

Subseasonal Prediction with and without a Well-Represented Stratosphere in CESM1[Ⓞ]

JADWIGA H. RICHTER,^a KATHY PEGION,^b LANTAO SUN,^c HYEMI KIM,^d JULIE M. CARON,^a ANNE GLANVILLE,^a EMERSON LAJOIE,^{e,f} STEPHEN YEAGER,^a WHO M. KIM,^a AHMED TAWFIK,^g AND DAN COLLINS^e

^a *Climate and Global Dynamics Laboratory, National Center for Atmospheric Research, Boulder, Colorado*

^b *Department of Atmospheric, Oceanic, and Earth Sciences, George Mason University, Fairfax, Virginia*

^c *Colorado State University, Fort Collins, Colorado*

^d *School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, New York*

^e *NOAA/NCEP/Climate Prediction Center, College Park, Maryland*

^f *Innovim, Inc., College Park, Maryland*

^g *Full Stack Science LLC, Cary, North Carolina*

(Manuscript received 18 February 2020, in final form 15 October 2020)

ABSTRACT: There is a growing demand for understanding sources of predictability on subseasonal to seasonal (S2S) time scales. Predictability at subseasonal time scales is believed to come from processes varying slower than the atmosphere such as soil moisture, snowpack, sea ice, and ocean heat content. The stratosphere as well as tropospheric modes of variability can also provide predictability at subseasonal time scales. However, the contributions of the above sources to S2S predictability are not well quantified. Here we evaluate the subseasonal prediction skill of the Community Earth System Model, version 1 (CESM1), in the default version of the model as well as a version with the improved representation of stratospheric variability to assess the role of an improved stratosphere on prediction skill. We demonstrate that the subseasonal skill of CESM1 for surface temperature and precipitation is comparable to that of operational models. We find that a better-resolved stratosphere improves stratospheric but not surface prediction skill for weeks 3–4.

SIGNIFICANCE STATEMENT: There is a growing demand in society for understanding sources of predictability on subseasonal to seasonal time scales. In this work we demonstrate that the CESM1 research Earth system model can be utilized as a subseasonal prediction model and show that its subseasonal prediction skill is comparable to that of operational models. We also show that the inclusion of a well-resolved stratosphere does not improve the subseasonal (week 3–4 averaged) forecast of temperature and precipitation at the surface.

KEYWORDS: Stratospheric circulation; Ensembles; Forecast verification/skill; Hindcasts; Seasonal forecasting; Tropical variability

1. Introduction

There is growing demand and interest in subseasonal prediction, specifically in the three to four weeks window, that would fill the gap between weather prediction (up to 14 days) and seasonal forecasts (3–6 months) (Robertson et al. 2015). Subseasonal prediction of extreme events such as heat waves, droughts, heavy rainfall, and cold outbreaks has applications in a variety of society-relevant sectors including agriculture, energy, water management, and public health (White et al. 2017). With the growing demand for improved subseasonal forecasts, there is also a need for better understanding of potential sources of predictability on subseasonal time scales (Mariotti et al. 2020; NAS 2016).

Predictability on subseasonal time scale is believed to come from processes that vary slowly relative to the atmosphere and interactions related to soil moisture, snowpack, ocean heat content and sea ice (NAS 2016). Modes of variability such as El Niño Southern Oscillation (ENSO) and the Madden–Julian oscillation

(MJO) (NAS 2016), as well as the stratosphere (see Butler et al. 2019) for detailed review) are additional predictability sources. In particular, the stratosphere is believed to lead to improved surface prediction via extreme stratospheric polar vortex events, such as major sudden stratospheric warmings (SSWs) (e.g., Baldwin and Dunkerton 2001; Baldwin et al. 2003). Marshall and Scaife (2010) showed one week of extended SSW prediction skill in a high-top version of the model in comparison to a low-top version. Also the stratospheric quasi-biennial oscillation (QBO) has been shown to influence subseasonal predictability via the modulation of the MJO (e.g., Yoo and Son 2016; Son et al. 2017; Nishimoto and Yoden 2017). Nie et al. (2019) have shown that stratospheric initial conditions contribute to the predictability of the North Atlantic Oscillation (NAO) and the Arctic Oscillation (AO). Although the stratosphere has been demonstrated to influence tropospheric variability, it is difficult to disentangle its contribution to subseasonal predictability from other predictability sources. The international subseasonal-to-seasonal (S2S) research project (Vitart and Robertson 2018) and the more recent Subseasonal Experiment (SubX) (Pegion et al. 2019) offer very extensive databases of subseasonal predictions. However, it can be difficult to isolate sources of predictability using these databases, including the stratosphere, as due to the operational nature of most participating models these databases are largely

[Ⓞ] Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/WAF-D-20-0029.s1>.

Corresponding author: Jadwiga H. Richter, jrichter@ucar.edu

ensembles of opportunity. Using the S2S database, Domeisen et al. (2019a,b) found that high-top prediction systems exhibit higher stratospheric prediction skill; however, higher surface prediction skill was difficult to demonstrate. The S2S and SubX models' setup vary vastly in terms of horizontal and vertical resolutions, model lid heights, number of ensemble members, reforecast dates and periods, varying initial conditions, etc., hence it is difficult to attribute improved skill to a particular model feature. We evaluate here the effectiveness of CESM1 as a subseasonal prediction model and demonstrate the research utility of this model in investigating the role of the stratosphere to subseasonal prediction by comparing the default 30-level (30L) atmospheric configuration to the 46-level (46L) stratosphere-resolving configuration.

