Analyzing precipitation projections: A comparison of different approaches to climate model evaluation

N. Schaller,[1] I. Mahlstein,[1] J. Cermak,[1] and R. Knutti[1]

Received 27 August 2010; revised 2 March 2011; accepted 10 March 2011; published 27 May 2011.

[1] Complexity and resolution of global climate models are steadily increasing, yet the uncertainty of their projections remains large, particularly for precipitation. Given the impacts precipitation changes have on ecosystems, there is a need to reduce projection uncertainty by assessing the performance of climate models. A common way of evaluating models is to consider global maps of errors against observations for a range of variables. However, depending on the purpose, feature-based metrics defined on a regional scale and for one variable may be more suitable to identify the most accurate models. We compare three different ways of ranking the CMIP3 climate models: errors in a broad range of climate variables, errors in global field of precipitation, and regional features of modeled precipitation in areas where pronounced future changes are expected. The same analysis is performed for temperature to identify potential differences between variables. The multimodel mean is found to outperform all single models in the global field-based rankings but performs only averagely for the feature-based ranking. Selecting the best models for each metric reduces the absolute spread in projections. If anomalies are considered, the model spread is reduced in a few regions, while the uncertainty can be increased in others. We also demonstrate that the common attribution of a lack of model agreement in precipitation projections to different model physics may be misleading. Agreement is similarly poor within different ensemble members of the same model, indicating that the lack of robust trends can be attributed partly to a low signal-to-noise ratio.

Citation: Schaller, N., I. Mahlstein, J. Cermak, and R. Knutti (2011), Analyzing precipitation projections: A comparison of different approaches to climate model evaluation, *J. Geophys. Res.*, *116*, D10118, doi:10.1029/2010JD014963.

1. Introduction

[2] In the discussion on climate change, trends in the hydrological cycle are of particular interest since they are expected to have severe consequences for societies and ecosystems. End users of climate model output with an interest in hydrological changes therefore need information about the quality of the predictions. However, model disagreement about precipitation is large, in particular on a regional scale. Although climate models are getting constantly more complex, unambiguous statements about future changes in precipitation patterns are still difficult to provide [*Trenberth et al.*, 2003]. The aim of this study is to define new metrics to evaluate the ability of current climate models to simulate regional precipitation and to investigate if future projection uncertainty can be reduced when considering the best models in these regions.

[3] The literature available on the evaluation of climate models is broad and many ways of assessing model performances have been proposed. Although each individual method provides interesting information, so far no widely accepted suite of metrics to evaluate the performance of climate models exists for precipitation or any given climate variable in general [*Räisänen*, 2007; *Intergovernmental Panel on Climate Change*, 2007; *Knutti et al.*, 2010a]. Several studies [*Lambert and Boer*, 2001; *Reichler and Kim*, 2008; *Gleckler et al.*, 2008; *Pincus et al.*, 2008] evaluate the performance of climate models for a range of climate variables and on a global scale by using statistical measures to quantify the errors. *Reichler and Kim* [2008] ranked the climate models based on a single performance index, defined as the aggregated errors in simulating the observed climatological mean states of several climate variables. *Gleckler et al.* [2008] and *Pincus et al.* [2008] used straightforward statistical measures (e.g., root-mean-square error, correlation, bias or standard deviation) to evaluate models against observations on a global scale for given variables. All three studies conclude that the MultiModel Mean (hereafter MMM) shows better agreement with the observations than any single model.

[4] However, evaluations on a global scale summarized for many variables are not useful in some specific cases. A model performing well for a given variable, season and

[1]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland.

0148-0227/11/2010JD014963

region might perform poorly for another variable, season and region [*Whetton et al.*, 2007]. *Gleckler et al.* [2008] also stress the fact that in their evaluation, the relative merits of each model in simulating individual processes or variables are lost. *Gleckler et al.* [2008] and *Pincus et al.* [2008] further state that all models have distinctive weaknesses in simulating specific variables.

[5] A range of studies concentrated their evaluation on precipitation. As a response to anthropogenic forcing, temperature is expected to increase in all regions of the globe while precipitation is expected to increase in the tropics and high latitudes and decrease in the midlatitudes [*Allen and Ingram*, 2002]. The regional character of the expected changes suggests a need for a model evaluation on that scale. *Giorgi and Mearns* [2002] divided the area over land into regions and calculated for each a measure combining information on model performance and convergence. *Tebaldi et al.* [2004] performed an evaluation based also on the criteria model bias and model convergence. Both studies aim at reducing the uncertainty range for future regional precipitation by weighting the models according to the criteria mentioned above. However, no information on individual model performance is delivered, which would be of interest for end users of climate model output from other scientific communities. For example, precipitation projections are needed as input for hydrological models and the large model disagreement is an issue. Information about the quality of the predictions/simulations of each model might be a way out, although recent studies tend to show that a good performance during a given time period does not guarantee a good performance in a future time period [*Jun et al.*, 2008; *Knutti et al.*, 2010b].

[6] Finally, some studies concentrated on one region of interest. *Phillips and Gleckler* [2006] evaluated the ability of the models to simulate the seasonal cycle of precipitation globally and in certain regions. They show that while the MMM outperforms any single model at simulating continental precipitation on a global scale, in some regions, this is less clearly the case. *Pierce et al.* [2009] found that over the western United States and for a detection and attribution purpose, forming the MMM is a better way to make use of the information than selecting the best models. Contrary to this, *Perkins and Pitman* [2009] as well as *Smith and Chandler* [2010] see a reduction in future projection uncertainty by selecting the best models for precipitation for regions over Australia.

[7] *Knutti et al.* [2010b] showed that most statistical metrics like the root mean square error do not correlate strongly with future projections, and they suggest that feature-based evaluations could provide additional useful information. A feature-based metric considers regional changes that are robust and can be understood physically. This is different from the approach chosen by several regional studies cited above, where the regions were defined quite arbitrarily to partition the land part of the Earth. Here, we define such feature-based metrics and evaluate the models' ability to reproduce them compared to observations. This study consequently aims to provide information on the individual performance of the CMIP3 models for the present climate, as well as information about the persistence of these performance measures in a future climate. Further, the results obtained for precipitation are compared with the ones obtained for temperature in order to identify potential differences between variables.

[8] The data are briefly presented in section 2, while section 3 provides a pointwise evaluation of the modeled precipitation during the observation period. Section 4 investigates reasons for the lack of model agreement in the future. The definition of the metrics used for the model evaluation along with the results presented as a ranking are shown in section 5. The projections of the corresponding precipitation indices are discussed in section 6 and conclusions are provided in section 7.

2. Data

[9] The model simulations for precipitation and temperature used in this study stem from 24 of the global coupled atmosphere ocean general circulation models (AOGCMs) made available by the World Climate Research Program (WCRP) Coupled Models Intercomparison Program Phase 3 (CMIP3) [*Meehl et al.*, 2007a] (see http://www-pcmdi.llnl.gov/about/index.php for further information). One ensemble member of each model of the precipitation field from the simulation of the 20th century and the scenario A1B is used, and they are equally weighted for the multimodel mean. During the observation period (1979–2004), the models are evaluated against a merged product of precipitation with global coverage, the Global Precipitation Climatology Project (GPCP) Version-2 monthly precipitation analysis [*Adler et al.*, 2003]. Another global precipitation product exists and is used as a secondary reference data set, the Climate Prediction Center's (CPC) Merged Analysis of Precipitation (CMAP) [*Xie and Arkin*, 1998]. GPCP is the reference data set for the evaluation performed in section 5 because CMAP uses atoll data over oceans, which leads to artifacts in trends [*Yin et al.*, 2004]. As a comparison, the same evaluation of the CMIP3 models is performed for temperature. Here, the ERA40 reanalysis data set [*Uppala et al.*, 2005] is used as reference for the time period 1979–2001.

3. Pointwise Evaluation

[10] Figure 1 summarizes the modeled and observed precipitation mean values as well as the bias of the MMM for boreal winter and summer. The mean precipitation of the MMM cannot be compared directly with mean precipitation of GPCP since the former is an average of multiple realizations and the latter represents only one realization. However, the main features of the precipitation patterns are captured by the MMM but with errors in their amplitude and exact location. The Spearman rank correlation coefficients between the MMM and GPCP are highly significant, $\rho = 0.9$ in DJF and $\rho = 0.89$ in JJA. However, the bias of the MMM, expressed in percent compared to the mean values of GPCP, is in general large over both oceans and land. Reasons for that are probably a combination of model errors and observational uncertainties plus a contribution of internal variability. As a comparison, the bias of the CMAP data set with respect to the GPCP data set is nonnegligible and in some regions, on the same order of magnitude as the one of the MMM [*Yin et al.*, 2004].

