AGU PUBLICATIONS

Geophysical Research Letters

Supporting Information for

Diagnosing secular variations in retrospective ENSO seasonal forecast skill using CMIP5 model-analogs

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

1. CIRES, University of Colorado, Boulder, Colorado
2. NOAA Earth Systems Research Laboratory, Boulder, Colorado
3. NOAA Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey

Contents of this file

Text S1 to S4
Figures S1 to S8
Tables S1 to S2

Introduction

The following gives additional details of the RMS distance metric used to choose analogs, how the trend component is included in the model-analog hindcasts, skill comparison with the NMME assimilation-initialized hindcasts, and how the CMIP5 "best-10" ensemble is chosen.

Text S1.

<u>The RMS distance</u>

Ding et al (2018) defined a root-mean-square (RMS) distance to choose analogs. The RMS distance between a target state $\mathbf{x}(t)$ and a library state $\mathbf{y}(t')$ is given by $d^2(t,t') = \sum_{i=1}^{2}\sum_{j=1}^{J}(\frac{x_j^i(t)}{\sigma_X^i} - \frac{y_j^i(t')}{\sigma_Y^i})^2$, with superscripts $i = 1$ indicating SSH anomalies (SSHAs) and $i = 2$ SST anomalies (SSTAs), and subscript j representing a gridpoint index with J total gridpoint indices within the training region. In the distance equation, σ_X^i and σ_Y^i indicate respective area averaged standard deviations for the target and library states. Readers can refer to Ding et al (2018) for more details on the model-analog technique.

Text S2.

The North American Multi-model Ensemble (NMME) hindcasts

We obtained retrospective forecast (hindcasts) from eight different models in the current North American Multi-model Ensemble (NMME) project (Kirtman et al. 2014); see Table S1 for details. To calculate anomalies, all hindcasts are “bias-corrected”: the mean hindcast drift as a function of lead and calendar month is removed separately for each ensemble member of each model, as is common practice with CGCM seasonal forecasts (Stockdale 1997; Saha et al. 2006; Kirtman and Min 2009). Following Barnston et al. (2015), the grand multi-model ensemble mean (NMME grand ensemble mean) forecasts were then determined using the bias-corrected ensemble members of all the models.

Text S3.

Accounting for externally-forced trends

First, we assume that the externally-forced component is separable from internal variability; then, following the method in Dai et al. (2015), any anomaly x is

$$x(j,t) = x_F(j,t) + x_I(j,t)$$

where j and t denote grid point and time, respectively, $x_{\boldsymbol{F}}(j,t)$ is the externally forced component, and $x_{\boldsymbol{I}}(j,t)$ is the internal climate anomaly. Suppose that $T(t)$ represents the best estimate of the evolving global mean surface temperature response to historical external forcing (Dai et al., 2015). The externally-forced component is then estimated by linear regression of $x(j,t)$ onto $T(t)$:

$$x_{\boldsymbol{F}}(j,t) = b(j) \times T(t) \ ,$$

where $b(j)$ is the regression slope at grid point j determined over the entire period. The initial internal (i.e., detrended) anomaly is then the residual

$$x_{\boldsymbol{I}}(j,t) = x(j,t) - b(j) \times T(t) \ .$$

Note that $x_{\boldsymbol{I}}(j,t)$is now used to determine analogs within the fixed climate control simulation, instead of $x(j,t)$ as was done by D18. Finally, the linear estimate of the externally-forced component is added back to each model-analog forecast ensemble member, resulting in the final forecast ensemble $\{y\left(\mathrm{t}_1' + \tau\right) + b \times T(t+\mathrm{\tau}), \ldots, y\left(\mathrm{t}_\mathrm{k}' + \tau\right) + b \times T(t+\mathrm{\tau}), \ldots, y(\mathrm{t}_\mathrm{K}' + \tau) + b \times T(t+\mathrm{\tau})\}$, which is then verified against $x(\mathrm{t}+\mathrm{\tau})$.

