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 the Coupled Model Intercomparison Project Phase 5 (CMIP5), using the ensemble of runs driven by observed sea surface temperatures (SSTs). Simulated seasonal precipitation teleconnection patterns 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 substantial departures from observations, with spatial correlation coefficients typically less than 0.5. However, the atmospheric models do considerably better in other measures. First, the amplitude of the precipitation response (root mean square deviation over each region) is well estimated by the mean of the amplitudes from the individual models. This is in contrast with the amplitude of the multi-model ensemble mean, which is systematically smaller (by about 30-40%) in the selected teleconnection regions. Second, high intermodel agreement on teleconnection sign provides a good predictor for high model agreement with observed teleconnections. The ability of the model ensemble to yield amplitude and sign measures that agree with the observed signal for ENSO precipitation teleconnections lends supporting evidence for the use of corresponding measures 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 substantial remote effects on climate (Horel and Wallace 1981; Ropelewski and Halpert 1987; Trenberth et al. 1998; Wallace et al. 1998; Dai and Wigley 2000).

In recent decades, measurable progress has been made in simulating ENSO dynamics and associated teleconnections within atmosphere-ocean coupled general circulation models (CGCMs) (Neelin et al. 1992; Delecluse et al. 1998; Davey et al. 2001; Latif et al. 2001; AchutaRao and Sperber 2006; Randall et al. 2007). A number of studies use the fully-coupled GCMs to assess 20th century ENSO variability and teleconnections against observations (Doherty and Hulme 2002; Capotondi et al. 2006; Joseph and Nigam 2006; Cai et al. 2009). Others examine the evolution of ENSO and these teleconnections under climate change (Doherty and Hulme 2002; van Oldenborgh et al. 2005; Merryfield et al. 2006; Meehl and Teng 2007; Coelho and Goddard 2009). Problems persist in the ability of the models to accurately represent the tropical Pacific mean state, annual cycle, and ENSO's natural variability (Guilyardi et al. 2009a; Cai et al. 2012). Additional uncertainties remain in the role of the atmospheric components of CGCMs in setting the dynamics of ENSO and its teleconnections (Guilyardi et al. 2004, 2009b; Lloyd et al. 2009; Sun et al. 2009; Weare 2012), as well as 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 (Dai 2006). In the tropics, equatorial wave dynamics spread tropospheric temperature anomalies, which induce feedbacks with 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). Examination of ENSO precipitation teleconnections can therefore contribute to assessing the accuracy of models for these pathways, though note this is distinct from the discussion in the literature that the tropical Pacific may experience "El Niño-like" climate change.

One difficulty with assessing teleconnections from coupled models is that errors in the ENSO dynamics (e.g., 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 (Joseph and Nigam 2006; Coelho and Goddard 2009). To isolate the atmospheric portion of the teleconnection pathway, it is useful to employ atmospheric component simulations forced by observed SSTs, referred to as Atmospheric Model Intercomparison Project (AMIP) runs (Gates et al. 1998). In coupled model runs, errors in position or amplitude of the main equatorial ENSO SST signal can have a substantial impact on the teleconnections (Cai et al. 2009), and it is quite challenging for the models to accurately simulate regional signals in precipitation, even when observed SSTs are specified.

A few studies use AMIP runs to examine ENSO teleconnections. Risbey et al. (2011) do so for teleconnections over Australia, noting errors in the modeled amplitude and pattern coherence. Spencer and Slingo (2003) find that issues in the sensitivity of precipitation to tropical Pacific SSTs lead to errors in the Aleutian low despite otherwise accurate tropical ENSO teleconnections. Cash et al. (2005) compare two uncoupled, atmospheric GCMs forced with identically prescribed SSTs, finding noticeable variations between the two models in the response of extratropical 500mb height and regional precipitation. They force these models with climatological SST fields and SSTs representative of a response to a CMIP2 CO_2 doubling experiment. They find that precipitation difference patterns between the two models are similar for either case, implying that the differences between the atmospheric GCMs are “relatively insensitive” to the prescribed SST fields.

Because challenges persist in correctly simulating a precipitation teleconnection response, analysis of the CMIP5 AMIP 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-January-February (DJF) ENSO precipitation teleconnections during 1979-2005 in the CMIP5 models, and we compare these to observations and to the earlier CMIP3 AMIP ensemble.

In standard evaluation measures of teleconnection patterns and amplitude, substantial differences exist among models and when compared to the observations. In light of such differences, we turn to other measures in which the multi-model ensemble may contain useful information. These include amplitude measures, a comparison of individual models to the multi-model ensemble mean (MMEM), and measures of sign agreement.

In these alternative measures, the CMIP5 model ensemble does unexpectedly well compared to observations. The performance on sign agreement measures is decent enough to motivate questions regarding the optimal way to apply significance tests within multi-model ensembles. We provide some explanation in the discussion section, noting that even though a full answer may not yet exist, such alternative measures are relevant to the evaluation of precipitation change in global warming.

2. Data sets and analysis

To produce ENSO precipitation teleconnection patterns, we use modeled and observed monthly mean SST and precipitation data during the DJF months for the years 1979-2005. For SST observations, we use the Extended Reconstructed Sea Surface Temperature (ERSST.v3) data set (Xue et al. 2003; Smith et al. 2008); for monthly precipitation rate observations, we employ the Climate Prediction Center Merged Analysis of Precipitation (CMAP) archive (Xie and Arkin 1997).

For modeled teleconnections, we use monthly AMIP precipitation (pr) and surface temperature (ts) data from the CMIP5 and CMIP3 archives, as detailed in Table 1 (for more information on AMIP runs, see Gates et al. 1998 and references therein). All modeled precipitation data are regridded to a 2.5º-by-2.5º grid prior to calculating teleconnection patterns. This is the native grid of the CMAP precipitation data set, and we use it to facilitate direct comparison of modeled teleconnections to the observations.

Linear regression and Spearman's rank correlation are used to calculate DJF precipitation teleconnections for the selected time period. Linear regression is widely used for assessing the relationship between global precipitation and tropical Pacific SSTs, where precipitation at a gridpoint is regressed against a spatially averaged SST time series (here, the Niño 3.4 index, defined from 5ºS to 5ºN and 190ºE to 240º E; see Trenberth 1997 for information on El Niño indices). One caveat is that linear regression assumes the precipitation data follow a Gaussian distribution, whereas in reality they are zero-bounded and exhibit non-Gaussian behavior. Spearman's rank correlation — in which the rank of the data is used to compute the correlation coefficient (Wilks 1995) — does not make such assumptions, and therefore we use it to provide a check on the sensitivity of teleconnection patterns to the statistical methods employed (for examples of studies that employ rank correlation, see Whitaker and Weickmann 2001 or Münnich and Neelin 2005).

Appropriate *t*-tests are used in both the linear and rank methods to resolve gridpoints that meet or pass certain confidence levels (von Storch and Zwiers 1999). The majority of this paper will focus on a *t*-test applied to teleconnections resolved via linear regression. This *t*-test is based on calculating a two-tailed *p*-value where the null hypothesis is a linear regression slope of zero. Note that our use of the Niño 3.4 index yields "standard" teleconnection patterns, which provide a good basis for comparison of models to observations. We recognize, however, that there is interesting work addressing 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 (Ashok et al. 2007; Kao and Yu 2009; Trenberth and Smith 2009).

3. Evaluating modeled spatial patterns and amplitudes of precipitation teleconnections

a. Teleconnection patterns resolved via linear regression and rank correlation

Figs. 1 and 2 show observed and modeled precipitation teleconnections for the DJF season as estimated by linear regression and Spearman's rank correlation, respectively. We show both methods to check that teleconnected rainfall patterns are robust against the statistical assumptions going into the calculation (ENSO composites, not shown, yield similar results). Spearman's rank correlation is insensitive to extreme values and so can bring regions with different amplitudes of variance on to common footing. This statistical method also offers a significance test that does not assume Gaussian statistics. Linear regression, by contrast, is easier to interpret in terms of a change of the physical variables, which in this case is precipitation rate per degree change of SST in the Niño 3.4 region. Beyond this, comparing modeled to observed teleconnections raises some interesting questions about the restrictions of the statistical significance tests. The most pertinent question to arise is how best to use the collective information offered by a multi-model ensemble. Substantial intermodel variations also occur, and they are discussed in subsections 3b, 3c, and 3d. Other aspects of the restrictive nature of these significance tests will be discussed in section 4

Figs. 1b and 2b show teleconnection patterns obtained from the model ensemble. Note that there are several ways to obtain a regression representative of all data contained in the 15-model ensemble. The option we choose provides a straightforward test of statistical significance. Specifically, we perform the regression over all 15 models simultaneously; a straightforward way to interpret (and program) this is as a concatenated time series of the 15 available models, and so we will refer to this as the concatenated multi-model ensemble ("CMME"), when it is necessary to distinguish it.

