Analyzing ENSO teleconnections in CMIP models as a measure of model fidelity in simulating precipitation

Authors: *Baird Langenbrunner[1], J. David Neelin[1]*

1. Department of Atmospheric and Oceanic Sciences, UCLA, Los Angeles, CA 90095

Corresponding author address: *Baird Langenbrunner, Dept. of Atmospheric and Oceanic Sciences, UCLA, 405 Hilgard Ave., Los Angeles, CA 90095-1565*

Email: *baird@atmos.ucla.edu*

Abstract

The accurate representation of precipitation is a recurring issue in climate models. El Niño-Southern Oscillation (ENSO) precipitation teleconnections provide a testbed for comparison of modeled to observed precipitation. We assess the simulation quality for the atmospheric component of models in Coupled Model Intercomparison Project 5 (CMIP5), using the ensemble of runs driven by observed sea surface temperatures. Simulated seasonal precipitation teleconnection patterns (defined using linear and rank regression) are compared to observations during 1979-2005 and to the CMIP3 ensemble. Within regions of strong observed teleconnections (equatorial South America, the western equatorial Pacific, and a southern section of North America), there is little improvement in the CMIP5 ensemble relative to CMIP3 in amplitude and spatial correlation metrics of precipitation. Spatial patterns within each region exhibit a substantial departure from observations, with spatial correlation coefficients typically less than 0.5, but the models do considerably better in other measures. The amplitude of the precipitation response (root mean square over each region) is well estimated by the mean of the amplitudes from the individual models. However, the amplitude of the multi-model ensemble mean is systematically smaller (by about 30-40% in the selected teleconnection regions). The models perform well at capturing the sign of observed teleconnections over broad regions. In this way, high intermodel agreement on teleconnection sign (positive or negative precipitation response) provides a good predictor for high model with observed teleconnection signals. This supports the usefulness of these measures for assessment of precipitation in global warming projections.

1. Introduction

The El Niño-Southern Oscillation (ENSO) is a leading mode of interannual climate variability originating in the tropical Pacific. ENSO teleconnections are a reflection of the strong coupling between the tropical ocean and global atmosphere, and SST anomalies in the equatorial Pacific can have remote effects on climate globally (e.g., Trenberth et al. 1998; Wallace et al. 1998; Ropelewski and Halpert 1987; Horel and Wallace 1981).

In recent decades, measurable progress has been made in simulating ENSO dynamics and associated teleconnections within ocean-atmosphere coupled general circulation models (CGCMs) (Neelin et al. 1992; Delecluse et al. 1998; Latif et al. 2001; Davey et al. 2001; Randall et al. 2007; AchutaRao and Sperber 2006). A number of studies use the fully coupled GCMs to assess 20th century ENSO variability and teleconnections against observations (Cai et al. 2009; Joseph and Nigam 2006; Capotondi et al. 2006). Others examine the evolution of ENSO and these teleconnections under climate change (Coelho and Goddard 2009; Meehl et al. 2007b; van Oldenborgh et al. 2005; Merryfield et al. 2006). Problems persist in the ability of the models to accurately represent the tropical Pacific mean state, annual cycle, and ENSO's natural variability (e.g., Guilyardi et al. 2009a), including the role of the atmospheric component in setting the dynamics (Guilyardi et al. 2009b; Sun et al. 2009; Lloyd et al. 2009; Guilyardi et al. 2004), as well as in uncertainties in how ENSO will behave under climate change (Collins et al. 2010).

The precipitation response to interannual climate variations like ENSO also continues to be a challenge for CGCMs. In the tropics, equatorial wave dynamics spreads tropospheric temperature anomalies, which induce feedbacks with the convection zones in surrounding regions (e.g., Chiang and Sobel 2002; Su et al. 2003). At mid-latitudes, wind anomalies generated by Rossby wave trains interact with storm tracks to create precipitation anomalies

(Held et al. 1989; Chen and van den Dool 1997; Straus and Shukla 1997). These moist teleconnection processes share physical mechanisms with feedbacks active in climate change (e.g., Neelin et al. 2003) and thus examination of these can contribute to assessing the accuracy of the models for these pathways.

One difficulty with assessing teleconnections from coupled models is that errors in the ENSO simulation, for instance in amplitude or spatial distribution of the main SST anomaly in the equatorial Pacific, degrade the quality of the simulation at the source region before the teleconnection mechanisms even begin (e.g., Joseph and Nigam 2006; Coelho and Goddard 2009). To isolate the atmospheric portion of the teleconnection pathway, it is common to employ atmospheric component runs forced by observed sea surface temperature (SST), often known as Atmospheric Model Intercomparison Project (AMIP)-style runs (Gates et al., 1998). Studies using AMIP runs to examine ENSO teleconnections include Risbey et al. (2011) for teleconnections over Australia, noting errors in the modeled amplitude and coherence and Spencer and Slingo (2003), noting issues in the sensitivity of precipitation to tropical Pacific SSTs lead to errors in the Aleutian low despite otherwise accurate tropical ENSO teleconnections.

Because the challenges of correctly simulating precipitation teleconnection response are so substantial, analysis of the CMIP5 AMIP-style ensemble can provide a way to gauge the fidelity of the current generation of models in simulating large-scale atmospheric processes leading to rainfall. In particular we evaluate December-February (DJF) ENSO precipitation teleconnections during 1979-2005 in the CMIP5 models against observations, and also compare to the similar set of AMIP-style runs in the earlier CMIP3 ensemble. Faced with substantial ongoing intermodel differences and differences with respect to observations in standard evaluation measures, we turn to other measures in which the multi-model ensemble may

78 contain useful information. These include amplitude measures, comparing assessment from
79 the full model ensemble to the multi-model ensemble mean (MMEM), and measures of
80 agreement on sign. In both of these the CMIP5 model ensemble does unexpectedly well
81 compared to observations. The performance on agreement-on-sign measures proves
82 sufficiently good that it motivates questions regarding the optimal way to apply significance
83 tests within multi-model ensemble, which the discussion section attempts to formulate
84 coherently, even if a full solution is not yet clear, since these are relevant to evaluation of
85 precipitation change in global warming.
86
87 **2. Data sets and analysis**
88 To produce ENSO precipitation teleconnection patterns, we use modeled and observed
89 monthly mean SST and precipitation data during the DJF months for the years 1979-2005. For
90 SST observations, we use the Extended Reconstructed Sea Surface Temperature (ERSST.v3)
91 data set (Smith et al. 2008; Xue et al. 2003); for monthly precipitation rate observations, we
92 use the Climate Prediction Center Merged Analysis of Precipitation (CMAP) archive (Xie and
93 Arkin 1997).
94 For modeled teleconnections, we use monthly AMIP run output from the CMIP5 and CMIP3
95 multi-model ensemble archives, available for download at http://pcmdi3.llnl.gov (for more
96 information on AMIP runs, see Gates et al. 1998 and references therein). Precipitation flux
97 and surface temperature data are used in teleconnection calculations. All modeled
98 precipitation data are regridded to a 2.5°x2.5° grid prior to calculating teleconnection
99 patterns. This is the native grid of the CMAP precipitation data, and we use it to facilitate
100 direct comparison of modeled teleconnections to the observations.
101 Linear regression and Spearman rank correlation methods are used to calculate DJF

precipitation teleconnections for the time period specified. Each method is carried out using the Nino3.4 SST index (defined by a spatial average from 5ºS to 5ºN, 190ºE to 240º E; see Trenberth and Stepaniak 2001 for discussion of El Niño indices) and gridpoint-by-gridpoint precipitation data on the common CMAP grid. The use of the Nino 3.4 index yields "standard" teleconnection patterns, which provide a good first basis for comparison of models to observations. We recognize, however, that there is interesting work that addresses the next level of distinction among different "flavors" of ENSO and the remote impacts of SST anomalies that have a central (rather than eastern) Pacific signature (Kao and Yu 2009; Trenberth and Smith 2009; Ashok et al. 2007). Appropriate two-tailed significance tests are used in both the linear and rank method to resolve gridpoints that meet or pass certain confidence levels.