Note that $T(t+\mathrm{\tau})$ is used instead of $T(t)$, to determine externally-forced contributions to the seasonal forecast. The trend clearly has a large impact on skill; for example, Fig. S1 shows that the skill coming from just the predicted trend component alone is quite large for SST hindcasts (although not for precipitation) in the Indian and west Pacific oceans. However, the impact of *predicting* the trend over the forecast lead time is very small, as is also shown in Fig. S1, which shows that the skill due to merely persisting the trend component (that is, using $T(t)$ in the forecast instead) is almost the same. That is, the primary benefit of including the trend component is to allow the initial state to better match observations, especially in regions where the trend dominates natural variability. In other words, within each 1-year forecast period, the evolution of the externally-forced trend component is slight. It is not the externally-forced trend over the next 1-year forecast period that matters, but the external-forcing induced trend that accumulated over the past fifty years.

This figure then also suggests that this skill is largely due to the skill metrics that are typically used for seasonal forecasts, so that a skill metric based on comparing detrended anomalies with detrended hindcasts would yield much lower values in the western Pacific and Indian ocean.

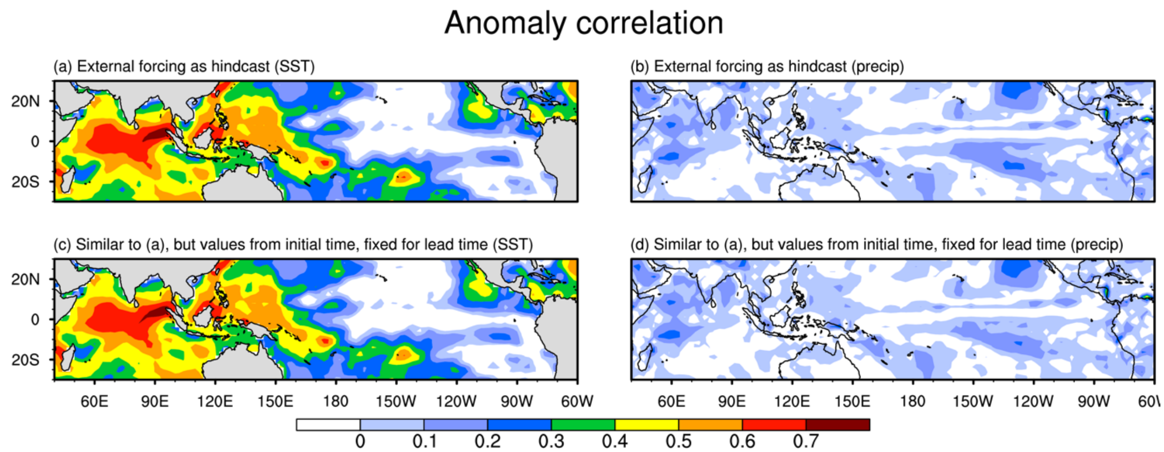

Figure S1. Evaluation of external forcing as hindcasts for SST (a, c) and precip (b, d). In (a, b), time evolution of $T(t+\tau)$, arising from τ, is taken into account; while in (c, d), $T(t)$ is fixed for all lead time τ. See Text S3 for details.

We determined T as the global and ensemble mean surface temperature of the CMIP5 multi-model historical (pre-2005) and RCP4.5 (post-2005) runs, rather than directly from observations, so that the model-analog technique can make forecasts as well as hindcasts. In essence, the CMIP5 ensemble mean predicts the externally-forced component, and the model-analog technique predicts the internal climate anomaly. We used all 45 CMIP5 historical simulations to estimate the externally-forced signal, although using just those models corresponding to our model-analog ensemble yielded essentially the same results. For each initial state at time t, the regression coefficient $b(j)$ was separately determined from the 1961-2015 period except for data from the interval [t, t + 5 yrs], which was withheld to ensure that the trend component of each hindcast was independent of the verification data.

Text S4.

