



AMERICAN METEOROLOGICAL SOCIETY

Journal of Climate

EARLY ONLINE RELEASE

This is a preliminary PDF of the author-produced manuscript that has been peer-reviewed and accepted for publication. Since it is being posted so soon after acceptance, it has not yet been copyedited, formatted, or processed by AMS Publications. This preliminary version of the manuscript may be downloaded, distributed, and cited, but please be aware that there will be visual differences and possibly some content differences between this version and the final published version.

The DOI for this manuscript is doi: 10.1175/JCLI-D-12-00592.1

The final published version of this manuscript will replace the preliminary version at the above DOI once it is available.

If you would like to cite this EOR in a separate work, please use the following full citation:

Sheffield, J., A. Barrett, B. Colle, D. Fernando, R. Fu, K. Geil, Q. Hu, J. Kinter, S. Kumar, B. Langenbrunner, K. Lombardo, L. Long, E. Maloney, A. Mariotti, J. Meyerson, K. Mo, D. Neelin, Z. Pan, A. Ruiz-Barradas, Y. Serra, A. Seth, J. Thibeault, J. Stroeve, Z. Yang, and L. Yin, 2013: North American Climate in CMIP5 Experiments. Part I: Evaluation of Historical Simulations of Continental and Regional Climatology. *J. Climate*. doi:10.1175/JCLI-D-12-00592.1, in press.



1 **North American Climate in CMIP5 Experiments.**

2 **Part I: Evaluation of Historical Simulations of Continental and Regional**

3 **Climatology**

4 Justin Sheffield, Andrew Barrett, Brian Colle, D. Nelun Fernando, Rong Fu, Kerrie L.
5 Geil, Qi Hu, Jim Kinter, Sanjiv Kumar, Baird Langenbrunner, Kelly Lombardo, Lindsey
6 N. Long, Eric Maloney, Annarita Mariotti, Joyce E. Meyerson, Kingse C. Mo, J. David
7 Neelin, Sumant Nigam, Zaitao Pan, Tong Ren, Alfredo Ruiz-Barradas, Yolande L. Serra,
8 Anji Seth, Jeanne M. Thibeault, Julianne C. Stroeve, Ze Yang, Lei Yin

9 Justin Sheffield, Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ
10 Brian Colle, Kelly Lombardo, School of Marine and Atmospheric Sciences, Stony Brook University -
11 SUNY
12 Rong Fu, Lei Yin, Ze Yang, Tong Ren, Jackson School of Geosciences, University of Texas at Austin, TX
13 D. Nelun Fernando, Jackson School of Geosciences, The University of Texas at Austin, Austin, TX and
14 Visiting Scientist Programs, University Corporation for Atmospheric Research, Boulder, Colorado, and
15 Surface Water Resource Division, Texas Water Development Board, Austin, TX
16 Kerrie L. Geil, Department of Atmospheric Sciences, University of Arizona, Tucson, AZ
17 Qi Hu, School of Natural Resources and Department of Earth and Atmospheric Sciences, University of
18 Nebraska-Lincoln, Lincoln, NE
19 Sanjiv Kumar, Jim Kinter, Center for Ocean-Land-Atmosphere Studies, Calverton, MD
20 Baird Langenbrunner, Joyce E. Meyerson, J. David Neelin, Department of Atmospheric and Oceanic
21 Sciences, University of California Los Angeles
22 Lindsey N. Long, Wyle Science, Technology and Engineering, College
23 Park, MD, and Climate Prediction Center/NCEP/NWS/NOAA, College Park, MD
24 Eric D. Maloney, Department of Atmospheric Science, Colorado State University, Fort Collins, CO

25 Annarita Mariotti, National Oceanic and Atmospheric Administration, Office of Oceanic and Atmospheric
26 Research (NOAA/OAR), Silver Spring, MD
27 Kingtse C. Mo, Climate Prediction Center/NCEP/NWS/NOAA, College Park, MD
28 Zaitao Pan, Saint Louis University, St. Louis, MO
29 Alfredo Ruiz-Barradas, Sumant Nigam, Department of Atmospheric and Oceanic Science, University of
30 Maryland, College Park, MD
31 Yolande L. Serra, Department of Atmospheric Sciences, University of Arizona, Tucson, AZ
32 Anji Seth and Jeanne M. Thibeault, Department of Geography, University of Connecticut, Storrs, CT
33 Andrew Barrett, Julienne C Stroeve, National Snow and Ice Data Center, Cooperative Institute for
34 Research in Environmental Sciences, University of Colorado, Boulder, CO

35
36

37 Journal of Climate

38 Submitted on July 30, 2012

39 Revised February 3, and May 15, 2013

40

41

42 *Corresponding author address: Justin Sheffield, Department of Civil and Environmental
43 Engineering, Princeton University, Princeton, NJ, 08540. Email: justin@princeton.edu

44

45 **Abstract**

46 This is the first part of a three-part paper on North American climate in CMIP5 that
47 evaluates the historical simulations of continental and regional climatology with a focus
48 on a core set of seventeen models. We evaluate the models for a set of basic surface
49 climate and hydrological variables and their extremes for the continent. This is
50 supplemented by evaluations for selected regional climate processes relevant to North
51 American climate, including cool season western Atlantic cyclones, the north American
52 monsoon, the US Great Plains low level jet, and Arctic sea ice. In general, the multi-
53 model ensemble mean represents the observed spatial patterns of basic climate and
54 hydrological variables but with large variability across models and regions in the
55 magnitude and sign of errors. No single model stands out as being particularly better or
56 worse across all analyses, although some models consistently outperform the others for
57 certain variables across most regions and seasons, and higher resolution models tend to
58 perform better for regional processes. The CMIP5 multi-model ensemble show a slight
59 improvement relative to CMIP3 models in representing basic climate variables, in terms
60 of the mean and spread, although performance has decreased for some models.
61 Improvements in CMIP5 model performance are noticeable for some regional climate
62 processes analyzed, such as the timing of the North American monsoon. The results of
63 this paper have implications for the robustness of future projections of climate and its
64 associated impacts, which are examined in the third part of the paper.

65

66

67 **1. Introduction**

68 This is the first part of a three-part paper on the Coupled Model Intercomparison
69 Project phase 5 (CMIP5; Taylor et al., 2012) model simulations for North America. The
70 first two papers evaluate the CMIP5 models in their ability to replicate the observed
71 features of North American continental and regional climate, and related climate
72 processes for the recent past. This first part evaluates the models in terms of continental
73 and regional climatology and the second part (Sheffield et al. 2013) evaluates intra-
74 seasonal to decadal variability. The third part (Maloney et al., 2013) describes the
75 projected changes for the 21st century.

76 The CMIP5 provides an unprecedented collection of climate model output data
77 for the assessment of future climate projections as well as evaluations of climate models
78 for contemporary climate, the attribution of observed climate change and improved
79 understanding of climate processes and feedbacks. As such, these data feed into the
80 Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5), and
81 other global, regional and national assessments. The goal of this study is to provide a
82 broad evaluation of CMIP5 models in their depiction of North American climate and
83 associated processes. The set of climate features and processes examined in this first part
84 were chosen to cover the climatology of basic surface climate and hydrological variables
85 and their extremes at daily to seasonal time scales, as well as selected climate features
86 that have regional importance. The second part of this study (Sheffield et al. 2013) covers
87 aspects of climate variability, such as intra-seasonal variability in the tropical Pacific, the
88 El Niño Southern Oscillation (ENSO) and the Atlantic Multidecadal Oscillation that play
89 major roles in driving North American climate variability. This study draws from

90 individual work by investigators within the CMIP5 Task Force of the US National
91 Oceanic and Atmospheric Administration (NOAA) Modeling Analysis and Prediction
92 Program (MAPP). This paper is part of a Journal of Climate special collection on North
93 America in CMIP5 models and we draw from individual papers within the special
94 collection, which provide detailed analysis of some of the climate features examined
95 here.

96 We begin in Section 2 by describing the CMIP5, providing an overview of the
97 models analyzed, the historical simulations and the general methodology for evaluating
98 the models. We focus on a core set of 17 CMIP5 models that represent a large set of
99 climate centers and model types, and synthesize model performance across all analyses
100 for this core set. Details of the observational datasets to which the climate models are
101 compared are also given in this section. The next two sections focus on different aspects
102 of North American climate and surface processes. Section 3 begins with an overview of
103 climate model depictions of continental climate, including seasonal precipitation, air
104 temperature, sea surface temperatures, and atmospheric and surface water budgets.
105 Section 4 evaluates the model simulations of extremes of temperature and surface
106 hydrology, and temperature-based biophysical indicators such as growing season length.
107 The next section 5 focuses on regional climate features such as north Atlantic winter
108 storms, the Great Plains low level jet, and Arctic sea ice. The results are synthesized in
109 Section 6 and compared to results from CMIP3 models for selected variables.

110

111 **2. CMIP5 Models and Simulations**

112 *2.1. CMIP5 Models*

113 We use data from multiple model simulations of the “historical” scenario from the
114 CMIP5 database. The scenarios are described in more detail below. The CMIP5
115 experiments were carried out by 20 modeling groups representing more than 50 climate
116 models with the aim of further understanding past and future climate change in key areas
117 of uncertainty (Taylor et al., 2012). In particular, experiments focus on understanding
118 model differences in clouds and carbon feedbacks, quantifying decadal climate
119 predictability and why models give different answers when driven by the same forcings.
120 The CMIP5 builds on the previous phase (CMIP3) experiments in several ways. Firstly a
121 greater number of modeling centers and models have participated. Secondly, the models
122 generally run at higher spatial resolution with some models being more comprehensive in
123 terms of the processes that they represent, therefore hopefully resulting in better skill in
124 representing current climate conditions and reducing uncertainty in future projections.
125 Table 1 provides an overview of the models used.

126 To provide a consistent evaluation across the various analyses, we focus on a core
127 set of 17 models, which are highlighted in the table by asterisks. The core set was chosen
128 to span a diverse set of modeling centers and model types (coupled atmospheric-ocean
129 models (AOGCM), Earth system models (ESM) and models with atmospheric chemistry
130 (ChemAO and ChemESM)), and includes an AOGCM and ESM from the same modeling
131 center for three centers (GFDL, Hadley Center, AORI/NIES/JAMSTEC). The set was
132 restricted by data availability and processing constraints, and so for some analyses (in
133 particular those requiring high temporal resolution data) a smaller subset of the core
134 models was analyzed. When data for non-core models were available, these were also
135 evaluated for some analyses and the results are highlighted if they showed better (or

136 particularly poor) performance. The specific models used for each individual analysis are
137 provided within the results section where appropriate.

138

139 *2.2. Overview of Methods*

140 Data from the “historical” CMIP5 scenarios are evaluated in this study. The
141 “historical” simulations are run in coupled atmosphere-ocean mode forced by historical
142 estimates of changes in atmospheric composition from natural and anthropogenic
143 sources, volcanoes, greenhouse gases (GHG) and aerosols, as well as changes in solar
144 output and land cover. Note that only anthropogenic GHGs and aerosols are prescribed
145 common forcings to all models, and each model differs in the set of other forcings that it
146 uses, such as land use change. For ESMs, the carbon cycle and natural aerosols are
147 modeled and are therefore feedbacks, and we take this into consideration when discussing
148 the results. For certain basic climate variables we also analyze model simulations from
149 the CMIP3 that provided the underlying climate model data to the fourth assessment
150 report (AR4) of the IPCC. Several models have contributed to both the CMIP3 and
151 CMIP5 experiments, either for the same version of the model, or for a newer version, and
152 this allows a direct evaluation of changes in skill in individual models as well as the
153 model ensemble.

154 Historical scenario simulations were carried out for the period from the start of
155 the industrial revolution to near present: 1850-2005. Our evaluations are generally carried
156 out for the most recent 30 years, depending on the type of analysis and the availability of
157 observations. For some analyses the only, or best available, data are from satellite remote
158 sensing which restricts the analysis to the satellite period, which is generally from 1979

159 onwards. For other analyses, multiple observational datasets are used to represent the
160 uncertainty in the observations. An overview of the observational datasets used in the
161 evaluations is given in Table 2, categorized by variable. Further details of these datasets
162 and any data processing are given in the relevant sub-sections and figure captions. Where
163 the comparisons go beyond 2005 (e.g. 1979-2008), model data from the RCP8.5 future
164 projection scenario simulation are appended to the model historical time series. Most of
165 the models have multiple ensemble members and in general we use the first ensemble
166 member. In some cases, the results for multiple ensembles are averaged where
167 appropriate or used to assess the variability across ensemble members. Results are
168 generally shown for the multi-model ensemble (MME) mean and for the individual
169 models using performance metrics that quantify the errors relative to the observations.

170

171 **3. Continental Seasonal Climate**

172 We begin by evaluating the seasonal climatologies of basic climate variables:
173 precipitation, near surface air temperature, sea surface temperature (SST), and
174 atmosphere-land water budgets.

175

176 *3.1. Seasonal Precipitation Climatology*

177 Figure 1 shows the model precipitation climatology and GPCP (Adler et al., 2003)
178 observations for December-February (DJF) and June-August (JJA) for 1979-2005. Table
179 3 shows the seasonal biases in precipitation for North America, the US and six regions.
180 Most of the models do reasonably well in producing essential large-scale precipitation
181 features and the bias in the MME mean seasonal precipitation over North America is

182 about 12% and -1% for DJF and JJA, respectively. However, there are substantial
183 differences among the models, and with observations at the regional scale (Table 3) and
184 generally an overestimation of precipitation in more humid and cooler regions, and
185 underestimation in drier regions. For the winter season (Fig. 1, left), the Pacific storm
186 track is very reasonably placed in latitude as it approaches the coast. One important
187 aspect of this, the angle of the storm track as it bends northward approaching the coast
188 from roughly Hawaii to Central California, is well reproduced in the models. The
189 intensity of the storm tracks off the West Coast compares reasonably well to the GPCP
190 product shown here. The model rainfall is not quite intense enough at the coast and
191 spreads slightly too far inland, as might be expected for the typical model resolution
192 which does not fully resolve mountain ranges and may help explain the overestimation by
193 all models for western North America (WNA, see Table 3 for region definitions). The
194 east coast storm tracks are well placed in DJF (see section 5.1 on winter-time extra-
195 tropical cyclones) and the multi-model ensemble mean does a good job in replicating the
196 Eastern Pacific Inter-Tropical Convergence Zone (ITCZ), although northern Mexico
197 receives too much rainfall. Figures 1c provides a model by model view of these features
198 using the 3 mm/day contour for each model to provide an outline of the major
199 precipitation features. If the models were in line with observations, all contours would lie
200 exactly along the boundary of the shaded observations. Taking into account the high
201 latitude precipitation excess in the Pacific storm track, individual models do quite well at
202 reproducing each of the main features of the DJF climatology, including the arrival point
203 at the North American West Coast of the southern edge of the Pacific storm track. Only a

204 few models exhibit the ITCZ extension feature that accounts for the northern Mexico
205 precipitation excess.

206 For the summer season (JJA; Fig. 1, right), the ITCZ and the Mexican monsoon
207 are reasonably well simulated in terms of position (see section 5.4 on the North American
208 monsoon), although the precipitation magnitude in parts of the Caribbean is
209 underestimated relative to GPCP. The East Coast storm track in the multi-model
210 ensemble mean is too spread out and less coherent than observed. This is due to
211 substantial differences in the placement of these storm tracks in the individual models
212 (Fig. 1d). The majority of the models exhibit excessive precipitation in at least some part
213 of the continental interior. While the bulk of the models do reasonably well at the
214 poleward extension of the monsoon over Central America, Mexico and the Inter-
215 Americas Seas region, a few models underestimate this extent, putting a split between the
216 poleward extension of the monsoon feature and the start of the East Coast storm track.
217 Overall, the models underestimate JJA precipitation over the Central America (including
218 Mexico), and Central North America regions (Table 3).

219

220 *3.2 Seasonal Surface Air Temperature Climatology*

221 Figure 2 compares the model simulated surface air temperature climatology to the
222 observation estimates from NCEP-DOE Reanalysis 2 and the CRU TS3.0 station-based
223 analysis. The MME mean compares well to the observations in most respects.
224 Differences from both observational estimates are less than or on the order of 1°C over
225 most of the continent except for certain regions (see Table 4). The multi-model ensemble
226 mean is cooler than both data sets over northern Mexico in DJF. In high latitudes,

227 differences between the observational estimates are large enough that the error patterns in
228 Fig. 2f and Fig. 2g differ substantially, especially in DJF (NCEP-DOE is also slightly
229 warmer than the North American Regional Reanalysis, not shown, in this region and
230 season). Beyond the overall simulation of the north-south temperature gradient and
231 seasonal evolution, certain regional features are well represented. In JJA, this includes the
232 regions of temperatures exceeding 30°C over Texas and near the Gulf of California, and
233 the extent of temperatures above 10°C, including the northward extension of this region
234 into the Canadian prairies. Individual model surface air temperature climatologies, shown
235 in the supplementary material (Figs. S1 and S2) and in terms of biases in Table 4, exhibit
236 substantial regional scatter, including excessive northward extent of the region above
237 30°C through the Great Plains in three of the models (CanESM2, CSIRO-Mk3.6.0 and
238 FGOALS-s2). In DJF, the multi-model ensemble mean does a good job of representing
239 the 0°C contour, while the 10°C contour extends slightly too far south, yielding slightly
240 cool temperatures over Mexico, with 15 out of 18 models with cold biases over the
241 broader CAM region (Table 4). The wintertime cold bias relative to both observational
242 estimates in very high latitudes is more pronounced in certain models such as HadGEM2-
243 ES, which is biased low by -7.0 and -5.1 °C over the ALA and NEC regions, respectively
244 (Table 4). The inter-model scatter in surface temperature simulations is summarized in
245 Fig. 2d for DJF and Fig. 2j for JJA using the inter-model standard deviation of the
246 ensemble (i.e., the standard deviation at each gridpoint among the 18 model climatologies
247 seen in Figs. S1 and S2). For DJF, the inter-model standard deviation is less than 2.5°C
248 through most of the contiguous US but increases toward high latitudes, exceeding 3.5°C
249 over much of the area north of 60°N. In JJA, there is a region of high inter-model

250 standard deviation, exceeding 3.5°C, roughly in the Great Plains region in the northern
251 US and southern Canada. This is a region with fairly high precipitation uncertainty in JJA
252 (Fig. 1f), and changes in surface temperature in this region have been linked to factors
253 affecting soil moisture, including pre-season snowmelt (e.g., Hall et al. 2008), so this
254 may be a suitable target for further study to reduce model uncertainty.

255

256 *3.3. Seasonal Sea Surface Temperature*

257 The annual cycle of sea surface temperature (SST) is shown in Figure 3 as winter-
258 to-spring (December-May) and summer-to-fall (June-November) means. We also show
259 precipitation over land, which is generally associated with SST variations in adjoining
260 ocean regions. Maps for individual models are shown in supplementary Figures S4 and
261 S5. The Western Hemisphere Warm Pool (WHWP), where temperatures are equal or
262 larger than 28.5°C, usually is absent from December to February, and appears in the
263 Pacific from March to May, while it is present in the Caribbean and Gulf of Mexico from
264 June to November (Wang and Enfield, 2001). The cooler part of the year is characterized
265 by the small extension of SST in excess of 27°C and a suggestion of a cold tongue in the
266 eastern equatorial Pacific, while during the warmer part of the year the extension of SSTs
267 in excess of 27°C is maximum and the cold tongue is well defined over the eastern
268 Pacific. High precipitation along the Mexican coasts, Central America, the Caribbean
269 Islands and the central-eastern US are associated with the warm tropical SSTs during the
270 warm half of the year. A decrease in the regional precipitation south of the equator is also
271 evident in this warm half of the year.

272 The MME mean shows the observed change in SST from cold to warm around the
273 WHWP region, however the warm pool is absent over the Caribbean and Gulf of Mexico
274 region. The change in precipitation from the cold to the warm parts of the year is
275 represented by the MME mean including the increase in precipitation over central US and
276 Mexico as well as the decrease south of the equator. The eastern Pacific in the models is
277 slightly cooler than observations in the cold part of the year but not in the form of weak
278 cold tongue from the Peruvian coast but rather as a confined equatorial cooling away
279 from the coast. The cold tongue along the eastern equatorial Pacific and along the coast
280 of Peru during the warmer part of the year is reasonably represented by the MME mean
281 although its extension is farther to the west. Differences with observations of the multi-
282 model mean indicate cool SST biases over the Pacific and Intra-American seas parts of
283 the WHWP in all models in both the cold and warm parts of the year. Warm biases are
284 evident close to the coasts of northeastern US, western Mexico and Peru. Precipitation
285 biases indicate a wet/dry bias to the west/east of $\sim 97^\circ\text{W}$ northward of 15°N over Mexico
286 and the US during both parts of the year (as well as the intense and extensive dry bias
287 over South America to the east of the Andes); the cold bias over the Intra-Americas sea
288 and the dry bias over the Great Plains in the US suggests a link between the two,
289 considering the former is a great source of moisture for the latter.

290 A set of error statistics for the mean annual SSTs are summarized in Table 5 for
291 the individual CMIP5 models, and the MME mean. The spatial correlations are > 0.9 for
292 all models, and are not able to quantitatively distinguish the performance of the models.
293 The MME mean maximizes the spatial correlation (0.97) and minimizes the RMSE
294 (0.77°C), but not the bias (-0.54°C). Eight of the models have RMSE values less than

295 1°C, and the largest biases ($> 1.3^{\circ}\text{C}$) are for CSIRO-Mk3.6.0, HadCM3, INMCM4, IPSL-
296 CM5A-LR, and MIROC-ESM. The biases, except for INMCM4 and CCSM4, are
297 negative, with the smallest bias for INMCM4, and the largest for CSIRO-Mk3.6.0.

298

299 *3.4. Seasonal Atmospheric and Land Water Budgets*

300 We next evaluate the climatologies of the atmospheric and land water budgets.
301 Seasonal changes in atmospheric water content are relatively small compared to the
302 moisture fluxes and so we focus on the latter. Variations in moisture divergence are
303 generally correlated with seasonal precipitation and so may help explain biases in model
304 precipitation. The vertically integrated moisture transport (vectors) and its divergence
305 (contours) are shown in Fig. 4 for five CMIP5 models (the number of models was limited
306 by the availability of high temporal resolution model data required to calculate the
307 moisture fluxes) and observational estimates from the 20CR for mean JJA and DJF for
308 1981-2000. In summer, the 20CR shows southerly transport from the North Atlantic
309 anticyclone that splits into two distinct branches: one flanking the Atlantic seaboard with
310 large scale convergence off the east coast and a second branch of moisture flows into the
311 interior central plains which is associated with convergence over the Rocky Mountains.
312 The western U.S. is dominated by divergence associated with the northerly component of
313 the North Pacific anticyclone. The five models show the two branches of moisture
314 transport, with associated convergence off the east coast and divergence in the plains,
315 albeit weaker. They also simulate the divergence in much of the west, but they do not
316 simulate the strong convergence over the Rockies and Mexican Plateau as seen in 20CR,
317 which is associated with the low bias in precipitation over these regions (Table 3; mean

318 biases for the five models shown here are -19.8% and -31.5% for CNA and CAM
319 regions, respectively). Spatial correlations for divergence in the North American region
320 range from 0.08 to 0.42, with MIROC5 and CNRM-CM5 performing the best out of the
321 five models according to this measure (Table 6). In winter, the 20CR shows a more zonal
322 transport than during summer, with weaker flow around the subtropical anticyclones and
323 moisture convergence across much of the continent. The models represent both the
324 moisture transport and divergence patterns well including the stronger convergence in the
325 Pacific Northwest and northern California and divergence in southern California,
326 although the magnitude of divergence is too strong along the coasts, most notably for the
327 CCSM4 and CNRM-CM5 models, and precipitation over the western North America is
328 overestimated by all five models examined here (WNA and ALA regions, Table 3), and
329 especially for the CCSM4. The improvement in winter over summer for the whole
330 domain is evident in the spatial correlations, which range between 0.60 and 0.76 for
331 winter, with a different set of models performing better than in summer (CanESM2,
332 CCSM4, and CNRM-CM5 – Table 6).

333 Evaluations of the model simulated terrestrial water budget are shown in Figures
334 5 and 6 against the off-line land surface model (LSM) simulations. Fig. 5 shows the
335 regional mean seasonal cycles of the components of the land surface water budget
336 (precipitation, evapotranspiration, runoff, change in water storage). In reality, water
337 storage includes soil moisture, surface water such as lakes, reservoirs, and wetlands,
338 groundwater and snowpack, but, in general, the climate models only simulate the soil
339 moisture and snowpack components. Figure 5 also separates out the snow component of
340 the water budget in terms of the snow water equivalent (SWE). Most models have a

341 reasonable seasonal cycle of precipitation and evapotranspiration but tend to overestimate
342 precipitation in the more humid and cooler regions (WNA, ENA, ALA, NEC) as noted
343 previously and overestimate evapotranspiration throughout the year and especially in the
344 cooler months. Runoff is generally underestimated, particularly in the central and eastern
345 North American regions (CNA and ENA) and in northeast Canada (NEC) and central
346 America (CAM). It also peaks earlier in the spring in some models (that can be linked to
347 a shortened snow season; see below), although the models generally replicate the spatial
348 variability in annual total runoff (Figure 6 and Figure S6 in the supplementary material).
349 The majority of models overestimate total runoff over dry regions and high latitudes,
350 particularly for the Pacific Northwest and Newfoundland. SWE is generally
351 overestimated by the multi-model ensemble for western North America, underestimated
352 in the east and overestimated in the Alaskan/Western Canada region, which are a
353 reflection of the precipitation biases. These biases are also reflected in the change in
354 storage, particularly for the Alaska region where many of the models show a large
355 negative change during late spring melt due to overestimation of SWE.

356 Figure 6 (Figure S6 for individual models) also shows the runoff ratio (runoff
357 divided by precipitation) over North America, which indicates the production of water at
358 the land surface that is subsequently potentially available as water resources. The
359 remaining precipitation is partitioned into evapotranspiration (assuming that storage does
360 not change much over long time periods). Overall the MME mean replicates the spatial
361 pattern from the observational estimate with higher ratios in humid and cooler regions,
362 and lower ratios in dry regions. However, the MME mean overestimates the ratios in
363 humid and cooler regions, especially in Alaska, western and northern Canada, and

364 underestimates the ratios in dry regions (Table 7). For North America overall, the models
365 overestimate the ratios. The biases in runoff are better explained by biases in runoff ratios
366 rather than biases in precipitation (not shown), especially in higher latitudes, highlighting
367 the importance of the land surface schemes in the climate models and whether they are
368 able to realistically partition precipitation into runoff and evapotranspiration, and
369 accumulate and melt snow.