CESM1 is the 6th generation climate model developed by NCAR, preceded recently by Community Climate System Model, v4 (CCSM4), primarily for studies of climate change and climate variability. Recently CESM1 has been used in studies of decadal prediction (Yeager et al. 2018) and seasonal prediction through the NMME project (Kirtman et al. 2014). The comparison of 30LCESM1 and 46LCESM1 presented here is specifically designed to compare the role of the stratosphere on subseasonal prediction, since all of the forecast system details, such as initialization, start dates, number of ensemble members, are the same between the model versions. Since the stratospheric phenomena such as the SSWs and QBO are improved in the 46L version of CESM1, we expect to see improved subseasonal prediction skill of those and related phenomena in this version of the model as compared to the 30L version.

2. Model and simulations

a. Model description

We use here the default 30LCESM1 and 46LCESM1, with the Community Atmosphere Model, version 5 (CAM5) as its atmospheric component, interactive ocean (LANL Parallel Ocean Program, version 2) (Danabasoglu et al. 2012; Hurrell et al. 2013), and the Community Land Model, version 4 (CLM4) (Lawrence et al. 2011). CAM5 and the physics are fully described in Neale et al. (2012). CAM5 used for this study utilizes the spectral element dynamical core (Dennis et al. 2012) with horizontal resolution of ~ 100 km. The 30LCAM5 and 46LCAM5 have identical levels below 100 hPa, and 16 additional levels in 46LCAM5 raise the model top from 2 to 0.3 hPa and allow for a better representation of stratospheric dynamics. The vertical grids for the two model versions are shown in Fig. S1 of Richter et al. (2015). Both model versions utilize the same tropospheric CAM5 physics. In addition, 46LCAM5 includes a parameterization of nonorographic gravity waves based on Richter et al. (2010) and Richter et al. (2014). The 46LCAM5 has a nearly identical climate in the troposphere as 30LCAM5, but an improved representation of the stratosphere as compared to 30LCAM5. In particular, the main differences between 46LCAM5 and 30LCAM5 are in the simulation of the QBO and SSWs.

b. AMIP simulations

As CESM1 is primarily used to study Earth's climate, multidecadal simulations have previously been carried out with

this model. We utilize here "AMIP" simulations, in which CESM1 was run with observed evolution of global monthly sea surface temperatures and sea ice conditions as lower boundary conditions from Hurrell et al. (2008) and observed external radiative forcings (solar, greenhouse gases, volcanoes, and aerosols) for the period of 1952–2001. A 10-member ensemble of these simulations was carried out (Richter et al. 2015). To compare the AMIP climatology to the climatology of the reforecasts, one of the ensemble members for 30LCAM5 and 46LCAM5 was extended here until 2015.

AMIP simulations clearly demonstrate the key differences between the 46-level and 30-level configurations of CESM1. In 30LCAM5 AMIP simulations, the zonal mean wind above 100 hPa is persistently easterly (between -5 and -10 m s^{-1}) and the observed westerlies are never reproduced, while 46LCAM5 produces an internally generated QBO with a mean period of 27 months, very similar to the mean observed period of 28 months (see Fig. S1 in the online supplemental material). The QBO in 46LCAM5 is somewhat deficient compared to observations, with the easterly phase a little weak, and neither phase reaching down to 100 hPa as seen in ERAI; however, the overall representation of the QBO in 46LCAM5 is significantly improved over 30LCAM5. The SSW frequency in 46LCAM5 and 30LCAM5 AMIP simulations are 0.6 and 0.55 yr^{-1} , respectively, based on the 10-member ensemble of AMIP simulations from 1952 to 2001 shown in Richter et al. (2015), and are similar to reanalysis (Charlton and Polvani 2007). However, the seasonal distribution of SSWs differs between the two models (Fig. S2). In 30LCAM5 the DJF SSW frequency is less than 0.2 yr^{-1} , whereas in 46LCAM5 SSW frequency is 0.36, which is very close to the frequency of ~ 0.4 found in observations.

c. S2S reforecasts

S2S reforecasts are carried out with the fully coupled version of 30LCESM1 and 46LCESM1 following the SubX protocol (Pegion et al. 2019) with weekly starts from 1999 to 2015. Hindcasts are initialized on Wednesdays, starting with 6 January 1999, with the exception of years in which 29 February occurs on a Wednesday, as CESM1 does not have leap years. In those instances the forecasts are initialized on Tuesday, 28 February. Reforecasts are carried out for a period of 45 days. All reforecasts consist of 10 ensemble members for both the 30LCESM1 and 46LCESM1. Output of hindcasts follows the SubX protocol (Pegion et al. 2019) and consists of daily means of selected variables. Some of the daily averages are created from all model timesteps, whereas others are averages of instantaneous 6-hourly output. The distinction is made based on NOAA/CPC guidelines.

CLM4 was used to produce land initial condition files for every day from 1979 to the end of 2015 that reflect the observed surface state and fluxes. This was done by performing a 60-yr stand-alone CLM4 simulation forced with 6-hourly precipitation, temperature, specific humidity, wind speed, lowest atmospheric level pressure, and incoming longwave and shortwave radiation from the Climate Research Unit–National Centers for Environmental Prediction joint dataset (CRU–NCEP; downloaded from <http://dods.ipsl.jussieu.fr/igcm/IGCM/BC/OOL/OL/CRU-NCEP>) from 1950 to 1979. To

achieve the desired 60-yr spinup period, the 30-yr CRU-NCEP (1950–79) forcing was iterated over twice. In the first cycle, soil states were initialized with a uniform value globally and the land surface was forced with CRU–NCEP reanalysis from 1950 to 1979. The process was then repeated, but with the land properties at the beginning of the second cycle starting from those taken from the final time step of the first spinup cycle. This process ensured that soil moisture and temperature stabilized with respect to the 1950–79 climate state. After completing the second spinup cycle the CLM4 simulation forced with CRU–NCEP was continued from 1979 to 2015 and restart files were output every day.