Choosing the "best-10" CMIP models

We note that a few models (e.g., CCSM4 and IPSL-CM5B-LR) are slightly more skillful than the 28-model mean in the central equatorial Pacific, perhaps because the 28-model mean skill is reduced by including some models with very low skill. For example, Table S1 shows that CMIP5 model Niño3.4 SST 6-month forecast skill ranges between 0.49-0.75. We explored the impact of adding models with less skill to the grand ensemble mean by ordering the models based on Nino3.4 SST 6-month forecast skill, and then including them one at a time in the multi-model grand ensemble mean (see Fig. S5). We found that the multi-model mean skill reached a maximum for an ensemble size of between 5-12 models, but beyond this point the forecast skill degraded as models with poorer performance were added. This suggests that only a subset of the 28 models is necessary to maximize forecast skill. Several recent studies also have found that an ensemble including only some models, determined using some suitable criteria, may yield higher forecast skill relative to using all available forecast models (e.g., Chen & van den Dool, 2017). However, note that the overall change in skill in predicting Nino3.4 SST is modest: rising from ~0.75 for the best model to ~0.775 for 5-12 models and then declining to only 0.73 for all 28 models (Fig. S5). As an example, in Fig. 1 we also show the skill of the model-analog multi-model ensemble determined from the ten most skillful CMIP5 models from Fig. S5 (marked by stars in Table S2). This “best-10” grand ensemble mean modestly improves skill over the 28-model mean in the tropical Pacific, with 0.7 correlation covering a larger area, but does not much improve skill elsewhere.

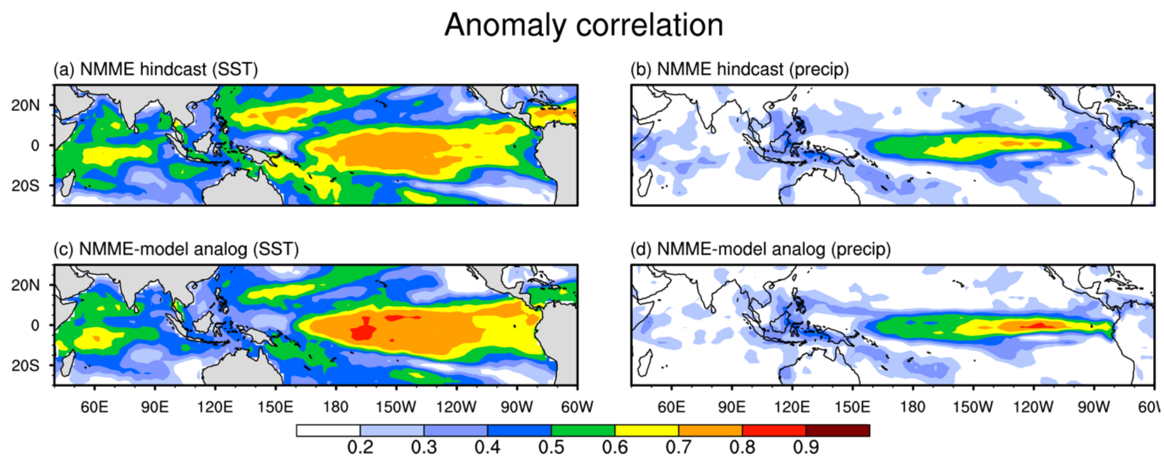

Figure S2. Skill of NMME (a,b) assimilation-initalized and (c,d) model-analog hindcasts of SST (a, c) and precipitation (b, d) anomalies at six-month lead, for the years 1982-2009. Only anomaly correlation is shown. Panels (a-b) show the NMME grand mean conducted by the same four models (Table S2); panels (c-d) show the grand mean of multi-model analogs, based on the four models: CM2.1, CM2.5 FLOR, CCSM4 and CESM1. The projected response to external radiative forcing is added to model-analog hindcasts.