370

371 **4. Continental Extremes and Biophysical Indicators**

372 This next section examines the performance of the models in representing
373 observed temperature and hydrological extremes. We first focus on temperature extremes
374 and temperature dependent biophysical indicators, and then persistent seasonal
375 hydrological extremes for precipitation and soil moisture. Regional extremes in
376 temperature and precipitation are evaluated in Section 5.

377

378 *4.1. Temperature Extremes and Biophysical Indicators*

379 Temperature extremes have important consequences for many sectors including
380 human health, ecosystem function, and agricultural production. We evaluate the models'
381 ability to replicate the observed spatial distribution over North America of the frequency
382 of extremes (Figure 7) for the number of summer days with maximum temperature
383 (T_{max}) $> 25^{\circ}\text{C}$ and the number of frost days with minimum temperature (T_{min}) $< 0^{\circ}\text{C}$
384 (Frich et al., 2002) and a set of biophysical indicators related to temperature: spring and
385 fall freeze dates and growing season length. We define the growing season length
386 following Schwartz et al. (2006) which is the number of days between the last spring

387 freeze of the year and the first hard freeze of the autumn in the same year. A hard freeze
388 is defined as when the daily minimum temperature drops below -2°C .

389 Overall, the models tend to underestimate the number of summer days by about
390 18 days over North America (Table 8), with regional underestimation of over 50 days in
391 the western US and Mexico, and parts of the eastern US, but otherwise are within 20 days
392 of the observations for most other regions. Several models (CanESM4, CCSM4, CNRM-
393 CM5, MIROC, MIROC-ESM) overestimate the number of summer days from the
394 northeastern US up to the Canadian Northern Territories, but tend to have smaller
395 underestimation in the western US and Mexico (see supplementary Figure S7). Nearly all
396 other models have low biases of up to 50 days in these drier regions, which, at least for
397 the western US, may be related to overestimation of precipitation and evapotranspiration
398 (as shown in Section 3.4) and thus a reduction in sensible heating of the atmosphere.
399 Several models have small biases for North America as a whole (Table 8), but often
400 because large regional biases cancel out, and only the BCC-CSM1-1, CSIRO-Mk3.6.0
401 and HadGEM2-ES models have reasonably low biases (< 30 days) across all regions. The
402 first two of these models also have relatively low runoff ratio biases for the western and
403 central North American regions (WNA, CNA) (HadGEM2-ES was not evaluated for
404 surface hydrology) suggesting that their simulation of warm summer days is not impeded
405 by biases in the surface energy budget. The number of frost days are better simulated in
406 terms of overall MME mean bias (-2.8 days) but there is a positive bias across the
407 Canadian Rockies and down into the US Rockies for most models suggesting that the
408 models' generally coarse resolution and differences in topographic heights are partly
409 responsible (see supplementary Figure S8). Some of the models are biased low in the

410 central US by over 50 days. Models with the least bias in frost days also tend to be the
411 least biased models for summer days, but again many of the regional biases cancel out for
412 the North America values.

413 The models do reasonably well at depicting the spatial distribution of growing
414 season length (MME mean bias = -8.5 days over North America). The largest biases of
415 between 30-50 days are in western Canada where the models underestimate and in the
416 central US where they overestimate. The former is mainly because the last spring freeze
417 is too late in western Canada and for the latter because of biases in both the last spring
418 freeze (too early) and the first autumn freeze (too late). The INMCM4 model has the
419 largest bias overall (-76 days), which is consistent over most of the continent (see Figure
420 S9). The MIROC5 and MIROC-ESM models have the largest overestimations of 33 and
421 38 days, respectively, and these biases are also consistent over much of the continent.

422

423 *4.2. Hydroclimate Extremes*

424 We examine the ability of CMIP5 models to simulate persistent drought and wet
425 spells in terms of precipitation and soil moisture (SM). We focus on the US because of
426 the availability of long-term estimates of SM from the NLDAS-US dataset.
427 Meteorological drought and wet spells are characterized by the 6-month Standardized
428 Precipitation Index (SPI6; McKee et al., 1993). Agricultural drought and wet spells are
429 evaluated in terms of soil moisture percentiles (Mo, 2008). The record length, N_{total} , is
430 defined as the total months from all ensemble simulations of a model or the total months
431 of the observed data set. At each grid point, an extreme negative (positive) event is
432 selected when the SPI6 index is below (above) -0.8 (0.8) for a dry (wet) event [Svoboda

433 *et al.* 2002]. For SM percentiles, the threshold is 20% (80%) for a dry (wet) event. At
434 each grid cell, the number of months that extreme events occur (N) is 20% of the record
435 length by construct ($N/N_{\text{total}} = 20\%$). Because a persistent drought event (wet event)
436 usually means persistent dryness (wetness), a drought (wet) episode is selected when the
437 index is below/above this threshold for 3 consecutive seasons (9 months) or longer. The
438 frequency of occurrence of persistent drought or wet spells (FOC) is defined as: $\text{FOC} =$
439 N_p/N , where N_p is the number of months that an extreme event persists for 9 months.

440 Figures 8 and 9 show the FOC averaged for persistent wet and dry events for SPI6
441 and SM, respectively for 15 of the core models (the GFDL-ESM2M and INMCM4 model
442 datasets only had a single ensemble member and the total record is therefore too short for
443 the analysis). The most noticeable feature is the east-west contrast of the FOC for both
444 SPI6 and SM as driven by the gradient in precipitation amount and variability (Mo and
445 Schemm, 2008). Persistent drought and wet spells are more likely to occur over the
446 western interior region, while extreme events are less likely to persist over the eastern US
447 and the west coast. The maxima of the FOC are located in two bands, one located over
448 the mountains and one extending from Oregon to Texas (Fig. 8-a). Persistent events are
449 also found over the Great Plains. The CanESM2, CCSM4 and MIROC5 models show the
450 east-west contrast, although the magnitudes of FOC are too weak for the CanESM2
451 model. The center of maximum FOC for MIROC5 is too far south.

452 Table 9 shows the performance of the models in representing the east-west
453 contrast in terms of a FOC index defined as the difference in the fraction of grid cells
454 with FOC greater than a given threshold between the western (32-48°N, 92-112°W) and
455 eastern (32-48°N, 70-92°W) regions. The thresholds are 0.2 for SPI and 0.3 for SM. The

456 FOC index values for the CCSM4 (0.35) and MIROC5 (0.34) models are closest to the
457 observations (0.37) for SPI6. The MPI-ESM-LR model also shows the east-west contrast
458 with one maximum located over Utah and another over the Great Plains, but the second
459 maximum is too spatially extensive. The MIROC-ESM, MRI-CGCM3, and NorESM1-M
460 models all show a band of maxima over the Southwest, but the FOC north of 35°N is too
461 weak. Other models such as CSIRO-Mk3.6.0, IPSL-CM5A-LR, CNRM-CM5, GISS-E2-
462 R and GFDL-CM3 have the maxima located over the Gulf region, which is too far south.
463 Finally, the HadCM3 and HadGEM2-ES (not shown) models do not have enough
464 persistent events.

465 For SM (Figure 9), the FOC from the NLDAS-UW shows that persistent
466 anomalies are located west of 90°W over the western interior region, with a FOC index of
467 0.68. Many of the models, such as BCC-CSM1.1, HadCM3, and IPSL-CM5A-LR, do not
468 have enough persistent events, and the CanESM2, GISS-E2-R, and MRI-CGCM3 models
469 shift the maxima to the central US. The CCSM4, GFDL-CM3, and NorESM1-M models
470 fail to replicate the east-west contrast because of their high FOC values throughout most
471 of the US. The best model for SM is the MPI-ESM-LR with a FOC index of 0.62,
472 because it represents the east-west contrast and also has realistic magnitudes. The
473 CSIRO-Mk3.6.0 model also simulates the east-west contrast, but the maximum is located
474 south of the NLDAS-UW analysis maximum.

475

476 **5. Regional Climate Features**

477 We next evaluate the CMIP5 models for a set of regional climate features that
478 have important regional consequences, either directly such as extreme temperature and

479 precipitation in the southern US, and the North American Monsoon, or indirectly such as
480 western Atlantic cool season cyclones and the US Great Plains low level jet. The last
481 analysis examines the simulation of Arctic sea ice, which is important locally but also has
482 implications for North American climate and elsewhere (Francis and Vavrus, 2012).

483

484 *5.1. Cool Season Western Atlantic Extratropical Cyclones*

485 Extratropical cyclones can have major impacts (heavy snow, storm surge, winds,
486 flooding) along the east coast of North America given the proximity of the western
487 Atlantic storm track. The Hodges (1994; 1995) cyclone tracking scheme was
488 implemented to track cyclones in 15 models (of which 12 were in the core set) for the
489 cool seasons (November to March) for 1979-2004. The CFSR reanalysis was used to
490 estimate observed cyclone tracks. Six-hourly mean sea-level pressure (MSLP) data were
491 used to track the cyclones, since it was found that including 850-hPa vorticity tracking
492 yielded too many cyclones. Since MSLP is strongly influenced by large spatial scales and
493 strong background flows, a spectral bandpass filter was used to preprocess the data.
494 Those wavelengths between 600 and 10,000 km were kept, and the MSLP pressure
495 anomaly had to persist for at least 24 hours and move at least 1000 km. Colle et al. (2013)
496 describes the details of the tracking approach and validation of the tracking procedure.

497 Figure 10 shows the cyclone density during the cool season for the CFSR, mean
498 and spread of the 15 models (see the legend of Fig. 11 for a complete listing), and select
499 models for eastern North America and the western and central North Atlantic. There is a
500 maximum in cyclone density in the CFSR over the Great Lakes, the western Atlantic
501 from east of the Carolinas northeastward to east of Canada, and just east of southern

502 Greenland (Figure 10a). The largest maximum over the western Atlantic (6-7 cyclones
503 per cool season per 50,000 km²) is located along the northern boundary of the Gulf
504 Stream current. The MME mean is able to realistically simulate the three separate
505 maxima locations (Figure 10b), but the amplitude is 10-20% underpredicted. The cyclone
506 density maximum over the western Atlantic does not conform to the boundary of the Gulf
507 Stream as much as observed. There is a large inter-model spread near the Gulf Stream,
508 since some models are able to better simulate western Atlantic density amplitude, such as
509 the CCSM4 and HadGEM2-CC (Figures 10e,f). However, the CCSM4 maximum is
510 shifted a few hundred kilometers to the north.

511 The distribution of cyclone central pressures at their maximum intensity were also
512 compared (Figure 11) between the CFSR, MME mean, and individual models for the
513 dashed box region in Figure 10b. There is a peak in cyclone intensity in both the CFSR
514 and MME mean around 900-1000 hPa, and there is large spread in the model intensity
515 distribution by almost a factor of two. The ensemble mean realistically predicts the
516 number of average strength to relatively weak cyclones; however, the intensity
517 distribution is too narrow compared to the CFSR, especially for the deeper cyclones <
518 980 hPa.

519 Colle et al. (2013) verified the 15 models by calculating the spatial correlation and
520 mean absolute errors of the cyclone track densities and central pressures. They ranked the
521 models and showed that 6 of the 7 best models were the higher resolution models (top
522 three: EC-Earth, MRI-CGM3, and CNRM-CM5), since many lower resolution models,
523 such as GFDL-ESM2M (Figures 10d), underpredict the cyclone density and intensity.
524 The MME mean calculated using the 12 core models has verification scores within 5% of

525 those from all 15 models (not shown), so it is likely that using all 17 core models in the
526 cyclone analysis would not have much impact on the results.

527

528 *5.2. Northeast Cool Season Precipitation*

529 We next examine regional precipitation in the highly populated northeast US,
530 which is expected to increase in the future (Maloney et al., 2013). The focus is on the
531 cool season, since extratropical cyclones provide much of the heavy precipitation in the
532 northeast. 14 of the core models (listed in Figure 11; daily precipitation data were not
533 available for three models) were evaluated for the cool seasons (November to March) of
534 1979-2004. The model daily precipitation was compared with the CPC-Unified daily
535 precipitation at 0.5 degree and CMAP monthly precipitation at 2.5 degree resolution.

536 Figures 12a-c shows the seasonal average precipitation for the two observational
537 analyses and the MME mean and spread. The heaviest precipitation (700-1000mm) is
538 over the Gulf Stream, which is associated with the western Atlantic storm track. This
539 maximum is well depicted in the multi-model mean, although it is underestimated by 50-
540 200mm, and there is a moderate spread between models (100-200mm). The precipitation
541 over the northeast US ranges from 375mm in the northwestern part to around 500mm at
542 the coast. The finer resolution CPC-Unified analysis has more variability downstream of
543 the Great Lakes (lake effect snow) as well as some terrain enhancements. The models
544 cannot resolve these smaller scale precipitation features, but the MME mean realistically
545 represents the north to south variation. However, the MME mean overestimates
546 precipitation by 25-75 mm (5-20%) over northern parts. Much of this overestimation is
547 for thresholds greater than 5 mm day⁻¹ over land. (Fig. 12d). The seasonal precipitation

548 MME spread over the northeast is 100-150mm (25-40%), and much of this spread is
549 reflected in the higher ($> 10 \text{ mm day}^{-1}$) thresholds, with the BCC-CMS1-1 simulating less
550 than the CPC Unified analysis, and a cluster of models, such as the INMCM4 and
551 MIROC5, having many more heavy precipitation events than observed.

552 The model precipitation was verified against the CPC-Unified analysis for the
553 black box region over the northeast US shown in Fig. 12b, and the models ranked in
554 terms of their mean absolute errors (MAE) (Table 10). The MME mean has the lowest
555 MAE. There is little relationship with resolution, since some relatively higher resolution
556 models (e.g., MIROC5 and MRI-CGCM3) perform worse than many other lower
557 resolution models. Most models have a 5-15% high bias in this region. There is little
558 correlation (~ 0.22) between the high biases in precipitation in this region and the cyclone
559 overestimation along the US East coast, thus suggesting the cyclone biases are coming
560 from other processes than diabatic heating errors from precipitation.

561

562 *5.3. Extreme Temperature and Rainfall over the Southern US*

563 The southern regions of the US are historically prone to extreme climate events
564 such as extreme summer temperatures, flood and dry spells. Previous CMIP and US
565 climate impact assessments (Karl et al. 2009) have projected a large increase of these
566 extreme events over regions of the south (southwest (SW), south central (SC), southeast
567 (SE)), especially for the SW and SC US. However, to what extent climate models can
568 adequately represent the statistical distributions of these extreme events over these
569 regions is still unclear. Figure 13 compares the model simulated precipitation and
570 temperature with observations as Taylor diagrams for 1) the annual number of heavy

571 precipitation days (precipitation $> 10\text{mm day}^{-1}$) and 2) the number of hot days ($T_{\text{max}} >$
572 32°C (90°F)). The observations are derived from the GHCN daily T_{max} and T_{min} gauge
573 data and the CPC US-Mexico daily gridded precipitation dataset. Results are shown for
574 15 models, 11 of which are core models, in terms of the spatial correlation with the
575 observations and standard deviation normalized by the observations. Table 11 also shows
576 the regional biases.

577 Overall, the spatial distribution of the number of heavy precipitation days is better
578 simulated in the SW and SC than the SE, for which the spatial correlations are below 0.5,
579 with many models having negative correlations. The normalized standard deviations are
580 less than observed indicating that the models cannot capture the high spatial variability in
581 this region. Part of the reason for this may be the severe underestimation of number of
582 tropical cyclones (Sheffield et al., 2013), although other factors are likely involved such
583 as the biases in summertime convective precipitation. For the SW and SC regions the
584 models do reasonably well at replicating the spatial variation, although with some spread
585 across models (correlation values of 0.56-0.91, and 0.59-0.97, for the SW and SC,
586 respectively). The MME mean simulated number of heavy precipitation is biased slightly
587 high for the SW (but note that the observed number of days, 8.5, is small), and low for
588 the SC and SE. For individual models, the GISS-E2-R model has a large high bias in the
589 SW and the CanESM2, GFDL-CM3, HadCM3, IPSL-CM5A-LR and MIROC5 models
590 have large low biases (> 10 days) in the SC and SE. Several models do reasonably well
591 for all regions (GFDL-ESM2G, GFDL-ESM2M, HadGEM2-CC, HadGEM2-ES,
592 MIROC4h, MPI-ESM-LR and MRI-CGCM3) in terms of their biases.

593 The number of hot days ($T_{\max} > 32^{\circ}\text{C}$) are underestimated by the MME mean for
594 all regions by between about 12-19 days, which is consistent with the underestimation of
595 summer days ($T_{\max} > 25^{\circ}\text{C}$) shown for N. America in Figure 7. Again the performance
596 of the models in terms of spatial patterns and variability, and regional bias is generally
597 worse for the SE. Interestingly, the three Hadley Center models considered here
598 (HadCM3, HadGEM2-CC and HadGEM2-ES) have the lowest biases for the SW and SC
599 (except for the CCSM4), and in the SE (except for the MIROC models). The models tend
600 to overestimate the spatial variability in the SW and underestimate it the SE, and the
601 spatial correlations for the $\text{SW} > \text{SC} > \text{SE}$. The MIROC4h, which is a very high-
602 resolution model (0.56 degree grid), stands out for all regions and both variables as
603 having high spatial correlation, and low bias for heavy precipitation days, although it
604 generally has too high spatial variability relative to the observations.

605

606 *5.4. North American Monsoon*

607 The North American Monsoon (NAM) brings rainfall to southern Mexico in May,
608 expanding northward to the southwest US by late June or early July. Monsoon rainfall
609 accounts for roughly 50-70% of the annual totals in these regions (Douglas et al., 1993;
610 Adams and Comrie, 1997), with the annual percentages decreasing northward where
611 winter rains become increasingly important. The annual cycle of precipitation from the
612 ITCZ through the NAM region is examined in Figure 14. The MME mean from the 17
613 core models (averaged for longitudes 102.5° - 115°W for 1979-2005) replicates the
614 northward migration of precipitation in the NAM region during the warm season, but is
615 biased low. However, the MME mean precipitation begins later, ends later, and is

616 stronger than the observed estimate from CMAP within the core monsoon region north of
617 20°N. Within the latitudes of the ITCZ (up to 12°N), the models strongly underestimate
618 the precipitation and fail to show the northward migration from stronger precipitation in
619 May south of 8°N to a maximum in July near 10°N. Instead, the models tend to place the
620 spring maximum at 10°N and have a late build up and late demise at all latitudes of the
621 ITCZ through boreal summer. Table 12 shows the RMSE for individual models over the
622 domain shown in Figure 14 and indicates that the CanESM2, HadCM3 and HadGEM2-
623 ES models have the lowest errors ($< 0.75 \text{ mm day}^{-1}$) and the BCC-CSM1-1, NorESM1-M
624 and MRI-CGCM3 the highest ($> 1.9 \text{ mm day}^{-1}$).

625 The seasonal cycle of monthly precipitation in the core NAM region of northwest
626 Mexico (23.875°-28.875°N, 108.875°-104.875°W) is also examined in Table 13 and
627 Figures 15 for the core models plus four other models. Our core domain is similar to that
628 used by the North American Monsoon Experiment (NAME; Higgins et al. 2006) and
629 related studies (e.g., Higgins and Gochis, 2007; Gutzler et al., 2009), but has been
630 reduced in size to ensure consistency of the monsoon precipitation signal at each grid
631 point. Following the methodology of Liang et al. (2008) for analysis of CMIP3 data, we
632 calculate a phase and RMS error of each model's seasonal cycle, where the phase error is
633 defined as the lag in months with the best correlation to the observations (Table 13). The
634 observations used are the P-NOAA, which is a recently developed gauge-based dataset
635 that is likely more accurate than the CMAP for this region. We additionally calculate
636 each model's annual bias as a percentage of the mean monthly climatological P-NOAA
637 value (1.66 mm day^{-1}). The seasonal cycles for models with small (lag=0), moderate

638 (lag=1) and large (lag=2-4) phase errors are shown in Figure 15a-c. Figure 15d shows the
639 MME mean for all phase errors, their spread and the observations.

640 Overall the small phase error models tend to overestimate rainfall in the core
641 NAM region compared to the two observational data sets throughout the year, with the
642 largest errors seen in fall, consistent with Figure 14. The overestimation of rainfall by the
643 models beyond the end of the monsoon season is also apparent in the small and large
644 phase error CMIP3 models (Liang et al. 2008). The similarity between the range of
645 RMSE values (0.46-2.23 mm day⁻¹) in their study of CMIP3 models and that of the
646 CMIP5 models in this analysis indicates that there has been no improvement in the
647 magnitude of the simulated annual cycle of monthly precipitation, with the lowest and
648 highest RMSE values having increased slightly since the previous generation of models.
649 On the other hand, there does seem to be improvement in the timing of seasonal
650 precipitation shifts, with 13 out of 21 (62%) CMIP5 models having a phase lag of zero
651 months as compared to 6 out of 17 (35%) CMIP3 models in Liang et al. (2008). The top
652 ranking models for phase, RMSE and bias shown in Table 13 (HadCM3, HadGEM2-ES,
653 CNRM-CM5, CanESM2, HadGEM2-CC) are also the models with the highest spatial
654 correlations of May-October 850hPa geopotential heights and winds when compared with
655 the ERA-Interim (Geil et al. 2013). The HadCM3, HadGEM2-ES and CanESM2 also
656 perform the best over the larger monsoon region (Table 12). Geil et al. (2013) find that
657 the models that best represent the seasonal shift of the monsoon ridge and subtropical
658 highs over the North Pacific and Atlantic tend to have the least trouble ending the
659 monsoon, suggesting there is room for improvement over the region through an improved
660 representation of the seasonal cycle in these large-scale features.

661

662 *5.5. Great Plains Low Level Jet*

663 An outstanding feature of the warm season (May-September) circulation in North
664 America is the strong and channeled southerly low-level flows, or the Great Plains low-
665 level jet (LLJ), from the Gulf of Mexico to the central US and the Midwest (Bonner and
666 Paegle 1970; Mitchell et al. 1995). The LLJ emerges in early May in the transition of the
667 circulation from the cold to the warm season. It reaches its maximum strength in June and
668 July. After August, the jet weakens and disappears in September when the cold season
669 circulation starts to set in. While many studies have examined specific processes
670 associated with the LLJ (Blackadar, 1957; Wexler, 1961; Holton, 1967), such as its
671 nocturnal peak in diurnal wind speed oscillation, as well as precipitation, the jet is a part
672 of the seasonal circulation shaped primarily by the orographic configuration in North
673 America, particularly the Rocky Mountain Plateau (e.g., Wexler, 1961). An important
674 climatic role of the LLJ is transporting moisture from the Gulf of Mexico to the central
675 and eastern US (Benton and Estoque, 1954; Rasmusson, 1967; Helfand and Schubert,
676 1995; Byerle and Paegle, 2003). Because the moisture is essential for development of
677 precipitation, even though additional dynamic processes are required for the latter to
678 happen (Veres and Hu 2013), correctly describing the LLJ and its seasonal cycle is
679 critical for simulating and predicting warm season precipitation and climate in central
680 North America.

681 Outputs from eight of the core models (CanESM2, CCSM4, CNRM-CM5,
682 GFDL-ESM2M, HadGEM2-ES, MIROC5, MPI-ESM-LR, and MRI-CGCM3) were
683 analyzed for their simulation of the LLJ. Figure 16 compares the spatial profile and

684 seasonal cycle between the MME mean and the NCEP/NCAR reanalysis in terms of the
685 summer 925hPa winds, the vertical structure of the summer meridional wind, and the
686 seasonal cycle of the LLJ. While the overall features of the simulated LLJ compare well
687 with the reanalysis results, several details differ. First of all the models produce a peak
688 meridional wind around 925hPa whereas the reanalysis result peaks around 850hPa. This
689 difference has little impact from the vertical resolution of the models and the reanalysis
690 because they share the same vertical resolution below 500hPa. For a few models that
691 have more model levels below 500hPa their vertical profile of the meridional wind shows
692 a similar peak at 925hPa. The vertical extent of the LLJ is shallower than that shown in
693 the reanalysis, as suggested by the differences in Fig. 16f, which may be related to the
694 peak wind being at a lower level in the troposphere. Secondly, the simulated LLJ extends
695 much further northwards in the Great Plains than the reanalysis (Fig. 16g,h,i). For the
696 seasonal cycle, the models show strong southerly winds that persist from mid-May to
697 near the end of July whereas the reanalysis shows that the LLJ weakens substantially in
698 early July (Fig. 16i). While these detailed differences exist, the error statistics in Table 14
699 indicate that these eight models simulate the LLJ satisfactorily.

700

701 *5.6. Arctic/Alaska Sea Ice*

702 Since routine monitoring by satellites began in late October 1978, Arctic sea ice
703 has declined in all calendar months (e.g. Serreze et al., 2007). Trends are largest at the
704 end of the summer melt season in September with a current rate of decline through 2012
705 of -14.3% per decade. Regionally, summer ice losses have been pronounced in the
706 Beaufort, Chukchi and East Siberian seas since 2002 causing a lengthening of the ice-free

707 season. The presence of sea ice helps to protect Alaskan coastal regions from wind-driven
708 waves and warm ocean water that can weaken frozen ground. As the sea ice has retreated
709 further from coastal regions, and ice-free summer conditions are lasting for longer
710 periods of time (in some regions by more than 2 months during the satellite data record),
711 wind-driven waves, combined with permafrost thaw and warmer ocean temperatures,
712 have led to rapid coastal erosion (Mars and Houseknecht, 2007; Jones, et al., 2009).

713 While the winter ice cover is not projected to disappear in the near future, all
714 models that contributed to the IPCC 2007 report showed that as temperatures rise, the
715 Arctic Ocean would eventually become ice-free in the summer (e.g. Stroeve et al., 2007).
716 However, estimates differed widely, with some models suggesting a transition towards a
717 seasonally ice-free Arctic may happen before 2050, and others, sometime after 2100. To
718 reduce the spread some studies suggest using only models that are able to reproduce the
719 historical sea ice extent (e.g. Overland et al., 2011; Wang and Overland, 2009).

