over southern Asia which is predominantly westerly (compared with the observed flow which is southerly around southern India and becomes south-westerly in the Bay of Bengal).

Other models which demonstrate errors in the circulation are *MIROC5* (flow is directed too southerly over continental southeast Asia) (Biases—B), *ACCESS1-3*, which underestimates the strength of the Somali jet (B), and *FGOALS-g2* and *IPSL-CM5A-LR* all have flow which is significantly too westerly across the Bay of Bengal (B, although we note that there are several other models which offer only a marginal improvement in this characteristic). All GISS models demonstrate a weak Somali jet and substantially too-strong southerly component of flow into the Bay of Bengal (not rated as not an LBC-avail model).

The annual cycles of temperature demonstrate a warm bias in most models (Fig. 4). The models *bcc-csm1-1-m* (B) and *ACCESS1-3* (B) have the largest warm biases, and this is consistent across most sub-regions. *EC-EARTH* (B) conversely demonstrates a cool bias throughout the region, but significantly has a much weaker seasonal cycle of temperature than observations. The models' representations of annual rainfall cycles are highly varied (Fig. 4). Whilst most models capture the July–August peak in rainfall over CSEA realistically (although with tendency to over-estimate the magnitude somewhat), the seasonal cycles in other regions such as MPS and BN are generally much poorer. Because the performance in these two sub-regions is unrealistic in so many of the models, we cannot differentiate enough between their performance to class any models as having 'biases' for these regions. In other regions, models which demonstrate particularly poor behaviour compared to other models are *MIROC-ESM* (B) and *MIROC-ESM-CHEM* (B), which capture the seasonal cycle poorly in JV (not capturing the drier JAS period at all) and PL (peak rainfall occurs in SON rather than JJA).

A thorough assessment of the monsoon and seasonal rainfall characteristics for a set of CMIP5 and CMIP3 models can be found in Sperber et al. (2013) a number of metrics describing the climatological and interannual/intra-seasonal variability are presented. The study does not specifically endeavour to highlight the underperforming models as we do, but does observe that no individual model can be identified as the 'best' considering all metrics, based on an analysis which highlights the 'best' 5 models in each category. We reverse their analysis of the listed metrics and ask whether any model(s) can be identified as significantly 'worse' by identifying which of the models analysed have the lowest scores. First, we identify those models with the lowest 5 scores across both CMIP3 and CMIP5 ensembles,

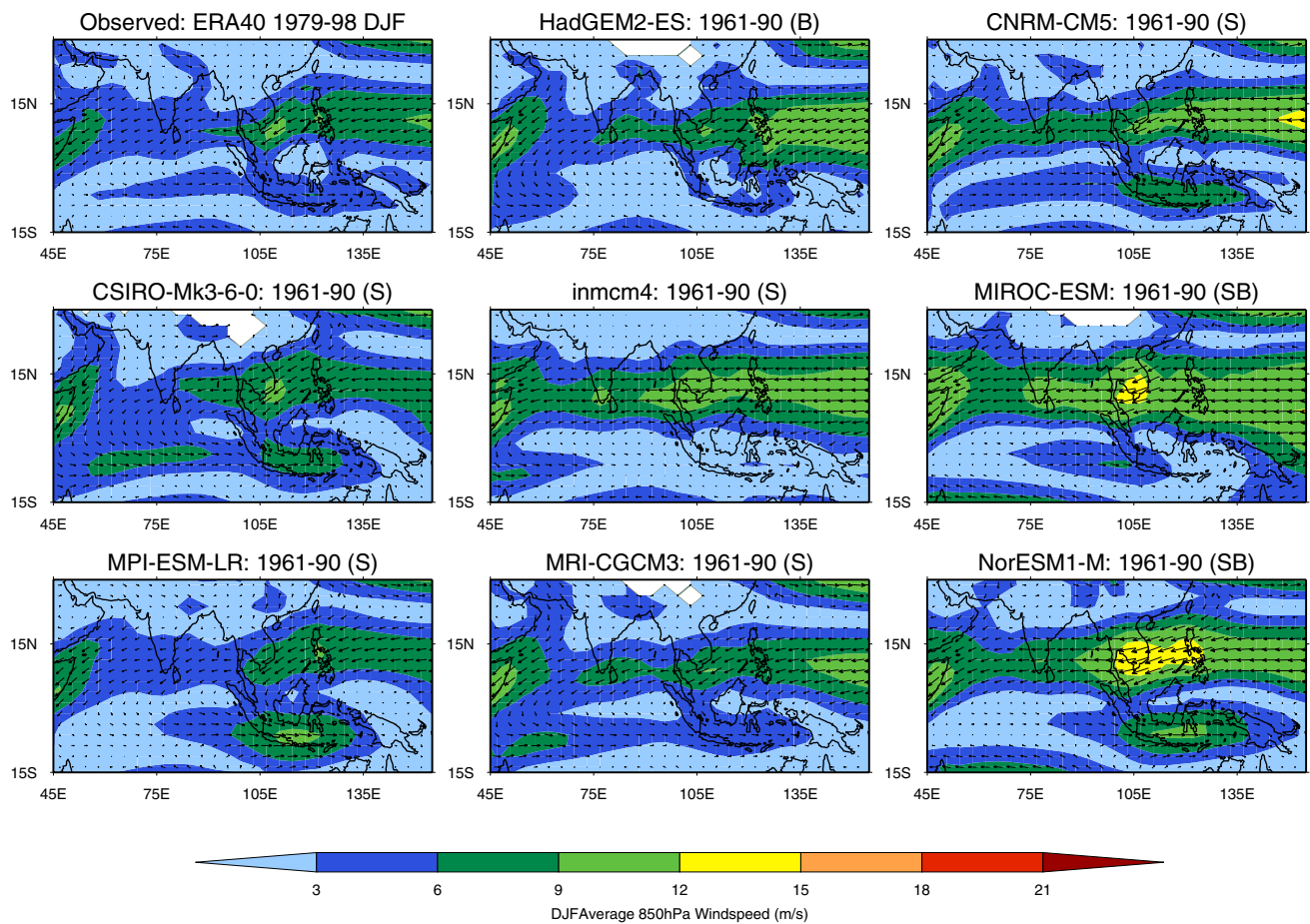


Fig. 2 The north-east (winter) Asian monsoon circulation in 850 hPa flow for ERA40 reanalyses (Uppala et al. 2006) and a sample of CMIP5 models. For equivalent plots for all available CMIP5 models see figures S1(a) and S1(b) in the online supplementary material

and then ask (1) are these lowest scores significantly lower than the scores for the majority of other models and (2) are other scores close to the values found for lowest scoring models, which should therefore be treated similarly? Those models with particularly low scores for any of the indices were assessed as having ‘biases’, as errors in their representation of present-day variability of the monsoon imply that they are unlikely to represent future change in monsoon variability realistically (The limited scope of this study to a set of available CMIP5 models means that we use only the ‘biases’ category). Values for selected metrics most relevant to this study are listed in Table 4.

The skills scores that describe the model climatology include the spatial pattern correlation of rainfall and 850 hPa flow with observations and a suite of metrics describing the annual cycle of rainfall including onset, peak, withdrawal and duration. We do not use these scores to contribute to our analysis due to the overlap with characteristics we have already assessed, but cite the outcomes here as supporting evidence. The lowest scoring CMIP5 models for these climatological metrics were found to be

MIROC-ESM, *MIROC-ESM-CHEM* and *MRI-CGCM3*. The two *MIROC* models scored badly on pattern correlation between model and observed precipitation and 850 hPa (as expected given our observations that the 850 hPa flow does not extend far enough east) and *MRI-CGCM3* scored badly on almost all metrics reflecting the annual rainfall cycle (this model was also identified as one of the worst models in our analysis of annual cycles of rainfall (Fig. 4) but we noted that it did not perform significantly worse than other CMIP5 models).

The indices relating to interannual variability provide an assessment of aspects of the model behaviour that we have not already assessed. Sperber et al. calculate the temporal correlation between anomalies of All-India Rainfall (AIR) and the Nino3.4 index, finding that only 11 of the 25 CMIP5 models simulate a statistically significant anti-correlation (whilst the observed relationship in observed datasets is around -0.5 , bounds of -0.3 to -0.75 reflect inter-decadal range of values). CMIP5 models with low scores were *MIROC-ESM*, *inmcm4*, *FGOALS-g2*, *GISS-E2-H* and *MIROC-ESM-CHEM* (B). Sperber et al. also

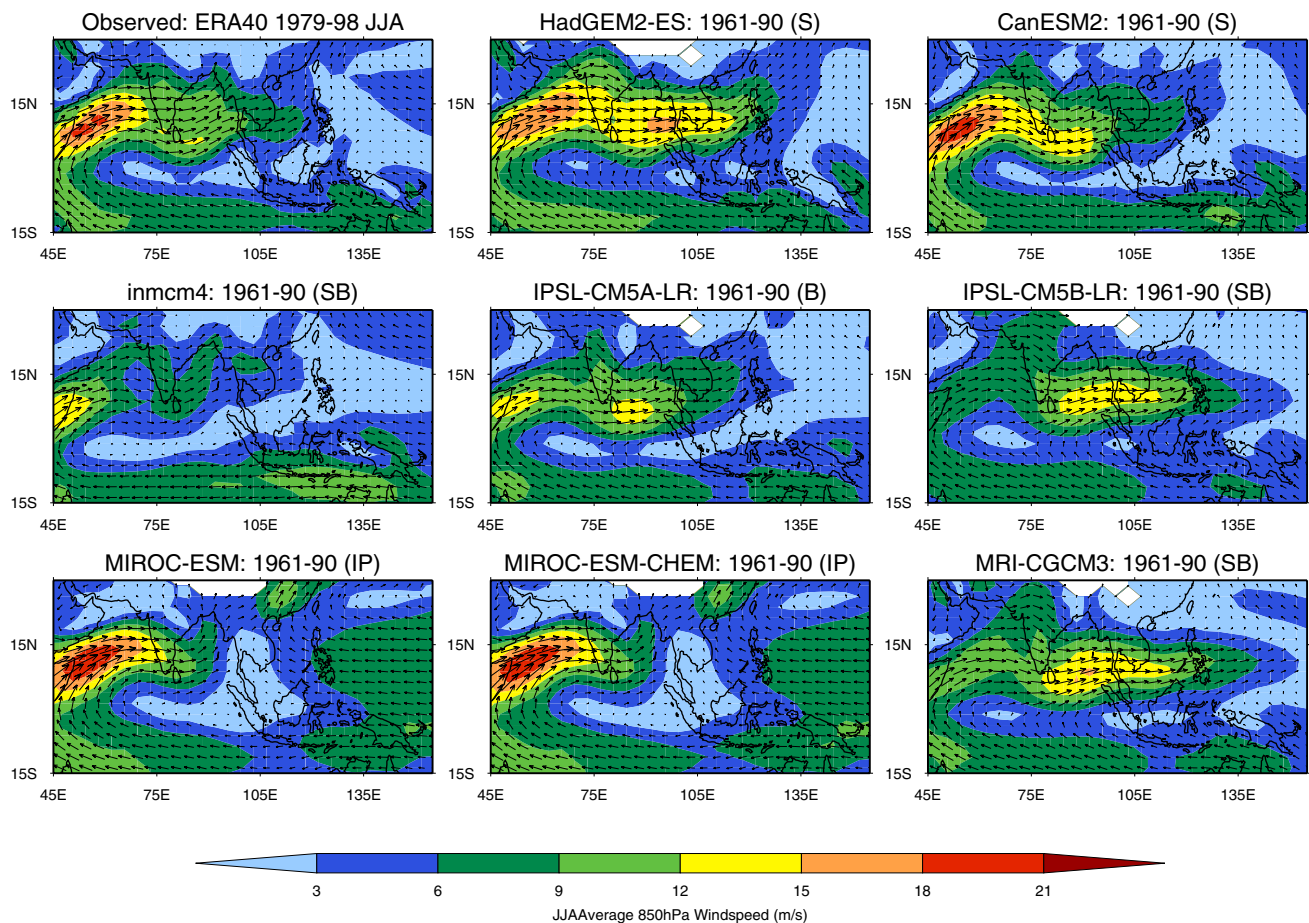


Fig. 3 The south-west (summer) Asian monsoon circulation in 850 hpa flow for ERA40 reanalyses and a sample of CMIP5 models. For equivalent plots for all available CMIP5 models see figures S2(a) and S2(b)

calculate the pattern correlation between anomalies of Nino3.4 and the AIR index. The strong positive correlation of 0.79 indicated by observations is underestimated by most models, with models *bcc-csm-1*, *CanESM2*, *HadGEM2-CC*, *MIROC-ESM*, *MIROC-ESM-CHEM* and *MIROC5* demonstrating very low or weakly negative correlation (B).