The more classical approach of obtaining a single map of teleconnections for a 15-model ensemble is to calculate the teleconnections for each model individually and average the 15 patterns together afterward, discussed previously as the "MMEM." While this is more widely used, obtaining a test of statistical significance becomes complicated, as one cannot easily take an average of significance tests across 15 models. Thus in Figs. 1 and 2, the variant shown is the first one, though note that the MMEM (not shown) and CMME patterns are nearly identical, with a global spatial correlation coefficient greater than ρ=0.999. The high correlation between these two methods is to be expected if the variance in each model is similar and stably estimated. In the remainder of this paper, we will focus on the ensemble patterns seen in both Figs. 1b and 1d, and we will refer to them using "MMEM" and "CMME" interchangeably.

In Fig. 1, we show CMME linear regression DJF teleconnection patterns (1b and 1d) alongside observations (1a and 1c). The ensemble pattern in Fig. 1b reproduces a number of observed features. 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 is also 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.

Figs. 1c and 1d show the same data as 1a and 1b, but with a two-tailed *t*-test test applied to the regression at each gridpoint. One can see in Fig. 1d that the CMME regression passes a 95% confidence level criterion over fairly broad areas in each major teleconnection region, thanks to the large amount of information available in the 15-model ensemble. Each of the areas discussed above passes this significance test, as do some smaller regions, such as southeastern Africa. Fig. 1c displays observed teleconnections masked to show only grid points that pass the 90% and 95% confidence levels, indicating a relatively limited area over which the gridpoint-based regressions meet these confidence criteria. Specifically, linear regressions in Fig. 1 produce statistically significant teleconnections at 36.8% of gridpoints across the globe in the CMME. The average of the individual 15 models is 17.6% of gridpoints, while that of the observations is 16.1%. Thus the local significance tests for individual models, not shown, are qualitatively similar to the spatial extent of the observations in Fig. 1c.

Given that the CMME yields a statistically significant prediction for the sign of the signal over the main teleconnection regions, a one-tailed *t*-test (on the side predicted by the CMME) could be used on the observations, in which case the 90% confidence level of a two-tailed test would correspond to the 95% confidence level of a one-tailed test. However, when loosening the confidence level restriction from 95% to 90% for observed teleconnections, we only see a small increase in the spatial extent of regions that pass the significance test. In comparing Figs. 1c and 1d, one can see that the CMME is significant at 95% confidence over a broader area than the observations.

Fig. 2 displays the same information as in Fig. 1, but for Spearman's rank correlation applied to the CMME and observations. 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 under the rank correlation of the observations (Fig. 2c) is comparable to that produced for the linear regression of the observations (Fig. 1c), and likewise for the CMME. We henceforth focus on linear regression teleconnection patterns, due to the simpler interpretation of the amplitudes.

b. Regional model disagreement

Another point that can be made with Figs. 1 and 2 is the large-scale agreement between teleconnected precipitation patterns in the CMME and in the observations. For reasons discussed in section 5, this agreement is apparent over broader regions where the CMME passes the *t*-test at 95% confidence, not just in the narrower regions where observations pass the *t*-test at 95% confidence. However, regional disagreement between observations and the CMME pattern is also seen, especially in regions where the observations have intense precipitation. In addition, the CMME exhibits a general “smoothing” of teleconnection patterns.

These overly smoothed teleconnection patterns in the CMME can be understood when examining individual model patterns. Fig. 3 shows teleconnections for one run of each model in CMIP5, displayed for the equatorial Americas; substantial regional variability is easily seen. Qualitatively similar figures highlighting regional disagreement have been produced in other studies that use CGCMs to examine ENSO teleconnections and precipitation characteristics (e.g., Dai 2006, his Fig. 9). Difficulties in simulating these teleconnections in CGCMs persist in the AMIP models shown here: variations in the location of the strongest precipitation anomaly in Fig. 3 are common from model to model, even though these are the areas that most easily pass significance criteria on an individual model basis. Over the region where the CMME regression passes a *t*-test at the 95% level, however, 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 CMME; and yet some possibility that there is useful information about the teleconnection patterns in the 15-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 (detailed below) that show a robust ENSO response; each region displays a continuous teleconnection signal significant at the 95% confidence level in observations (see 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 region and is directly forced by the largest ENSO-related SST anomalies. We consider the other three regions the “teleconnection regions,” since to accurately simulate teleconnected rainfall in each, the models must capture the pathways leading to remote precipitation change. The Taylor diagrams in Fig. 4 show the spatial correlations between the observations and each model plotted against the spatial root mean square deviation of each model’s pattern (i.e., the standard deviation σ_{mod}) normalized by observations (σ_{obs}); we refer to this measure as the teleconnection amplitude. For models with multiple runs, correlations and amplitudes are calculated for each run first and then averaged among them; each individual model is given equal weight in the MMEM. Note we use the MMEM here, and not the CMME, though Taylor diagrams using the latter (not shown) are nearly identical. Additionally, some of the individual models have small negative correlations with observations in certain regions. These models are used in calculating the MMEM, though for diagrammatic simplicity the domain of the Taylor diagrams is not extended to display these points.

Fig. 4 allows easy comparison between CMIP3 and CMIP5 AMIP runs. There is little (if any) improvement from CMIP3 to CMIP5 in reproducing teleconnected rainfall patterns in these regions. Additionally, models exhibit generally low correlations (ranging from less than 0.2 to a few instances exceeding 0.7, with an average correlation coefficient of about 0.40) with observations. In every region, one can also see that the MMEM is typically more accurate than the majority of individual models in reproducing spatial patterns. However, the MMEM amplitude is substantially lower than that of the individual ensemble members, and it underestimates the observations in every region outside of the central equatorial Pacific. As a final point, we note that Taylor diagrams of the corresponding rank correlation method (not shown) also indicate consistent results.

d. Teleconnection amplitude in major impact regions

The varied agreement in amplitude measures from Fig. 4 suggests that it may be more reasonable to use amplitude information from individual ensemble members, rather than using that of the MMEM. To get a better sense of how teleconnection amplitude of individual models might be affected by internal variability within the models themselves, we take advantage of AMIP models with multiple realizations, and we assess the internal variability among these runs for each model. We then compare this to the amplitude range of the 15-model ensemble. Fig. 5 displays the radial axis from the Taylor diagrams discussed previously, but where multiple runs from each model are available, we plot them individually (43 total runs for 15 models in CMIP5; 26 total runs for 13 models in CMIP3; see Table 1).

The vertical extent of the black lines in Fig. 5, representing ± one standard deviation of the amplitudes for the runs of a given model, is a measure of internal variability for that model. The vertical extent of each green bar is ± one standard deviation of the MMEM amplitude, and it serves as a measure of intermodel variability. Notable points from this diagram include: (1) The MMEM systematically underestimates the spread and central tendency of intermodel variability, with a low bias of about 20-40% outside of the immediate ENSO region; (2) the regional disagreement among models owes itself partly to internal model variability, but intermodel variability contributes to the majority of the regional disagreement seen in Fig. 3; (3) individual models are overestimating the amplitude in the immediate ENSO region for CMIP5, even though their spread is more symmetric about the observations in remote regions;

(4) when comparing CMIP5 to CMIP3, CMIP5 shows no consistent improvement or change due to model development. Although the MMEM may fall closer to observed amplitudes in some regions for CMIP5, this comes at the expense of a tendency for individual models to overestimate rainfall teleconnections in the central ENSO region.

Fig. 5 suggests that serious errors can result from considering only information available in the MMEM. While its spatial patterns correlate better with observations than most individual models, the MMEM teleconnection amplitude is routinely 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 an argument for agreement plots of precipitation change in global warming scenarios

Agreement plots for the sign of precipitation change under global warming scenarios are commonly used in multi-model studies (e.g., Randall et al. 2007; Meehl et al. 2007), often as complementary information to the MMEM. Agreement-on-sign tests can be viewed as relatively weak statements regarding the precipitation change at individual gridpoints for the model ensemble, and it has been argued that sign agreement should be used in conjunction with requirements on individual models that gridpoints pass statistical significance tests for change in mean precipitation (e.g., Neelin et al. 2006; Tebaldi et al. 2011, hereafter N06 and T11, respectively).

Here we examine agreement-on-sign measures based on the ENSO precipitation regression patterns for each model. Because we can assess these against observations, we can use this to examine the procedure as a means of inferring its usefulness. If a procedure that identifies high model agreement at a gridpoint *also* correctly predicts the sign of the observations at that gridpoint, it can help build confidence in using corresponding procedures for the global warming case.