720 Historical sea ice extent (1953-2005) from 26 models during September and
721 March is presented as box and whisker plots (Figure 17), constructed from all ensemble
722 members of all models, with the width of the box representing the number of ensemble
723 members. Table 15 shows the biases for the individual models. Five climate models
724 (CanESM2, EC-EARTH, GISS-E2-R, HadGEM2-AO and MIROC4h) have mean
725 September extents that fall below the minimum observed value, with EC-EARTH, GISS-
726 E2-R and CanESM2 having more than 75% of their extents below the minimum observed
727 value. Three models (CSIRO-Mk3.6.0, FGOALS-s2, NorESM1-M) have more than 75%
728 of their extents above the maximum observed value. Overall, 14 models have mean
729 extents below the observed 1979-2005 mean September extent. During March, several

730 models fall outside the observed range of extents, with 16 models having more than 75%
731 of their extents outside the observed maximum and minimum values (8 above, 8 below).
732 Six models essentially straddle the mean observed March sea ice extent.

733 Spatial maps of March CMIP5 sea ice thickness averaged from 1993 to 2005 are
734 shown in Figure 18 together with thickness estimates from ICESat (2003-2009; Kwok
735 and Cunningham, 2008). Table 15 shows the biases for the individual models relative to
736 the ICESat data. While we do not expect the models to be in phase with the observed
737 natural climate variability and therefore accurately represent the magnitude of the ICESat
738 thickness fields, it is important to assess whether or not the models are able to reproduce
739 the observed spatial distribution of ice thickness. Data from ICESat and IceBridge, as
740 well as earlier radar altimetry missions (ERS1/2) and submarine tracks indicate that the
741 thickest ice is located north of Greenland and the Canadian Archipelago (> 5m thick)
742 where there is an onshore component of ice motion resulting in strong ridging.
743 Thicknesses are smaller on the Eurasian side of the Arctic Ocean where there is persistent
744 offshore motion of ice and divergence, leading to new ice growth in open water areas.
745 Most models fail to show thin ice close to the Eurasian coast and thicker ice along the
746 Canadian Arctic Archipelago and north coast of Greenland. Instead, many models show a
747 ridge of thick ice that spans north of Greenland across the Lomonosov Ridge towards the
748 East Siberian Shelf, with thinner ice in the Beaufort/Chukchi and the Kara/Barents seas.
749 In large part this is explained in terms of biases in the distribution of surface winds; for
750 example, if a model fails to produce a well-structured Beaufort Sea High, this will
751 adversely affect the ice drift pattern and hence the thickness pattern. Nevertheless, when
752 we compare mean thickness fields from ICESat with thickness fields from the CMIP5

753 models, we find that for the Arctic Ocean as a whole, the thickness distributions from the
754 models overlap with those from the satellite product. However, for the North American
755 side of the Arctic Ocean model thicknesses tend to be smaller than thicknesses estimated
756 from ICESat. This in part explains the low bias in September ice extent for some of the
757 models, as thinner ice is more prone to melting out in summer. Models with extensively
758 thick winter ice (e.g. NorESM1-M and MIROC5) on the other hand tend to overestimate
759 the observed September ice extent.

760

761 **6. Discussion and Conclusions**

762

763 *6.1. Synthesis of Model Performance*

764 This study evaluates the CMIP5 models for a set of basic climate and surface
765 hydrological variables for annual and seasonal means and extremes, and selected regional
766 climate features. Evaluations of model performance are not straightforward because of
767 the broad range of uses of climate model data (Gleckler et al., 2008) and therefore there is
768 not an accepted universal set of performance metrics. Issues relevant to performance are
769 dependent on several elements including decadal variability, observational uncertainties,
770 and that some models are tuned to certain processes, often at the expense of other aspects
771 of climate. The performance metrics evaluated here are generally focused on basic
772 climate variables and standard statistical measures such as bias, RMSE and spatial
773 correlation. One of the strengths of this study is the broad range of evaluations that test
774 multiple aspects of the model simulations at various time and space scales, and for
775 specific important regional features that we do not necessarily expect coarse resolution

776 models to simulate well. Independently these metrics indicate much better performance
777 by certain models relative to the ensemble, whilst some models have poor performance in
778 that a feature is not simulated at all, such as lack of persistence in extreme hydrological
779 events, or the errors are unacceptably large. However, it is not clear whether one model
780 or set of models performs better than others for the full set of climate variables.

781 Figure 19 shows a summary ranking of model performance across all continental
782 and US domain analyses presented in sections 3 and 4 in terms of biases with the
783 observational estimates. We choose not to show results for the regional processes as these
784 are generally for fewer models and only provide one sample of important features of
785 North American climate. Other metrics, such as the RMSE, could have been used, but the
786 bias values were available for all continental analyses. Model performance is shown by
787 two methods: the first is the normalized bias, calculated as the difference of the absolute
788 model bias from the lowest absolute bias value, divided by the range in absolute bias
789 values across all models. A value of 0.0 indicates the lowest absolute bias and a value of
790 1.0 indicates the highest value. The values reflect the distribution of values across models
791 such that outlier models are still identified. The second method is the rank of the sorted
792 absolute bias values, which is uniformly distributed.

793 The first thing to note is the difference in spread between the two panels in Figure
794 19 for individual metrics, which is a reflection of the two types of distribution of values
795 (model ensemble dependent and uniform). The normalized metrics highlight outlier
796 models that perform much better than the rest of the models (e.g. BCC-CSM1.1 for
797 runoff ratios (Q/P)) or much worse (e.g. INMCM4 for DJF precipitation or CanEMS2 for
798 JJA temperature). Another example is persistent precipitation events (P persist US), for

799 which there is a cluster of eight models that do equally poorly compared to the rest of the
800 models. No single model stands out as being better or worse for multiple metrics. Some
801 models do relatively well for the same variable and a single/both season across all
802 regions, such as HadGEM2-ES and MIROC5 for precipitation.

803 The rankings in the lower panel are more clustered across analyses, such as for
804 the Hadley Center models for DJF temperature, although the actual biases are generally
805 not that different to the other models. The MRI-CGCM3 is consistently ranked low for
806 runoff ratios in all regions and for the number of summer/frost days and growing season
807 length (and in terms of the normalized metrics). The INMCM4 model is consistently
808 ranked low for precipitation in both seasons and DJF temperature, although again its
809 normalized values are generally not very different from the other models. It is tempting to
810 provide an overall ranking or weighted metric across all analyses for each model, but
811 there is no obvious way of doing this for a diverse set of metrics, although this has been
812 attempted in other studies (e.g. Reichler and Kim, 2008). Nevertheless, it is useful to
813 identify those models that are ranked highly for multiple metrics. For example, the
814 following models are ranked in the top 5 for at least 12 metrics (approximately one third
815 of the total number of metrics): MPI-ESM-LR (16 metrics), GISS-E2-R (15), CCSM4
816 (14), CSIRO-Mk3.6.0 (14), BCC-CSM1-1 (12); with the bottom two models: GFDL-
817 CM3 (6) and INMCM4 (4).

818

819 *6.2. Changes in Performance between CMIP3 and CMIP5 for Basic Climate Variables*

820 A key question is whether the CMIP5 results have improved since CMIP3, and
821 why. As mentioned in the introduction, the CMIP5 models generally have higher

822 horizontal resolution and have improved parameterizations and additional process
823 representations since CMIP3. Several of the analyses presented here indicate improved
824 results since CMIP3, for example for the North American monsoon, by comparison with
825 earlier studies. Here we show a direct comparison of CMIP5 with CMIP3 results for
826 basic climate variables in Figure 20, which shows RMSE values for CMIP5 and CMIP3
827 models for seasonal precipitation and surface air temperature over North America and
828 SSTs over the surrounding oceans. 14 of the 17 core CMIP5 models have an equivalent
829 CMIP3 model, that is the same model (HadCM3), a newer version, or an earlier related
830 version, and so a direct comparison of any improvements since CMIP5 is feasible.

831 Overall, the MME mean performance has improved slightly in CMIP5 for nearly
832 all variables. For example, there is a reduction in the MME mean RMSE for summer
833 precipitation (0.90 mm day^{-1} for CMIP3, 0.86 mm day^{-1} for CMIP5), and for winter SSTs
834 ($1.72 \text{ }^{\circ}\text{C}$ to $1.55 \text{ }^{\circ}\text{C}$). The largest percentage reduction in RSME for the MME mean is for
835 summer temperatures (11.8% reduction in RMSE). The spread in model performance (as
836 quantified by the standard deviation) has remained about the same for precipitation,
837 increased for temperature and decreased for SSTs. The increase in spread for temperature
838 is due to both increases and decreases in model performance relative to the CMIP3
839 models. Several models have improved considerably and across nearly all variables and
840 seasons, such as the CCSM4, INMCM4, IPSL-CM5A-LR, and MIROC5. Reductions in
841 performance for individual models are less prevalent across variables, but are large for
842 CSIRO-Mk3.6.0, HadCM3, and MRI-CGCM3 for SSTs in both seasons. The CanESM2
843 has worse performance than its CMIP3 equivalent (CGCM3.1) for all variables, although
844 it is unclear how the two models are related. Interestingly the HadCM3 model, which is

845 used for both the CMIP3 and CMIP5 simulations, appears to have degraded in
846 performance for SSTs.

847

848 *6.3. Summary and Conclusions*

849 We have evaluated the CMIP5 multi-model ensemble for its depiction of North
850 American continental and regional climatology, with a focus on a core set of models.
851 Overall, the multi-model ensemble does reasonably well in representing the main features
852 of basic surface climate over North America and the adjoining seas. Regional
853 performance for basic climate variables is highly variable across models, however, and
854 this can bias the assessment of the ensemble because of outlier models and therefore the
855 median value may be a better representation of the central tendency of model
856 performance (Liepert and Previdi, 2012). No particular model stands out as performing
857 better than others across all analyses, although some models perform much better for sets
858 of metrics, mainly for the same variable across different regions. Higher resolution
859 models tend to do better at some aspects than others, especially for the regional features
860 as expected, but not universally so and not for basic climate variables. ESMs that
861 simulate the coupled carbon cycle and therefore have different atmospheric GHG
862 concentrations do not stand out as performing better or worse than other models.

863 There are systematic biases in precipitation with overestimation for more humid
864 and cooler regions and underestimation for drier regions. Biases in precipitation filter
865 down to biases in the surface hydrology, although this is also related to the representation
866 of the land surface in many models, with implications for assessment of water resources
867 and hydrological extremes. The poor performance in representing observed seasonal

868 persistence in precipitation and soil moisture is a reflection of this. As many of the errors
869 are systematic across models, there is potential for diagnosing these further based on a
870 multi-model analysis.

871 The models have a harder time representing extreme values, such as those based
872 on temperature and precipitation. The biases in temperature means and extremes may be
873 related to those in land hydrology that affects the surface energy balance and therefore
874 can impact how much energy goes into heating the near surface air during dry periods
875 and in drier regions. Biases in precipitation and its extremes are likely related to
876 differences in large-scale circulation and SST patterns, as well as problems in
877 representing regional climate features. Hints of this are shown in the some of the analyses
878 presented here, such as the errors in regional moisture divergence over North America,
879 but linkages between other regional climate features and terrestrial precipitation biases
880 are not apparent, such as for western Atlantic winter cyclones, and further investigation is
881 required to diagnose these. Part 2 of this paper (Sheffield et al., 2013) indicates that most
882 models have trouble representing teleconnections between modes of climate variability
883 (such as ENSO) and continental surface climate variables, and this may also reflect the
884 representation of mean climate.

885 Overall, the performance of the CMIP5 models in representing observed climate
886 features has not improved dramatically compared to CMIP3, at least for the set of models
887 and climate features analyzed here. There are some models that have improved for certain
888 features (e.g. the timing of the North American Monsoon), but others that have become
889 worse (e.g. continental seasonal surface climate).

890 The results of this paper have implications for the robustness of future projections
891 of climate and its associated impacts. Part three of this paper (Maloney et al., 2013)
892 evaluates the CMIP5 models for N. America in terms of the future projections for the
893 same set of climate features as evaluated for the 20th century in this first part and the
894 second part of the paper (Sheffield et al., 2013). Whilst model historical performance is
895 not sufficient for credible projections, the depiction of at least large scale climate features
896 is necessary. Overall, the models do well in replicating the broad scale climate of N.
897 America and some regional features, but biases in some aspects are of the same
898 magnitude as the projected changes (Maloney et al., 2013). For example, the low bias in
899 daily maximum temperature over the southern US in some models is similar to the future
900 projected changes. Furthermore, the uncertainty in the future projections across models
901 can also be of the same magnitude the model spread for the historic period.

902

903 **Acknowledgements.** We acknowledge the World Climate Research Programme's
904 Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the
905 climate modeling groups for producing and making available their model output. For
906 CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and
907 Intercomparison provides coordinating support and led development of software
908 infrastructure in partnership with the Global Organization for Earth System Science
909 Portals. The authors acknowledge the support of NOAA Climate Program Office
910 Modeling, Analysis, Predictions and Projections (MAPP) Program as part of the CMIP5
911 Task Force.

912 **References**

- 913 Adler, R. F., G. J. Huffman, A. Chang, R. Ferraro, P. Xie, J. Janowiak, B. Rudolf, U.
914 Schneider, S. Curtis, D. Bolvin, A. Gruber, J. Susskind, and P. Arkin, 2003: The
915 Version 2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation
916 Analysis (1979–Present). *J. Hydrometeor.*, **4**, 1147–1167.
- 917 Arora, V. K., J. F. Scinocca, G. J. Boer, J. R. Christian, K. L. Denman, G. M. Flato, V. V.
918 Kharin, W. G. Lee, and W. J. Merryfield, 2011: Carbon emission limits required to
919 satisfy future representative concentration pathways of greenhouse gases, *Geophys.*
920 *Res. Lett.*, **38**, L05805, doi:10.1029/2010GL046270.
- 921 Bao, Q., and co-authors, 2012: The Flexible Global Ocean-Atmosphere-Land System
922 model Version: FGOALS-s2. *Adv. Atmos. Sci.*, submitted.
- 923 Bi, D., M. Dix, S. Marsland, T. Hirst, S. O’Farrell and coauthors, 2012: ACCESS: The
924 Australian Coupled Climate Model for IPCC AR5 and CMIP5. AMOS conference,
925 2012, Sydney, Australia (available online at
926 <https://wiki.csiro.au/confluence/display/ACCESS/ACCESS+Publications>)
- 927 Blackadar, A. K., 1957: Boundary layer wind maxima and their significance for their
928 growth of nocturnal inversions. *Bull. Amer. Meteor. Soc.*, **38**, 283–290.
- 929 Bonner, W. D., 1968: Climatology of the low level jet. *Mon. Wea. Rev.*, **96**, 833-850.
- 930 Byerle, L. A., and J. Paegle, 2003: Modulation of the Great Plains low-level jet and
931 moisture transports by orography and large scale circulations. *J. Geophys. Res.*, **108**,
932 8611, doi:10.1029/2002JD003005.

933 Caesar, J., L. Alexander, and R. Vose, 2006: Large-scale changes in observed daily
934 maximum and minimum temperatures: Creation and analysis of a new gridded data
935 set. *J. Geophys. Res.*, **111**, D05101, doi:10.1029/2005JD006280

936 Chylek, P., J. Li, M. K. Dubey, M. Wang, and G. Lesins, 2011: Observed and model
937 simulated 20th Century Arctic temperature variability: Canadian Earth System Model
938 CanESM2. *Atmospheric Chemistry and Physics Discussions*, **11** (8), 22,893–22,907,
939 doi: 10.5194/acpd-11-22893-2011

940 Colle, B. A., Z. Zhang, K. Lombardo, P. Liu, E. Chang, M. Zhang, and S. Hameed, 2013:
941 Historical and future predictions of eastern North America and western Atlantic
942 extratropical cyclones in CMIP5 during the cool season. *J. Climate*, submitted.

943 Collins, M., S. F. B Tett, and C. Cooper, 2001: The internal climate variability of
944 HadCM3, a version of the Hadley Centre Coupled Model without flux adjustments.
945 *Climate Dynamics*, **17** (1), 61-81.

946 Compo, G. P., J. S. Whitaker, P. D. Sardeshmukh, N. Matsui, R. J. Allan, X. Yin, B. E.
947 Gleason, R. S. Vose, G. Rutledge, P. Bessemoulin, S. Brönnimann, M. Brunet, R. I.
948 Crouthamel, A. N. Grant, P. Y. Groisman, P. D. Jones, M. Kruk, A. C. Kruger, G. J.
949 Marshall, M. Maugeri, H. Y. Mok, Ø. Nordli, T. F. Ross, R. M. Trigo, X. L. Wang, S.
950 D. Woodruff, and S. J. Worley, 2011: The Twentieth Century Reanalysis Project.
951 *Quarterly J. Roy. Meteorol. Soc.*, **137**, 1-28. DOI: 10.1002/qj.776.

952 Donner, L. J., with 28 co-authors, 2011: The dynamical core, physical parameterizations,
953 and basic simulation characteristics of the atmospheric component AM3 of the GFDL
954 Global Coupled Model CM3. *J. Climate*, **24**, doi:10.1175/2011JCLI3955.1.

955 Dufresne, J-L., and 58 co-authors, N, 2012: Climate change projections using the IPSL-
956 CM5 Earth System Model: from CMIP3 to CMIP5, *Clim. Dyn.*, submitted.

957 Fetterer, F., K. Knowles, W. Meier, and M. Savoie, 2002, updated 2009. /Sea Ice Index/.
958 Boulder, Colorado USA: National Snow and Ice Data Center. Digital media.

959 Francis, J. A., and S. J. Vavrus, 2012: Evidence linking Arctic amplification to extreme
960 weather in mid-latitudes. *Geophys. Res. Lett.*, **39**, L06801,
961 doi:10.1029/2012GL051000.

962 Frich, P., L. V. Alexander, P. Della-Marta, B. Gleason, M. Haylock, A. M. G. Klein
963 Tank, and T. Peterson, 2002: Observed coherent changes in climatic extremes during
964 the second half of the twentieth century. *Climate Res.*, **19**, 193–212.

965 Geil, K. L., Y. L. Serra, and X. Zeng, 2013: Assessment of CMIP5 model simulation of
966 the North American Monsoon System. *J. Climate*, submitted.

967 Gent, Peter R., and Coauthors, 2011: The Community Climate System Model Version 4.
968 *J. Climate*, **24**, 4973–4991. doi: <http://dx.doi.org/10.1175/2011JCLI4083.1>

969 Giorgi, F. and R. Francisco, 2000: Uncertainties in regional climate change prediction: a
970 regional analysis of ensemble simulations with the HADCM2 coupled AOGCM.
971 *Climate Dynamics*, **16** (2-3), 169–182, doi:10.1007/PL00013733.

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

974 Hall, A., Qu, X., and Neelin, J. D., 2008: Improving predictions of summer climate
975 change in the United States, *Geophys. Res. Lett.*, **35**, L01702,
976 DOI:10.1029/2007GL032012

977 Hazeleger, W., and 31 co-authors, 2010: EC-Earth: A seamless Earth system prediction
978 approach in action. *Bull. Amer. Meteor. Soc.*, **91**, 1357-1363, doi:
979 10.1175/2010BAMS2877.1

980 Helfand, H. M., and S. D. Schubert, 1995: Climatology of the simulated Great Plains
981 low-level jet and its contribution to the continental moisture budget of the United
982 States. *J. Climate*, **8**, 784–806.

983 Higgins, R. W., J. E. Janowiak and Y.-P. Yao, 1996: A gridded hourly precipitation data
984 base for the United States (1963-1993). NCEP/Climate Prediction Center Atlas No. 1,
985 U. S. Department of Commerce, National Oceanic and Atmospheric Administration,
986 National Weather Service.

987 Hodges, K. I., 1994: A general method for tracking analysis and its application to
988 meteorological data. *Mon. Wea. Rev.*, **122**, 2573–2586.

989 Hodges, K. I., 1995: Feature tracking on the unit sphere. *Mon. Wea. Rev.*, **123**, 3458–
990 3465.

991 Holton, J. R., 1967: The diurnal boundary layer wind oscillation above sloping terrain.
992 *Tellus*, **19**, 199–205.

993 Huffman, G. J., R. F. Adler, D. T. Bolvin, G. Gu, E. J. Nelkin, K. P. Bowman, Y. Hong,
994 E. F. Stocker, D. B. Wolff, 2007: The TRMM Multi-satellite Precipitation Analysis:
995 Quasi-Global, Multi-Year, Combined-Sensor Precipitation Estimates at Fine Scale.
996 *J. Hydrometeor.*, **8** (1), 38–55.

997 Jones, B. M., C. D. Arp, M. T. Jorgenson, K. M. Hinkel, J. A. Schmutz, and P. L. Flint ,
998 2009: Increase in the rate and uniformity of coastline erosion in Arctic Alaska.
999 *Geophys. Res. Lett.*, **36**, L03503, doi:10.1029/2008GL036205

1000 Jones, C. D., and others, 2011: The HadGEM2-ES implementation of CMIP5 centennial
1001 simulations, *Geosci. Model Dev.*, **4**, 543-570, doi:10.5194/gmd-4-543-2011.

1002 Kalnay E., Kanamitsu M., Kistler R., Collins W., Deaven D., Gandin L., Iredell M., Saha
1003 S., White G., Woollen J., Zhu Y., Chelliah M., Ebisuzaki W., Higgins W., Janowiak
1004 J., Mo K.C., Ropelewski C., Wang J., Leetman A., Reynolds R., Jenne R., Joseph D.,
1005 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.*, **77**,
1006 437–471.

1007 Kanamitsu M., W. Ebisuzaki, J. S. Woollen, S.-K. Yang, J. J. Hnilo, M. Fiorino, and G. L.
1008 Potter, 2002: NCEP–DOE AMIP II Reanalysis (R-2). *Bull. Amer. Meteor. Soc.*, **83**,
1009 1631-1643.

1010 Kim, D., A. H. Sobel, A. D. Del Genio, Y. Chen, S. Camargo, M.-S. Yao, M. Kelley, and
1011 L. Nazarenko, 2012: The tropical subseasonal variability simulated in the NASA
1012 GISS general circulation model, *J. Clim.*, in press.

1013 Kwok, R., and G. F. Cunningham, 2008: ICESat over Arctic sea ice: Estimation of snow
1014 depth and ice thickness. *J. Geophys. Res.*, **113**, C08010, doi:10.1029/2008JC004753.

1015 Legates D. R., and G. J. McCabe, 1999: Evaluating the use of “goodness-of-fit” measures
1016 in hydrologic and hydroclimatic model validation. *Water Resources Research*, **35**:
1017 233-241.

1018 Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges, 1994: A simple
1019 hydrologically based model of land surface water and energy fluxes for GCMs. *J.*
1020 *Geophys. Res.*, **99**, 14,415-14,428.

1021 Liang, X.-Z., J. Zhu, K. E. Kunkel, M. Ting, and J. X. L. Wang, 2008: Do CGCMs
1022 simulate the North American monsoon precipitation seasonal-interannual variations.
1023 *J. Climate*, **21**, 3755-3775.

1024 Liepert, B. G., and M. Previdi, 2012: Inter-model variability and biases of the global
1025 water cycle in CMIP3 coupled climate models. *Environ. Res. Lett.*, **7** 014006
1026 doi:10.1088/1748-9326/7/1/014006

1027 Maloney, E. D., S. J. Camargo, E. Chang, B. Colle, R. Fu, K. L. Geilw, Q. Hu, X. Jiang,
1028 N. Johnson, K. B. Karnauskas, J. Kinter, B. Kirtman, S. Kumar, B. Langenbrunner,
1029 K. Lombardo, L. Long, A. Mariotti, J. E. Meyerson, K. Mo, J. D. Neelin, Z. Pan, R.
1030 Seager, Y. Serraw, A. Seth, J. Sheffield, J. Thibeault, S.-P. Xie, C. Wang, B. Wyman,
1031 and M. Zhao, 2011: North American Climate in CMIP5 Experiments: Part III:
1032 Assessment of 21st Century Projections. *J. Climate*, submitted.

1033 Mars, J. C., and D. W. Houseknecht, 2007: Quantitative remote sensing study indicates
1034 doubling of coastal erosion rate in past 50 yr along a segment of the Arctic coast of
1035 Alaska. *Geology*, **35** (7), 583-586. doi: 10.1130/G23672A.1

1036 McCabe, G. J., T. R. Ault, B. I. Cook, J. L. Betancourt, and M. D. Schwartz, 2012:
1037 Influences of the El Nino Southern Oscillation and the Pacific Decadal Oscillation on
1038 the times of the North American spring. *Int. J. Climatol.*, **32**: 2301-2310.

1039 Merryfield, W. J., W.-S. Lee, G. J. Boer, V. V. Kharin, J. F. Scinocca, G. M. Flato, and S.
1040 Polavarapu, 2012: The Canadian seasonal to interannual prediction system. Part I:
1041 Models and initialization. *Monthly Weather Review*, submitted.

1042 Mitchell, M. J., R. W. Arritt, and K. Labas, 1995: A climatology of the warm season
1043 Great Plains low-level jet using wind profiler observations. *Wea. Forecasting*, **10**,
1044 576–591.

1045 Mitchell, T. D., and P. D. Jones, 2005: An improved method of constructing a database of
1046 monthly climate observations and associated high-resolution grids. *Int. J. Climatol.*,
1047 **25**, 693–712.

1048 Overland, J. E., 2011: Potential Arctic change through climate amplification processes.
1049 *Oceanography*, **24** (3), 176-185. <http://dx.doi.org/10.5670/oceanog.2011.70>

1050 Rasmusson, E.M., 1967: Atmospheric water vapor transport and the water balance of the
1051 North America. Part I: Characteristics of the water vapor flux field. *Mon. Wea. Rev.*,
1052 **95**, 403-426.

1053 Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell,
1054 E. C. Kent, and A. Kaplan, 2003: Globally complete analyses of sea surface
1055 temperature, sea ice and night marine air temperature, 1871-2000. *J. Geophys. Res.*,
1056 **108** (4407).

1057 Reichler, T., and J. Kim, 2008: How well do coupled models simulate today's climate?
1058 *Bull. Amer. Meteor. Soc.*, 89, 303-311

1059 Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull.*
1060 *Amer. Meteor. Soc.*, **91**, 1015–1057. doi: <http://dx.doi.org/10.1175/2010BAMS3001.1>

1061 Sakamoto, T. T., Y. Komuro, T. Nishimura, M. Ishii, H. Tatebe, H. Shiogama, A.
1062 Hasegawa, T. Toyoda, M. Mori, T. Suzuki, Y. Imada, T. Nozawa, K. Takata, T.
1063 Mochizuki, K. Ogochi, S. Emori, H. Hasumi and M. Kimoto, 2012: MIROC4h - a

1064 new high-resolution atmosphere-ocean coupled general circulation model. *J. Meteor.*
1065 *Soc. Japan*, **90** (3), 325-359.