3. Evaluating models

a. Teleconnection patterns in linear regression and rank correlation

Figures 1 and 2 show observed and modeled precipitation teleconnections for the DJF season as estimated by linear regression and Spearman rank correlation, respectively, against the Nino3.4 SST index. The rank correlation offers an appealing alternate method, because it tends to be robust against outliers and brings regions with different amplitudes of variance on to common footing. Furthermore, for a rank correlation, the rainfall at each gridpoint is mapped onto a uniform distribution, offering a significance test that does not assume Gaussian statistics (Whitaker and Weickmann 2001; Wilks 1995). Rank correlation can also be considered a form of regression between the rank change of two variables. The linear regression, by contrast, is easier to interpret in terms of a change of the physical variables, in this case precipitation rate per unit change of SST. Linear regression and rank correlation are

both employed (as well as composites, which are not shown here but yield similar results) to check that the teleconnection patterns are robust to the estimation method. Beyond this check, the model-to-observation comparison raises some interesting questions about the restrictions of the statistical significance tests, as will be discussed in section 4.

One of the questions to arise is how best to use the information from a multi-model ensemble. Figs. 1b and 2b show linear and rank regression teleconnection patterns obtained from the multi-model ensemble. In order to obtain a single regression map for the entire CMIP5 ensemble, we can either (a) perform the regression over all 15 models simultaneously, or (b) average all teleconnection patterns from the ensemble of models. These two regression variants have been tested to arrive at the multi-model ensemble regression pattern. In Figs. 1 and 2, the variant shown is the first one – a regression performed over the time series of all 15 models simultaneously – which offers a straightforward significance test for the results of the entire ensemble. A straightforward way to interpret (and program) this is as a concatenated time series of available models, and so we will refer to this as the concatenated multi-model ensemble (CMME), when it is necessary to distinguish it.

Figs. 1b and 1d show the CMME linear regression DJF teleconnection patterns compared to observations in Figs. 1a and 1c. The multi-model ensemble pattern in Fig. 1b reproduces a number of features of the observations. A broad region of reduced precipitation over equatorial South America, stretching out through the Atlantic Intertropical Convergence Zone (ITCZ), is qualitatively simulated, although the region of the most intense anomalies is slightly displaced spatially from the observations. The region of increased precipitation starting off the coast of California and extending through Mexico, the Gulf States and beyond Florida into the Atlantic storm track similarly is qualitatively reflected in the CMME regression. In the western Pacific and surrounding the main ENSO region to the north and south, there is a broad

"horseshoe pattern" of reduced precipitation, which the CMME captures reasonably well in terms of the low amplitude parts, although the location of the most intense anomalies is off. Figures 1c and 1d show the same data as Figs. 1a and 1b but with a two-tailed significance test applied to the regression at each gridpoint. In Fig. 1d, the CMME regression passes a 95% confidence level criterion over fairly broad areas in each major teleconnection region, thanks to the large amount of model information available. Each of the areas discussed above is considered significant at this level, as are some smaller regions, such as southeastern Africa. Fig. 1c displays observed teleconnections masked to show only grid points that pass at the 90% and 95% confidence levels. This point-by-point test indicates a relatively limited area over which the regression meets these confidence criteria.

Given that the CMME yields a statistically significant prediction for the sign of the signal over the main teleconnection regions, one might consider that a one-sided t-test could be used on the observations, in which case the 90% confidence level of a two-sided test would correspond to the 95% confidence level of a one-sided test. However, when loosening the restriction from a 95% to a 90% confidence level, we only see a small increase in the areal extent of regions that pass the significance test. In comparing Figs. 1c and 1d, one can see that the multi-model ensemble has a broader area over which teleconnections are significant at the 95% confidence level, relative to the observations.

Fig. 2 displays the same information as in Fig. 1, but for a Spearman rank correlation over the CMME. The teleconnection patterns that result using either the linear or rank method are similar overall, implying that ENSO precipitation teleconnections are robust despite assumptions made about the distribution of rainfall events a priori. Differences may be noted between the two methods in particular regions, such as the rank correlation deemphasizing the narrow band along the equator in South America in the CMME (Fig. 2b) relative to the

linear regression (Fig. 1b), although not in the observations (Fig. 2a). The region passing significance criteria at the 95% level using rank correlations for the observed teleconnections (Fig. 2c) is comparable to that produced for the linear regression (Fig. 1c), and likewise for the CMME.

Teleconnection patterns for the entire ensemble were also produced by creating a multi-model ensemble mean (MMEM). This was done by performing a separate regression for each model and then averaging all 15 patterns. The MMEM pattern that results (not shown) is nearly identical to that of the CMME (global correlation coefficient greater than 0.999). (Note that a significance test for the MMEM, however, is not as straightforward.) The high correlation between these two methods is to be expected if the variance in each model is similar and is stably estimated. This implies that averaging 15 models' linear regression maps is nearly identical to performing a regression over those 15 models concatenated into a single time series; as such, in the remainder of this paper, we take teleconnection patterns for both the MMEM and CMME to be interchangeable. Additionally, we henceforth focus on teleconnection patterns estimated via linear regression due to the simpler interpretation of the amplitudes (units of mm day^{-1} C^{-1}).

b. Regional model disagreement

Another point that can be made with Figs. 1 and 2 is the large-scale agreement between CMME teleconnections and observations; surprisingly, this is more apparent when 95% significance criteria are not imposed. However, regional disagreement between observations and the MMEM pattern may also be seen, especially in regions where the observations have intense precipitation, while the MMEM exhibits a general "smoothing" of teleconnection patterns.

These smoothed teleconnection patterns in the MMEM can be understood when examining patterns from individual models. Fig. 3 shows teleconnections for one run of each model, displayed for the equatorial Americas. There is very substantial regional variability among them. Differences in the location of the strongest precipitation anomaly are common from model to model, even though these are the areas that most easily pass significance criterion on a model by model basis (local significance tests for individual models, not shown, are qualitatively similar to the areal extent in Fig. 1c). However, over the region where the CMME passes the t-test at the 95% level, one can see the overall teleconnection pattern is plausible at large scales in each of the models. Thus, Fig. 3 provides a visual sense of the trade-offs to be quantified: disagreement among models at regional scales; excessive smoothing relative to observations in the MMEM; and yet some possibility that there is useful information about the teleconnection pattern in the model ensemble, if it can be suitably extracted.

c. Taylor diagram analysis of modeled teleconnections

The regional variation among AMIP models leads to a distinction between their ability (1) to reproduce spatial patterns of teleconnections, and (2) to represent the amplitudes of these patterns. To examine individual model fidelity in simulating patterns and amplitude of rainfall teleconnections, we look at four regions that show a robust ENSO response; each region displays a continuous teleconnection signal significant at the 95% confidence level, when examining Fig. 1c.

These four regions include (a) the equatorial Pacific (the “cold tongue” region; positive DJF ENSO signal), (b) the horseshoe-shaped region in the western Pacific (negative signal), (c) equatorial South America (negative signal), and (d) a southern section of North America (positive signal). The equatorial Pacific region is shown for reference, since this is the source

222 region, directly forced by the largest ENSO-related SST anomalies. We consider the other
223 three regions the "teleconnection regions," since to accurately simulate teleconnected
224 rainfall in each of them, the models must capture the pathways leading to remote
225 precipitation change. The Taylor diagrams in Fig. 4 show the spatial correlations between the
226 observations and each model (radial axis) plotted along with the spatial root mean square
227 (RMS) amplitude of each model pattern, normalized by the observed RMS. Correlations and
228 amplitudes are averaged among runs for models with multiple realizations; each individual
229 model is given equal contribution in the MMEM.
230 As seen in Fig. 4, there is little (if any) improvement from CMIP3 to CMIP5 in reproducing
231 teleconnected rainfall patterns in these regions. In addition, note the generally low
232 correlations (ranging from less than 0.2 to a few instances exceeding 0.7, with an average
233 correlation coefficient of about 0.40) between each model and observations. In every region,
234 one can also see that the MMEM is typically more accurate than the majority of individual
235 models in reproducing these patterns. However, the normalized amplitude measure shows
236 that the MMEM amplitude is substantially lower than teleconnection amplitudes of the
237 individual ensemble members. Furthermore, in every region outside of the central equatorial
238 Pacific, the MMEM underestimates the observations. As a final point, we note that Taylor
239 diagrams of the corresponding rank correlation method (not shown) indicate consistent
240 results.
241
242 *d. Teleconnection amplitude in major impact regions*
243 The varied agreement in amplitude measures from Fig. 4 suggests that it may be more
244 reasonable to use amplitude information from individual ensemble members, rather than
245 using that of the MMEM. To get a better sense of how teleconnection amplitude of individual