A further pair of indices describes the interannual variability associated with the east Asian summer monsoon, calculated by regressing the June–August Wang–Fan (WFN) zonal wind shear index (Wang and Fan 1999) against anomalies of observed rainfall and 850hpa wind. One CMIP5 model clearly behaves less realistically than the others with a negative pattern correlation of rainfall anomalies—*inmcm4* (B), whilst all other models captured the observed positive correlation. Correlations between 850hpa wind anomalies were consistently positive and above 0.65 in CMIP5 models, such that lowest values were not considered to be significantly lower than values for other models.

4.2 Europe

European weather and climate is dominated by the passage of frontal systems, dominantly from the south west. The passage of these systems (the North Atlantic storm track) is sensitive to the local atmospheric dynamics (for example, the position of the jet stream, the impact of blocking anticyclones which disrupt the path of these systems, and, on inter-annual to multi-decadal timescales, the North Atlantic Oscillation) as well as ocean currents, specifically the Gulf Stream. While current models simulate European climate with considerable skill, common errors such as a tendency to underestimate blocking frequencies (Woollings 2010) highlight limitations that should be considered when interpreting their projections. We assess the models' representation of the large scale climatological flow at 850 hpa, the position and frequency of storm tracks and the annual cycles of surface temperature and precipitation in order to identify large scale behaviour in the models which is represented too poorly to be considered credible.

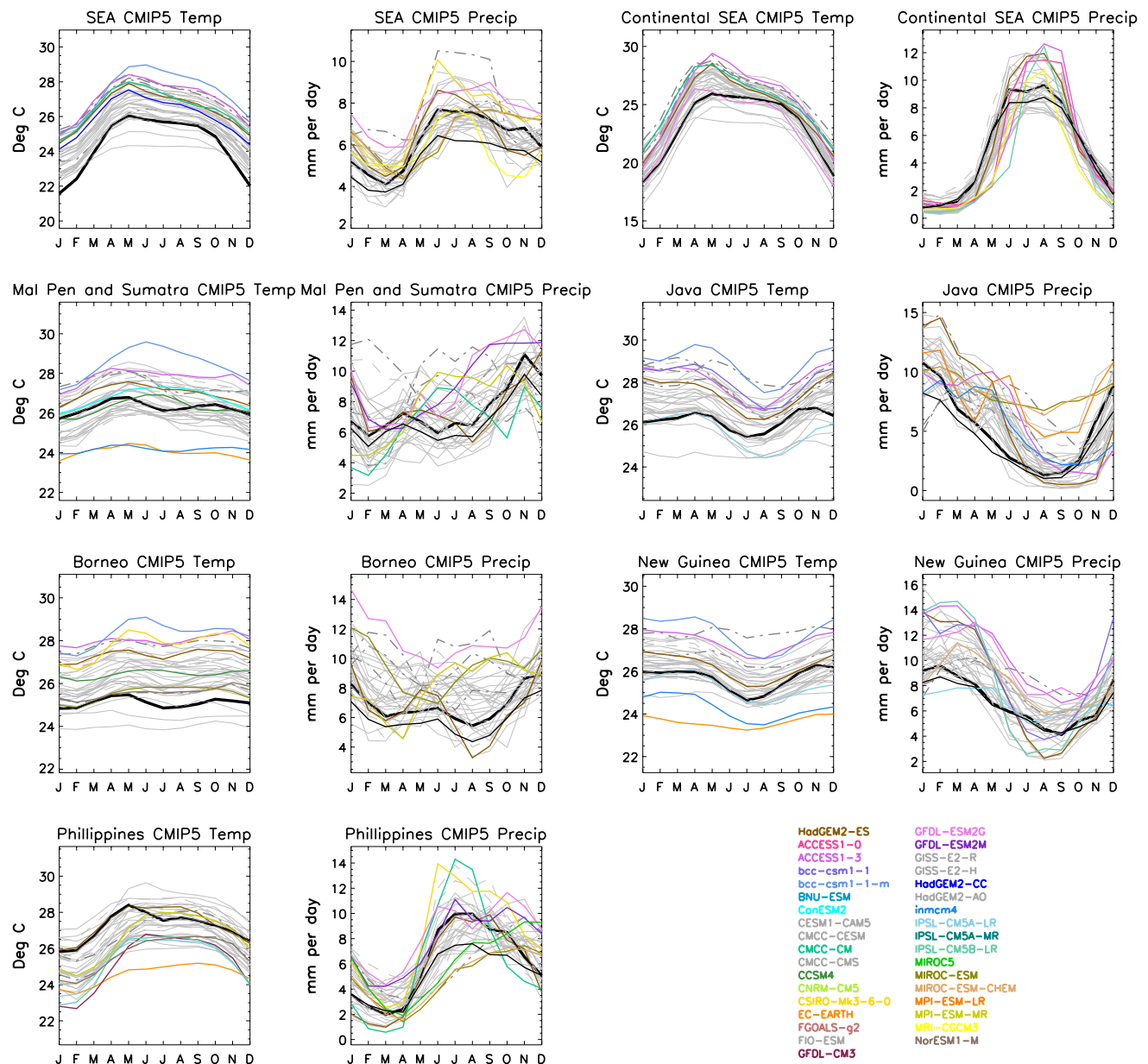


Fig. 4 Annual cycle of temperature and precipitation in CMIP5 models for southeast Asia and its sub-regions (see Table 3 for definition of subregions). Colours models where at least one of the 2 metrics (r , $rmse$) lies in the lowest 5 across all available models, and are LBC-avail. Dark grey 'dash-dot' one of the metrics lies in the lowest 5,

but data are not LBC-avail. Pale grey solid line models are LBC-avail and metrics do not lie in lowest 5, pale-grey dotted line, models are not LBC-avail and do not have metrics in lowest 5. Black observations from CRU (temperature) and GPCP and CMAP (precipitation) datasets

The climatological seasonal circulation is shown in Figs. 5 and 6. The stronger winter flow is often more westerly in models than is observed—this is most notable in models *ACCESS1-3*, *bcc-csm1-1*, *BNU-ESM*, *CanESM2*, *CCSM4*, *CSIRO-mk3-6-0*, *FGOALS-g2*, *IPSL-CM5A-LR*, *NorESM1-M* and *NorESM1-ME* (B). In several models the region of strongest flow sits considerably too far south towards Spain and the Mediterranean rather than the UK—this is particularly so in *FGOALS-g2* and *CSIRO-mk3-6-0* (SB).

During summer the flow is weaker and more westerly, which is replicated by most models. However, in *FGOALS-g2*, the flow remains significantly too far south, such that there is no clear westerly flow across the UK (SB) and in *IPSL-CM5B-LR* and *MIROC5*, there is no clear westerly flow across Europe from the Atlantic (IP) (this is also the case for all 4 GISS models, but these are not rated as are not LBC-avail).

Storm frequency in the Northern Hemisphere mid-latitudes is shown for zonal means in Fig. 7 for CMIP5 models

Table 4 Indices of summer monsoon variability for CMIP5 models from Sperber et al. (2013)

	Indian Monsoon		E. Asian Summer Monsoon	
	<i>AIR/N34</i>	<i>Pr pattern corr</i>	<i>Pr</i>	<i>850hpa</i>
Observations	-0.533	0.798	0.959	0.989
BCC-CSM-1	-0.250	-0.140	0.695	0.93
CanESM2	-0.273	<i>0.014</i>	0.672	0.861
CCSM4	-0.556	0.337	0.789	0.947
CNRM-CM5	-0.307	0.245	0.642	0.894
CSIRO-Mk3.6.0	-0.487	0.162	0.346	0.858
FGOALS-g2	<i>-0.052</i>	0.238	0.739	0.936
GFDL-CM3	-0.442	0.192	0.315	0.867
GFDL-ESM2G	-0.289	0.251	0.458	0.972
GFDL-ESM2M	-0.187	0.251	0.606	0.955
GISS-E2-H	<i>-0.094</i>	0.254	0.586	0.918
GISS-E2-R	-0.366	0.379	0.656	0.906
HadCM3	-0.299	0.180	0.773	0.897
HadGEM2-CC	-0.335	-0.068	0.787	0.935
HadGEM2-ES	-0.344	0.216	0.839	0.949
INM-CM4	-0.033	0.110	-0.047	0.816
IPSL-CM5A-LR	-0.700	0.611	0.450	0.708
IPSL-CM5A-MR	-0.763	0.636	0.532	0.749
MIROC-ESM	0.088	<i>0.061</i>	0.596	0.694
MIROC-ESM-CHEM	<i>-0.104</i>	<i>0.045</i>	0.687	0.882
MIROC4h	-0.327	0.529	0.723	0.921
MIROC5	-0.321	<i>0.010</i>	0.567	0.946
MPI-ESM-LR	-0.291	0.401	0.283	0.899
MRI-CGCM3	-0.274	0.338	0.819	0.937
NorESM1-M	-0.690	0.522	0.811	0.959

Values in bold and italics are those which lie in the lowest 5 across all CMIP3 and CMIP5 models studied, values in italics were close to those found in the 'lowest' models. Highlighted in grey are those models which were allocated a 'biases' rating as a result of these indices. 'AIR/N34': Correlation between anomalies of Nino3.4 index and All-India rainfall. 'Pr Pattern Corr': Spatial correlation of JJAS precipitation anomalies obtained from regression with the Nino3.4 SST. 'E.Asian Pr': Negative of Wang–Fan zonal wind shear index regressed against JJA precipitation anomalies. 'E.Asian 850 hpa': Negative of Wang and Fan (1999) zonal wind shear index regressed against JJA 850 hpa wind anomalies

and ERA-Interim storm tracks identified using TRACK (Hodges et al. 2011). Observations indicate a clear tri-modal pattern with peaks in storm tracks at around 40°N (Mediterranean), 55–60°N (northern Europe) and 70°N (Arctic Circle). Most models overestimate the number of tracks in southern Europe (the 'trough' in between the Mediterranean and northern Europe peaks) and under-estimate the number of tracks northern-most latitudes, but most broadly capture the tri-modal pattern. Five models clearly behave less realistically than others, mis-representing this geographical distribution in all seasons—*MIROC-ESM*, *MIROC-ESM-CHEM*, *FGOALS-g2*, *BNU-ESM* and *bcc-csm1-1* (IP).

The annual cycles of precipitation and temperature (Fig. 8) show a notable negative bias in *IPSL-CM5B-LR*

(SB) in winter months of around 5 degrees in UK and Scandinavia. In the case of the *IPSL-CM5B-LR* cool bias, this takes the winter temperature well below 0 °C over the UK, and is therefore likely to have significant implications for the extent of sea-ice in the northern oceans (SB). All three IPSL models demonstrate poor realism of the annual cycle of rainfall in most regions, but particularly MED, WEU and EEU, where erroneous peaks in JJA rainfall occur (B).