1066 Sheffield, J., G. Goteti, and E. F. Wood, 2006: Development of a 50-yr high-resolution
1067 global dataset of meteorological forcings for land surface modeling, *J. Climate*, **19**
1068 (13), 3088-3111.

1069 Sheffield, J. and E. F. Wood, 2007: Characteristics of global and regional drought, 1950-
1070 2000: Analysis of soil moisture data from off-line simulation of the terrestrial
1071 hydrologic cycle. *J. Geophys. Res.*, **112** (D17), doi:10.1029/2006JD008288656

1072 Sheffield, J., S. J. Camargo, B. Colle, Q. Hu, X. Jiang, N. Johnson, S. Kumar, K.
1073 Lombardo, B. Langenbrunner, E. Maloney, J. E. Meyerson, J. D. Neelin, Y. L. Serra,
1074 D.-Z. Sun, C. Wang, S.-P. Xie, J.-Y. Yu, T. Zhang, 2012: North American Climate in
1075 CMIP5 Experiments: Part II: Evaluation of 20th Century Intra-Seasonal to Decadal
1076 Variability, *J. Climate*, submitted.

1077 Stroeve, J., M. M. Holland, W. Meier, T. Scambos, and M. Serreze, 2007: Arctic sea ice
1078 decline: Faster than forecast. *Geophys. Res. Lett.*, **34**, L09501,
1079 doi:10.1029/2007GL029703

1080 Stroeve, J. C., V. Kattsov, A. Barrett, M. Serreze, T. Pavlova, M. Holland, and W. N.
1081 Meier, 2012: Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations,
1082 *Geophys. Res. Lett.*, doi: 10.1029/2012GL052676R, in press.

1083 Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the
1084 experiment design. *Bull. Am. Meteorol. Soc.*, **93**, 485–498, doi:10.1175/BAMS-D-11-
1085 00094.1.

1086 Universidad Nacional Autónoma de México (UNAM), 2007: Gridded precipitation and
1087 temperature analysis from the Centro de Ciencias de la Atmósfera, Mexico; available
1088 from the International Research Institute for Climate and Society
1089 (<http://ingrid.ldeo.columbia.edu/SOURCES/UNAM/gridded/monthly/v0705/>).

1090 Veres, M. C., and Q. Hu, 2012: AMO-forced regional processes affecting summertime
1091 precipitation variations in the central United States. *J. Climate*, in press.

1092 Voltaire, A., and others, 2012: The CNRM-CM5.1 global climate model: Description
1093 and basic evaluation, *Clim. Dyn.*, doi:10.1007/s00382-011-1259-y, in press.

1094 Volodin, E. M., N. A. Diansky, and A. V. Gusev, 2010: Simulating Present-Day Climate
1095 with the INMCM4.0 Coupled Model of the Atmospheric and Oceanic General
1096 Circulations. *Izvestia, Atmospheric and Oceanic Physics*, **46**, 414-431

1097 Vose, R. S., R. L. Schmoyer, P. M. Steurer, T. C. Peterson, R. Heim, T. R. Karl, and J.
1098 Eischeid, 1992: The Global Historical Climatology Network: long-term monthly
1099 temperature, precipitation, sea level pressure, and station pressure data.
1100 ORNL/CDIAC-53, NDP-041. Carbon Dioxide Information Analysis Center, Oak
1101 Ridge National Laboratory, Oak Ridge, Tennessee.

1102 Wang, C., and D. B. Enfield, 2001: The tropical Western Hemisphere warm pool.
1103 *Geophys. Res. Lett.*, **28**, 1635-1638.

1104 Wang, M. and J. E. Overland (2009), A sea ice free summer Arctic within 30 years?,
1105 *Geophys. Res. Lett.*, **36**, L07502, doi:10.1029/2009GL037820.

1106 Wang, A., T. J. Bohn, S. P. Mahanama, R. D. Koster, and D. P. Lettenmaier, 2009:
1107 Multimodel ensemble reconstruction of drought over the continental United States. *J.*
1108 *Climate*, **22**, 2694–2712.

1109 Watanabe, M., and Coauthors, 2010: Improved Climate Simulation by MIROC5: Mean
1110 States, Variability, and Climate Sensitivity. *J. Climate*, **23**, 6312–6335.

1111 Wexler, H., 1961: A boundary layer interpretation of the low level jet. *Tellus*, **13**, 368–
1112 378.

1113 Xie, P., and P.A. Arkin, 1997: Global precipitation: A 17-year monthly analysis based on
1114 gauge observations, satellite estimates, and numerical model outputs. *Bull. Amer.*
1115 *Meteor. Soc.*, **78**, 2539 - 2558.

1116 Xie, P., M. Chen, and W. Shi, 2010: CPC unified gauge-based analysis of global daily
1117 precipitation. Preprints, *24th Conf. on Hydrology*, Atlanta, GA, Amer. Meteor. Soc.,
1118 2.3A

1119 Xin X., Wu T., Zhang J., 2012: Introductions to the CMIP 5 simulations conducted by the
1120 BCC climate system model (in Chinese). *Advances in Climate Change Research*.
1121 submitted.

1122 Yukimoto, S., et al., 2012: A new global climate model of the Meteorological Research
1123 Institute: MRI-CGCM3—Model description and basic performance, *J. Meteorol. Soc.*
1124 *Jpn.*, **90a**, 23–64.

1125 Zanchettin, D., A. Rubino, D. Matei, O. Bothe, and J. H. Jungclaus, 2012: Multidecadal-
1126 to-centennial SST variability in the MPI-ESM simulation ensemble for the last
1127 millennium. *Clim. Dyn.*, **39**, 419-444 doi:10.1007/s00382-012-1361-9.

1128 Zhang, Z. S., Nisancioglu, K., Bentsen, M., Tjiputra, J., Bethke, I., Yan, Q.,
1129 Risebrobakken, B., Andersson, C., and Jansen, E., 2012: Pre-industrial and mid-
1130 Pliocene simulations with NorESM-L. *Geosci. Model Dev.*, **5**, 523-533,
1131 doi:10.5194/gmd-5-523-2012.

1132 **Figure Captions**

1133 **Figure 1.** Precipitation climatology (mm/day) for (left) December-February and (right)
1134 June-August (1979-2005). a) GPCP estimate of observed precipitation for DJF. b) MME
1135 mean over the 18 models for DJF; for models with multiple runs, all runs are averaged
1136 before inclusion in the multi-model ensemble. c) Comparison of individual models to
1137 observations using the 3 mm day⁻¹ contour as an index of the major precipitation features:
1138 half the models are shown in each of sub-panel I and II with the legend giving the color-
1139 coding for the models in each. Shading shows the regions where GPCP exceeds 3 mm
1140 day⁻¹; a model with no error would have its contour fall exactly along the edge of the
1141 shaded region. d)-f) As in a)-c), respectively, except for JJA. The model data were
1142 regridded to the resolution of the GPCP observations (2.5-degree).

1143 **Figure 2.** Surface air temperature climatology (°C) for (top) December-February and
1144 (bottom) June-August (1979-2005). a) MME mean (over the 17 core models plus
1145 FGOALS-s2) for DJF. b) NCEP-DOE Reanalysis 2 estimate of observed surface air
1146 temperature climatology for DJF. c) As in b) but for CRU. d) Standard deviation of
1147 surface air temperature among the 18 model DJF climatological values at each point. e)
1148 Difference between the MME mean climatology in a) and the NCEP-DOE Reanalysis 2.
1149 f) As in e) but for CRU. g)-l) As in a)-f) but for JJA. All observation and model data were
1150 interpolated to the same 2.5° grid.

1151 **Figure 3.** Climatological sea surface temperature and precipitation in observations from
1152 HadISSTv1.1 and GPCPv2.2 data sets, and historical simulations from 17 CMIP5 models
1153 for 1979-2004. (a) Observations, (b) MME mean, and (c) MME mean minus
1154 observations, for winter-to-spring (December to May). (d-f) as in (a-c) but for summer-
1155 to-fall (April to November). Temperatures are shaded blue/red for values equal or
1156 lower/larger than 23/24°C; the thick black line highlights the 28.5°C isotherm as
1157 indicator of the Western Hemisphere Warm Pool. Precipitation is shaded green for values
1158 equal or larger than 2 mm day⁻¹. Contour intervals are 1°C and 1 mm day⁻¹ for the mean
1159 values and 0.2°C and 2 mm day⁻¹ for the differences. The SST/precipitation fields were
1160 regridded to a common 5°×2.5°/2.5°×2.5° grid.

1161 **Figure 4.** Vertically integrated moisture transport (vectors) and its divergence (contours)
1162 for the 20CR reanalysis (a,g) and five CMIP5 models for mean JJA (b-f) and DJF (h-l)
1163 for 1981-2000. Vertically integrated moisture transport is computed to 500 hPa using 6-
1164 hourly data from the 20CR and one realization each from the historical experiments for
1165 CanESM2, CCSM4, CNRM-CM5, GFDL-ESM2M, and MIROC5 models. The model
1166 data are shown at their original resolution.

1167 **Figure 5.** Mean seasonal cycle (1979-2005) of North American regional land water
1168 budget components for 12 CMIP5 models (CanESM2, CSIRO-Mk3.6.0, GFDL-ESM2G,
1169 GISS-E2-H, GISS_E2-R, IPSL-CM5A-LR, IPSL-CM5A-MR, MIROC-ESM, MIROC-
1170 ESM-CHEM, MPI-ESM-LR, MRI-CGCM3, NorESM1-M) compared to the average of
1171 the two off-line LSM simulations (VIC and GLDAS2 Noah). Regions are Western North
1172 America (WNA), Central North America (CNA), Eastern North America (ENA), Alaska
1173 and Western Canada (ALA), Northeast Canada (NEC), and Central America (CAM) as

1174 modified from Giorgi and Francisco (2000) and shown in supplementary Fig. S3. All data
1175 were interpolated to 2.5-degree resolution.

1176 **Figure 6.** Mean annual runoff (mm/year) (top) and runoff (Q/P) ratio (bottom) for 1979-
1177 2004 from observations (Q from VIC and GLDAS2 Noah, and P from GPCP) and the
1178 multi-model average from 15 CMIP5 climate models (BCC-CSM1-1, CanESM2,
1179 CCSM4, CNRM-CM5, CSIRO-Mk3.6.0, GFDL-CM3, GFDL-ESM2M, GISS-E2-R,
1180 HadCM3, INMCM4, MIROC5, MIROC-ESM, MPI-ESM-LR, MRI-CGCM3,
1181 NorESM1-M). All data were interpolated to 2.5-degree resolution.

1182 **Figure 7.** Comparison of biophysical indicators between observations and the CMIP5
1183 ensemble. Biophysical indicators are (top row) number of summer days, (middle row)
1184 number of frost days, and (bottom row) growing season length averaged over 1979-2005.
1185 Left column shows the observations (left column) from the HadGHCND dataset; middle
1186 column is the multi-model ensemble mean of the 17 core models, and the right column is
1187 their difference (MME – obs.). The frequencies were calculated on the model grid and
1188 then interpolated to 2.0-degree resolution for comparison with the observational
1189 estimates.

1190 **Figure 8.** The frequency of occurrence of persistent extreme precipitation events defined
1191 by SPI6 averaged over positive and negative events for (a) observed precipitation based
1192 on the CPC and UW datasets, (b) BCC-CSM1-1, (c) CanESM2, (d) CCSM4, (e) CNRM-
1193 CM5.1, (f) CSIRO-Mk3.6.0, (g) GFDL-CM3, (h) GISS-E2-R, (i) HadCM3, (j) IPSL-
1194 CM5A-LR, (k) MIROC5, (l) MIROC-ESM, (m) MPI-ESM-LR, (n) MRI-CGCM3 and
1195 (o) NorESM1-M. The HadCM3 and HadGEM2-ES results are similarly weak and so the
1196 former are shown only. Each data set is treated as one member of the ensemble and
1197 frequencies were calculated at the resolution of the dataset.

1198 **Figure 9.** Same as Figure 8 but for persistent soil moisture events. Estimates of observed
1199 soil moisture are taken from the multi-model NLDAS-UW dataset.

1200 **Figure 10.** (a) Cyclone density for the CFSR analysis showing the number of cyclones
1201 per cool season (November to March) per 50,000 km² for 1979-2004. (b) Same as (a)
1202 except for the mean (shaded) and spread (contoured every 0.3) of 15 CMIP5 models
1203 ordered from higher to lower spatial resolution: CanESM2, EC-EARTH, MRI-
1204 CGCM3, CNRM-CM5, MIRCO5, HadGEM2-ES, HadGEM2-CC, INMCM4, IPSL-
1205 CM5A-MR, MPI-ESM-LR, NorESM1-M, GFDL-ESM2M, IPSL-CM5A-LR, BCC-
1206 CSM1, MIROC-ESM-CHEM. Same as (a) except for the (c) MPI-ESM-LR, (d) GFDL-
1207 ESM2M, (e) HadGEM2-CC, and (f) CCSM4 models. The cyclone densities were
1208 calculated on the native grid of the dataset.

1209 **Figure 11.** Number of cyclone central pressures at their maximum intensity (minimum
1210 pressure) for the 1979-2004 cool seasons within the dashed box region in Fig. 10 for a 10
1211 hPa range centered every 10 hPa showing the CFSR (bold blue), (b) CMIP5 MME mean
1212 (bold red), and individual CMIP5 models.

1213 **Figure 12.** (a) CPC merged precipitation analysis at 2.5 deg resolution showing cool
1214 seasonal average precipitation (shaded every 75 mm) for the 1979-2004 cool seasons

1215 (November – March). (b) Same as (a) except for the CPC Unified precipitation at 0.5 deg
1216 resolution. (c) Same as (a) except for the mean of 14 of the 17 CMIP5 members listed in
1217 (d) and spread (in mm). (d) Number of days that the daily average precipitation (in
1218 mm/day) for the land areas in the black box in (b) occurred within each amount bin for
1219 select CMIP5 members, CMIP5 mean, and the CPC Unified. The CPC and model data
1220 were regridded to 1.0-degree resolution.

1221 **Figure 13.** Comparison of precipitation and temperature extremes for southern US
1222 regions between the CMIP5 models and CPC and GHCN observations, respectively, for
1223 1979-2005. (left column) Taylor diagram of the spatial pattern of annual number of days
1224 when precipitation $> 10\text{mm day}^{-1}$ over the southwest (SW), south central (SC) and
1225 southeastern (SE) US. (right column) Taylor diagram of the spatial pattern of annual
1226 number of days when $T_{\text{max}} > 32^{\circ}\text{C}$ (90°F) for the three regions. The standard deviations
1227 have been normalized relative to the observed values. (A: CanESM2, B: CCSM4, C:
1228 GFDL-CM3, D: GFDL-ESM2G, E: GFDL-ESM2M, F: GISS-E2-R, G: HadCM3, H:
1229 HadGEM2-CC, I: HadGEM2-ES, J: IPSL-CM5A-LR, K: MIROC4h, L: MIROC5, M:
1230 MPI-ESM-LR, N: MRI-CGCM3). SW is defined as the contiguous US south of 40°N
1231 between 125°W and 110°W ; SC is the contiguous US south of 40°N between 110°W and
1232 90°W ; SE is the contiguous US south of 40°N between 90°W and 70°W . All data were
1233 regridded to 2.5-degree resolution.

1234 **Figure 14.** Average monthly precipitation for 1979-2005 shown by latitude in the North
1235 American monsoon region (longitudes 102.5 to 115°W) from the CMAP observational
1236 estimate (a), the MME mean for the 17 core CMIP5 models (b) and their difference (c),
1237 all in units of mm day^{-1} . The CMAP and model data were regridded to T42 resolution
1238 (~ 2.8 -degree).

1239 **Figure 15.** Annual cycle in rainfall for the North American monsoon region for the
1240 historical (1979-2005) period of 21 CMIP5 models compared to the P-NOAA AND
1241 CMAP observational datasets for (a) small (phase error = 0), (b) moderate (phase error =
1242 1), (c) large (phase error = 2-4) phase errors, and (d) the MME mean and range (shading)
1243 for all models. The CMAP and model data were regridded to the P-NOAA resolution
1244 (0.5-degree).

1245 **Figure 16.** (a)-(c) Averaged summer 925hPa wind during 1971-2000 for NCEP-NCAR
1246 reanalysis, eight-model CMIP5 ensemble mean for the same period, and the reanalysis
1247 minus MME mean, respectively. (d)-(f) Lower troposphere mean vertical profile of
1248 meridional wind averaged over 95° - 100°W for the reanalysis, MME mean, and the
1249 reanalysis minus MME mean, respectively. (g)-(i) Seasonal cycle of the 925hPa
1250 meridional wind averaged over 27.5° - 32.5°N for the reanalysis, MME mean, and the
1251 reanalysis minus MME mean. All units are m s^{-1} . Shading indicates wind speeds greater
1252 than 3.0 m s^{-1} in the figures of the first and second columns and wind speeds greater than
1253 1.0 m s^{-1} in the figures of the third column. The data were regridded to 2.5-degree
1254 resolution.

1255 **Figure 17.** September and March sea ice extent from 26 CMIP5 models compared to
1256 observations from the NSIDC from 1953 to 2005. For each model, the boxes represent

1257 inter-quartile ranges (25th to 75th percentiles). Median (50th percentile) extents are
1258 shown by the thick horizontal bar in each box. The width of each box corresponds to the
1259 number of ensemble members for that model. Whiskers (vertical lines and thin horizontal
1260 bars) represent the 10th and 90th percentiles. Mean monthly extents are shown as
1261 diamonds. Corresponding mean, minimum and maximum observed extends are shown as
1262 red and green lines, respectively.

1263 **Figure 18.** Observed March ice thickness (m) from ICESat for 2003 to 2009 (top),
1264 ensemble mean ice thickness from 26 CMIP5 models (see Table 15 for list of models)
1265 averaged over 1993 to 2005 with contours showing the MME standard deviation (middle)
1266 and the difference between the modeled and observed (CMIP5 – observed) thickness
1267 (bottom). The model data were regridded to the ICESat resolution.

1268 **Figure 19.** Comparison of CMIP5 models across a set of continental performance metrics
1269 based on bias values given in Tables 3-8. (top) Biases normalized relative to the range of
1270 bias values across models, with lower values indicating lower bias. (bottom) Models
1271 ranked according to bias values, with 1 indicating the model with the lowest bias and 17
1272 the model with the highest bias. Results for models without available data are indicated in
1273 white. The bias metrics shown (in order from left to right) are for regional precipitation
1274 (P) for DJF and JJA, regional temperature (T) for DJF and JJA, annual SSTs for
1275 surrounding oceans (see Figure 3), annual runoff ratios (Q/P), the annual number of
1276 summer days (SuDays), frost days (FrDays) and growing season length (GSL), and east-
1277 west gradient in the number of persistent precipitation (P Persist) and soil moisture (SM
1278 Persist) events.

1279 **Figure 20.** Comparison of CMIP5 and CMIP3 model performance for seasonal (DJF and
1280 JJA) precipitation (P), surface air temperature (T) and SST. Results are shown as RMSE
1281 values calculated for 1971-1999 relative to the GPCP, CRU and HadISST observational
1282 datasets. Precipitation and temperature RMSE values are calculated over North America
1283 (130-60W, 0-60N) and SST RMSE values are calculated over neighboring oceans (170-
1284 35W, 10S-40N). The core set of CMIP5 models and their equivalent CMIP3 models
1285 where available (otherwise indicated by N/A) are shown. The MME mean values are also
1286 shown.

1287

1288

1289

1290

1291

1292 **Table 1.** CMIP5 models evaluated and their attributes. The core models are highlighted
1293 with an asterix. Model types are: Ocean-Atmosphere coupled (OA), Ocean-Atmosphere-
1294 Chemistry coupled (ChemOA), Earth System Model (ESM), and Earth System Model
1295 Chemistry coupled (ChemESM).

Model	Center	Type	Atmospheric Horizontal Resolution (lon. x lat.)	Number of model levels	Reference
ACCESS1-0	Commonwealth Scientific and Industrial Research Organization/Bureau of Meteorology, Australia	AO	1.875 x 1.25	38	Bi et al. (2012)
BCC-CSM1.1*	Beijing Climate Center, China Meteorological Administration, China	ESM	2.8 x 2.8	26	Xin et al. (2012)
CanCM4	Canadian Centre for Climate Modelling and Analysis, Canada	AO	2.8 x 2.8	35	Merryfield et al. (2012)
CanESM2*	Canadian Center for Climate Modelling and Analysis, Canada	ESM	2.8 x 2.8	35	Arora et al. (2011)
CCSM4*	National Center for Atmospheric Research, USA	AO	1.25 x 0.94	26	Gent et al. (2011)
CESM1-CAM5-1-FV2	Community Earth System Model Contributors (NSF-DOE-NCAR)	AO	1.4 x 1.4	26	Gent et al. (2011)
CNRM-CM5.1*	National Centre for Meteorological Research, France	AO	1.4 x 1.4	31	Voltaire et al. (2011)
CSIRO-Mk3.6.0*	Commonwealth Scientific and Industrial Research Organization/Queensland Climate Change Centre of Excellence, AUS	AO	1.8 x 1.8	18	Rotstayn et al. (2010)
EC-EARTH	EC-EARTH consortium	AO	1.125 x 1.12	62	Hazeleger et al. (2010)
FGOALS-S2.0	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences	AO	2.8 x 1.6	26	Bao et al. (2012)
GFDL-CM3*	NOAA Geophysical Fluid Dynamics Laboratory, USA	AO	2.5 x 2.0	48	Donner et al. (2011)
GFDL-ESM2G/M*	NOAA Geophysical Fluid Dynamics Laboratory, USA	ESM	2.5 x 2.0	48	Donner et al. (2011)

GISS-E2-H/R*	NASA Goddard Institute for Space Studies, USA	ChemAO	2.5 x 2.0	40	Kim et al. (2012)
HadCM3*	Met Office Hadley Centre, UK	AO	3.75 x 2.5	19	Collins et al. (2001)
HADGEM2-CC	Met Office Hadley Centre, UK	ESM	1.875 x 1.25	60	Jones et al. (2011)
HadGEM2-ES*	Met Office Hadley Centre, UK	ChemESM	1.875 x 1.25	60	Jones et al. (2011)
INMCM4*	Institute for Numerical Mathematics, Russia	AO	2 x 1.5	21	Volodin et al. (2010)
IPSL-CM5A-LR*	Institut Pierre Simon Laplace, France	ChemESM	3.75 x 1.8	39	Dufresne et al. (2012)
IPSL-CM5A-MR	Institut Pierre Simon Laplace, France	ChemESM	2.5 x 1.25	39	Dufresne et al. (2012)
MIROC4h	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology, Japan	AO	0.56 x 0.56	56	Sakamoto et al. (2012)
MIROC5*	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology, Japan	AO	1.4 x 1.4	40	Watanabe et al. (2010)
MIROC-ESM*	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies	ESM	2.8 x 2.8	80	Watanabe et al. (2010)
MIROC-ESM-CHEM	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies	ChemESM	2.8 x 2.8	80	Watanabe et al. (2010)
MPI-ESM-LR*	Max Planck Institute for Meteorology, Germany	ESM	1.9 x 1.9	47	Zanchettin et al. (2012)
MRI-CGCM3*	Meteorological Research Institute, Japan	AO	1.1 x 1.1	48	Yukimoto et al. (2011)
NorESM1-M*	Norwegian Climate Center, Norway	ESM	2.5 x 1.9	26	Zhang et al. (2012)

1296

1297

1298 **Table 2.** Observational and reanalysis datasets used in the evaluations

Dataset	Type	Spatial Domain	Temporal Domain	Reference
<i>Precipitation</i>				
CMAP v2	Gauge/satellite	2.5 deg, global	Monthly/pentad, 1979-present	Xie and Arkin, 1997
GPCP v2.1	Gauge/satellite	1.0 deg, global	Monthly, 1979-2009	Adler et al., 2003
CRU TS3.1	Gauge	0.5 deg, global land	Monthly, 1901-2008	Mitchell and Jones (2005)
CMAP v2	Gauge/satellite	2.5 deg, global	Monthly/pentad, 1979-present	Xie and Arkin, 1997
CPC unified	Gauge	0.5 deg, US	Daily, 1948-2010	Xie et al., 2010
CPC-US-Mexico	Gauge	1.0 deg, US/Mexico	Daily, 1948-present	Higgins et al. (1996)
UW	Gauge	0.5 deg, US	Daily, 1916-2009	Maurer et al. (2002)
P-NOAA	Gauge	0.5 deg, North America	Monthly, 1895-2010	Cook and Vose, 2011*
<i>Temperature</i>				
CRU TS3.1	Gauge	0.5 deg, global land	Monthly, 1901-2008	Mitchell and Jones (2005)
GHCN	Gauge	2.5 degree, global land	Daily, varies	Vose et al. (1992)
HadGHCND	Gauge	2.5x3.75 degree,	Daily, 1950-2000	Caesar et al. (2006)

global land

Sea Surface Temperature and Sea Ice

HadISSTv1.1	In-situ/satellite	1.0 deg, global oceans	Monthly, 1870-present	Rayner et al. (2003)
NSIDC Sea Ice Index	Satellite	Arctic Basin	Monthly, 1979-present	Fetterer et al., 2002
IceSAT	Satellite	25km, Arctic Basin	Monthly, 2003-2008	Kwok and Cunningham (2008)

Land Surface Hydrology

NLDAS-UW	Multiple LSMs	0.5 deg US	Daily, 1916-2009	Wang et al. (2009)
VIC	VIC LSM	1.0 deg, global land	3-hourly, 1948-2008	Sheffield et al. (2007)
GLDAS	Noah LSM	1.0 deg, global land	3-hourly, 1979-2008	Rodell et al. (2004)

Reanalyses

NCEP-NCAR	Model reanalysis	~1.9 deg, global	6-hourly, 1948-present	Kalnay et al. (1996)
NCEP-DOE	Model reanalysis	~1.9 deg, global	6-hourly, 1979-present	Kanamitsu et al. (2002)
CFSR	Model reanalysis	~0.3 deg, global	6-hourly, 1979-2010	Saha et al. (2010)
20CR	Model reanalysis	2.0 deg, global	6-hourly, 1871-present	Compo et al. (2011)

1299 *P-NOAA dataset provided by Drs. Russ Vose and Ed Cook.