246 models might be affected by internal variability, we take advantage of AMIP models with
247 multiple realizations, and we assess the internal variability among these runs for each model.
248 We then compare this to the amplitude of the entire ensemble of models. Fig. 5 displays the
249 radial axis from the Taylor diagrams discussed previously, but where multiple runs from each
250 model are available, we plot them individually (43 total runs for 15 models in CMIP5; 26 total
251 runs for 13 models in CMIP3).
252 The vertical extent of the black lines in Fig. 5, representing ± one standard deviation among
253 the multiple runs for one model (when available), is a measure of internal variability for a
254 given model's runs, while that of the green bars (± one standard deviation of the multi-model
255 ensemble) is a measure of intermodel variability. Notable points from this diagram include:
256 (1) The MMEM systematically underestimates the spread and central tendency of intermodel
257 variability, with a low bias of about 20-40% outside of the immediate ENSO region; (2) the
258 regional disagreement among models owes itself partly to internal model variability, but
259 intermodel variability contributes to the majority of the regional disagreement seen in Figure
260 3; (3) models are overestimating the amplitude in the immediate ENSO region for CMIP5, even
261 though their spread is more symmetric about the observations in remote regions; (4) when
262 comparing CMIP5 to CMIP3, CMIP5 shows no consistent improvement or change due to model
263 development. Although the MMEM may fall closer to observed amplitudes in CMIP5, this
264 comes at the expense of a tendency to overestimate rainfall teleconnections in the central
265 ENSO region.
266 Fig. 5 suggests that serious errors can result from considering only information available in the
267 MMEM. While its spatial patterns tend to do less poorly in spatial correlation with
268 observations than most individual models, the MMEM teleconnection amplitude is routinely
269 too low in the remote regions considered. It is therefore useful to consider measures of

teleconnection amplitude and spread from individual models, in addition to the MMEM, in situations where regional disagreement can dampen the MMEM amplitudes due to averaging varied model signals.

4. Sign agreement plots in ENSO teleconnections and their relation to precipitation change in global warming scenarios

Agreement plots for the sign of precipitation change under global warming scenarios have been commonly used in multi-model studies (e.g., see Randall et al. 2007; Meehl et al. 2007), often being used as complementary information to the MMEM. Agreement-on-sign tests can be viewed as relatively weak statements regarding the precipitation change, and it has been argued that these should be used in conjunction with requirements that the signal pass local tests for significant difference from zero on a model-by-model basis (e.g., Neelin et al. 2006; Tebaldi et al. 2011, hereafter N06 and T11).

Here we examine agreement-on-sign measures based on the ENSO precipitation regression patterns for each model. Because we can also assess these against observations, we can use this to examine the procedure as a means of inferring its usefulness for global warming, where observations of large precipitation change will not be feasible for at least some decades.

Fig. 6a shows the traditional agreement-on-sign plot, but for precipitation teleconnections in the CMIP5 AMIP ensemble. At each grid point, we count the number of models that agree on a negative (positive) DJF teleconnection signal, so that the agreement plot shows the integer value of models which agree on a dry (wet) regression slope during ENSO. Note areas with high agreement on sign cover a much larger spatial region than those passing the two-tailed significance test at the 95% level applied to observations (see Fig. 1c).

Fig. 6b provides one way of summarizing the significance information from a traditional two-tailed t-test applied to each model of the CMIP5 AMIP ensemble. The number of models that pass a traditional t-test at 95% confidence is counted at each gridpoint, so that the color bar represents the number of models meeting this criterion. This figure highlights areas in which 80% of the ensemble members pass this test. Fig. 6b, along with the agreement in Fig. 6a, provides a sense of what the N06 and T11 tests would indicate. Essentially, the former would count only sign agreement among models passing a t-test at a defined level at each gridpoint, and the latter would screen out regions where less than a certain fraction of models individually pass a t-test. It is immediately apparent that either of these will be highly restrictive in the regions considered, despite high agreement on sign. By contrast, note the similarity between the regions passing the significance test on the full CMME (Fig. 1d) and the areal extent of high agreement in Figs. 6a, 7a, and 7b.

Most of the signal that survives such a model-by-model t-test criterion in Fig. 6b occur in gridpoints with a positive precipitation anomaly. This is perhaps consistent with the hypothesis that a distribution of rainfall events may be affecting the applicability of the t-test, as rainfall event distributions can have a long tail at larger values but are cut off at zero at the lower end. However, using Spearman rank correlation values and an associated significance test that does not require Gaussian assumptions (not shown) does not differ strongly from the counting results in Fig. 6b.

The models thus exhibit high agreement on sign over much more extensive spatial regions than what a local t-test criterion would imply. In comparing Figures 6a and 6b, it is useful to consider to what extent sign agreement can provide an alternative measure of confidence. The fact that there are regions exhibiting very high agreement on sign suggests that there is information contained within the ensemble that would be excluded by these criteria. This

318 leads us to conjecture that the N06 and T11 criteria may both be too restrictive. Whereas N06
319 and T11 were making statements about predicted precipitation change, we have the option of
320 checking the modeled teleconnections against observations. In particular, we can produce an
321 agreement plot that counts model agreement on the sign of observed rainfall teleconnection
322 patterns (Fig. 7a). This agreement plot clearly indicates that the models agree with the sign
323 of the observed teleconnection signal over broad regions.
324 Fig. 7a is suggestively similar to the traditional agreement on sign plot in Fig. 6a, but the two
325 computations are not directly comparable. To make the comparison to the model ensemble
326 more direct, we use a variant of the agreement plot: the number of models that agree with
327 the sign of the MMEM (Fig. 7b). The MMEM serves as a model-based hypothesis used to check
328 the ensemble members. One detail in this computation is that we exclude each model from
329 the MMEM when determining agreement on sign, so as to avoid inflating the agreement count.
330 The plot for agreement on sign between the CMIP5 AMIP models and the corresponding MMEM
331 is shown in Fig. 7b. The information on regions of high agreement is largely the same as that
332 in Fig. 6a. Note that where all 15 models agree on sign, the count will be 15 in both
333 measures, although the meaning of these measures will differ in the middle of the range.
334 Areas are not shaded where fewer than 80% of models agree with one another (Fig. 6a) or
335 with the MMEM or observations (Figs. 7b or 7a).
336 A measure of the usefulness of the model-to-MMEM agreement plot (Fig. 7b) is how well it
337 compares to the plot of model-to-observations agreement (Fig. 7a). Broad spatial areas
338 exhibit high agreement with the sign of the observations in Fig. 7a, both for teleconnected
339 precipitation increases and decreases. It is notable that these areas are substantially more
340 widespread than would be suggested by the t-test in Fig. 1b. Comparing regions of high
341 model-to-MMEM agreement (Fig. 7b) to the corresponding agreement with observations (Fig.

342 7a), one can see many similarities. Regions where more than 80% of models agree with the
343 MMEM on sign are reflected in model agreement with observations. In short, high model-to-
344 MMEM agreement on sign does well at predicting where the models will be in high agreement
345 with the sign of the observations.
346
347 **5. Discussion**
348 As discussed in the previous section, Figs. 6 and 7 suggest that there are substantial regions
349 where models from the CMIP5 AMIP ensemble are providing useful information on the sign of
350 rainfall teleconnections, despite both individual models and the observations failing to meet
351 significance criteria in some regions.
352 A first conjecture for this discrepancy would be that an inherent non-Gaussianity of the
353 rainfall distribution might affect the applicability of the t-test, even at seasonal timescales.
354 However, applying a significance test that does not depend on Gaussian assumptions for the
355 corresponding teleconnection patterns estimated with rank correlation yields results that are
356 not very different from those of the linear regression t-test. A second postulate is that the
357 agreement on sign both uses information from the full model ensemble and tests a slightly
358 different hypothesis than difference from zero. This is evident in comparing Fig. 7b to Fig. 1d,
359 one can see that areas of high agreement on sign tend to coincide with areas that pass a t-
360 test at 95% confidence.
361 Even taking this into account, there remains the intriguing question of why the models agree
362 so well with the observations on the sign of the teleconnection pattern. There are two
363 aspects to this question: one statistical, and the other physical. The statistical aspect is that
364 only areas of high amplitude tend to pass the t-test at high confidence in the satellite
365 observational record. The physical aspect is that the models appear to perform better at