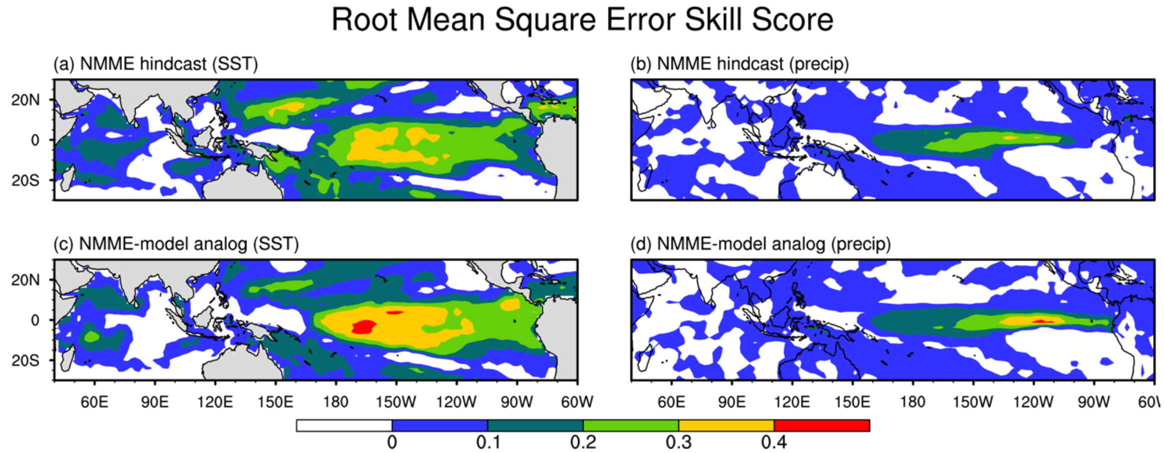

Figure S3. The same as Fig. S2 except for RMS skill score.

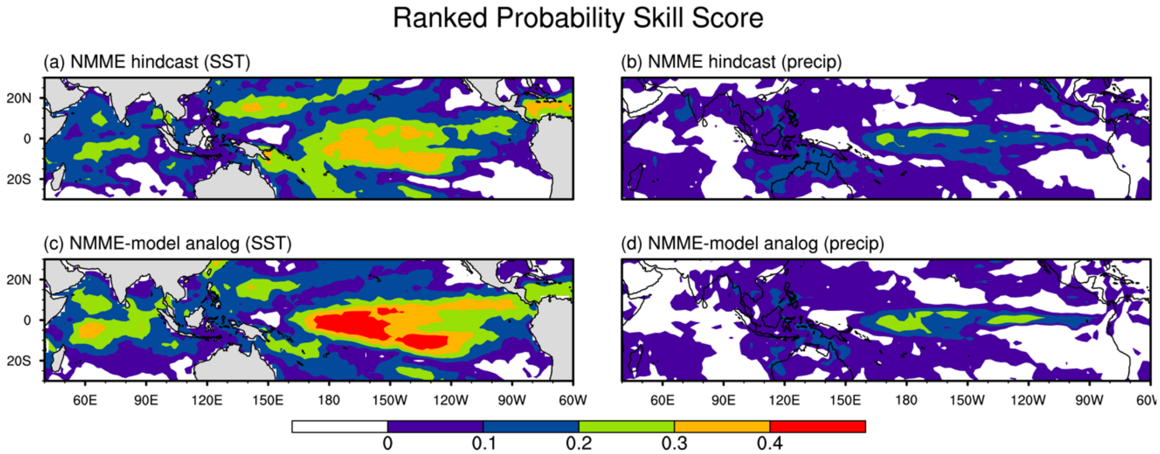

Figure S4. The same as Fig. S2 except for ranked probability skill score (RPSS).