1300

1301 **Table 3.** DJF and JJA bias (% of observed mean) in CMIP5 continental and regional
1302 precipitation relative to the GPCP observations. The mean and standard deviation of the
1303 biases across the multimodel ensemble are also given. North America (NA): 10N to 72N
1304 and 190E to 305E; contiguous US (conUS): 25N to 50N and 235W to 285W; the regions
1305 are Alaska (ALA), North East Canada (NEC), Eastern North America (ENA), Central
1306 North America (CNA), West North America (WNA), and Central America (CAM), as
1307 modified from Giorgi and Francisco (2000) and shown in supplementary Fig. S3.

Model	NA	conUS	ALA	NEC	ENA	CNA	WNA	CAM
DJF								
BCC-CSM1-1	14.84	21.39	40.32	-1.40	-10.24	-16.38	47.51	97.71
CanESM2	-3.80	-5.53	1.24	9.81	-2.23	-13.22	5.26	-21.09
CCSM4	17.13	10.24	49.31	16.52	-7.87	-20.21	48.87	64.19
CNRM-CM5	0.44	-1.58	4.50	-0.05	-5.71	-25.48	27.34	1.55
CSIRO-Mk3.6.0	2.82	5.35	16.78	7.36	-11.15	-8.46	20.41	13.45
GFDL-CM3	18.92	27.41	19.57	7.98	4.46	-2.60	52.70	59.52
GFDL-ESM2M	10.09	14.70	4.50	11.77	0.03	-16.67	40.63	45.80
GISS-E2-R	23.68	25.18	32.55	15.27	1.45	-6.81	47.96	110.22
HadCM3	2.59	14.43	-9.81	0.84	10.82	-4.61	27.56	-25.27
HadGEM2-ES	1.38	4.26	-4.11	-9.20	6.66	-4.44	13.90	-4.44
INMCM4	40.75	25.34	75.24	56.06	14.36	-7.40	71.31	97.49
IPSL-CM5A-LR	16.64	12.38	53.18	18.24	4.84	-14.11	42.28	25.43
MIRCO5	7.51	7.94	14.83	11.28	-1.46	-13.67	29.22	28.43
MIROC-ESM	16.82	8.73	66.33	31.07	-2.38	-37.19	70.34	-12.75
MPI-ESM-LR	11.87	14.27	11.71	21.56	3.17	-6.02	34.18	18.59
MRI-CGCM3	24.97	34.43	27.12	-6.09	8.94	-6.16	70.00	85.93
NorESM1-M	2.17	-6.86	59.74	4.10	-28.15	-39.07	25.49	65.74
mean	12.28	12.48	27.24	11.48	-0.85	-14.26	39.70	38.27
stdev	11.31	11.66	26.07	15.40	10.06	10.92	19.55	44.03
JJA								
BCC-CSM1-1	-14.95	-20.74	22.05	-8.61	-20.63	-30.17	-10.90	-16.91
CanESM2	-25.72	-39.23	40.26	-10.17	-22.26	-49.66	-25.63	-49.05
CCSM4	-7.04	3.47	23.70	0.92	7.67	-5.36	4.06	-39.06
CNRM-CM5	-0.13	-4.73	42.30	8.02	10.39	-29.25	31.69	-25.13
CSIRO-Mk3.6.0	-3.52	-29.14	54.57	4.81	-17.30	-38.92	-3.67	6.61
GFDL-CM3	7.65	15.95	34.39	20.85	-4.38	7.45	54.74	-31.10
GFDL-ESM2M	7.35	12.03	61.15	21.49	-5.20	2.66	49.86	-31.24
GISS-E2-R	20.81	32.35	14.93	11.93	22.82	24.24	66.36	-9.00
HadCM3	-3.44	1.16	26.83	8.19	0.57	0.73	19.82	-40.07
HadGEM2-ES	-1.41	-15.89	72.29	14.40	-0.04	-32.63	4.59	-16.90
INMCM4	4.00	-1.74	43.12	31.19	17.89	-19.67	44.51	-44.06

IPSL-CM5A-LR	-13.67	-1.62	16.88	3.27	9.85	-13.26	13.83	-63.80
MIRCO5	3.32	0.72	38.38	-3.87	10.09	-17.38	36.93	-13.17
MIROC-ESM	-4.37	4.37	47.01	6.18	1.62	-27.95	69.36	-56.12
MPI-ESM-LR	9.05	11.02	37.85	19.95	14.30	8.21	10.36	-14.15
MRI-CGCM3	10.98	16.85	23.16	9.44	6.35	0.48	61.05	-10.33
NorESM1-M	-10.01	11.48	12.31	-9.71	2.99	12.31	9.47	-51.51
mean	-1.24	-0.22	35.95	7.55	2.04	-12.25	25.67	-29.70
stdev	11.22	17.88	16.77	11.80	12.89	20.63	28.70	19.53

1308

1309

1310 **Table 4.** DJF and JJA bias in CMIP5 continental and regional near surface air
1311 temperature ($^{\circ}\text{C}$) relative to the CRU TS3.0 observations. The mean and standard
1312 deviation of the biases across the multimodel ensemble are also given. North America
1313 (NA): 10N to 72N and 190E to 305E; contiguous US (conUS): 25N to 50N and 235W to
1314 285W; the regions are Alaska (ALA), North East Canada (NEC), Eastern North America
1315 (ENA), Central North America (CNA), West North America (WNA), and Central
1316 America (CAM), as modified from Giorgi and Francisco (2000) and shown in
1317 supplementary Fig. S3.

1318 .

Model	NA	conUS	ALA	NEC	ENA	CNA	WNA	CAM
	DJF							
BCC-CSM1-1	-0.97	-1.98	1.95	-1.61	-1.29	-1.30	-1.88	-1.20
CanESM2	2.09	2.04	1.46	3.13	4.79	4.06	0.74	-0.84
CCSM4	0.01	-0.53	1.81	-0.30	1.03	-0.22	-0.37	-2.04
CNRM-CM5	-1.87	-1.95	-3.88	-0.99	-0.54	-0.55	-2.39	-2.54
CSIRO-Mk3.6.0	-1.61	-2.05	-2.31	-1.51	-1.04	-2.27	-1.31	-1.92
GFDL-CM3	1.37	-0.03	4.07	3.04	1.15	1.25	0.27	-2.31
GFDL-ESM2M	1.67	-0.84	5.05	5.75	1.35	0.75	-0.40	-3.31
GISS-E2-R	-0.33	-1.21	0.72	1.00	-0.56	-0.51	-1.13	-1.23
HadCM3	-2.88	-3.46	-3.72	-2.16	-1.51	-3.53	-3.48	-2.00
HadGEM2-ES	-3.81	-2.98	-7.00	-5.12	-1.61	-3.15	-3.47	-1.54
INMCM4	1.71	0.26	3.73	3.38	1.87	1.48	1.51	-3.71
IPSL-CM5A-LR	0.15	-1.14	4.16	-0.12	0.00	-0.64	-0.64	-2.27
MIRCO5	1.02	0.65	0.74	2.14	1.78	1.16	0.27	0.39
MIROC-ESM	3.29	1.74	4.22	6.74	3.90	3.00	1.35	0.73
MPI-ESM-LR	-0.08	0.30	-1.85	1.50	0.64	1.28	-0.93	-0.10
MRI-CGCM3	-0.37	-0.37	1.78	-2.63	-1.33	0.38	0.56	-2.26
NorESM1-M	-0.89	-1.60	2.12	-2.14	-0.34	-1.44	-1.48	-1.81
mean	-0.09	-0.77	0.77	0.59	0.49	-0.01	-0.75	-1.65
stdev	1.85	1.52	3.41	3.15	1.86	2.02	1.49	1.19
	JJA							
BCC-CSM1-1	-0.76	0.54	-3.24	-1.40	0.16	1.80	-0.78	-0.20
CanESM2	3.14	4.11	1.77	3.17	3.14	5.76	3.60	0.62
CCSM4	1.10	1.37	-0.13	1.55	1.21	2.04	1.78	-0.82
CNRM-CM5	0.41	-0.06	1.57	1.23	-0.72	1.16	0.08	-1.37
CSIRO-Mk3.6.0	1.30	2.97	-1.25	-0.11	1.38	5.18	1.45	1.46
GFDL-CM3	-1.96	-2.24	-1.47	-2.60	-1.37	-2.27	-2.04	-1.65
GFDL-ESM2M	-0.36	-0.59	-0.29	0.25	-0.34	-0.19	-0.75	-0.67
GISS-E2-R	-0.64	-2.09	1.71	0.79	-0.46	-1.74	-2.10	-1.25

HadCM3	-1.12	-0.74	-1.41	-1.73	-1.70	0.08	-1.30	-0.23
HadGEM2-ES	1.17	1.79	0.98	0.29	0.95	2.56	2.22	-1.36
INMCM4	-0.82	-2.07	0.78	1.23	-1.68	-1.02	-2.05	-2.52
IPSL-CM5A-LR	0.38	-1.22	2.60	2.45	0.35	0.01	-1.43	-1.40
MIRCO5	2.63	2.30	3.88	2.56	2.15	3.11	2.79	0.26
MIROC-ESM	2.76	1.89	2.91	4.09	3.03	3.31	1.67	2.20
MPI-ESM-LR	-0.84	-0.66	-1.31	-0.40	-0.81	-0.09	-1.55	0.02
MRI-CGCM3	-0.37	-1.77	2.10	1.23	-0.77	-0.97	-1.80	-2.04
NorESM1-M	-1.02	-0.51	-3.05	-0.90	-0.87	-0.64	0.03	-1.35
mean	0.29	0.18	0.36	0.69	0.21	1.06	-0.01	-0.61
stdev	1.51	1.93	2.09	1.80	1.54	2.34	1.89	1.24

1319

1320 **Table 5.** CMIP5 error statistics for annual average SSTs. The mean and standard
 1321 deviation of the statistics across the multimodel ensemble are given, as well as the
 1322 statistics of the MME mean SSTs. Statistics are calculated over neighboring oceans to
 1323 North America (170-35°W, 10S-40°N; domain displayed in Fig. 3) for average annual
 1324 values for 1979-2004.

Model	Spatial Correlation	RMSE (°C)	Bias (°C)
BCC-CSM1.1	0.95	0.98	-0.15
CanESM2	0.97	0.87	-0.16
CCSM4	0.96	0.88	0.37
CNRM-CM5.1	0.96	0.94	-0.63
CSIRO-Mk3.6.0	0.94	1.39	-1.80
GFDL-CM3	0.94	1.05	-0.68
GFDL-ESM2M	0.94	1.08	-0.46
GISS-E2-R	0.95	0.98	-0.15
HadCM3	0.94	1.53	-0.64
HadGEM2-ES	0.96	0.95	-0.96
INMCM4	0.94	1.06	-0.01
IPSL-CM5A-LR	0.93	1.32	-0.78
MIROC5	0.95	0.95	-0.65
MIROC-ESM	0.91	1.32	-0.60
MPI-ESM-LR	0.96	0.95	-0.49
MRI-CGCM3	0.95	1.25	-0.49
NorESM1-M	0.92	1.21	-0.78
mean	0.95	1.10	-0.54
stdev	0.02	0.20	0.47
MME mean	0.97	0.77	-0.54

1325

1326

1327 **Table 6.** Spatial correlations between simulated and observed estimates of column
 1328 integrated moisture divergence for summer (JJA) and winter (DJF) seasons for the North
 1329 American region. The CMIP5 model data were regridded to the 20CR grid (~200km) for
 1330 this calculation. The mean and standard deviation of the correlations across the
 1331 multimodel ensemble are also given.

Model	Spatial Correlation
	Summer (JJA)
CanESM2	0.28
CCSM4	0.18
CNRM-CM5	0.39
GFDL-ESM2M	0.08
MIROC5	0.42
mean	0.27
stdev	0.14
	Winter (DJF)
CanESM2	0.76
CCSM4	0.72
CNRM-CM5	0.75
GFDL-ESM2M	0.66
MIROC5	0.60
mean	0.70
stdev	0.07

1332

1333

1334 **Table 7.** Bias (model minus observational estimates) in annual runoff ratio (total runoff /
1335 precipitation) averaged over 1979-2004 for the North American continent, the contiguous
1336 US and the six regions defined in Table 3. The mean and standard deviation of the
1337 statistics across the multimodel ensemble are also given. Observed runoff is estimated
1338 from VIC and GLDAS2 Noah; precipitation is from GPCP. All data were interpolated to
1339 2.5-degree resolution.

Model	NA	conUS	ALA	NEC	ENA	CNA	WNA	CAM
BCC-CSM1-1	0.09	0.06	0.36	-0.10	-0.04	-0.03	0.11	0.21
CanESM2	0.40	0.26	0.69	0.56	0.35	0.20	0.43	0.12
CCSM4	0.36	0.19	0.75	0.51	0.31	0.10	0.37	0.15
CNRM-CM5	0.38	0.26	0.54	0.52	0.39	0.17	0.42	0.22
CSIRO-Mk3.6.0	0.30	0.19	0.53	0.37	0.28	0.16	0.26	0.26
GFDL-CM3	0.33	0.24	0.47	0.47	0.37	0.18	0.35	0.16
GFDL-ESM2M	0.32	0.25	0.36	0.38	0.36	0.16	0.36	0.30
GISS-E2-R	0.22	0.19	0.15	0.28	0.28	0.14	0.27	0.20
HadCM3	0.37	0.29	0.57	0.54	0.43	0.31	0.31	0.03
INMCM4	0.37	0.21	0.64	0.55	0.33	0.12	0.40	0.11
MIRCO5	0.33	0.21	0.58	0.49	0.40	0.17	0.29	0.23
MIROC-ESM	0.37	0.25	0.56	0.49	0.44	0.12	0.36	0.10
MPI-ESM-LR	0.31	0.21	0.55	0.46	0.33	0.18	0.30	0.11
MRI-CGCM3	0.42	0.32	0.64	0.55	0.46	0.28	0.40	0.28
NorESM1-M	0.35	0.18	0.77	0.51	0.30	0.11	0.34	0.15
mean	0.33	0.22	0.54	0.44	0.33	0.16	0.33	0.18
stdev	0.08	0.06	0.16	0.17	0.12	0.08	0.08	0.08

1340

1341

1342

1343 **Table 8.** Bias and spatial correlation between the HadGHCND observations and the
1344 CMIP5 ensemble for number of summer days, number of frost days and growing season
1345 length averaged over 1979-2005. The mean and standard deviation of the statistics across
1346 the multimodel ensemble are given, as well as the statistics of the MME mean. The
1347 frequencies were calculated on the model grid and then interpolated to 2.0-degree
1348 resolution for comparison with the observational estimates.

Model	Number of summer days		Number of frost days		Growing season length (days)	
	Bias (days)	Spatial correlation	Bias (days)	Spatial correlation	Bias (days)	Spatial correlation
BCC-CSM1-1	-14.0	0.95	-4.7	0.96	-7.8	0.91
CanESM2	17.1	0.96	-16.4	0.93	-12.1	0.90
CCSM4	0.0	0.88	-3.5	0.95	-9.0	0.92
CNRM-CM5	-7.4	0.92	12.6	0.92	-14.2	0.89
CSIRO-Mk3.6.0	-8.2	0.98	3.7	0.95	-4.3	0.90
GFDL-CM3	-39.5	0.93	0.6	0.97	-24.6	0.93
GFDL-ESM2M	33.0	0.92	-7.8	0.96	-5.6	0.92
GISS-E2-R	33.5	0.94	-12.8	0.96	7.4	0.96
HadCM3	-21.9	0.98	21.6	0.95	-38.2	0.88
HadGEM2-ES	-6.9	0.92	2.2	0.97	-14.9	0.95
INMCM4	-28.8	0.94	17.0	0.85	-76.1	0.53
IPSL-CM5A-LR	-39.3	0.85	4.1	0.97	-6.5	0.95
MIROC5	1.3	0.91	-15.1	0.98	33.4	0.96
MIROC-ESM	-5.7	0.90	-34.7	0.97	38.5	0.96
MRI-CGCM3	-34.8	0.87	-6.8	0.95	4.0	0.95
MPI-ESM-LR	-30.4	0.92	-12.5	0.94	5.5	0.93
NorESM1-M	-21.6	0.89	-3.7	0.96	-19.7	0.93
mean	-10.21	0.92	-3.31	0.95	-8.5	0.90
stdev	22.56	0.04	13.56	0.03	25.65	0.10
MME mean	-18.1	0.96	-2.8	0.97	-8.5	0.95

1349

1350

1351 Table 9. Frequency of occurrence of persistent extreme precipitation and soil moisture
 1352 events over the US for the CPC observations/NLDAS analysis and 15 CMIP5 models.
 1353 The mean and standard deviation of the statistics across the multimodel ensemble are also
 1354 given.

Model	SPI6	SM
Obs/Analysis	0.37	0.68
BCC-CSM1.1	0.00	0.05
CanESM2	0.29	0.32
CCSM4	0.35	0.14
CNRM-CM5.1	0.16	0.30
CSIRO-Mk3.6.0	0.02	0.70
GFDL-CM3	0.02	0.01
GISS-E2-R	0.04	0.47
HadCM3	0.00	0.32
HadGEM2-ES	0.00	0.37
IPSL-CM5A-LR	0.21	0.20
MIROC5	0.34	0.25
MIROC-ESM	0.26	0.27
MPI-ESM-LR	0.23	0.62
MRI-CGCM3	0.16	0.01
NorESM1-M	0.01	0.08
mean	0.14	0.27
stdev	0.13	0.21

1355

1356

1357

1358 **Table 10:** Error statistics for the CMIP5 model precipitation over the northeastern US.
 1359 The mean absolute error (mm per season), RMSE (mm day⁻¹), and mean bias
 1360 (model/observed) for 14 CMIP5 models verified using the daily CPC-Unified
 1361 precipitation within the black box in Fig. 10b. The mean and standard deviation of the
 1362 statistics across the multimodel ensemble are given, as well as the statistics of the MME
 1363 mean precipitation.
 1364

Model	Mean Absolute Error (mm/season)	Root Mean Square Error (mm/day)	Mean Bias (model/obs)
CanESM2	94.02	1.08	1.04
CCSM4	101.07	1.06	1.10
GFDL-ESM2M	101.32	1.15	1.13
GFDL-CM3	103.53	1.14	1.14
BCC-CSM1-1	104.62	1.16	1.08
CNRM-CM5	104.99	1.12	1.16
HadGEM-ES	105.41	1.18	1.16
MIROC-ESM	111.21	1.25	0.92
NorESM1-M	112.70	1.23	1.08
CSIRO-Mk3_6_0	114.49	1.46	1.03
IPSL-CM5A-LR	115.96	1.27	1.03
MIROC5	118.35	1.28	1.20
INMCM4	123.83	1.34	1.20
MRI-CGCM3	126.15	1.48	1.12
mean	109.83	1.23	1.10
stdev	9.25	0.13	0.08
MME mean	84.55	0.89	1.10

1365
 1366
 1367

1368 **Table 11.** Annual bias in the number of heavy precipitation days (precipitation > 10mm
1369 day⁻¹) and hot days (Tmax > 32°C (90°F)) for the southern US regions (defined in Figure
1370 13) for 1979-2005. The mean and standard deviation of the biases across the multimodel
1371 ensemble are given, as well as the statistics of the MME mean heavy precipitation and
1372 hot days. Observed actual values from the GHCN and CPC datasets are shown in
1373 parentheses. All data were regridded to 2.5-degree resolution.

	Number of heavy precipitation days			Number of hot days		
	SW	SC	SE	SW	SC	SE
Obs	(8.5)	(23.4)	(37.5)	(55.8)	(59.5)	(40.1)
CanESM2	-6.8	-16.9	-24.3	13.8	8.0	34.7
CCSM4	2.4	-9.6	-11.1	-8.3	-5.6	-20.6
GFDL-CM3	0.1	-16.1	-23.1	-39.9	-49.4	-37.4
GFDL-ESM2G	7.8	3.0	1.3	-35.4	-31.0	-23.6
GFDL-ESM2M	8.2	3.1	2.4	-33.1	-26.4	-19.4
GISS-E2-R	18.9	7.3	11.4	-34.9	-41.3	-35.5
HadCM3	-6.3	-20.9	-29.2	5.0	8.3	-15.0
HadGEM2-CC	2.0	-0.5	1.3	7.8	-5.5	-18.2
IHadGEM2-ES	-3.5	-5.3	-1.0	9.6	-0.3	-15.8
IPSL-CM5A-LR	-3.9	-20.5	-29.0	-49.0	-35.5	-38.3
MIROC4h	1.3	-5.0	-2.5	-17.2	14.3	10.7
MIROC5	-3.2	-13.1	-13.5	-14.1	13.8	1.6
MPI-ESM-LR	2.6	-8.9	-4.3	-26.9	-19.0	-27.6
MRI-CGCM3	9.6	-4.9	-1.3	-39.6	-46.1	-38.7
mean	2.09	-7.74	-8.78	-18.73	-15.41	-17.36
stdev	7.12	9.01	13.04	21.25	22.90	20.85
MME mean	2.1	-7.7	-8.8	-18.7	-12.6	-17.4

1374

1375

1376

1377

1378 **Table 12.** Annual mean RMSE for precipitation (mm day^{-1}) for each of the 17 core
 1379 CMIP5 models compared with CMAP observed estimates for the North American
 1380 Monsoon region $20\text{-}35^{\circ}\text{N}$, $102.5\text{-}115^{\circ}\text{W}$. The mean and standard deviation of the RMSE
 1381 across the multimodel ensemble are given also. The CMAP and model data were
 1382 regridded to T42 resolution ($\sim 2.8\text{-degree}$).

1383

Model	RMSE (mm day^{-1})
BCC-CSM1-1	1.92
CCSM4	1.53
CNRM-CM5	1.29
CSIRO-Mk3.6.0	1.09
CanESM2	0.44
GFDL-CM3	1.54
GFDL-ESM2M	1.72
GISS-E2-R	1.46
HadCM3	0.63
HadGEM2-ES	0.75
INMCM4	1.11
IPSL-CM5A-LR	0.99
MIROC-ESM	1.32
MIROC5	1.58
MPI-ESM-LR	1.09
MRI-CGCM3	2.08
NorESM1-M	1.96
mean	1.32
stdev	0.47

1384

1385

1386 **Table 13.** CMIP5 model error statistics for the simulation of the North American
 1387 monsoon in the core region of northwest Mexico (23.875°-28.875°N, 108.875°-
 1388 104.875°W), calculated with respect to the P-NOAA observational dataset. The mean and
 1389 standard deviation of the statistics across the multimodel ensemble are given also. The
 1390 model data were regridded to the P-NOAA resolution (0.5-degree).

Model	RMSE (mm day ⁻¹)	Bias (%)	Lag (months)
BCC-CSM1-1	1.96	83.6	0
CanESM2	1.11	-41.6	0
CCSM4	1.24	70.7	0
CNRM-CM5	0.75	40.4	0
CSIRO-Mk3.6.0	0.77	24.6	0
GFDL-CM3	1.57	74.6	1
GFDL-ESM2G	2.74	137.7	0
GFDL-ESM2M	2.48	117.8	1
GISS-E2-R	1.90	23.5	4
HadCM3	0.83	0.2	0
HadGEM2-CC	0.88	44.8	0
HadGEM2-ES	0.85	37.9	0
INMCM4	1.67	-9.7	4
IPSL-CM5A-LR	1.26	29.6	1
IPSL-CM5A-MR	0.92	1.4	1
MIROC-ESM	1.64	40.0	2
MIROC4h	1.32	65.4	0
MIROC5	1.71	91.4	0
MPI-ESM-LR	1.36	72.8	0
MRI-CGCM3	1.40	79.4	0
NorESM1-M	2.33	110.4	1
mean	1.46	52.10	0.71
stdev	0.58	44.87	1.23

1391

1392

1393

1394 Table 14. Error statistics for the simulation of the Great Plains low-level jet (GPLLJ).
 1395 The statistics are calculated over the regions shown in Fig. 17 and are the RMSE and the
 1396 Index of Agreement (Legates and McCabe 1999; McCabe et al. 2002). The mean and
 1397 standard deviation of the statistics across the multimodel ensemble are also given. The
 1398 data were regridded to 2.5-degree resolution before calculating the statistics.
 1399

Model	RMSE				Index of Agreement			
	Intensity (vertical)	Seasonal cycle	Spatial extent	Average	Vertical structure	Seasonal cycle	Spatial extent	Average
CanESM2	0.87	0.85	0.87	0.87	0.95	0.96	0.94	0.95
CCSM4	0.91	0.88	0.86	0.88	0.95	0.96	0.94	0.95
CNRM-CM5	0.66	0.74	0.85	0.75	0.97	0.97	0.93	0.96
GFDL-ESM2M	0.90	1.01	0.85	0.92	0.93	0.93	0.92	0.93
HadGEM2-ES	1.06	0.98	0.84	0.96	0.92	0.95	0.95	0.94
MIROC5	1.12	0.65	0.88	0.88	0.91	0.98	0.93	0.94
MPI-ESM-LR	0.76	0.85	0.77	0.79	0.95	0.96	0.94	0.95
MRI-CGCM3	1.04	1.28	0.92	1.08	0.90	0.89	0.90	0.90
mean	0.92	0.91	0.86	0.89	0.94	0.95	0.93	0.94
stdev	0.16	0.13	0.04	0.07	0.02	0.02	0.01	0.01

1400

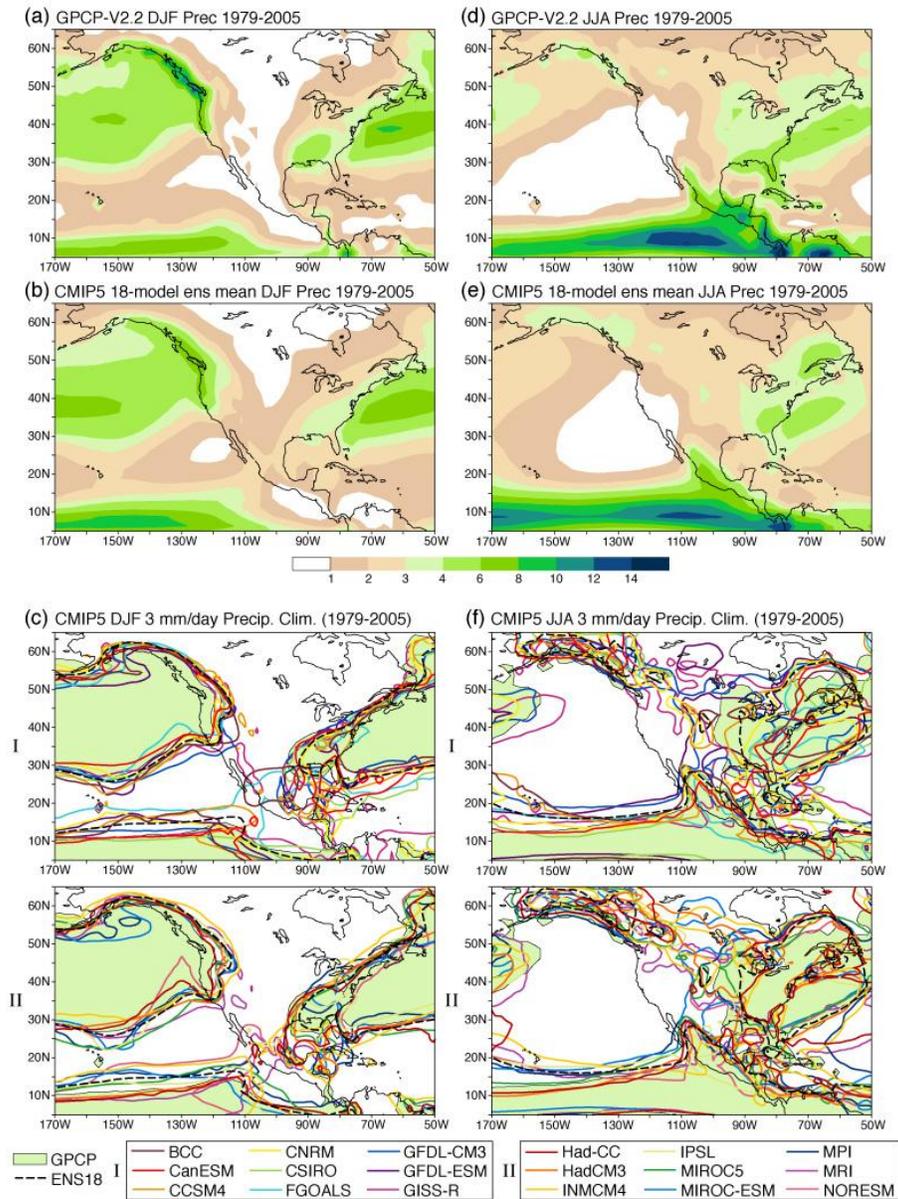
1401

1402 **Table 15.** Biases in CMIP5 model Arctic sea ice extent and thickness. Biases are based
 1403 on the ensemble mean for each model that has more than one ensemble member and
 1404 computed relative to the observed value. The mean and standard deviation of the statistics
 1405 across the multimodel ensemble are also given. September extent bias for in 10^6 km^2 .
 1406 March ice thickness bias in meters. For the calculation of March thickness bias, the
 1407 model data for 1993-2005 were regridded to the ICESat resolution and compared to the
 1408 ICESat data for 2003-2009.