4.3 Africa

The wide range of climate conditions encountered across Africa are strongly influenced by the seasonal migration of the Inter-Tropical Convergence Zone (ITCZ) and the associated seasonal rainfalls. For example, the west-African monsoon brings rainfall to the southern coast of west Africa and into the Sahel in summer months. The climate of many regions is affected by strong interannual variability; teleconnections with major modes of variability in sea-surface temperatures (SST) such as ENSO and the IOD are found in observations, and are described in greater detail in Rowell (2013). The three key aspects of African climate that we assess here are the west African monsoon circulation, the climatological annual cycles of temperature and precipitation for African sub-regions and key teleconnections.

Mapped 850 hpa flow indicates that most models capture the significant features in the flow during DJF (not shown) and JJA (Fig. 9). One aspect where we can differentiate between the models is their ability to capture the west African monsoon—the reversal of flow onto the west-African coast from the south and west during JJA. While there is considerable variation in the strength and direction of this return flow, it is notable that the feature does not appear at all in any of the three *IPSL* models or *MRI-CGCM3* (as well as non LBC-avail models *GISS-E2-H*, *GISS-E2-R-CC*, *GISS E2-H-CC*, and *FIO-ESM*) (SB). We further note that the flow across the African region is exaggerated in magnitude in *MIROC-ESM* and *MIROC-ESM-CHEM* (B).

Assessing the annual cycles of rainfall and temperature, models tend to simulate, on average, too much rainfall, and average temperature tends to be too cool (Fig. 10). The three *IPSL* models and *EC-EARTH* suffer the largest cool biases (although these are not significantly larger than in other models). Sub-regionally, *EC-Earth* shows particularly large cool biases in the two Sahel sub-regions ESH and WSH and WTA (B). Rainfall realism is very mixed. In the Sahelian regions, models can overestimate peak rainfall by 100 % (e.g. *GFDL-ESM2G*, *GFDL-ESM2 M* models, *CSIRO-mk3-6-0*, in WSH, and *MROC5/MIROC-ESM-CHEM* in ESH), or similarly underestimate by almost 80 % (e.g. *MRI-CGCM3*, *inmcm4* in WSH, B). Most however,

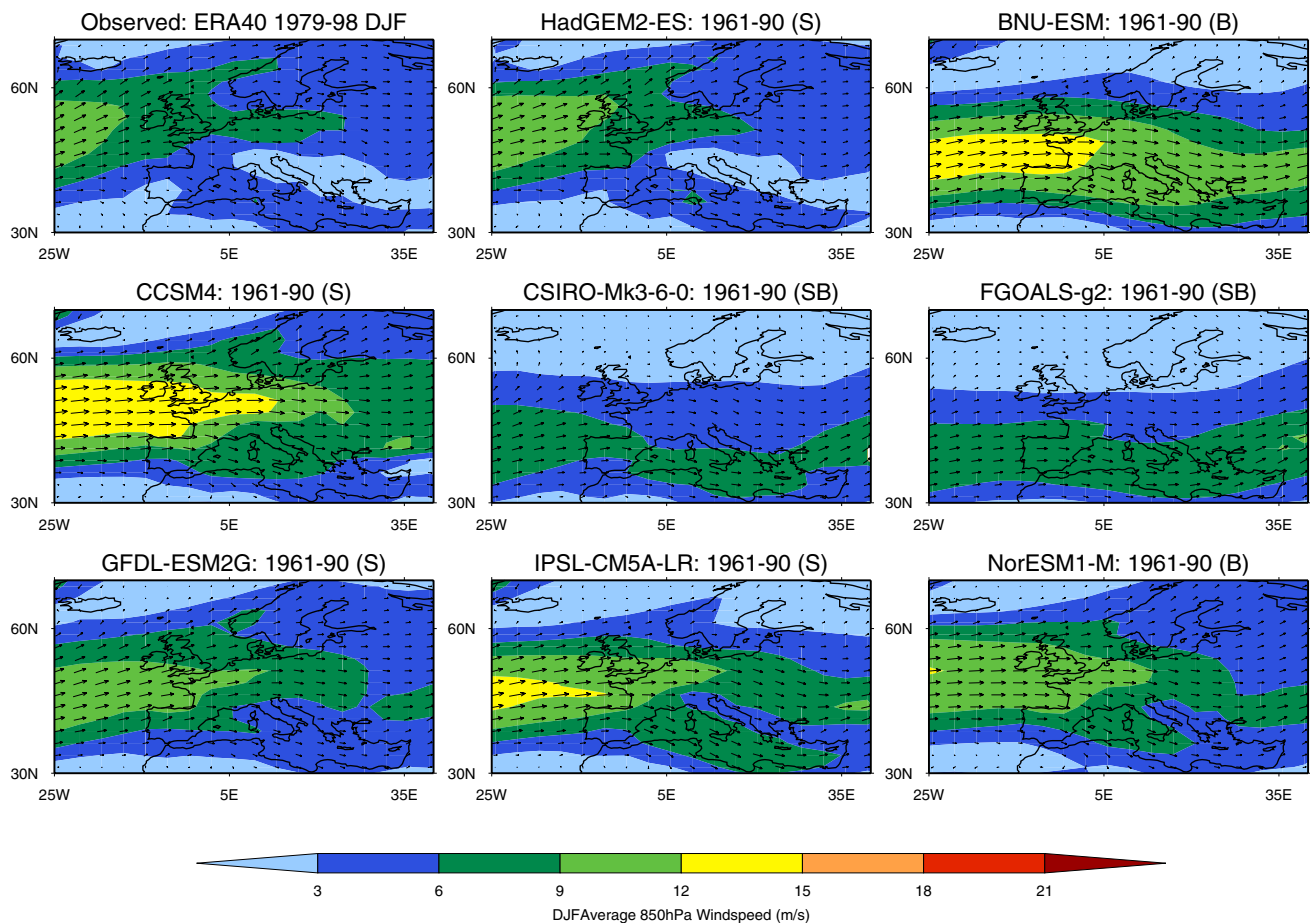


Fig. 5 Winter 850 hPa flow for ERA40 reanalyses and a sample of CMIP5 models for Europe. For equivalent plots for all available CMIP5 models see figures S3(a) and S3(b)

replicate the timing of the rainfall reasonably, with the notable exception of *FGOALS-S2* (rainy season is late in both ESH and WSH, B), *NorESM-1* (flat peak in both WSH and ESH, B) and *bcc-csm1-1* which has no peak JJA rainfall in ESH (B). For other regions, WTA, HA and SA performance is mixed and no models emerge as significantly ‘worse’ than others.

In Table 5 we draw on results of an assessment of 36 key teleconnections affecting Africa manifest as 6 coherent rainfall anomalies with statistically significant correlations to one or more of 6 modes of sea surface temperature (SST) variability (Rowell 2013). The results are summarised for each model as the proportion of those teleconnections that do not differ significantly from those observed at the 10 % level. This metric therefore incorporates information from a range of teleconnections, including whether the models correctly replicate the lack of relationship between some of the SST modes and regional rainfall. The five models with the lowest overall scores are *CanESM2*, *IPSL-CM5A-LR*, *INMCM4*, *IPSL-CM5B-LR* and *CMCC-CMS* (B).

5 Completing the decision-making matrix for model elimination

Having assessed the models against a range of criteria in each region, we summarise this information to provide an overall outcome across the three regions (Table 6). The overall outcome is allocated based on (a) the lowest rating across all criteria assessed and (b) whether those low ratings are given based on more than one of the sub-regions, according to the following criteria:

- **Implausible:** Must score ‘implausible’ in at least one region and at least ‘significant biases’ in another.
- **Significant biases:** Scores ‘implausible’ in one region, but not in any others; or, scores ‘significant biases’ in one or more regions, and/or biases in another region.
- **Biases:** Scores ‘significant biases’ in just one region, or ‘biases’ in 2 or more regions.

It is clear that the performance of models varies between regions, and that applying an ‘overall’ score on a

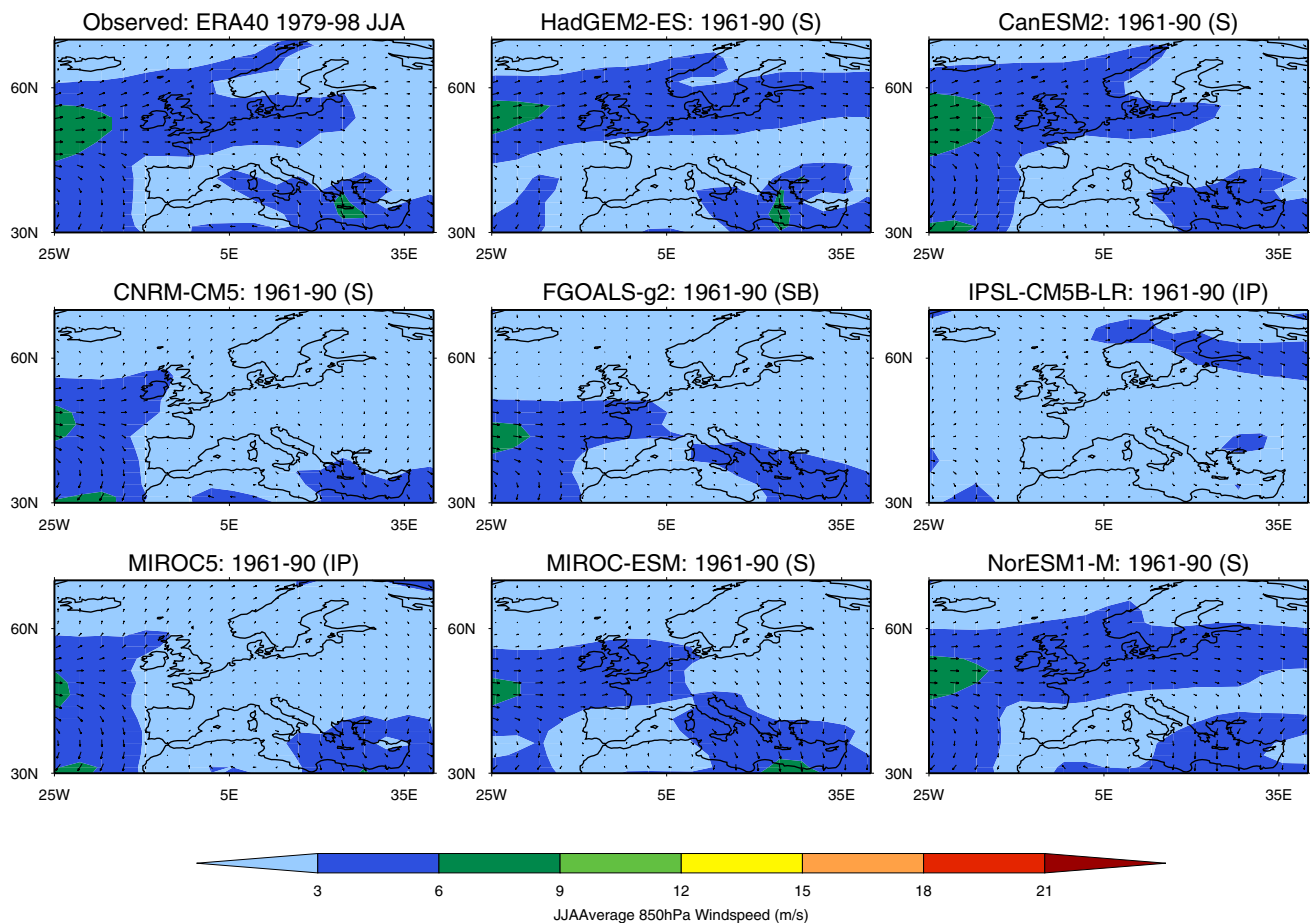


Fig. 6 Summer 850 hpa flow for ERA40 reanalyses and a sample of CMIP5 models for Europe (see For equivalent plots for all available CMIP5 models see figures S4(a) and S4(b))

multi-region basis using the above criteria involves making some compromises on the outcomes for individual regions. Notably, we do not exclude all models which contain an ‘implausible’, only those for which there is evidence of ‘significant biases’ or ‘implausible’ characteristics in another of the three regions.

