Geophysical Research Letters

RESEARCH LETTER
10.1029/2019GL086765

Relating CMIP5 Model Biases to Seasonal Forecast Skill in the Tropical Pacific

Hui Ding[1,2], Matthew Newman[1,2], Michael A. Alexander[2], and Andrew T. Wittenberg[3]

[1]CIRES, University of Colorado Boulder, Boulder, CO, USA, [2]NOAA Earth Systems Research Laboratory, Boulder, CO, USA, [3]NOAA Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA

Key Points:
- Skill of monthly ENSO hindcasts (1961–2015) of 28 CMIP5 models, determined from model-analogs, is related to tropical Pacific model biases
- Models with poorer precipitation hindcast skill generally have larger cold and dry biases in the equatorial cold tongue region
- Model-analog ENSO hindcast skill may be a useful metric for CMIP6 models, given its correspondence with CGCM mean and variance errors

Supporting Information:
- Supporting Information S1

Correspondence to:
H. Ding,
hui.ding@noaa.gov

Citation:
Ding, H., Newman, M., Alexander, M. A., & Wittenberg, A. T. (2020). Relating CMIP5 model biases to seasonal forecast skill in the tropical Pacific. *Geophysical Research Letters*, *47*, e2019GL086765. https://doi.org/10.1029/2019GL086765

Received 27 DEC 2019
Accepted 12 FEB 2020
Accepted article online 14 FEB 2020

©2020. American Geophysical Union. All Rights Reserved. This article has been contributed to by US Government employees and their work is in the public domain in the USA.

Abstract We examine links between tropical Pacific mean state biases and El Niño/Southern Oscillation forecast skill, using model-analog hindcasts of sea surface temperature (SST; 1961–2015) and precipitation (1979–2015) at leads of 0–12 months, generated by 28 different models from the fifth phase of the Coupled Model Intercomparison Project (CMIP5). Model-analog forecast skill has been demonstrated to match or even exceed traditional assimilation-initialized forecast skill in a given model. Models with the most realistic mean states and interannual variability for SST, precipitation, and 10-m zonal winds in the equatorial Pacific also generate the most skillful precipitation forecasts in the central equatorial Pacific and the best SST forecasts at 6-month or longer leads. These results show direct links between model climatological biases and seasonal forecast errors, demonstrating that model-analog hindcast skill—that is, how well a model can capture the observed evolution of tropical Pacific anomalies—is an informative El Niño/Southern Oscillation metric for climate simulations.

Plain Language Summary Research institutions around the world have developed computer models to conduct simulations of Earth's climate to support the Intergovernmental Panel on Climate Change Assessment. Therefore, it is necessary to assess aspects of these models using historical observations. Here, we develop a new model metric that involves making forecasts over several seasons, using existing long simulations of Earth's preindustrial climate conducted by many research institutions and provided to the community for use in climate studies. Within each simulation, we find the best matches, or "model-analogs," to historical observations of tropical Indo-Pacific Ocean surface temperatures and sea level. The evolutions of these analogs over the next several months within the simulation are then used as forecasts. Long-term accuracy of the forecasts indicates how well a model reproduces the time evolution of the observations. Precipitation forecast accuracy (regardless of lead) and sea surface temperature forecast skill (at 6-month or longer leads) emerge as key identifiers of the models that best simulate tropical Indo-Pacific climate.

1. Introduction

The El Niño/Southern Oscillation (ENSO) is the dominant mode of interannual climate variability in the tropical Pacific (e.g., Philander, 1990; Timmermann et al., 2018). ENSO events are characterized by deviations of 1–4 °C from the sea surface temperature (SST) mean state in the central and eastern equatorial Pacific. These SST anomalies (SSTAs) exert a significant influence on convective activity and the atmospheric circulation, affecting weather and climate worldwide (Alexander et al., 2002; Trenberth et al., 1998). Therefore, skillful forecasts of ENSO events, typically made using coupled general circulation models (CGCMs), are of great socioeconomic interest.

Many CGCMs simulate an excessive westward extension of the cold tongue into the tropical Pacific warm pool. The resulting equatorial cold SST bias (e.g., Bellenger et al., 2014; Davey et al., 2002; Lin, 2007) has been associated with poor representation of other tropical Pacific mean state variables, including excessive surface easterly winds and deficient rainfall, and these errors can be mutually intensified via ocean-atmosphere interactions including those associated with ENSO (e.g., Li & Xie, 2014; Lin, 2007; Wittenberg et al., 2018; Zheng et al., 2011), causing climate variability errors (e.g., Chen et al., 2017; Choi et al., 2015; Graham et al., 2017; He et al., 2018).

Some studies have found direct links between these CGCM biases and ENSO seasonal forecast skill, while others have not (e.g., Gualdi et al., 2005; Lee et al., 2010; DelSole & Shukla, 2010; Magnusson et al., 2013;

Smith et al., 2013; Vecchi et al., 2014; Ding et al., 2015; Kim et al., 2017; Scaife et al., 2019; Dippe et al., 2019). For example, Manganello and Huang (2009) found that a CGCM warm SST bias off the South American coast degrades ENSO forecast skill. Lee et al. (2010) noted pattern correlations in the global tropics between mean climate and 1-month lead forecast skill. Magnusson et al. (2013) obtained slightly improved accuracy for 6-month lead SST forecasts in the eastern equatorial Pacific, by correcting mean state biases using heat flux adjustments. In contrast, in a recent study of 14 CGCMs used for seasonal prediction, Scaife et al. (2019) found no intermodel relationship between precipitation forecast skill and local precipitation mean bias in the tropical Indo-Pacific. Richter et al. (2018) concluded that reducing SST mean biases did not necessarily improve an atmospheric model's ability to simulate the zonal wind and precipitation responses to prescribed observed SSTAs. Thus, there is no consensus on connections between model mean biases and seasonal forecast skill, including how the cold tongue bias may impact ENSO forecast skill. Such connections may have been obscured in past studies for at least two reasons: (1) They may require large sets of retrospective forecasts (or "hindcasts") of ENSO from a large number of models to identify, especially in the presence of decadal variations in forecast skill, and (2) initialized forecasts are subject to model initialization shock and drift, which also affect the bias-skill relationship.