Ocean and sea ice initial conditions come from a forced ocean-sea ice configuration of CESM1 that employs adjusted Japanese 55-year reanalysis project (JRA-55) state fields and fluxes (referred to as JRA55-do forcing; Tsujino et al. 2018) as surface boundary conditions. This simulation was spun up through five consecutive cycles of 1958–2009 forcing, with the last cycle extended through 2016. The simulation followed the protocols of the OMIP-2 component of the Coupled Model Intercomparison Project 6 (CMIP6; Griffies et al. 2016; Tsujino et al. 2020), but was an early test of the protocol and hence was not contributed to CMIP6. Preliminary analysis of official OMIP-2 contributions (Tsujino et al. 2020) indicates that the historical ocean and sea ice states obtained are generally as good or better than traditional CORE-II-forced simulations (Griffies et al. 2014), such as the CORE-II ocean–sea ice reconstruction used to initialize the CESM decadal prediction system (Yeager et al. 2018). In particular, the wind fields in the tropical Pacific are substantially better in JRA55-do than in CORE-II (Yeager et al. 2018), leading to more realistic tropical Pacific thermocline and SST variability. Furthermore, the JRA55-do dataset is updated to near–real time, permitting consistent ocean and sea ice state estimation from the late 1950s up to the present.

The atmospheric component of the model was initialized using ERA-Interim reanalysis (Dee et al. 2011) interpolated to CAM5's grid. For climate and decadal prediction, ensemble spread in CESM is typically generated using a very small, round-off level, temperature perturbation in the initial atmospheric condition (e.g., Kay et al. 2015). Such a method does not generate enough spread on the subseasonal time scale (as will be demonstrated below), hence for S2S reforecasts the ensemble spread was generated by implementing the random field (RF) perturbation initialization described in Magnusson et al. (2009) and applied to the atmospheric initial state. RF perturbation involves taking a difference between two random initial conditions, from the same season as the starting month of the simulation, scaling it by a perturbation scaling factor (PSF), and adding and subtracting it from the initial state of the model. Because the perturbations are taken between two prior model states, they are in approximate flow balance. Each random difference therefore generates two ensemble members, and the method can be applied to generate as many ensemble members as are needed. Magnusson et al. (2009) demonstrated that RF initialization is as effective in generating model spread as several other, more sophisticated methods, such as the ensemble transform and singular vectors. We adopt

here the Magnusson et al. (2009) approach by generating 500 initial condition perturbations for each month of the year from daily initial conditions output from the AMIP simulations. For each reforecast date, five initial condition perturbations are randomly chosen out of the 500, scaled by a PSF of 0.15, and added/subtracted to the ERAI initial state to generate the 10 initial conditions used to initialize the ensemble members for the S2S reforecasts. Figure 1a compares the root-mean-square error (RMSE) to the ensemble spread or the ensemble standard deviation of 500-hPa geopotential height anomalies for hindcasts started with the simple temperature perturbation method and the RF perturbation method implemented in CESM1. The figure is shown for 46LCEM1, but it is nearly identical for the 30LCEM1. The figure shows averages over four 45-day hindcasts, each one performed with 10 ensemble members. Only four start dates are averaged as simulations with the simple temperature perturbation were only performed for those dates to demonstrate the difference in spread. The model initial condition interpolated from ERAI reanalysis consists of the zonal and meridional wind, temperature, specific humidity, surface geopotential and surface pressure. All other variables are derived by the model. Time “0” in Fig. 1 shows the first available model output after the model has adjusted to the initial condition. Values at subsequent timesteps represent the daily mean.

Figure 1 clearly illustrates that the simple temperature perturbation method does not provide enough ensemble spread between days 14–25, a key period for subseasonal prediction as the standard deviation of 500-hPa geopotential height (Fig. 1, solid blue curve) is much lower than the RMSE (Fig. 1, blue dashed curve) for that time period. RMSE should match the ensemble spread if the system is statistically consistent (e.g., Leutbecher and Palmer 2008). The ensemble spread using the RF method is significantly bigger (Fig. 1, red solid line) and closer to RMSE (Fig. 1, red dashed line) in the first three weeks of the hindcasts as compared to the small temperature perturbation method; however, the ensemble is still underdispersed. This level of underdispersiveness is a common feature of present day S2S models (e.g., Robertson et al. 2015). Figure 1 also shows that the RMSE of Z500 anomalies, even in the first few days of the reforecasts is quite large despite the fact that the verifying reanalysis is the same as reanalysis used for initialization. This initialization shock is a common problem among S2S systems that do not start from reanalysis created with their own modeling system and undergo interpolation and model adjustment in the first model time step.