Fig. 6a shows the traditional agreement-on-sign plot for ENSO teleconnections in the CMIP5 AMIP ensemble. At each gridpoint, we count the number of models that agree on a positive (negative) DJF teleconnection signal for the linear regression over Niño 3.4, so that the plot shows the integer value of models which agree on a wet (dry) response during ENSO. The sign of the regression slope at each gridpoint is equivalent to the sign of the expected DJF precipitation response during an El Niño event. Areas with 12 or more models agreeing on sign are shaded based on a binomial test. Specifically, if we consider the null hypothesis that the value of an ENSO precipitation signal for a given point is equally likely to be positive or negative, i.e. drawn from a binomial distribution with a probability of p=0.5, then when 12 or more models agree on sign, the null hypothesis for this 50-50 probability can be rejected at a confidence level greater than 95% (for 15 models, the 95% confidence level falls between an agreement count of 10 and 11).

The gridpoints with high sign agreement that pass the binomial test at the 95% level in Fig. 6a cover a spatial region similar to the areas passing the two-tailed t-test applied to the CMME (Fig. 1d) at the 95% level. However, the areas of high sign agreement cover a much larger spatial region than those passing the t-test at the 95% level for individual model realizations, which are similar to the areas passing the t-test at this level for observations (see Fig. 1c and the discussion in section 3a).

This last point suggests two comparisons. First, we can contrast regions of high sign agreement identified by the binomial test with examples of criteria that have been considered in the global warming literature that combine t-tests on individual models with sign agreement criteria from the ensemble. Second, in this ENSO teleconnection testbed, we can evaluate the model ensemble's sign prediction against observations. These results are displayed in Figs. 6b and 6c. These panels display hatching according to the N06 or T11 criteria, respectively, overlaid on a plot that assesses the prediction of the model ensemble for the sign of the teleconnection signal; details of these criteria are outlined below.

To produce the cross-hatching in Fig. 6b, we follow the N06 procedure: (1) at each gridpoint, count the number of models in the ensemble that have a slope significantly different from zero at the 95% confidence interval; (2) cross-hatch grid points where greater than 50% of models are significant and also agree on the sign of the precipitation teleconnection. The N06 criteria impose a requirement that at least half of models both be significant and agree on sign.

To produce the cross-hatching in Fig. 6c, we follow the T11 procedure: (1) at each gridpoint, count the number of models with a teleconnection significant at the 95% confidence interval (as in N06); (2) for gridpoints where more than 50% of models show a significant rainfall response, cross-hatch if 80% or more of significant models agree on the sign of the response; (3) if fewer than 50% of models agree on the sign, shade the gridpoint black.

The underlying color shading in Figs. 6b and 6c is identical and evaluates the sign prediction of the AMIP CMME for the teleconnection signal, produced in the following way: (1) take the regions of high sign agreement passing the binomial test at the 95% significance level in Fig. 6a as a prediction of the sign of the observed teleconnection pattern and compare that to the observations at the same gridpoint; (2) if the observations and the model prediction agree on sign, shade blue (red) for a positive (negative) ENSO precipitation signal, representing a correct prediction by the intermodel agreement plot (Fig. 6a); (3) if the observations and the Fig. 6a disagree on the sign, shade the gridpoint purple to indicate an erroneous prediction;

(4) if the agreement on sign does not pass the binomial test criterion of Fig. 6a, no prediction is made and the gridpoint is left unshaded.

When examining Figs. 6b and 6c, the most important point is that the model ensemble prediction of sign does very well when assessed against observations. In major regions for which model agreement passes the binomial test at 95% confidence, almost the whole area yields the correct sign. The scattered, incorrect gridpoints tend to be either isolated or at the edges of correct regions, such that a scientific assessment of likely areas of increase or decrease based on the predicted areas (color shading in Figs. 6a and 6b) would be highly accurate. Potential physical mechanisms for the success of the sign prediction are discussed in the next section.

Another obvious point in Fig 6b and 6c is the similarity between the N06 and T11 approaches. In practice, the T11 test employed here is equivalent to the N06 test defined at a 40% threshold (80% x 50% = 40%). The one difference is that T11 further specify those grid points where more than 50% of models are significant but fewer than 80% agree on sign, which they classify as “no prediction.” This last T11 criterion may be useful in evaluating precipitation change under global warming, where at a given gridpoint, statistical significance of the precipitation change for individual models does not necessarily mean they will agree on sign.

In comparing the N06 and T11 procedures to the regions over which the models correctly predict sign of the observations, it is immediately apparent that the N06 and T11 tests are highly conservative. Though they do remove the modest fraction of points for which the sign would have been incorrectly predicted based on high agreement (passing the binomial test at the 95% level), they do so at the cost of excluding substantial regions that are correctly predicted. This is evident in Figs. 6b and 6c, where the hatched areas are restricted in spatial extent relative to the broader shaded regions.

To show the sign agreement of the model ensemble with observations in more detail, we display in Fig. 7a the number of individual ensemble members that agree on sign with observations for ENSO teleconnections. The same criterion for displaying high model agreement (12 or more models) is used as in Fig. 6a. Within this region, it may be seen that there are large portions in which the number of models agreeing on sign with observations is even higher, including substantial areas where 100% of models agree with the sign of the observations.

To obtain a counterpart of this plot from the model ensemble, Fig. 7b shows the number of models agreeing with the sign of the MMEM. Note that in producing this, we exclude each model's contribution to the MMEM when determining agreement, so as to avoid inflating the count. The similarities between Figs. 7a and 7b indicate that high sign agreement with the MMEM can serve as a predictor for sign agreement with the observations.

5. Discussion

As discussed in the previous section, Figs. 6 and 7 suggest that there are substantial regions where models from the CMIP5 AMIP ensemble are providing useful information on the sign of rainfall teleconnections, despite individual models and the observations failing to meet *t*-test criteria at the 95% level in parts of these regions. We argue below that this is a combined consequence of the larger size of the model ensemble relative to individual runs, the nature of the quantity being tested (the sign), and the models' skill in predicting the observed sign. Before addressing this, we consider the possibility that the broader region of skill at sign prediction in the ensemble (relative to individual model runs) could simply be an issue with applicability of the *t*-test due to the inherent non-Gaussianity of the rainfall distribution, even at seasonal timescales. This was addressed in Fig. 2 by repeating the teleconnection calculations using Spearman's rank correlation, which makes no assumptions of Gaussianity for the gridpoint rainfall distributions, and an accompanying statistical significance test. This yields results similar to those of the linear regression *t*-test.

We now consider an explanation based on the fact that the sign agreement both uses information from the full model ensemble and tests a different hypothesis than difference from zero. Because the collective 15-model ensemble contains a much larger set of realizations of internal variability, it is natural that regions of smaller signal should pass a given significance criteria in measures that use all 15 models. This is evident in comparing Fig. 6a to Fig. 1d, where areas of high sign agreement (passing the binomial test at the 95% level) tend to coincide with areas that pass a *t*-test on the CMME at 95% confidence. In both cases the broad regions of statistical significance come from using all 15 models.

Taking this into account, we consider the question of why the models agree so well with the observations on the sign of the teleconnection patterns, despite doing poorly at detailed spatial distribution. There are two aspects to this question: one statistical, and the other physical. The statistical aspect is that where the models exhibit sign agreement of 80%, the best estimate of the parameter p in the binomial distribution is 0.8. While it is beyond the scope of the paper to establish Bayesian posterior probability density functions or other measures of margin of error on the inferred p, the point needed to interpret the results here is straightforward: if the models are sufficiently good representations of observations such that the observed signal can be considered to be drawn from a binomial distribution with a similar value of p at each point, then one would expect the high level of agreement seen. Thus the 15-model ensemble shows success at predicting the sign of the observations in broader regions than those where teleconnection signals pass *t*-tests applied to individual models or observations. If we consider the fact that these broader regions are those that pass the 95% confidence level of the binomial test, this success of the ensemble at sign prediction is completely consistent with expectations and with the statement that the models are doing well at simulating the observed sign.

The ability of models to provide information beyond what a particular significance test may suggest is not a new concept in modeled precipitation studies. Risbey et al. (2011) resolve significant teleconnections in an AMIP model using a 30-year record and a two-tailed t-test. The authors note that the number of gridpoints passing a 95% significance criterion is much fewer than the same method applied to a century of historical data. As a result, they loosen their restriction to an 80% confidence interval, noting that the associated teleconnection patterns are similar for records of either length. Power et al. (2012) evaluate projected precipitation changes from the coupled CMIP3 model ensemble, and they demonstrate using the binomial distribution that model consensus on the sign of end-of-century rainfall anomalies is itself a strong argument for confidence in ensemble agreement patterns.

That the ensemble does, in fact, get broad areas of small amplitude change correct in our teleconnection analysis adds to the discussion in the literature that projected change is worth assessing even in regions that do not meet *t*-test criteria applied to individual runs (Tebaldi et al. 2011, Power et al. 2012) if these regions do meet significance tests applied to the ensemble. This is particularly relevant in global warming studies, where a modest regional precipitation anomaly in a MMEM could mean substantial changes in regional precipitation budgets.