Note that for models *EC-EARTH* and *FGOALS-g2* their overall scores are lowered by a category reflecting the fact that the 850hpa wind components were not available at the time of analysis. We now combine this with information about the projections in order to complete our decision making matrix Table 1).

We can gain an impression of the projections from the ensemble, and each models position within the ensemble using scatter plots of the change in mean temperature between the two 30 year periods of 1961–1990 and 2070–2099 under RCP 8.5 (ΔT) and similarly the change in mean precipitation (ΔP) (Fig. 11). Further, in order to capture the variations in spatial patterns of precipitation change within each region, we follow the methodology used in McSweeney et al. (2012) by calculating the pattern

correlation between the ΔP fields from each model with that of the ensemble median ΔP , providing an indication of whether the spatial patterns of change are ‘typical’ (highly correlated with ensemble mean) or ‘atypical’ (lowest or anti-correlation with the ensemble mean change). This is plotted against the root-mean-square of the precipitation change, representing the average magnitude of rainfall response (Figure S6).

Amongst the ‘implausible’ models that are excluded based on performance, we see that *MIROC-ESM* and *MIROC-ESM-CHEM* can be identified as ‘outliers’ over Europe in JJA and SON (the warmest and wettest projections), and also in Africa in JJA as the wettest (Fig. 11). The remaining model classed as ‘implausible’, *IPSL-CM5B-LR*, does not lie outside the ensemble range in any of our analyses (Fig. 11). So the implication of excluding these models is greatest for Europe and Africa. For Europe we exclude projections with the largest precipitation increases in summer and autumn, but given the poor representation of the geographical distribution of the storms with which a large proportion of the European

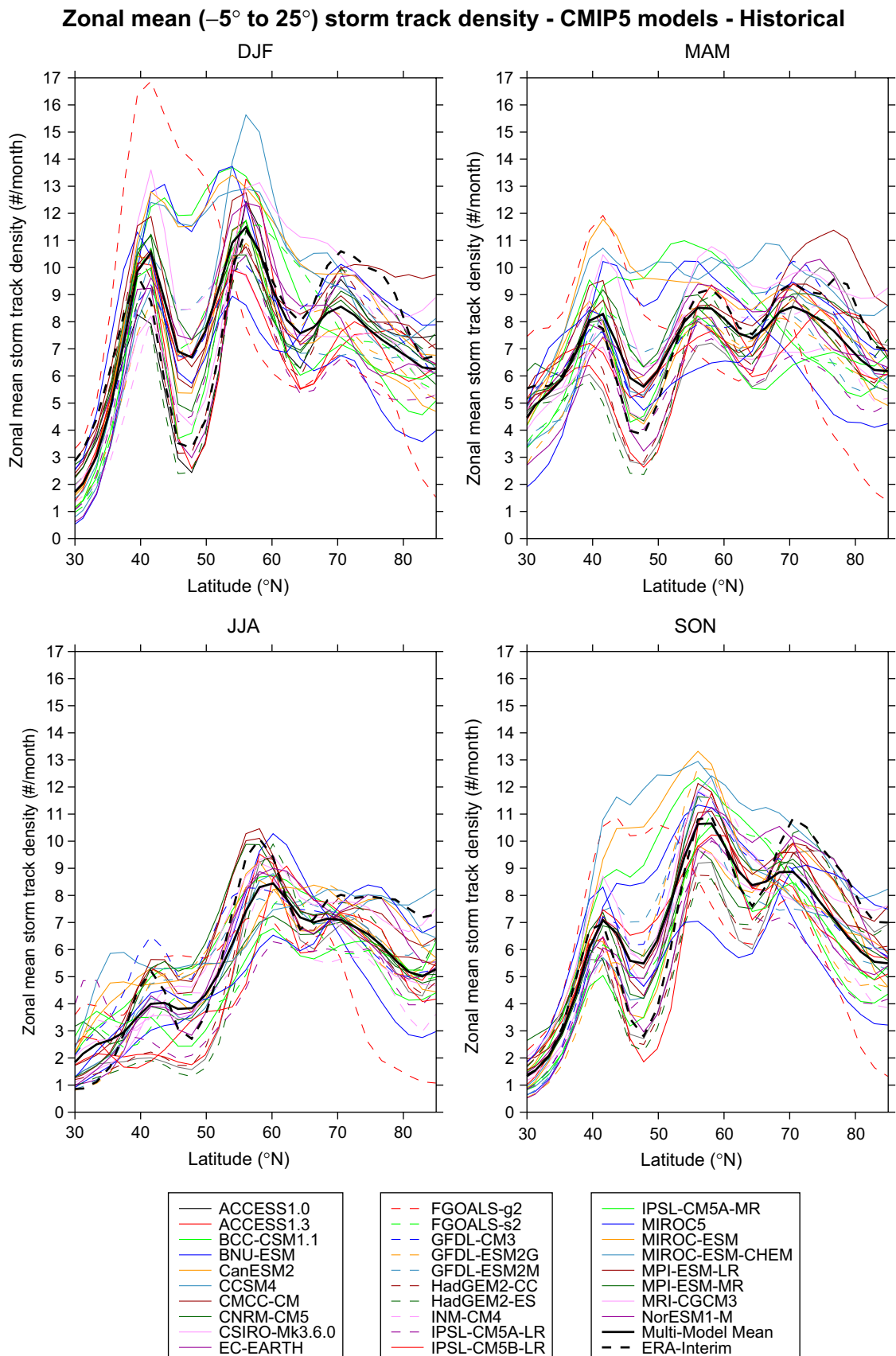


Fig. 7 Zonal mean storm track density in the CMIP5 models

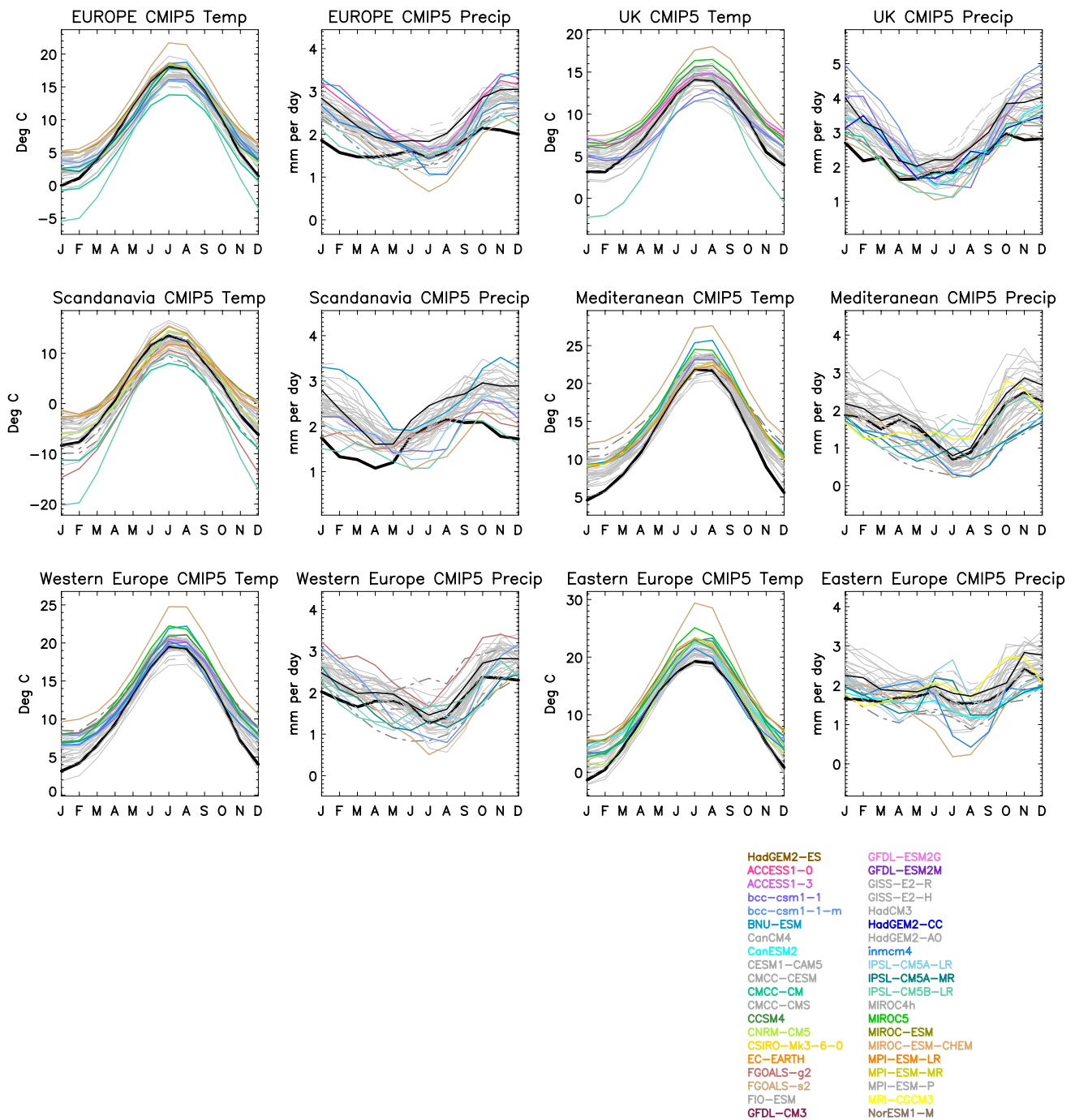


Fig. 8 Annual cycle of temperature and temp and precipitation in CMIP5 models for Europe and its sub-regions (see Table 3 for sub-region definitions). Colours models where at least one of the 2 metrics (r, rmse) lies in the lowest 5 across all available models, and are LBC-avail. Dark grey 'dash-dot' one of the metrics lies in the lowest

5, but data are not LBC-avail. Pale grey solid line Models are LBC-avail and metrics do not lie in lowest 5, pale-gray dotted line, models are not LBC-avail and do not have metrics in lowest 5. Black Observations from CRU (temperature) and GPCP and CMAP (precipitation) datasets

rainfall is associated, we argue that excluding these models, and therefore narrowing the range of projections is justifiable. For Africa, the elimination of these models does narrow the range of projections in one season (JJA) slightly.