Ensemble spread is also important to consider in the stratosphere, especially when looking at stratospheric influences on subseasonal prediction. Figure 1b shows the time evolution of the zonal mean zonal wind at 60°N, 10 hPa for a 1 January 2013 reforecast using the simple temperature perturbation initialization method and the RF perturbation method. When the simple temperature perturbation method is used, there is little difference between the ensemble members (thin blue lines) until day 13 of the forecast, and the spread between ensemble members is quite small through week 4 of

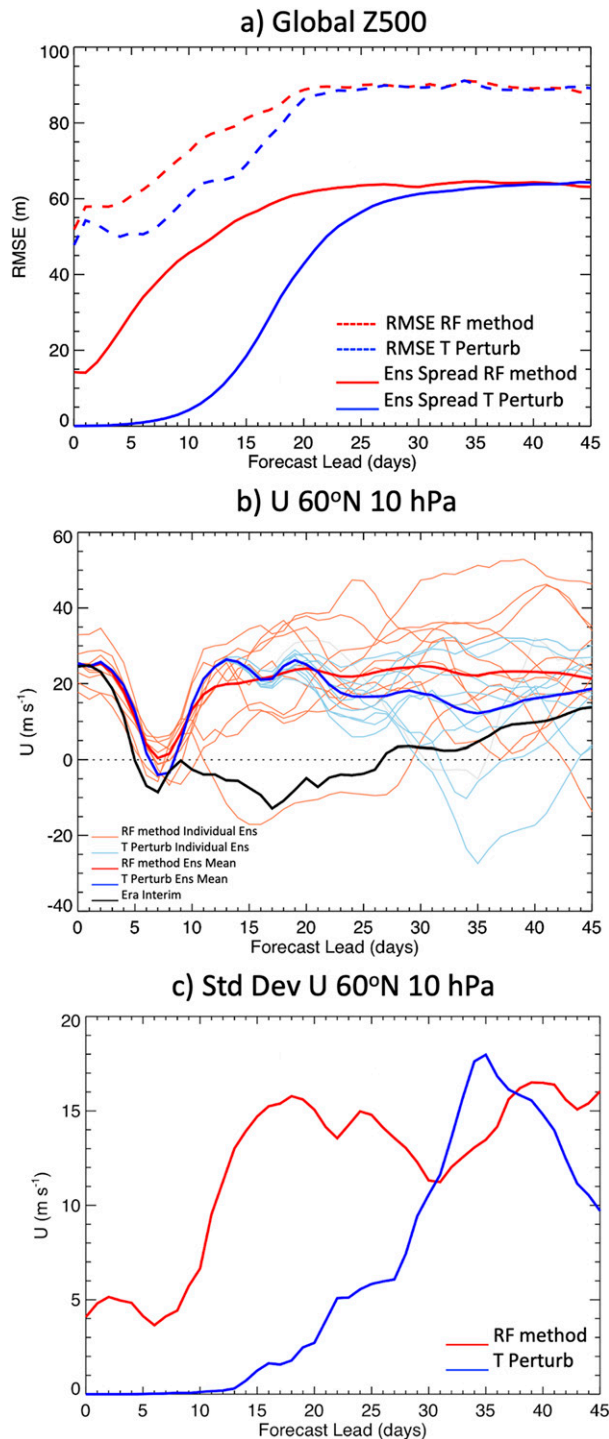


FIG. 1. (a) Ensemble mean root-mean-square error (RMSE; dashed lines) and ensemble spread (or ensemble standard deviation; solid lines) for 500-hPa geopotential height anomalies for 46LCESM1 reforecasts using a temperature perturbation (blue lines) and using the RF method (red lines) to generate ensemble members. Variables are calculated over 10 ensemble members of 45-day-long simulations with start dates of 6 Nov 2012, 5 Dec 2012, 1 Jan 2013, and 5 Feb 2013 (total of 40 simulations). (b) Zonal mean zonal wind at 60°N, 10 hPa for a reforecast with 46LCESM1 started

the forecast as indicated by the blue line in Fig. 1c, which remains at $\sim 0 \text{ m s}^{-1}$ through day 10 of the reforecasts. When the RF perturbation method is applied, there is visible spread between the ensemble members (Fig. 1b, orange thin lines) from the beginning of the reforecast and the ensemble spread, as illustrated by the standard deviation of U zonal wind at 60°N, 10 hPa (Fig. 1c, red line) is much greater between days 15 and 28, as compared to the ensemble spread using simple temperature method.

3. Subseasonal prediction skill

We examine here the subseasonal prediction skill in the S2S reforecasts at the surface (temperature and precipitation) as well as in the stratosphere. We also examine the prediction skill of key modes of variability, the NAO and the MJO.

a. Surface temperature and precipitation

The week 3–4 (day 15–28 averaged) anomaly correlation coefficient (ACC) for 2-m temperature and precipitation over all world land areas for 30LCESM1 is shown in Figs. 2 and 3. CPC Global Daily Temperature dataset and NOAA CPC Global Daily Unified Gauge-Based Analysis of Precipitation dataset, both over land at the horizontal resolution of $0.5^\circ \times 0.5^\circ$, are used as verifying observational datasets. The average daily temperature is calculated as the average of the daily maximum and minimum temperature. For each season, ACC is calculated from anomalies from climatology averaged over days 15–28 for hindcasts that begin during a given season. Very similar ACC values are derived when using ERAI reanalysis as the verifying dataset (not shown). ACC values are shown in colors when they are significantly different from zero at the 5% level (i.e., $\text{ACC} > 0.2$). The significance level is arrived at by considering that for each season there are 13 start dates per year and we examined 17 years, hence a total of 221 samples. We assume a 2-week decorrelation time, and hence the number of independent samples is 110.5. There are hence 108 degrees of freedom, leading to a correlation equal or greater than 0.2 being significant at the 5% level using a two-tailed Student's t test (Wilks 2011). Bottom panels of Figs. 2 and 3 show the ACC of persistence forecasts. Top panels of Fig. 2 show that prediction skill of week 3–4 surface temperature is much higher in DJF than in JJA. In DJF, over all the land areas except for Greenland, ACC for surface temperature is higher than 0.2, and in most areas higher than 0.3. ACC exceeding 0.5 is a common threshold for useful skill (e.g., Murphy and Epstein 1989; Jones et al. 2000). ACC values greater than 0.5 in DJF are found in South East Asia and the northern part of South

←

on 1 Jan 2013 using the temperature perturbation method (light blue lines for individual ensemble members and dark blue for ensemble mean) and RF method (orange lines for individual ensemble members and red line for ensemble mean). Black curve shows ERAI observations. (c) Standard deviation of the zonal mean zonal wind at 60°N, 10 hPa for the same simulations as in (b) (blue: temperature perturbation; red: RF initialization method).