Model	September extent bias 10^6 km^2	March thickness bias (m)
BCC-CSM1-1	-0.439	-0.05
CanCM4	-1.881	-0.76
CanESM2	-2.221	-0.44
CCSM4	0.537	0.08
CESM1-CAM5	0.698	0.11
CNRM-CM5	-0.638	-0.27
CSIRO-Mk3.6.0	3.952	0.31
FGOALS-s2	2.309	-0.06
GFDL-CM3	0.231	-0.37
GISS-E2-H	-2.979	-
GISS-E2-R	-2.653	-0.31
HadCM3	-0.772	-0.20
HadGEM2	-1.845	-1.06
HadGEM2-CC	-0.035	-0.52
HadGEM2-ES	-1.318	-0.90
INMCM4	-1.468	-0.04
IPSL-CM5A-LR	0.708	0.02
IPSL-CM5A-MR	-0.718	-0.32
MIROC-ESM-CHEM	-0.465	-0.81
MIROC-ESM	-0.783	-0.85
MIROC4h	-1.673	-0.71
MIROC5	-0.918	-0.04
MPI-ESM-LR	0.070	-0.57
MRI-CGCM3	-0.931	-0.16
NorESM1-M	1.205	-0.30
mean	-0.48	-0.34
stdev	1.54	0.36

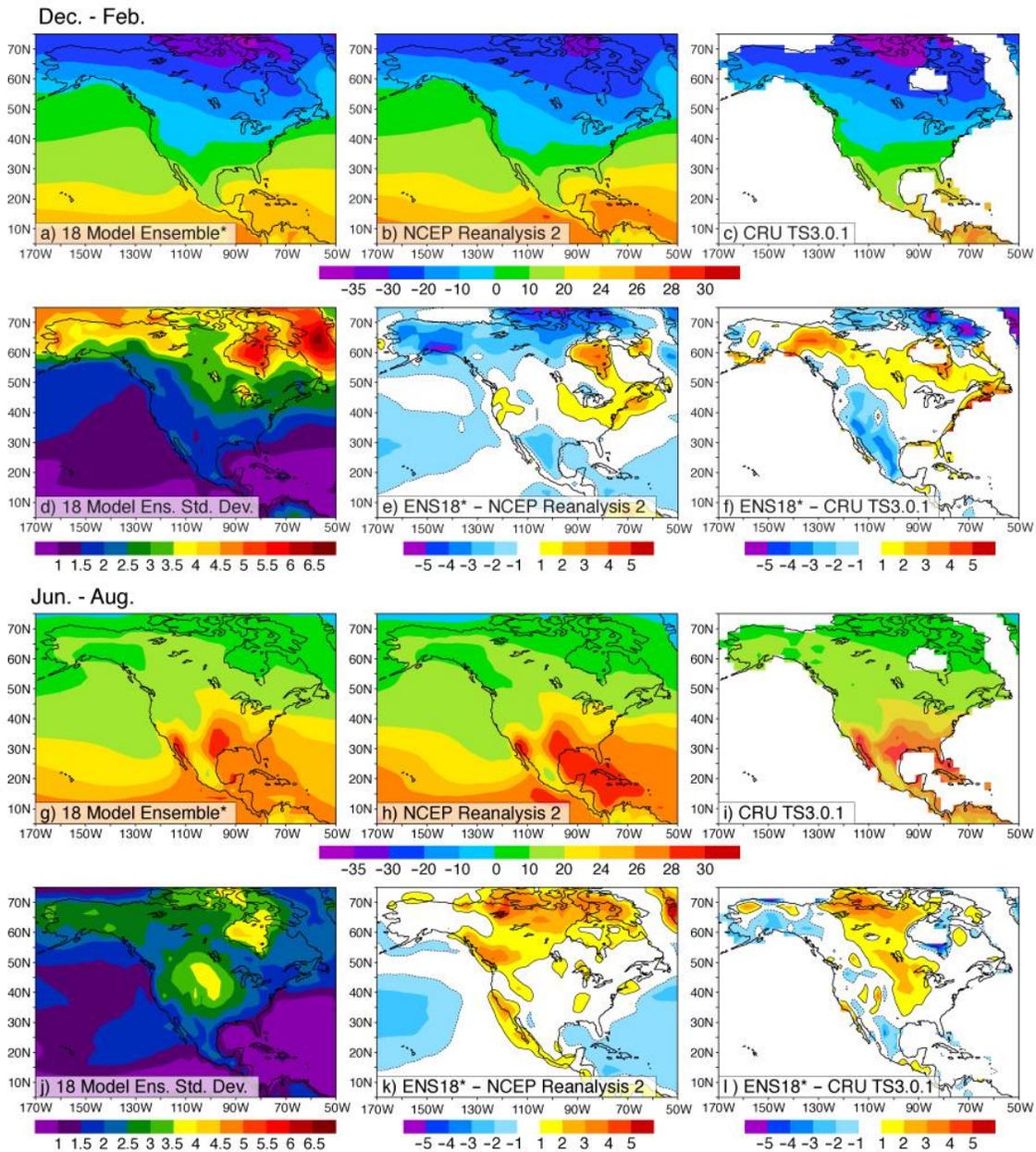
1409

1410



1411

1412 **Figure 1.** Precipitation climatology (mm/day) for (left) December-February and (right)
 1413 June-August (1979-2005). a) GPCP estimate of observed precipitation for DJF. b) MME
 1414 mean over the 18 models for DJF; for models with multiple runs, all runs are averaged
 1415 before inclusion in the multi-model ensemble. c) Comparison of individual models to
 1416 observations using the 3 mm day⁻¹ contour as an index of the major precipitation features:
 1417 half the models are shown in each of sub-panel I and II with the legend giving the color-
 1418 coding for the models in each. Shading shows the regions where GPCP exceeds 3 mm
 1419 day⁻¹; a model with no error would have its contour fall exactly along the edge of the
 1420 shaded region. d)-f) As in a)-c), respectively, except for JJA. The model data were
 1421 regridded to the resolution of the GPCP observations (2.5-degree).

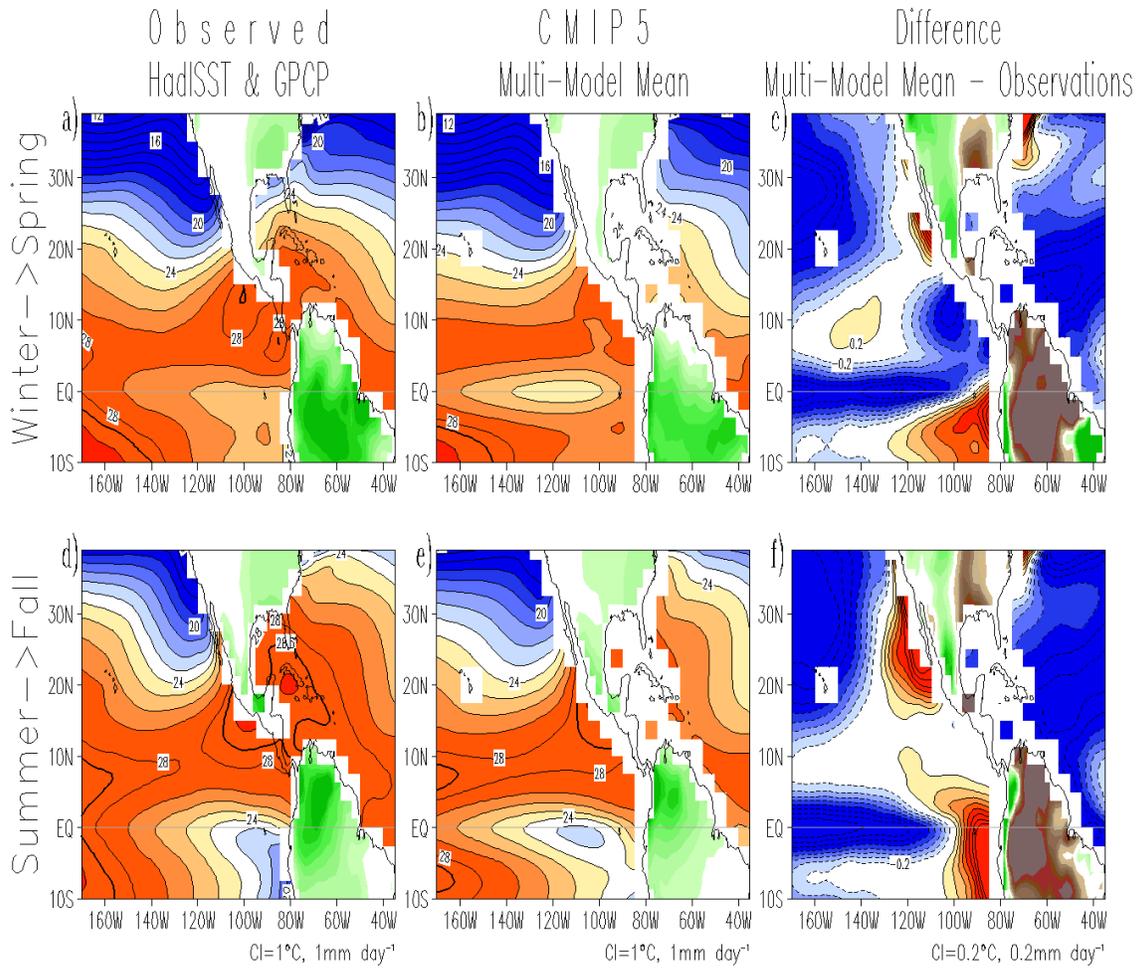


1423

1424 **Figure 2.** Surface air temperature climatology ($^{\circ}\text{C}$) for (top) December-February and
 1425 (bottom) June-August (1979-2005). a) MME mean (over the 17 core models plus
 1426 FGOALS-s2) for DJF. b) NCEP-DOE Reanalysis 2 estimate of observed surface air
 1427 temperature climatology for DJF. c) As in b) but for CRU. d) Standard deviation of
 1428 surface air temperature among the 18 model DJF climatological values at each point. e)
 1429 Difference between the MME mean climatology in a) and the NCEP-DOE Reanalysis 2.
 1430 f) As in e) but for CRU. g)-l) As in a)-f) but for JJA. All observation and model data were
 1431 interpolated to the same 2.5° grid.

1432

1433



1434

1435

1436

1437

1438

1439

1440

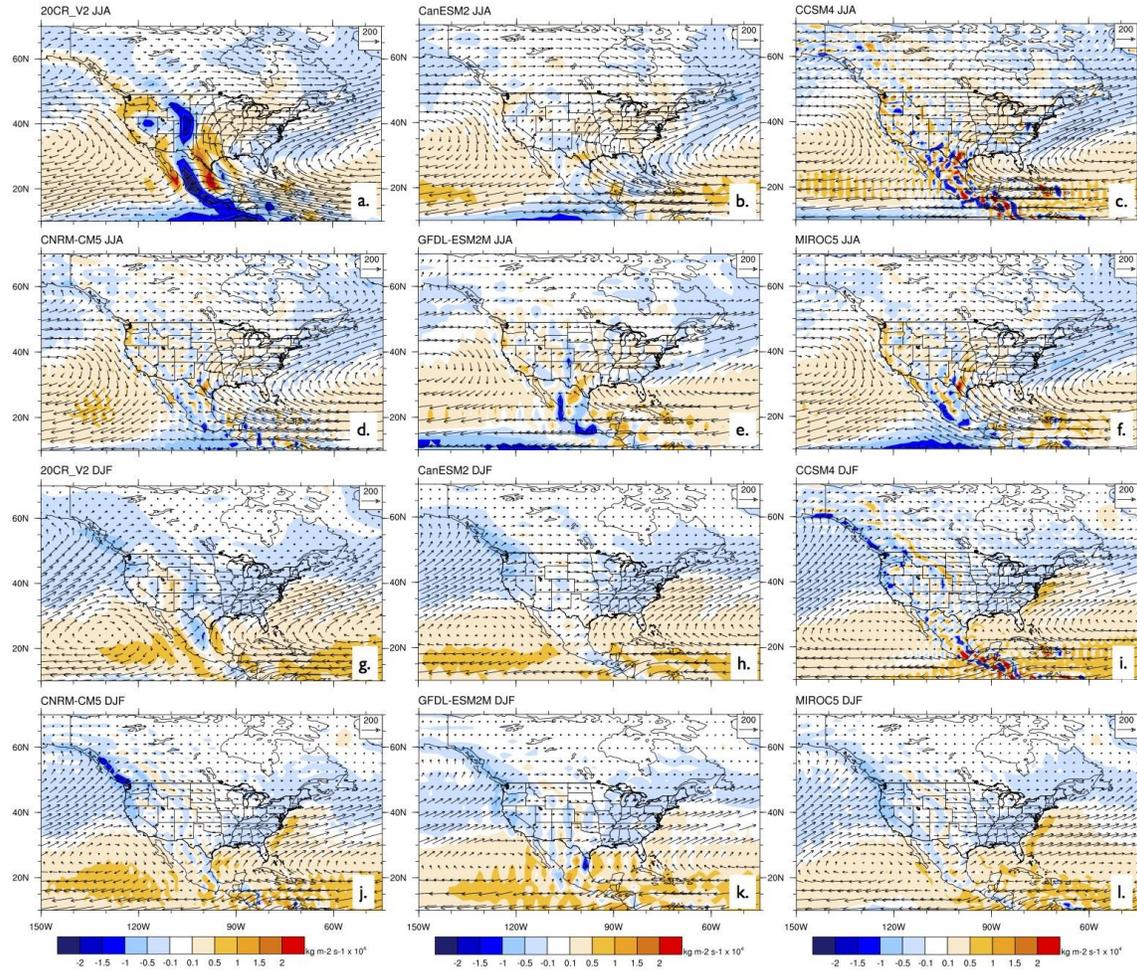
1441

1442

1443

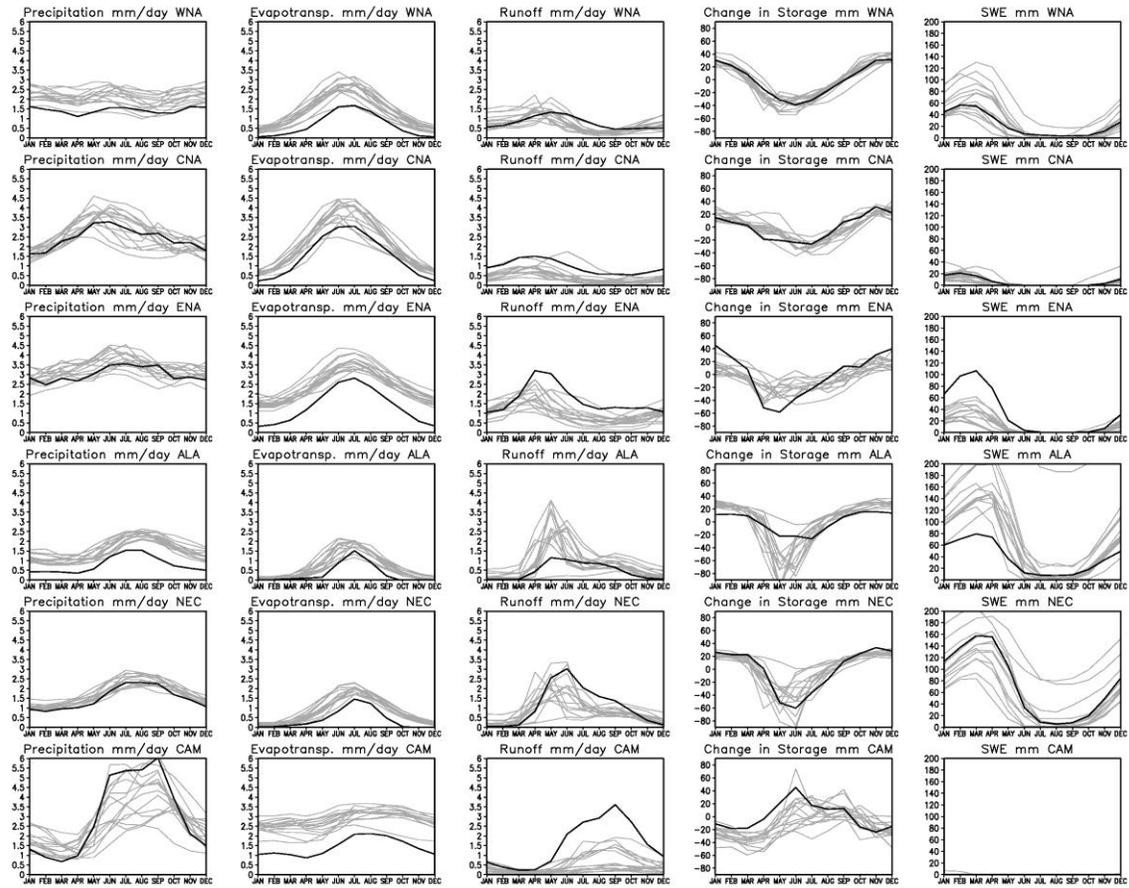
1444

Figure 3. Climatological sea surface temperature and precipitation in observations from HadISSTv1.1 and GPCPv2.2 data sets, and historical simulations from 17 CMIP5 models for 1979-2004. (a) Observations, (b) MME mean, and (c) MME mean minus observations, for winter-to-spring (December to May). (d-f) as in (a-c) but for summer-to-fall (April to November). Temperatures are shaded blue/red for values equal or lower/larger than 23/24°C; the thick black line highlights the 28.5°C isotherm as indicator of the Western Hemisphere Warm Pool. Precipitation is shaded green for values equal or larger than 2 mm day⁻¹. Contour intervals are 1°C and 1 mm day⁻¹ for the mean values and 0.2°C and 2 mm day⁻¹ for the differences. The SST/precipitation fields were regridded to a common 5°×2.5°/2.5°×2.5° grid.



1445

1446 **Figure 4.** Vertically integrated moisture transport (vectors) and its divergence (contours)
 1447 for the 20CR reanalysis (a,g) and five CMIP5 models for mean JJA (b-f) and DJF (h-l)
 1448 for 1981-2000. Vertically integrated moisture transport is computed to 500 hPa using 6-
 1449 hourly data from the 20CR and one realization each from the historical experiments for
 1450 CanESM2, CCSM4, CNRM-CM5, GFDL-ESM2M, and MIROC5 models. The model
 1451 data are shown at their original resolution.

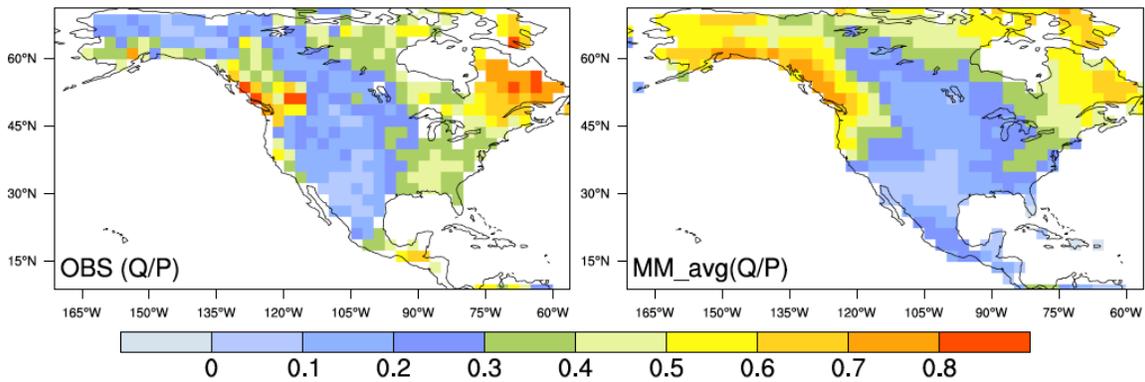
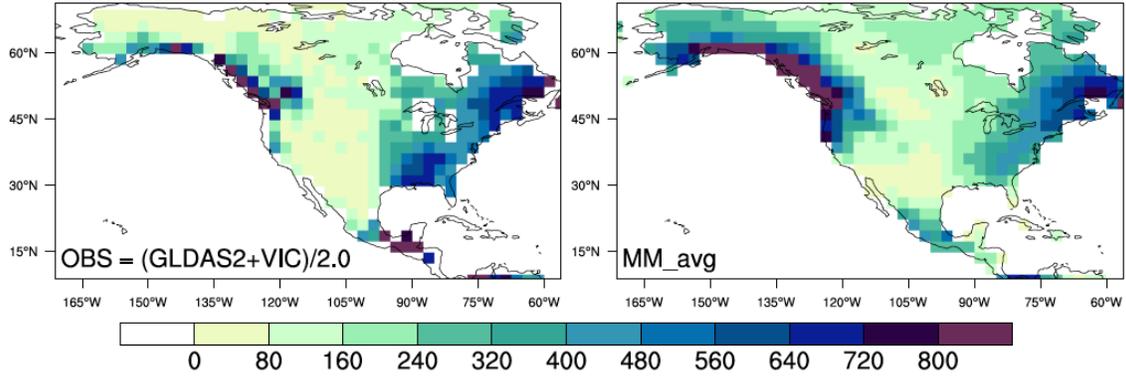


1452

1453 **Figure 5.** Mean seasonal cycle (1979-2005) of North American regional land water
 1454 budget components for 12 CMIP5 models (CanESM2, CSIRO-Mk3.6.0, GFDL-ESM2G,
 1455 GISS-E2-H, GISS_E2-R, IPSL-CM5A-LR, IPSL-CM5A-MR, MIROC-ESM, MIROC-
 1456 ESM-CHEM, MPI-ESM-LR, MRI-CGCM3, NorESM1-M) compared to the average of
 1457 the two off-line LSM simulations (VIC and GLDAS2 Noah). Regions are Western North
 1458 America (WNA), Central North America (CNA), Eastern North America (ENA), Alaska
 1459 and Western Canada (ALA), Northeast Canada (NEC), and Central America (CAM)
 1460 as modified from Giorgi and Francisco (2000) and shown in supplementary Fig. S3. All data
 1461 were interpolated to 2.5-degree resolution.