An important physical question that arises from the present teleconnection results is: why does the 15-model ensemble perform better at predicting the sign of the observed signal (including in broad areas of modest precipitation amplitude response) and at yielding the amplitude of the observed response than the individual models do at reproducing detailed spatial patterns of observed teleconnections? The unimpressive spatial correlations (Fig. 4) are affected by poor individual model skill in positioning high amplitude signals. We suggest that this may be associated with the multiple physical processes operating in ENSO teleconnections. Specifically, there are atmospheric processes at work that will have smaller intermodel uncertainty and smaller internal variability but are widespread spatially. Examples for these processes include an increase in tropospheric temperature driving changes in radiative fluxes, as well as driving an increase in water vapor and a corresponding increase in the threshold for convection (the thermodynamic process sometimes referred to as the "rich-get-richer" mechanism; Chou and Neelin 2004; Held and Soden 2006; Trenberth 2011). At the same time, feedbacks associated with dynamical changes in moisture convergence can produce large excursions from expected values of precipitation, both in intermodel and temporal variability. The models contain reasonable approximations to each of these processes, but the location of strong precipitation changes can be highly sensitive to factors such as model convection parameterizations, including the threshold for convective onset (Kanamitsu et al. 2002; Neelin et al. 2010).

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 seasonal precipitation anomalies. We focus on standard teleconnection patterns associated with the ENSO Niño 3.4 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 three methods: linear regression, Spearman’s rank correlation, and compositing techniques (not shown), all with similar results. The rank correlation method provides an alternative significance test, which is useful in narrowing some of the questions that arise for regions of low amplitude signal. Teleconnection patterns defined with linear regression are useful for questions that involve the amplitude of the signal; as such, we focus on results from the linear regression.

How well the models perform at reproducing the observed teleconnection patterns (amplitudes and spatial patterns) depends strongly on the quantity for which they are assessed. In standard measures of spatial correlation, taken over the regions outlined above, 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 varying substantially from model to model and with respect to observations, yielding low spatial correlations between modeled and observed teleconnection patterns (average correlation coefficients on the order of 0.40 in the defined regions).

The MMEM performs marginally better than most individual models in spatial correlation measures, largely because the regions of strongest and varying change have been smoothed. However, the MMEM systematically underestimate amplitude measures of the regional precipitation response by 30-40%, typically falling more than one standard deviation below the central tendency of the 15-model ensemble. This underestimation is again associated with regional disagreement among ensemble members, a well-documented artifact in precipitation studies of GCM ensembles (e.g., N06; Räisänen 2007; Knutti et al. 2010; Neelin et al. 2010; Schaller et al. 2011). The average of individual CMIP5 AMIP amplitudes, by contrast, is an accurate predictor for the observations in all regions but the central ENSO region, where models overestimate the precipitation response. Sizeable internal variability of precipitation teleconnections is also shown to exist within each model, though it does not dominate the intermodel spread.

One thing underlined by the low spatial correlations in individual models is that even in AMIP experiments, where only the atmospheric components of CGCMs are being compared, simulation of ENSO teleconnections is fairly challenging for the models. While coupled models will have additional feedbacks, the AMIP experiments provide a first line of assessment. Furthermore, because we can compare AMIP simulations to observations, we can assess how the model simulations fare under other metrics commonly used in assessment of ensemble patterns and intermodel agreement

Sign agreement measures for 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 consensus on sign, passing a binomial test (to reject the null hypothesis of 50-50 probability of either sign) at the 95% level. These regions are more spatially extensive than the regions for which individual models (or observations) would pass a two-tailed *t*-test at the 95% (or even the 90%) level. Furthermore, the regions passing the binomial test correspond well to the set of points passing a *t*-test (at the 95% level) applied to the 15-model ensemble. Thus the larger region with high agreement on sign, relative to regions passing criteria (e.g., N06 or T11) that make use of *t*-tests on individual models, is simply the result of the sign agreement test making use of the 15-model ensemble.

For these teleconnection patterns, the sign prediction can be tested against observations. The models exhibit high sign agreement with observations over similarly broad regions, implying that high sign agreement within the model ensemble (gridpoints passing the binomial test at the 95% level) is a good predictor for sign agreement with observations. One can infer from this that the model ensemble is producing useful information regarding the teleconnected precipitation signal in regions that do not pass a *t*-test at the 95% level for individual models, provided they pass a significance test that makes use of information from the full ensemble.

The evaluation of the model simulations for ENSO teleconnections may be used, with due caution, to draw inferences for assessment of precipitation in global warming projections. Many of the physical processes leading to rainfall teleconnections are analogous to the global warming case. In particular, widespread tropospheric warming initiates tropical dynamics that cause similar global precipitation change in both teleconnections and global warming. In both cases, one can trace localized precipitation anomalies with high amplitude and sizeable intermodel spread back to tropical regions of strong convergence feedbacks and regions where large-scale wave dynamics interacts with mid-latitude storm tracks.

The unimpressive skill of models at capturing the precise regional distribution of large-amplitude rainfall teleconnections compared to observations is consistent with poor intermodel agreement on a precise pattern of precipitation change in global warming. However, the skill of individual models at reproducing the observed teleconnection signal amplitude (assessed from the mean of the individual model amplitudes, *not* the MMEM) suggests that corresponding measures for global warming precipitation change may be trustworthy. Furthermore, sign agreement plots for the AMIP ensemble prove skillful at predicting the sign of observed teleconnections. While agreement plots for end-of-century precipitation change obviously have different spatial patterns than the signals considered here, the fact that sign agreement plots are skillfull at predicting spatially extensive ENSO remote precipitation impacts — which are challenging simulation targets that share physical pathways with global warming precipitation signals — provides a supporting argument in favor of using sign agreement plots in global warming studies to make predictions of change from an ensemble of models.

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. Münnich 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 provided coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. Finally, we thank J. Meyerson for her significant help in data analysis and plotting.

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.

Cai, W., M. Lengaigne, S. Borlace, M. Collins, T. Cowan, M. J. McPhaden, A. Timmermann, S. Power, J. Brown, C. Menkes, A. Ngari, E. M. Vincent, and M. J. Widlansky, 2012: More extreme swings of the South Pacific convergence zone due to greenhouse warming. *Nature*, **488**, 365-369.

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.

Cash, B. A., E. K. Schneider, and L. Bengtsson, 2005: Origin of regional climate differences: role of boundary conditions and model formulation in two GCMs. *Climate Dyn.*, **25**, 709-723.

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 L. 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.

Dai, A. and T. M. L. Wigley, 2000: Global patterns of ENSO-induced Precipitation. *Geophys. Res. Lett.*, **27**, 1283-1286.

Dai, A., 2006: Precipitation Characteristics in Eighteen Couple Climate Models. *J. Climate*, **19**, 4605-4630.

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.

DeWeaver, E., and S. Nigam, 2004: On the forcing of ENSO teleconnections by anomalous heating and cooling. *J. Climate*,**17,** 3225-3235.

Doherty, R. and M. Hulme, 2002: The relationship between the SOI and extended tropical precipitation in simulations of future climate change. *Geophys. Res. Lett.*, **29**, 1475.

Gates, W. L., and Coauthors, 1998: An overview of the results of the Atmospheric Model Intercomparison Project (AMIP I). *Bull. Amer. Meteor. Soc.*, **73**, 1962–1970.

Guilyardi, E., and Coauthors, 2004: Representing El Niño in coupled ocean–atmosphere GCMs: The dominant role of the atmospheric component. *J. Climate*, **17**, 4623–4629.

Guilyardi, E., A. Wittenberg, A. Fedorov, M. Collins, C. Wang, A. Capotondi, G. van Oldenborgh, and T. Stockdale, 2009a: Understanding El Niño in ocean–atmosphere general circulation models: Progress and challenges. *Bull. Amer. Meteor. Soc.*, **90**, 325–340.

Guilyardi, E., P. Braconnot, F.-F. Jin, S. T. Kim, M. Kolasinski, T. Li, and I. Musat, 2009b: Atmosphere feedbacks during ENSO in a coupled GCM with a modified atmospheric convection scheme. *J. Climate*, **22**, 5698–5718.

Held, I. M., S. W. Lyons, and S. Nigam, 1989: Transients and the extratropical response to El Niño. *J. Atmos. Sci.*, **46**, 163-174.

Held, I. M., and B. J. Soden, 2006: Robust responses of the hydrological cycle to global warming. *J. Climate*, **19**, 5686– 5699.

Horel, J. D., and J. M. Wallace, 1981: Planetary-scale atmospheric phenomena associated with the Southern Oscillation. *Mon. Wea. Rev.*, **109,** 813–829.