366 capturing broad areas of modest amplitude than they do at the spatial correlation (a measure
367 that is affected by poor model skill in correctly positioning high amplitude signals).
368 We postulate that this may be associated with the physics of the system, in which more than
369 one physical process operates in ENSO teleconnections. Specifically, there are processes that
370 will have smaller intermodel uncertainty and smaller internal variability but are widespread
371 spatially. An example for this could be a increase in tropospheric temperature driving an
372 increase in water vapor and a corresponding increase in the threshold for convection (a
373 process sometimes referred to as the "rich-get-richer" mechanism; Chou and Neelin 2004;
374 Held and Soden 2006; Trenberth 2011).
375 At the same time, feedbacks associated with dynamical changes in moisture convergence can
376 produce large excursions from expected values of precipitation, both in intermodel and
377 temporal variability. The models contain reasonable approximations to each of these
378 processes, and they may very well provide more information in regions where a test based on
379 a limited characterization of the distribution would suggest limited significance. In other
380 words, it is possible that different physical processes at work in the atmospheric response are
381 producing a convolution of different distributions for which a t-test with a null hypothesis of
382 zero signal is less than optimal.
383 Another example of models providing better than expected estimates of teleconnection
384 pattern from a short record is found in Risbey et al. (2011), for composite plots for rainfall
385 teleconnections using a 30-year record with a two-tailed t-test. The authors note that the
386 number of gridpoints passing a 95% significance criterion is much fewer than the same method
387 applied to a century of historical data. As a result, they loosen their restriction to an 80%
388 confidence interval, noting that the teleconnection patterns are similar for records of either
389 length.

While we cannot undertake a comprehensive examination here of the available significance tests that could illuminate the information a model ensemble provides, the cases here suggest that further work on this would be fruitful. In particular, it appears that previous criteria for assessing significance on an individual model basis (e.g., N06 and T11) may be excessively restrictive, excluding information that would be admitted by t-tests based on the full model ensemble or by agreement-on-sign criteria that prove to be correct when compared to observations.

6. Summary and conclusions

AMIP runs from the CMIP3 and CMIP5 ensembles provide one standard by which we can judge the ability of the CGCMs' atmospheric components to reproduce dynamic feedback processes that lead to remote precipitation. We focus on standard teleconnection patterns associated with the ENSO Nino3.4 index. Comparisons among the ensemble of models and with the observations are made using precipitation teleconnection patterns for the DJF for the years 1979-2005. The spatial patterns and amplitudes of these teleconnections are analyzed in several regions with robust ENSO feedbacks, including the eastern tropical Pacific, the "horseshoe" region in the western tropical Pacific, a southern section of N. America, and equatorial S. America. Teleconnection patterns are examined using both regression (linear and rank) and compositing techniques, all with similar results. The rank method provides an alternate significance test, which is useful in narrowing some of the questions that arise for regions of low amplitude signal. The patterns defined with linear regression are useful for questions that involve the amplitude of the signal, since the amplitude of the precipitation change per SST change has an easier physical interpretation than amplitude given in terms of rank.

How well the models perform at reproducing the observed teleconnection pattern depends strongly on the quantity for which they are assessed. In standard measures of spatial correlation and amplitude, here taken over a set of regions where the teleconnection signal is reasonably strong, the CMIP3 and CMIP5 AMIP models exhibit strong regional disagreement with one another and with observations. Comparing patterns visually, this is associated with regions of strong precipitation change differing substantially from model to model and also with respect to observations. This yields low spatial correlations between modeled and observed teleconnections, with correlation coefficients on the order of 0.40 on average in the defined teleconnection regions. The MMEM performs marginally better than the majority of models in spatial correlation measures, largely because the regions of strongest change have been smoothed.

The MMEM systematically underestimates amplitude measures of the regional precipitation response, typically falling more than one standard deviation below the mean of the model ensemble. This low amplitude in the MMEM is again associated with regional disagreement among ensemble members in the placement of high amplitude precipitation change. It is expected that in presence of intermodel variability, the MMEM amplitude will be lower than the individual model amplitudes (a fact noted in other analyses of GCM ensembles, e.g., N06; Knutti et al. 2010b; Neelin et al., 2010; Schaller et al. 2011; Räisänen 2007).

Here, we quantify this, and ask whether better information can be obtained from the ensemble than that permitted by the MMEM. This includes assessment of the degree to which internal atmospheric variability contributes to regional disagreement among models. Teleconnection amplitudes of individual CMIP5 models distribute accurately about the observed values in all regions but the central ENSO region. Internal variability of this amplitude is significant within each model, but it does not dominate the intermodel spread.

Thus it is intermodel variability – and not the internal variability within each model – that is the major factor in causing the MMEM to perform poorly in amplitude measures. In sum, the MMEM underestimates the observed by 30-40% in the teleconnection regions, but the mean of the individual model amplitudes provides a good estimate of the observed amplitude.

Measures of the agreement on the sign of a precipitation response in model ensembles are often used for assessing global warming precipitation changes. Examining sign agreement for the teleconnection patterns, the model ensemble has broad spatial regions with high levels of agreement. These regions are more spatially extensive than the spatial regions for which individual models (or the observations) would pass a two-tailed t-test for a signal significantly different from zero at the 95% (or even the 90%) level. Surprisingly, the models exhibit high agreement on sign with the *observations* over similarly broad regions. In other words, high agreement on sign within the model ensemble is a good predictor for agreement on sign with observations for ENSO teleconnections.

Based on this agreement-on-sign analysis, it may be inferred that the model ensemble is producing useful information regarding the teleconnection precipitation signal even in regions that do not pass a t-test at the 95% level for individual models. This may in part be due to the fact that the full ensemble is being used in the sign test, and so this measure benefits from more information. A t-test for regression patterns using the full multi-model ensemble indicates comparably large regions that pass the significance test at the 95% level. It may also in part be associated with the models correctly producing some of the multiple mechanisms that contribute to the precipitation signal, such as skill at simulating mechanisms that lead to low amplitude signal occurring, despite issues with reproduction of intense, localized precipitation change.

The evaluation of the model simulations for ENSO teleconnections may be used, with due

462 caution, to draw inferences for assessment of precipitation in global warming projections.
463 Many of the processes producing the precipitation change are analogous to the global warming
464 case. In particular, widespread tropospheric warming initiates precipitation change
465 mechanisms in the tropics in both teleconnections and global warming. Regions of strong
466 convergence feedbacks in certain tropical regions, and regions where large-scale wave
467 dynamics interacts with mid-latitude storm tracks, producing localized precipitation
468 anomalies with high amplitude and high intermodel variation, are analogous in both cases.
469 The unimpressive skill of the models at capturing the precise regional distribution of the high
470 intensity changes compared to teleconnection observations is consistent with the poor
471 intermodel agreement on a precise pattern of precipitation change in global warming.
472 However, the skill of the ensemble at reproducing the observed teleconnection signal
473 amplitude (provided it is not assessed from the MMEM) suggests that corresponding measures
474 for global warming precipitation change may be trustworthy. Furthermore, the surprisingly
475 good skill of agreement-on-sign measures from the model ensemble at predicting the sign of
476 the observed teleconnection bodes well for the usefulness of such measures in anticipating
477 the sign of precipitation changes associated with global warming.

Acknowledgements. This work was supported in part by the NOAA Climate Program Office Modeling, Analysis, Predictions and Projections (MAPP) Program under grant NA11OAR4310099 as part of the CMIP5 Task Force and National Science Foundation grant AGS-1102838. We thank M. Munnich for insights into the behavior of rank correlation estimates of teleconnections. CMAP precipitation data and NOAA_ERSST_V3 SST data are provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their website at http://www.esrl.noaa.gov/psd/. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. For CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

References

AchutaRao, K., and K. Sperber, 2006: ENSO simulations in coupled ocean-atmosphere models: Are the current models better? *Climate Dyn.*, **27**, 1–16.

Ashok, K., S. K. Behera, S. A. Rao, H. Weng, and T. Yamagata, 2007: El Niño Modoki and its possible teleconnection. *J. Geophys. Res.*, **112**, C11007.

Cai, W., A. Sullivan, and T. Cowan, 2009: Rainfall teleconnections with Indo-Pacific variability in the WCRP CMIP3 models. *J. Climate*, **22**, 5046–5071.

Capotondi, A., A. Wittenberg, and S. Masina, 2006: Spatial and temporal structure of Tropical Pacific interannual variability in 20th century coupled simulations. *Ocean Modell.*, **15**, 274.