1462

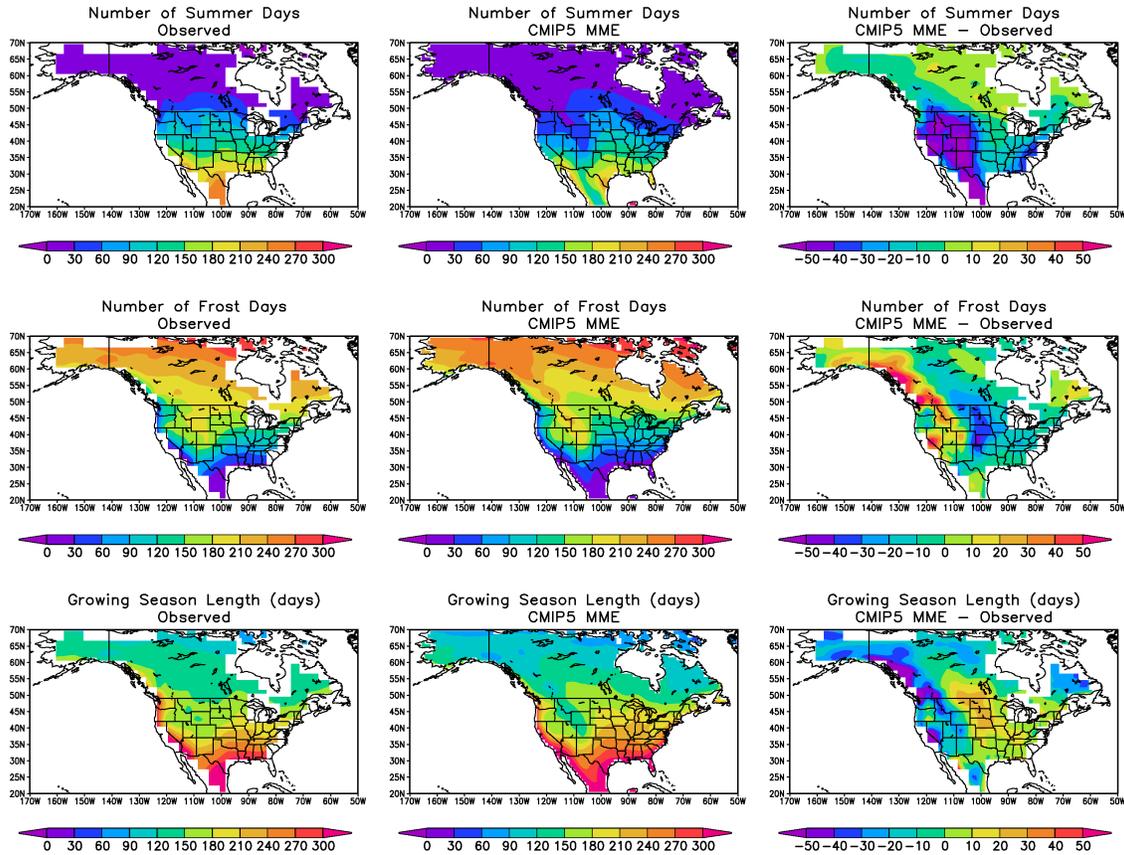
1463



1466 **Figure 6.** Mean annual runoff (mm/year) (top) and runoff (Q/P) ratio (bottom) for 1979-
1467 2004 from observations (Q from VIC and GLDAS2 Noah, and P from GPCP) and the
1468 multi-model average from 15 CMIP5 climate models (BCC-CSM1-1, CanESM2,
1469 CCSM4, CNRM-CM5, CSIRO-Mk3.6.0, GFDL-CM3, GFDL-ESM2M, GISS-E2-R,
1470 HadCM3, INMCM4, MIROC5, MIROC-ESM, MPI-ESM-LR, MRI-CGCM3,
1471 NorESM1-M). All data were interpolated to 2.5-degree resolution.

1472

1473



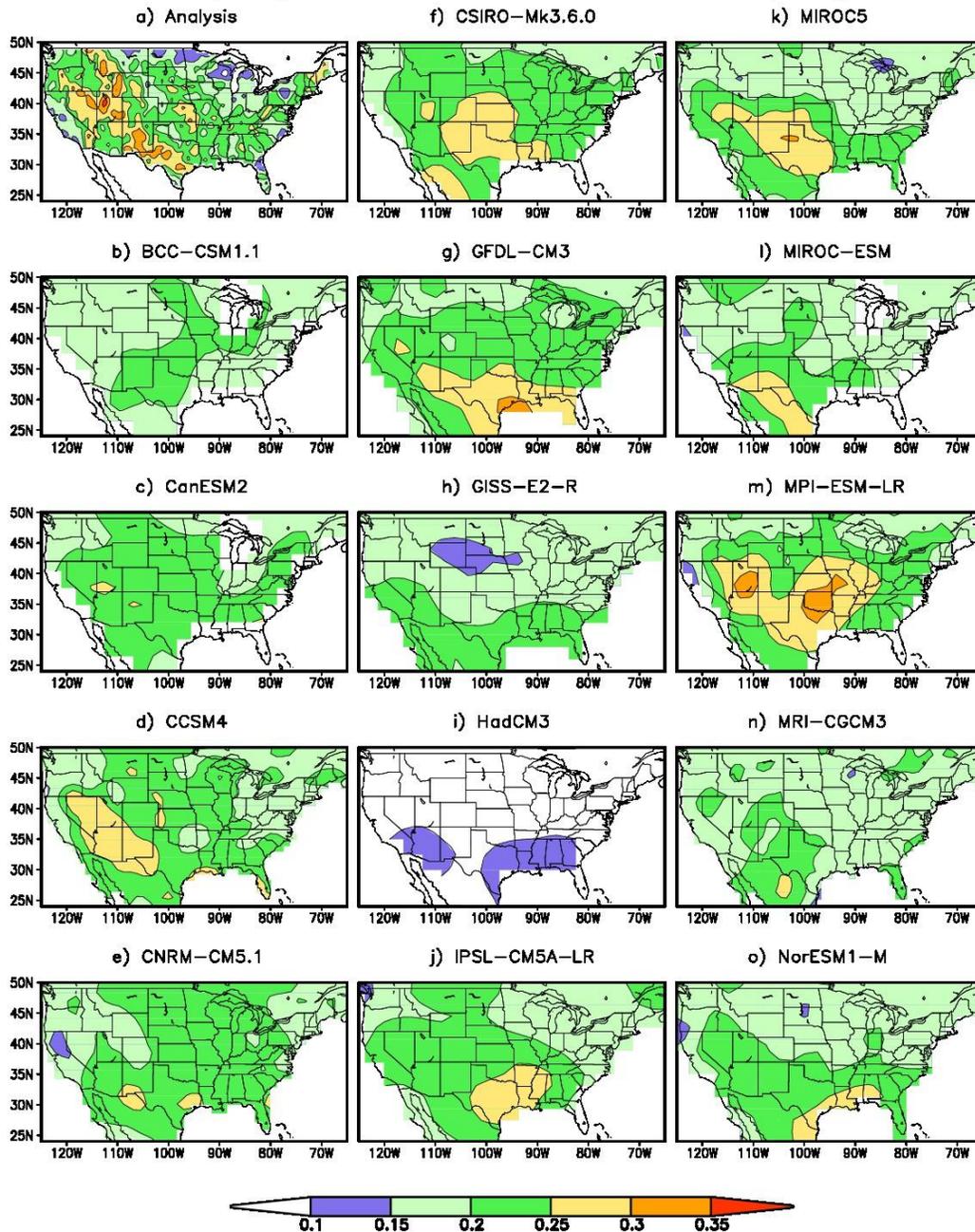
1474

1475 **Figure 7.** Comparison of biophysical indicators between observations and the CMIP5
 1476 ensemble. Biophysical indicators are (top row) number of summer days, (middle row)
 1477 number of frost days, and (bottom row) growing season length averaged over 1979-2005.
 1478 Left column shows the observations (left column) from the HadGHCND dataset; middle
 1479 column is the multi-model ensemble mean of the 17 core models, and the right column is
 1480 their difference (MME – obs.). The frequencies were calculated on the model grid and
 1481 then interpolated to 2.0-degree resolution for comparison with the observational
 1482 estimates.

1483

1484

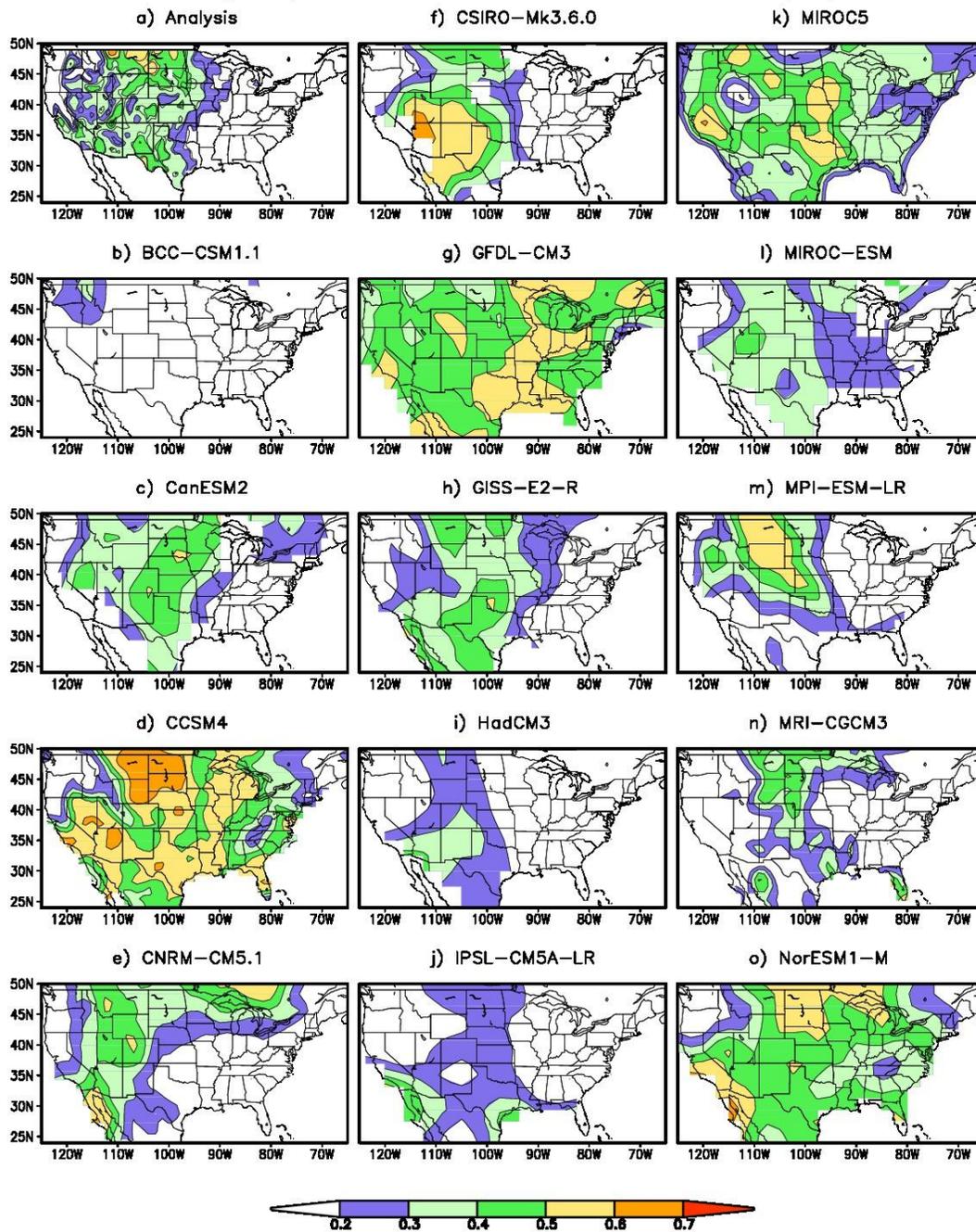
Frequency of Occurrence of Extreme Events (SPI6)



1485

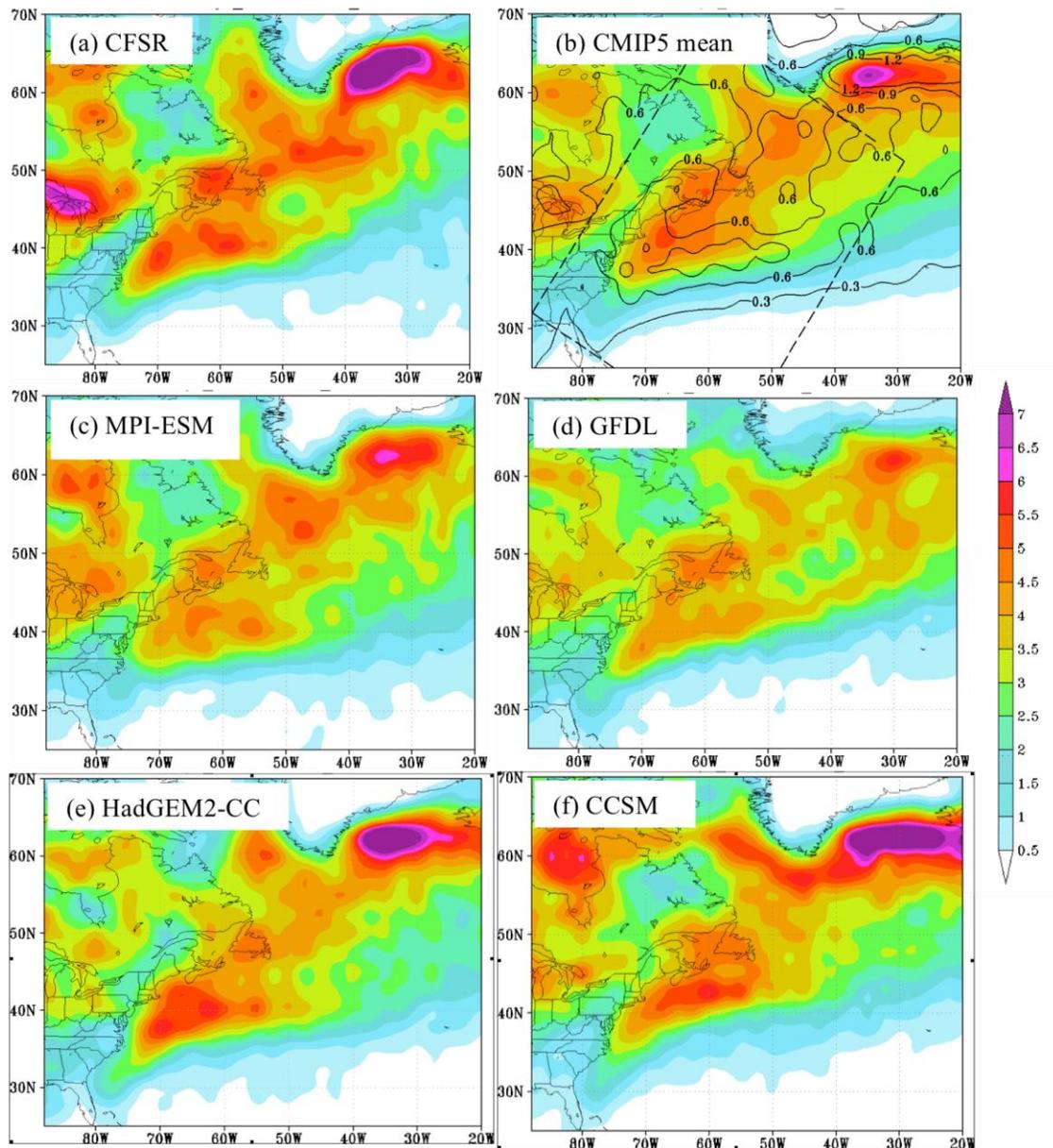
1486 **Figure 8.** The frequency of occurrence of persistent extreme precipitation events defined by SPI6 averaged over positive and negative events for (a) observed precipitation based on the CPC and UW datasets, (b) BCC-CSM1-1, (c) CanESM2, (d) CCSM4, (e) CNRM-
 1487 CM5.1, (f) CSIRO-Mk3.6.0, (g) GFDL-CM3, (h) GISS-E2-R, (i) HadCM3, (j) IPSL-
 1488 CM5A-LR, (k) MIROC5, (l) MIROC-ESM, (m) MPI-ESM-LR, (n) MRI-CGCM3 and
 1489 (o) NorESM1-M. The HadCM3 and HadGEM2-ES results are similarly weak and so the
 1490 former are shown only. Each data set is treated as one member of the ensemble and
 1491 frequencies were calculated at the resolution of the dataset.
 1492
 1493

Frequency of Occurrence of Extreme Events (SM)



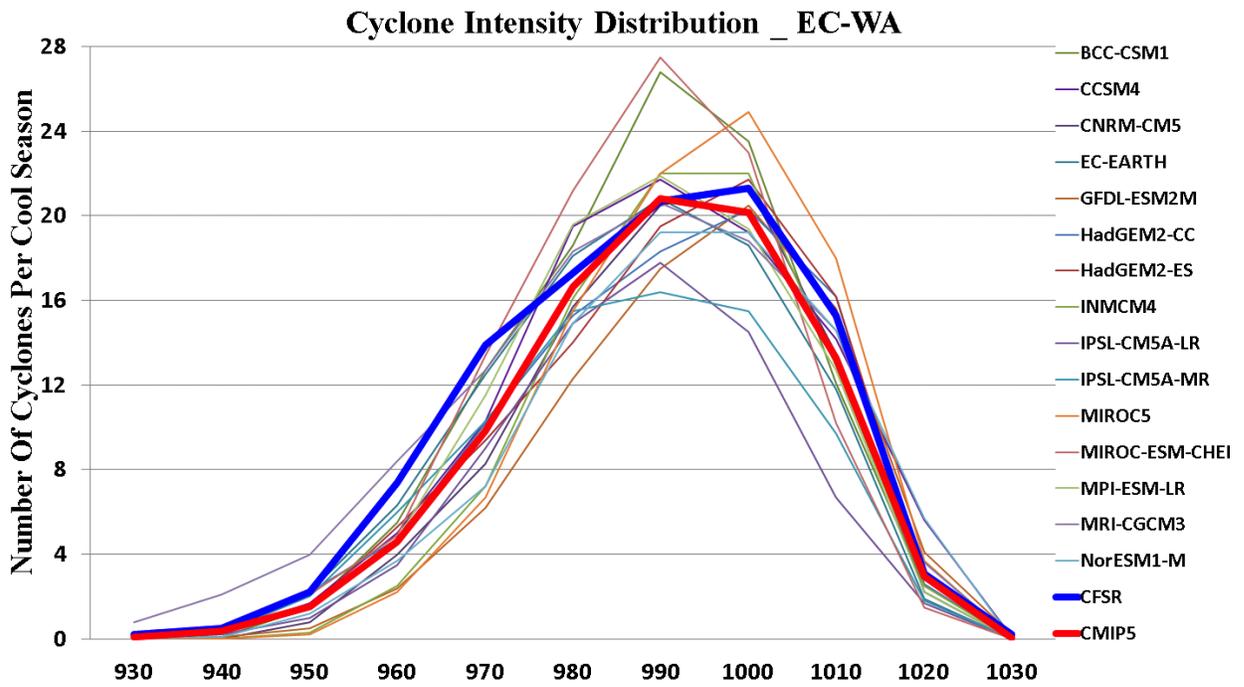
1494

1495 **Figure 9.** Same as Figure 8 but for persistent soil moisture events. Estimates of observed
 1496 soil moisture are taken from the multi-model NLDAS-UW dataset.



1497

1498 **Figure 10.** (a) Cyclone density for the CFSR analysis showing the number of cyclones
 1499 per cool season (November to March) per 50,000 km² for 1979-2004. (b) Same as (a)
 1500 except for the mean (shaded) and spread (contoured every 0.3) of 15 CMIP5 models
 1501 ordered from higher to lower spatial resolution: CanESM2, EC-EARTH, MRI-
 1502 CGCM3, CNRM-CM5, MIRCO5, HadGEM2-ES, HadGEM2-CC, INMCM4, IPSL-
 1503 CM5A-MR, MPI-ESM-LR, NorESM1-M, GFDL-ESM2M, IPSL-CM5A-LR, BCC-
 1504 CSM1, MIROC-ESM-CHEM. Same as (a) except for the (c) MPI-ESM-LR, (d) GFDL-
 1505 ESM2M, (e) HadGEM2-CC, and (f) CCSM4 models. The cyclone densities were
 1506 calculated on the native grid of the dataset.

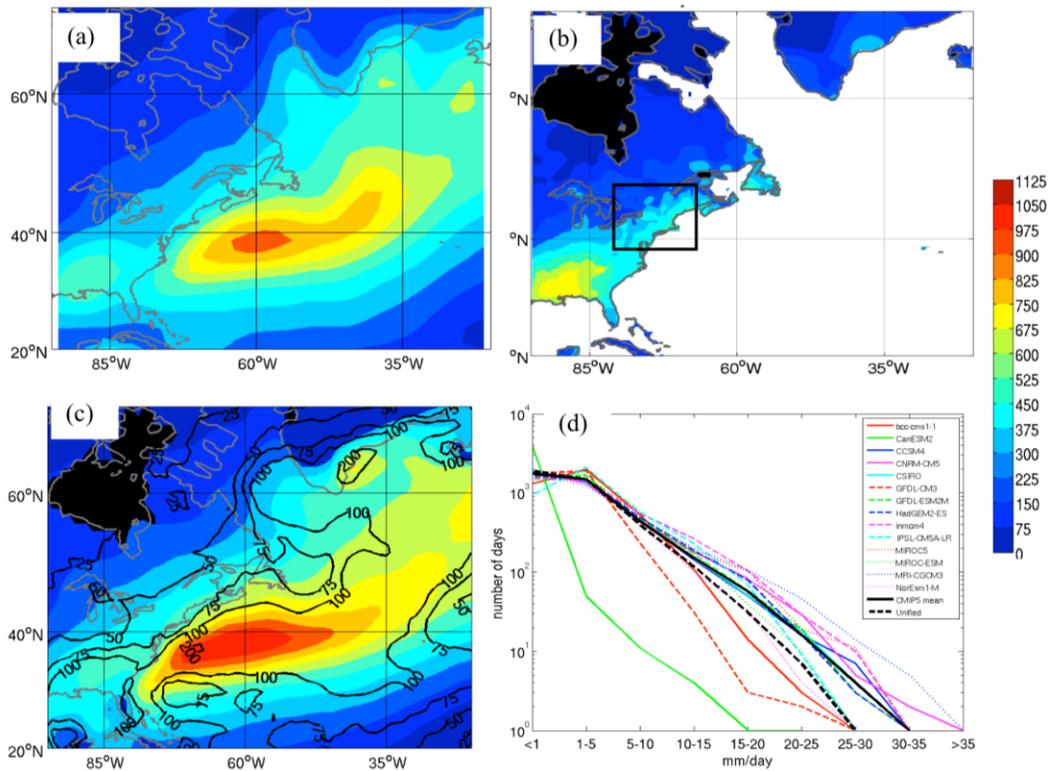


1507

1508 **Figure 11.** Number of cyclone central pressures at their maximum intensity (minimum
 1509 pressure) for the 1979-2004 cool seasons within the dashed box region in Fig. 10 for a 10
 1510 hPa range centered every 10 hPa showing the CFSR (bold blue), (b) CMIP5 MME mean
 1511 (bold red), and individual CMIP5 models.

1512

1513



1514

1515 **Figure 12.** (a) CPC merged precipitation analysis at 2.5 deg resolution showing cool
 1516 seasonal average precipitation (shaded every 75 mm) for the 1979-2004 cool seasons
 1517 (November – March). (b) Same as (a) except for the CPC Unified precipitation at 0.5 deg
 1518 resolution. (c) Same as (a) except for the mean of 14 of the 17 CMIP5 members listed in
 1519 (d) and spread (in mm). (d) Number of days that the daily average precipitation (in
 1520 mm/day) for the land areas in the black box in (b) occurred within each amount bin for
 1521 select CMIP5 members, CMIP5 mean, and the CPC Unified. The CPC and model data
 1522 were regridded to 1.0-degree resolution.

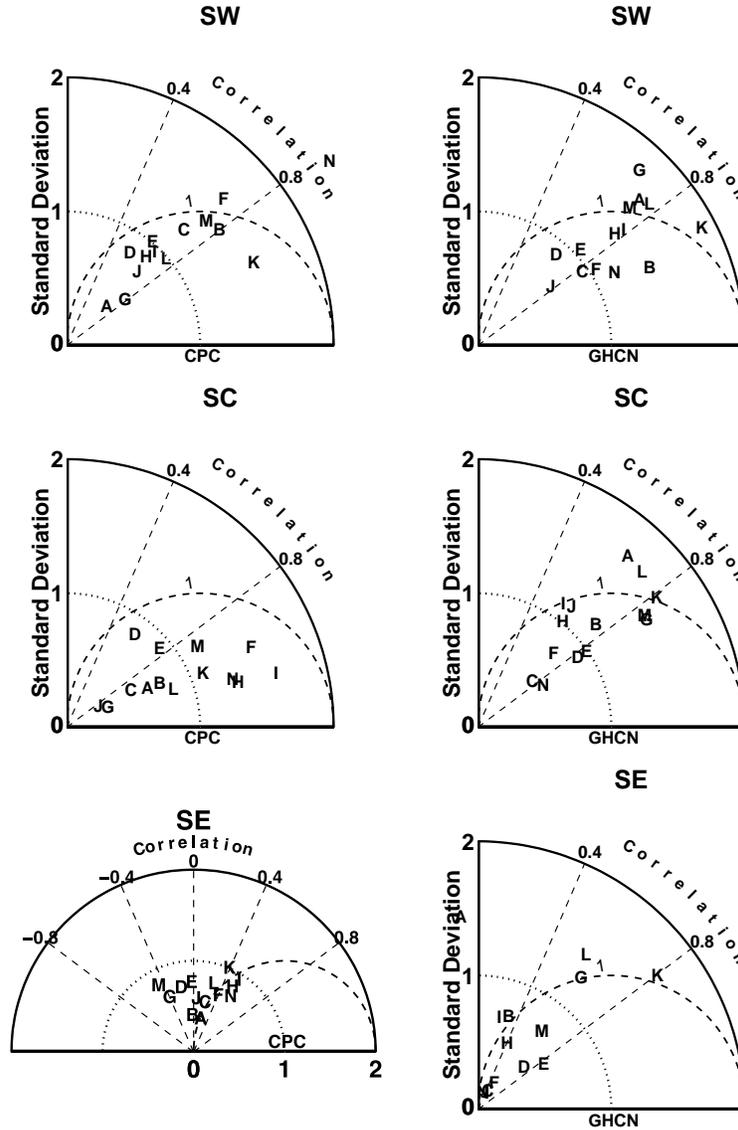
1523

1524

1525

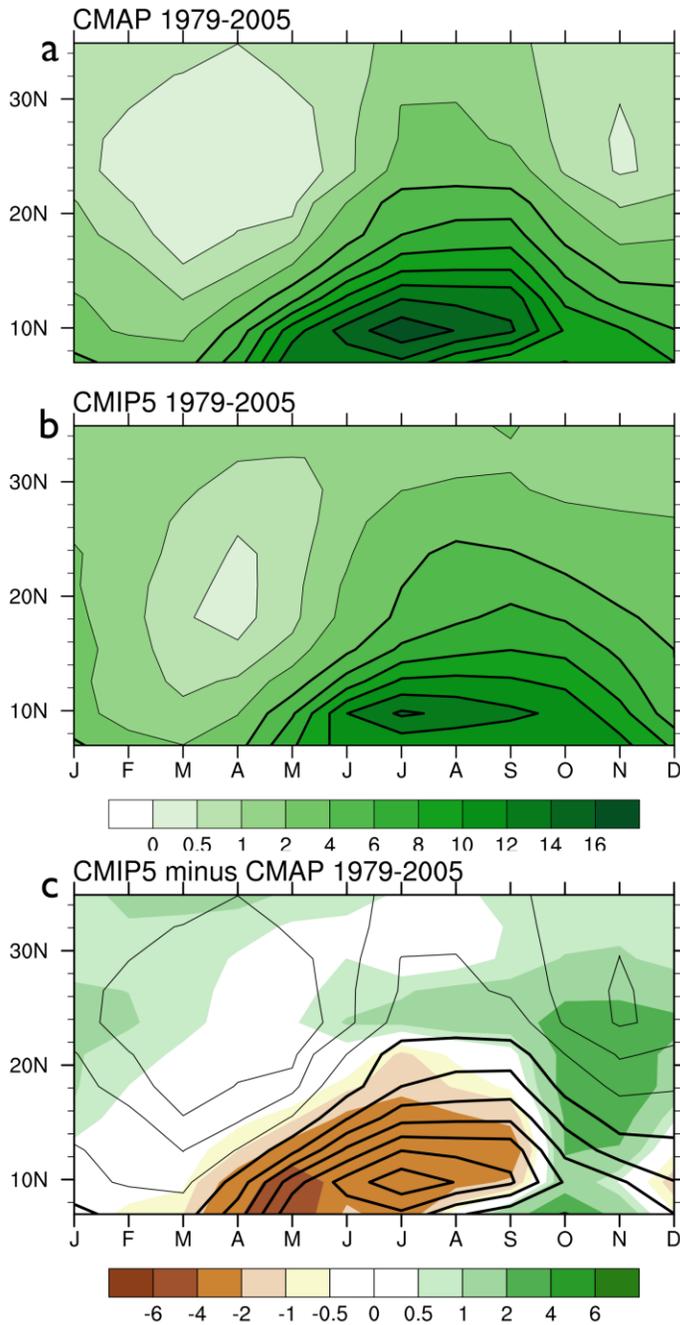
1526

1527



1528 **Figure 13.** Comparison of precipitation and temperature extremes for southern US
 1529 regions between the CMIP5 models and CPC and GHCN observations, respectively, for
 1530 1979-2005. (left column) Taylor diagram of the spatial pattern of annual number of days
 1531 when precipitation $> 10\text{mm day}^{-1}$ over the southwest (SW), south central (SC) and
 1532 southeastern (SE) US. (right column) Taylor diagram of the spatial pattern of annual
 1533 number of days when $T_{\text{max}} > 32^{\circ}\text{C}$ (90°F) for the three regions. The standard deviations
 1534 have been normalized relative to the observed values. (A: CanESM2, B: CCSM4, C:
 1535 GFDL-CM3, D: GFDL-ESM2G, E: GFDL-ESM2M, F: GISS-E2-R, G: HadCM3, H:
 1536 HadGEM2-CC, I: HadGEM2-ES, J: IPSL-CM5A-LR, K: MIROC4h, L: MIROC5, M:
 1537 MPI-ESM-LR, N: MRI-CGCM3). SW is defined as the contiguous US south of 40°N
 1538 between 125°W and 110°W ; SC is the contiguous US south of 40°N between
 1539 110°W and 90°W ; SE is the contiguous US south of 40°N between 90°W and
 1540 70°W . All data were regridded to 2.5-degree resolution.