Among the models identified as having 'biases' and 'significant biases', *inmcm4* frequently lies at one end of the ensemble range, consistently projecting the least warming in almost all regions and seasons. In Europe this model also gives the driest projections, whilst in SEA it

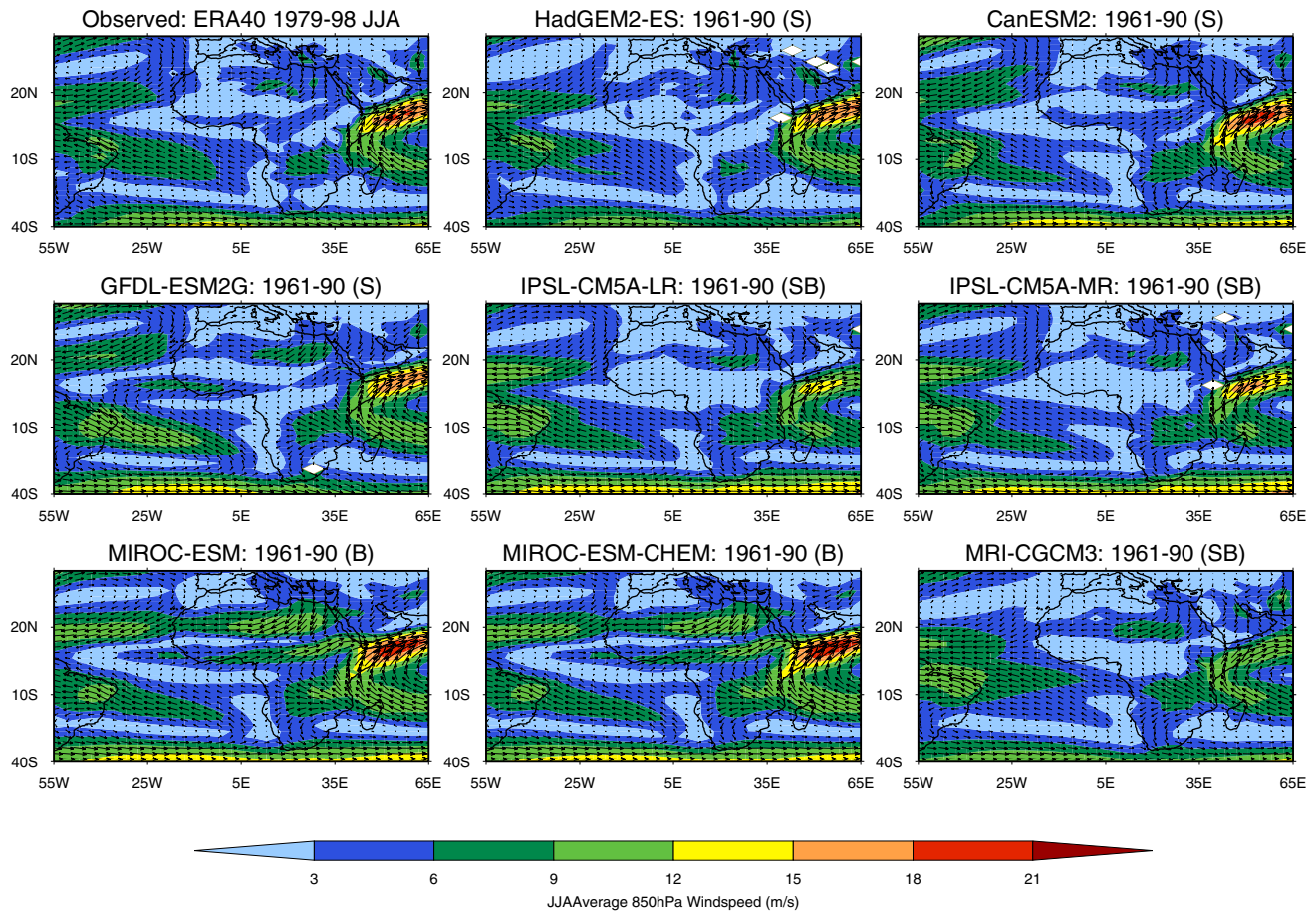


Fig. 9 Summer 850 hpa flow for ERA40 reanalyses and a sample of CMIP5 models for Africa. For equivalent plots for all available CMIP5 models see figures S5(a) and S5(b)

tends to project amongst the wettest (Fig. 11). Change in rainfall projections from the model *inmcm4* also appear to show relatively low correlations with the ensemble mean change particularly for Africa (Figure S6). *BNU-ESM* is out-lying amongst projections for Europe suggesting drier projections than most other models (notably in DJF it is the only model to project negative mean precipitation change), and also indicates amongst the largest increases in rainfall for Africa in JJA and SON (this is important given that models *MIROC-ESM* and *MIROC-ESM-CHEM* have similar projections but are considered ‘implausible’ do to their poor performance in other regions). *IPSL-CM5A-MR* and *IPSL-CM5A-LR* tend to behave similarly to one another in future projections and for southeast Asia comprise the models with the largest temperature and precipitation increases annually. We therefore exclude only one of these IPSL models and we retain *IPSL-CM5A-MR* with a ‘biases’ rating in preference to LR which has a ‘significant biases’ rating. *CSIRO-mk3-6-0* is an outlier as a model projecting drying in southeast Asia. We therefore retain four models classed as having ‘biases’ or ‘significant

biases’—*CSIRO-mk3-6-0*, *BNU-ESM*, *IPSL-CM5A-MR* and *inmcm4*.

The remaining models with ‘biases’ or ‘significant biases’—*bcc-csm-1*, *CanESM2*, *EC-EARTH*, *FGOALS-g2*, *IPSL-CM5B-LR*, *NorESM1-M* and *MPI-ESM-LR*—do not lie outside the range of outcomes according to our analysis, and therefore we conclude that we can exclude these models. These outcomes are summarised in the completed decision-making matrix (Table 7).

6 Sampling ranges of future projections

6.1 Sampling methodology

In order to identify the ‘optimal’ sample of n models from the remaining 16 models which are both LBC-avail, and have not been rejected as a result of analysis in Sect. 4, we explore how randomly selected samples span the range of outcomes across the 3 regions and all seasons. While we have a remit to select 8–10 models, it is useful to

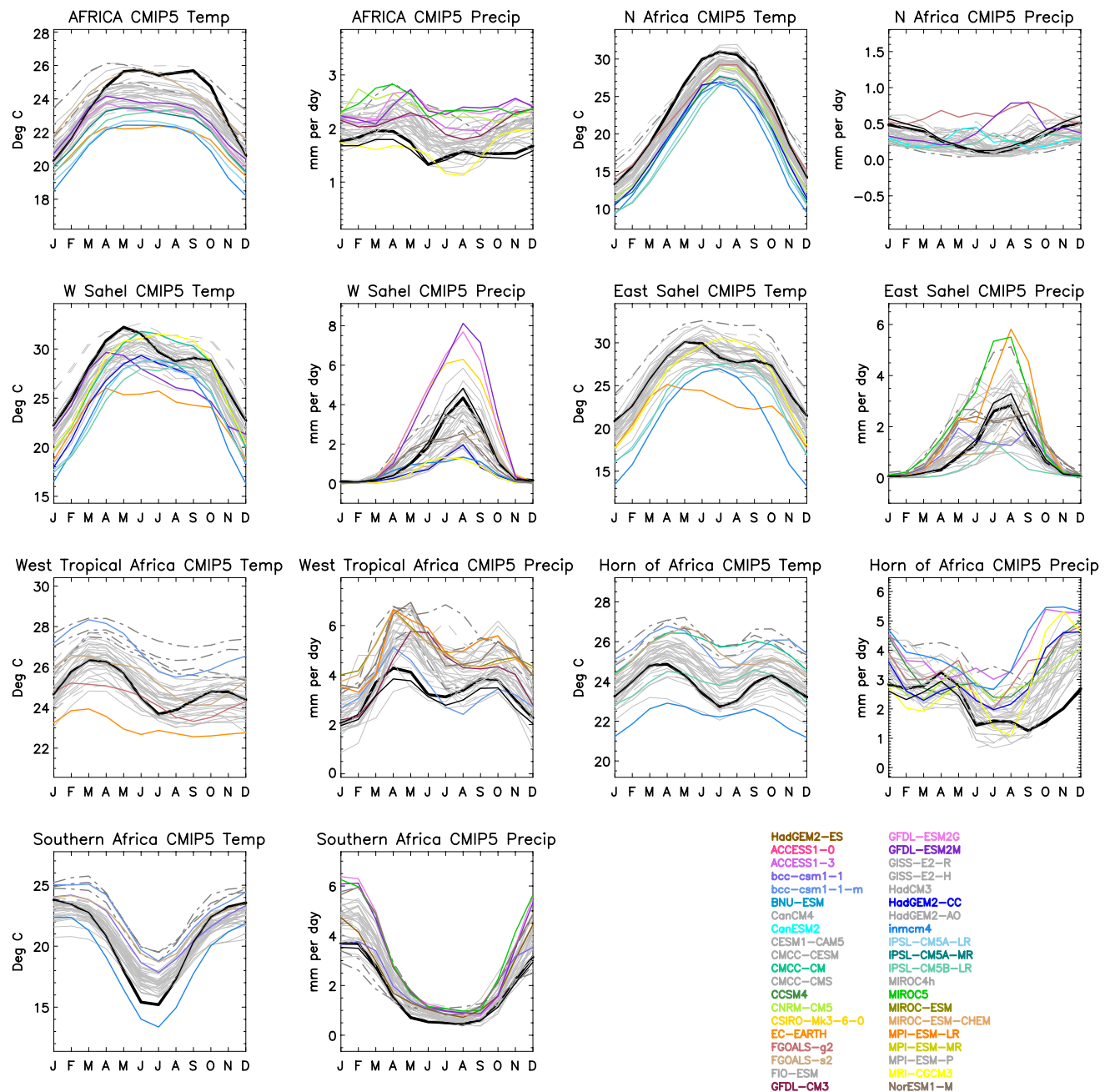


Fig. 10 Annual cycle of temperature and precipitation in CMIP5 models for Africa and its sub-regions (see Table 3 for definition of subregions). Colours models where at least one of the 2 metrics (r , $rmse$) lies in the lowest 5 across all available models, and are LBC-avail. Dark grey 'dash-dot' one of the metrics lies in the lowest 5,

but data are not LBC-avail. Pale grey solid line Models are LBC-avail and metrics do not lie in lowest 5, pale-gray dotted line, models are not LBC-avail and do not have metrics in lowest 5. Black Observations from CRU (temperature) and GPCP and CMAP (precipitation) datasets

understand the potential added benefit of additional models, or how well a smaller subset might perform, so we explore values of n between 3 and 13.

For each value of n , we randomly sample 500 unique combinations of n models, and calculate the fraction of the range of changes in surface temperature and precipitation that are spanned by the subset at each $3.75 \times 2.5^\circ$ grid-cell. This

metric is from hereon described as the Fractional Range Coverage (FRC). A regional average of the FRC is calculated for each season, and annually. The range of regional FRC values at each sample size is shown in Fig. 12. As would be expected, for smaller sample sizes, the range of coverage varies more with sample, highlighting the greater need for careful selection when the number of models to be downscaled is very small.

Table 5 Summary of CMIP5 model representation of key teleconnection relationships in Africa

Model	Number of Teleconnections represented (of 36)
MPI-ESM-P	30
HadGEM2-AO	29
GFDL-ESM2G	29
NorESM1-ME	29
MPI-ESM-LR	29
HadGEM2-CC	29
MPI-ESM-MR	28
MIROC5	28
HadGEM2-ES	27
ACCESS1.0	27
BNU-ESM	26
ACCESS1.3	26
bcc-csm1-1-m	25
CMCC-CESM	25
NorESM1-M	25
IPSL-CM5A-MR	25
GISS_MODEL-E-R	25
GFDL-CM3	24
FIO-ESM	24
CMCC-CM	24
MIROC-ESM	24
MIROC-ESM-CHEM	24
FGOALS-g2	24
CNRM-CM5	24
bcc-csm1-1	23
CSIRO-Mk3.6.0	23
GISS-E2-R	22
GISS-E2-H	22
GFDL-ESM2M	22
MRI-CGCM3	22
<i>CMCC-CMS</i>	<i>21</i>
<i>IPSL-CM5B-LR</i>	<i>20</i>
<i>INM-CM4</i>	<i>20</i>
<i>IPSL-CM5A-LR</i>	<i>19</i>
<i>CanESM2</i>	<i>19</i>

After Rowell (2013). The lowest scoring 5 models are highlighted in italics

Typically, a sample size of 8–10 of the remaining 16 models yields a precipitation FRC value of around 0.9–0.95 in the ‘best samples’ (i.e. the top whisker for each sample size) in the SEA and Europe regions and can be as low as 0.85 for Africa (the lower values for this region may reflect it relatively large size); the equivalent FRC for temperature is 0.95–1.0 for all regions. In a smaller sample of models (5–6) these values drop to around 0.75–0.9, and 0.7–0.85 for precipitation (SEA/Europe and Africa, respectively) and

remain high at above 0.95 for temperature across regions SEA and Europe, and around 0.9–0.95 for Africa. Maximum FRC across samples at each size reaches close to 1.0 for temperature when $n = 8$ –10, but for precipitation only when n is 12 or 13.