To address these challenges, we here use a "model-analog" method developed by Ding et al. (2018; hereafter D18), where seasonal forecasts are initialized by searching a CGCM's preexisting long control simulation for anomalous model states that are similar to the initial anomalous observed state. (This is in contrast to traditional forecast initialization methods, which use full-field assimilation of observations into the forecast model.) The subsequent evolution of these states in the preexisting control simulation then provides forecasts at various lead times. Tropical Pacific SST and precipitation forecast skill from the North American Multi-Model Ensemble (NMME; Kirtman et al., 2014), in which models are initialized using the traditional approach, is generally matched or exceeded by applying the model-analog method to the corresponding long control runs from each NMME model (D18). Ding et al. (2019; hereafter D19) subsequently showed that model-analogs could be used to estimate forecast skill for the years 1961–2015, from 28 different models drawn from the fifth phase of the Coupled Model Intercomparison Project (CMIP5). Many of the CMIP5 models had skill comparable to the NMME hindcasts. That is, D19 used model-analogs to nearly double both the length of the available hindcast period, and the number of seasonal forecast models.

Here, we leverage D19's large hindcast data set to assess how forecast skill relates to model biases. First, we use the model-analogs to determine each model's tropical Pacific SST and precipitation hindcast skill at leads of 0–12 months and use this skill as a metric for ENSO, measuring how well each CMIP5 model captures the spatiotemporal evolution of observed ENSO events. We then examine the relationship between the cold tongue bias and the model-analog seasonal hindcast skill for SST and precipitation, within a multimodel context of 28 CMIP5 CGCMs.

2. Model-Analog Forecasts

The model-analog technique (D18; D19) is simply analog forecasting (e.g., Barnett & Preisendorfer, 1978; Lorenz, 1969; Van den Dool, 1994), except that the closest matches to the observed initial conditions are drawn from a long CGCM control simulation, typically several hundred years in length, rather than from the much shorter observational record. To make ENSO forecasts, the K anomalous model states that best match (i.e., have the smallest root-mean-square [RMS] distance to) the initial observed SST and sea surface height (SSH) anomalies within the tropical Indo-Pacific (30°E to 70°W, 30°S to 30°N), are chosen as the initial ensemble. The subsequent evolution of these model-analogs then provides the forecast ensemble. This method initializes the model on its own attractor and therefore avoids any initialization shock. The model-analog technique can be used to make forecasts of not only tropical Indo-Pacific SSH and SST but also any other model quantity. For further details, including sensitivity to the analog-selection method, size of the model-analog ensemble, and length of the CGCM simulation, along with the skill comparison to traditionally initialized hindcasts, see D18 and D19.

As in D19, initial observed states are obtained from monthly mean SST and SSH anomalies during 1961–2015, where SST is from the Hadley Sea Ice and Sea Surface Temperature v1.1 data set (Rayner et al., 2003) and SSH is from the European Centre for Medium-Range Weather Forecasts Ocean Reanalysis System 4 data set (Balmaseda et al., 2013). These were also used for hindcast verification, along with

precipitation from the Global Precipitation Climatology Project Version 2.3 (Adler et al., 2003). The "library" data sets from which the model-analogs are drawn consist of monthly mean anomalies from preindustrial control simulations of 28 different CMIP5 models (Table S1 in the supporting information; Taylor et al., 2012). All model and observed data are interpolated onto a common 2° longitude by 2° latitude grid, prior to analysis. For each simulation, monthly anomalies are calculated by subtracting each model's own monthly climatology determined using the full data set; that is, the forecasts have their mean biases removed by construction.

We use the tropical Indo-Pacific model-analog hindcasts from D19 for both SST (1961–2015) and precipitation (1979–2015), at leads of 0–12 months. The contribution of the external radiatively forced trend, determined by D19 from the ensemble mean of all the CMIP5 model ensembles, is removed; including this trend did not improve ENSO forecasts, nor is there a detectable secular trend in ENSO forecast skill (D19). An ensemble of 10 analogs ($K = 10$) was drawn from each simulation, with little change in skill for $K = 15$. All ensemble members were equally weighted.

Hindcast skill was evaluated using two deterministic skill measures that relate the ensemble mean forecast anomaly to the observed (verification) anomaly: (1) anomaly correlation (AC) and (2) standardized RMS error (SRMSE), where the RMS error is expressed as a fraction of the observed climatological standard deviation.

3. Results

3.1. Model-Analog Forecast Skill as a Model-Evaluation Metric

Figures 1a and 1b show the 6-month lead SST and precipitation hindcast skill for the grand 28-model ensemble mean. The hindcasts clearly have skill for central and eastern equatorial Pacific SSTs, where the correlation reaches 0.7. The grand ensemble mean also displays tropical Pacific precipitation skill, with correlations of about 0.5–0.6 in the central and eastern equatorial Pacific (Figure 1b), despite precipitation not being used to select model-analogs. Individual CMIP5 models also generate hindcasts with skill in the central and eastern equatorial Pacific (Figures S1 and S2), depending on model and region. These skill maps generally resemble those determined from operational models, such as the NMME (Kirtman et al., 2014; D18), although some CMIP5 models exhibit less skill (D19).

To develop a uniform ENSO skill metric, we evaluate CMIP5 SST and precipitation hindcast skill within the Niño3.4 region (5°S to 5°N, 170–120°W) for forecast leads of 0–12 months. Niño3.4 SST is a commonly used index to measure ENSO (e.g., Trenberth & Stepaniak, 2001), and a majority of the models display maximum forecast skill within the Niño3.4 region (Figures S1 and S2). Following Newman and Sardeshmukh (2017), we display skill for a particular model at a specific lead time by using the forecast amplitude (normalized by the observed climatological standard deviation) and the arccosine of the AC between forecasted and observed time series as the radius and azimuth, respectively, of a point on a Taylor diagram (Taylor, 2001). The distance from this point to the reference point (REF) is then, by construction, the SRMSE of the forecast (Newman & Sardeshmukh, 2017; Taylor, 2001). This is repeated for each lead time to yield 13 points (Month 0 to Month 12) for each model, connected with line segments to produce a curve indicating the evolution of model skill with forecast lead time. Note that "Month 0" skill is not a forecast measure but instead measures the match between the model-analog ensemble mean initial state and observations; that is, how well observed anomalies can be reconstructed within the model space ("reconstruction skill"). This permits a comprehensive evaluation of the spatiotemporal structure of model variability relative to observations.