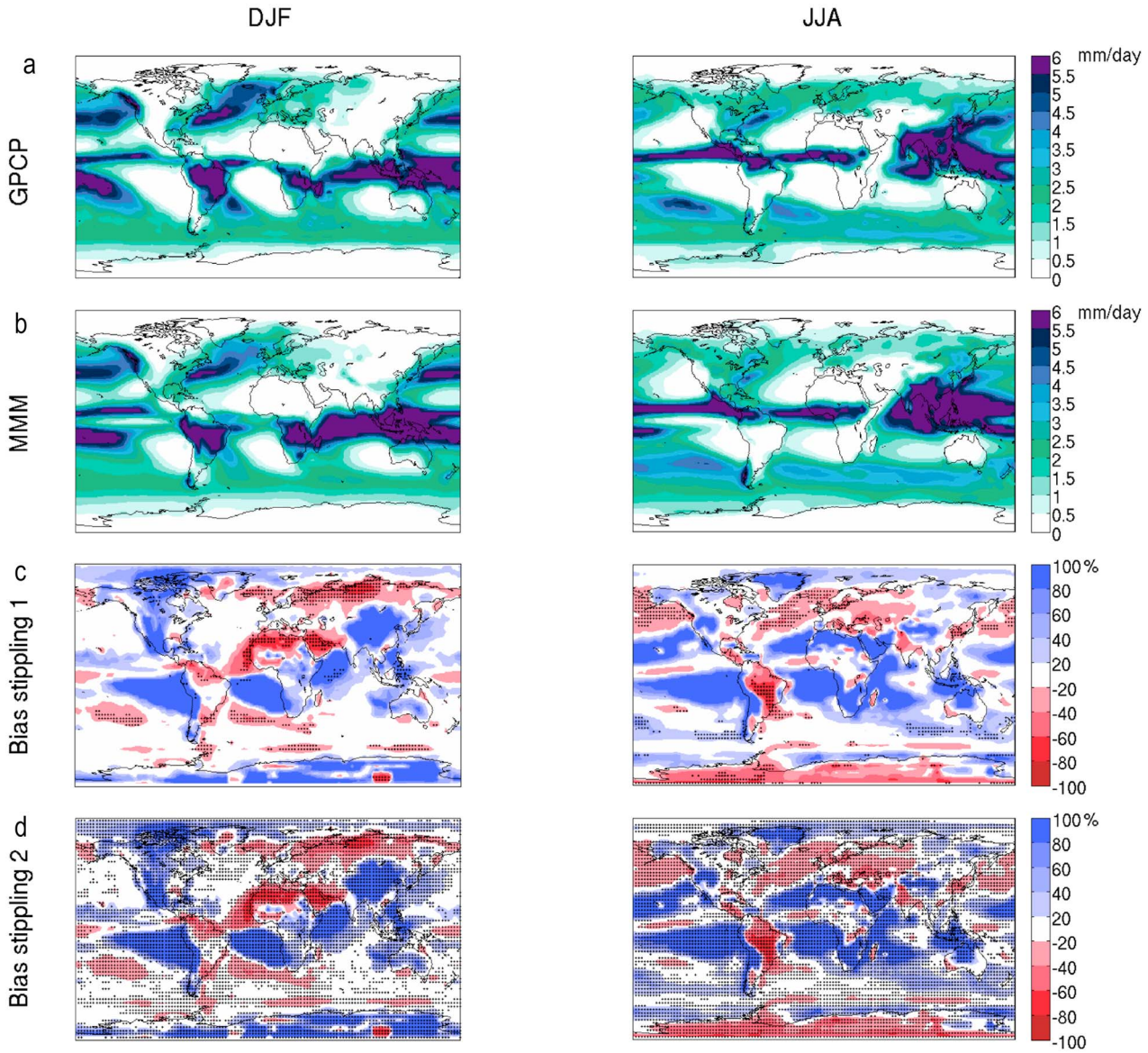

Figure 1. (a) Mean precipitation (1979–2004) values during December to February (DJF) and June to August (JJA) in GPCP. (b) Mean precipitation values (1979–2004) during DJF and JJA for the MMM. (c) Bias of the MMM in percent compared to GPCP (1979–2004) during DJF and JJA. Grid points are stippled when GPCP lies outside 2 standard deviations of the CMIP3 models. (d) Bias of the MMM in percent compared to GPCP (1979–2004) during DJF and JJA. Grid points are stippled when the bias is larger than twice the internal variability estimated from those CMIP3 models that provided at least 4 ensemble members (see text).

[11] The biases are shown twice on the third and fourth rows of Figure 1, each time with a different criterion for stippling. In the third row of Figure 1, grid points are stippled where the observations lie outside ±2 standard deviations of the 24 CMIP3 models ("Bias stippling 1"). For those grid points, the biases are larger than one would expect given the internal variability of the models and their structural differences. The stippled area is 6.1% of the globe in DJF and 6.7% in JJA, which is only slightly more than what one would expect to occur by chance with the criteria of 2 standard deviations. This outcome tends to indicate that the observations are indistinguishable from the models (see discussion in section 6.2) given the large model errors. The criterion for stippling in the fourth row of Figure 1 is defined from an estimate of the internal variability of the models for the period 1979–2004. The CMIP3 models that have more than 4 ensemble members (i.e., CGCM3.1(T47), CCSM3, ECHAM5/MPI-OM, MRI-CGCM2.3.2 and PCM) are selected, and for each of these models, the standard deviation of their ensemble members is calculated. Then the average of these standard deviations is used as a measure of internal variability. Grid points are stippled where the absolute value of the bias is at least twice as large as this average standard deviation ("Bias stippling 2"). The fact that

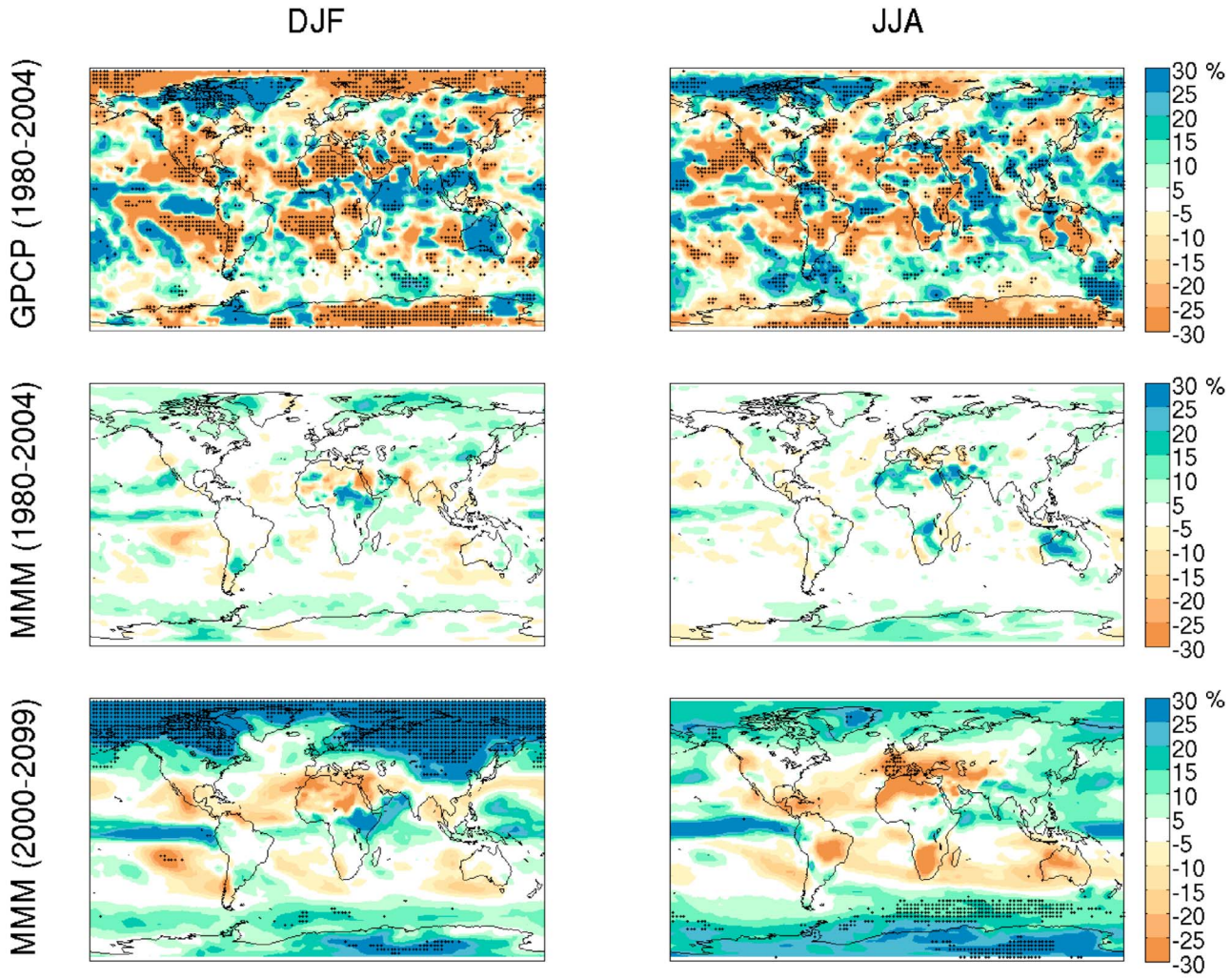

Figure 2. (top) Precipitation trends (1980–2004) during December to February (DJF) and June to August (JJA) in GPCP. Grid points are stippled when the trends are significant at the 95% confidence level. (middle) Precipitation trends (1980–2004) during DJF and JJA for the MMM. (bottom) Precipitation trends (2000–2099) during DJF and JJA for the MMM. For MMM panels, grid points are stippled if at least 18 out of 24 models agree on a significant trend (95% confidence level) with the same sign. Note that the variability in the MMM panels is strongly reduced compared to observations due to the averaging of many ensemble members.

72.5% in DJF and 82.6% in JJA of the whole globe is stippled according to this criterion implies that the observations are inconsistent in many areas with respect to the modeled range of natural internal variability, either because of observational errors, model errors or because the models underestimate internal variability.