6.2 How well do the optimal n models span the range of outcomes?

For each sample size, region and season, the upper whisker of the box-plots might be considered to represent the optimal subset of that size, for that specific region and season. However, here we require a single sample at each size which provides the greatest average FRC across all regions, seasons and both variables. In order to identify an optimal sample we normalise the fractional-coverage values at each grid-cell by the mean and standard deviation of the 500 samples for each value of n (Normalised Fractional Range Coverage—NFRC), and re-calculate the regional averages. Each sample thus has a score which is the averaged NFRC across all regions, seasons and the 2 variables. Due to the lesser probability of capturing the range of changes in precipitation compared with temperature noted above, the precipitation values were weighted $\times 2$ compared with those for temperature. The ‘optimal sample’ for each value of n is simply the sample with the largest average NFRC.

An optimal sample might be that which gives the best coverage over a region, multiple regions, or globally. We may find a lower level FRC for each region within the pan-regional or global optimal samples than in the optimal sample chosen specifically for that region, and all three values are indicated in Fig. 12. Note that our remit here is to identify a set specifically for the 3 regions (the pan-regional optimum), but we explore the implications of extending this to a global optimum for context.

The loss of FRC for any region that is incurred by broadening our requirement for an optimum from the regional to pan-regional, and then to satisfy a global range is explored here. The pan-regional ‘optimal sample’ typically achieves a value for each region, season and variable well-within the upper quartile of the 500 samples. Extending our optimisation criteria to a global generally leads to a small loss of fractional range coverage at each region, as would be expected. This reduction is relatively larger in temperature than for precipitation; for temperature, there is relatively little loss in FRC between the regional and pan-regional optimal sample, and then a greater loss between pan-regional and global. This is reversed for precipitation, where we see a greater disparity between regional and pan-regional than between pan regional and global, which reflects the differences in the spatial characteristics between rainfall and temperature.

Table 6 CMIP5 model performance summary

	South East Asia				Europe			Africa			Overall
	North-east Monsoon Annual Cycles of T and P	Summer Monsoon	Summer Monsoon Indices	T/P Annual Cycles of T and P	Circulation	Storm tracks	Annual Cycles of T and P	W AM	Tele-Connections (Rowell, 2013)		
ACCESS1-0											ACCESS1-0
ACCESS1-3											ACCESS1-3
bcc-csm1-1											bcc-csm1-1
bcc-csm1-1-m											bcc-csm1-1-m
BNU-ESM											BNU-ESM
CanESM2											CanESM2
CCSM4											CCSM4
CMCC-CM											CMCC-CM
CNRM-CM5											CNRM-CM5
CSIRO-Mk3-6-0											CSIRO-Mk3-6-0
EC-EARTH											EC-EARTH*
FGOALS-g2											FGOALS-g2
GFDL-CM3											GFDL-CM3
GFDL-ESM2G											GFDL-ESM2G
GFDL-ESM2M											GFDL-ESM2M
HadGEM2-CC											HadGEM2-CC
HadGEM2-ES											HadGEM2-ES
inmcm4											inmcm4
IPSL-CM5A-LR											IPSL-CM5A-LR
IPSL-CM5A-MR											IPSL-CM5A-MR
IPSL-CM5B-LR											IPSL-CM5B-LR
MIROC5											MIROC5
MIROC-ESM											MIROC-ESM
MIROC-ESM-CHEM											MIROC-ESM-CHEM
MPI-ESM-LR											MPI-ESM-LR
MPI-ESM-MR											MPI-ESM-MR
MRI-CGCM3											MRI-CGCM3
Nor-ESM1-M											Nor-ESM1-M

Colours signify aspects of performance identified as 'Satisfactory' (green), 'Biases' (Yellow), 'Significant Biases' (Orange) and 'Implausible' (Pink). Grey shading indicates that the data were not available at the time of analysis

Typically, the FRC values that we expect to find for a regional-optimum sample of 8–10 models is 0.9–0.95 in precipitation, dropping to 0.8–0.9 in a pan-regional or global-optimal sample (FRC values are consistently a little lower than these for Africa). For temperature, the regional, pan-regional and global optima are all above or close to 0.95.

At the other end of the range of sample FRC values, we can see that, conversely, if we were to choose our sample randomly rather than optimising our selection, the range of values we can expect to capture in our random sample varies widely. Particularly at small sample sizes (e.g. 5), we have a high probability of capturing a very reduced range of outcomes—the median FRC value across all randomly

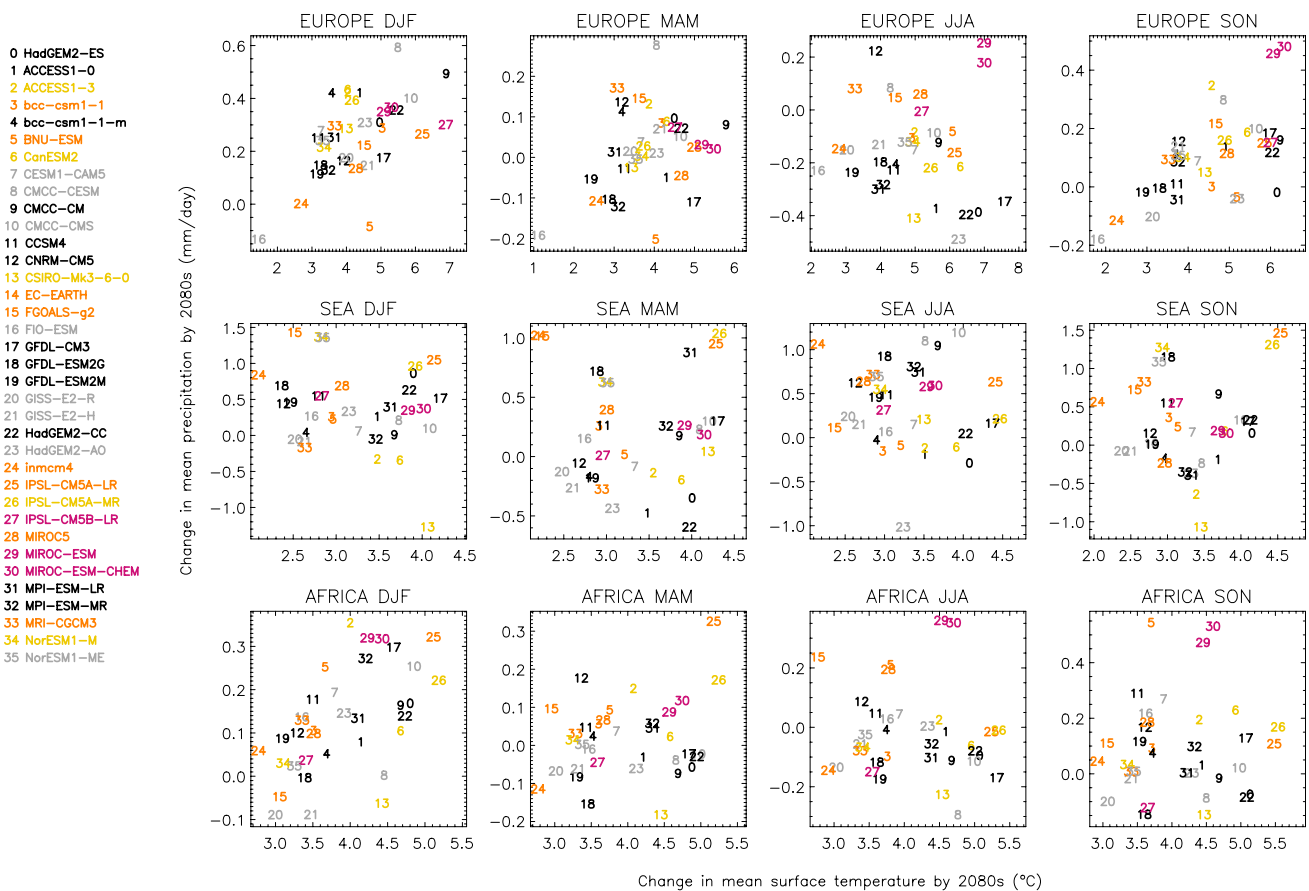


Fig. 11 Scatter plots indicating regional average change in mean temperature and precipitation for all 3 major regions. Models shown are those for which monthly temperature and precipitation data were available for historical and RCP8.5 simulations up to 2100. Colours

indicate the overall outcome from model validation, whereby pink 'implausible', orange 'significant biases' and yellow 'biases'. Grey models are not LBC-avail, remaining models are shown in black

Table 7 Decision-making matrix for model selection

	Outlier	Other models predict similar outcomes
Model suffers sufficient shortcoming that it significantly reduce our confidence in its projections ('Implausible')	<i>MIROC-ESM</i> <i>MIROC-ESM-CHEM</i>	<i>IPSL-CM5B-LR</i>
Model suffers significant shortcomings which we cannot clearly link to confidence in its projections ('Biases/Significant Biases')	<i>inmcm4</i> <i>BNU-ESM</i> <i>CSIRO-Mk3-6-0</i> <i>IPSL-CM5A-MR</i>	<i>ACCESS1-3</i> <i>bcc-csm1-1</i> <i>CanESM2</i> <i>EC-EARTH</i> <i>FGOALS-g2</i> <i>IPSL-CM5A-LR</i> <i>MIROC5</i> <i>MRI-CGCM3</i> <i>NorESM1-M</i>
Model performance is satisfactory ('Satisfactory')	<i>HadGEM2-ES</i> <i>ACCESS1-0</i> <i>bcc-csm1-1-m</i> <i>CMCC-CM</i> <i>CCSM4</i> <i>CNRM-CM5</i>	<i>GFDL-CM3</i> <i>GFDL-ESM2G</i> <i>GFDL-ESM2M</i> <i>HadGEM2-CC</i> <i>MPI-ESM-MR</i> <i>MPI-ESM-LR</i>