The resulting Taylor diagrams are shown in Figure 1c for SST skill and Figure 1d for precipitation skill. For each curve, skill generally recedes monotonically away from the perfect skill reference point (REF) with increasing forecast lead, with (roughly) counterclockwise evolution (Newman & Sardeshmukh, 2017). Each curve may therefore be viewed as a "skill trajectory," a metric capturing the evolution of model hindcast skill as a function of forecast lead. At long leads the curves spiral toward the origin, as each model's analog ensemble disperses toward the climatological probability density function and the ensemble mean anomaly tends toward 0. Although there is a wide spread of SST skill trajectories, they all have similar shapes and are distributed about the intermodel ensemble mean skill (heavy black line). In contrast, the

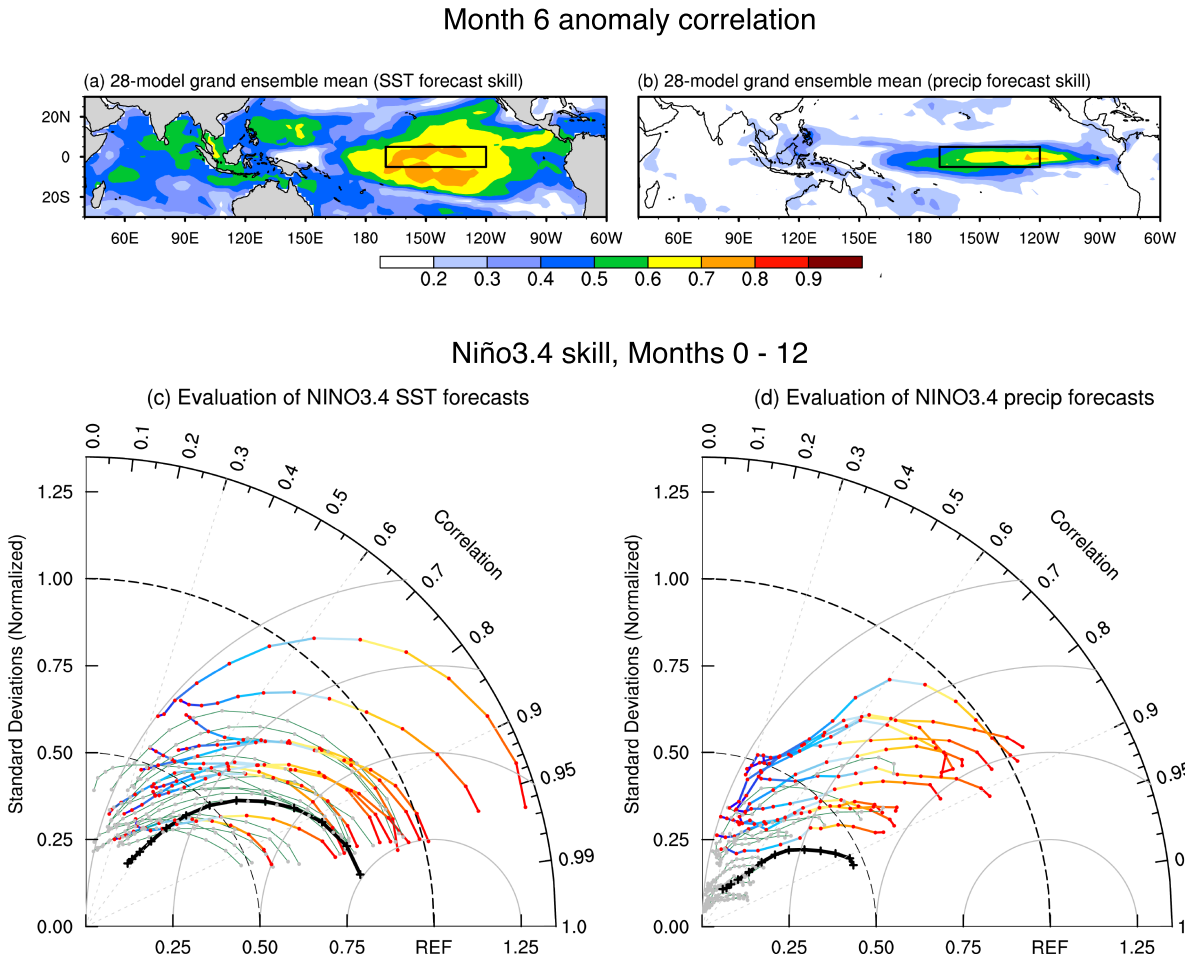

Figure 1. (a) SST and (b) precipitation model-analog hindcast anomaly correlation skill for observed variations at 6-month lead, calculated from the 28-model grand ensemble means. Black boxes indicate the Niño3.4 region. (c and d) Taylor diagrams of model-analog forecast skill at leads of 0–12 months for (c) Niño3.4 SST and (d) Niño3.4 precipitation. Each curve represents a "skill trajectory" from one CMIP5 model, with one dot per forecast lead month. In these diagrams, the distance of a point from the origin (indicated by dashed gray semicircles) is the forecast standard deviation, and the angular coordinate represents anomaly correlation. The radial distance from the reference point (REF, the verification field) (indicated by thin solid gray semicircles) represents the RMS forecast error, which is mathematically related to the other two quantities. Both the forecast amplitude and error are normalized by the observed standard deviation. Colored curves/red dots indicate the twelve models with the best initial representation of Niño3.4 precipitation in (c) and (d), respectively; remaining models are indicated with thin gray lines.

precipitation skill trajectories are clustered into two groups (one representing generally good skill and the other representing very poor skill), with the ensemble mean skill trajectory lying between them. Interestingly, there is no obvious correspondence between each model's SST and precipitation skill trajectory. That is, knowing that a model is in either the high (rainbow colored) or low (gray) precipitation skill group does not tell us what its SST skill trajectory looks like, as can be seen by comparing the trajectories of these groups in the two panels.

3.2. Relating Model-Analog Forecast Skill to Model Biases in Mean States and Variability

We now use our ENSO skill metric to compare each model's hindcast error to its time mean bias. In Figure 2 we plot each model's model-analog forecast error (SRMSE, the distance to the REF point in Figure 1) for either Niño3.4 precipitation (Figures 2a and 2b) or SST (Figures 2c and 2d), against the corresponding pre-industrial control run's climatological cold tongue bias, measured by the Niño3.4 SST bias. The SST bias is strongly related to both the reconstruction error and 6-month lead hindcast error of precipitation anomalies, with a correlation of about −0.7 (Figures 2a and 2b), significant at the 99% level. However, the relationship of the SST bias to SSTA forecast error is relatively weak and depends on lead time. There is no obvious relationship between the SST bias and the initial reconstruction error for SSTAs (Figure 2c), although the correlation