[12] The observed and modeled precipitation trends for the observation period (1980–2004) as well as the modeled precipitation trends for a 100 year time period in the future (2000–2099) are represented in Figure 2. In the observations (GPCP), many small-scale structures in the trends can be seen and finding a physical explanation for them is not obvious. Significant drying (at the 95% confidence level) seems to dominate in the polar regions as well as over the west coasts of the continents while significant wettening is mostly located over Greenland, the Northern Territories and in the Indian ocean. The trends of the MMM from 1980 to 2004 are shown in Figure 2 (middle). If 18 out of the 24 CMIP3 models agree on the sign of significant change, the grid point is stippled, which is never the case during the observation period. On the one hand, this criterion minimizes the possibility that models have the same sign of the trend just by chance and on the other hand, it is not too stringent to prevent that the criterion is never met, which would not be very informative in the case of future trends. For the MMM, a weak wettening of the high latitudes and the equatorial region can be recognized, while some regions in the midlatitudes experience a slight drying but these features are not robustly simulated by the models. Again, the trends in precipitation of the MMM are not expected to agree perfectly with the trends of GPCP for the same reasons described above. The amplitude in the trend patterns is smaller for the MMM than in the observations due to the fact that natural variability is reduced in the MMM because

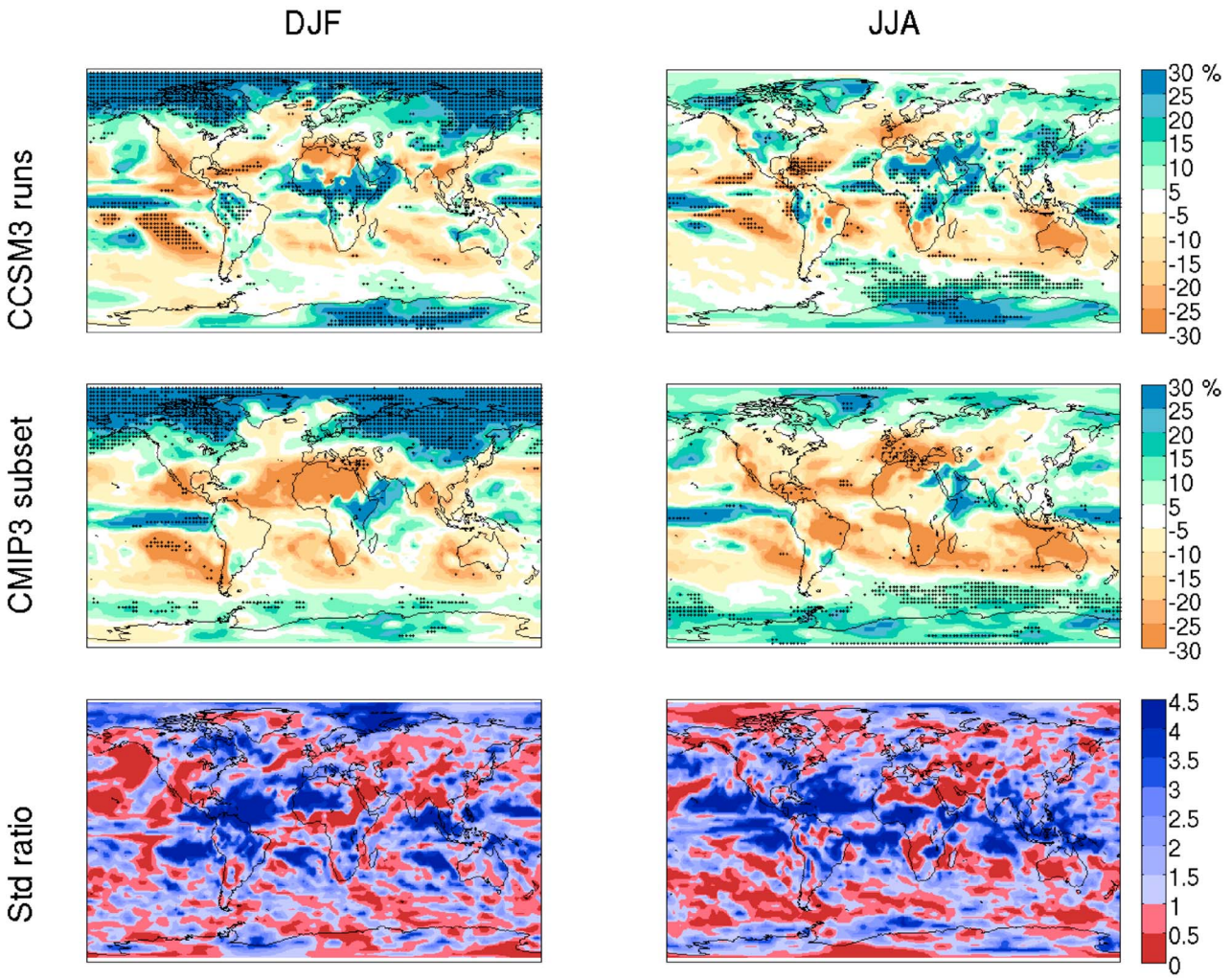

Figure 3. (top) Trends of the seven runs of CCSM averaged (2000–2099) for December to February (DJF) and June to August (JJA). Grid points are stippled if at least six out of seven runs agree on a significant trend with the same sign. (middle) Trends of seven CMIP3 models averaged together (2000–2099) for DJF and JJA. Grid points are stippled if at least six out of seven models agree on a significant trend with the same sign. (bottom) The ratio of the standard deviations of the seven CMIP3 models and the seven runs of CCSM.

of model averaging [*Räisänen*, 2007]. Possible reasons for a discrepancy between observed and modeled precipitation trends can be many fold: low signal-to-noise ratio, observation uncertainties, inadequate parameterizations in the models as well as incomplete representation of the forcings or too low spatial resolution. In addition, it must be stressed that the precipitation trends are nonsignificant in many regions for GPCP and the CMIP3 models, which indicates that at such short time scales, natural variability dominates.

[13] For the future time period, a wettening of the high latitudes and of the equatorial region, along with a drying of the midlatitudes can be recognized. Model agreement (stippling if 18 out of the 24 CMIP3 models agree on the sign of significant change) is generally confined to the high latitudes during the cold season. It is further interesting to note that the drying is a less robust feature than the wettening. The agreement criterion is rarely reached in regions were a drying is expected because there, variability is large and mean precipitation low which leads to highly variable percentage changes in the CMIP3 models. However, the chosen agreement criterion is severe, and robust precipitation changes can still be expected in areas where there is no stippling in the bottom row of Figure 2 [see *Meehl et al.*, 2007b, Figure 10.9] as an alternative criterion for model agreement).

4. Model Agreement

[14] As shown in section 3, large uncertainties are associated with changes in the precipitation patterns in a warmer climate since model agreement is poor compared to temperature for example. Spread in projections is caused by the differences between the models and by internal variability. To investigate whether the poor model agreement is actually due to differences between the models or rather caused by the nature of precipitation itself, the trends from 2000 to 2099 of the 7 available runs of the CCSM3 model and of a subset of 7 reasonably independent CMIP3 models

Table 1. Definition of the Precipitation Indices[a]

Index Name	Definition	Domain
African index	$AFI = \overline{Pr}_{JJA}$	13°S–35°S, 14°E–42°E
Amazonian index	$AMI = \overline{Pr}_{JJA}$	24°S–1°N, 31°W–59°W, land only
Asian index	$ASI = \overline{Pr}_{JJA} - \overline{Pr}_{DJF}$	10°N–29°N, 70°E–118°E
Australian index	$AUI = \overline{Pr}_{JJA}$	10°S–40°S, 107°E–138°E
Central American index	$CAI = \overline{Pr}_{JJA}$	10°N–29°N, 110°W–62°W
High-latitudes index	$HLI = \overline{Pr}_{DJF}$	52°N–71°N, land only
Mediterranean index	$MEI = \overline{Pr}_{JJA}$	29°N–49°N, 11°W–37°E
Storm tracks index	$STI = \overline{Pr}_{DJFB} - \overline{Pr}_{DJFA}$	zone A (35°S–46°S)/zone B (49°S–60°S)

[a] $\overline{Pr}_{DJF}$ ($\overline{Pr}_{JJA}$) denotes the mean precipitation during DJF (JJA) in the corresponding domain.

(CGCM3.1(T47), CSIRO-Mk3.5, GFDL-CM2.1, INGV-SXG, MIROC3.2(medres), ECHAM5/MPI-OM, MRI-CGCM2.3.2) are computed (see Figure 3). The conclusions however do not depend on the exact choice of the subset but likely hold for all possible subsets. While the exact location of spatial patterns of significant precipitation change are slightly different between the 7 CCSM3 runs and the 7 CMIP3 models, the wettening of the high latitudes and the equatorial region along with drying in some areas in the midlatitudes are captured by both. Again, a model agreement criterion is defined: grid points are stippled if at least 6 out of 7 runs/models agree on a significant sign of change. The percentage of area stippled is larger in the 7 CCSM3 runs (13.6% in DJF and 11.8% in JJA) compared to the 7 CMIP3 models (9% in DJF and 5.8% in JJA), as can be expected. Nevertheless, the area stippled for the 7 CCSM3 runs is surprisingly small and still in the same range as for the 7 CMIP3 models. This result indicates that even if the uncertainty caused by model differences is eliminated, internal variability still contributes strongly to the lack of agreement in precipitation projections.

[15] The relative importance of internal variability compared to model differences can be further quantified. The ratio of the standard deviations of the CMIP3 subset compared to the CCSM runs is computed in Figure 3 (bottom). The global average of this ratio is roughly 3 (2.7 in DJF and 2.95 in JJA), meaning that the contribution of model differences is around 3 times larger than that of internal variability in terms of standard variation. While model differences dominate, this does not imply that reducing model uncertainty in future projections will necessarily improve the significance of the projected trends. For single grid points where variability is large, the signal may not be significant even in a perfect model.