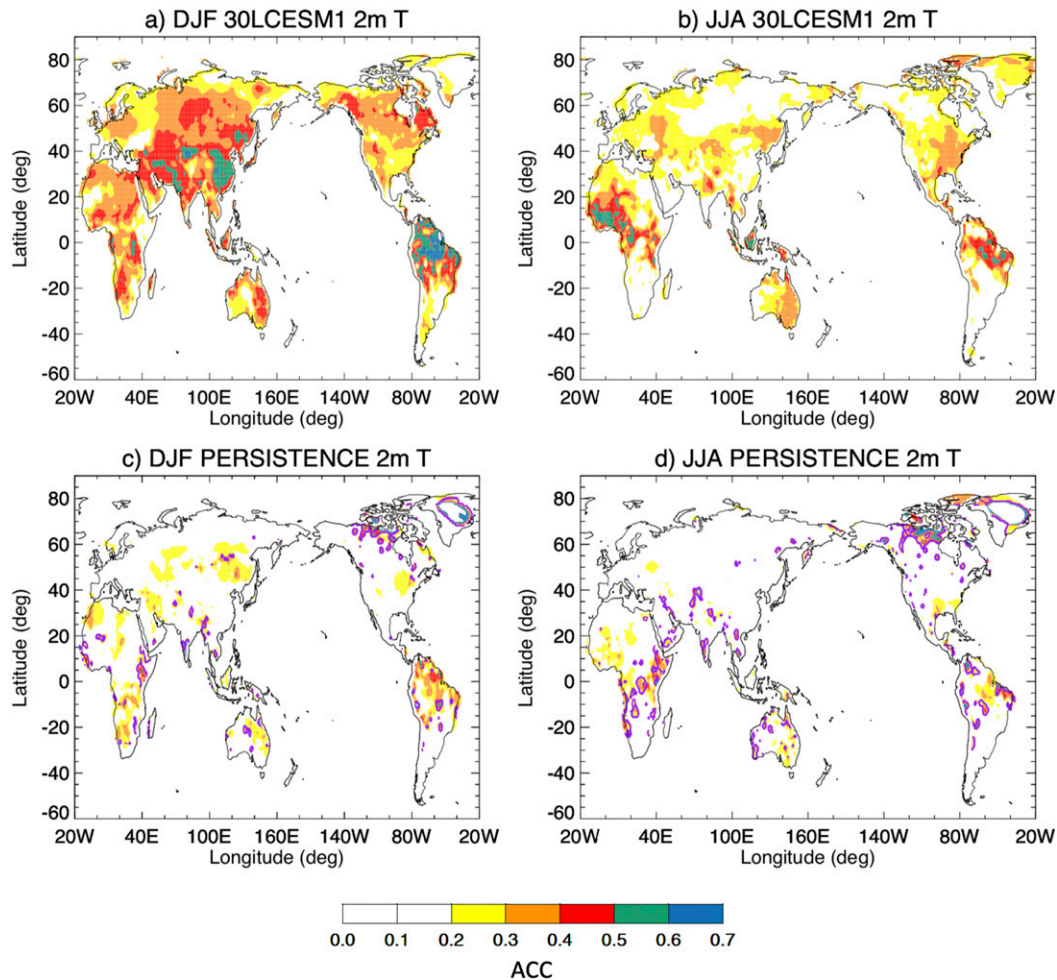


FIG. 2. (a),(c) DJF and (b),(d) JJA 2-m (2m) temperature ACC weeks 3–4 for (top) 30LCEM1 and (bottom) persistence. In (c) and (d) purple contours denote regions for which persistence ACC exceeds the ACC of 30LCEM1.

America. In JJA, areas with highest week 3–4 prediction skill are western Africa, and northern part of South America. Skill in eastern Australia and the southeastern United States is also higher than in most places. In JJA, CESM1 shows the largest 2-m temperature ACC in central Africa and northern part of South America, and part of Australia. There is little skill during this season over Eurasia and large portions of North America. For both DJF and JJA, ACC of 30LCEM1 greatly exceeds that of a persistence except over parts of Greenland. As previous studies have found, precipitation prediction for week 3–4 is quite difficult. ACC in most areas of the world remains below 0.2. In DJF small regions of central Asia, western United States, western and northernmost Australia, and northern South America exhibit most skill (Fig. 2c). In JJA, precipitation skill is lower than 0.2 over the majority of land areas, except for southeastern Australia, and eastern South America. Similarly to 2-m temperature, the persistence skill for precipitation is typically smaller than that of 30LCEM1, with the exception of precipitation over parts of Greenland, which has an ACC of ~ 0.4 in the persistence forecast. ACC was also calculated for 46LCEM1, however,

the difference between surface temperature and precipitation skill calculated over the hindcast period for the two models is negligible (differences of less than 0.05 over majoring of the globe), and the ACC maps are indistinguishable from each other (not shown). This suggests that despite the improved representation of stratospheric variability in 46LCEM1, the seasonally averaged week 3–4 predictive skill of the two CESM1 versions is virtually the same.