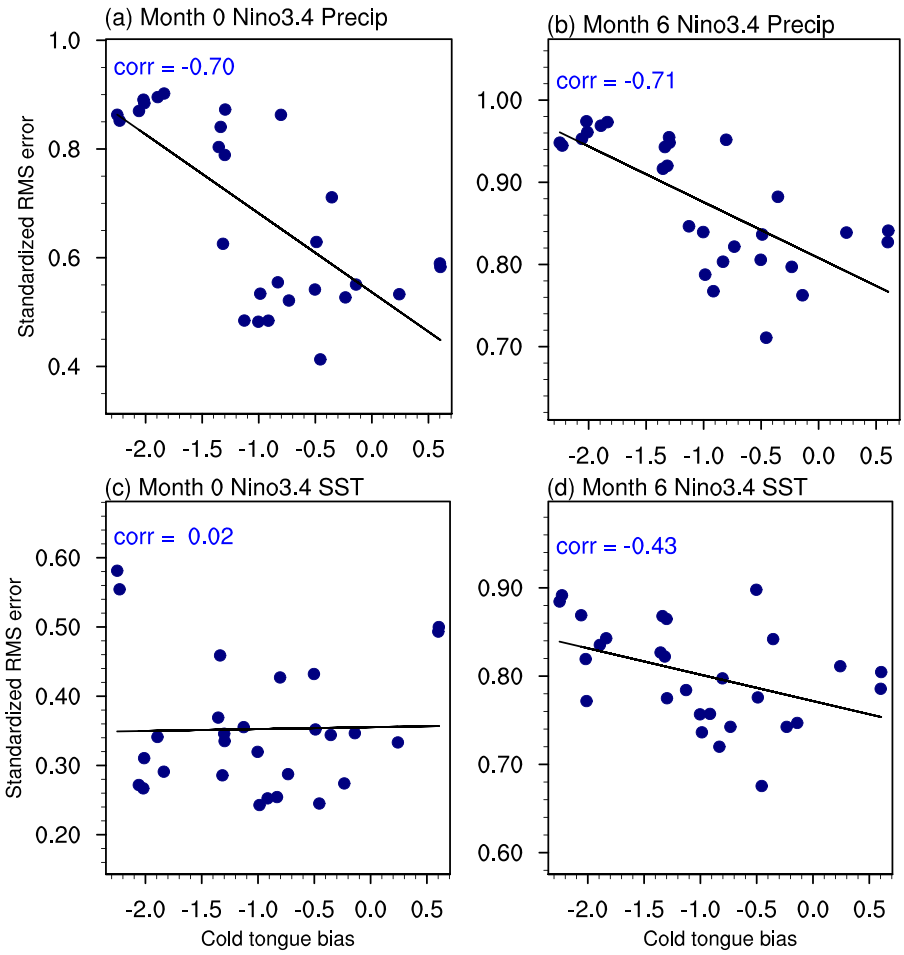

Figure 2. Scatterplots of model-analog standardized root-mean-square forecast error (SRMSE) versus control-run Niño3.4 SST bias (°C), for forecasts of Niño3.4 precipitation at leads of (a) 0 and (b) 6 months, and for forecasts of Niño3.4 SST at leads of (c) 0 and (d) 6 months. Black lines indicate linear regressions. The correlation between bias and error is at the top left of each panel.

between the SST bias and the 6-month lead forecast error reaches about −0.43, significant at the 95% level (Figure 2d). Note that if the group of models with very large precipitation forecast error and mean SST bias colder than −1.7°C is removed, a significant relationship still remains between precipitation forecast error and mean SST bias, while the relationship between SST forecast error and mean SST bias remains weak. We also obtained similar results when relating the SST bias to AC between model-analog forecasts and observations (Figure S3).

We next examine links between Niño3.4 forecast skill and tropical Pacific model biases in both mean states and variances. Again, we define two groups of 12 models each—a "high skill" group and a "low skill" group —using Niño3.4 SRMSE for either SST or precipitation, and either reconstruction skill at lead 0 or forecast skill at lead 6 months. (Results are similar if the models are instead classified using AC, not shown.) For each group, the ensemble mean bias for the corresponding preindustrial control runs is then determined for the means and standard deviations of four fields: SST, SSH, precipitation, and 10-m zonal wind.

The models with the lowest initial precipitation reconstruction skill suffer from severe climatological biases at the equator (Figures 3a and 3e), including an SST bias of −2 °C, rainfall biases of −2 to −3 mm/day, and an easterly wind bias of about −3 m/s in the western equatorial Pacific, with a corresponding positive SSH bias in the west with maxima straddling the equator (not shown). In addition, they also display cold SST biases in the north and south tropical Pacific, a warm SST bias near the coast of South America, and excessive rainfall off the equator (Lin, 2007; Zheng et al., 2011; Li & Xie, 2014). The systematic errors at the equator are