5. Ranking

5.1. Method

[16] In this section, the climate models are evaluated on a regional scale using feature-based metrics. These metrics are designed to focus on areas that reveal a clear signal of change in precipitation over the time period considered. This has also been the motivation, at least to some extent, of previous studies [*Pitman et al.*, 2004; *Pierce et al.*, 2009; *Perkins and Pitman*, 2009] but with the difference that they concentrated only on one region of interest. Here the aim is to go one step further by defining metrics in different regions of the globe over land and ocean parts and to compare the performance of the individual models using several feature-based metrics.

[17] The selected features are regions where the predicted precipitation change is robust. They are identified with the map of the future trends in precipitation of the MMM discussed in section 4 for two different seasons (DJF and JJA; see Figure 2, bottom). Eight metrics, which we refer to as precipitation indices (see Table 1 for definitions) are chosen, based on the significance of the trends and on the scientific understanding of the physical processes responsible for these changes. It is important to emphasize that the eight precipitation indices have to be regarded as examples and not as the only set of feature-based metrics possible.

[18] The eight precipitation indices are defined in Table 1. The storm tracks index (STI) is designed to detect the poleward shift of the storm tracks in the Southern Hemisphere, where zone A refers to the preferred region of cyclone activity in the past and zone B to the region where the storms are expected to pass by in the future [*Hoskins and Hodges*, 2005; *Previdi and Liepert*, 2007]. The African index (AFI) and the Australian index (AUI) capture the precipitation decrease over the midlatitudes of the Southern Hemisphere that are related to the positive trend of the Southern Annular Mode (SAM) index prevailing since the climate shift of the mid-1970s [*Thompson and Solomon*, 2002]. The Asian index (ASI) depicts the expected decrease of precipitation during the dry season and the increase of precipitation during the wet season in Southeast Asia. In the context of global warming, more warming over land than over the ocean is expected leading to a northward shift of the lower tropospheric monsoon circulation and consequently to an increase in mean precipitation during the Asian summer monsoon [*Dairaku and Emori*, 2006; *Sun and Ding*, 2010]. Changes in the location of the ITCZ are also expected to reduce precipitation during June, July and August (JJA) over the Amazon Basin (the Amazonian index, AMI) [*Christensen et al.*, 2007]. For the Northern Hemisphere, the high-latitudes index (HLI) captures the increase in precipitation during December, January and February (DJF) over the continents [*Previdi and Liepert*, 2007]. In a warmer climate moisture convergence toward the convection zones will increase and as a consequence, moisture divergence in the midlatitudes will be enhanced, causing a decrease in precipitation [*Neelin et al.*, 2006]. The most prominent features of this subtropical/lower midlatitude drying in the Northern hemisphere are the JJA precipitation decrease over the Caribbean/Central American region (captured by the Central American index, CAI) and the one over the Mediterranean region captured by the Mediterranean index (MEI), which is also associated with the soil moisture feedback over land [*Rowell and Jones*, 2006; *Seneviratne et al.*, 2006].

Table 2. Definition of the Temperature Indices[a]

Index Name	Definition	Domain
African index	$AFI = \overline{T}_{JJA}$	13°S–35°S, 14°E–42°E
Amazonian index	$AMI = \overline{T}_{JJA}$	24°S–1°N, 31°W–59°W, land only
Asian index	$ASI = \overline{T}_{JJA}$	10°N–29°N, 70°E–118°E
Australian index	$AUI = \overline{T}_{JJA}$	10°S–40°S, 107°E–138°E
Central American index	$CAI = \overline{T}_{JJA}$	10°N–29°N, 110°W–62°W
High-latitudes index	$HLI = \overline{T}_{DJF}$	52°N–71°N, land only
Mediterranean index	$MEI = \overline{T}_{JJA}$	29°N–49°N, 11°W–37°E
Storm tracks index	$STI = \overline{T}_{DJFAB}$	zone AB (35°S–60°S)

[a] $\overline{T}_{DJF}$ ($\overline{T}_{JJA}$) denotes the mean temperature during DJF (JJA) in the corresponding domain.

[19] Due to the heterogeneity of primary data, the quality of the merged gauge-satellite monthly precipitation products, GPCP and CMAP, cannot be expected to be equally good in the 8 regions where the feature-based metrics are defined. In general, GPCP and CMAP are more similar over land than over oceans, simply due to the availability of gauge measurements [*Yin et al.*, 2004]. Consequently, the quality of the observation data sets is expected to be better for precipitation indices defined mainly over land, like the AFI, AMI AUI and MEI. As can be seen on Figure 5, GPCP is slightly different from CMAP for the ASI and CAI since these indices are mainly defined over oceans. The largest differences between both data set are however encountered for the two metrics defined in the high latitudes, hence the HLI and STI, where both data sets use different input data [*Yin et al.*, 2004]. Despite the inherent uncertainties, the GPCP and CMAP data sets can be regarded as best estimate data sets of precipitation patterns.

[20] The precipitation indices allow to identify whether some models clearly perform better than average in regions where significant changes are expected and where the physical processes responsible for the changes are thought to be understood. The spatial pattern of precipitation within a region is however not evaluated. It is further interesting to investigate if the good models in a given region also perform well in this region but for another variable [*Whetton et al.*, 2007]. We therefore compare the results obtained for precipitation with temperature. Temperature is chosen because its signal of change does not strongly depend on the region considered and the field is relatively homogeneous. For this reason, temperature indices can be defined in the same region and for the same season as the precipitation indices and still be meaningful. The ASI and STI are exceptions because they describe processes that exist for precipitation but not for temperature. The ASI and STI for temperature are therefore simply defined as a temperature average for one season (see Table 2).

[21] The eight index trends of the MMM from 1980 to 2079 are significant on the 0.01 level for both variables. Unfortunately, the observational period is short and in case of precipitation, trends are not significant as discussed in section 3, making an evaluation of the trends meaningless. Consequently, the CMIP3 models and the MMM are ranked according to their ability to simulate the mean value of each index during the observation period (precipitation index ranking and temperature index ranking hereafter). The errors of each model at simulating the mean index value are simply calculated as difference between the observed index mean and the modeled index mean and do not include information about discrepancies in the spatial structure within the index domain. To compare and aggregate these performance metrics, they are converted to a common ranking system. A rank of 1 is attributed to the model with the smallest error on the metric considered, a rank of 2 to the second-best model, etc. While some quantitative information is lost in this ranking method, it has the advantage that indices with different scales and units can readily be compared in an aggregated form. Finally, to summarize the performances of the models over the eight indices, the ranks obtained for each of the eight indices are summed up, and this sum is ranked again (lines "Prec ALL" and "Temp ALL" in Figure 4). The sum of the ranks for the precipitation indices and the temperature indices is finally ranked in Figure 4c ("indices"). The motivation for summing the ranks over different regions and variables is to test if the index-based results gradually converge to the widely used broad-brush metrics that summarize performances for a large range of climate variables on a global scale.

[22] The index ranking is first compared to a ranking performed on global scale, again for both variables, precipitation and temperature, only. The root mean square error (rmse) of each model with respect to the observations and the spatial correlation between simulated and observed precipitation and temperature (referred to as the rmse/corr ranking hereafter) are calculated separately for each variable. The model having the lowest rmse (highest correlation coefficient) ranks first. In Figure 4c, the "rmse/corr" depicts a ranking of the sum of the ranks obtained for both variables on the rmse and the corr ranking, again to identify if by doing so, the outcomes of the broad-brush metrics can be reproduced.

[23] Finally, the index and the rmse/corr rankings are compared to a ranking performed with a broad-brush metric, which is a version of the ranking on a broad range of climate variables performed by *Reichler and Kim* [2008] (RK08 ranking hereafter), updated with more variables and using four seasons (T. Reichler, personal communication, 2009).

[24] In summary, the RK08 ranking identifies the model performance on a global scale summarized for different climate variables, the rmse/corr ranking provides a picture of the models' spatial error with respect to the precipitation and temperature data on a global scale and the index ranking allows for the identification of the models that best simulate local precipitation and temperature features expected to change in the future due to anthropogenic forcing.

5.2. Results and Discussion

[25] The results of the index rankings for both precipitation and temperature are summarized in Figure 4a. At first glance, none of the CMIP3 models appears to consistently outperform the rest. This is particularly obvious in the precipitation index ranking. Here, each model performs at least once better and worse than the average, while the MMM performs average for all indices. The results for the temperature index ranking are only slightly different: again the models can perform better and worse than average for different indices, except ECHAM5/MPI-OM and the MMM, which always perform better than average. It is further interesting to note that there are no significant correlations among the eight indices for each variable nor between