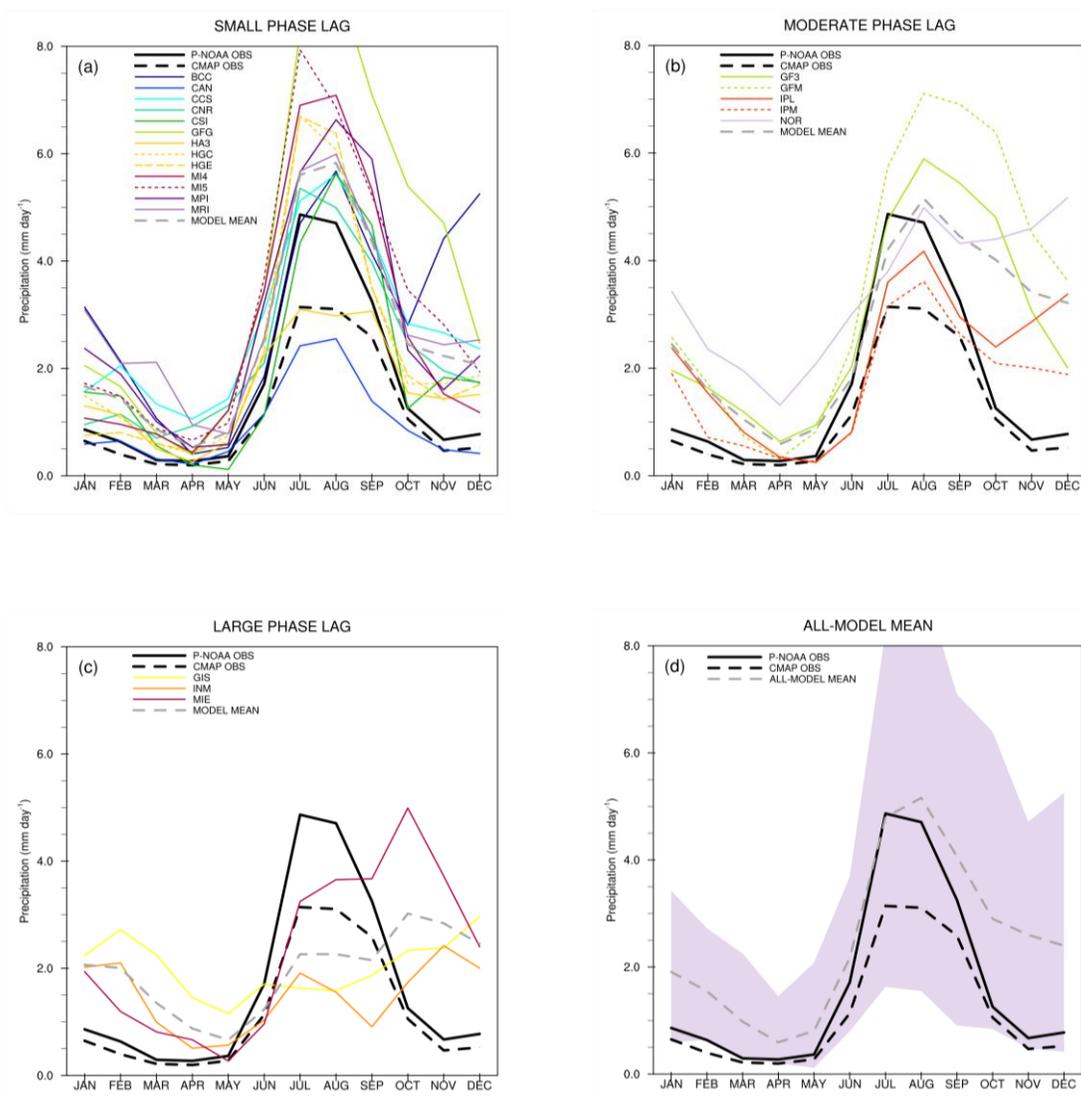
1541



1542

1543 **Figure 14.** Average monthly precipitation for 1979-2005 shown by latitude in the North
 1544 American monsoon region (longitudes 102.5 to 115W) from the CMAP observational
 1545 estimate (a), the MME mean for the 17 core CMIP5 models (b) and their difference (c),
 1546 all in units of mm day⁻¹. The CMAP and model data were regridded to T42 resolution
 1547 (~2.8-degree).

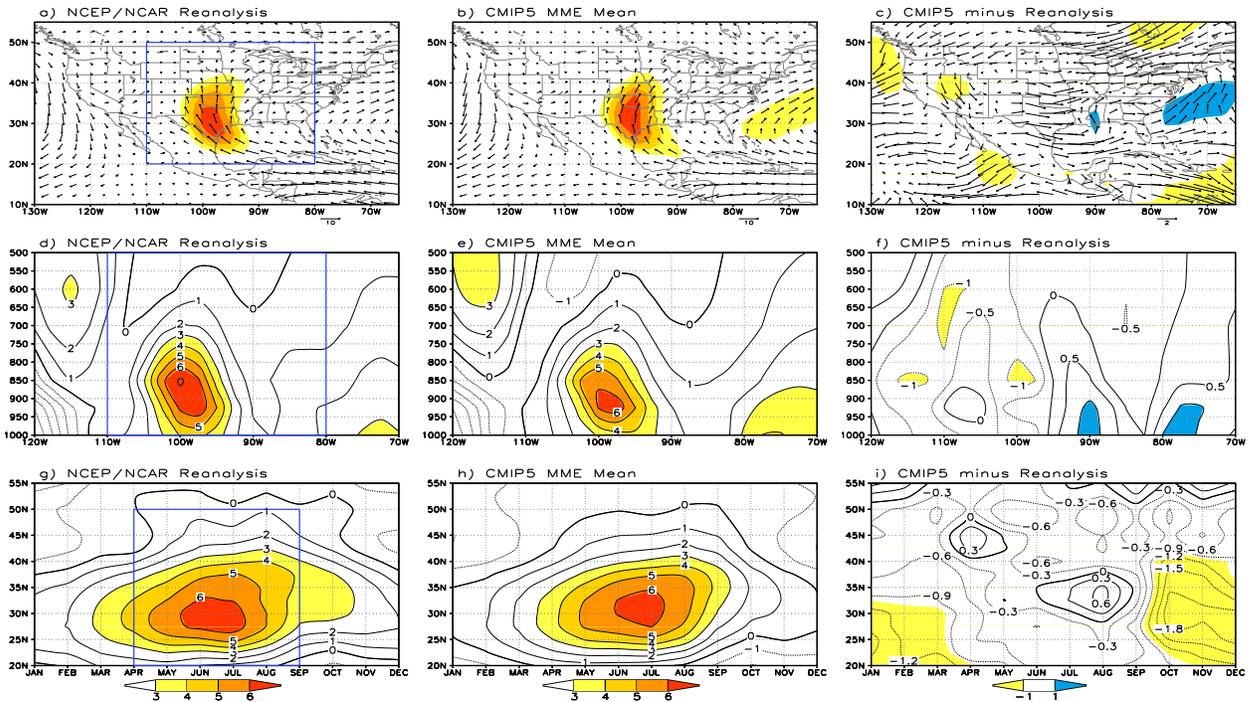
1548



1549 **Figure 15.** Annual cycle in rainfall for the North American monsoon region for the
 1550 historical (1979-2005) period of 21 CMIP5 models compared to the P-NOAA AND
 1551 CMAP observational datasets for (a) small (phase error = 0), (b) moderate (phase error =
 1552 1), (c) large (phase error = 2-4) phase errors, and (d) the MME mean and range (shading)
 1553 for all models. The CMAP and model data were regridded to the P-NOAA resolution
 1554 (0.5-degree).

1555

1556



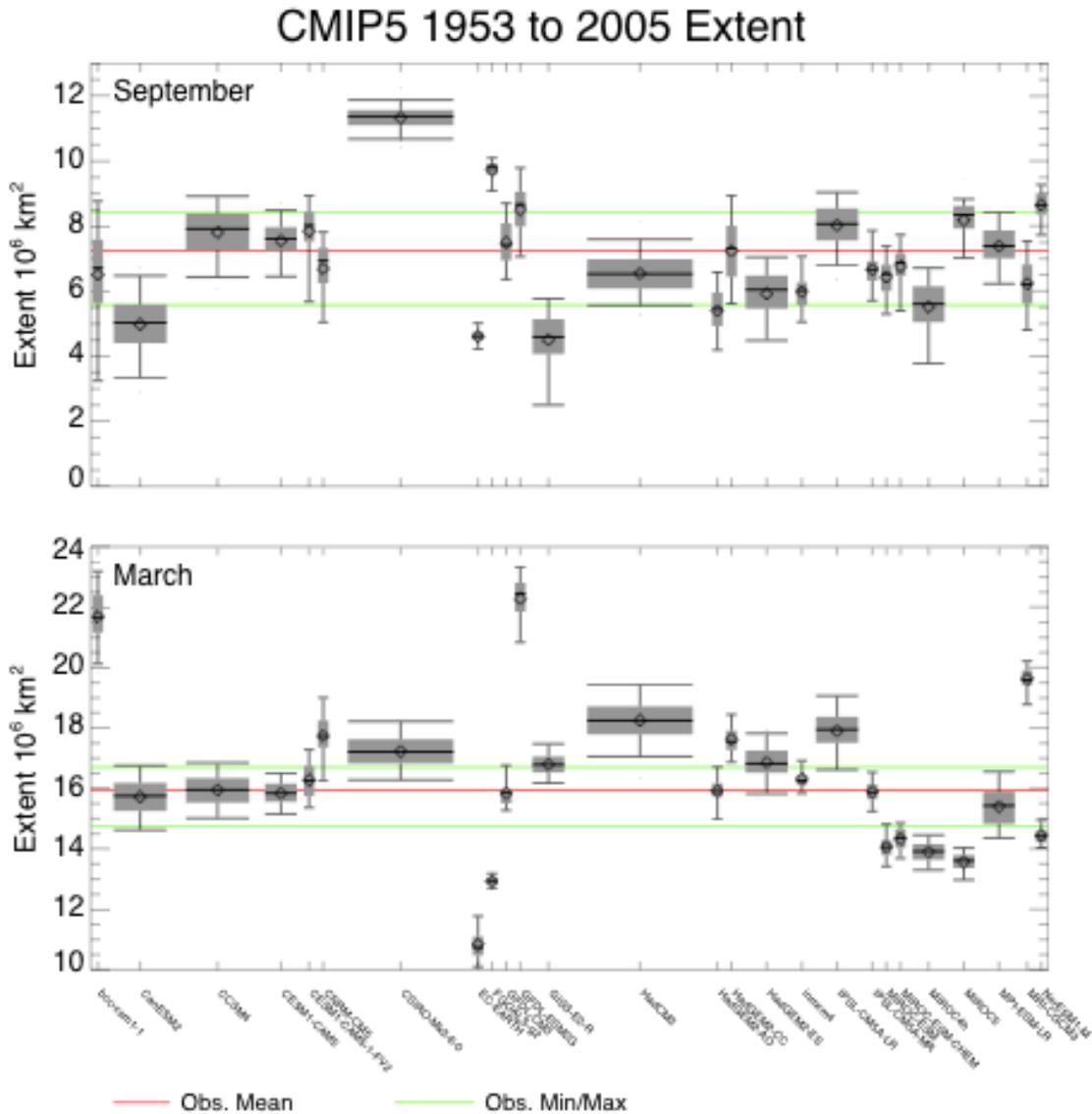
1557

1558 **Figure 16.** (a)-(c) Averaged summer 925hPa wind during 1971-2000 for NCEP-NCAR
 1559 reanalysis, eight-model CMIP5 ensemble mean for the same period, and the reanalysis
 1560 minus MME mean, respectively. (d)-(f) Lower troposphere mean vertical profile of
 1561 meridional wind averaged over 95°-100°W for the reanalysis, MME mean, and the
 1562 reanalysis minus MME mean, respectively. (g)-(i) Seasonal cycle of the 925hPa
 1563 meridional wind averaged over 27.5°-32.5°N for the reanalysis, MME mean, and the
 1564 reanalysis minus MME mean. All units are m s^{-1} . Shading indicates wind speeds greater
 1565 than 3.0 m s^{-1} in the figures of the first and second columns and wind speeds greater than
 1566 1.0 m s^{-1} in the figures of the third column. The data were regridded to 2.5-degree
 1567 resolution.

1568

1569

1570

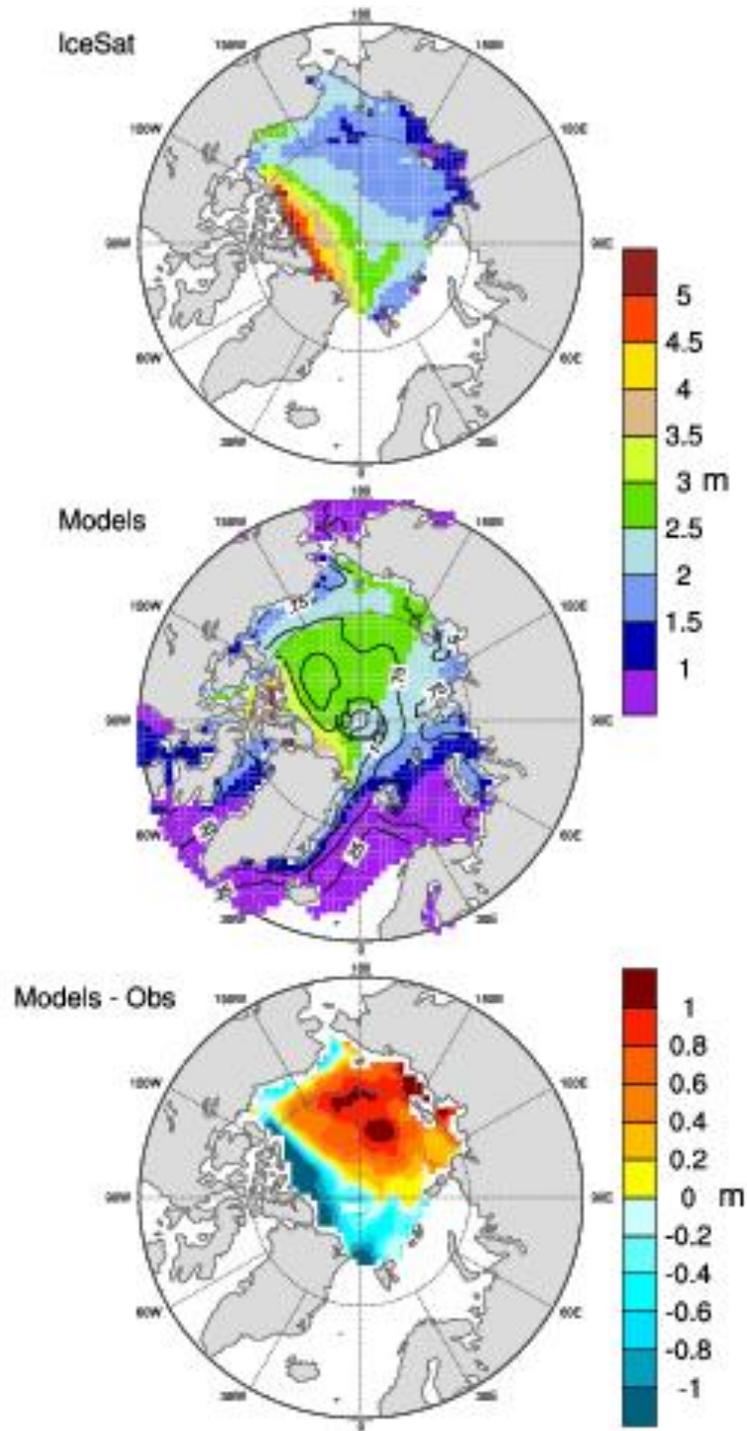


1572

1573 **Figure 17.** September and March sea ice extent from 26 CMIP5 models compared to
 1574 observations from the NSIDC from 1953 to 2005. For each model, the boxes represent
 1575 inter-quartile ranges (25th to 75th percentiles). Median (50th percentile) extents are
 1576 shown by the thick horizontal bar in each box. The width of each box corresponds to the
 1577 number of ensemble members for that model. Whiskers (vertical lines and thin horizontal
 1578 bars) represent the 10th and 90th percentiles. Mean monthly extents are shown as
 1579 diamonds. Corresponding mean, minimum and maximum observed extends are shown as
 1580 red and green lines, respectively.

1581

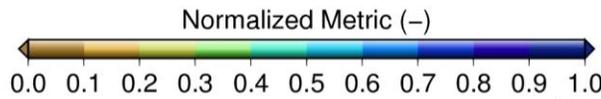
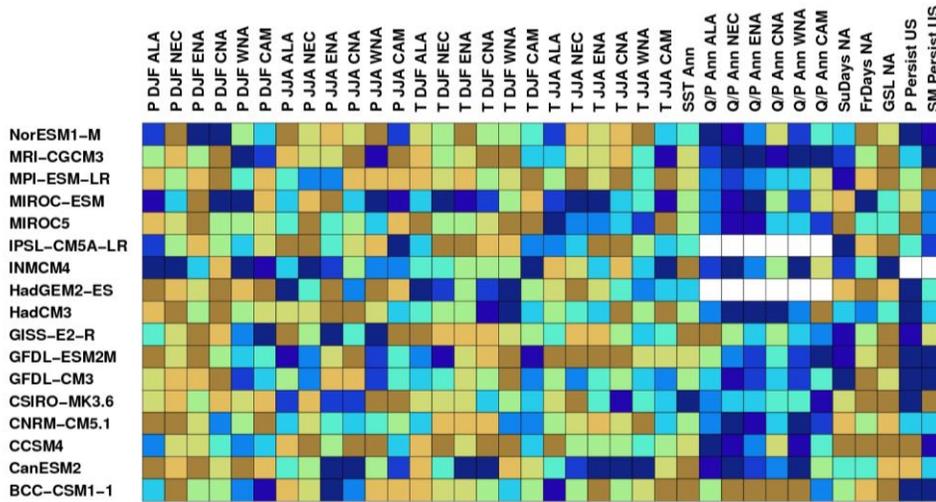
1582



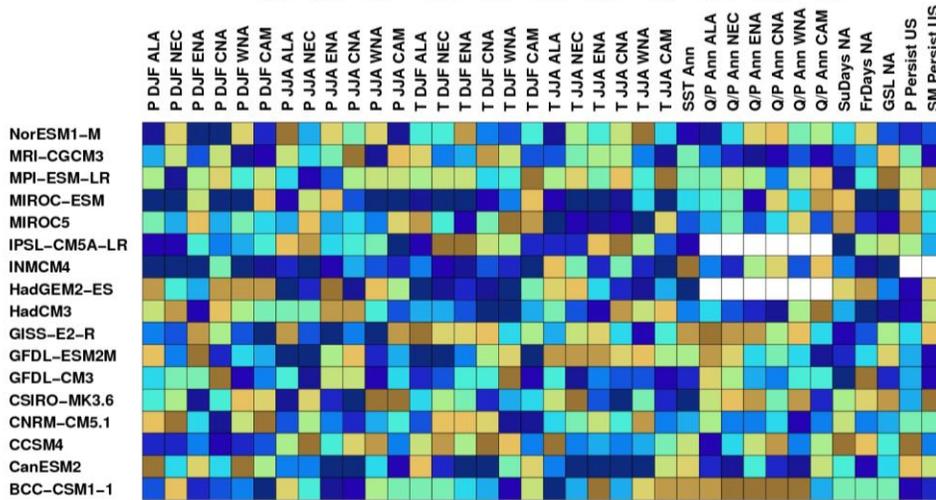
1583

1584 **Figure 18.** Observed March ice thickness (m) from ICESat for 2003 to 2009 (top),
 1585 ensemble mean ice thickness from 26 CMIP5 models (see Table 15 for list of models)
 1586 averaged over 1993 to 2005 with contours showing the MME standard deviation (middle)
 1587 and the difference between the modeled and observed (CMIP5 – observed) thickness
 1588 (bottom). The model data were regridded to the ICESat resolution.

1589

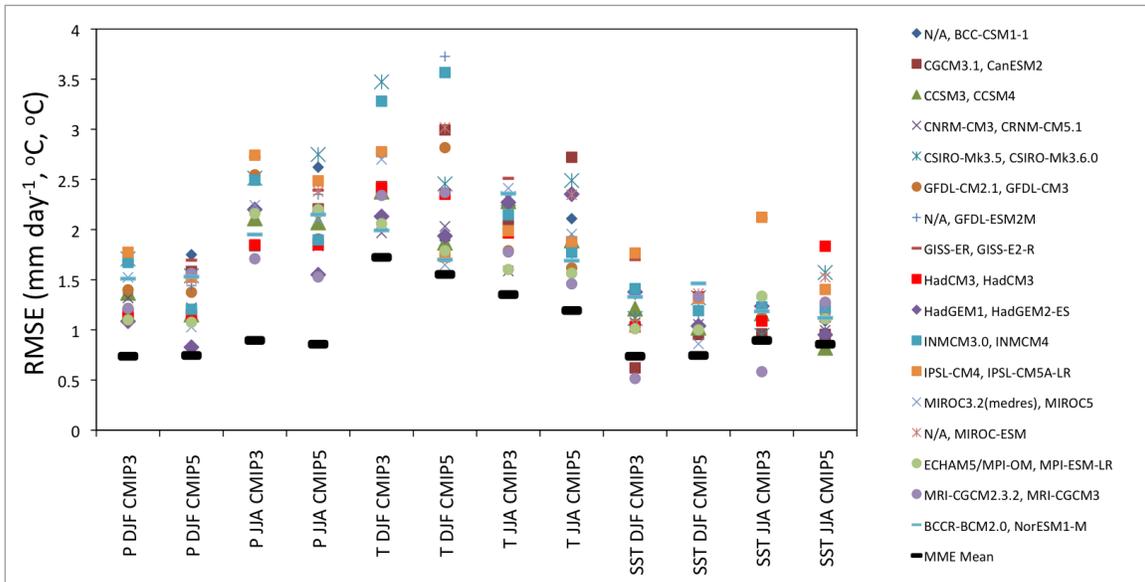


1590



1591

1592 **Figure 19.** Comparison of CMIP5 models across a set of continental performance metrics
 1593 based on bias values given in Tables 3-8. (top) Biases normalized relative to the range of
 1594 bias values across models, with lower values indicating lower bias. (bottom) Models
 1595 ranked according to bias values, with 1 indicating the model with the lowest bias and 17
 1596 the model with the highest bias. Results for models without available data are indicated in
 1597 white. The bias metrics shown (in order from left to right) are for regional precipitation
 1598 (P) for DJF and JJA, regional temperature (T) for DJF and JJA, annual SSTs for
 1599 surrouding oceans (see Figure 3), annual runoff ratios (Q/P), the annual number of
 1600 summer days (SuDays), frost days (FrDays) and growing season length (GSL), and east-
 1601 west gradient in the number of persistent precipitation (P Persist) and soil moisture (SM
 1602 Persist) events.



1603

1604 **Figure 20.** Comparison of CMIP5 and CMIP3 model performance for seasonal (DJF and
 1605 JJA) precipitation (P), surface air temperature (T) and SST. Results are shown as RMSE
 1606 values calculated for 1971-1999 relative to the GPCP, CRU and HadISST observational
 1607 datasets. Precipitation and temperature RMSE values are calculated over North America
 1608 (130-60W, 0-60N) and SST RMSE values are calculated over neighboring oceans (170-
 1609 35W, 10S-40N). The core set of CMIP5 models and their equivalent CMIP3 models
 1610 where available (otherwise indicated by N/A) are shown. The MME mean values are also
 1611 shown.

1612

1613