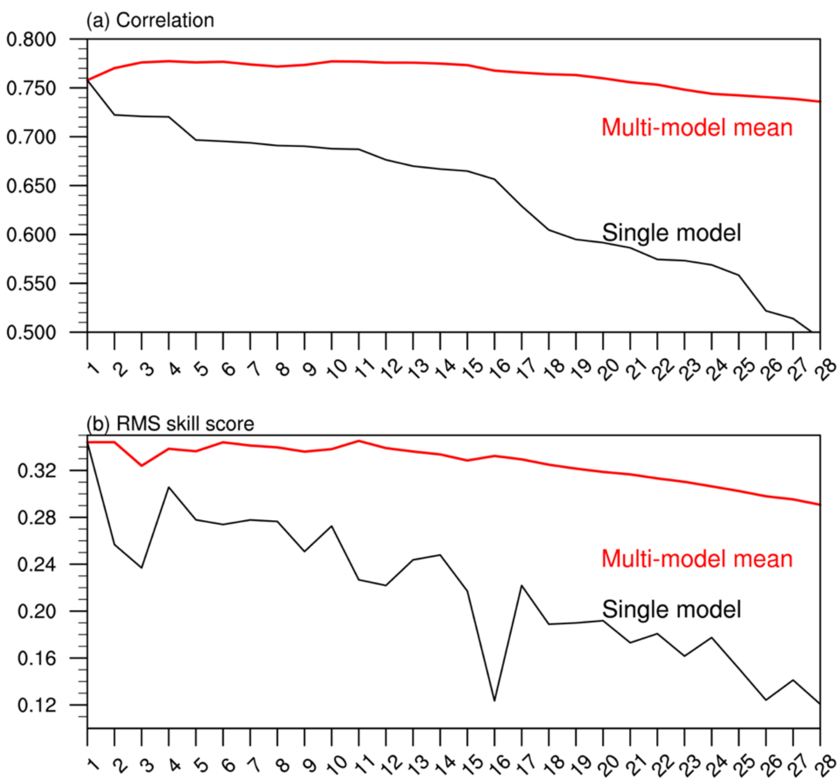

Figure S5. The red curves show model-analog multi-model mean 6-month lead forecast skill of Nino3.4 SST anomalies, measured by (a) correlation and (b) RMS skill score, as a function of number of models (abscissa). The 28 grand ensembles are made as follows: 1) individual model-analog ensemble mean correlation skill of Nino3.4 SST is calculated at 6-month lead; 2) correlation values are ranked in descending order; 3) grand model-analog ensembles are constructed by beginning with the model with highest correlation, which is model 1 in abscissa and then adding one more model, with lower correlation than the previous models, to the grand ensemble until all the models are added. The "best-10" grand ensemble denotes 10 in abscissa. The black curves show the evaluation of individual model ensemble mean, and the models are shown in the descending order in 6-month lead correlation skill of Nino3.4 SST.