Joseph, R., and S. Nigam, 2006: ENSO evolution and teleconnections in IPCC's Twentieth-Century climate simulations: Realistic representation? *J. Climate*, **19**, 4360-4377.

Kao, H.-Y., and J.-Y. Yu, 2009: Contrasting eastern-Pacific and central-Pacific types of ENSO. *J. Climate*, **22**, 615–632.

Kanamitsu, M., and Coauthors, 2002: NCEP dynamical seasonal forecast system 2000. *Bull. Amer. Meteor. Soc*, **83,** 1019–1037.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2010: Challenges in combining projections from multiple climate models. *J. Climate*, **23**, 2739–2758.

Latif, M., and Coauthors, 2001: ENSIP: The El Niño Simulation Intercomparison Project. *Climate Dyn.*, **18**, 255–272.

Lloyd, J., E. Guilyardi, H. Weller, and J. Slingo, 2009: The role of atmosphere feedbacks during ENSO in the CMIP3 models. *Atmos. Sci. Lett.*, **10**, 170–176.

Meehl, G. A., and Coauthors, 2007: Global climate projections. *Climate Change 2007: The Physical Science Basis*, S. Solomon, et al., Eds., Cambridge University Press, 747–845.

Meehl, G. A., and H. Teng, 2007: Multi-model changes in El Niño teleconnections over North America in a future warmer climate. *Climate Dyn.*, **29**, 779-790.

Merryfield, W., 2006: Changes to ENSO under CO_2 doubling in a multimodel ensemble. *J. Climate*, **19**, 4009–4027.

Münnich, M., and J. D. Neelin, 2005: Seasonal influence of ENSO on the Atlantic ITCZ and equatorial South America. *Geophys. Res. Lett.*, **32**, L21709.

Neelin, J. D., and Coauthors, 1992: Tropical air–sea interaction in general circulation models. *Climate Dyn.*, **7,** 73–104.

Neelin, J. D., C. Chou, and H. Su, 2003: Tropical drought regions in global warming and El Niño teleconnections. *Geophys. Res. Lett.*, **30,** 2275.

Neelin, J. D., M. Münnich, H. Su, J. E. Meyerson, and C. E. Holloway, 2006: Tropical drying trends in global warming models and observations. *Proc. Natl. Acad. Sci.*, **103,** 6110-6115.

Neelin, J. D., A. Bracco, H. Luo, J. C. McWilliams, and J. E. Meyerson, 2010: Considerations for parameter optimization and sensitivity in climate models. *Proc. Natl. Acad. Sci.*, **107**, 21 349–21 354.

Oldenborgh, G.J. van, and T. Stockdale, 2009: Understanding El Niño in Ocean-Atmosphere General Circulation Models: Progress and challenges. *Bull. Amer. Met. Soc.*, **90**, 325-340.

Power, S. B., F. Delage, R. Colman, and A. Moise, 2012: Consensus on Twenty-First-Century Rainfall Projections in Climate Models More Widespread than Previously Thought. *J. Climate*, **25**, 3792-3809.

Räisänen, J., 2007: How reliable are climate models? *Tellus*, **59A**, 2–29.

Randall, D. A., and Coauthors, 2007: Climate models and their evaluation. *Climate Change 2007: The Physical Science Basis*, S. Solomon et al., Eds., Cambridge University Press, 589–662.

Risbey, J. S., P. C. McIntosh, M. J. Pook, H. A. Rashid, and A. C. Hirst, 2011: Evaluation of rainfall drivers and teleconnections in an ACCESS AMIP run. *Aust. Meteor. Oceanogr. J.*, **61,** 91-105.

Ropelewski, C. F., and M. S. Halpert, 1987: Global and regional scale precipitation patterns associated with the El Niño/ Southern Oscillation. *Mon. Wea. Rev.*, **115**, 1606–1626.

Schaller, N., I. Mahlstein, J. Cermak, and R. Knutti, 2011: Analyzing precipitation projections: A comparison of different approaches to climate model evaluation. *J. Geophys. Res.*, **116**, D10118.

Smith, T. M., R. W. Reynolds, T. C. Peterson, and J. Lawrimore, 2008: Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880-2006). *J. Climate*, **21,** 2283-2296.

Spencer, H., and J. M. Slingo, 2003: The simulation of peak and delayed ENSO teleconnections. *J. Climate*, **16**, 1757–1774.

Straus, D. M., and J. Shukla, 1997: Variations of midlatitude transient dynamics associated with ENSO. *J. Atmos. Sci.*, **54**, 777-790.

Su, H., J. D. Neelin, and J. E. Meyerson, 2003: Sensitivity of tropical tropospheric temperature to sea surface temperature forcing. *J. Climate,* **16,** 1283–1301.

Sun, D.-Z., Y. Yu, and T. Zhang, 2009: Tropical water vapor and cloud feedbacks in climate models: A further assessment using coupled simulations. *J. Climate*, **22**, 1287–1304.

Tebaldi, C., J. Arblaster, and R. Knutti, 2011: Mapping model agreement on future climate projections. *Geophys. Res. Lett.*, **38**, L23701.

Trenberth, K. E., 1997: The Definition of El Niño. *Bull. Amer. Met. Soc.*, **78**, 2771-2777.

Trenberth, K. E., G. W. Branstator, D. Karoly, A. Kumar, N.-C. Lau, and C. Ropelewski, 1998: Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *J. Geophys. Res.,* **103,** 14 291–14 324.

Trenberth, K. E., and L. Smith, 2009: Variations in the three-dimensional structure of the atmospheric circulation with different flavors of El Niño. *J. Climate*, **22**, 2978–2991.

Trenberth, K. E., 2011: Changes in precipitation with climate change. *Clim. Res.*, **47,** 123-

138.

von Storch, H., and F. W. Zwiers, 1999: Statistical analysis in climate research. *Cambridge University Press*, Cambridge.

Weare, B. C., 2012: El Niño teleconnections in CMIP5 models. *Climate Dyn.*, published online, doi:10.1007/s00382-012-1537-3.

Whitaker, J. S., and K. M. Weickmann, 2001: Subseasonal variations of tropical convection and week-2 prediction of wintertime western North American rainfall. *J. Climate*, **14**, 3279–3288

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 467 pp.

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

Xue, Y., T. M. Smith, and R. W. Reynolds, 2003: Interdecadal changes of 30-yr SST normals during 1871-2000. *J. Climate*, **16,** 1601-1612.

Table 1. CMIP5 and CMIP3 modeling centers and models used, and the number of AMIP runs available at the time of our analysis. Data are available for download at http://pcmdi3.llnl.gov.

Modeling center or group (institute ID)	**CMIP5 AMIP model**	**runs**	**CMIP3 AMIP model**	**runs**
Beijing Climate Center, China Meteorological Administration (BCC)	BCC-CSM1.1	3		
Canadian Centre for Climate Modelling and Analysis (CCCMA)	CanAM4	4		
National Center for Environmental Research (NCAR)	CCSM4	1	CCSM3	1
			PCM	1
Centro Euro-Mediterraneo per I Cambiamente Climatici (CMCC)	CNRM-CM5	1	CNRM-CM3	1
Commonwealth Scientific and Industrial Research Organization in collaboration with Queensland Climate Change Centre of Excellence (CSIRO-QCCCE)	CSIRO-Mk3.6.0	1		
LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences (LASG-CESS)	FGOALS-s2	3	FGOALS-g1.0	3
NOAA Geophysical Fluid Dynamics Laboratory (NOAA GFDL)	GFDL-HIRAM-C180	3	GFDL-CM2.1	1
NASA Goddard Institute for Space Studies (NASA GISS)	GISS-E2-R	5	GISS-ER	4
Met Office Hadley Centre (MOHC)	HadGEM2-A	5	UKMO-HadGEM1	1
Institute for Numerical Mathematics (INM)	INM-CM4	1	INM-CM3.0	1
Institut Pierre-Simon Laplace (IPSL)	IPSL-CM5A-LR	5	IPSL-CM4	5
Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology (MIROC)	MIROC5	2	MIROC3.2(hires)	1
			MIROC3.2(medres)	3
Max Planck Institute for Meteorology (MPI-M)	MPI-ESM-LR	3	ECHAM5/MPI-OM	3
Meteorological Research Institute (MRI)	MRI-CGCM3	3	MRI-CGCM2.3.2	1
Norwegian Climate Centre (NCC)	NorESM1-M	3		