The 2-m temperature and precipitation skill of 30LCEM1 and 46LCEM1 averaged over the globe and North America is compared to seven SubX models in Fig. 4 for four seasons: September–October–November (SON), December–January–February (DJF), March–April–May (MAM), and June–July–August (JJA). Figures 4a and 4c illustrate that the 2-m temperature skill of CESM1 for weeks 3–4 is comparable to the NOAA operational model, CFSv2, and higher than the skill of most other models contributing to SubX except for in the JJA season. We arrive at a similar finding for other areas of the world (not shown). Figures 4a and 4c also show that the ACC for 2-m temperature for 30LCEM1 and 46LCEM1 is very similar calculated over the reforecast period, again indicative

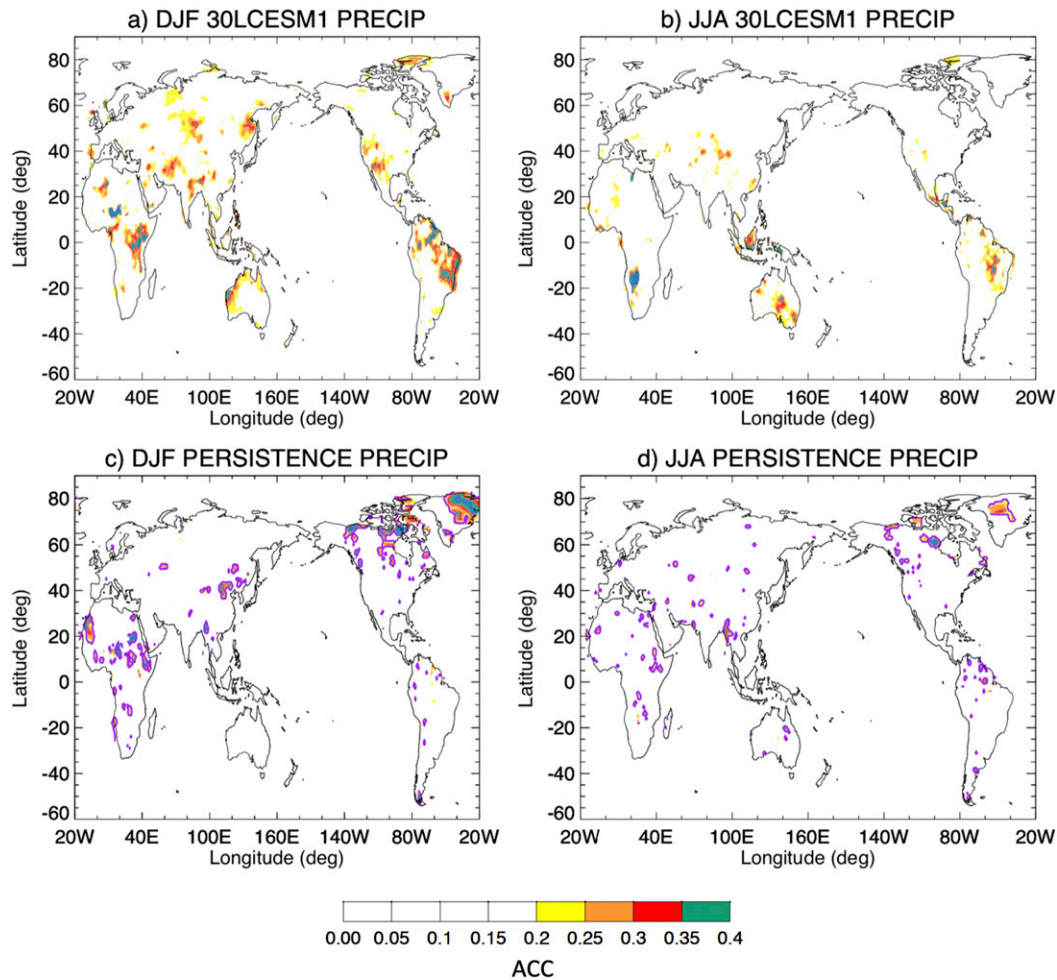


FIG. 3. As in Fig. 2, but for precipitation.

of no overall surface prediction skill gained from the addition of a better resolved stratosphere. Pegion et al. (2019) discussed that the predictive skill of a multimodel ensemble is typically higher than of individual models. The skill of the SubX ensemble is shown in Fig. 4 as MME7. When both versions of CESM1 are added to the multimodel mean, the skill of the multimodel ensemble further increases as illustrated by ACC values for MME9.