Chen, W. Y, and H. M. van den Dool, 1997: Asymmetric impact of tropical SST anomalies on atmospheric internal variability over the North Pacific. *J. Atmos. Sci.*, **54**, 725-740.

Chiang, J. C. H., and A. H. Sobel, 2002: Tropical tropospheric temperature variations caused by ENSO and their influence on the remote tropical climate. *J. Climate*, **15**, 2616–2631.

Chou, C., and J. D. Neelin, 2004: Mechanisms of global warming impacts on regional tropical precipitation. *J. Climate*, **17**, 2688–2701.

Coelho, Caio A. S., and Lisa Goddard, 2009: El Niño–Induced Tropical Droughts in Climate Change Projections. *J. Climate*, **22**, 6456–6476.

Collins, M., and Coauthors, 2010: The impact of global warming on the tropical Pacific Ocean and El Niño. *Nat. Geosci.*, **3**, 391–397.

Davey, M., and Coauthors, 2001: STOIC: A study of coupled model climatology and variability in tropical regions. *Climate Dyn.*, **18**, 403–420.

Delecluse, P., M. K. Davey, Y. Kitamura, S. G. H. Philander, M. Suarez, and L. Bengtsson, 1998: Coupled general circulation modeling of the tropical Pacific. *J. Geophys. Res.*, **103** (C7), 14 357–14 373.

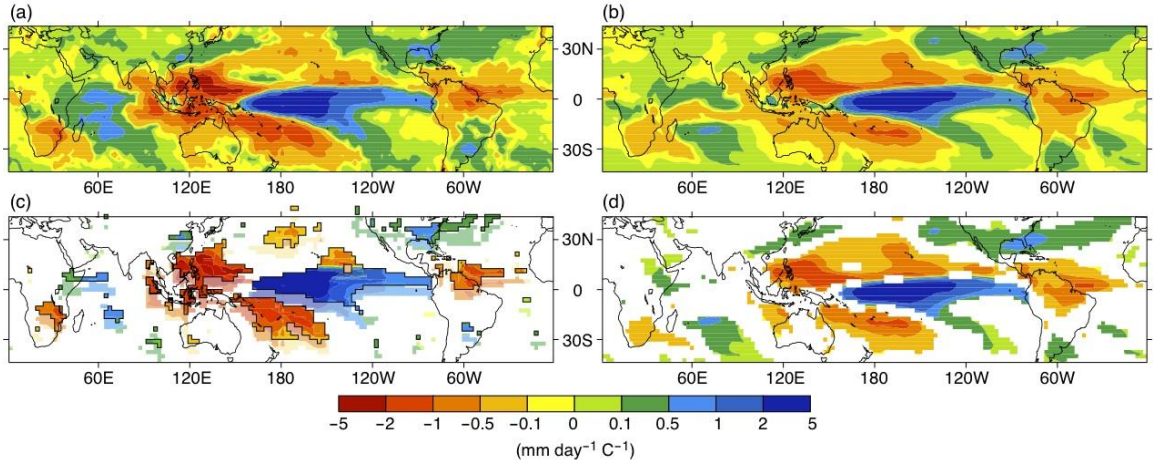
514 DeWeaver, E., and S. Nigam, 2004: On the forcing of ENSO teleconnections by anomalous
515 heating and cooling. *J. Climate*,**17,** 3225-3235.
516 Gates, W. L., and Coauthors, 1998: An overview of the results of the Atmospheric Model
517 Intercomparison Project (AMIP I). *Bull. Amer. Meteor. Soc.*, **73**, 1962–1970.
518 Guilyardi, E., A. Wittenberg, A. Fedorov, M. Collins, C. Wang, A. Capotondi, G. van
519 Oldenborgh, and T. Stockdale, 2009a: Understanding El Niño in ocean–atmosphere general
520 circulation models: Progress and challenges. *Bull. Amer. Meteor. Soc.*, **90**, 325–340.
521 Guilyardi, E., P. Braconnot, F.-F. Jin, S. T. Kim, M. Kolasinski, T. Li, and I. Musat, 2009b:
522 Atmosphere feedbacks during ENSO in a coupled GCM with a modified atmospheric convection
523 scheme. *J. Climate*, **22**, 5698–5718.
524 Guilyardi, E., and Coauthors, 2004: Representing El Niño in coupled ocean–atmosphere GCMs:
525 The dominant role of the atmospheric component. *J. Climate*, **17**, 4623–4629.
526 Held, I. M., and B. J. Soden, 2006: Robust responses of the hydrological cycle to global
527 warming. *J. Climate*, **19**, 5686– 5699.
528 Held, I. M., S. W. Lyons, and S. Nigam, 1989: Transients and the extratropical response to El
529 Niño. *J. Atmos. Sci.*, **46**, 163-174.
530 Horel, J. D., and J. M. Wallace, 1981: Planetary-scale atmospheric phenomena associated
531 with the Southern Oscillation. *Mon. Wea. Rev.*, **109,** 813–829.
532 Joseph, R., and S. Nigam, 2006: ENSO evolution and teleconnections in IPCC's Twentieth-
533 Century climate simulations: Realistic representation? *J. Climate*, **19**, 4360-4377.
534 Kao, H.-Y., and J.-Y. Yu, 2009: Contrasting eastern-Pacific and central-Pacific types of ENSO.
535 *J. Climate*, **22**, 615–632.
536 Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2010: Challenges in combining
537 projections from multiple climate models. *J. Climate*, **23**, 2739–2758.

538 Latif, M., and Coauthors, 2001: ENSIP: The El Niño Simulation Intercomparison Project.
539 *Climate Dyn.*, **18**, 255–272.
540 Lloyd, J., E. Guilyardi, H. Weller, and J. Slingo, 2009: The role of atmosphere feedbacks
541 during ENSO in the CMIP3 models. *Atmos. Sci. Lett.*, **10**, 170–176.
542 Meehl, G. A., and Coauthors, 2007a: Global climate projections. *Climate Change 2007: The*
543 *Physical Science Basis*, S. Solomon, et al., Eds., Cambridge University Press, 747–845.
544 Merryfield, W., 2006: Changes to ENSO under CO2 doubling in a multimodel ensemble. *J.*
545 *Climate*, **19**, 4009–4027.
546 Neelin, J.D., A. Bracco, H. Luo, J. C. McWilliams, and J. E. Meyerson, 2010: Considerations
547 for parameter optimization and sensitivity in climate models. *Proc. Natl. Acad. Sci.*, **107**, 21
548 349–21 354.
549 Neelin, J. D., M. Munnich, H. Su, J. E. Meyerson, and C. E. Holloway, 2006: Tropical drying
550 trends in global warming models and observations. *Proc. Natl. Acad. Sci.*, **103,** 6110-6115.
551 Neelin, J.D., C. Chou, and H. Su, 2003: Tropical drought regions in global warming and El Niño
552 teleconnections. *Geophys. Res. Lett.*, **30,** 2275.
553 Neelin, J. D., and Coauthors, 1992: Tropical air–sea interaction in general circulation models.
554 *Climate Dyn.*, **7,** 73–104.
555 Oldenborgh, G.J. van, and T. Stockdale, 2009: Understanding El Niño in Ocean-Atmosphere
556 General Circulation Models: Progress and challenges. *Bull. Amer. Met. Soc.*, **90**, 325-340.
557 Räisänen, J., 2007: How reliable are climate models? *Tellus*, **59A**, 2–29.
558 Randall, D. A., and Coauthors, 2007: Climate models and their evaluation. *Climate Change*
559 *2007: The Physical Science Basis*, S. Solomon et al., Eds., Cambridge University Press, 589–
560 662.
561 Risbey, J. S., P. C. McIntosh, M. J. Pook, H. A. Rashid, and A. C. Hirst, 2011: Evaluation of