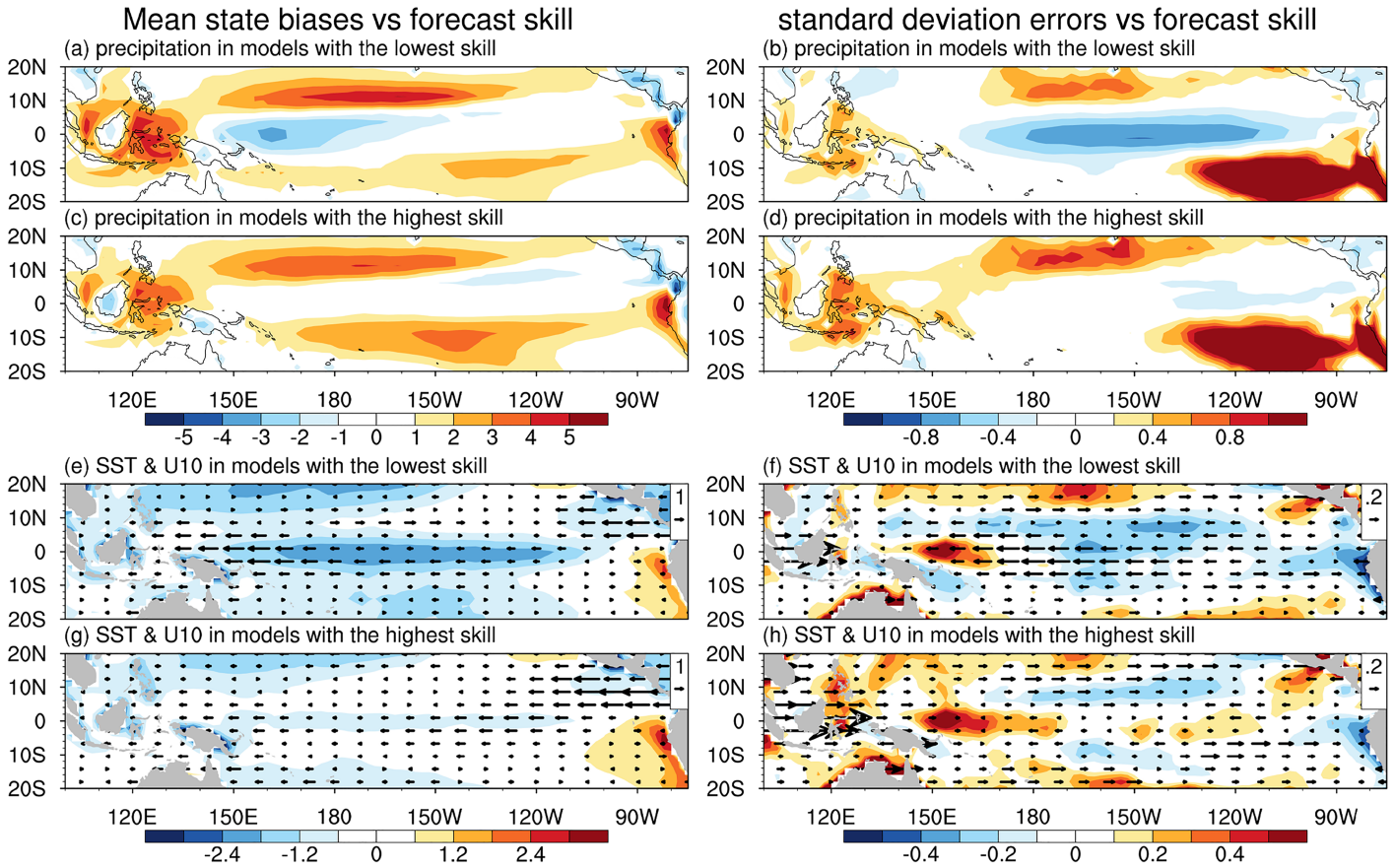

Figure 3. Multimodel biases in mean states (left) and standard deviations (right) for (a–d) precipitation and (e–h) SST (shading) and 10-m zonal winds (vectors). Biases are calculated from the preindustrial control simulations of the 12 models with the (a, b, e, and f) worst and (c, d, g, and h) best initial model-analog reconstruction error (SRMSE) for Niño3.4 precipitation. In the right column, all quantities are normalized by the observed climatological standard deviation, and vectors pointing east (west) indicate standard deviation values greater (less) than observed for the 10-m zonal winds.

mitigated for the high-skill models (Figures 3c and 3g), which exhibit a modest cold bias of around −0.6 °C, close-to-observed rainfall in the western and central equatorial Pacific, and a much weaker easterly wind bias and low-SSH bias in the western Pacific. The differences between the two ensembles are significant at the 95% level in the equatorial Pacific (Figure 4a). (Regressing the model mean states onto the model-analog Niño3.4 precipitation reconstruction skill gives similar results, not shown.) The high-skill models also exhibit significantly less central equatorial cold SST bias year-round than the low-skill models, with some seasonal variation (Figure S4). However, the high-skill models still have some off-equatorial Pacific biases, and warm SST biases off the west coast of South America, which are almost as severe as in the low-skill models, suggesting that the varying ENSO prediction skill among the models is mainly driven by their variations in mean states within the ocean equatorial waveguide.

The models with the lowest Niño3.4 precipitation reconstruction skill also exhibit too little climate variability in the central/east equatorial Pacific (Figures 3b and 3f). Compared with observations, the SST variability is 20% weaker in the central/east Pacific and 40–50% stronger in the west (Figure 3f), indicative of a westward displacement of the ENSO SSTA pattern in the presence of a cold SST bias within the cold tongue (e.g., Guilyardi et al., 2009; Joseph & Nigam, 2006; Newman & Sardeshmukh, 2017). Correspondingly, the variability of precipitation and 10-m zonal winds are around 60% and 50% weaker than observed, respectively (Figures 3b and 3f). In contrast, the high-skill models simulate much more realistic variability of these quantities at the equator (Figures 3d and 3h), though they still have too much SST variability in the western equatorial Pacific (Figure 3h). The difference between the high- and low-skill models is significant at the 95% level in the equatorial Pacific (Figure 4b). The high-skill models also more realistically simulate the seasonal cycle of equatorial SST variability (Figure S5).