Figures and captions

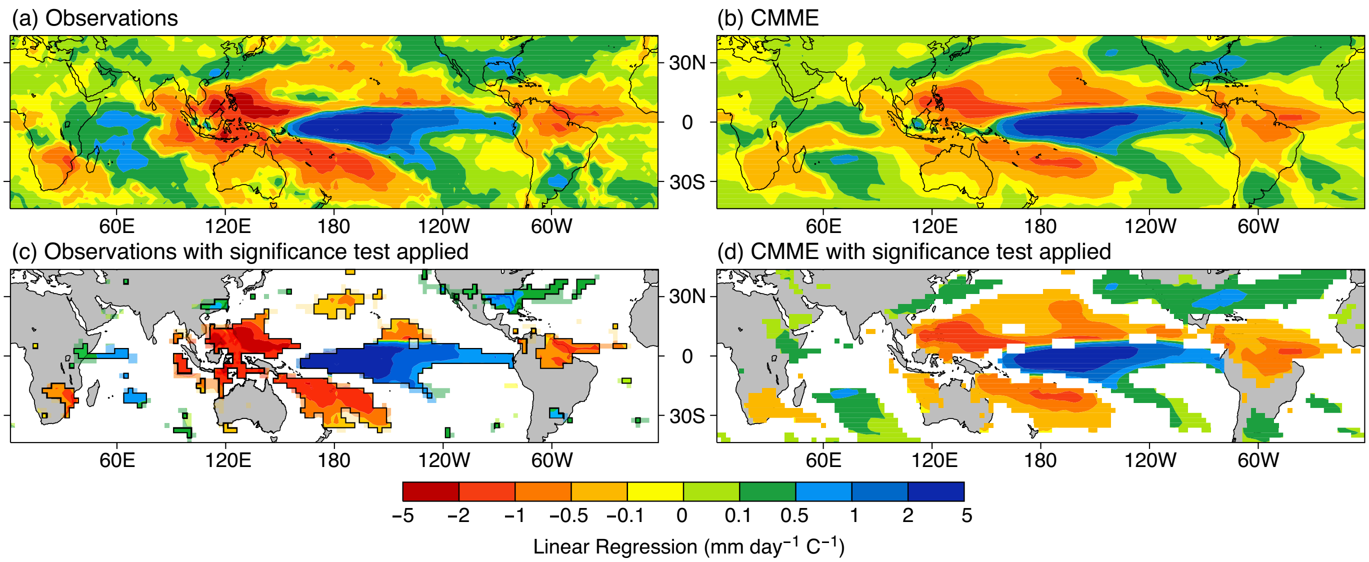

Figure 1. DJF precipitation teleconnections for the years 1979-2005, as diagnosed through a linear regression analysis of precipitation against the Niño 3.4 index (units of mm day^{-1} C^{-1}). (a) Observed teleconnections. (b) Concatenated multi-model ensemble (CMME) teleconnections for the CMIP5 AMIP 15-model ensemble. (c) Same as in (a), but with a two-tailed *t*-test applied to the regression values and shaded at 95% confidence (black outline) and 90% confidence (lighter shading). (d) Same as in (b) but shaded only where a *t*-test yields gridpoints significant at or above the 95% confidence level.

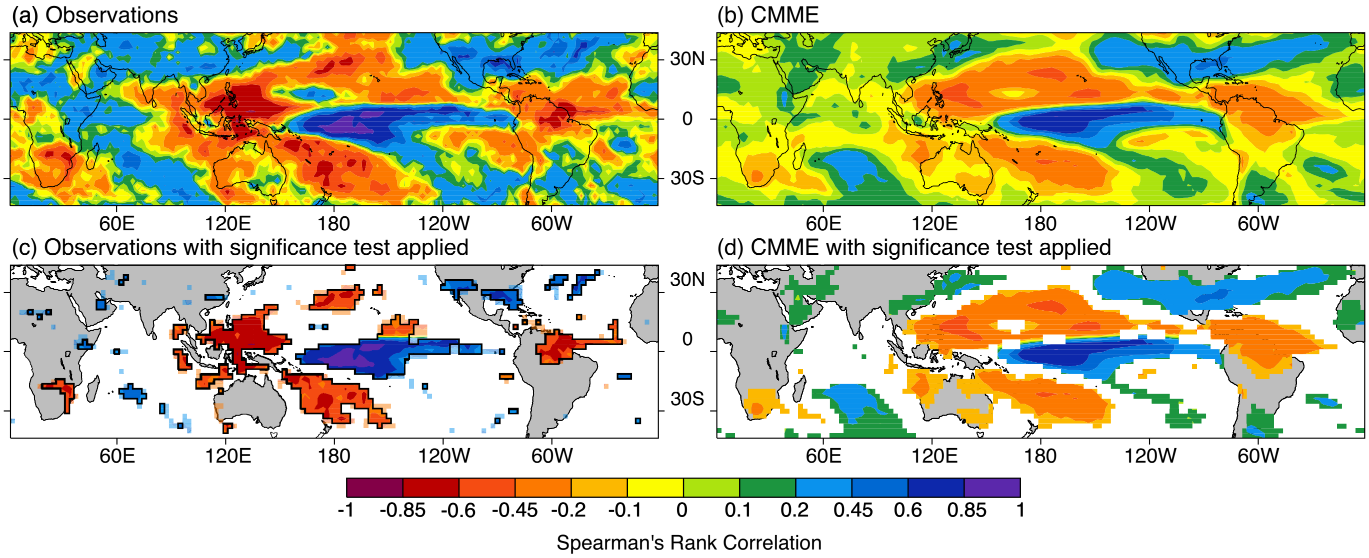

Figure 2. As in Fig. 1, but for Spearman's rank correlation analysis between gridpoint precipitation and the Niño 3.4 index. Note here that the color bar is unitless and corresponds to the Spearman's rank correlation coefficient, with a minimum of -1.0 and a maximum of +1.0. Panels (a) and (b) show the teleconnection patterns from the rank correlation applied to the observations and CMME, respectively. (c) Same as in (a) but shaded only where gridpoints pass the 95% confidence level (black outline) and the 90% confidence level (lighter shading) of a 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.

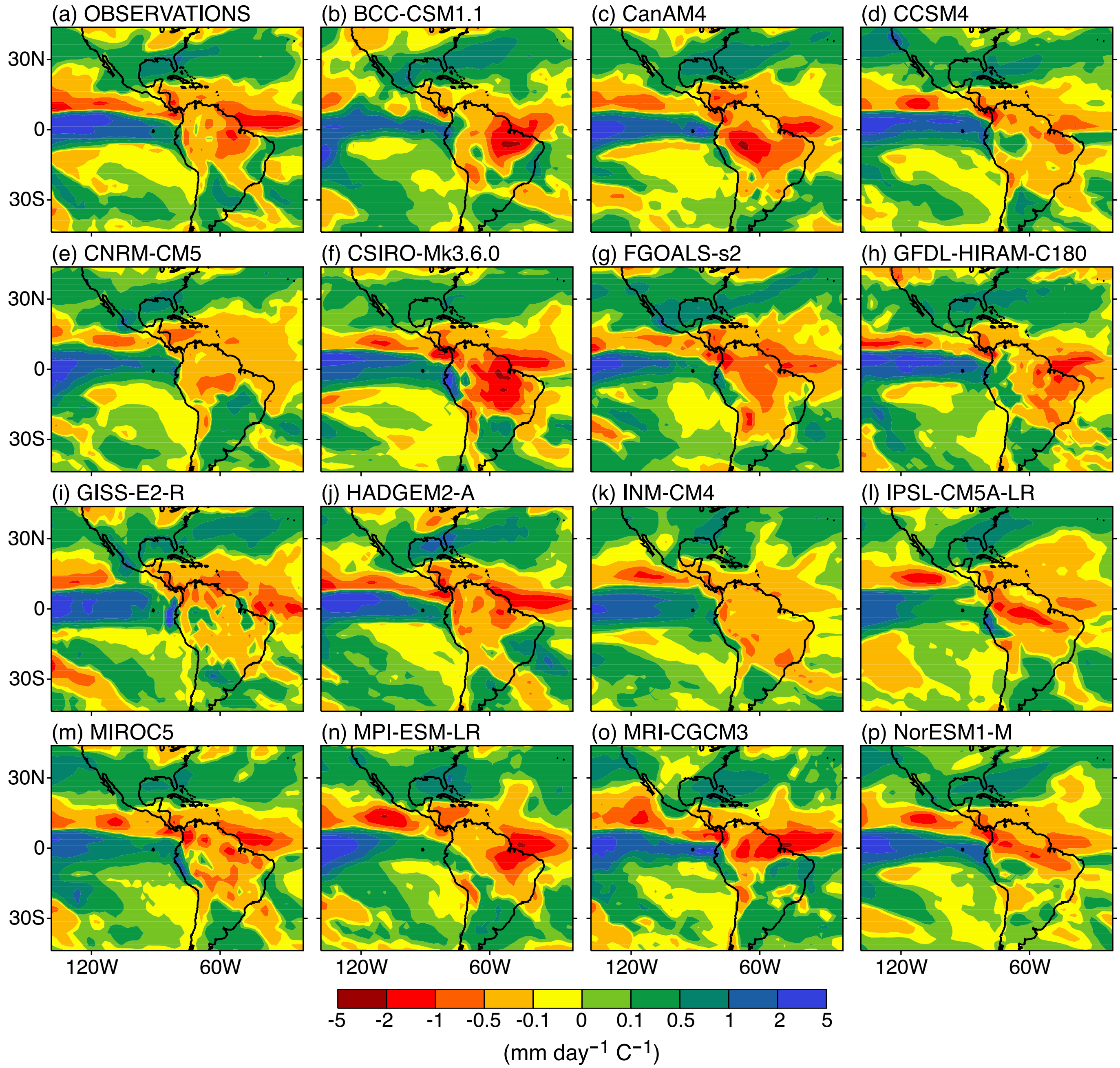

Figure 3. DJF precipitation teleconnections shown for (a) the observations, top left, and (b)-(p) one run from each of 15 available CMIP5 AMIP models (listed alphabetically by model acronym). Teleconnections here are resolved via the linear regression analysis as in Fig. 1, with an identical color bar 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.