562 rainfall drivers and teleconnections in an ACCESS AMIP run. *Australian Meteorological and*
563 *Oceanographic Journal*, **61,** 91-105.
564 Ropelewski, C. F., and M. S. Halpert, 1987: Global and regional scale precipitation patterns
565 associated with the El Niño/ Southern Oscillation. *Mon. Wea. Rev.*, **115**, 1606–1626.
566 Schaller, N., I. Mahlstein, J. Cermak, and R. Knutti, 2011: Analyzing precipitation projections:
567 A comparison of different approaches to climate model evaluation. *J. Geophys. Res.*, **116**,
568 D10118.
569 Smith, T. M., R. W. Reynolds, T. C. Peterson, and J. Lawrimore, 2008: Improvements to
570 NOAA's historical merged land-ocean surface temperature analysis (1880-2006). *J. Climate*,
571 **21,** 2283-2296.
572 Spencer, H., and J. M. Slingo, 2003: The simulation of peak and delayed ENSO
573 teleconnections. *J. Climate*, **16**, 1757–1774.
574 Straus, D. M., and J. Shukla, 1997: Variations of midlatitude transient dynamics associated
575 with ENSO. *J. Atmos. Sci.*, **54**, 777-790.
576 Su, H., J. D. Neelin, and J. E. Meyerson, 2003: Sensitivity of tropical tropospheric
577 temperature to sea surface temperature forcing. *J. Climate,* **16,** 1283–1301.
578 Sun, D.-Z., Y. Yu, and T. Zhang, 2009: Tropical water vapor and cloud feedbacks in climate
579 models: A further assessment using coupled simulations. *J. Climate*, **22**, 1287–1304.
580 Tebaldi, C., J. Arblaster, and R. Knutti, 2011: Mapping model agreement on future climate
581 projections. *Geophys. Res. Lett.*, **38**, L23701.
582 Trenberth, K. E., 2011: Changes in precipitation with climate change. *Clim. Res.*, **47,** 123-
583 138.
584 Trenberth, K.E., and D. P. Stepaniak, 2001: Indices of El Niño evolution. *J. Climate*, **14**, 1697–
585 1701.

586 Trenberth, K.E., G. W. Branstator, D. Karoly, A. Kumar, N.-C. Lau, and C. Ropelewski, 1998:
587 Progress during TOGA in understanding and modeling global teleconnections associated with
588 tropical sea surface temperatures. *J. Geophys. Res.,* **103,** 14 291–14 324.
589 Trenberth, K.E., and L. Smith, 2009: Variations in the three-dimensional structure of the
590 atmospheric circulation with different flavors of El Niño. *J. Climate*, **22**, 2978–2991.
591 Whitaker, J. S., and K. M. Weickmann, 2001: Subseasonal variations of tropical convection
592 and week-2 prediction of wintertime western North American rainfall. *J. Climate*, **14**, 3279–
593 3288
594 Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*.
595 Academic Press, 467 pp.
596 Xie, P., and P. A. Arkin, 1997: Global Precipitation: A 17-year monthly analysis based on
597 gauge observations, satellite estimates, and numerical model outputs. *Bull. Amer. Meteor.*
598 *Soc.,* **78,** 2539– 2558.
599 Xue, Y., T. M. Smith, and R. W. Reynolds, 2003: Interdecadal changes of 30-yr SST normals
600 during 1871-2000. *J. Climate*, **16,** 1601-1612.
601

602



603

604 Figure 1. DJF precipitation teleconnections for the years 1979-2005, as diagnosed through a
605 linear regression analysis of precipitation against the Nino3.4 index (units of mm day^{-1} C^{-1}).
606 (a) Observed teleconnections; (b) concatenated multi-model ensemble teleconnections
607 (CMME) for 15 CMIP5 AMIP models; (c) same as in (a), but with a two-tailed significance test
608 applied to the regression values, shown at 95% confidence (black outline) and 90% confidence
609 (lighter shading); (d) same as in (b) but shaded only where the regression is significant at or
610 above the 95% confidence level.

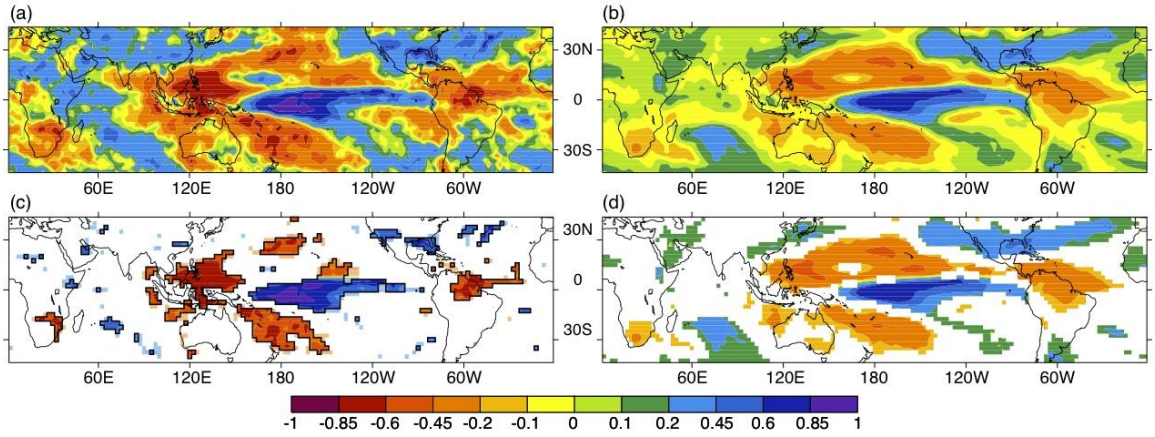


Figure 2. As in Fig. 1, but for a Spearman rank correlation analysis between precipitation and the Niño3.4 index. Note here that the color bar is in units of the Spearman rank correlation coefficient, with a minimum value of -1.0 and a maximum of +1.0. (a) and (b) show the teleconnection patterns from the rank correlation applied to the observations and CMME, respectively. Patterns plotted in (c) are as in (a) but shaded only where gridpoints pass the 95% confidence level (black outline) and the 90% confidence level (lighter shading) of a two-tailed statistical significance test for the rank correlation analysis. (d) The CMME teleconnections shaded for gridpoints that pass at the 95% significance level in the rank correlation analysis.