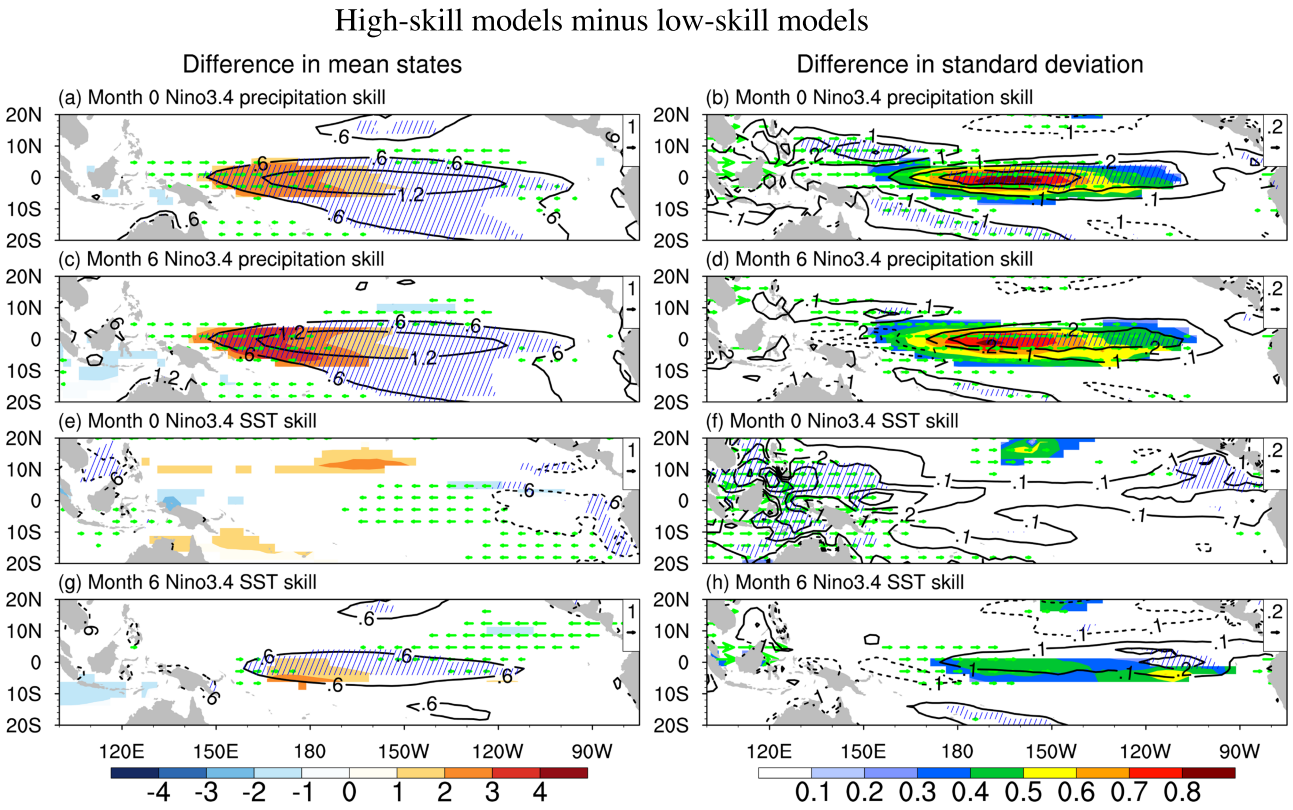

Figure 4. Differences in (a, c, e, and g) mean states and (b, d, f, and h) standard deviations between the model groups with the highest and lowest reconstruction/forecast skill for Niño3.4 precipitation at leads (a, b) 0 and (c, d) 6 months, and for Nino3.4 SST at (e, f) 0 and (g, h) 6 months. Differences in precipitation (shaded) and 10-m zonal winds (vectors) are shown only at grid cells exceeding 95% significance. Differences in SST (contours) are hatched when they exceed 95% significance. The Student's *t* test is used to test significance, with each model as an independent sample. In the right column, all quantities are normalized by the observed climatological standard deviation, and vectors pointing east (west) indicate greater (less) standard deviations for the 10-m zonal winds.

Grouping models based on 6-month lead Niño3.4 precipitation skill again results in the high-skill models having much more realistic mean states and variability (Figure S6). The high- and low-skill model equatorial Pacific differences are again significant at the 95% level for both mean states (Figure 4c) and variability (Figure 4d). Similar results are obtained for other forecast leads (not shown).

We further examined systematic errors in CMIP5 historical simulations, driven by observed changes in radiative forcing over the twentieth century, for each set of models used to generate Figure 4. The low- and high-skill models display nearly the same errors in the historical simulations as those in the preindustrial control runs (not shown). Results were also generally similar when using Niño4 rather than Niño3.4 precipitation skill (not shown).

Finally, we also grouped models using high and low SST skill. As suggested by Figure 2c, Figure 4e confirms that there is little significant difference in the mean states of either SST or precipitation between models with high or low SST reconstruction skill, although low-skill models have a positive SSH bias in the eastern part of the basin (not shown). In contrast, models with high 6-month lead SST skill display significantly more realistic mean states than the low-skill models (Figure 4g) and more realistic variability of precipitation and 10-m zonal wind (Figure 4h). However, these differences are generally much smaller than when grouping by precipitation skill (Figures 4a–4d).

4. Summary and Discussion

In this study, model-analog hindcasts of SST (1961–2015) and precipitation (1979–2015) anomalies, generated from the preindustrial control runs of 28 different CMIP5 models, demonstrate a direct relationship between seasonal forecast skill and model systematic errors. There are significant intermodel correlations

between the equatorial Pacific cold tongue bias and the model-analog initial reconstruction/forecast skill for precipitation anomalies. Models with the best model-analog precipitation anomaly skill showed the most realistic mean states and interannual variability for precipitation, SST, and 10-m zonal winds in the equatorial Pacific. Surprisingly, the cold tongue bias was only weakly linked to SSTA forecast skill. Models with better hindcast skill for SSTAs at Month 6 or longer leads also simulated more realistic mean states and variability than models with lower hindcast skill, but this relationship did not extend to initial reconstruction skill for SSTAs. Moreover, the difference between the two model groups with the lowest and highest 6-month lead SSTA hindcast skill is less than that between the two model groups with the lowest and highest reconstruction/hindcast skill for precipitation anomalies. Thus, even though precipitation is not used to select the model-analogs, model-analog precipitation hindcast skill emerges as a key identifier of better tropical Indo-Pacific simulations. Our results also suggest that reproducing the time evolution of SSTAs is an important additional metric beyond simply reproducing static aspects of SST variability.

Recent studies have noted that the SST bias in the equatorial Pacific cold tongue strongly impacts the air-sea coupled "Bjerknes feedbacks" that are important for ENSO (Bayr et al., 2018, 2019). A cold equatorial Pacific SST bias induces a westward displacement of the rising branch of the Walker circulation, thereby weakening the atmospheric convective response to equatorial Pacific SSTAs and degrading precipitation forecast skill as shown in the present study. While our study and others (e.g., Bayr et al., 2019; Dommenget et al., 2014) have found that some climate models can simulate realistic SST variability despite a severe cold SST bias, the SST variability in those models is prominently driven by unrealistic positive shortwave cloud feedbacks associated with excessive stratus cloud, rather than by Bjerknes feedbacks (Bayr et al., 2019; Dommenget et al., 2014).

Standard metrics (Bellenger et al., 2014; Guilyardi et al., 2009) have been proposed to diagnose, understand, and evaluate ENSO in CGCMs, to support CMIP and the Intergovernmental Panel on Climate Change Assessment Reports (Meehl et al., 2007; Taylor et al., 2012). These metrics assess the main features and processes important for ENSO, including the tropical Pacific mean state, annual cycle, anomaly patterns and evolution in the central/eastern equatorial Pacific, teleconnections to remote regions, and key dynamic and thermodynamic feedbacks. Additional ENSO metrics have been proposed, for example, to measure ENSO nonlinearity in CMIP5 simulations (Karamperidou et al., 2017) and to assess ENSO's spatiotemporal diversity (Capotondi et al., 2015; Kug et al., 2010; Lee et al., 2014; Lemmon & Karnauskas, 2019). Our study introduces two further metrics—model-analog reconstruction skill and forecast skill—which measure the ability of a model to represent observed ENSO events and their spatiotemporal evolution, offering a powerful new way to evaluate a model and its potential utility for seasonal predictions.

Acknowledgments
This work has been funded by the NOAA Earth System Modeling. M. N. was partially supported by DOE (Grant 0000238382). We thank Andrew Hoell for his constructive comments. We acknowledge the CMIP5 climate modeling groups for producing and making available their model output. The CMIP5 data output is available online (https://esgf-node.llnl.gov/search/cmip5/).