Figures 4b and 4d show the ACC for precipitation averaged over the globe and North America for CESM1 and SubX models. Over North America, there is very little skill in the week 3–4 forecast. The ACC is over 0.1 only for DJF for the NOAA CFSv2 model, and the multimodel mean of the seven SubX models, and the multimodel mean including SubX models and CESM1. The values are a little higher when the entire globe is considered. Although there are differences in ACC between 30LCEM1 and 46LCEM1 in MAM, we do not examine these further due to the overall low predictive skill of precipitation.

b. Comparing forecast skill

The sign test is one way to compare forecast skill and is considered a “fair” test in the sense that there is no expectation

one model will do better than another model (DelSole and Tippett 2014). The Brier skill score (BSS) for each forecast in the reforecast period was compared between CFSv2 and 30LCEM1 and 46LCEM1, between CFSv2 and each individual SubX model, as well as between CFSv2 and the SubXMME. If CFSv2 had the better BSS, a +1 was assigned. If CFSv2 had a lower BSS, a –1 was assigned. This process was repeated for all of the forecasts in the reforecast database and the cumulative count of the results can be summarized as probabilities for providing a skillful forecast. Figure 5 shows the probability of providing a skillful forecast (given as a percentage on the y axis) for 30LCEM1 and 46LCEM1, each individual SubX model, and the SubXMME, relative to CFSv2. The percentages for models marked by blue bars are considered statistically significantly likely to provide a more skillful forecast and the remaining values indicate that those models are equally or less skillful, relative to CFSv2. These results are statistically significant at the 95% confidence interval and show that 30LCEM1 (46LCEM1) has a 62% (65%) chance of providing a more skillful temperature forecast than CFSv2, demonstrating that this system is a reliable tool for subseasonal

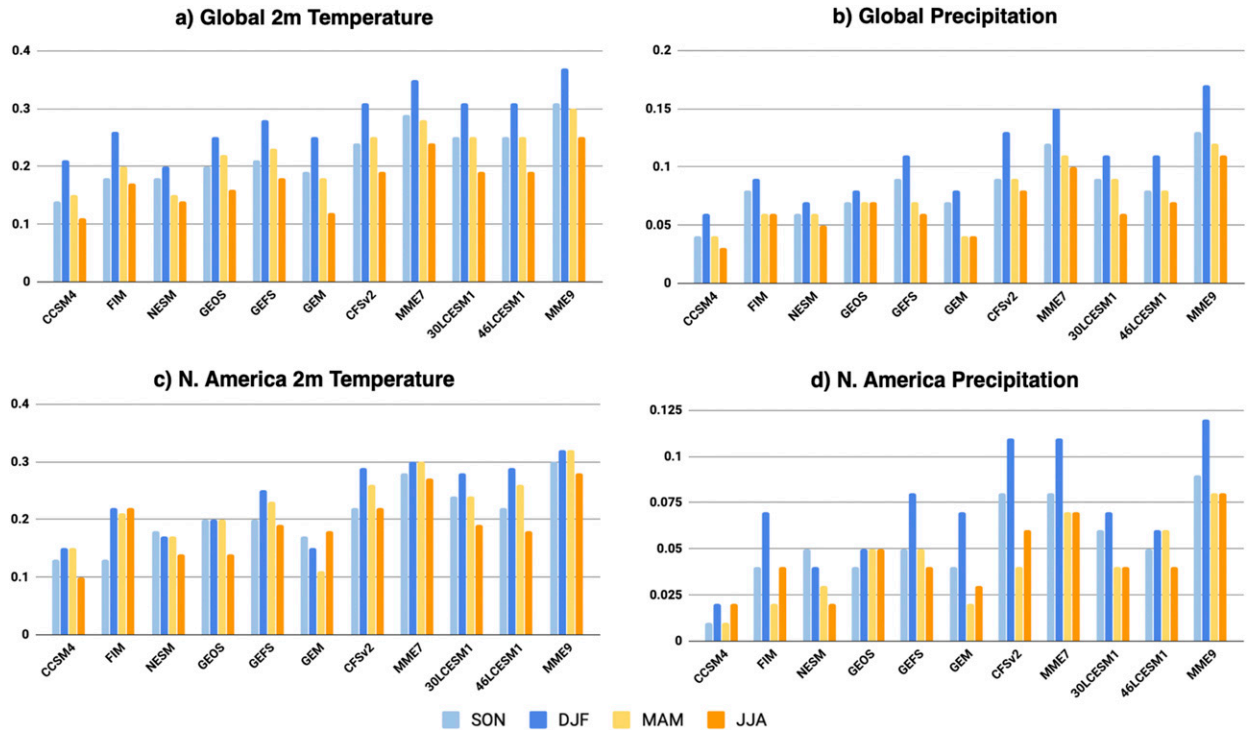


FIG. 4. ACC for week 3 and 4 (day 15–28) averaged (a),(c) 2-m temperature and (b),(d) precipitation averaged (top) over the globe and (bottom) North America for the SubX models (CCSM4, FIM, NESM, GEOS, GEF5, GEM, CFSv2), multimodel mean of the seven SubX models (MME7), 30LCEM1, 46LCEM1, and multimodel mean of SubX models and both CESM models (MME9) averaged over four seasons: SON (light blue), DJF (blue), MAM (yellow), and JJA (orange). SubX models are described in Pegion et al. (2019).

predictability research, often exceeding the operational forecast. Note that the number of ensemble members used by each modeling center is different (e.g., only 4 for CFSv2), which can affect the skill of the forecasts.

c. Stratosphere

We expected the 46LCEM1 to have higher prediction skill than 30LCEM1 due to a better-resolved stratosphere; however, so far we have not seen evidence of this. We compare here the prediction skill of the dominant modes of stratospheric variability and stratospheric-tropospheric coupling, the QBO, and the NH winter polar vortex along with SSWs in 30LCEM1 and 46LCEM1 to see whether the stratosphere itself is predicted better in the 46LCEM1. We use the ERAI dataset (Dee et al. 2011) for verification. We expect the predictability of the QBO to be improved in 46LCEM1 due to the ability of that model to produce a QBO in a climate simulation. SSWs arise from the breaking of planetary waves in the polar vortex (e.g., McIntyre 1982, Baldwin et al. 2018) and their occurrence, although still not fully understood, depends both on the stratosphere and troposphere (Hitchcock and Haynes 2016). We expect the predictability of SSWs to be improved in 46LCEM1 as compared to 30LCEM1, especially between December and January, as in AMIP simulations the SSW frequency is more realistic in that model.