Rejected models are highlighted in italics

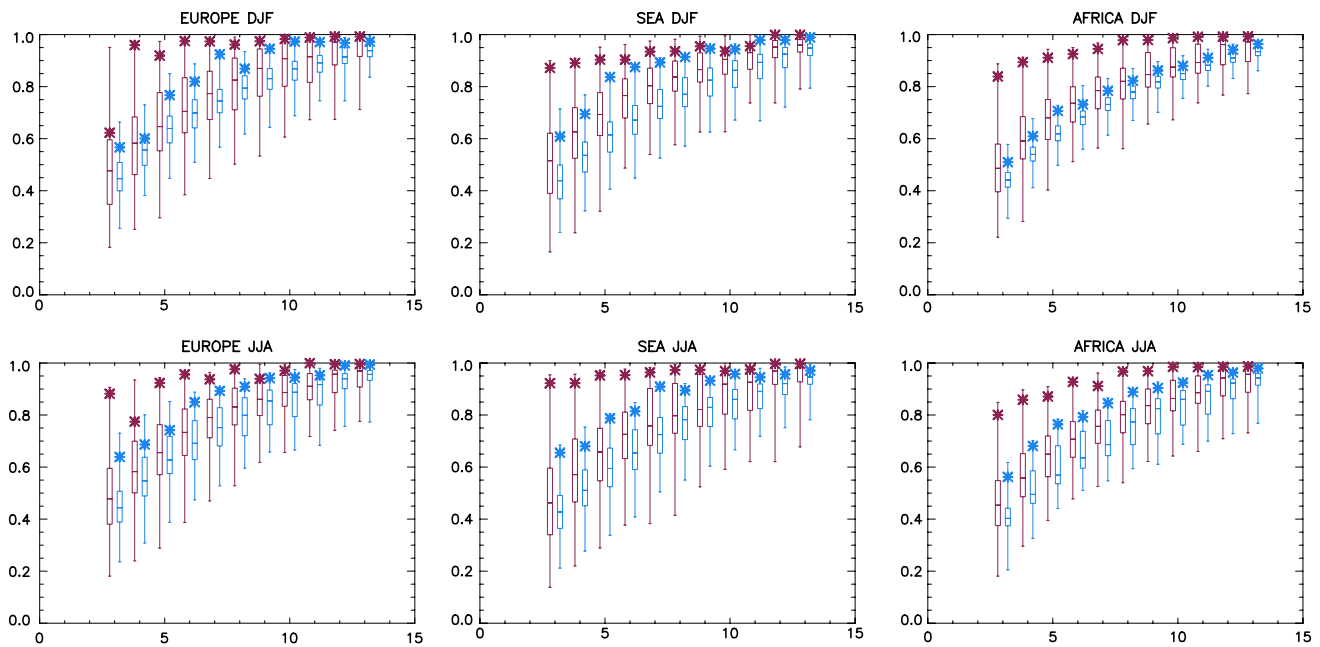


Fig. 12 Boxplots indicating the fraction of the range of outcomes (FRC) by the 2080s under RCP8.5 spanned by 500 samples of n models, where n is 3–13 (out of an available 16 models which are both LBC-avail and have not been excluded for poor realism). Boxplots depict the median, and inter-quartile range of the 500 samples,

with whiskers indicating the full range. Filled circles indicate the sample identified as ‘optimal’ across all 3 regions and 4 seasons, square optimal sample for each region across all seasons and cross indicates the globally optimal sample. Pink temperature, blue precipitation. Examples shown for seasons DJF and JJA

selected samples is around 0.6–0.7 indicating that approximately 50 % of samples we might choose would yield FRC values lower than this.

We note that in this stage scores might be weighted similarly in order to prioritise other aspects of change—for example, to prioritise one season if it is more relevant to impacts studies. In this case, however, our remit is broad so all seasons are weighted evenly.

For our preferred sample size ($n = 8$), we show the geographical coverage of the pan-regional optimal sample in Figs. 13 and 14, and a summary of the regional-average change in Fig. 15. Figure 13 demonstrates near-complete coverage of the range of projections in temperature (>0.9) quite consistently throughout the region, exceptions are small and scattered, with no large coherent regions of poorer coverage. For precipitation (Fig. 14) the gaps in coverage are larger and some are coherent, but this is expected due to lower average coverage and higher spatial variability of changes in rainfall. A region of lower FRC is evident in the south-east corner of the SEA region in SON.

Figure 15 shows the coverage of regional-average changes by the pan-regional optimal set of 8 models. We can see that although the range of outcomes is almost fully covered by the set in most regions and seasons, the selected set notably excludes the only available and not excluded model (CNRM-CM5) that indicates increases in European rainfall in JJA. We might therefore recommend the addition

of this model in order to capture this range more fully, thus recommending the use of a set of 9 models for studies in these regions.

7 Discussion and conclusions

We have demonstrated a methodology for selecting a set of available GCMs which (a) avoids including models which are least realistic and (b) simultaneously captures the maximum possible range of changes in mean temperature and precipitation for three continental-scale regions. Such a subset should represent the full range of GCM-simulated future climate outcomes sufficiently well to provide a set of projections for any of the three regions that, while based on a smaller number of models, still provides sufficiently representative ranges of future climate outcomes to remain policy-relevant.

Using a set of GCMs rather than the full ensemble might be considered a compromise required due to resource limitations for downscaling. However, we demonstrate that a strategically selected set of models can capture a representative range of changes in mean temperature and precipitation. We also identify models which are less realistic and we can avoid including in our subset without affecting the range of outcome significantly. In only a few cases do we eliminate models which are outliers affecting the range of outcomes.

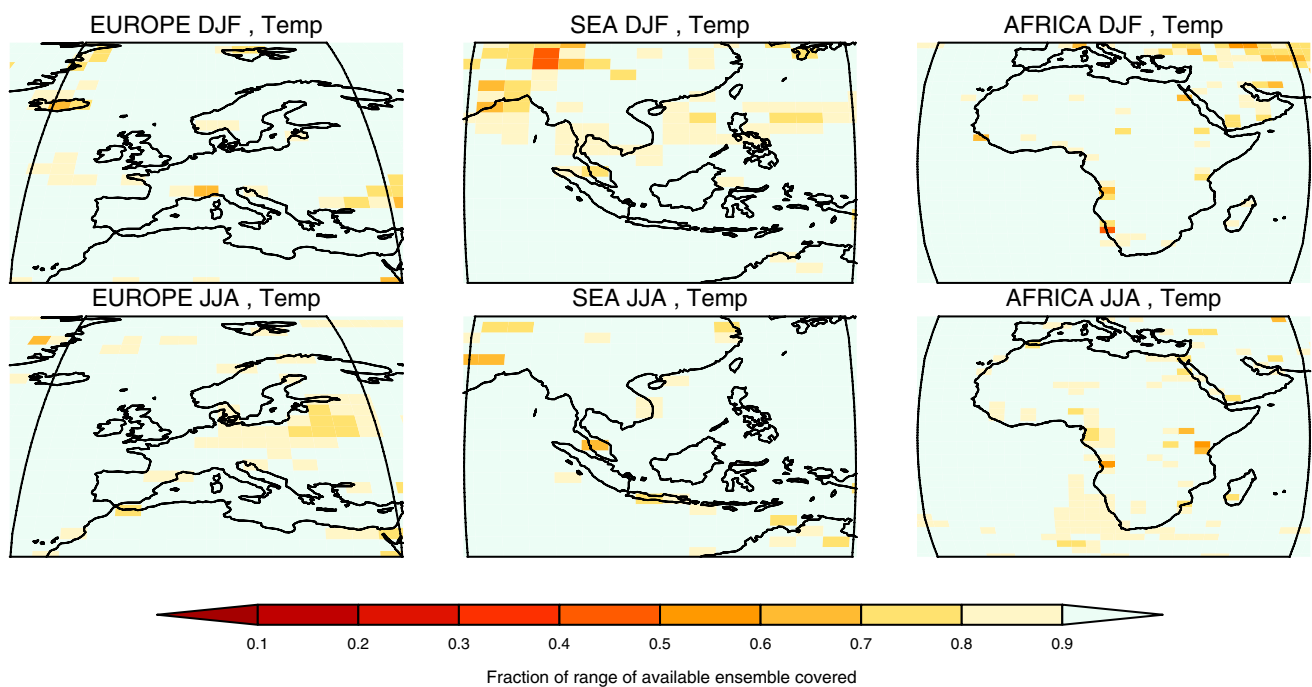


Fig. 13 Fraction of the range of projections in temperature spanned by the selected 8-member subset across each of the 3 regions and seasons DJF and JJA

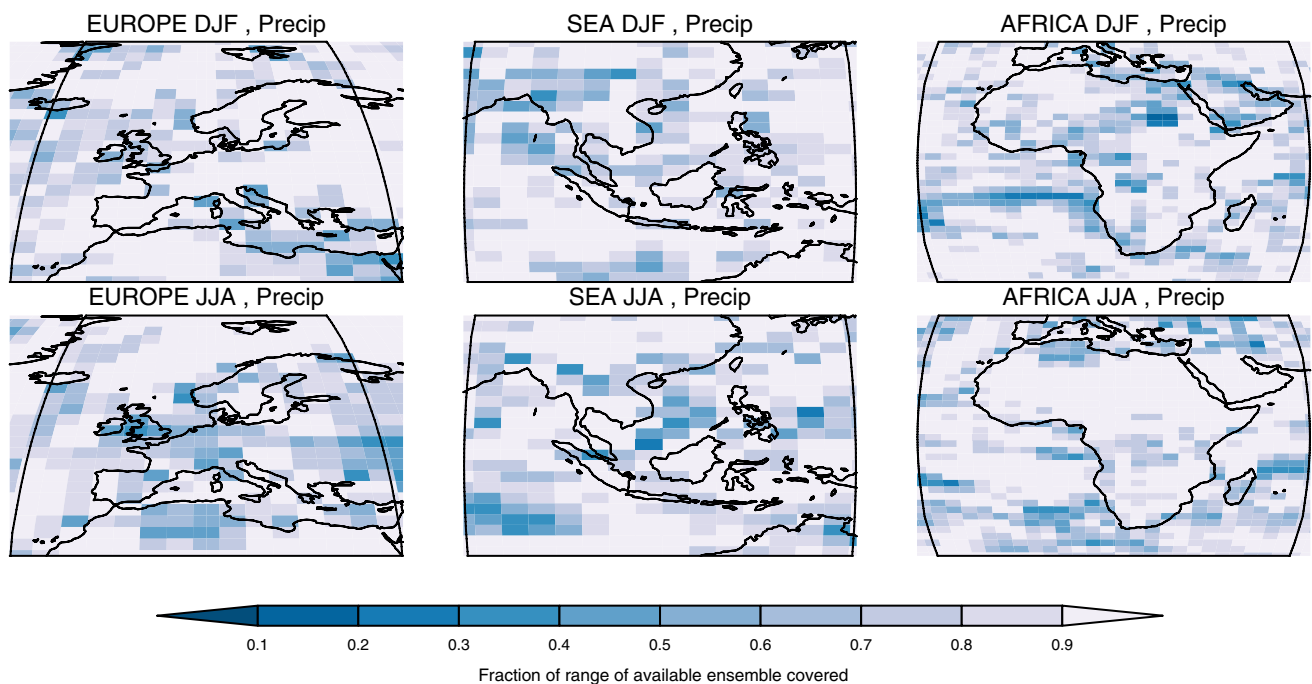


Fig. 14 Fraction of the range of projections in precipitation spanned by the selected 8-member subset across each of the 3 regions and seasons DJF and JJA

The criteria we have applied in selection has necessarily included a number of subjective (judgement-based) decisions. The use of metrics can provide an objective

source of assessing models, and help to reduce the volume of information that we assess by summarising important aspects of model behaviour in one or two indices. However,

difficulties in identifying meaningful indices mean that these metric-based approaches may not provide all of the information that might be useful and in data-sparse regions may not give a good guide to model performance. Further subjectivity is introduced in such approaches by the definition of thresholds for ‘good/bad’ models, or appropriate weighting schemes, as well as the choice of metric used. Here a combination of metrics and assessment by visual inspection of mapped fields has been used. Assessing by visual inspection allows us potentially to identify a wide range of characteristics of error, but introduces a high level of subjectivity. We manage this subjectivity by clearly justifying decisions with clear lines of evidence in order to make these decisions transparent.

Our assessment has also necessarily been limited to a restricted number of criteria as it is not feasible to undertake full assessments of all CMIP5 models. As an increasing body of literature on both the performance and projections of CMIP5 models emerges, others will be able to draw on a wider range of well documented evidence for selection. Efforts to gather and share well-documented metrics from CMIP5 models, such as those currently being undertaken by the WCRP Climate Model metrics Panel, could provide a valuable basis for the informed selections which are likely to become an increasingly important stage in the development of regional climate change projections.

Linking the baseline behaviour of a model to the credibility of its projections remains a key difficulty in considering elimination or downweighting of ensemble members. It is notable that those models which have a tendency to show less realistic behaviour are often those for which projections lie on the margins of, or outside the, range of the majority of the ensemble (for example, *MIROC-ESM*, *MIROC-ESM-CHEM*, *inmcm4*, *IPSL-CM5A-LR* and *BNU-ESM* are all models which were flagged with ‘implausible’ or ‘significant biases’ ratings). This methodology provides a useful mechanism for flagging these cases for further investigation, but there is clearly potential for deeper scientific investigation into the plausibility of these projections.