References

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P.-P., Janowiak, J., et al. (2003). The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *Journal of Hydrometeorology*, *4*(6), 1147–1167. https://doi.org/10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2

Alexander, M., Bladé, I., Newman, M., Lanzante, J., Lau, N., & Scott, J. (2002). The atmospheric bridge: The influence of ENSO teleconnections on air-sea interaction over the global oceans. *Journal of Climate*, *15*(16), 2205–2231. https://doi.org/10.1175/1520-0442(2002)015<2205:TABTIO>2.0.CO;2

Balmaseda, M. A., Mogensen, K., & Weaver, A. T. (2013). Evaluation of the ECMWF ocean reanalysis system ORAS4. *Quarterly Journal of the Royal Meteorological Society*, *139*, 1132–1161. https://doi.org/10.1002/qj.2063

Barnett, T., & Preisendorfer, R. (1978). Multifield analog prediction of short-term climate fluctuations using a climate state vector. *Journal of the Atmospheric Sciences*, *35*(10), 1771–1787. https://doi.org/10.1175/1520-0469(1978)035<1771:MAPOST>2.0.CO;2

Bayr, T., Latif, M., Dommenget, D., Wengel, C., Harlaß, J., & Park, W. (2018). Mean-state dependence of ENSO atmospheric feedbacks in climate models. *Climate Dynamics*, *50*, 3171–3194. https://doi.org/10.1007/s00382-017-3799-2

Bayr, T., Wengel, C., Latif, M., Dommenget, D., Luebbecke, J., & Park, W. (2019). Error compensation of ENSO atmospheric feedbacks in climate models and its influence on simulated ENSO dynamics. *Climate Dynamics*, *53*, 155–172. https://doi.org/10.1007/s00382-018-4575-7

Bellenger, H., Guilyardi, E., Leloup, J., Lengaigne, M., & Vialard, J. (2014). ENSO representation in climate models: From CMIP3 to CMIP5. *Climate Dynamics*, *42*, 1999–2018. https://doi.org/10.1007/s00382-013-1783-z

Capotondi, A., Wittenberg, A. T., Newman, M., di Lorenzo, E., Yu, J. Y., Braconnot, P., et al. (2015). Understanding ENSO diversity. *Bulletin of the American Meteorological Society*, *96*, 921–938. https://doi.org/10.1175/BAMS-D-13-00117.1

Chen, C., Cane, M. A., Wittenberg, A. T., & Chen, D. (2017). ENSO in the CMIP5 simulations: Life cycles, diversity, and responses to climate change. *Journal of Climate*, *30*, 775–801. https://doi.org/10.1175/JCLI-D-15-0901.1

Choi, K.-Y., Vecchi, G. A., & Wittenberg, A. T. (2015). Nonlinear zonal wind response to ENSO in the CMIP5 models: Roles of the zonal and meridional shift of the ITCZ/SPCZ and the simulated climatological precipitation. *Journal of Climate*, *28*, 8556–8573. https://doi.org/10.1175/JCLI-D-15-0211.1

Davey, M., Huddleston, M., Sperber, K., Braconnot, P., Bryan, F., Chen, D., et al. (2002). STOIC: A study of coupled model climatology and variability in tropical ocean regions. *Climate Dynamics*, *18*(5), 403–420. https://doi.org/10.1007/s00382-001-0188-6

DelSole, T., & Shukla, J. (2010). Model fidelity versus skill in seasonal forecasting. *Journal of Climate*, *23*(18), 4794–4806. https://doi.org/10.1175/2010JCLI3164.1

Ding, H., Greatbatch, R. J., Latif, M., & Park, W. (2015). The impact of sea surface temperature bias on equatorial Atlantic interannual variability in partially coupled model experiments. *Geophysical Research Letters*, *42*, 5540–5546. https://doi.org/10.1002/2015GL064799

Ding, H., Newman, M., Alexander, M. A., & Wittenberg, A. T. (2018). Skillful climate forecasts of the tropical Indo-Pacific Ocean using model-analogs. *Journal of Climate*, *31*, 5437–5459. https://doi.org/10.1175/JCLI-D-17-0661.1

Ding, H., Newman, M., Alexander, M. A., & Wittenberg, A. T. (2019). Diagnosing secular variations in retrospective ENSO seasonal forecast skill using CMIP5 model-analogs. *Geophysical Research Letters*, *46*, 1721–1730. https://doi.org/10.1029/2018GL080598

Dippe, T., Greatbatch, R. J., & Ding, H. (2019). Seasonal prediction of equatorial Atlantic sea surface temperature using simple initialization and bias correction techniques. *Atmospheric Science Letters*, *20*, e898. https://doi.org/10.1002/asl.898

Dommenget, D., Haase, S., Bayr, T., & Frauen, C. (2014). Analysis of the Slab Ocean El Niño atmospheric feedbacks in observed and simulated ENSO dynamics. *Climate Dynamics*, *42*, 3187–3205. https://doi.org/10.1007/s00382-014-2057-0

Graham, F. S., Wittenberg, A. T., Brown, J. N., Marsland, S. J., & Holbrook, N. J. (2017). Understanding the double peaked El Niño in coupled GCMs. *Climate Dynamics*, *48*, 2045–2063. https://doi.org/10.1007/s00382-016-3189-1

Gualdi, S., Alessandri, A., & Navarra, A. (2005). Impact of atmospheric horizontal resolution on El Niño Southern Oscillation forecasts. *Tellus A: Dynamic Meteorology and Oceanography*, *57*(3), 357–374. https://doi.org/10.3402/tellusa.v57i3.14662

Guilyardi, E., Wittenberg, A., Fedorov, A., Collins, M., Wang, C., Capotondi, A., et al. (2009). Understanding El Niño in ocean-atmosphere general circulation models: Progress and challenges. *Bulletin of the American Meteorological Society*, *90*, 3. https://doi.org/10.1175/2008BAMS2387.1

He, J., Johnson, N. C., Vecchi, G. A., Kirtman, B., Wittenberg, A. T., & Sturm, S. (2018). Precipitation sensitivity to local variations in tropical sea surface temperature. *Journal of Climate*, *31*, 9225–9238. https://doi.org/10.1175/JCLI-D-18-0262.1

Joseph, R., & Nigam, S. (2006). ENSO evolution and teleconnections in IPCC's twentieth-century climate simulations: Realistic representation? *Journal of Climate*, *19*, 17. https://doi.org/10.1175/JCLI3846.1

Karamperidou, C., Jin, F.-F., & Conroy, J. L. (2017). The importance of ENSO nonlinearities in tropical pacific response to external forcing. *Climate Dynamics*, *49*, 2695–2704. https://doi.org/10.1007/s00382-016-3475-y

Kim, S. T., Jeong, H.-I., & Jin, F.-F. (2017). Mean bias in seasonal forecast model and ENSO prediction error. *Scientific Reports*, *7*, 6029. https://doi.org/10.1038/s41598-017-05221-3

Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L. III, Paolino, D. A., Zhang, Q., et al. (2014). The North American multimodel ensemble: Phase-1 seasonal-to-interannual prediction; Phase-2 toward developing intraseasonal prediction. *Bulletin of the American Meteorological Society*, *95*, 585–601. https://doi.org/10.1175/BAMS-D-12-00050.1

Kug, J.-S., Choi, J., An, S.-I., Jin, F.-F., & Wittenberg, A. T. (2010). Warm pool and cold tongue El Niño events as simulated by the GFDL CM2.1 coupled GCM. *Journal of Climate*, *23*(5), 1226–1239. https://doi.org/10.1175/2009JCLI3293.1

Lee, J.-Y., Wang, B., Kang, I.-S., Shukla, J., Kumar, A., Kug, J.-S., et al. (2010). How are seasonal prediction skills related to models' performance on mean state and annual cycle? *Climate Dynamics*, *35*, 2–3. https://doi.org/10.1007/s00382-010-0857-4

Lee, S.-K., DiNezio, P. N., Chung, E. S., Yeh, S. W., Wittenberg, A. T., & Wang, C. (2014). Spring persistence, transition and resurgence of El Niño. *Geophysical Research Letters*, *41*, 8578–8585. https://doi.org/10.1002/2014GL062484

Lemmon, D. E., & Karnauskas, K. B. (2019). A metric for quantifying El Niño pattern diversity with implications for ENSO-mean state interaction. *Climate Dynamics*, *52*, 7511–7523. https://doi.org/10.1007/s00382-018-4194-3

Li, G., & Xie, S.-P. (2014). Tropical biases in CMIP5 multimodel ensemble: The excessive equatorial Pacific cold tongue and double ITCZ problems. *Journal of Climate*, *27*, 1765–1780. https://doi.org/10.1175/JCLI-D-13-00337.1

Lin, J.-L. (2007). The double-ITCZ problem in IPCC AR4 coupled GCMs: Ocean-atmosphere feedback analysis. *Journal of Climate*, *20*, 18. https://doi.org/10.1175/JCLI4272.1

Lorenz, E. N. (1969). Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric Sciences*, *26*(4), 636–646. https://doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2

Magnusson, L., Alonso-Balmaseda, M., Corti, S., Molteni, F., & Stockdale, T. (2013). Evaluation of forecast strategies for seasonal and decadal forecasts in presence of systematic model errors. *Climate Dynamics*, *41*, 2393–2409. https://doi.org/10.1007/s00382-012-1599-2

Manganello, J. V., & Huang, B. (2009). The influence of systematic errors in the Southeast Pacific on ENSO variability and prediction in a coupled GCM. *Climate Dynamics*, *32*,1015. https://doi.org/10.1007/s00382-008-0407-5

Meehl, G. A., Covey, C., Taylor, K. E., Delworth, T., Stouffer, R. J., Latif, M., et al. (2007). The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bulletin of the American Meteorological Society*, *88*, 9. https://doi.org/10.1175/BAMS-88-9-1383

Newman, M., & Sardeshmukh, P. D. (2017). Are we near the predictability limit of tropical Indo-Pacific sea surface temperatures? *Geophysical Research Letters*, *44*, 8520–8529. https://doi.org/10.1002/2017GL074088

Philander, S. (1990). *El Niño, La Niña, and the Southern Oscillation*. New York, NY: Academic Press.

Rayner, N., Parker, D. E., Horton, E., Folland, C., Alexander, L., Rowell, D., et al. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research*, *108*(D14), 4407. https://doi.org/10.1029/2002JD002670

Richter, I., Doi, T., Behera, S. K., & Keenlyside, N. (2018). On the link between mean state biases and prediction skill in the tropics: An atmospheric perspective. *Climate Dynamics*, *50*, 3355–3374. https://doi.org/10.1007/s00382-017-3809-4

Scaife, A. A., Ferranti, L., Alves, O., Athanasiadis, P., Baehr, J., Deque, M., et al. (2019). Tropical rainfall predictions from multiple seasonal forecast systems. *International Journal of Climatology*, *39*, 974–988. https://doi.org/10.1002/joc.5855

Smith, D. M., Eade, R., & Pohlmann, H. (2013). A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction. *Climate Dynamics*, *41*, 3325–3338. https://doi.org/10.1007/s00382-013-1683-2

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, *106*(D7), 7183–7192. https://doi.org/10.1029/2000JD900719

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*, 4. https://doi.org/10.1175/BAMS-D-11-00094.1

Timmermann, A., An, S. I., Kug, J. S., Jin, F. F., Cai, W., Capotondi, A., et al. (2018). El Niño-Southern Oscillation complexity. *Nature*, *559*, 535–545. https://doi.org/10.1038/s41586-018-0252-6

Trenberth, K., Branstator, G., & Karoly, D. (1998). Progress during TOGA in under- standing and modeling global teleconnections associated with tropical. *Journal of Geophysical Research*, *103*(C7), 14–291. https://doi.org/10.1029/97JC01444

Trenberth, K. E., & Stepaniak, D. P. (2001). Indices of El Niño evolution. *Journal of Climate*, *14*(8), 1697–1701. https://doi.org/10.1175/1520-0442(2001)014<1697:LIOENO>2.0.CO;2

Van den Dool, H. (1994). Searching for analogues, how long must we wait? *Tellus A*, *46*(3), 314–324. https://doi.org/10.1034/j.1600-0870.1994.t01-2-00006.x

Vecchi, G. A., Delworth, T., Gudgel, R., Kapnick, S., Rosati, A., Wittenberg, A. T., et al. (2014). On the seasonal forecasting of regional tropical cyclone activity. *Journal of Climate*, *27*, 7994–8016. https://doi.org/10.1175/JCLI-D-14-00158.1

Wittenberg, A. T., Vecchi, G. A., Delworth, T. L., Rosati, A., Anderson, W., Cooke, W. F., et al. (2018). Improved simulations of tropical Pacific annual-mean climate in the GFDL FLOR and HiFLOR coupled GCMs. *Journal of Advances in Modeling Earth Systems*, *10*, 3176–3220. https://doi.org/10.1029/2018MS001372

Zheng, Y., Shinoda, T., Lin, J.-L., & Kiladis, G. N. (2011). Sea surface temperature biases under the stratus cloud deck in the southeast Pacific Ocean in 19 IPCC AR4 coupled general circulation models. *Journal of Climate*, *24*, 15. https://doi.org/10.1175/2011JCLI4172.1