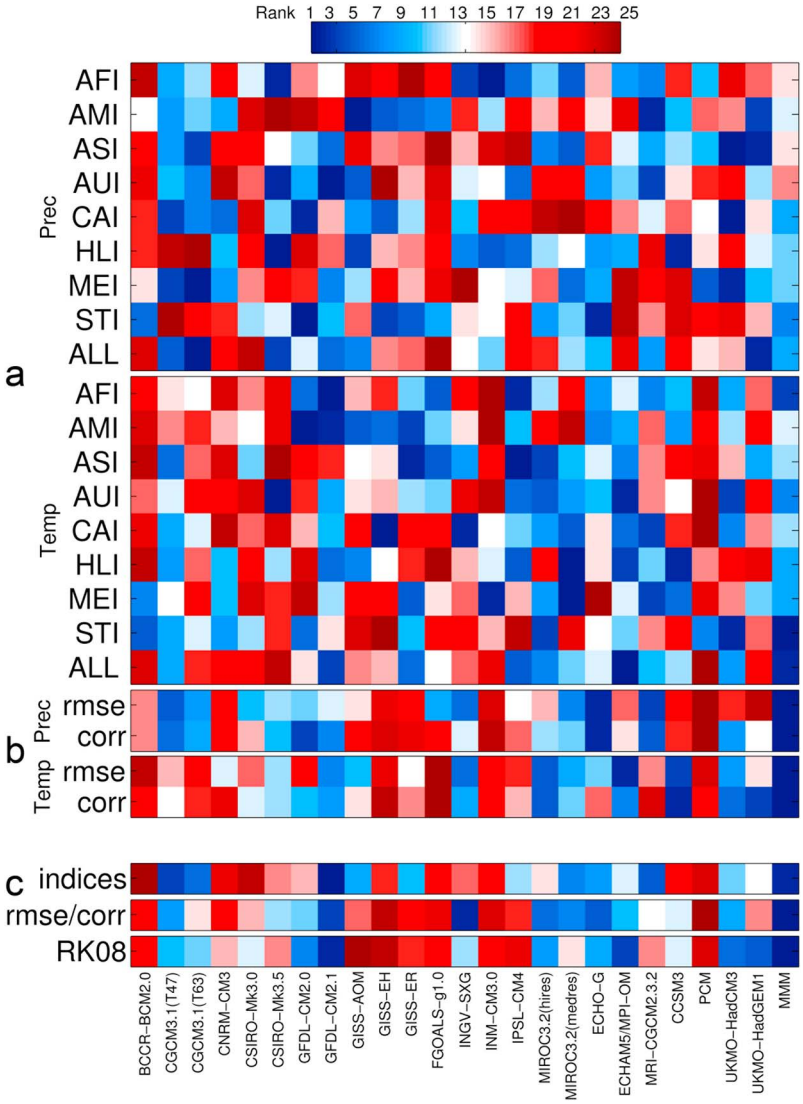

Figure 4. (a) Ranks obtained by the CMIP3 models for the eight (top) precipitation and (bottom) temperature indices, where the "ALL" line summarizes the ranks obtained for all eight precipitation and temperature indices. (b) The ranks obtained by the CMIP3 models for the rmse/corr ranking for (top) precipitation and (bottom) temperature. (c) Summary of the performance of the CMIP3 models for the indices and rmse/corr ranking for both precipitation and temperature as well as an updated version of the model ranking performed by *Reichler and Kim* [2008]. Blue and red indicate above- and below-average performance, respectively.

precipitation and temperature for each index (not shown). The performances of each model are summarized in the lines Prec ALL and Temp ALL (see Figure 4). For Prec and Temp ALL, the MMM ranks eleventh and third, respectively.

[26] For the interpretation of the results, it is important to keep in mind that even though in evaluation studies the MMM is often considered as just another model, it is actually not. By definition, the MMM can perform only from average up to best but cannot be worse than average, while each individual model can occupy any place from the worst up to the best (see also Figures 5 and 6). The results depicted in the lines Prec ALL and Temp ALL illustrate the fact that the more indices are included, the better the performance of the MMM. This outcome is similar to the findings of *Pierce et al.* [2009]. This continuously improving performance of the MMM is partly due to the fact that it never performs below average, in contrast to the individual models. However, this does not mean that the MMM is better at simulating individual index mean values, but is rather an artefact that arises when more indices are considered. The MMM never has to compensate for a below-average performance plus it is favored by the fact that there is no correlation between the different indices for both variables. The above-average models are therefore difficult to identify because when considering regional features for a given variable, none of the CMIP3 models consistently outperforms the rest.

[27] Figure 4 also shows the results of the ranking performed on the global scale for precipitation and temperature, termed as the rmse/corr ranking. Here, and in agreement

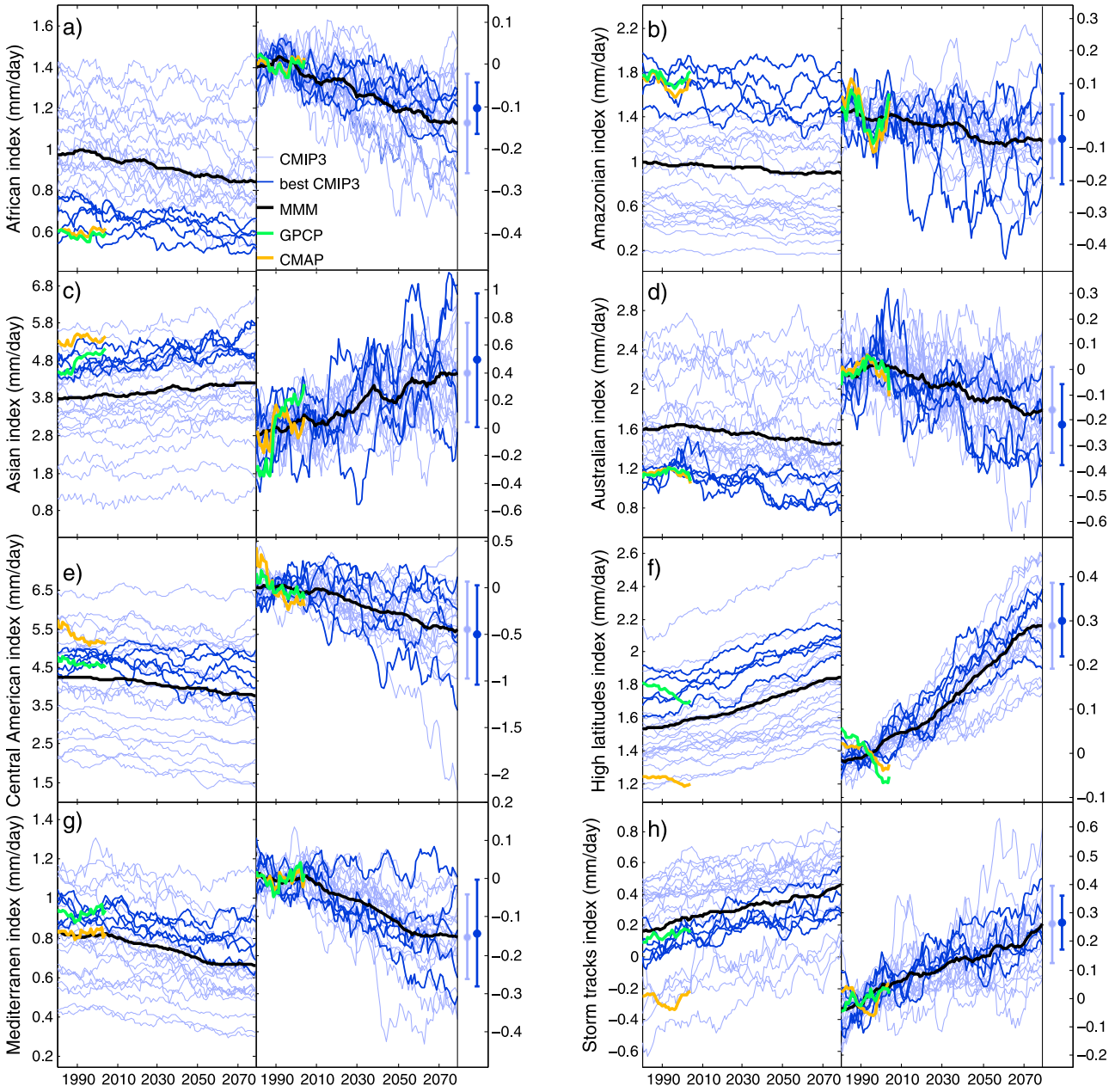

Figure 5. (a–h) (left) Time series and (right) anomalies relative to the observational period of the CMIP3 models, CMAP, and GPCP for the eight precipitation indices. The five best models for each index are given as dark blue lines, while the black line represents the MMM. An 11 year average is applied to the time series. The mean value and ±1 standard deviation in the year 2079 are shown in light blue (dark blue) for all CMIP3 models (five best models).

with previous studies [*Phillips and Gleckler*, 2006; *Gleckler et al.*, 2008; *Pincus et al.*, 2008], the MMM clearly performs best for both variables. Averaging the individual models smoothes out variations and small-scale biases of the precipitation field, so that errors partly cancel out in the MMM [*Phillips and Gleckler*, 2006; *Pierce et al.*, 2009]. Consequently, the MMM is favored by global statistical metrics because of its relatively small magnitude of biases over the whole globe and its good representation of the spatial pattern, while feature-based metrics favor a single model capable of displaying the area mean precipitation over a given region. Except for the performance of the MMM, the rmse/corr ranking is quite different for precipitation and temperature, illustrating that also globally, a model performing well for a variable might perform poorly for an other. Further, it is interesting to compare the index ranking with the rmse/corr ranking for each variable individually. The Spearman's rank correlation coefficient between Prec ALL and Prec rmse is nonsignificant at the 95% level while it is significant between Prec ALL and Prec corr ($\rho = 0.49$). For temperature, the correlations are also low but significant: $\rho = 0.56$ between Temp ALL and Temp rmse and

