# On the correspondence between seasonal forecast biases and long-term climate biases in sea surface temperature

**Hsi-Yen Ma[1], A. Cheska Siongco[1], Stephen A. Klein[1], Shaocheng Xie[1], Alicia R. Karspeck[2,3], Kevin Raeder[3], Jeffrey L. Anderson[3], Jiwoo Lee[1], Ben P. Kirtman[4], William J. Merryfield[5], Hiroyuki Murakami[6,7] and Joseph J. Tribbia[3]**

[1]Lawrence Livermore National Laboratory, Livermore, California

[2]Jupiter, Boulder, Colorado

[3]National Center for Atmospheric Research, Boulder, Colorado

[4]Rosenstiel School for Marine and Atmospheric Science, University of Miami, Miami, Florida

[5]Canadian Centre for Climate Modelling and Analysis, Environment and Climate Change Canada, Victoria, British Columbia, Canada

[6]National Oceanic and Atmospheric Administration/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey

[7]University Corporation for Atmospheric Research, Boulder, Colorado

(Submitted to Journal of Climate)

Corresponding author: Hsi-Yen Ma, Lawrence Livermore National Laboratory, 7000 East Avenue, L-103, Livermore, CA 94551, USA (ma21@llnl.gov)





**Early Online Release:** This preliminary version has been accepted for publication in *Journal of Climate*, may be fully cited, and has been assigned DOI 10.1175/JCLI-D-20-0338.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

## Abstract

The correspondence between mean sea surface temperature (SST) biases in retrospective seasonal forecasts (hindcasts) and long-term climate simulations from five global climate models is examined to diagnose the degree to which systematic SST biases develop on seasonal time scales. The hindcasts are from the North American Multi-Model Ensemble and the climate simulations are from the Coupled Model Intercomparison Project. The analysis suggests that most robust climatological SST biases begin to form within 6 months of a realistically initialized integration, although the growth rate varies with location, time, and model. In regions with large biases, interannual variability and ensemble spread is much smaller than the climatological bias. Additional ensemble hindcasts of the Community Earth System Model with a different initialization method suggest that initial conditions do matter for the initial bias growth, but the overall global bias patterns are similar after 6 months. A hindcast approach is more suitable to study biases over the tropics and sub-tropics than over the extra-tropics because of smaller initial biases and faster bias growth. The rapid emergence of SST biases makes it likely that fast processes with times scales shorter than the seasonal time scales in the atmosphere and upper ocean are responsible for a substantial part of the climatological SST biases. Studying the growth of biases may provide important clues to the causes and ultimately the amelioration of these biases. Further, initialized seasonal hindcasts can profitably be used in the development of high-resolution coupled ocean-atmosphere models.






Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

# 1 Introduction

As more than two-thirds of Earth's surface is covered by ocean, sea surface temperature (SST hereafter) plays a crucial role in the Earth system because it regulates ocean surface water and energy budgets and affects the weather and climate. However, recent generations of state-of-the-art Global Climate Models (GCMs) or Earth System Models (ESMs) have often suffered from similar systematic biases in SST (Mechoso et al. 1995; Lin 2007; De Szoeke and Xie 2008; Wang et al 2014; Richter 2015; Zhang et al. 2015; Zuidema et al. 2016; Lee et al. 2019). Here, we define the bias as the difference between model simulations and observations (or analyses). Figure 1 shows the annual multi-model mean SST biases from the historical simulations of the fifth and sixth phase of the Coupled Model Intercomparison Project (CMIP5, Taylor et al. 2012; CMIP6, Eyring et al. 2016). Large systematic cold biases are present over the equatorial Pacific, sub-tropical north and south Pacific Ocean, sub-tropical Indian Ocean, and sub-tropical and north Atlantic Ocean. Large warm biases are present over the eastern equatorial Pacific and Atlantic, southeastern Pacific and Atlantic, northeastern Pacific, northwestern Atlantic, and Southern Ocean. From CMIP5 to CMIP6, similar bias patterns with comparable bias magnitudes are still present in the latest generation of ESMs. These biases can hinder model prediction skills on seasonal to decadal time scales, and potentially impact the fidelity of simulated future temperature changes (e.g., Palmer et al., 2008; Richter et al., 2018).

Identifying the causes of SST biases from these long-term, fully coupled model simulations is challenging because a bias over a particular region may be due to either local oceanic or atmospheric processes, or both. A bias may also come from remote locations through teleconnections. There are also feedback processes from different component models (atmosphere, land and sea ice) involving different time scales. Furthermore, reducing a bias



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

by model tuning or changing model parameterizations may or may not resolve the root problem because of compensating biases. As a result, a similar SST bias may appear in future model versions. Therefore, it is critical to identify and understand the underlying causes of a model's bias.

While identifying causes of a particular regional bias is challenging from the long-term, fully coupled model simulations, significant progress has been made in recent years through diagnosing SST biases in realistically initialized seasonal or decadal coupled hindcasts (Huang et al. 2007, 2014; Liu et al. 2012; Hazeleger et al. 2013; Toniazzo and Woolnough 2014; Vannière et al. 2013, 2014; Voldoire et al. 2014, 2019; Sanchez-Gomez et al. 2016; Shonk et al. 2018; da Silveira et al. 2019; Siongco et al. 2020). With the Met Office Unified Model, there is a strong qualitative similarity in tropical and sub-tropical SST bias patterns between short-term forecasts (a few days to weeks) and long-term climate runs from the same model (Brown et al. 2012). The use of a seamless modeling approach to diagnose and correct initial bias growth can lead to improving long-term coupled model climate, particularly if biases in the short-term forecasts or hindcasts can be better understood (Hurrell et al. 2009; Martin et al. 2010). The initialized coupled model framework is a potentially powerful approach in that it can address: (a) over what time-scales do biases develop; (b) how much of the bias is due to atmospheric biases, oceanic biases, or biases from coupled ocean-atmospheric feedbacks; and (c) whether a given atmospheric or oceanic physical parameterization will be suitable as a core component of the model.

A number of studies have used this framework to diagnose the origin of tropical SST biases. For example, Vannière et al. (2013; 2014), Shonk et al. (2019) and Siongco et al. (2020) investigate the biases in the equatorial Pacific cold tongue and the double intertropical convergence zone (ITCZ) with seasonal hindcasts. The cold tongue bias is present within a



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

few weeks or months of lead time in all the models except one (IPSLCM5A-LR, Vannière et al. 2014), and the initial development of atmospheric surface wind bias, which drives the excessive equatorial upwelling, is the primary cause for the cold bias. Over the tropical eastern and southeastern Atlantic, the warm bias appears to be a combination of the excessive surface shortwave radiative forcing due to the insufficient low-cloud cover, weak upwelling resulting from the too weak equatorial trade winds, too weak coastal upwelling, and the deficiencies in the regional wind-SST-precipitation coupling, based on series of seasonal and decadal hindcast studies (Huang et al. 2007; Toniazzo and Woolnough 2014; Voldoire et al. 2014, 2019). A recent study by Hermanson et al., (2018) examined drifts in SST from the initial state with two seasonal forecast systems (Beijing Climate Center-Climate Prediction System and Met Office Global Seasonal forecast system version 5), and compared the drifts to the long-term bias in the free-running version of each model at eight selected locations. They found that SST drifts on seasonal time scales vary between the two forecasting systems at the selected locations with the former often showing larger mean forecast bias than the long-term bias, and the latter often showing the opposite relation between mean forecast bias and the long-term bias While these studies demonstrate the utility of the initialized coupled framework for model biases, it is not yet clear whether the hindcast approach is useful across the global oceans in a multi-model context.

To that end, we systematically diagnose in a multi-model context, the correspondences between short- and long-term SST biases around the global oceans, which has not been systematically addressed in previous studies. This is similar to the diagnosis performed on systematic biases in atmosphere models in Xie et al. (2012) and Ma et al. (2013, 2014). To do so, we analyze six sets of seasonal hindcasts and their corresponding long-term climate simulations from models participating in the North American Multi-Model Ensemble (NMME, Kirtman et al. 2014) project. Our objectives are to identify whether there is good



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

correspondence between short- and long-term SST biases over regions of large biases shown in Figure 1, and whether the correspondence is consistent across all the models analyzed in this study. Furthermore, we propose a set of criteria for identifying whether a hindcast approach is useful for a specific regional bias study. The remainder of this manuscript is organized into four sections. Section 2 describes the model experiments and validation datasets. Section 3 examines the correspondence between short- and long-term biases from the seasonal hindcasts and long-term fully coupled climate simulations. Section 4 summarizes our findings and draws conclusions.

# 2 Model experiments and validation datasets

## 2.1 Models and experiments

### 2.1.1 Ensemble seasonal hindcasts

Ensemble seasonal hindcasts are obtained from Phase-II of the NMME project (Kirtman et al. 2014). NMME is an intra-seasonal to seasonal to interannual multi-model ensemble prediction experiment for characterizing forecast uncertainty and improving predictability for operational needs. The project is coordinated by various U.S. and Canadian modeling centers. The NMME hindcasts are available from the Earth System Grid Federation (ESGF, https://www.earthsystemgrid.org/search.html?Project=NMME). Table 1 lists information about five GCMs that we selected for seasonal hindcasts from the NMME project; these five models were selected because long-term coupled climate simulations were available for analysis. The five models are the Coupled Climate Model versions 3 and 4 (CanCM3 and CanCM4, respectively, Merryfield et al. 2013) from the Canadian Centre for Climate Modeling and Analysis (CCCma); the Community Climate System Model version 4 (CCSM4, Gent et al. 2011) and Community Earth System Model version 1 (CESM1, Hurrell et al. 2013) from the National Center for Atmospheric Research (NCAR); the Forecast-



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

Oriented Low Ocean Resolution version of the Coupled Model version 2.5 (FLORB01, Vecchi et al. 2014) from the National Oceanic and Atmospheric Administration (NOAA) / Geophysical Fluid Dynamics Laboratory (GFDL). We also performed an additional set of ensemble seasonal hindcasts for the year 2005 with CESM1 using a different initialization method for all model components as part of the Lawrence Livermore National Laboratory Cloud-Associated Parameterizations Testbed (CAPT) project (Phillips et al. 2004, Ma et al. 2015). A description of the procedure of how these coupled hindcasts were performed is described in Appendix B. We refer to the CESM1 hindcasts from the NMME project as CESM1-NMME and hindcasts from the CAPT as CESM1-CAPT hereafter.

Figure 2 illustrates the hindcast procedure for these NMME models. For each modeling group, a 10-member ensemble of 12-month long hindcasts were performed starting at 00Z on the first day of each month between January 1980 and December 2014. For the hindcast month 1 (Mon1 or the 0-month lead), SSTs are averaged for the first month of the hindcasts over all the ensemble members. SSTs of the hindcast month 2 (Mon2 or the 1-month lead) are averaged for the second month and so on for hindcast month 3 to month 12 (Mon3 to Mon12, or 2-month to 11-month lead).

Initialization and ensemble generation procedures differ among modeling centers as the NMME project did not ask for its synchronization. For CanCM3, CanCM4, and FLORB01, initial conditions are from their own data assimilation systems (Merryfield et al. 2013 for CanCM3 and CanCM4; Zhang et al. 2007 for FLORB01). For CCSM4, initial conditions are taken from the National Centers for Environmental Prediction (NCEP) Climate Forecast System Reanalysis (CFSR). For CESM1-NMME, the ocean and sea ice initial conditions are from a Coordinated Ocean-Ice Reference Experiment (CORE, Griffies et al 2009) and the atmosphere and land models are initialized from a long-term, spun-up climatology. We assume that these initialization procedures provide model initial conditions



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

close to observed. For CESM-CAPT, the procedure of generating initial conditions for atmospheric and land initial conditions of CESM1 is described in Ma et al. (2015). The ocean initial conditions were generated by applying a forty-eight-member ensemble adjustment Kalman filter data assimilation system (Karspeck et al. 2013) from the Data Assimilations Research Testbed (DART, Anderson et al. 2009) at NCAR (see Appendix B). In the case of SST, we will test below how close each model's initial condition is to the observations.

### 2.1.2 Long-term climatological coupled simulations

To obtain long-term climatological SST mean biases, we also used the corresponding historical simulations from Phase 5 of the Coupled Model Intercomparison Project (CMIP5, Taylor et al. 2012). The CMIP5 historical simulations for CCSM4 (six ensemble members) and CESM1 (three ensemble members) cover the years from 1850 to 2005. The CMIP5 historical simulations for CanCM4 (ten ensemble members) are only available from 1961 to 2005. Climatological SST of CanCM3 and FLORB01 with the same models was obtained directly from the CCCma and NOAA/GFDL modeling groups, respectively. For CanCM3, SST is taken from a 40-year long coupled climate simulation from 1971 to 2010 using historical forcings (Merryfield et al. 2013). For FLORB01 model, SST is taken from the last 300 years of a 1500-year long coupled climate simulation using radiative forcing and land-use conditions representative of the year 1990 (Vecchi et al. 2014, Murakami et al. 2015). We believe that the SST mean biases from these simulations should be very representative of the climatological mean biases.

## 2.2 Validation datasets

Global analyses of monthly SST are from the Met Office Hadley Centre's SST data set (HadISST, Rayner et al. 2003, https://www.metoffice.gov.uk/hadobs/hadisst/). This dataset is on a 1° latitude × 1° longitude horizontal grid from 1870 to present. To assess SST



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

uncertainty from different analysis products, we also compared our hindcasts to the National Oceanic and Atmospheric Administration (NOAA) Optimum Interpolation SST Version 2 (OISST, https://www.esrl.noaa.gov/psd/). The OISST has two versions, a monthly resolution version on a 1° latitude × 1° longitude horizontal grid (Reynolds et al. 2002) and a daily resolution version on a 0.25° latitude × 0.25° longitude horizontal grid (Reynolds et al. 2007). Both versions cover from 1981 to present. The overall annual mean bias patterns in SST from the hindcasts are very similar with HadISST and OISST (not shown here). The differences in the bias magnitude of using different SST analyses are generally small (< 1º C) in the low latitudes. There are some large differences (~3º C) in the bias magnitude over high latitudes, especially near the storm tracks or sea ice. Nevertheless, most systematic SST biases are much larger than the differences from different SST datasets so the results we present in the later sections are not affected by the choice of either observed SST dataset. We decided to use the HadISST for most of our analysis as it covers the much longer historical period from 1870 to present. The daily OISST is only used to determine biases in the initial conditions for the hindcasts. Observed and modeled SSTs are linearly interpolated to a resolution of 1° longitude by 1° latitude for comparison.

# 3 Correspondence between short- and long-term systematic SST biases

## 3.1 Biases in the initial SST

Our primary goal is to identify regions where large SST biases in the long-term climatological simulations appear within the first few months of seasonal hindcasts. One can then use initialized seasonal hindcasts to diagnose SST bias growth over those particular regions. The ideal scenario for doing this is when the initial conditions, especially the upper ocean state, are as close to observations as possible. If the initial SST bias is already large compared to a model's climatological SST bias, attribution of SST bias to certain processes



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

would be challenging[1]. Therefore, an important first step is to check whether there are large biases in the initial state.