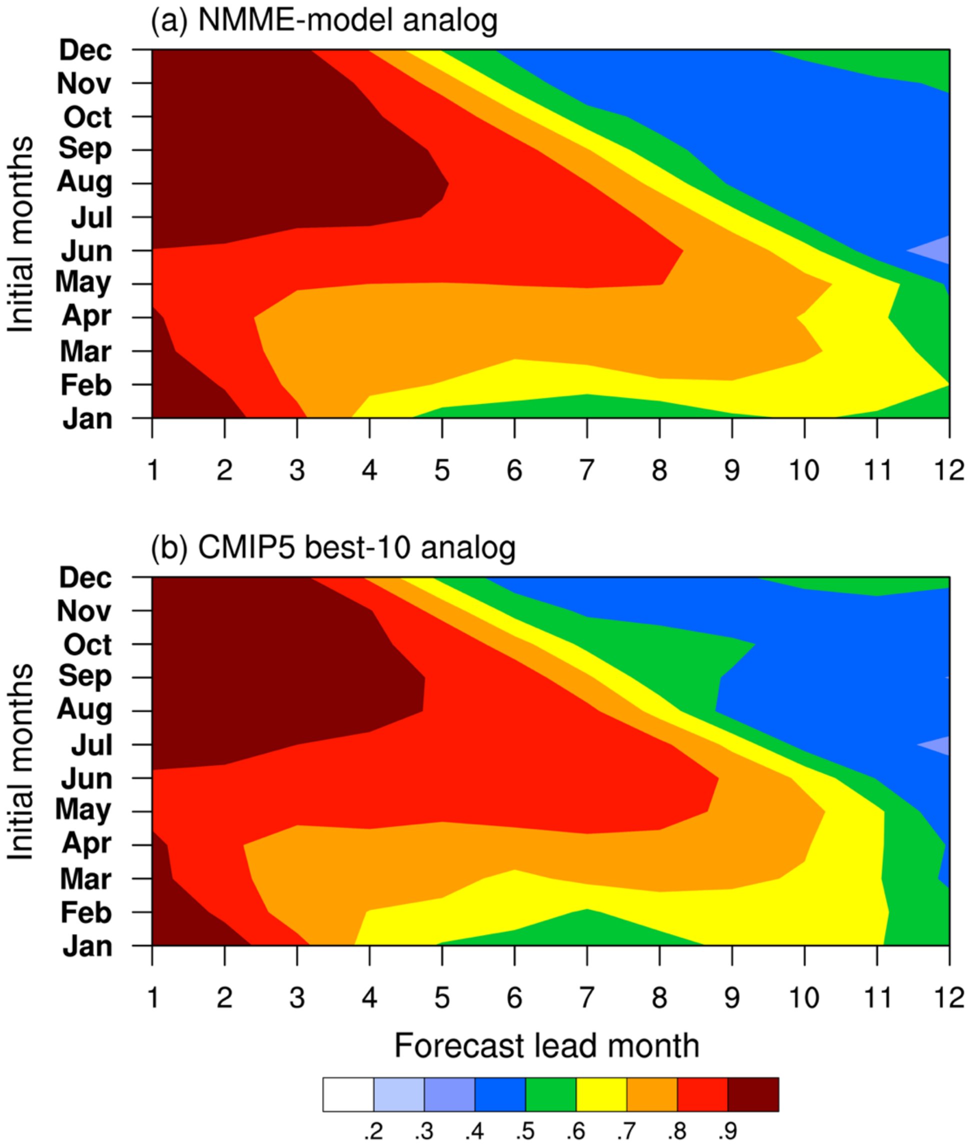

Figure S6. Model-analog hindcast skill for Nino3.4 SST during 1961-2015, as a function of forecast lead time (abscissa) and initial month (ordinate). Shading denotes correlation. Analog hindcasts are based on (a) the NMME models and (b) the “best-10” CMIP5 models, respectively. In all model-analog hindcasts, the projected response to external radiative forcing is included.