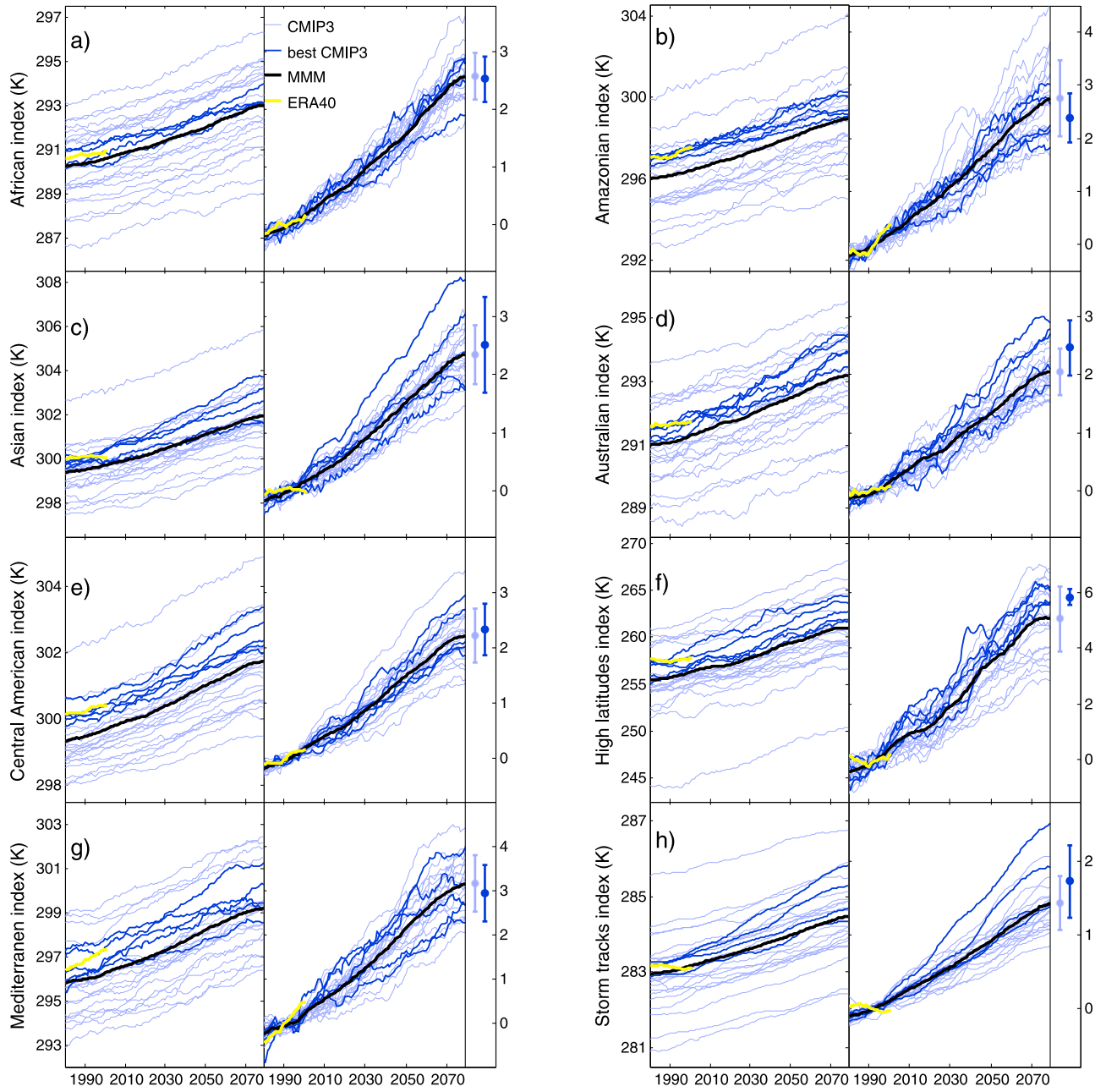

Figure 6. (a–h) (left) Time series and (right) anomalies relative to the observational period of the CMIP3 models and ERA-40 for the eight temperature indices. The five best models for each index are given as dark blue lines, while the black line represents the MMM. An 11 year average is applied to the time series. The mean value and ±1 standard deviation in the year 2079 are shown in light blue (dark blue) for all CMIP3 models (five best models).

ρ = 0.4 between Temp ALL and Temp corr. This indicates that when the performances of the individual models on a few chosen regional features are summarized, the results of a ranking based on the rmse or correlation on a global scale can be approached.

[28] Finally, the errors obtained for all precipitation and temperature indices are summed to obtain the line "indices" on Figure 4c, which can be compared with the rmse/corr ranking and the broad-brush metric ranking RK08 (Figure 4c). In this final index ranking, the MMM ranks fifth and it is reasonable to assume that including more regional features and/or more variables will contribute to improve the rank of the MMM, which will eventually rank first. It is obvious that the MMM ranks first for the rmse/cor ranking since it was already the case for the rmse and corr ranking of each individual variable. The MMM also ranks first in the RK08 ranking for the same reason. By definition the error of the MMM at each grid point cannot be larger than the mean error of the models and consequently, the errors of the MMM are the smallest when averaged globally. Nevertheless, the three ways of ranking presented here share similarities. The Spearman's rank correlations are significant

between the three rows in Figure 4c: $\rho = 0.44$ between "indices" and "RK08," $\rho = 0.61$ between "indices" and "rmse/corr" and $\rho = 0.78$ between "RK08" and "rmse/corr." This shows that when summarizing the errors at simulating the mean of several feature-based metrics for different variables, the performance of the individual models is partly the same as when evaluating the models with measures of the global spatial distribution and including more (e.g., RK08 ranking) or less (e.g., rmse/corr ranking) climate variables.

[29] To summarize, an evaluation of the models using global statistical measures like the RK08 ranking does not capture the average performance of the MMM at simulating the mean precipitation amounts in a given region. Such evaluation techniques rather reflect the fact that the MMM has the smallest errors as soon as the domain size exceeds several grid points because it cannot have per definition the maximal error on a grid point (in contrary to the individual models). When the ranks obtained for the eight precipitation and temperature indices are summed up and ranked again, a similar result is seen: the MMM does not have to compensate for poor rankings and the performance of the MMM becomes gradually better the more regions and variables are summed up. The information that single models are better than the MMM at simulating regional mean precipitation amounts for a given season can be relevant for impact studies but is hidden in evaluations using a global broad-brush approach. In addition, the results obtained for temperature suggests that the worse performance of the MMM for the feature-based metrics compared to global summary statistics is not a particularity of precipitation but is likely to hold for most variables. The interpretation of the MMM is further discussed at the end of section 6.

6. Future Projections

6.1. Method

[30] Once the models performing best for a given regional feature are identified, the question arises whether these models will still be the best performing ones in the future. The assumption that the models simulating the present climate accurately will also simulate well the future climate is often made [e.g., *Tebaldi et al.*, 2004]. While it is impossible for obvious reasons to perform a model evaluation with feature-based metrics for the future to check if this assumption is correct, investigating the convergence of the models on future predictions can partly answer this question. If a subset of models (chosen based on agreement with observations) shows considerably smaller spread, then the observations can be regarded as useful to distinguish between models. This is equivalent to a correlation between biases in the present-day simulation and the predicted change. The assumption is that such correlation is not just an artifact of all models making similar assumptions, but rather that it reflects an underlying physical process or feedback that influences both the base state of a model as well as the simulated change. In practice such correlations unfortunately are relatively low in many cases [*Knutti et al.*, 2010b; *Whetton et al.*, 2007], probably partly as a consequence of the observations being used already in the model development process.

[31] The time evolution of the absolute values and the anomalies of the eight precipitation and temperature indices for the 100 year period 1980–2079 is shown in Figures 5 and 6 using an 11 year average. For the precipitation indices, the time series of GPCP and CMAP for 1980–2004 are also presented. Similarly, the index time series of ERA-40 are shown besides the modeled index time series of temperature. In addition, for each variable and each index, the five models performing best are identified and represented by dark blue lines. The model spread by the end of 2079 for all models and the five best models is represented by an error bar at the right of each panel. The error bars represent the mean value ±1 standard deviation.

6.2. Results and Discussion

[32] Figure 5 shows the modeled absolute and anomaly time series of each precipitation index from 1980–2079 along with the observed absolute and anomaly time series from 1980–2004. The model spread is large for all indices. For example, in case of the absolute values of the ASI, the projections vary by a factor of 5, hence the difficulty to give clear statements about future precipitation amounts in this region. The reason for the average performance of the MMM in the index ranking presented above becomes evident by looking at the time series. The MMM by definition lies in the middle of the model spread, while the observation data sets lie in most indices at one end of the model spread. Many models have similar biases and averaging models therefore does not reduce the biases, which explains why the MMM cannot perform best. In addition, the individual models capture better the natural variability of regional precipitation patterns than the MMM. This is due to the fact that by averaging all 24 CMIP3 models to construct the MMM, natural variability is automatically removed.

[33] As already mentioned, regional trends over the relatively short observational period (25 years) are often dominated by natural variability which is why the evaluation is only performed on the ability of the models to simulate the index mean value. Still, it is central that climate models are able to correctly simulate the trends. A source of concern in the case of the HLI is the inability of most CMIP3 models to reproduce the DJF precipitation decrease during the observational period. Further, the discrepancies between the two observational data sets CMAP and GPCP are very large for the HLI and STI. While for the rmse/corr ranking these differences have only a marginal influence (not shown), using CMAP as the reference data set for the feature-based ranking described above will lead to different outcomes. In certain regions it is therefore currently ambiguous to identify the best models due partly to uncertainties in the observational data sets. The implication is that the difficulties in defining model performance are not only a problem of agreeing on a metric, but is seriously limited by observational uncertainties. This underscores the need for continuous, global and homogeneous observations at high resolution.

[34] Considering only the five best models for each index narrows the range of predicted absolute values (dark blue lines in Figure 5), as expected. However, if anomalies are considered (see right-hand plots in Figures 5a–5h), the model spread is only reduced for 3 indices (AFI, HLI and STI; see error bars in Figure 5), it remains approximately the same for the AUI and CAI and even increases for the AMI,

ASI and MEI. The results of the temperature index time series are shown on Figure 6. In contrast to precipitation, the signal of temperature change dominates the natural variability and model agreement is larger. However, in terms of anomalies, only a minority of indices (AMI and HLI) see a reduction of model spread.