To examine the magnitude of initial biases in SST, Figure 3 shows the SST ensemble mean biases averaged over the first day of model integrations with the starting dates of January 1, 2005 and July 1, 2005 (Results are not shown for CESM1-NMME and FLORB01 because daily SSTs were not available). Table 2 lists the root mean square errors (RMSEs) of SST calculated over the tropics and sub-tropics (0°–360°E, 30°S–30°N), as well as extra-tropics (0°–360°E, 30°N–60°N or 60°S–30°S) for the same starting dates. We examined 1-day biases because the initial conditions for ocean models are not available for analysis from the NMME project. Since the ocean heat capacity is large, we assume that large biases in the SST after one day of integration are representative of biases in the initial state for all the models. We have verified that this is the case for CESM-CAPT as the differences in SST between the initial state and 1-day hindcast are generally less than 0.1°C.

The day-1 tropical SST biases are in general smaller than the extra-tropical SST biases, regardless of season or hemisphere. SST biases are within ±0.5°C in most tropical and sub-tropical oceans for both starting dates, especially in CCSM4 (Figure 3). For the starting date of January 1, 2005, CanCM3, CanCM4, and CESM1-CAPT show warm biases with magnitude of ~1-2°C over the Indian Ocean and subtropical north Atlantic. For the starting date of July 1, 2005, these three models show cold biases with magnitudes of ~1-2°C over the Indian Ocean and sub-tropical north Atlantic. There are also warm biases with magnitude of ~1-3°C near the coast of Peru, Chile, Angola and Namibia. In the extra-tropics, large biases with magnitude of ~3-4°C are mostly found over the storm tracks in the North Pacific, North Atlantic and Southern Ocean. This is consistent with the RMSEs in Table 2. The RMSEs are

[1] Biases in the initial conditions may provide information if a model was using its own data assimilation system.



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

smaller than 0.5°C between 30°S and 30°N for both starting dates. Slightly larger RMSEs are found in mid-latitudes with CanCM3, CanCM4, and CESM1-CAPT showing ~1.1-1.3°C in magnitudes in northern hemisphere, and ~0.7-1.2°C in southern hemisphere in both January and July. Large biases in the initial states may come from the imperfect data assimilation systems or initialization procedures, imperfect model physics, or insufficient observations to assimilate.

## 3.2 Bias correspondence in SST mean state

To evaluate the bias correspondence between seasonal hindcasts and long-term climate biases over the global oceans across all the models, we first present in Figures 4 and 5 the ensemble annual mean biases for SST at selected hindcast lead times along with the SST annual mean biases from the corresponding long-term climatological simulations. This is to identify regions where large SST biases in the long-term climatological simulations appear within the first few months of seasonal hindcasts, and to examine how similar the bias pattern and magnitude in the hindcasts are to the long-term climatology. All the calculations in the rest of the text were done with all ensemble members and all hindcast years listed in Table 1 with the corresponding observed SST unless otherwise noted.

The biases present at Mon1 show little resemblance to the systematic biases shown in Figure 1 or even the climatological bias of the same model (Figure 5, right column), and they are considerably smaller than the biases at longer lead times. We start to see the growth of SST bias magnitude and extent for hindcasts in all models between Mon2 and Mon4 in the tropics and sub-tropics, such as the equatorial Pacific cold tongue bias (except for CCSM4 and FLORB01), warm biases over the coastal southeastern Pacific and Atlantic, and cold biases over the sub-tropical Pacific and Atlantic in both hemispheres (except CESM1 southern hemisphere). From Mon6 to Mon12 for any model, the bias pattern and magnitudes are very similar indicating the saturation of initial growth of bias in SST after Mon6. When



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

comparing hindcast biases at Mon6 or later hindcast lead times to climate biases (Figure 5) for any given model, we observe a similar bias pattern of SST to the corresponding climatological bias pattern globally as we will discuss in detail later using Taylor diagrams (Figure 7). For CESM1-CAPT, there is only one year (2005) for analysis. Nevertheless, the bias pattern looks very similar to that of CESM1-NMME. This suggests that one year of hindcasts with enough ensemble members would be enough to exhibit the mean bias pattern of the long-term climatology, especially over the tropics and sub-tropics. We will have further discussion in Section 3.3.

The bias magnitudes, however, are quite variable across models. Figure 6 shows the RMSE of annual mean SST calculated over the tropics and sub-tropics, as well as extra-tropics from hindcasts with different hindcast lead times and the long-term climatology. We find that the RMSEs in the tropics and sub-tropics are generally smaller than those in the extra-tropics for all the models. Compared to the RMSEs in the tropics and sub-tropics, the hindcast RMSEs in the northern mid-latitudes for all the models are ~ 0.2 – 1°C larger. The hindcast RMSEs in the southern mid-latitudes are also ~ 0.2 – 0.5°C larger for CanCM4, CCSM4, CESM1-NMME, and CESM-CAPT, while CanCM3 and FLORB01 show comparable RMSEs. The Mon12 bias magnitudes for all the models are ~ 80 – 105% the size of climatological bias except for FLORB01 in the southern mid-latitudes where much larger warm biases are present in the long-term climatology.

To quantitively demonstrate the bias correspondence, Figure 7 shows the pattern statistics of spatial correlations and normalized spatial standard deviations between the average hindcast bias and the long-term climatological bias on a Taylor diagram. Unlike the canonical Taylor diagram that uses the observation field as the reference field, the reference SST field for each model in the diagrams is the annual mean bias from its corresponding long-term climatology with respect to HadISST. The analysis is separated into four domains with



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

one covering nearly the entire global oceans (60°S - 60°N), one covering the tropical and sub-tropical oceans (30°S - 30°N), and another two for mid-latitude oceans in both hemispheres (30°N - 60°N and 30°S - 60°S). For most lead times, all four domains have spatial standard deviations of the hindcasts (i.e. bias magnitudes) that are smaller than their long-term climatological counterparts. Standard deviations also increase with hindcast lead time to values close to that of the climatological bias indicating the growth of average bias magnitude to near full amplitude by the end of Mon12. The standard deviations in the low-latitudes (30°S to 30°N) are much closer to the climatological values than those in the mid-latitudes (30°N to 60°N and 60°S to 30°S), which are about 70% of the climatological values.

While it is subject to expert judgement and it can vary from one to another, we consider that small initial SST biases in the hindcasts, high bias correlation coefficients (~ >0.6) by Mon6, and comparable spatial standard deviation (~ 1 normalized spatial standard deviation) or bias magnitude constitutes a good bias correspondence in our global analysis. For bias correlation globally (Figure 7a), all six experiments for nearly all lead times from the five models have correlations larger than 0.4 with their corresponding long-term climate biases. CanCM3, CanCM4, CESM1-NMME, CESM1-CAPT and FLORB01 have bias correlations larger than 0.8 by Mon6. The low correlations in CCSM4 hindcasts in Figure 7a are mostly due to the large warm bias between 40°-60°N over the north Atlantic Ocean where a cold bias is present in the long-term climatology (see Figure 5). This causes the very small (less than 0.3) or even negative correlations (at later lead times) in the northern mid-latitudes (Figure 7c). The bias correlations are much larger (> 0.8) for CCSM4 and other models if the analysis domain is only the tropics and sub-tropics (Figure 7b). This is consistent with all the bias attribution studies mentioned in the introduction, which focused on regions within this domain. It is also worth noting that bias correlations in all hindcasts increase with lead time (except for CCSM4 over northern mid-latitudes), suggesting the biases gradually evolve



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

toward the long-term climate biases with lead time. Other than CCSM4, all other models also show bias correlations between 0.7 and 0.95 in mid-latitudes in both hemispheres.

The bias correlations and standard deviations do not change much after Mon7. With the correlation coefficients at 7 months typically around 0.8 in the tropics and sub-tropics, this suggests that a large portion of climatological bias pattern is associated with fast upper ocean and atmospheric processes given the relatively small SST biases in the initial conditions. Here, we define fast processes as processes in the atmosphere or ocean with time scales shorter than the seasonal time scales. Examples include the wind-driven Ekman upwelling over the equatorial Pacific which occurs at time scales of about one month (Nigam and Chao, 1996; Neelin 1996), or the surface heat flux and heat storage balance in mid-latitudes at seasonal time scales (Wang and Carton, 2002). For specific regions, the fast response of the equatorial Pacific cold tongue biases is mainly associated with biases in the atmospheric surface wind stress (e.g., Shonk et al. 2019; Siongco et al. 2020), and the fast response of the tropical eastern and southeastern Atlantic warm biases is mainly associated with biases in the low clouds, surface shortwave radiation and ocean upwelling (e.g., Huang et al. 2007; Toniazzo and Woolnough 2014; Voldoire et al. 2014, 2019). For the extra-tropics in both hemispheres, the correlation coefficients at 7 months are typically ~ 0.5-0.6, this suggests that the climatological bias pattern is also influenced by teleconnections and feedbacks over a longer timescale. Two examples are that the cold SST biases over the mid-latitude Atlantic Ocean may be affected by the biases in the Atlantic meridional overturning circulation, while the warm SST biases over the Southern Ocean may be affected by slower vertical mixing due to a deeper mixed layer. Both processes have time scales much longer than a season.

While we examined the mean biases of ensemble seasonal hindcasts over a thirty-year period, one important question to ask is whether these SST biases are robust and whether there is large interannual variability in the mean biases. To explore this, we can examine the



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

magnitude of the ensemble spread of SST in the hindcasts. Figure 8 and 9 shows the ensemble monthly mean SST biases for January and July 2005 from the hindcasts at Mon6 (i.e. from the hindcasts starting on August 1, 2004 and February 1, 2005, respectively). Also shown are the standard deviations of the ensemble mean biases at Mon6 and the ratios of the mean biases to the standard deviations. While the procedures for generating the ensembles in each of the models are different, larger values of standard deviations ~1.2-1.6 ºC are only found over the mid-latitudes and within a narrow region along the equator. The large standard deviations in these regions are consistent with the active upper ocean dynamics in the mid-latitudes and equatorial wave guide. It is also clear that the standard deviations are larger in mid-latitudes associated with storm tracks. However, for many other regions of large SST mean biases shown in Figures 1 and 5, such as cold biases over the equatorial Pacific, sub-tropical north and south Pacific Ocean, sub-tropical Indian Ocean, sub-tropical and north Atlantic Ocean, or warm biases over the eastern equatorial Pacific and Atlantic, southeastern Pacific and Atlantic, Northeastern Pacific, and northwestern Atlantic, the ratios of mean bias to standard deviation are also larger than one. This suggests that the mean biases over these regions are robust.

To verify whether there is large interannual variability in the mean biases, we examine in Figure 10 the interannual standard deviations from the annual mean bias of SST at Mon12, and the ratios of the annual mean bias of SST to the interannual standard deviation of SST. The average SST bias from all hindcasts within a given year were first calculated, and then at each point the standard deviation across years in the hindcast period (Table 1) were calculated. The largest interannual standard deviations are over the Equatorial Pacific associated with the El Niño-Southern Oscillation (ENSO), and over the mid-latitude oceans particularly near the Northwest Atlantic to the north of the Gulf Stream separation point. Comparing the ratios of mean bias to the standard deviations (right column of Figure 10) with



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

the Mon12 biases in Figure 5 shows that regions of large SST biases generally show ratios larger than 2 with some regions larger than 10. This suggests that the average of an ensemble of hindcasts from a single year can be representative of the long-term bias in most places with the exception of the Equatorial Pacific and the northwest Atlantic where bias magnitude and extent may depend on the year.

## 3.3 Impact of initialization procedure on SST biases

As we have observed in Figures 4 and 7, the bias pattern in CESM1-NMME looks similar to that of CESM1-CAPT although the bias magnitudes vary between the two sets of hindcasts. While the reasons for such differences mainly come from the simulated years in the composites and the initialization procedures (see Section 2.1 and Appendix B), we can further investigate the impact of the latter by examining the hindcasts from the same year for both runs.

As we will demonstrate shortly, the initial conditions do matter for the initial bias growth, but the overall global bias patterns are similar after 6 months. The annual mean SST biases at Mon1, Mon6, and Mon12 for 2005 hindcasts only are shown in Figure 11. The choice of the year is due to the availability of initial conditions from the ocean for CESM1-CAPT, and limited by computational resources. We also present in Table 3 the SST bias correlation coefficients between the two simulations from CESM1, which allows for quantitative comparison of bias pattern. We further separate the calculations into tropics/sub-tropics, and extra-tropics. The correlations in Table 3 indicate whether SST bias patterns from CESM-NMME and CESM-CAPT for the year of 2005 are similar to each other. Table 3 also indicates whether SST bias pattern of one-year hindcasts from either NMME or CAPT procedure would be representative to that from the full 30 years of hindcasts from CESM1-NMME. At Mon1, there are substantially larger biases (by up to 2 ºC) in CESM1-NMME in the northern mid-latitudes (Figure 11), which are likely due to differences in initialization



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

procedures. This is mainly because DART will adjust ocean states closer to observations especially for the upper ocean where the quality of observations is better, whereas the CORE forced ocean does not adjust ocean states directly. Nevertheless, the bias correlations between the NMME and CAPT is still large (0.65 in northern mid-latitudes) as indicated in Table 3. At Mon6 and Mon12, the bias patterns in the tropics and sub-tropics qualitatively look even more similar to each other with bias correlation of 0.84 and 0.78 respectively. The hindcast bias patterns also look similar to the model's climatological bias with correlations of 0.88 and 0.89 for CAPT, and 0.81 and 0.73 for NMME (Figure 5). There are, however, still large differences in bias magnitude in mid-latitudes, especially in the Southern Ocean, where CESM1-NMME shows more cold biases and CESM1-CAPT shows more warm biases. Other than the initialization procedures, some of the differences in the SST mean biases in the mid-latitudes may also come from natural variability of the atmospheric synoptic waves or the ocean circulation. It is, however, difficult to quantify their relative impact. Nevertheless, the bias magnitudes are sensitive to the initialization procedure, while the SST bias pattern is less sensitive, especially over tropics and sub-tropics.

The bias correlation is ~0.9 or higher between NMME-2005 and NMME-all (30-year, Table 3) for any hindcast lead times in any latitude bands. This further suggests that the SST bias pattern does not vary too much from year to year. While comparing the bias correlations between hindcasts and the long-term climatology (the last 9 rows of Table 3), one single year of hindcasts does show high bias correlations of 0.8 at Mon6 over low latitudes and northern mid-latitudes (0.1 or 0.2 smaller in southern mid-latitudes). It is not surprising that the bias correlations increase by ~0.1 to 0.2 when all the NMME years are compared to the long-term climatological biases. Nevertheless, to diagnose systematic SST biases using initialized coupled hindcast approach, one year of coupled hindcasts can provide substantial and robust information.



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

## 3.4 Applicability of an initialized hindcast approach for diagnosing regional SST biases

Our earlier analysis focuses on the bias correspondence over tropical, sub-tropical and extra-tropical domains in both hemispheres We further explore whether an initialized coupled hindcast approach would be suitable for specific regional bias diagnosis and attribution studies, such as those regions with large systematic climate biases identified in Figure 1. We will also present a quantitative set of criteria to identify such a region.