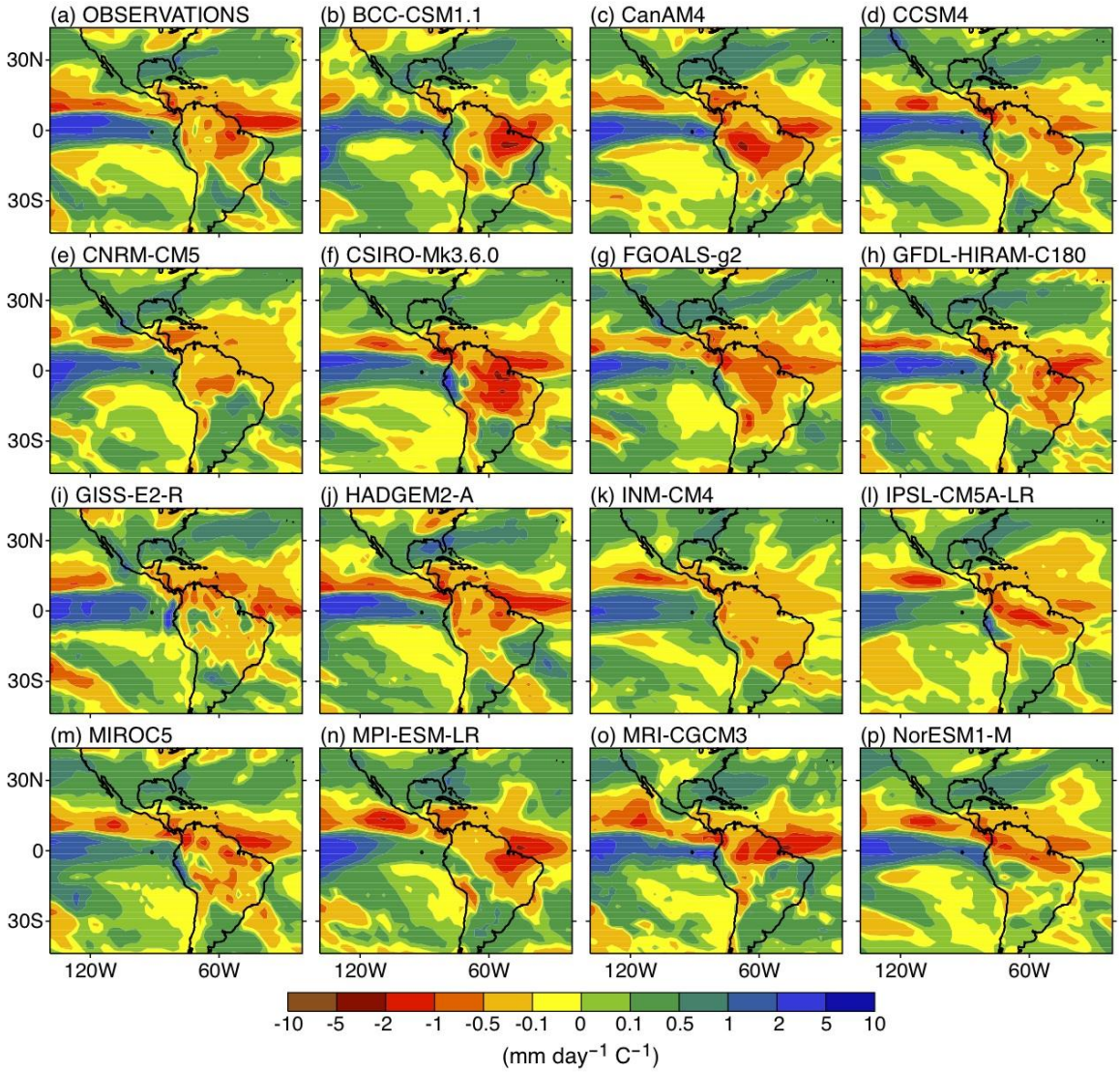
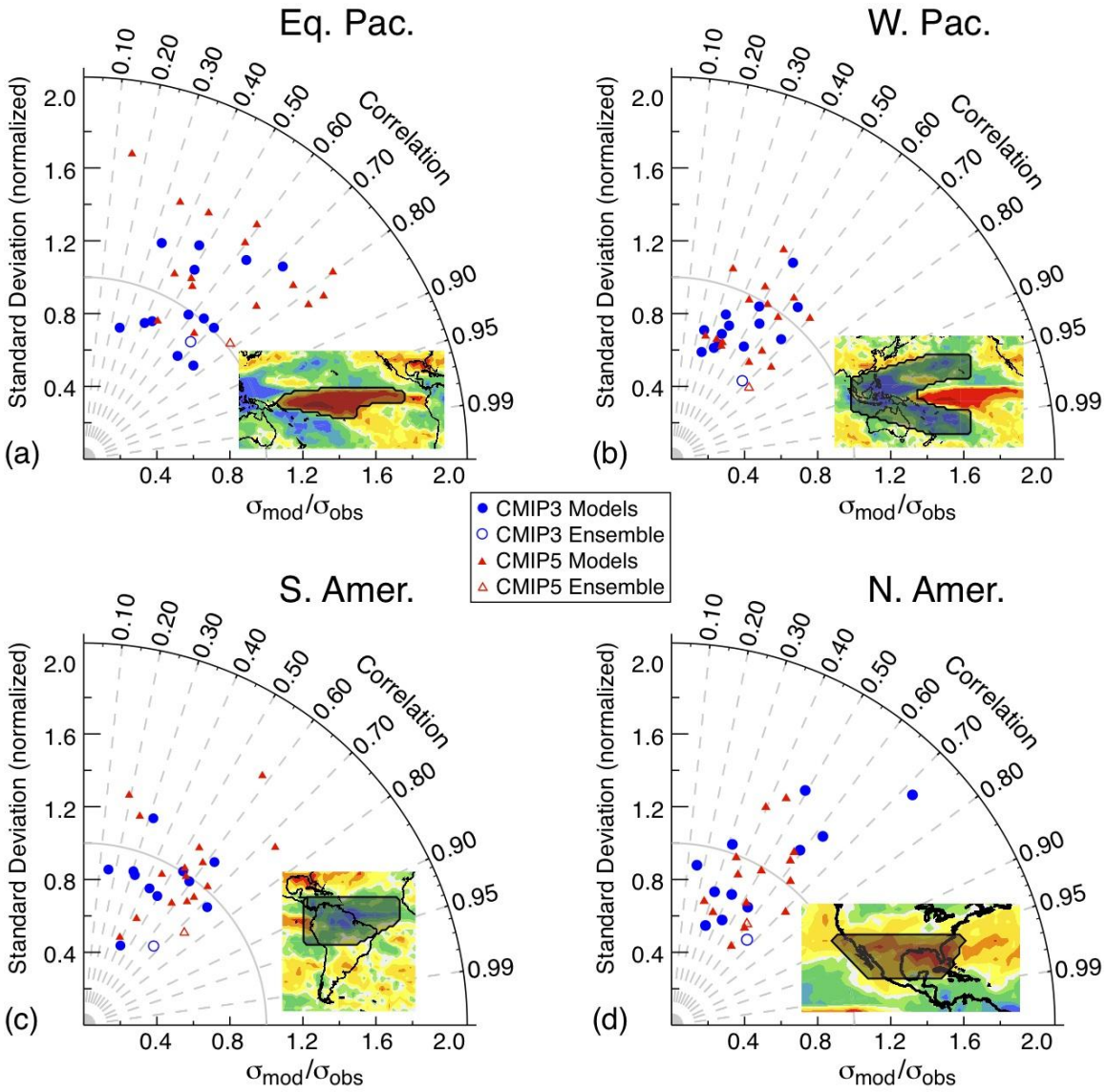


Figure 3. DJF precipitation teleconnections shown for (a) the observations (top left) and for (b)-(p) one run from each of 15 available CMIP5 AMIP models (listed alphabetically by model acronym). Teleconnections here are resolved via a linear regression analysis as in Fig. 1, with an analogous color bar that has units of mm day^{-1} C^{-1}. Patterns are plotted for the equatorial Americas to highlight regional (intermodel) disagreement among the ensemble members.



628

629 Figure 4. Taylor diagrams for the standardized amplitude and spatial correlation of
630 precipitation teleconnections in four selected regions, as indicated in the inset of each panel:
631 (a) the equatorial Pacific (central ENSO) region, (b) the "horseshoe" region in the western
632 equatorial Pacific; (c) an equatorial section of South America, and (d) a southern section of
633 North America. On the Taylor diagrams, angular axes show spatial correlations between
634 modeled and observed teleconnections; radial axes show spatial standard deviations of the
635 teleconnection signals in each area, normalized against that of the observations. Shaded red

636 triangles (15 total) and blue circles (11 total) denote each of the CMIP5 and CMIP3 AMIP
637 models, respectively. The unshaded red triangle is the CMIP5 MMEM; the unshaded blue circle
638 is the CMIP3 MMEM.
639