[35] The way the MMM was calculated in this study can be referred to as an "equal weighting" because each model has one "vote." A more sophisticated approach consists in assigning more weight to the "good" models. Several studies see some improvement in future projections when using "optimum weighting" approaches [e.g., *Perkins and Pitman*, 2009; *Räisänen et al.*, 2010]. On the other hand, *Santer et al.* [2009] find that an "optimum weighting" does not affect the results of their detection and attribution study for water vapor. For the feature-based metrics presented here, applying an "optimum weighting" to the models according to the ranking presented in section 6.1 will likely lead to a reduction of the uncertainty for only a few indices but these indices are different for precipitation (AFI, HLI and STI) and temperature (AMI and HLI). However, the problem is that an "optimum weighting" would keep the model uncertainty constant or even increase it for the rest of the indices. In addition, the differences in model spread found between the five best models and all models are highly time dependent: calculating the standard deviation by the year 2059 or 2099 would have lead to slightly different outcomes in terms of the indices showing a reduction of spread but the conclusion would remain the same. A further critical issue is the sampling of small subsets. The standard deviation may also change by picking a random subset of the models, even if the criteria for picking the models has no relevance at all. For 5 out of 24 models, there is a probability of about 5% for the spread (standard deviation) to increase or decrease by 50% or more in a random subset. In other words, at least a 50% change in the spread can be considered significant and very unlikely to arise by chance. Only the AFI for precipitation and the HLI for temperature show such large changes. In most indices the change in the spread after selecting the subset of models is well within what one would expect from randomly picking a subset. The results presented here are in agreement with *Weigel et al.* [2010], who argued that even if for some cases the "optimum weighting" outperforms the "equal weighting," the risk that the former is worse than the latter is large. In cases where there is currently no agreement on which skill measure to use in order to identify the best models, it is indeed more transparent to weight the models equally. However, *Weigel et al.* [2010] also showed that not considering those models known for lacking key mechanisms needed to provide meaningful projections might be justified in some cases.

[36] Further, *Knutti et al.* [2010b] showed that means and trends are generally not well correlated. In the case of the precipitation indices, a significant Pearson correlation coefficient between index mean (1980–2004) and trend (2020–2079) is only found for the STI ($\rho = -0.43$), an index for which the model uncertainty is reduced when considering only the five best models. For the temperature indices, a significant correlation is found for the ASI ($\rho = 0.43$) and the CAI ($\rho = 0.48$), which are indices that do not experience reduction of model spread when selecting only the five best models. However, given that these correlations are low and that there is no obvious physical explanation for them, they should not be overinterpreted. Rather, the fact that significant correlations between means and trends for an index do not always correspond to those indices with a reduction of model spread when considering the five best models indicates that feature-based metrics are not more useful to reduce the uncertainty in terms of anomalies than other metrics. Nevertheless, many end users are interested in absolute precipitation amounts and in this case, feature-based metrics are a simple way to identify the models that have some skill in a region but also to identify those who have obviously no skill.

[37] Finally, it should be noted that the interpretation of the MMM has been the subject of some debate. In particular, different interpretations of model independence, model robustness and of the ensemble of models itself are possible and lead to different interpretations of future model uncertainty [*Pirtle et al.*, 2010; *Knutti et al.*, 2010b, 2010a; *Annan and Hargreaves*, 2010]. On the one hand, climate models can be considered as "random samples from a distribution of possible models centered around the true climate" [*Jun et al.*, 2008]. Consequently, when averaging all models to construct the MMM, the errors are expected to decrease and the MMM to approach the truth [*Tebaldi et al.*, 2004]. The statistically indistinguishable ensemble paradigm is an alternative way to interpret ensembles, where the truth is a sample from the same distribution as each model of the ensemble [e.g., *Tebaldi and Sanso*, 2009]. *Annan and Hargreaves* [2010] compared both paradigms and find that the CMIP3 ensemble generally provides a good sample under the statistically indistinguishable paradigm. Assessing the statistical nature of the CMIP3 ensemble is beyond the scope of this study however, results from section 3 as well as in the case of the eight feature-based metrics for precipitation, it seems that the ensemble of models is not centered around the truth but appears biased. Therefore, the MMM is not closer than any other model to the observations, which seem to be statistically indistinguishable from the ensemble members. For temperature, the CMIP3 ensemble also appears biased but to a smaller extent than for precipitation. Nevertheless, there is a need for further studies focusing on how to interpret results from multiple models.

7. Conclusion

[38] The motivation for ranking the models is to specify which one(s) can provide the most reliable projections. Until now, model simulations have often been evaluated with statistical measures and on large spatial scales, where the MMM was found to perform best. As an alternative evaluation method, we provide eight feature-based performance metrics for precipitation and temperature. Feature-based metrics are designed to capture a robust signal of change in a particular variable that can be explained physically. As a first step, the causes behind the large projection uncertainty for precipitation are investigated. In large regions of the world, differences between the models contribute more to the total spread in projections than internal variability. However, agreement in the sign of trend among several runs of the same model is only slightly larger than among different models, indicating that even if differences between

models are reduced, internal variability will still cause a large lack of agreement in precipitation projections.

[39] For the regional feature-based metrics, the models performing best are different for each region and variable, and the choice of the observational data set is important in the case of precipitation. Averaging the models is more effective on aggregated metrics than on small scales and features. This is illustrated by the fact that when summarizing the performances of the models for all indices and both variables, the MMM ranks better than for each index individually. When the performances for the feature-based metrics are summarized, they correlate with the ranking obtained with statistical measures of errors and with the global field-based measures of *Reichler and Kim* [2008]. This is agreement with earlier studies [*Boer*, 1993; *Gleckler et al.*, 2008; *Pincus et al.*, 2008] who find that the MMM outperforms any individual model if enough metrics or grid points are evaluated and aggregated. We also tested a further way of ranking the models based on their ability to simulate the spatial correlation between the mean precipitation and temperature pattern in the index regions (not shown). It was found that the performance of the MMM for this regional correlation ranking is between its performance in the index ranking and the corr ranking for both precipitation and temperature. The MMM ranked first in ~35% of the cases, which is better than for the index ranking where it ranks average, but worse than for the corr ranking where it clearly ranks first. These findings confirm our hypothesis that the more grid cells, metrics or variables are aggregated, the better the performance of the MMM becomes.

[40] In a second part, the convergence of the projections of the best performing models for each index is investigated. On one hand and in particular for precipitation, the projections of the five best models in terms of absolute values appear more realistic than the ones performing below average since for most indices, the observations lie at one end of the model spread. However, when considering the anomalies, it is found that regardless of the variable, the majority of the indices see no reduction or even an increase in future uncertainty. These results suggest that on a regional scale, weighting the models might improve the projections only in few cases. In the absence of a process based argument, given the small number of existing models and the chosen subsets of 5 models, only a reduction in model spread by more than 50% is an indication of a successful constraint (see section 6.2). Model weighting should therefore be performed carefully. Our results tend to support previous findings showing that a good performance in the present does not guarantee skill in the future [*Jun et al.*, 2008; *Reifen and Toumi*, 2009]. On the other hand, there are a few cases where past and future performance in models are clearly related and physically well understood, for example, past greenhouse gas attributable warming scaling linearly with future transient greenhouse gas warming [*Allen and Ingram*, 2002; *Stott et al.*, 2006]. Such relationships are routinely used and widely accepted to constrain or calibrate projections with simple and intermediate complexity models [e.g., *Knutti et al.*, 2002; *Forest et al.*, 2002; *Meinshausen et al.*, 2009]. Another prominent example is the Arctic, where models underestimating past sea ice decline also show much weaker sea ice loss in the future [*Boe et al.*, 2009b] and where performance in simulating the current Arctic climate is related to projected future response in that region [*Boe et al.*, 2009a; *Mahlstein and Knutti*, 2011]. In such obvious cases we argue that observed evidence should not be ignored when synthesizing models.

[41] Evaluating the models is a central task in climate science and the reason why there is currently no agreement on a standard way to perform an evaluation reflects the fact that on the one hand, the connection between present-day and future performance is poorly understood and on the other hand, it also depends on the purpose. While hydrologists need assessments of the best performing models on a regional scale and primarily for precipitation and temperature, some model developers are more interested in summarizing the performance of climate models for many variables and over all regions of the globe as for example in work by *Reichler and Kim* [2008]. For specific applications and predictions, defining metrics not only based on mean biases but also on regional or temporal characteristics (e.g., distributions of daily rainfall) or on physical processes [e.g., *Eyring et al.*, 2005] may be more promising. It is evident that the index ranking presented here is partly subjective due to the choice of the eight indices. The indices should therefore be regarded as examples and depending on the purpose, other sets of indices can be defined. We also point out that the results are at most valid for precipitation and temperature and do not allow for any evaluations of the model performance on other variables or on a global scale. Further considerations of alternative ways of evaluating climate models in order to make best use of their predictions are encouraged.

[42] **Acknowledgments.** We acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modeling (WGCM) for their roles in making available the WCRP CMIP3 multimodel data set. Support for this data set is provided by the Office of Science, U.S. Department of Energy.

References

Adler, R. F., et al. (2003), The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present), *J. Hydrometeorol.*, *4*(6), 1147–1167.

Allen, M. R., and W. J. Ingram (2002), Constraints on future changes in climate and the hydrologic cycle, *Nature*, *419*, 224–232, doi:10.1038/nature01092.

Annan, J. D., and J. C. Hargreaves (2010), Reliability of the CMIP3 ensemble, *Geophys. Res. Lett.*, *37*, L02703, doi:10.1029/2009GL041994.

Boe, J., A. Hall, and X. Qu (2009a), Deep ocean heat uptake as a major source of spread in transient climate change simulations, *Geophys. Res. Lett.*, *36*, L22701, doi:10.1029/2009GL040845.

Boe, J. L., A. Hall, and X. Qu (2009b), September sea-ice cover in the Arctic Ocean projected to vanish by 2100, *Nat. Geosci.*, *2*(5), 341–343, doi:10.1038/NGEO467.