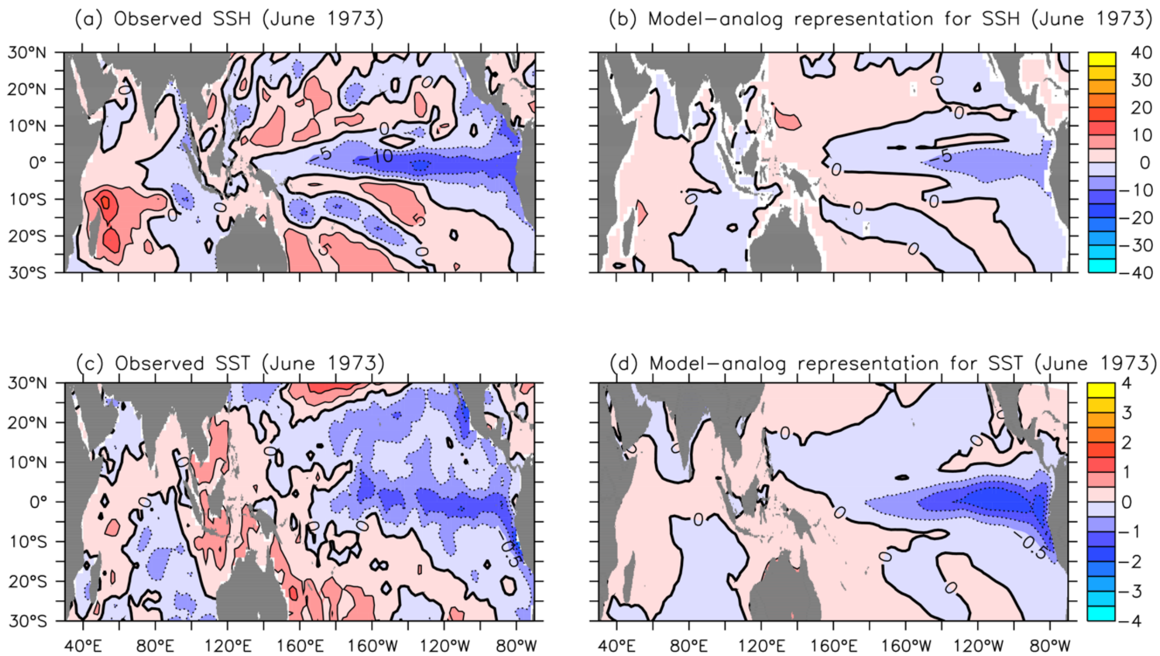

Figure S7. (top row) SSH and (bottom row) SST anomalies in June 1973 calculated from (a, c) observations and (b, d) corresponding CMIP5 CCSM4 model-analog ensemble-mean representation, respectively. The units for SSH and SST are cm and Celsius, respectively.

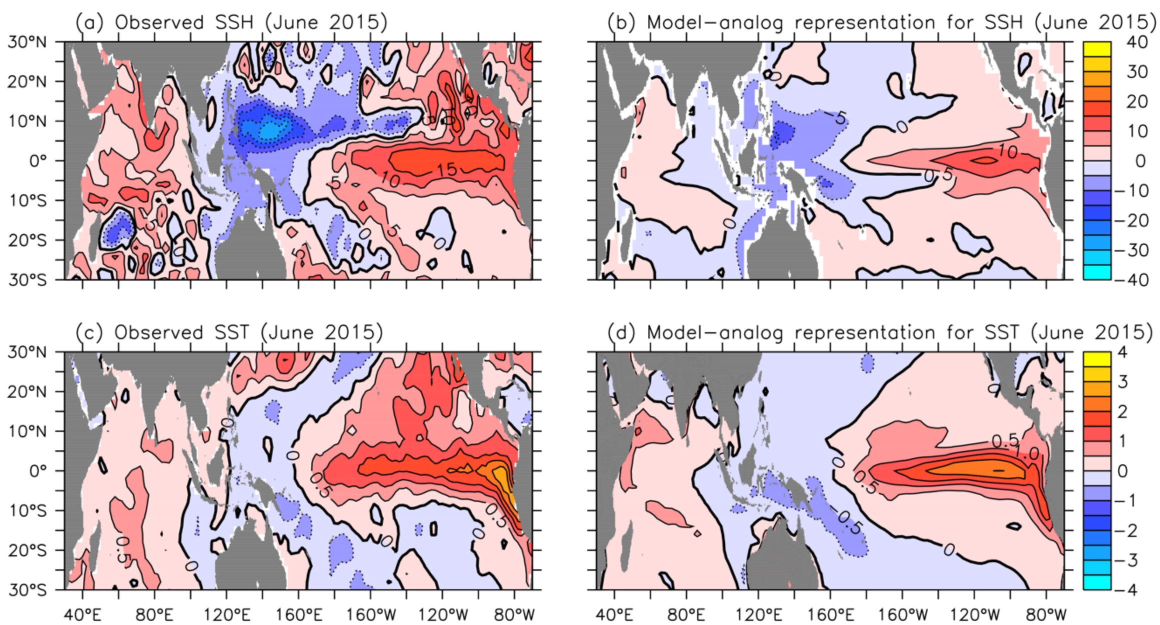

Figure S8. The same as Fig. S7 except for June 2015.

Model	Expanded model name	Number of ensemble members	Maximum lead month
COLA-RSMAS-CCSM4*	COLA–University of Miami–NCAR coupled model	10	12
NCAR-CESM1*	NCAR Community Earth System Model 1	10	12
GFDL-CM2pl-aer04*	Modified version of the GFDL CM2.1 coupled model	10	12
GFDL-CM2p5-FLOR*	GFDL Forecast-oriented Low Ocean Resolution version of CM2.5	24	12
CMC1-CanCM3	Canadian coupled model 1	10	12
CMC2-CanCM4	Canadian coupled model 2	10	12
NASA-GMAO-062012	Modified version of the NASA coupled model	12	10
NCEP-CFSv2	NOAA/NCEP coupled model	24	10

Table S1. A list of the eight NMME models whose hindcast experiments were employed in the NMME grand ensemble. The grand ensemble mean of all the eight NMME models is shown in Figure 4. Model-analog method is applied to the four models, marked by an asterisk, in order to compare with model-analog hindcasts with the corresponding NMME hindcasts (see Figs. S2-S4).