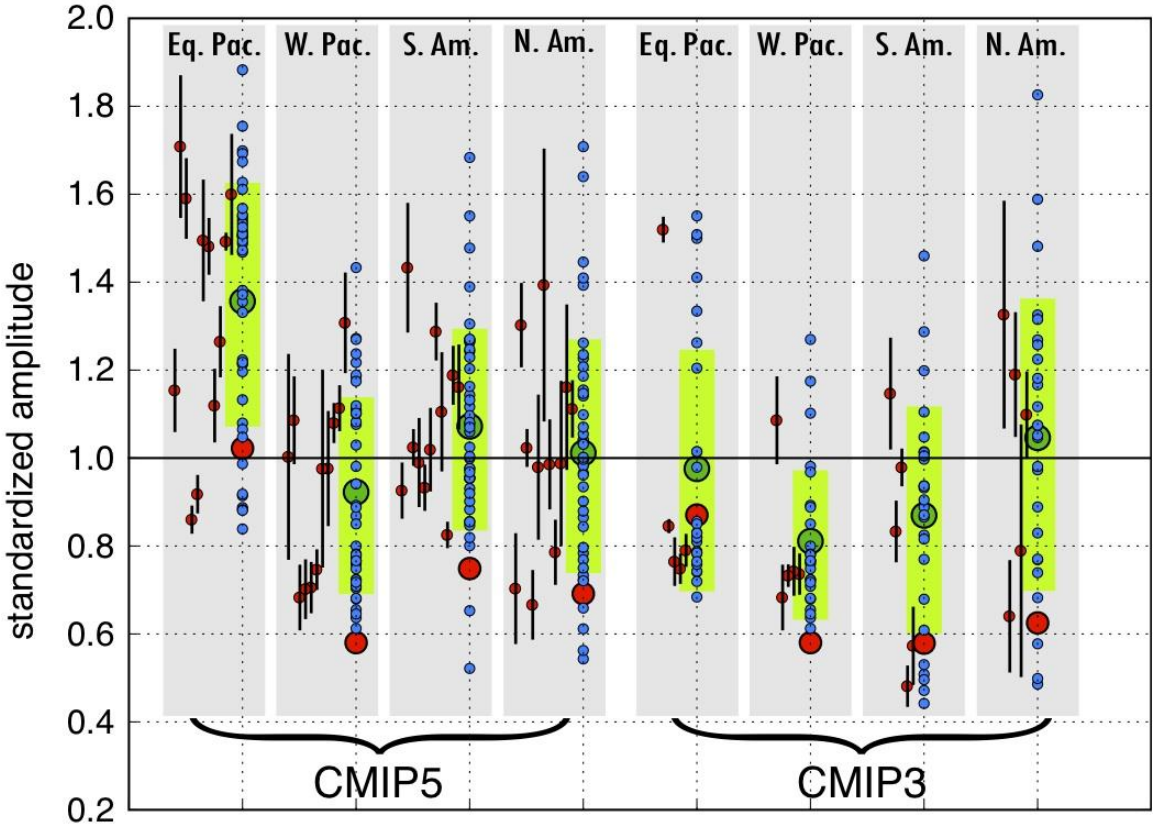


Figure 5. Standardized amplitude of precipitation teleconnections in each of the four regions identified in Fig. 4 (spatial standard deviation of each model's precipitation teleconnections divided by that of the observations). CMIP5 models are shown on the left, CMIP3 models on the right. Each blue dot represents a separate model run, and where multiple runs are available for a given model, a blue dot is plotted for each. Black bars represent the spread among the multiple runs for one model (where available), centered at that model's average amplitude among the multiple runs (±1 standard deviation). The green dots and green bars denote the average teleconnection amplitude (±1 standard deviation) for the entire ensemble, for each region. The red dot is the weighted MMEM.

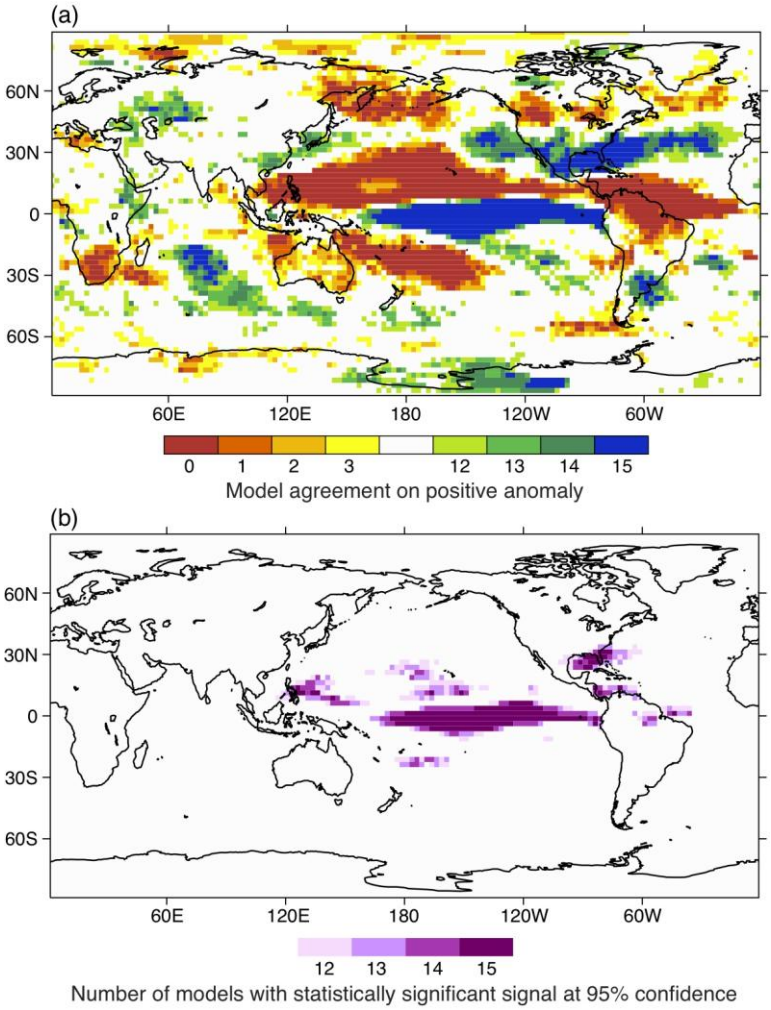


Figure 6. (a) Agreement on sign plot for 15 modeled teleconnection patterns (linear regression). Blue colors represent high agreement on a positive precipitation response during ENSO events; red colors represent high agreement on a negative precipitation response. Note that in an ensemble of 15 models, an agreement count of 12 implies that 80% of models agree on the sign of the precipitation teleconnection at that gridpoint. (b) Number of models that pass a two-tailed statistical significance test for the linear regression being significantly different from a slope of 0. Note again that 12 models implies 80% agreement in the ensemble.

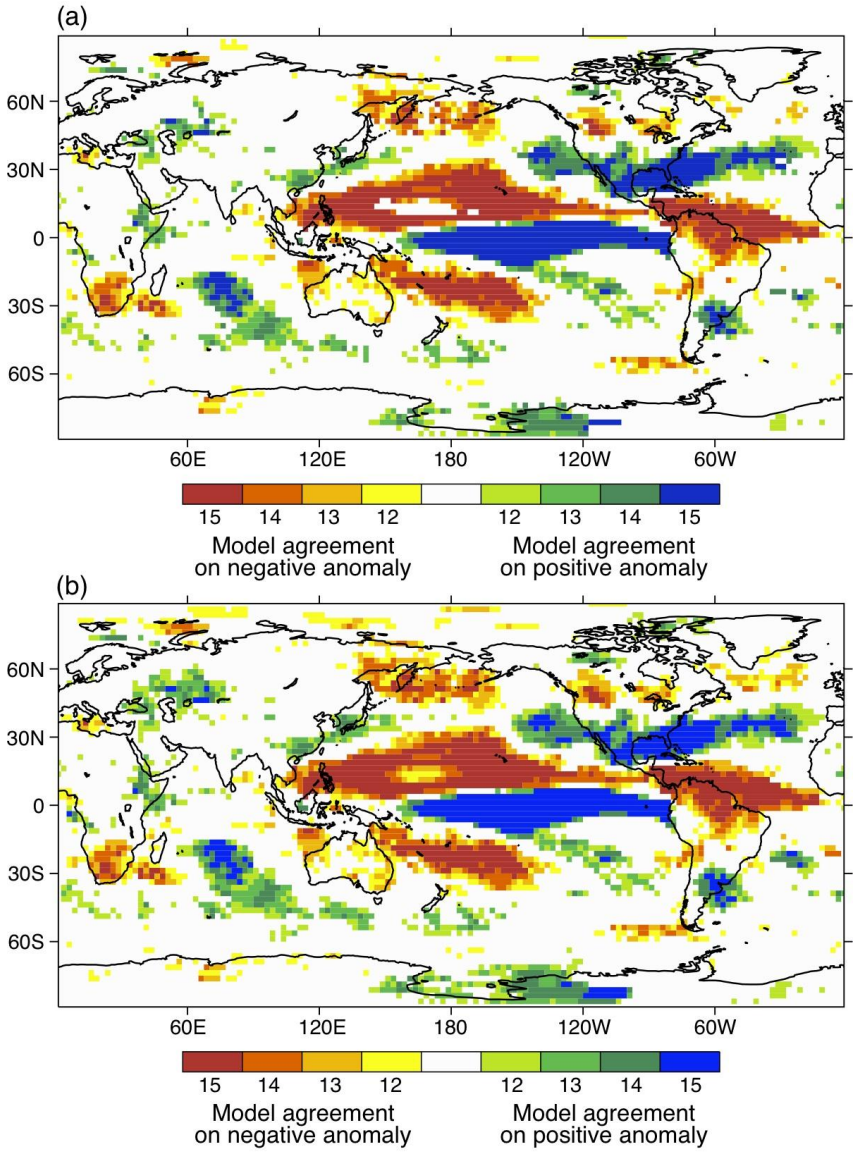


Figure 7. (a) Agreement on sign of precipitation teleconnections between each of 15 CMIP5 AMIP models and the observations. (b) Agreement on sign of precipitation teleconnections between the CMIP5 AMIP models and the MMEM, calculated by first subtracting the influence of each model from the MMEM when determining the agreement count. Both (a) and (b) use teleconnection patterns diagnosed via linear regression. Red areas denote models that agree with the observations or MMEM on a negative precipitation signal during ENSO events; blue areas imply agreement on a positive precipitation signal.