To measure how close the SST mean bias magnitudes in the Mon12 hindcasts are to their climatological counterpart, we present in Figure 12 the ratio of SST annual mean bias for Mon12 hindcasts to their corresponding climatological annual mean bias (i.e., bias ratio). Positive values of the ratio indicate the hindcast and climatology share the same bias sign, with value closer to 1 being better correspondence. Ratios that are negative or far from 1 suggest longer time scale feedback processes at work. We also calculated the RMSE of annual mean SST from hindcasts with different lead times and long-term climatology over ten selected regions (Figure 13). The RMSEs from different hindcast lead times indicate the growth of bias magnitude over that particular region. Here, we use RMSEs rather than the mean bias because using regional mean bias may be misleading. If both positive and negative biases co-exist in the same region, the area mean bias could be small while the RMSE could be large. This is the case in some of our regions. As we will demonstrate, the rapid emergence of SST biases in the hindcasts with comparable bias magnitudes to their climate biases in most regions over the tropics and sub-tropics make them good candidates for using an initialized coupled hindcast approach.

A coupled hindcast approach would be useful for the equatorial Pacific cold tongue bias (EQ Pacific, 150-260°E, 2°S -2°N). All models except CCSM4 show a positive bias ratio of ~ 0.6–2 (Figure 12). The CCSM4 actually shows a negative bias ratio because CCSM4





Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

does not simulate a cold bias over this region in the hindcasts while there is a weak cold SST bias in the long-term climatology. We also see smaller RMSEs for other models at Mon1 and a growth of SST RMSEs from the first month (Fig. 13a).

For the systematic warm biases over the sub-tropical northeastern (NE Pacific, 110-130°W, 20-30°N) and southeastern Pacific (SE Pacific, 70-90°W, 10-25°S), and sub-tropical southeastern Atlantic (SE Atlantic, 0-15°E, 5-25°S), the bias extent and magnitude are quite variable across the models (Figure 4 and 5). For example, CESM1 shows warm biases confined to the coasts of California and Peru, while FLORB01 shows a much broader extent of warm biases in all three regions in both hindcasts and long-term climatology. For models with warm biases over these three regions, we see a bias ratio between ~0.6 and 2 (Figure 12). The RMSEs in most models are already greater than 1ºC at Mon1, and we also see a growth of RMSEs in the hindcasts for all these regions although the magnitude of RMSEs do not change much after 6-Mon lead (Figures 13b, 13c and 13d). There are large initial warm biases in some of the models (CanCM3, CAM4, and CESM1-CAPT) near the coastal regions that spread westward from the ocean boundaries (Figure 3). Therefore, it is better to diagnose the causes of warm biases over regions away from the coasts if better initial conditions cannot be obtained.

For regions over the sub-tropical Pacific and Atlantic oceans in both hemispheres (N Pacific, 160-210°E, 20-35°N; S Pacific, 130-170°W, N Atlantic, 15-25°S; 30-60°W, S Atlantic, 15-25°N; 15-40°W, 15-25°S), not all models show large cold bias extent (Figures 4 and 5) as suggested in the multi-model mean (Figure 1). The cold biases are small or not significant in the CCSM4 and FLORB01 climatology in the southern hemisphere although the hindcasts do simulate cold biases over these regions. The bias ratios are also quite variable for these regions with values between ~0.4 and 2 (Figure 12), mostly due to the different extent of cold bias in each model between hindcasts and long-term climatology. The cold biases in



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

these regions are less severe with smaller RMSEs (< 1.5 ºC for most hindcasts lead times, Figures 13e, 13f, 13g, and 13h) compared to the previous four regions. Nevertheless, these regions are also suitable for using hindcast approach for bias diagnosis.

For the cold bias over the mid-latitude north Atlantic (Mid-Lat N Atlantic, 20-45°W, 40-55°N), all models except CCSM4 show a positive bias ratio between ~0.4 and 2. CanCM3, CanCM4, and FLORB01 all show a growth of RMSEs with hindcast lead time. For CESM1-NNME or CESM-CAPT, the RMSEs are relatively constant from Mon1. This is mostly because the cold biases already exist in the initial states. We also see an initial cold bias in CanCM3, and CanCM4 with magnitude of ~1-2°C from Figure 3. Even though we see the growth of biases in some of the models, the large biases in the initial states would make the bias attribution difficult.

For SST biases over the Southern Ocean (0-120°E, 60-45°S), regional SST variability is quite large due to the large natural variability in the atmospheric surface winds associated with storm tracks and their interactions with the upper ocean circulation and currents. Only FLORB01 shows a warm bias in most Southern Ocean regions in both hindcasts and long-term climatology (see Figures 4 and 5). Other models simulate both warm and cold biases. Although the bias correlations between the hindcasts and long-term climatology in Figure 7 show an average of 0.5-0.6 for the Southern Ocean (0°–360°E, 60°S–30°S) and the bias ratios in Figure 12 are closer to 1 for some regions in the Southern Ocean, there are still large initial biases (Figure 3). This also makes using a coupled hindcast approach to diagnose SST bias in this region challenging.

To quantitively examine whether an initialized coupled hindcast approach would be suitable for a certain regional bias diagnosis, we propose the following criteria: (1) the RMSE of the climatological SST is > 0.5 ºC, (2) the initial SST RMSE is < 0.5 ºC, (3) the RMSE magnitude at Mon12 is at least 60% of the climatological RMSE (indicating a growth of SST



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

bias), and (4) the Mon12 hindcast and the climatological bias have the same bias sign. As we have pre-selected ten regions of large SST bias (Figure 1), most regions in the models satisfy the first criterion. For the second criterion, we used the Mon1 RMSEs instead of the initial SST RMSEs because the latter are not available for every model. We understand that using Mon1 RMSEs is not ideal and can affect the results. Nevertheless, the regional RMSEs (or the mean biases) at Mon1 are still small in many regions as shown in Figure 13 (or Figure 4). Table 4 summarizes the suitability for these 10 regions. Consistent with our analysis in Section 3.2 and here, regions in the tropics and sub-tropics are generally suitable for using an initialized coupled hindcast approach to diagnose SST bias growth as long as biases in the initial states are small. Only CCSM4 is suitable for warm bias study over the sub-tropical northeastern and southeastern Pacific, and sub-tropical southeastern Atlantic, because large Mon1 bias near the coastal regions are present in other models. Over the extra-tropics, using a hindcast approach is more challenging for all the models given large initial biases, large natural variability (Figures 8-10), and possible bias contribution from slow atmosphere or ocean feedback processes indicated by the relatively low bias correlations (Figure 7).

Comparing our findings here and those in Hermanson et al., (2018) for the common regions (EQ Pacific, SE Atlantic, Mid-Lat N Atlantic, and Southern Ocean), one model in Hermanson et al., (2018) could be used to study initial growth of SST biases in EQ Pacific, SE Atlantic and Mid-Lat N Atlantic based on the criteria proposed here and in their study. The other model in Hermanson et al., (2018) generally shows the opposite signs between mean forecast bias and the long-term bias on the seasonal time scales, which makes an initialized coupled hindcast approach not suitable for studying long-term SST biases in these regions. Note that Hermanson et al., (2018) only examined ensemble hindcasts with starting dates from May 1st and November 1st of each year. The results may change if more starting dates were used in their study.



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

## 4 Summary and discussion

In this study, we examine the SST bias correspondence between ensemble mean seasonal hindcasts and long-term climate simulations from five GCMs to diagnose over what time scales the systematic SST biases develop, and to identify regions where an initialized coupled hindcast approach would be useful for bias diagnosis and attribution. We also propose a set of criteria for the latter. The seasonal hindcasts are from the NMME Project and the long-term climate simulations are from the CMIP5 experiments (CanCM3 from a 40-year long historical coupled simulation and FLORB01 from a 300-year long coupled climate simulation). An additional set of hindcast experiments are performed with CESM1 using a different initialization approach to further diagnose the impact of initialization procedure on the simulated SST biases in the hindcasts.

Our analysis suggests that most robust climatological SST mean biases shown in Figure 1 form within 6 months of the hindcasts. The bias patterns and magnitudes do not change much after Mon7. With the correlation coefficients at 7 months typically around 0.8 in the tropics and sub-tropics, this suggests that a significant portion of the climatological bias pattern is associated with fast upper ocean and atmospheric processes given the relatively small SST biases in the initial conditions. For the extra-tropics in both hemispheres, the correlation coefficients at 7 months are typically ~ 0.5-0.6, suggesting that the climatological bias pattern is less associated with fast upper ocean and atmospheric processes.

The ensemble spread and the interannual variability in the SST mean bias are much smaller than the mean bias over regions with large SST biases, which suggests the robustness of these systematic biases in the hindcasts. Comparing hindcasts from CESM1-NMME and CESM1-CAPT suggests that initialization procedures do matter for bias magnitude and pattern in the first few months of a hindcast. Nevertheless, the overall global mean bias patterns are still similar to the climatological bias patterns by six months.



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

We further investigated ten selected locations with large systematic SST climate biases and identify whether an initialized coupled hindcast approach is suitable for bias diagnosis and attribution. The rapid emergence of SST biases in the hindcasts with comparable bias magnitudes to their climate biases in most regions over the tropics and sub-tropics, such as the cold bias over the equatorial Pacific, warm biases over the subtropical eastern oceans, and cold biases over the subtropical Pacific and Atlantic in both hemispheres, make them good candidates for using an initialized coupled hindcast approach. Over the extra-tropics, such as the Southern Ocean or north Atlantic, a hindcast approach is more challenging given large initial biases, large natural variability, and possible bias contribution from slow atmosphere or ocean feedback processes indicated by the lower bias correlations (~0.5). Studying the growth of biases using initialized hindcasts over regions with good bias correspondence and smaller biases in the initial states can provide important clues to the causes and ultimately the amelioration of these systematic biases. This was demonstrated in many recent studies using initialized seasonal or decadal hindcasts (Huang et al. 2007, 2014; Liu et al. 2012; Hazeleger et al. 2013; Toniazzo and Woolnough 2014; Vannière et al. 2013, 2014; Voldoire et al. 2014, 2019; Sanchez-Gomez et al. 2016; Shonk et al. 2019; Siongco et al. 2020). While the hindcast approach is useful in diagnosing fast growing biases within the hindcast time scales, it is an approach to complement the CMIP-type long-term coupled experiments. In many cases, to further understand the causes of a particular bias usually requires additional hypothesis-testing experiments, either in the hindcast mode or long-term climate mode.

In addition to diagnosing biases in the fully coupled models, another application for the short-duration coupled hindcasts is the development of new model parameterizations, especially those relevant to atmospheric moist processes. During the development phase of a scheme, it is usually tested in a stand-alone model component (e.g., atmospheric-only model



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

configuration). When the scheme is ready to be tested in the fully coupled configuration, it is usually challenging to assess the impact of the new scheme from the coupled feedbacks in the long-term climatological simulations. The initial drift of the model state in the ocean from the hindcasts, however, can help determine the impacts of the new scheme as most drift in the ocean is likely connected to the atmospheric moist processes, such as convection (Song and Zhang 2018). Another benefit of using coupled model hindcasts is that one can perform hindcasts for more recent years when more observations are available, especially in the ocean.

Finally, as it becomes more common for modeling centers and groups to use high-resolution coupled models to simulate more processes in detail and on a smaller spatial scale, initialized coupled model hindcasts can be used to greatly reduce computation costs for model evaluation (e.g., da Silveira et al. 2019), and possibly tuning, which is usually challenging to do for very high-resolution coupled GCMs.

## Acknowledgements

HM, AS, SK, SX, JL were funded by the Regional and Global Model Analysis program area (RGMA) and Atmospheric System Research (ASR) program of the U.S. Department of Energy and their work was performed under the auspices of the U.S. DOE by LLNL under contract DE-AC52-07NA27344. Computing resources and simulations archive were provided from the Livermore Computing Center at LLNL and the National Energy Research Scientific Computing Center (NERSC), contract number DE-AC02-05CH11231.

## Data Availability Statement

The simulations will be available online through the NERSC Science Gateways (details provided on https://docs.nersc.gov/services/science-gateways/). Due to the large volume of datasets and limited disk space, raw data will be shared online upon request. Post-processed data and plotting scripts for figures in this manuscript are available over https://portal.nersc.gov/archive/home/h/hyma/www/CAPT/NMME/.



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

## Appendix A: List of CMIP5 and CMIP6 models

In Figure 1, the multi-model ensemble mean SST is computed from twenty-five (25) CMIP5 models (ACCESS1.3, BCC-CSM1.1, CanCM4, CanESM2, CCSM4, CESM1-CAM5, CMCC-CM, CNRM-CM5, CSIRO-Mk3.6.0, FGOALS-g2, GFDL-CM2p1, GFDL-CM3, GISS-E2-H, GISS-E2-R, HadCM3, HadGEM2-ES, INMCM4, IPSL-CM5A-LR, IPSL-CM5A-MR, MIROC5, MIROC-ESM, MPI-ESM-LR, MPI-ESM-MR, NorESM1-ME, NorESM1-M), and thirty-four (34) CMIP6 models (ACCESS-CM2, ACCESS-ESM1-5, BCC-CSM2-MR, BCC-ESM1, CAMS-CSM1.0, CanESM5, CESM2, CESM2-FV2, CESM2-WACCM, CESM2-WACCM-FV2, CIESM, CNRM-CM6.1, CNRM-ESM2.1, E3SM-1-0, E3SM-1-1, EC-Earth3, EC-Earth3-Veg, FIO-ESM-2-0, GFDL-CM4, GFDL-ESM4, GISS-E2-1-G, GISS-E2-1-H, HadGEM3-GC31-LL, INM-CM4-8, IPSL-CM6A-LR, MCM-UA-1-0, MIROC-ES2L, MIROC6, MPI-ESM-1-2-HAM, MRI-ESM2-0, NESM3, NorCPM1, SAM0-UNICON, UKESM1-0-LL). The SSTs are averaged over the historical period, 1850-2005 for CMIP5 and 1850-2014 for CMIP6. SST biases are calculated using HadISST as observational reference, covering the period from 1870 to present. Statistical significance was calculated with t-tests using yearly, multi-model data and yearly observations.

## Appendix B: Initialization procedure for the Coupled Cloud-Associated Parameterizations Testbed

The Cloud-Associated Parameterizations Testbed (CAPT) is a joint project between the U.S. Department of Energy (DOE) Lawrence Livermore National Laboratory (LLNL) and National Center for Atmospheric Research (NCAR), designed to diagnose and improve representation of cloud-associated physical processes in climate models by applying a weather forecast technique to climate models (e.g., Phillips et al. 2004; Williams et al. 2013; Ma et al. 2015). As there are no real-time operational constraints, the CAPT is conducted



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

using retrospective forecasts (hindcasts) for which the verifying observations are already available.