Boer, G. J. (1993), Climate change and the regulation of the surface moisture and energy budgets, *Clim. Dyn.*, *8*(5), 225–239.

Christensen, J., et al. (2007), Regional climate projections, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., pp. 847–940, Cambridge Univ. Press, Cambridge, U. K.

Dairaku, K., and S. Emori (2006), Dynamic and thermodynamic influences on intensified daily rainfall during the Asian summer monsoon under doubled atmospheric CO_2 conditions, *Geophys. Res. Lett.*, *33*, L01704, doi:10.1029/2005GL024754.

Eyring, V., et al. (2005), A strategy for process-oriented validation of coupled chemistry-climate models, *Bull. Am. Meteorol. Soc.*, *86*, 1117–1133, doi:10.1175/BAMS-86-8-1117.

Forest, C. E., P. H. Stone, A. P. Sokolov, M. R. Allen, and M. D. Webster (2002), Quantifying uncertainties in climate system properties with the use of recent climate observations, *Science*, *295*(5552), 113–117.

Giorgi, F., and L. O. Mearns (2002), Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the reliability ensemble averaging (REA) method, *J. Clim.*, *15*(10), 1141–1158.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux (2008), Performance metrics for climate models, *J. Geophys. Res.*, *113*, D06104, doi:10.1029/2007JD008972.

Hoskins, B. J., and K. I. Hodges (2005), A new perspective on Southern Hemisphere storm tracks, *J. Clim.*, *18*(20), 4108–4129.

Intergovernmental Panel on Climate Change (2007), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., 996 pp., Cambridge Univ. Press, Cambridge, U. K.

Jun, M., R. Knutti, and D. W. Nychka (2008), Spatial analysis to quantify numerical model bias and dependence: How many climate models are there?, *J. Am. Stat. Assoc.*, *103*(483), 934–947, doi:10.1198/016214507000001265.

Knutti, R., T. F. Stocker, F. Joos, and G. K. Plattner (2002), Constraints on radiative forcing and future climate change from observations and climate model ensembles, *Nature*, *416*, 719–723.

Knutti, R., G. Abramowitz, M. Collins, V. Eyring, P. J. Gleckler, B. Hewitson, and L. Mearns (2010a), Good practice guidance paper on assessing and combining multi model climate projections, in *Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections*, edited by T. F. Stocker et al., 13 pp., IPCC Working Group I Tech. Supp. Unit, Univ. of Bern, Bern, Switzerland.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl (2010b), Challenges in combining projections from multiple climate models, *J. Clim.*, *23*(10), 2739–2758.

Lambert, S. J., and G. J. Boer (2001), CMIP1 evaluation and intercomparison of coupled climate models, *Clim. Dyn.*, *17*(2-3), 83–106.

Mahlstein, I., and R. Knutti (2011), Ocean heat transport as a cause for model uncertainty in projected Arctic warming, *J. Clim.*, *24*, 1451–1460, doi:10.1175/2010JCLI3713.1.

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor (2007a), The WCRP CMIP3 multimodel dataset—A new era in climate change research, *Bull. Am. Meteorol. Soc.*, *88*, 1383–1394, doi:10.1175/BAMS-88-9-1383.

Meehl, G. A., et al. (2007b), Global climate projections, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., pp. 747–845, Cambridge Univ. Press, Cambridge, U. K.

Meinshausen, M., N. Meinshausen, W. Hare, S. C. B. Raper, K. Frieler, R. Knutti, D. J. Frame, and M. R. Allen (2009), Greenhouse-gas emission targets for limiting global warming to 2 degrees C, *Nature*, *458*, 1158–1162, doi:10.1038/nature08017.

Neelin, J. D., M. Munnich, H. Su, J. E. Meyerson, and C. E. Holloway (2006), Tropical drying trends in global warming models and observations, *Proc. Natl. Acad. Sci. U. S. A.*, *103*(16), 6110–6115.

Perkins, S. E., and A. J. Pitman (2009), Do weak AR4 models bias projections of future climate changes over Australia?, *Clim. Change*, *93*(3-4), 527–558, doi:10.1007/s10584-008-9502-1.

Phillips, T. J., and P. J. Gleckler (2006), Evaluation of continental precipitation in 20th century climate simulations: The utility of multimodel statistics, *Water Resour. Res.*, *42*, W03202, doi:10.1029/2005WR004313.

Pierce, D. W., T. P. Barnett, B. D. Santer, and P. J. Gleckler (2009), Selecting global climate models for regional climate change studies, *Proc. Natl. Acad. Sci. U. S. A.*, *106*(21), 8441–8446, doi:10.1073/pnas.0900094106.

Pincus, R., C. P. Batstone, R. J. P. Hofmann, K. E. Taylor, and P. J. Gleckler (2008), Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models, *J. Geophys. Res.*, *113*, D14209, doi:10.1029/2007JD009334.

Pirtle, Z., R. Meyer, and A. Hamilton (2010), What does it mean when climate models agree? A case for assessing independence among general circulation models, *Environ. Sci. Policy*, *13*(5), 351–361.

Pitman, A. J., G. T. Narisma, R. A. Pielke, and N. J. Holbrook (2004), Impact of land cover change on the climate of southwest Western Australia, *J. Geophys. Res.*, *109*, D18109, doi:10.1029/2003JD004347.

Previdi, M., and B. G. Liepert (2007), Annular modes and Hadley cell expansion under global warming, *Geophys. Res. Lett.*, *34*, L22701, doi:10.1029/2007GL031243.

Räisänen, J. (2007), How reliable are climate models?, *Tellus, Ser. A*, *59*(1), 2–29.

Räisänen, J., L. Ruokolainen, and J. Ylhäisi (2010), Weighting of model results for improving best estimates of climate change, *Clim. Dyn.*, *35*(2-3), 407–422.

Reichler, T., and J. Kim (2008), How well do coupled models simulate today's climate?, *Bull. Am. Meteorol. Soc.*, *89*, 303–311, doi:10.1175/BAMS-89-3-303.

Reifen, C., and R. Toumi (2009), Climate projections: Past performance no guarantee of future skill?, *Geophys. Res. Lett.*, *36*, L13704, doi:10.1029/2009GL038082.

Rowell, D. P., and R. G. Jones (2006), Causes and uncertainty of future summer drying over Europe, *Clim. Dyn.*, *27*(2-3), 281–299.

Santer, B. D., et al. (2009), Incorporating model quality information in climate change detection and attribution studies, *Proc. Natl. Acad. Sci. U. S. A.*, *106*(35), 14,778–14,783, doi:10.1073/pnas.0901736106.

Seneviratne, S. I., D. Lüthi, M. Litschi, and C. Schär (2006), Land-atmosphere coupling and climate change in Europe, *Nature*, *443*, 205–209, doi:10.1038/nature05095.

Smith, I., and E. Chandler (2010), Refining rainfall projections for the Murray Darling Basin of south-east Australia—The effect of sampling model results based on performance, *Clim. Change*, *102*(3-4), 377–393, doi:10.1007/s10584-009-9757-1.

Stott, P. A., J. F. B. Mitchell, M. R. Allen, T. L. Delworth, J. M. Gregory, G. A. Meehl, and B. D. Santer (2006), Observational constraints on past attributable warming and predictions of future global warming, *J. Clim.*, *19*(13), 3055–3069.

Sun, Y., and Y. H. Ding (2010), A projection of future changes in summer precipitation and monsoon in East Asia, *Sci. China Earth Sci.*, *53*(2), 284–300.

Tebaldi, C., and B. Sanso (2009), Joint projections of temperature and precipitation change from multiple climate models: A hierarchical Bayesian approach, *J. R. Stat. Soc., Ser. A*, *172*, 83–106.

Tebaldi, C., L. O. Mearns, D. Nychka, and R. L. Smith (2004), Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations, *Geophys. Res. Lett.*, *31*, L24213, doi:10.1029/2004GL021276.

Thompson, D. W. J., and S. Solomon (2002), Interpretation of recent Southern Hemisphere climate change, *Science*, *296*(5569), 895–899.

Trenberth, K. E., A. G. Dai, R. M. Rasmussen, and D. B. Parsons (2003), The changing character of precipitation, *Bull. Am. Meteorol. Soc.*, *84*, 1205–1217, doi:10.1175/BAMS-84-9-1205.

Uppala, S. M., et al. (2005), The ERA-40 re-analysis, *Q. J. R. Meteorol. Soc.*, *131*(612), 2961–3012, doi:10.1256/qj.04.176.

Weigel, A. P., R. Knutti, M. A. Liniger, and C. Appenzeller (2010), Risks of model weighting in multimodel climate projections, *J. Clim.*, *23*(15), 4175–4191, doi:10.1175/2010JCLI3594.1.

Whetton, P., I. Macadam, J. Bathols, and J. O'Grady (2007), Assessment of the use of current climate patterns to evaluate regional enhanced greenhouse response patterns of climate models, *Geophys. Res. Lett.*, *34*, L14701, doi:10.1029/2007GL030025.

Xie, P. P., and P. A. Arkin (1998), Global monthly precipitation estimates from satellite-observed outgoing longwave radiation, *J. Clim.*, *11*(2), 137–164.

Yin, X. G., A. Gruber, and P. Arkin (2004), Comparison of the GPCP and CMAP merged gauge-satellite monthly precipitation products for the period 1979–2001, *J. Hydrometeorol.*, *5*(6), 1207–1222.

J. Cermak, R. Knutti, I. Mahlstein, and N. Schaller, Institute for Atmospheric and Climate Science, ETH Zurich, Universitätstrasse 16, CH-8092 Zurich, Switzerland. (nathalie.schaller@env.ethz.ch)