Model name	Expanded model name	Length of run (yr)	Month 6 correlation skill of Nino3.4 SST
ACCESS1-0	Australian Community Climate and Earth System Simulator Coupled Model	500	0.667
ACCESS1-3	Australian Community Climate and Earth System Simulator Coupled Model	500	0.569
CanESM2*	Second Generation Canadian Earth System Model	995	0.720
CCSM4*	Community Climate System Model, version 4	1050	0.758
CMCC-CESM	CMCC Carbon Earth System Model	277	0.656
CMCC-CM	CMCC Climate Model	330	0.629
CMCC-CMS*	CMCC Climate Model with a resolved Stratosphere	500	0.691
CNRM-CM5*	Centre National de Recherches M_et_eorologiques Coupled Global Climate Model, version 5	850	0.688
GFDL-CM3*	Geophysical Fluid Dynamics Laboratory, Climate Model versions 3.0	500	0.695
GFDL-ESM2G*	Geophysical Fluid Dynamics Laboratory Earth System Model with Generalized Ocean Layer	500	0.690

	Dynamics (GOLD) component		
GFDL-ESM2M	Geophysical Fluid Dynamics Laboratory Earth System Model with Modular Ocean Model 4 (MOM4) component	500	0.687
GISS-E2-R*	Goddard Institute for Space Studies Model E2, coupled with the Russell ocean model	550	0.721
GISS-E2-R-CC	Goddard Institute for Space Studies Model E2, coupled with the Russell ocean model, Interactive Carbon Cycle	251	0.676
HadGEM2-CC	Hadley Centre Global Environment Model, version 2–Carbon Cycle	240	0.595
HadGEM2-ES	Hadley Centre Global Environment Model, version 2-Earth System	575	0.514
INMCM4	Institute of Numerical Mathematics Coupled Model, version 4.0	500	0.558
IPSL-CM5A-LR	L'Institut Pierre-Simon Laplace Coupled Model, version 5, coupled with Nucleus for European Modelling of the Ocean (NEMO), low	1000	0.605

	resolution		
IPSL-CM5A-MR	L'Institut Pierre-Simon Laplace Coupled Model, version 5, coupled with NEMO, mid resolution	300	0.670
IPSL-CM5B-LR*	L'Institut Pierre-Simon Laplace Coupled Model, version 5, coupled with NEMO, new atmospheric physics low resolution	300	0.722
MIROC-ESM	Model for Interdisciplinary Research on Climate, Earth System Model	630	0.494
MIROC-ESM-CHEM	Model for Interdisciplinary Research on Climate, Earth System Model, an atmospheric chemistry coupled version	255	0.522
MIROC5	Model for Interdisciplinary Research on Climate, version 5	670	0.587
MPI-ESM-LR	Max Planck Institute Earth System Model, low resolution	1000	0.573
MPI-ESM-MR	Max Planck Institute Earth System Model, medium resolution	1000	0.574
MPI-ESM-P	Max Planck Institute Earth System Model, low	1155	0.592

	resolution, and paleo mode		
MRI-CGCM3	Meteorological Research Institute Coupled Atmosphere–Ocean General Circulation Model, version 3	500	0.665
NorESM1-M*	Norwegian Earth System Model 1, medium resolution	500	0.694
NorESM1-ME*	Norwegian Earth System Model 1, medium resolution with capability to be fully emission driven	252	0.697

Table S2. A list of the 28 CMIP5 models whose preindustrial control simulations were employed as the data library for model-analogs. Models, marked by an asterisk, are employed in the "best-7" grand ensemble.