The current procedure for obtaining CAPT initial conditions for atmospheric and land initial conditions of CESM1 is described in Ma et al. (2015). In short, CAPT has primarily used initial conditions for the atmosphere that are generated at operational numerical weather prediction (NWP) centers like the National Centers for Environmental Prediction (NCEP) and the European Centre for Medium Range Weather Forecasting (ECMWF). In the present study, atmospheric state variables (horizontal velocity, specific humidity, and temperatures) are from the ECMWF ERA-Interim reanalysis (Dee et al. 2011). A nudging simulation was performed first to acquire necessary variables (e.g., cloud and aerosol fields), which are not available from the reanalysis for the atmospheric initial conditions. Land initial conditions are taken from an offline land model simulation forced by reanalysis and observations including precipitation, surface winds, and surface radiative fluxes.

Initial conditions for the ocean model (POP2) were provided by the Data Assimilation Research Testbed (DART, Anderson et al. 2009; Karspeck et al. 2013) at NCAR. The ocean reanalysis/initial conditions are generated by applying a forty-eight-member ensemble adjustment Kalman filter (EAKF) data assimilation system with POP2 (Karspeck et al. 2013). Observations of subsurface temperature and salinity from the World Ocean Database 2009 (Johnson et al. 2009) are assimilated into the ocean model at a daily frequency from 1998 to 2005. The atmospheric forcing for the ocean model comes from an independently generated EAKF analysis with the Community Atmosphere Model version 4 (Raeder et al. 2012) forced by SST and sea ice prescribed from the NOAA Optimally Interpolated SST version 2 (OISST, Reynolds et al. (2002). Since the sea ice initial conditions were not available from the POP2-DART reanalysis, we simply used the observed sea ice concentration from the NOAA Optimally interpolated sea ice product



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

(https://www.esrl.noaa.gov/psd/data/gridded/data.noaa.oisst.v2.html). If one's focus is on the tropical and sub-tropical SST on seasonal time scales, the impact from the sea ice is expected to be minimal and we have verified this.

With the initial conditions from the above procedures, 12-month long ensemble coupled hindcasts started every first date of each month at 00Z between February 1, 2004 and December 1, 2005 were performed. For each start date, 24 ensemble members are generated, based on the ensemble initial conditions from the POP2-DART system. There is only one set of initial conditions for other model components.

## References


Anderson, J. L., T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. O. Avellano, 2009: The data assimilation research testbed: A community facility. *Bulletin of the American Meteorological Society*, **90**, 1283–1296.

Brown, A., S. Milton, M. Cullen, B. Golding, J. Mitchell, and A. Shelly, 2012: Unified modeling and prediction of weather and climate. *Bull. Amer. Meteor. Soc.*, **93**, 1865–1877.

De Szoeke, S. P. and Xie, S.-P., 2008: The Tropical Eastern Pacific Seasonal Cycle: Assessment of Errors and Mechanisms In IPCC AR4 Coupled Ocean–Atmosphere General Circulation Models. *Journal of Climate*, **21**(11):2573–2590.

Dee, D. P. and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–828.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937-1958.

Gent, P. R., and Coauthors, 2011: The Community Climate System Model version 4. *J. Climate*, **24**, 4973–4991, https://doi.org/ 10.1175/2011JCLI4083.1.




Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

Griffies, S. M. and coauthors, 2009: Coordinated Ocean-ice Reference Experiments (COREs). *Ocean Modelling*, **26**, 1–46.

Hazeleger, W., and Coauthors, 2013: Multiyear climate predictions using two initialization strategies. *Geophys. Res. Lett.*, **40**, 1794–1798, doi:10.1002/grl.50355.

Hermanson, L., Ren, H., Vellinga, M. et al., 2018: Different types of drifts in two seasonal forecast systems and their dependence on ENSO. *Clim. Dyn.*, **51**, 1411–1426. https://doi.org/10.1007/s00382-017-3962-9

Huang, B., Z.-Z. Hu, B. Jha, 2007: Evolution of model systematic errors in the tropical Atlantic Basin from coupled climate hindcasts. *Clim. Dyn.*, **28**, 668–682.

Huang, B., and Coauthors, 2014: Climate drift of AMOC, North Atlantic salinity and artic sea ice in CFSv2 decadal predictions. *Clim. Dyn.*, **44**, 559-583. https://doi.org/10.1007/s00382-014-2395-y.

Hurrell, J., G. A. Meehl, D. Bader, T. L. Delworth, B. Kirtman, and B. Wielicki, 2009: A unified modeling approach to climate system prediction. *Bull. Amer. Meteor. Soc.*, **99**, 1819–1832.

Hurrell, J. M., and Coauthors, 2013: The Community Earth System Model: A framework for collaborative research. *Bull. Amer. Meteor. Soc.*, **94**, 1339–1360, https://doi.org/10.1175/BAMS-D- 12-00121.1.

Karspeck. A. R., S. Yeager, G. Danabasoglu, T. Hoar, N. Collins, K. Raeder, J. L. Anderson, and J. Tribbia, 2013: An ensemble adjustment kalman filter for the ccsm4 ocean component. *Journal of Climate*, **26**, 7392–7413.

Lee, J., Y. Xue, F. D. Sales, I. Diallo, L. Marx, M. Ek, K. R. Sperber, and P. J. Gleckler, 2019: Evaluation of multi-decadal UCLA-CFSv2 simulation and impact of interactive atmospheric-ocean feedback on global and regional variability. *Climate Dynamics*, **52**, 3683-3707. https://doi.org/10.1007/s00382-018-4351-8.





Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

Lin, J.-L., 2007: The double-ITCZ problem in IPCC AR4 coupled GCMS: Ocean–atmosphere feedback analysis. *J. Climate*, **20**, 4497–4525, https://doi.org/10.1175/JCLI4272.1.

Liu, H., M. Zhang, and W. Lin, 2012: An investigation of the initial development of the double -ITCZ warm SST biases in the CCSM. *Journal of Climate*, **25**, 140–155.

Ma, H.-Y., S. Xie, J. S. Boyle, S. A. Klein, and Y. Zhang, 2013: Metrics and diagnostics for precipitation-related processes in climate model short-range hindcasts. *J. Climate*, **26**, 1516–1534.

Ma, H.-Y., and coauthors, 2014: On the correspondence between mean forecast errors and climate errors in CMIP5 models. *J. Climate*, **27**, 1781–1798.

Ma, H.-Y., C. C. Chuang, S. A. Klein, M.-H. Lo, Y. Zhang, S. Xie, X. Zheng, P.-L. Ma, Y. Zhang, and T. J. Phillips, 2015: An improved hindcast approach for evaluation and diagnosis of physical processes in global climate models, *J. Adv. Model. Earth Syst.*, **7**, 1810–1827, doi:10.1002/2015MS000490.

Martin, G. M., S. F. Milton, C. A. Senior, M. E. Brooks, S. Ineson, T. Reichler, and J. Kim, 2010: Analysis and reduction of systematic errors through a seamless approach to modeling weather and climate. *J. Climate*, **23**, 5933–5957.

Mechoso, C. R., Robertson, A. W., Barth, N., Davey, M., Delecluse, P., Gent, P., Ineson, S., Kirtman, B., Latif, M., Treut, H. L., et al. (1995). The seasonal cycle over the tropical pacific in coupled ocean–atmosphere general circulation models. *Monthly Weather Review*, **123**(9):2825–2838.

Merryfield, W. J., W.-S. Lee, G. J. Boer, V. V. Kharin, J. F. Scinocca, G. M. Flato, R. S. Ajayamohan, J. C. Fyfe, Y. Tang, and S. Polavarapu (2013), The Canadian Seasonal to Interannual Prediction System. Part I: Models and Initialization, *Mon. Weather Rev.*, **141**, 2910-2945, doi:10.1175/MWR-D-12-00216.1.

Murakami, H., and coauthors (2015), Simulation and prediction of category 4 and 5 hurricanes in the high-resolution GFDL HiFLOR coupled climate model, *J. Climate*, **28**, 9058–9079.

Neelin, J. D., 1991: The slow sea surface temperature mode and the fast-wave limit: Analytic theory for tropical interannual oscillations and experiments in a hybrid coupled model. *Journal of the Atmospheric Sciences*, **48**, 584–606.

Nigam, S. and Y. Chao, 1996: Evolution dynamics of tropical ocean-atmosphere annual cycle variability. *Journal of Climate*, **9**, 3187–3205.

Palmer, T. N., F. J. Doblas-Reyes, A. Weisheimer, M. J. Rodwell, 2008: Toward seamless prediction: calibration of climate change projections using seasonal forecasts. *Bull. Amer. Meteor. Soc.*, 89, 459-470.

Phillips, T.J., G.L. Potter, D.L. Williamson, R.T. Cederwall, J. S. Boyle, M. Fiorino, J.J. Hnilo, J.G. Olson, S. Xie, and J.J. Yio (2004), Evaluating parameterizations in general circulation models: Climate simulation meets weather prediction. *Bull. Amer. Meteor. Soc.*, **85**, 1903–1915.

Raeder, K., J. L. Anderson, N. Collins, T. Hoar, J. Kay, P. Lauritzen, and R. Pincus, 2012: DART/CAM: An ensemble data assimilation system for CESM atmospheric models. *J. Climate*, **25**, 6304–6317.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan (2003), Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, *J. Geophys. Res.*, **108**, 4407, doi:10.1029/2002JD002670, D14.

Reynolds, R.W., N.A. Rayner, T.M. Smith, D.C. Stokes, and W. Wang (2002), An improved in situ and satellite SST analysis for climate. *J. Climate*, **15**, 1609-1625.



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

Reynolds, R. W., T. M. Smith, C. Liu, D. B. Chelton, K. S. Casey, and M. G. Schlax, 2007: Daily High-Resolution-Blended Analyses for Sea Surface Temperature. *J. Climate,* **20**, 5473-5496.

Richter, I., 2015: Climate model biases in the eastern tropical oceans: Causes, impacts and ways forward. Wiley Interdiscip. *Rev: Climate Change*, **6**, 345–358, doi:10.1002/wcc.338.

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

Sanchez-Gomez, E., and coauthors (2016), Drift dynamics in a coupled model initialized for decadal forecasts. *Clim. Dyn.*, **46**, 1819–1840.

Shonk, J. K. P., and coauthors (2018) Identifying causes of Western Pacific ITCZ drift in ECMWF System 4 hindcasts. *Clim. Dyn.*, **50**, 939–954.

da Silveira, I. P., P. Zuidema, and B. P. Kirtman, 2019: Fast SST error growth in the southeast Pacific Ocean: comparison between high and low-resolution CCSM4 retrospective forecasts. *Climate Dynamics,* 10.1007/s00382-019-04855-5.

Siongco, A. C., H.-Y. Ma, S. A. Klein, S. Xie, A. R. Karspeck, K. Raeder, J. L. Anderson, 2020: A hindcast approach to diagnosing the equatorial Pacific cold tongue SST bias in CESM1. *J. Climate*, **33**, 1437-1453.

Song, X., and G. J. Zhang, 2018: The roles of convection parameterization in the formation of double ITCZ syndrome in the NCAR CESM: I. Atmospheric processes. *Journal of Advances in Modeling Earth Systems,* **10**, 842–866. https://doi.org/10.1002/ 2017MS001191.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498.



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

Toniazzo, T. and S. Woolnough, 2014: Development of warm SST errors in the southern tropical Atlantic in CMIP5 decadal hindcasts. *Climate Dynamics*, **43**(11):2889–2913.

Vannière, B., Guilyardi, E., Madec, G., Doblas-Reyes, F. J., and S. Woolnough, 2013: Using Seasonal Hindcasts to Understand the Origin of the Equatorial Cold Tongue Bias in CGCMs and its Impact on ENSO. *Climate Dynamics*, **40**(3-4):963–981.

Vannière, B., E. Guilyardi, T. Toniazzo, G. Madec, and S. Woolnough, 2014: A systematic approach to identify the sources of tropical SST errors in coupled models using the adjustment of initialized experiments. *Climate Dynamics*, **43**, 2261-2282.

Vecchi, G. A., and Coauthors, 2014: On the seasonal forecasting of regional tropical cyclone activity. *J. Climate*, **27**, 7994–8016, https://doi.org/10.1175/JCLI-D-14-00158.1.

Voldoire, A., M. Claudon, G. Caniaux, H. Giordani, and R. Roehrig, R., 2014: Are atmospheric biases responsible for the tropical Atlantic SST biases in the CNRM-CM5 coupled model? *Climate Dynamics*, 43, 2963–2984.

Voldoire, A. and Coauthors, 2019: Role of wind stress in driving SST biases in the Tropical Atlantic. *Climate Dynamics*, https://doi.org/10.1007/s00382-019-04717-0.

Wang, C., L. Zhang, S.-K. Lee, L. Wu, and C. R. Mechoso, 2014: A global perspective on CMIP5 climate model biases. *Nat. Climate Change*, **4**, 201–205, doi:https://doi.org/10.1038/nclimate2118.

Wang, J. and J. A. Carton, 2002: Seasonal heat budgets of the North Pacific and North Atlantic Oceans. *J. Phys. Oceanogr.*, **32**, 3474–3489

Williams, K. D., A. Bodas-Salcedo, M. Déqué, S. Fermepin, B. Medeiros, M. Watanabe, C. Jakob, S. A. Klein, C. A. Senior, and D. L. Williamson (2013), The Transpose-AMIP II Experiment and Its Application to the Understanding of Southern Ocean Cloud Biases in Climate Models. *J. Climate*, **26**, 3258–3274.



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

Xie, S., H.-Y. Ma, J. S. Boyle, S. A. Klein, and Y. Zhang (2012), On the correspondence between short- and long- timescale systematic errors in CAM4/CAM5 for the Years of Tropical Convection. *J. Climate*, **25**, 7937–7955.

Zhang, S., M. J. Harrison, A. Rosati, and A. T. Wittenberg, 2007: System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies. *Mon. Wea. Rev.*, **135**, 3541–3564

Zhang, X., H. Liu, and M. Zhang, 2015: Double ITCZ in Coupled Ocean-Atmosphere Models: From CMIP3 to CMIP5. *Geophys. Res. Lett.*, **42**, 8651–8659, doi:10.1002/2015GL065973.

Zuidema, P., and Coauthors, 2016: Challenges and prospects for reducing coupled climate model SST biases in the eastern tropical Atlantic and Pacific oceans: The US CLIVAR eastern tropical oceans synthesis working group. *Bulletin of the American Meteorological Society*, **97**(12):2305–2328.





Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

Table 1. Summary of model information analyzed in this study. Note that CESM1-CAPT hindcasts were performed locally and are not part of the NMME project.

| Model Name | Modeling Group | Hindcast Period | Ensemble Size | References |
|---|---|---|---|---|
| CanCM3 | Canadian Centre for Climate Modelling and Analysis | 1982-2011 (30 years) | 10 | *Merryfield et al.* (2013) |
| CanCM4 | Canadian Centre for Climate Modelling and Analysis | 1982-2011 (30 years) | 10 | *Merryfield et al.* (2013) |
| CCSM4 | University of Miami | 1983-2014 (32 years) | 10 | *Gent et al.* (2011) |
| CESM1-NMME | National Center for Atmospheric Research | 1981-2010 (30 years) | 10 | *Hurrell et al.* (2013) |
| *CESM1-CAPT | Lawrence Livermore National Laboratory | 2005 (1 year) | 24 | *Hurrell et al.* (2013) |
| FLORB01 | National Oceanic and Atmospheric Administration / Geophysical Fluid Dynamics Laboratory | 1981-2010 (30 years) | 10 | *Vecchi et al.* (2014) |





Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

Table 2. Root mean square errors of SST (°C) averaged over the first day of model integrations with the starting dates of January 1, 2005 and July 1, 2005. The reference SST is the daily OISST v2.

| | January 1, 2005 | | | July 1, 2005 | | |
|---|---|---|---|---|---|---|
| | 0°–360°E, 30°S–30°N | 0°–360°E, 30°N–60°N | 0°–360°E, 60°S–30°S | 0°–360°E, 30°S–30°N | 0°–360°E, 30°N–60°N | 0°–360°E, 60°S–30°S |
| CanCM3 | 0.42 | 1.08 | 0.88 | 0.39 | 1.04 | 0.74 |
| CanCM4 | 0.42 | 1.06 | 0.87 | 0.4 | 1.03 | 0.75 |
| CCSM4 | 0.14 | 0.27 | 0.24 | 0.15 | 0.54 | 0.26 |
| CESM1-CAPT | 0.39 | 1.31 | 1.23 | 0.44 | 1.24 | 0.93 |





Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

Table 3. Spatial correlation coefficients of SST annual mean biases (°C) for different CESM simulation pairs. The reference SST is the monthly HadISST.

| | 0°–360°E, 30°S–30°N | 0°–360°E, 30°N–60°N | 0°–360°E, 60°S–30°S |
|---|---|---|---|
| 1-Mon-NMME-2005 vs 1-Mon-CAPT-2005 | 0.64 | 0.65 | 0.78 |
| 6-Mon-NMME-2005 vs 6-Mon-CAPT-2005 | 0.94 | 0.77 | 0.85 |
| 12-Mon-NMME-2005 vs 12-Mon-CAPT-2005 | 0.87 | 0.81 | 0.86 |
| 1-Mon-NMME-2005 vs 1-Mon-NMME-all | 0.95 | 0.96 | 0.96 |
| 6-Mon-NMME-2005 vs 6-Mon-NMME-all | 0.95 | 0.97 | 0.96 |
| 12-Mon-NMME-2005 vs 12-Mon-NMME-all | 0.88 | 0.97 | 0.95 |
| 1-Mon-CAPT-2005 vs CLIM | 0.49 | 0.65 | 0.39 |
| 6-Mon-CAPT-2005 vs CLIM | 0.88 | 0.85 | 0.57 |
| 12-Mon-CAPT-2005 vs CLIM | 0.89 | 0.85 | 0.69 |
| 1-Mon-NMME-2005 vs CLIM | 0.63 | 0.62 | 0.54 |
| 6-Mon-NMME-2005 vs CLIM | 0.9 | 0.81 | 0.69 |
| 12-Mon-NMME-2005 vs CLIM | 0.81 | 0.86 | 0.77 |
| 1-Mon-NMME-all vs CLIM | 0.69 | 0.74 | 0.65 |
| 6-Mon-NMME-all vs CLIM | 0.92 | 0.89 | 0.75 |
| 12-Mon-NMME-all vs CLIM | 0.94 | 0.92 | 0.81 |





Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

Table 4. Summary of whether an initialized coupled hindcast approach would be suitable for diagnosing a certain regional bias based on the following criteria: (1) the RMSE of the climatological SST is > 0.5 ºC, (2) the RMSE of Mon1 SST is < 0.5 ºC, (3) the RMSE of Mon12 is at least 60% of the climatological RMSE (indicating a growth of SST bias), and (4) the Mon12 hindcast and the climatological SST have the same bias sign. "✓" indicates a specific region satisfy all the criteria, while "1", "2", "3", or "4" indicates the first, second, third or fourth criterion is not satisfied.

| | CanCM3 | CanCM4 | CCSM4 | CESM1-NMME | CESM1-CAPT | FLORB01 |
|---|---|---|---|---|---|---|
| EQ Pacific (180°–240°E, 2°S–2°N) | ✓ | ✓ | 1,4 | ✓ | ✓ | ✓ |
| NE Pacific (110°–130°W, 20°–30°N) | 2 | 4 | ✓ | ✓ | ✓ | 2 |
| SE Pacific (70°–90°W, 10°–25°S) | 2 | 2 | ✓ | 2 | 2 | 2 |
| SE Atlantic (0°–15°E, 5°–25°S) | 2 | 2 | ✓ | 2 | 2 | 2 |
| N Pacific (160°–210°E, 20°–35°N) | ✓ | ✓ | ✓ | 2,3 | 3 | ✓ |
| S Pacific (130°–170°W, 15°–25°S) | ✓ | 3 | 1 | 4 | 4 | ✓ |
| N Atlantic (30°–60°W, 15°–25°N) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| S Atlantic (15°–40°W, 15°–25°N) | ✓ | ✓ | 1,4 | ✓ | 3,4 | 1,4 |
| Mid-Lat N Atlantic (20°–45°W, 40°–55°N) | 2 | 2,3 | 2,4 | 2 | 2 | 2 |
| Southern Ocean (0°–120°E, 45°–60°S) | 2 | 2 | 4 | 2 | 2 | 2,3 |



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

## Figure Captions

Figure 1. Annual multi-model mean biases (°C) of SST from the CMIP5 (25 models) and CMIP6 (34 models) historical simulations. Regions where mean biases are statistically significant at the 95% confidence level are color shaded. The observational reference is the HadISST. See Appendix A for more information about the CMIP models. The boxes indicate the regions discussed in Section 3.4.

Figure 2. Schematic diagram for the seasonal hindcast procedure for the NMME project. Each set of 10-member ensemble hindcasts started from 00Z on the first day of each month between January 1980 and December 2014 (abscissa). The duration of each hindcast is 12 months (ordinate). For the hindcast month 1 (Mon1 or the 0-month lead), SSTs are averaged for the first month of the hindcasts over all the ensemble members. SSTs of the hindcast month 2 (Mon2 or the 1-month lead) are averaged for the second month and so on for hindcast month 3 to month 12 (Mon3 to Mon12, or 2-month to 11-month lead). See text for more details.

Figure 3. SST biases (°C) averaged over the first day of model integrations with starting dates of January 1, 2005 (left panels) and July 1, 2005 (right panels). The observational reference is the NOAA OISST.

Figure 4. Annual mean SST biases (°C) for hindcast ensembles of Mon1, Mon2, Mon3, and Mon4. Regions where annual mean biases are statistically significant different from zero at the 95% confidence level are color shaded in all the models except for CESM1-CAPT for which this test is not possible because only one year of hindcasts was performed. The observational reference is the HadISST and the hindcast period for each model is listed in Table 1.



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

Figure 5. Annual mean SST biases (°C) for hindcast ensembles of Mon6, Mon9 and Mon12, as well as for the corresponding annual long-term mean biases (a 40-year long historical coupled simulation for CanCM3; CMIP5/historical for CanCM4, CCSM4, and CESM1; a 300-year long control coupled simulation for FLOR). Regions where annual mean biases are statistically significant at the 95% confidence level are color shaded in all the models except for CESM1-CAPT hindcasts for which this test is not possible because only one year of hindcasts was performed.

Figure 6. Root mean square errors of annual mean SST (ºC) calculated over tropical and extra-tropical domains with different hindcast lead times and the long-term climatology (CLIM).

Figure 7. Taylor diagrams illustrating the spatial correlation and normalized spatial standard deviation of SST annual mean biases from the seasonal hindcasts. The reference fields (REF) are the corresponding biases in the long-term climatological runs. Data are analyzed over (a) 0°–360°E, 60°S–60°N, (b) 0°–360°E, 30°S–30°N, (c) 0°–360°E, 30°N–60°N, and (d) 0°–360°E, 60°S–30°S. The observational reference is the HadISST. Only grid points with biases that are statistically significant at 95% confidence level in both the hindcasts and climate runs are used for pattern statistics calculation in the Taylor diagrams.

Figure 8. Ensemble monthly mean SST biases (ºC) of January 2005 from the seasonal hindcasts for Mon6 (left panels). Also shown are the standard deviations of the ensemble mean biases (ºC) for Mon6 (middle panels) and the ratios of the monthly mean biases to the standard deviation of ensemble mean biases (right panels). The observational reference is the HadISST.

Figure 9. Ensemble monthly mean SST biases (ºC) of July 2005 from the seasonal hindcasts for Mon6 (left panels). Also shown are the standard deviations of the ensemble mean biases



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

(ºC) for Mon6 (middle panels) and the ratios of the monthly mean biases to the standard deviation of ensemble mean biases (right panels). The observational reference is the HadISST.

Figure 10. Interannual standard deviation of annual mean SST biases (ºC) for Mon12 (left panels). Also shown on the right panels are the ratios of annual mean SST biases to the interannual standard deviation of SST mean biases. The observational reference is the HadISST.

Figure 11. SST annual mean biases (ºC) of year 2005 from seasonal hindcasts for Mon1, Mon6, and Mon12 from CESM1-NMME (left panels) and CESM1-CAPT (right panels). The observational reference is the HadISST.

Figure 12. Ratio of annual mean SST biases for Mon 12 seasonal hindcasts to their corresponding climatological annual mean SST biases. Only regions where the annual mean SST biases of the long-term climatology are statistically significant at the 95% confidence level and their absolute values are larger than 0.5 ºC, are color shaded

Figure 13. Root mean square errors (RMSE ºC) of annual mean SST over ten selected locations (longitudes and latitudes are indicated on the top of each panel) with different hindcast lead times and the long-term climatology (CLIM). The RMSE for each model and hindcast month is calculated in reference to the HadISST.





Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

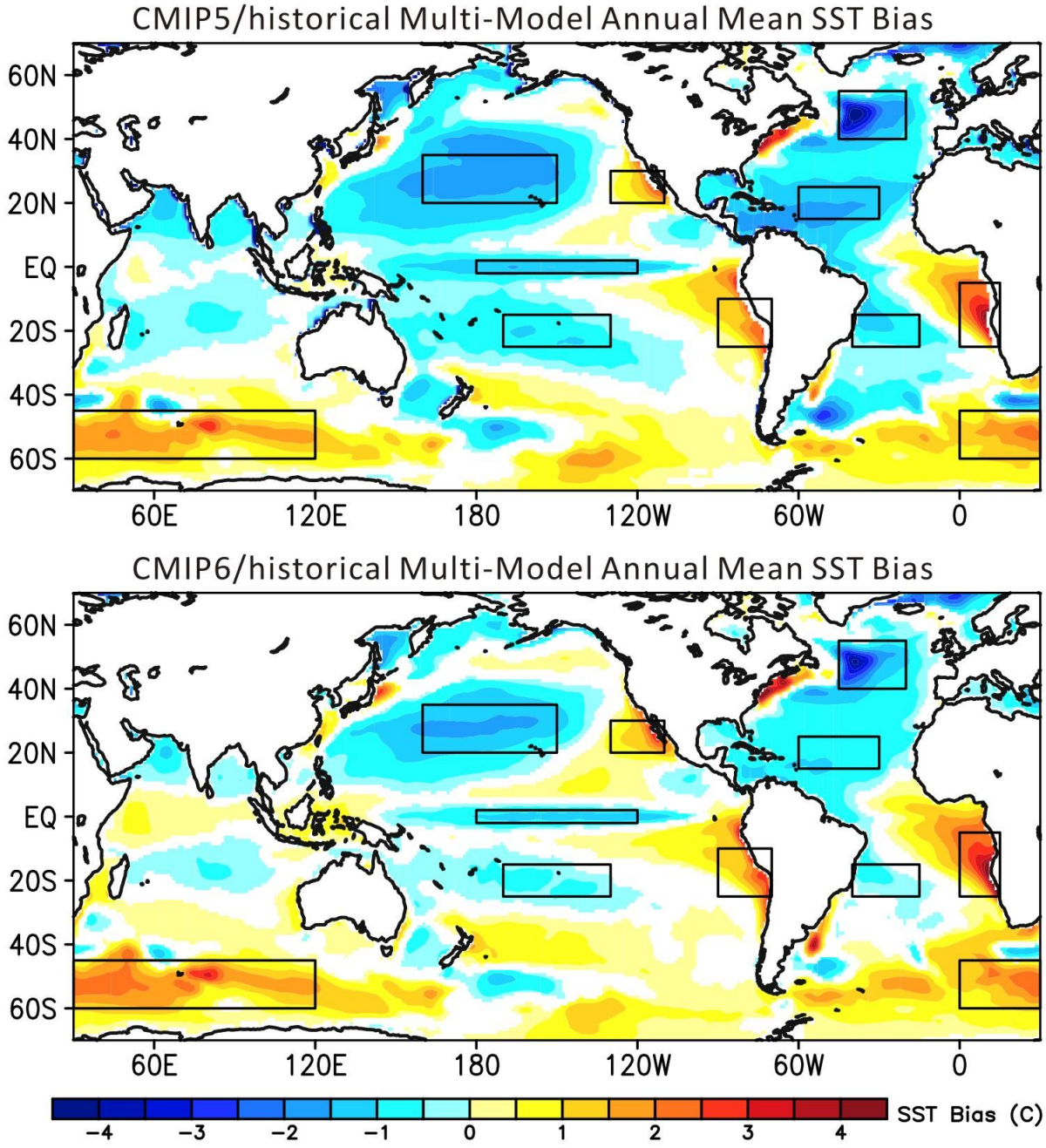


Figure 1. Annual multi-model mean biases (°C) of SST from the CMIP5 (25 models) and CMIP6 (34 models) historical simulations. Regions where mean biases are statistically significant at the 95% confidence level are color shaded. The observational reference is the HadISST. See Appendix A for more information about the CMIP models. The boxes indicate the regions discussed in Section 3.4.