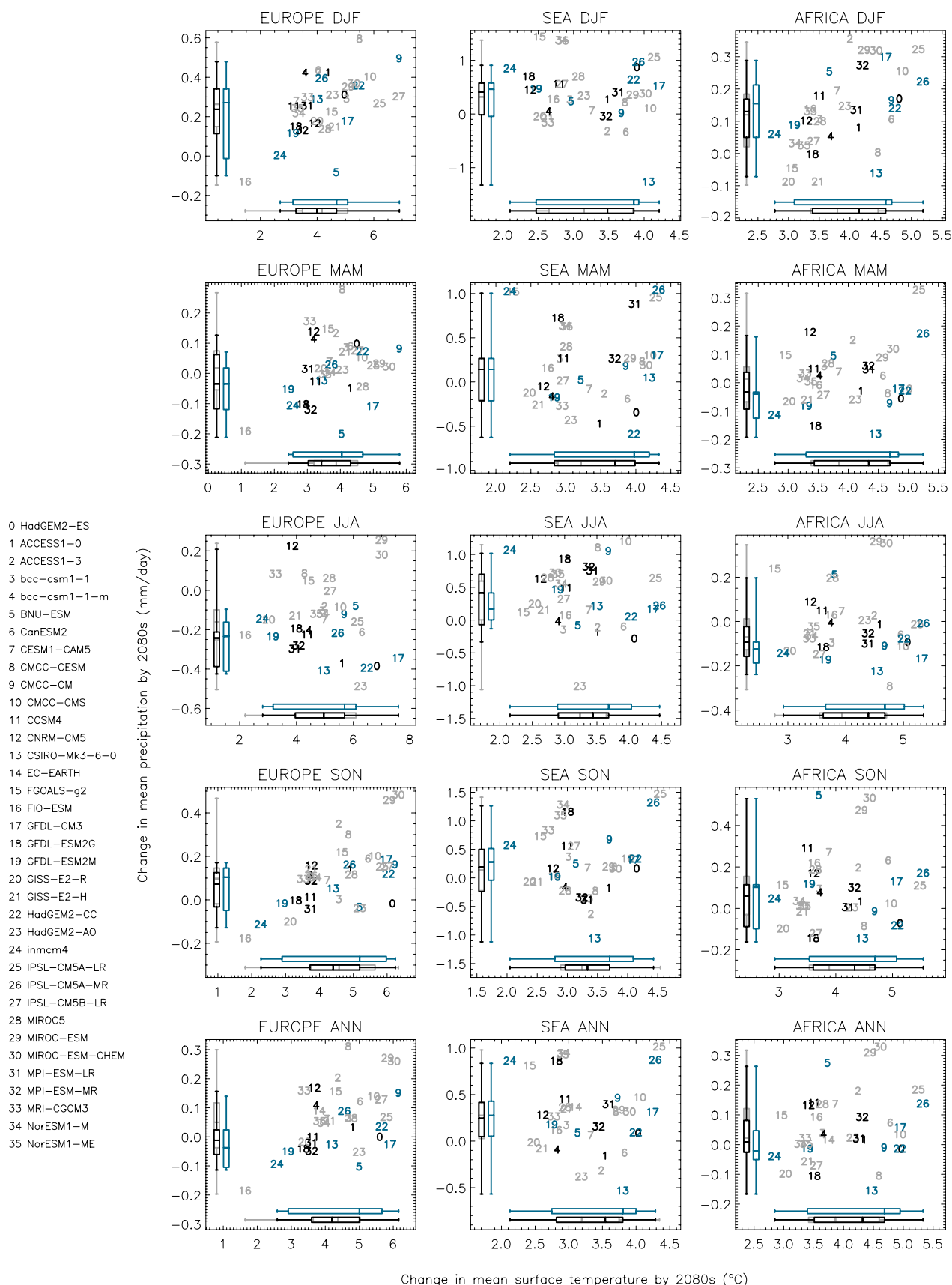
The process of systematically assessing models’ baseline behaviour before downscaling has the further benefit of providing very useful contextual information for interpretation and appropriate application of the projected changes for the region(s) of interest. We know that some aspects of climate are poorly represented in all models due to common errors—for example, Hung et al. (2013), find spatial characteristics of the Madden Julian Oscillation are poorly represented in all but one CMIP5 model, which mean that we can have only limited confidence in their representation of intra-seasonal variability for tropical regions affected by this feature. While common deficiencies such as this may not provide information that is useful for differentiating between models for selection, their identification as a part

Fig. 15 The range of regionally-averaged projections in precipitation and temperature spanned by the selected 8-member subset across each of the 3 regions and all seasons. *Blue* selected models, *Black* available models not selected, *Grey* models which were not considered for inclusion because they were either eliminated on grounds of performance or were not LBC-avail

of the selection process may lead to improved understanding of the limitations of all of the available climate projections in realistically representing changes in some of the more complex aspects of a region’s climate.

We have demonstrated how a selection approach might be applied to the identification of a set of GCMs which is suitable for use across multiple regions. We have shown that the identifying an ‘optimal’ subset to span changes in mean temperature and precipitation across three large continental-scale regions does not reduce substantially the proportion of the full range that we would hope to capture compared with selecting ‘optimal’ subsets for each region. However, the potential exclusion of some models based on very poor performance presents a more difficult problem for multi-region selection. The models *MIROC-ESM-CHEM* and *MIROC-ESM* were found to perform poorly enough in Europe and southeast Asia that we considered their projections for those regions as ‘implausible’. However, no such performance issues were found in Africa, and the exclusion of those models based on performance in other regions leads us exclude the projections with largest JJA rainfall increases in Africa from the subset. While there are clearly differences in the relative skill of models from one region to another, there are direct and indirect dependencies between phenomena from one region to the next which could justify a multi-region approach. A single region approach might overlook remote processes with indirect relevance. At the other end of the spatial scale spectrum, there may users of climate information who are interested in climate data at the single-grid-cell level. FRC is calculated and shown at the grid-box level in order to show the geographical variations within sub-regions. However, grid-scale information from either the GCM or downscaled GCM is of very limited value in isolation due to known errors in climate models in resolving processes at the model’s highest spatial resolutions (e.g. Masson and Knutti 2011).

The method described in this paper uses selection to address the problem of capturing GCM uncertainty which is known to be large, and in the case of precipitation, sometimes contradictory between models for some regions (Knutti and Sedlacek 2012), while downscaling with only a single RCM. Our approach does not account for all sources of uncertainty involved in projecting future climate. The choice of RCM of course represents a further important contribution to the range of climate outcomes in a region—for



example, for Africa the contribution of uncertainty by using multiple RCMs has been shown in some studies to be larger than that of multiple GCMs (Patricola and Cook 2010; Paeth et al. 2011). The development of approaches to the strategic selection of RCMs, the design of GCM-RCM combination matrices and the interpretation of projections generated via ‘ad-hoc’ combinations of different GCM-RCM pairs and statistical downscaling models are all issues in the design and interpretation of modelling experiments designed to generate regional climate information that require further development. Further exploration of these issues will be of great interest to those involved in generating regional climate projections for impacts and vulnerability applications, particularly those involved in the CORDEX (Giorgi et al. 2009) and ISIMIP (Warszawski et al. 2013) experiments.

Acknowledgments This work was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101). The work undertaken by DPR is part of the output from a project funded by the UK Department for International Development (DFID) for the benefit of developing countries. The views expressed are not necessarily those of DFID. 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 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, and to James Murphy for useful comments on the manuscript.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

- Adler RF et al (2003) The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present). *J Hydrometeorol* 4:1147–1167
- Giorgi F, Jones C, Asrar G (2009) Addressing climate information needs at the regional level: the CORDEX framework. *WMO Bull*, vol 58, World Meteorological Organization, Geneva, Switzerland, pp 175–183
- Greene AM, Goddard L, Lall U (2006) Probabilistic multimodel regional temperature change projections. *J Clim* 19:4326–4343
- Hodges KI, Lee RW, Bengtsson L (2011) A comparison of extratropical cyclones in recent reanalyses ERA-Interim, NASA MERRA, NCEP CFSR, and JRA-25. *J Clim* 24:4888–4906
- Hung M, Lin J, Wang W, Kim D, Shinoda T, Weaver S (2013) MJO and convectively coupled equatorial waves simulated by CMIP5 climate models. *J Clim*. doi:10.1175/JCLI-D-12-00541.1
- IPCC (2007) The physical science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, New York
- IPCC (2013) Climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, UK and New York, p 1535
- Knutti R (2010) The end of model democracy? *Clim Change* 102:395–404
- Knutti R, Sedlacek J (2012) Robustness and uncertainties in the new CMIP5 climate model projections. *Nat Clim Change* 3:369–373. doi:10.1038/nclimate1716
- Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA (2010) Challenges in combining projections from multiple climate models. *J Clim* 23:2739–2758
- Masson D, Knutti R (2011) Spatial-scale dependence of climate model performance in the CMIP3 ensemble. *J Clim* 24:2680–2692
- McSweeney CF, Jones RG, Booth BBB (2012) Selecting ensemble members to provide regional climate change information. *J Clim* 25:7100–7121
- Mitchell TD, Jones PD (2005) An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int J Climatol* 25(6):693–712
- Overland JE, Wang MY, Bond NA, Walsh JE, Kattsov VM, Chapman WL (2011) Considerations in the selection of global climate models for regional climate projections: the arctic as a case study. *J Clim* 24:1583–1597
- Paeth H, Hall NMJ, Gaertner MA, Alonso MD, Moumouni S, Polcher J, Ruti PM, Fink AH, Gosset M, Lebel T, Gaye AT, Rowell DP, Moufouma-Okia W, Jacob D, Rockel B, Giorgi F, Rummukainen M (2011) Progress in regional downscaling of West African precipitation. *Atmos Sci Lett* 12:75–82
- Patricola CM, Cook KH (2010) Northern African climate at the end of the twenty-first century: an integrated application of regional and global climate models. *Clim Dyn* 35:193–212
- Rowell DP (2013) Simulating large-scale teleconnections to Africa: what is the state of the art? *J Clim* 26:5397–5417
- Sexton DMH, Murphy JM, Collins M, Webb MJ (2012) Multivariate probabilistic projections using imperfect climate models part I: outline of methodology. *Clim Dyn* 38:2513–2542
- Sperber KR, H Annamalai, I-S Kang, A Kitoh, A Moise, A Turner B Wang T Zhou (2013) the asian summer monsoon: an intercomparison of CMIP5 vs. CMIP3 simulations of the late century. *Clim Dyn* 41(9–10):2711–2744
- Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. *Bull Am Meteorol Soc* 93:485–498
- Tebaldi C, Sanso B (2009) Joint projections of temperature and precipitation change from multiple climate models: a hierarchical Bayesian approach. *J R Stat Soc Ser Stat Soc* 172:83–106
- Tebaldi C, Smith RW, Nychka D, Mearns LO (2005) Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multi-model ensembles. *J Clim* 18:1524–1540
- Uppala SM et al (2006) The ERA-40 re-analysis. *Quart J R Meteorol Soc* 131:2961–3012
- Wang B, Fan Z (1999) Choice of south Asian summer monsoon indices. *Bull Am Meteorol Soc* 80:629–638
- Warszawski L et al (2013) The inter-sectoral impact model intercomparison project (ISI-MIP): project framework. *Proc Natl Acad Sci* 111(9):3228–3232. doi:10.1073/pnas.1312330110
- Watterson IG, Whetton PH (2011) Distributions of decadal means of temperature and precipitation change under global warming. *J Geophys Res Atmos* 116:13
- Woollings T (2010) Dynamical influences on European climate: an uncertain future. *Philos Trans R Soc Math Phys Eng Sci* 368:3733–3756
- Xie PP, Arkin PA (1997) Global precipitation: a 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bull Am Meteorol Soc* 78:2539–2558