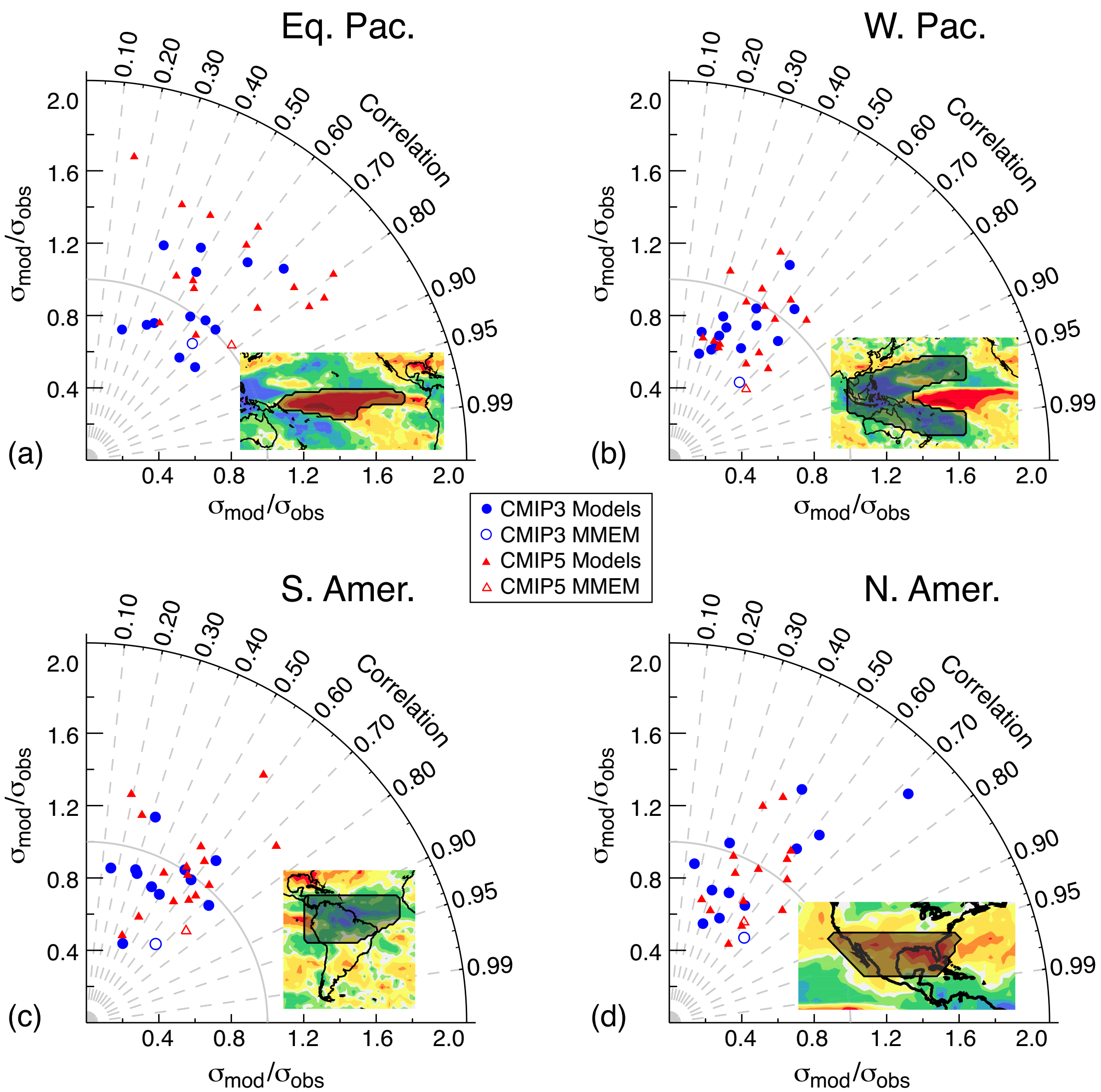

Figure 4. Taylor diagrams for the standardized amplitude and spatial correlation of precipitation teleconnections in four selected regions, as indicated in the inset of each panel: (a) the equatorial Pacific (central ENSO) region, (b) the “horseshoe” region in the western equatorial Pacific, (c) an equatorial section of South America, and (d) a southern section of North America. On the Taylor diagrams, angular axes show spatial correlations between modeled and observed teleconnections; radial axes show spatial standard deviation (root mean square deviation) of the teleconnection signals in each area, normalized against that of the observations. Shaded red triangles (15 total) and blue circles (11 total) denote each of the CMIP5 and CMIP3 AMIP models, respectively. The unshaded red triangle is the CMIP5 MMEM; the unshaded blue circle is the CMIP3 MMEM. Note that some models have negative correlations with the observed teleconnections in a few regions, and while we include them in the MMEM, we do not plot them individually in the diagrams.