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

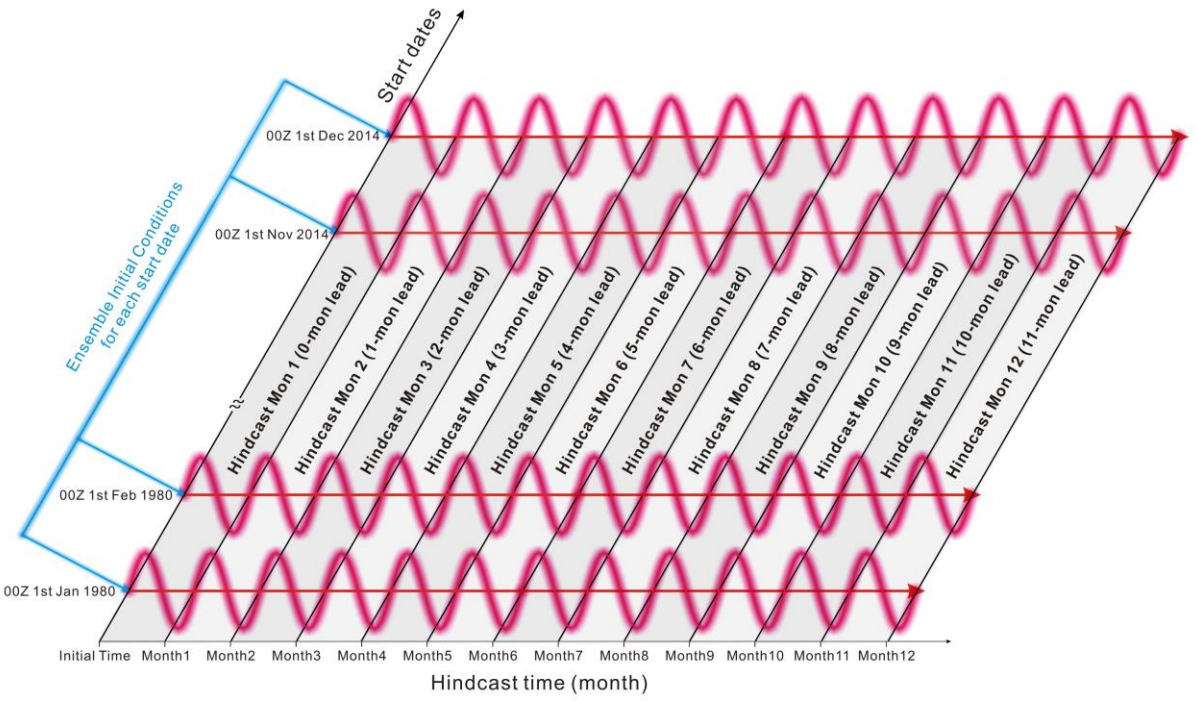


Figure 2. Schematic diagram for the seasonal hindcast procedure for the NMME project. Each set of 10-member ensemble hindcasts started from 00Z on the first day of each month between January 1980 and December 2014 (abscissa). The duration of each hindcast is 12 months (ordinate). For the hindcast month 1 (Mon1 or the 0-month lead), SSTs are averaged for the first month of the hindcasts over all the ensemble members. SSTs of the hindcast month 2 (Mon2 or the 1-month lead) are averaged for the second month and so on for hindcast month 3 to month 12 (Mon3 to Mon12, or 2-month to 11-month lead). See text for more details.



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

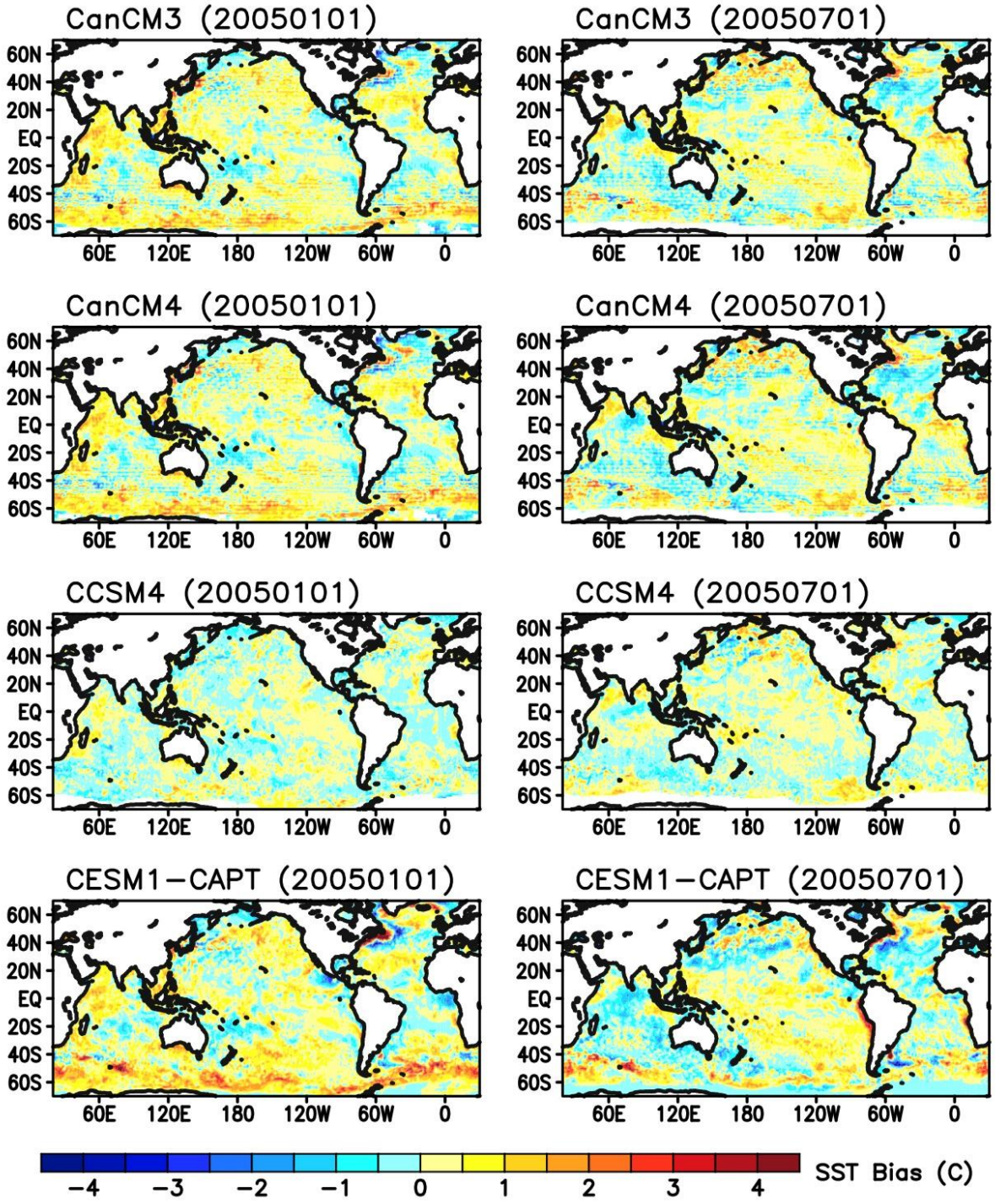


Figure 3. SST biases (°C) averaged over the first day of model integrations with starting dates of January 1, 2005 (left panels) and July 1, 2005 (right panels). The observational reference is the NOAA OISST.



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

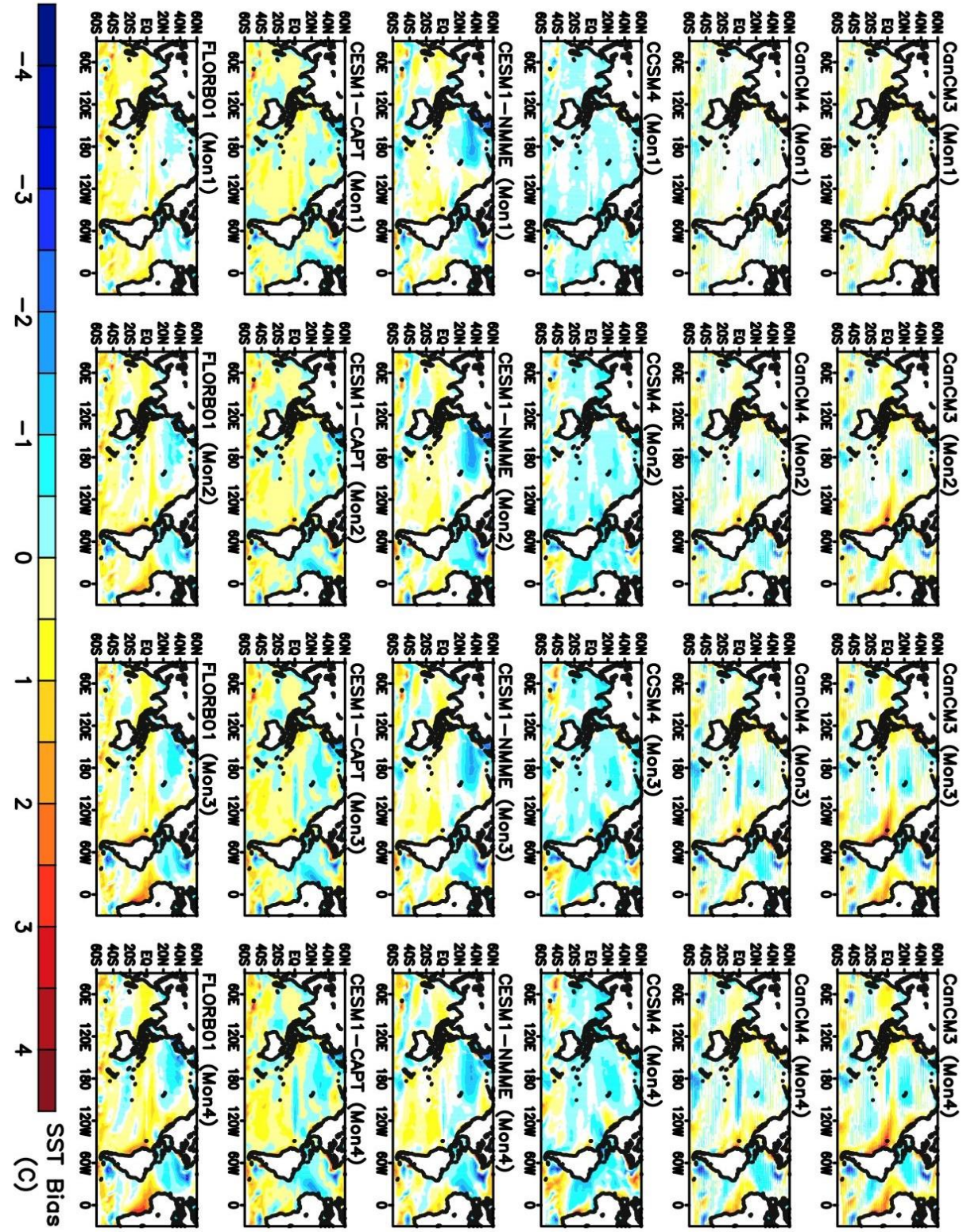


Figure 4. Annual mean SST biases (°C) for hindcast ensembles of Mon1, Mon2, Mon3, and Mon4. Regions where annual mean biases are statistically significant different from zero at the 95% confidence level are color shaded in all the models except for CESM1-CAPT for which this test is not possible because only one year of hindcasts was performed. The observational reference is the HadISST and the hindcast period for each model is listed in Table 1.



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

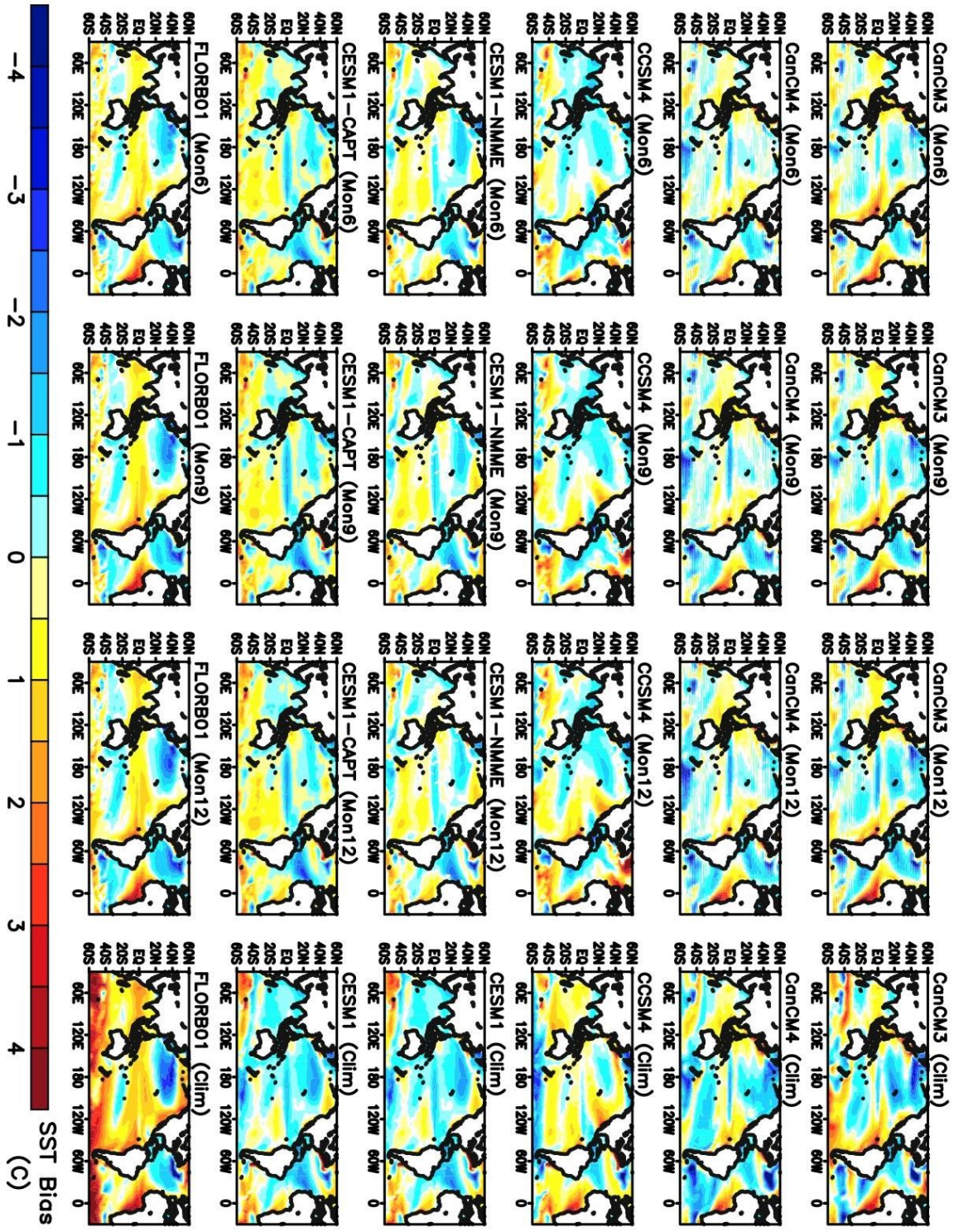


Figure 5. Annual mean SST biases (°C) for hindcast ensembles of Mon6, Mon9 and Mon12, as well as for the corresponding annual long-term mean biases (a 40-year long historical coupled simulation for CanCM3; CMIP5/historical for CanCM4, CCSM4, and CESM1; a 300-year long control coupled simulation for FLOR). Regions where annual mean biases are statistically significant at the 95% confidence level are color shaded in all the models except for CESM1-CAPT hindcasts for which this test is not possible because only one year of hindcasts was performed.



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

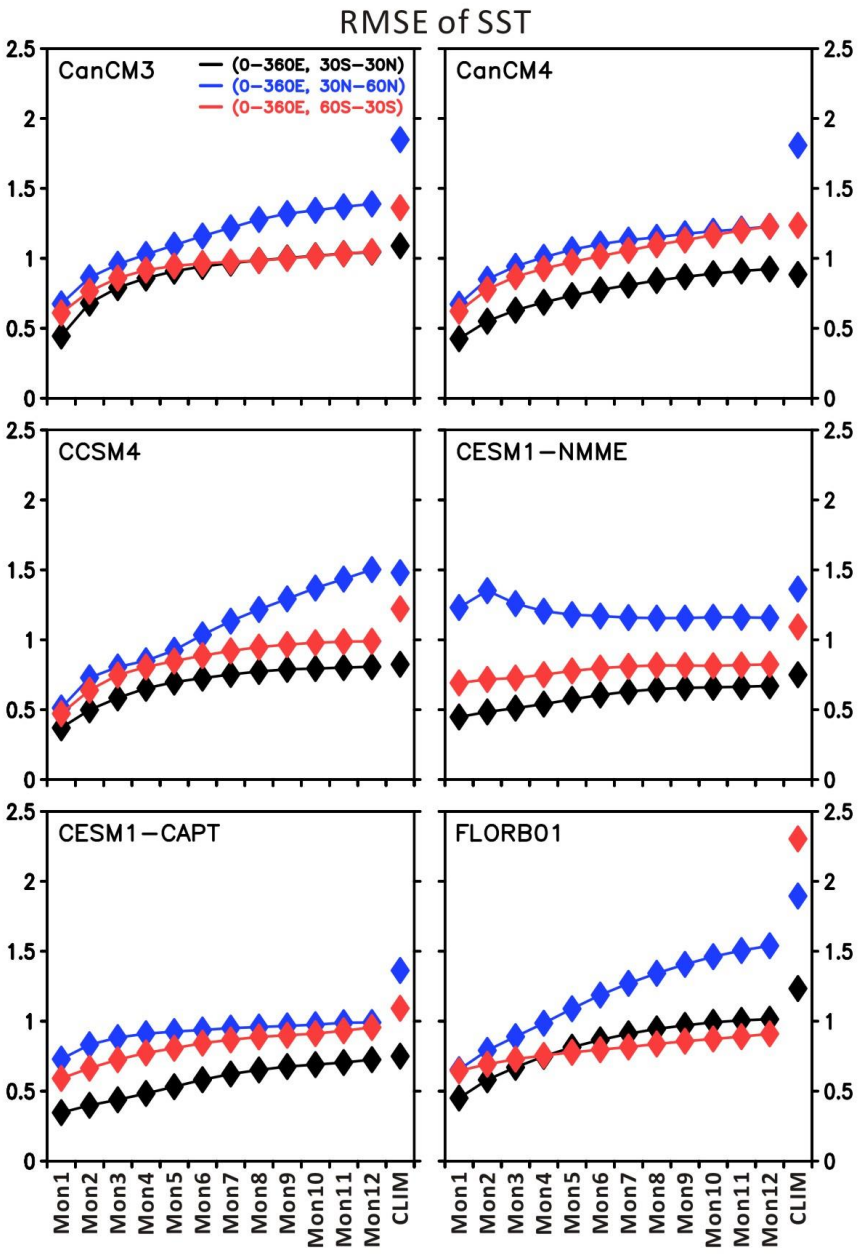


Figure 6. Root mean square errors of annual mean SST (ºC) calculated over tropical and extra-tropical domains with different hindcast lead times and the long-term climatology (CLIM).



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

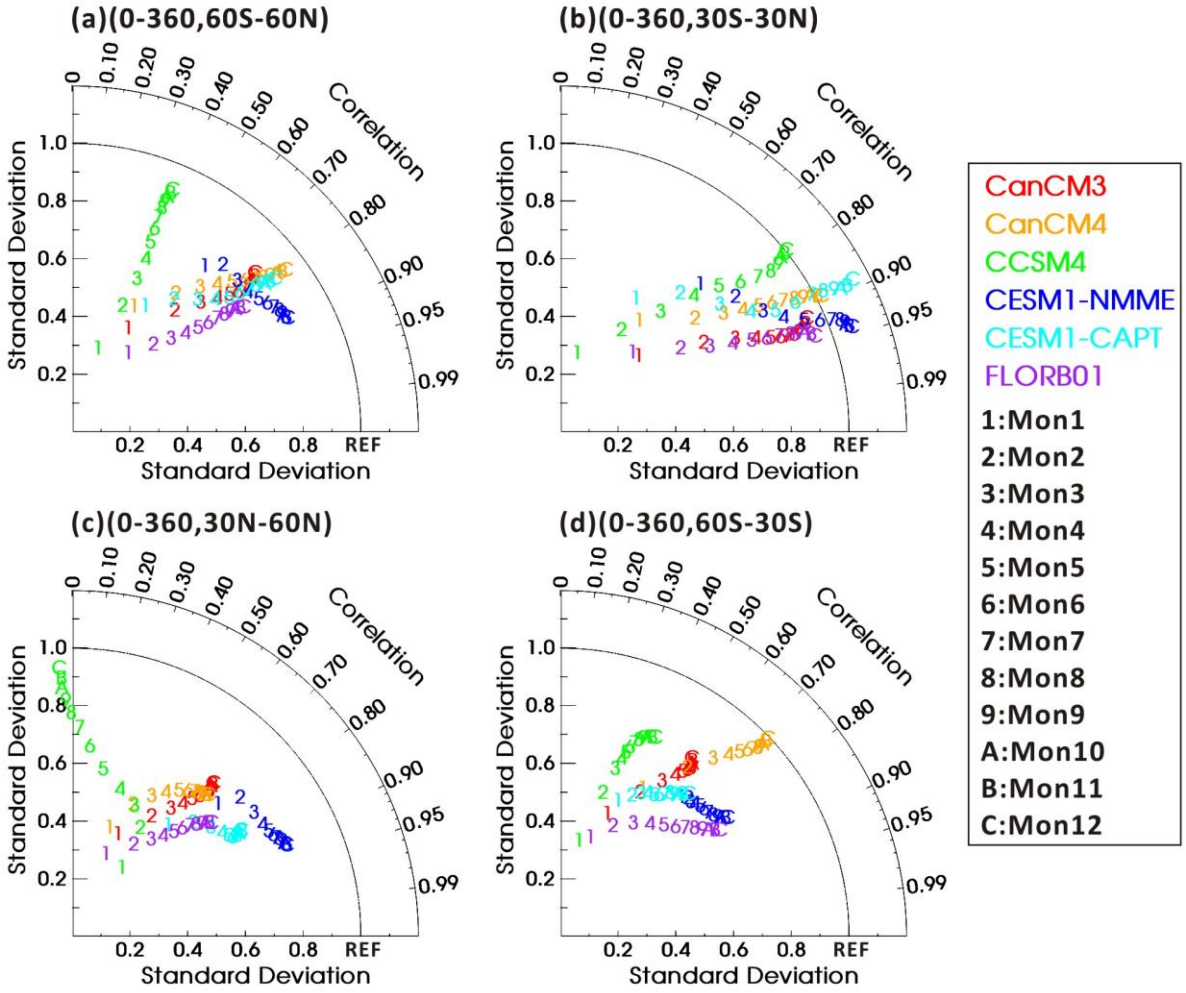


Figure 7. Taylor diagrams illustrating the spatial correlation and normalized spatial standard deviation of SST annual mean biases from the seasonal hindcasts. The reference fields (REF) are the corresponding biases in the long-term climatological runs. Data are analyzed over (a) 0°–360°E, 60°S–60°N, (b) 0°–360°E, 30°S–30°N, (c) 0°–360°E, 30°N–60°N, and (d) 0°–360°E, 60°S–30°S. The observational reference is the HadISST. Only grid points with biases that are statistically significant at 95% confidence level in both the hindcasts and climate runs are used for pattern statistics calculation in the Taylor diagrams.





Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

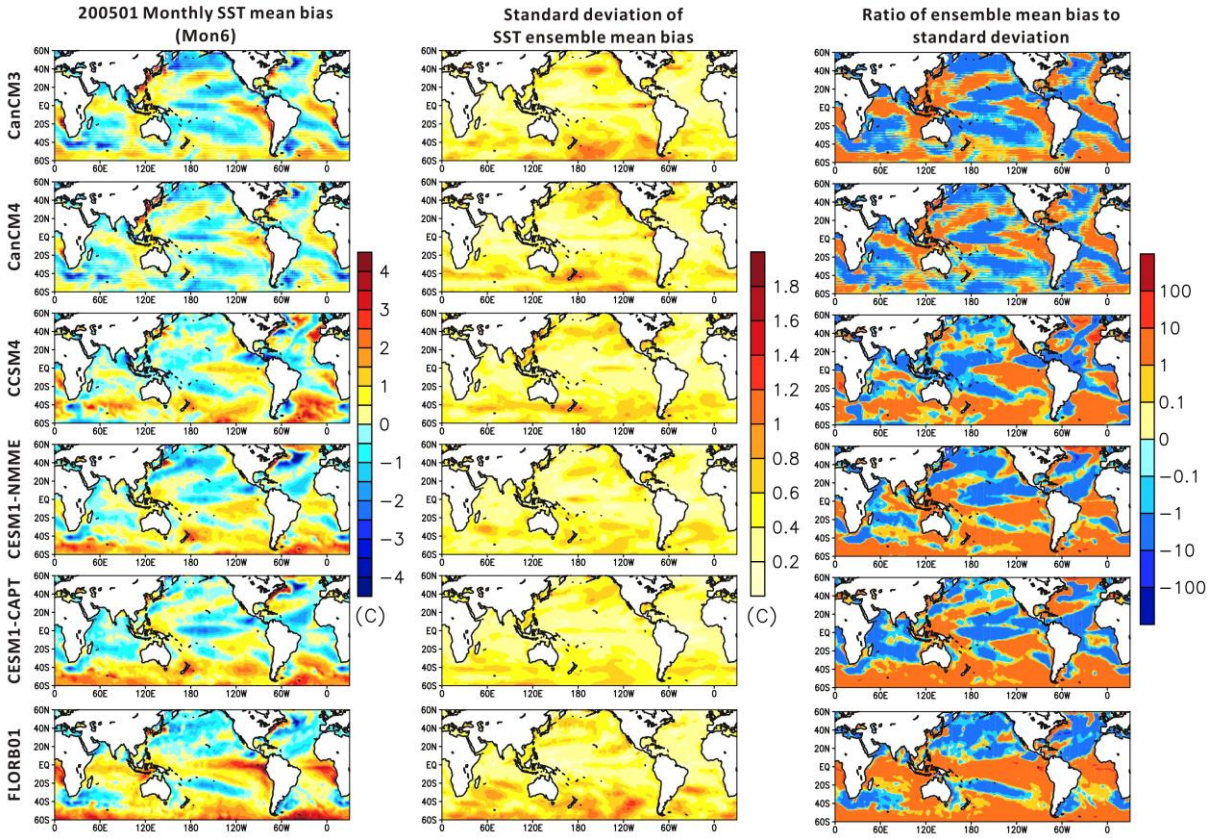


Figure 8. Ensemble monthly mean SST biases (ºC) of January 2005 from the seasonal hindcasts for Mon6 (left panels). Also shown are the standard deviations of the ensemble mean biases (ºC) for Mon6 (middle panels) and the ratios of the monthly mean biases to the standard deviation of ensemble mean biases (right panels). The observational reference is the HadISST.





Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

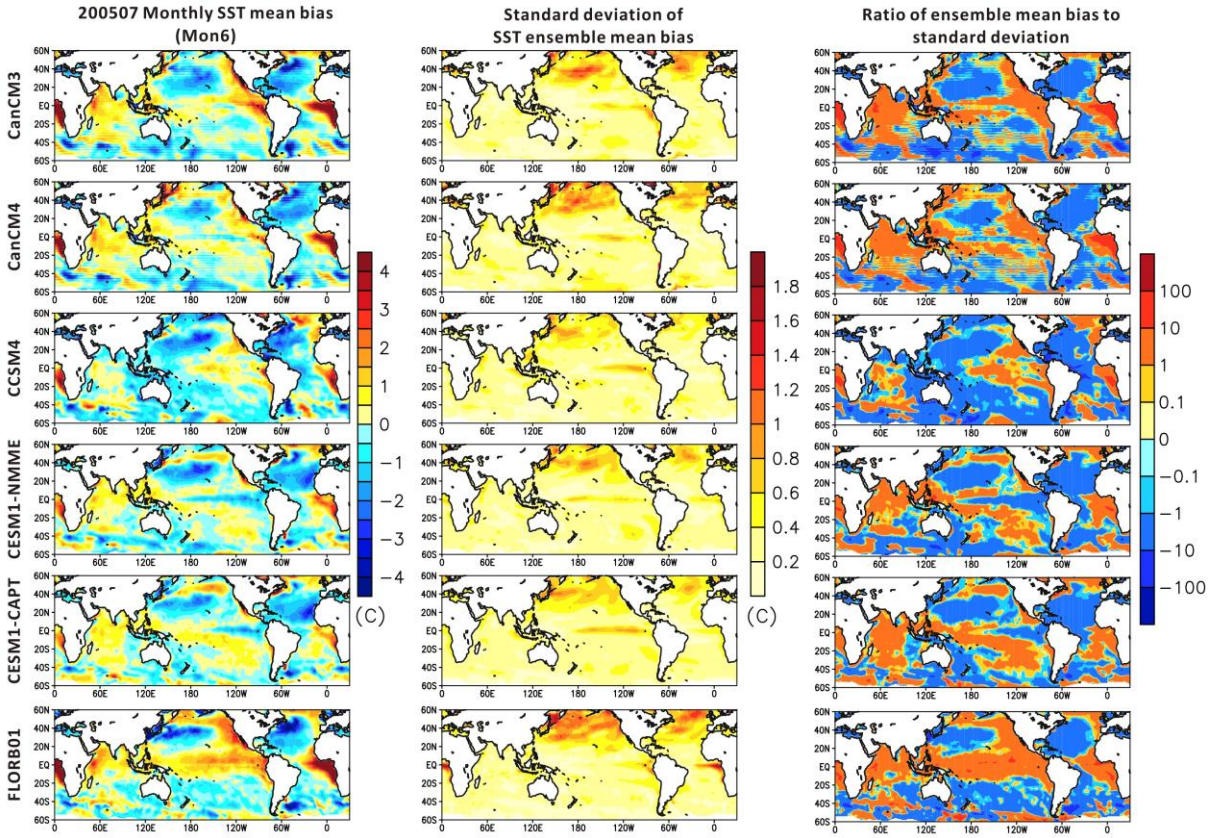


Figure 9. Ensemble monthly mean SST biases (ºC) of July 2005 from the seasonal hindcasts for Mon6 (left panels). Also shown are the standard deviations of the ensemble mean biases (ºC) for Mon6 (middle panels) and the ratios of the monthly mean biases to the standard deviation of ensemble mean biases (right panels). The observational reference is the HadISST.





Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

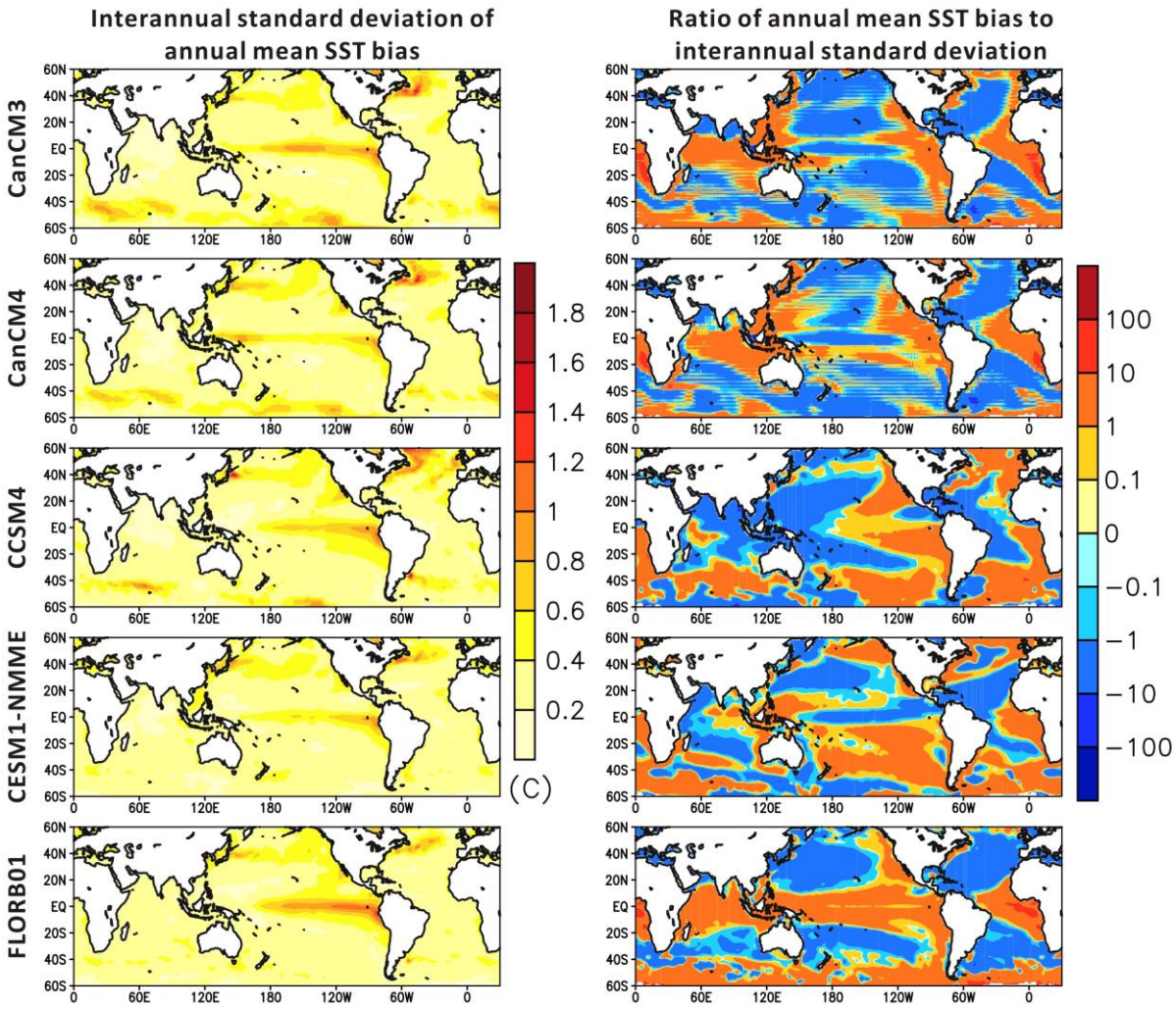


Figure 10. Interannual standard deviation of annual mean SST biases (ºC) for Mon12 (left panels). Also shown on the right panels are the ratios of annual mean SST biases to the interannual standard deviation of SST mean biases. The observational reference is the HadISST.





Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

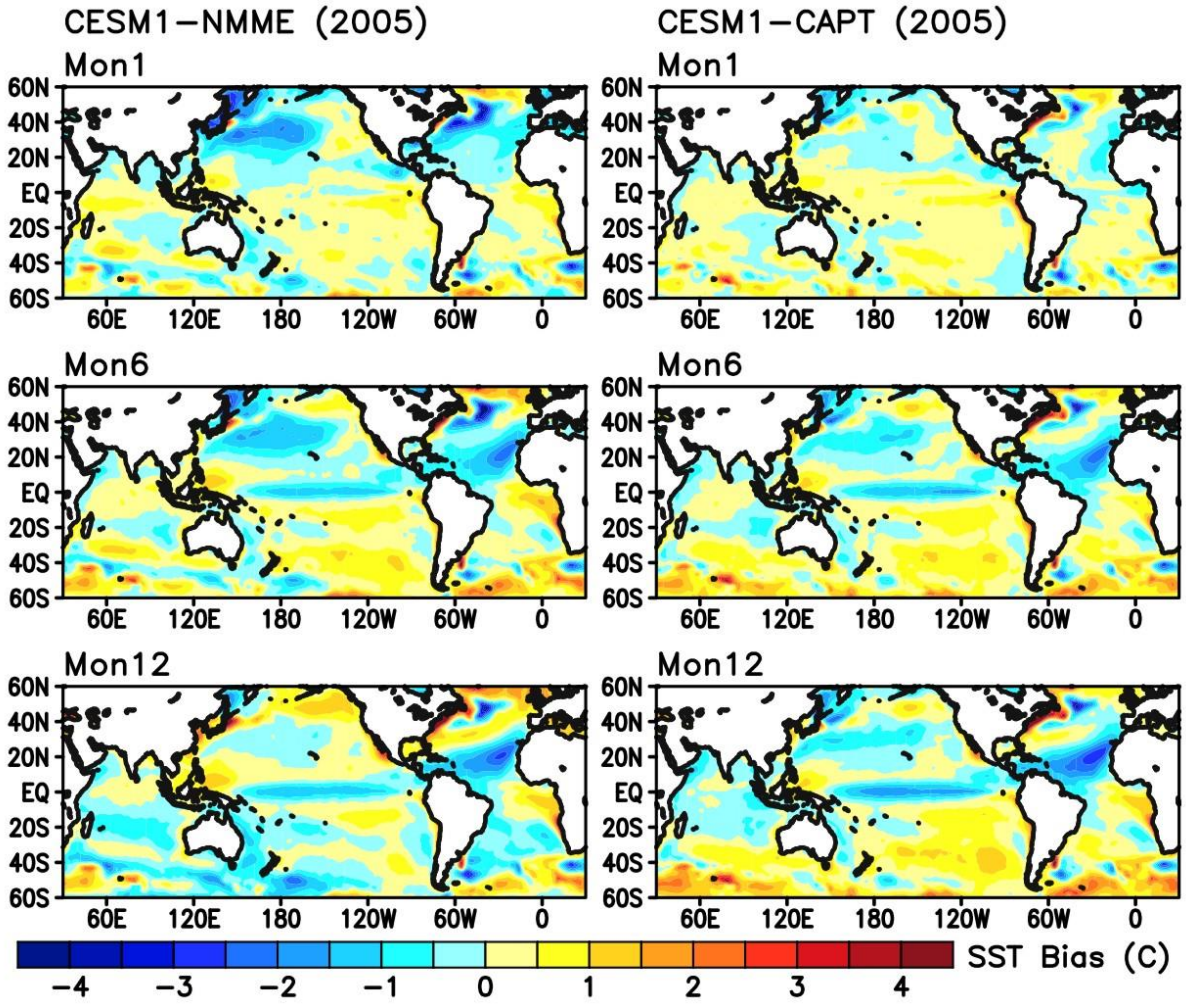


Figure 11. SST annual mean biases (ºC) of year 2005 from seasonal hindcasts for Mon1, Mon6, and Mon12 from CESM1-NMME (left panels) and CESM1-CAPT (right panels). The observational reference is the HadISST.



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

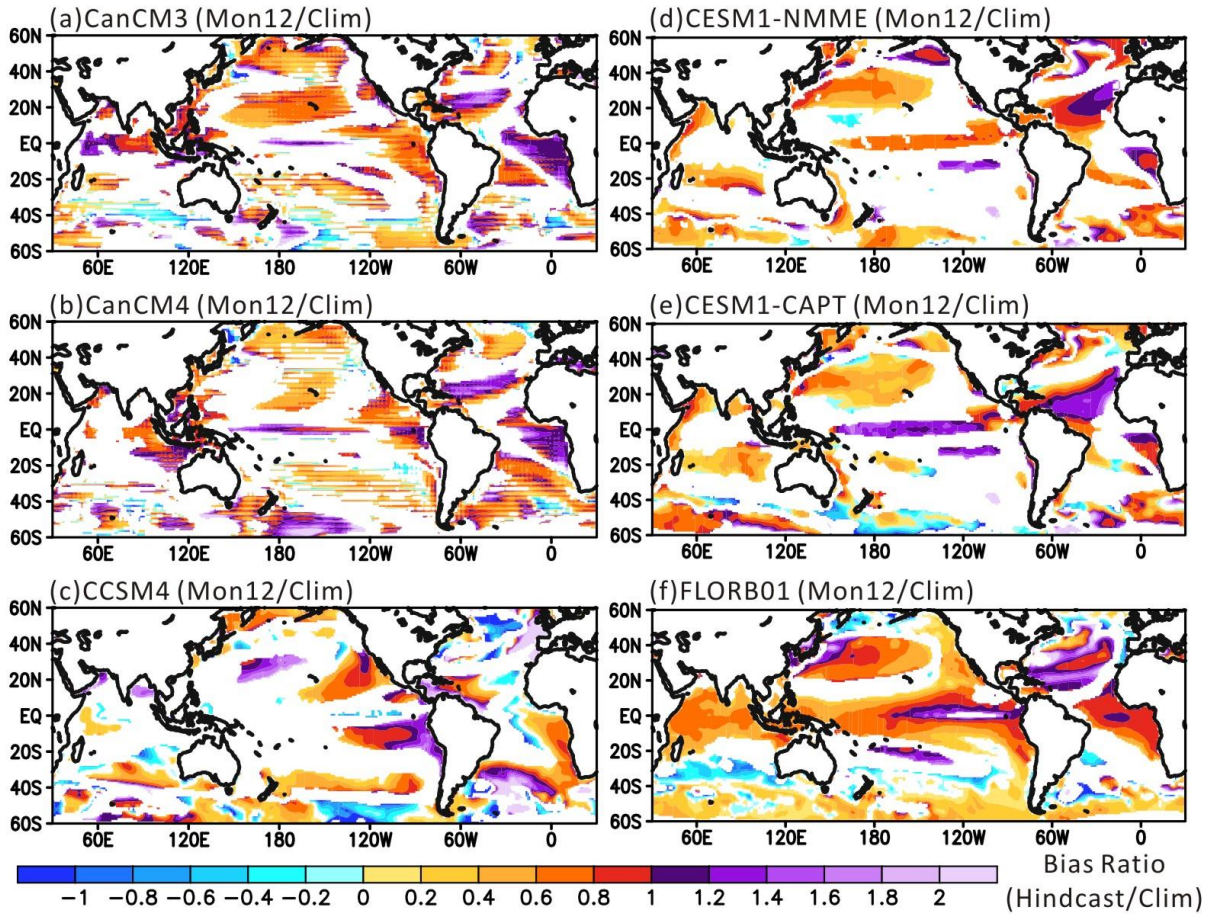


Figure 12. Ratio of annual mean SST biases for Mon 12 seasonal hindcasts to their corresponding climatological annual mean SST biases. Only regions where the annual mean SST biases of the long-term climatology are statistically significant at the 95% confidence level and their absolute values are larger than 0.5 ºC, are color shaded.





Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.

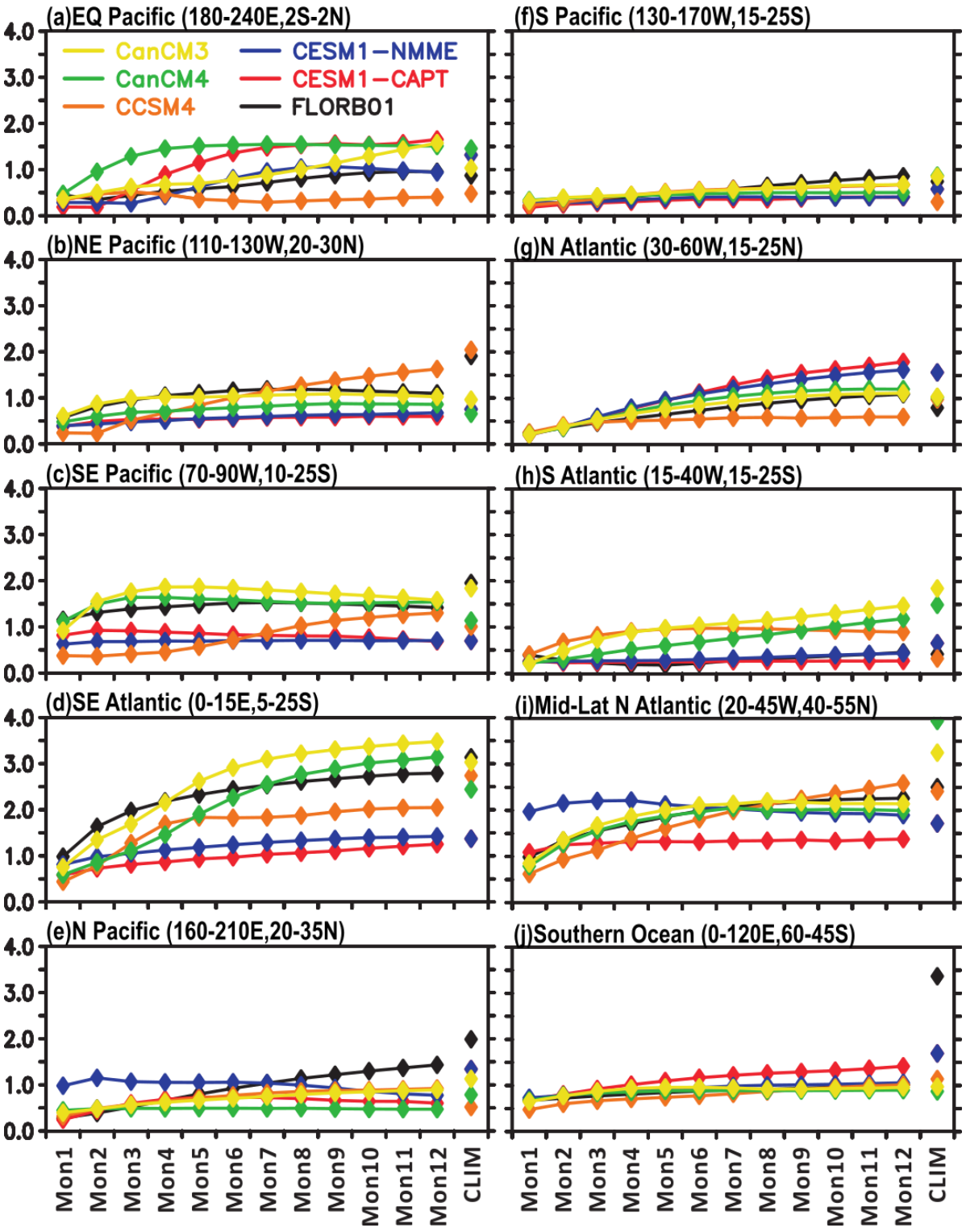


Figure 13. Root mean square errors (RMSE ºC) of annual mean SST over ten selected locations (longitudes and latitudes are indicated on the top of each panel) with different hindcast lead times and the long-term climatology (CLIM). The RMSE for each model and hindcast month is calculated in reference to the HadISST. Note that the CLIM is the same for both CESM1-NMME and CESM-CAPT.



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0338.1.