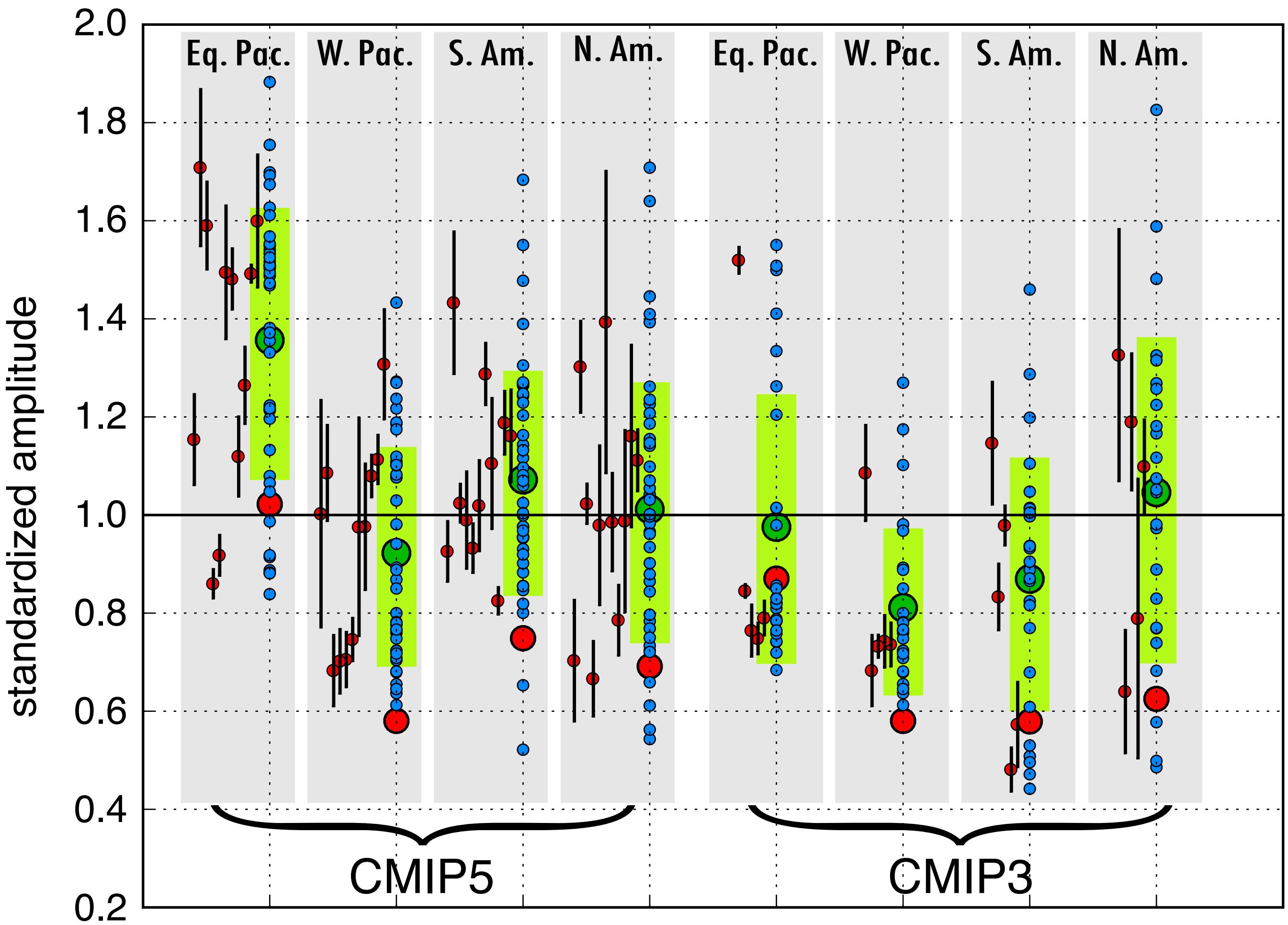

Figure 5. Standardized amplitude of precipitation teleconnections in each of the four regions identified in Fig. 4. The calculation for this amplitude is discussed in the caption of Fig. 4 and in the text. CMIP5 models (15 models, 43 runs) are shown on the left; CMIP3 models (13 models, 26 runs) on the right; see Table 1 for models used. 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 (when available), centered at that model's average amplitude among the multiple runs (±1 standard deviation of the amplitude measure). The green dots and green bars denote the average teleconnection amplitude and its spread (±1 standard deviation) for the entire ensemble, in each region. The red dot is the MMEM including all available models and runs, weighted so that each separate model contributes equally.

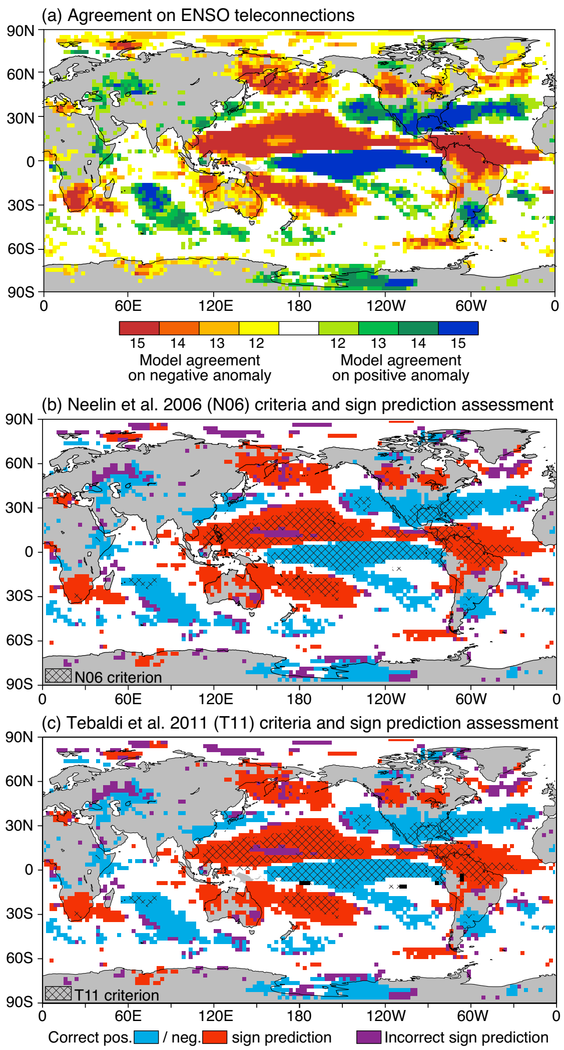

Figure 6. (a) Agreement on a positive teleconnection signal (linear regression) within the 15-model ensemble. Blue (red) colors represent high agreement on a positive (negative) precipitation response during ENSO events. Note that in an ensemble of 15 models, an agreement count of 12 implies that 80% of models agree on the sign of the precipitation teleconnection at that gridpoint, which is the area passing a binomial test at greater than the 95% confidence level (discussed in text). (b) Neelin et al. 2006 (N06) significance criteria (cross-hatching) overlaid on the sign prediction of the 15-model ensemble (colored shading). (c) Tebaldi et a. 2011 (T11) significance criteria (cross-hatching) overlaid on the sign prediction of the ensemble, as in (b). Details of the N06 and T11 cross-hatching criteria and sign prediction shading are outlined in the text. The cross-hatching is shown as an overlay in (b) and (c) to highlight the restrictive nature of the N06 and T11 criteria relative to the more extensive spatial coverage over which the 15-model ensemble passes the binomial test at the 95% level *and* exhibits an accurate prediction of the observed teleconnection signals.

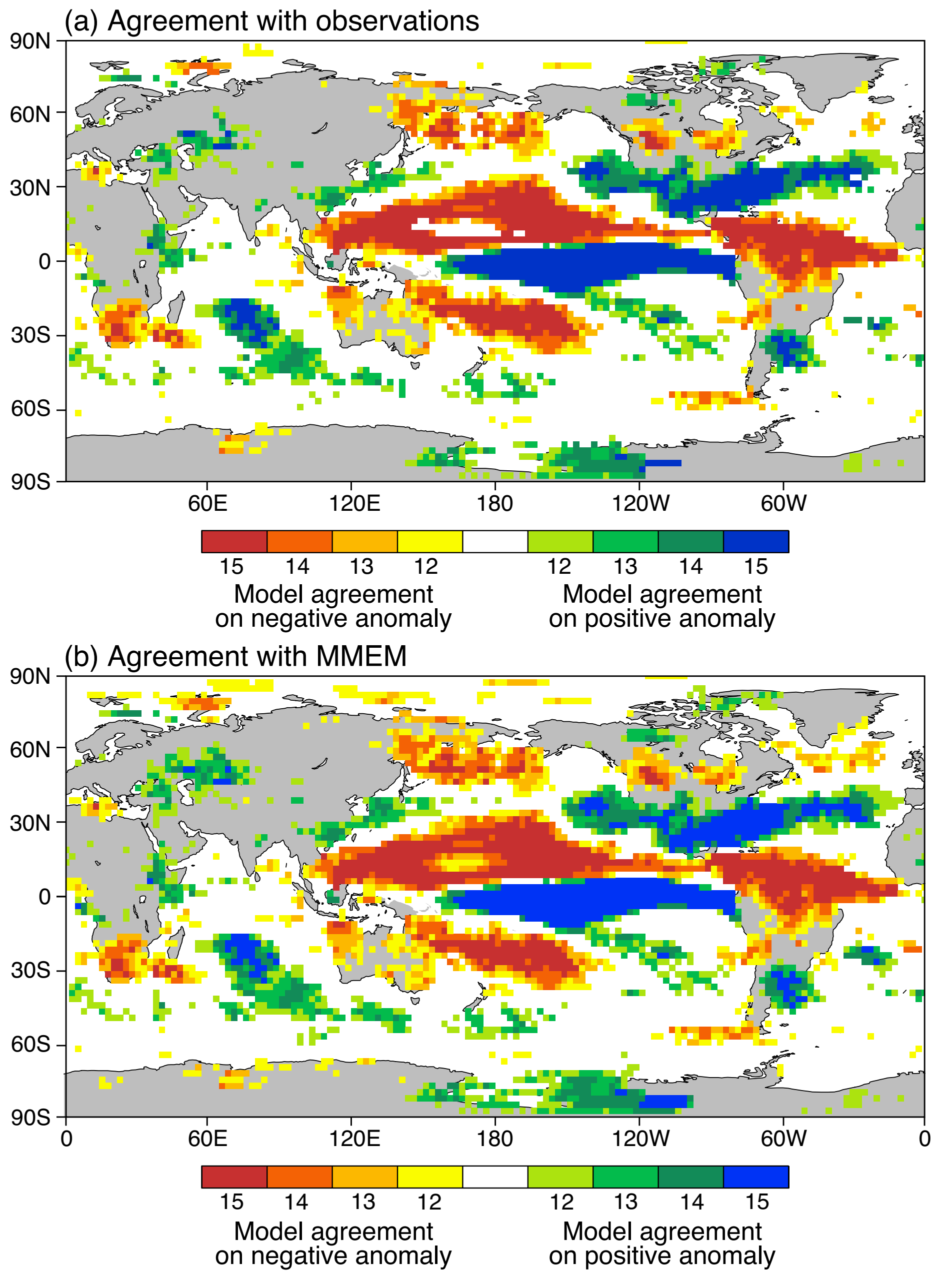

Figure 7. (a) Sign agreement of precipitation teleconnections between each of 15 CMIP5 AMIP models and the observations. (b) Sign agreement of precipitation teleconnections between the CMIP5 AMIP models and the MMEM, calculated using one run from each model. For (b), each model is individually removed from the MMEM before determining its sign agreement. Both (a) and (b) use Niño 3.4 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.