We define the QBO index as the zonal mean zonal wind anomalies from climatology at 50 hPa averaged between 5°S

and 5°N. The mean zonal mean wind at 50 hPa between 5°S and 5°N is approximately 1 m s^{-1} hence there is little difference as to whether the QBO is defined on the raw zonal mean wind or anomalies. We calculate the ACC for all years as well as for QBO westerly and QBO easterly dates separately (Fig. 6). This is done by calculating the ACC from daily data for forecast lead of 1–45 days for either all the available start dates, or only the dates during which the QBO was in the westerly (easterly) phase during initialization. QBO westerly (easterly) years are defined as years in which the zonal mean wind at 50 hPa are above (below) 0 m s^{-1} . Different thresholds to define the QBO phase yield similar results (not shown). Figure 6a shows that the overall prediction skill of the QBO on the subseasonal time scale is quite high, with the ACC for all QBO phases at day 30 at 0.97 for the 46LCEM1 and 0.94 for 30LCEM1. Due to the lack of internally generated QBO in 30LCEM1, we expect the QBO to be predicted better in the 46LCEM1, which it is (Fig. 6a). However, 30LCEM1, despite faster decay of QBO amplitude, still has very high skill ($\text{ACC} > 0.9$) throughout the entire forecast period. This implies that the initialized state of the QBO persists throughout the forecast on a subseasonal time scale and remains close to observations regardless of whether the model used is capable of internally generating a QBO. The prediction skill of the QBO by CESM1 varies depending on its phase. Figure 6b shows that the prediction skill of the QBO is lower in the westerly phase than in the QBO easterly phase especially for 30LCEM1: the ACC is ~ 0.96

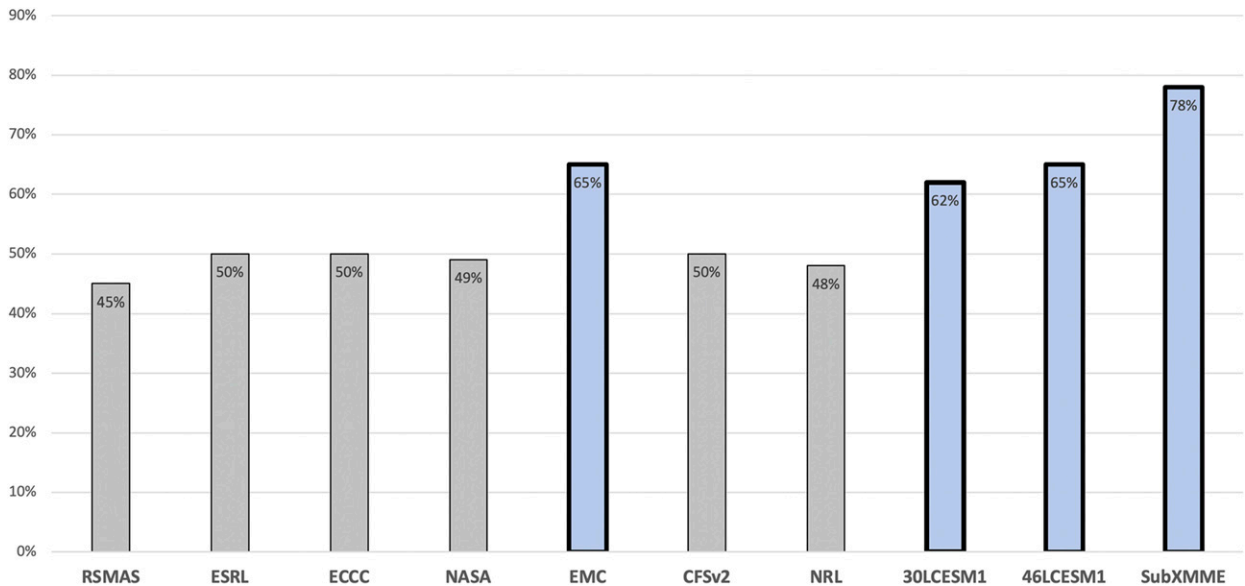


FIG. 5. Percent-chance a model will have a *more*, *less*, or *equally* skillful 2-m temperature forecast relative to the CFSv2. Results are based on comparing Brier skill scores and are significant at the 95% confidence interval. Values in blue bars indicate that the model has a statistically significant chance of providing a more skillful forecast than CFSv2. The 30LCESM1 (46LCESM1) has a 62% (65%) chance of providing a more skillful temperature forecast than CFSv2.

(0.9) for 46LCESM1 (30LCESM1) in QBO westerly and ~ 0.98 (0.97) in QBO easterly at day 30. The 30LCESM1 hence has difficulties sustaining the westerly phase of the QBO. Our findings on the prediction skill of the QBO on subseasonal time scale are consistent with a recent study by Lim et al. (2019) who found a 1-month QBO correlation between 0.95 and 1.0 for high-top models, and a lower prediction skill 0.85–0.95 for models with a lower top and poor representation of the stratosphere.

Figure 7 shows the ACC of DJF zonal mean wind at 60°N and 10 hPa. Similarly to the QBO, the ACC was calculated for forecast lead days 1–45 for each hindcast start date during the DJF season. The figure shows that the strength of the polar vortex winds averaged over the entire winter season is predictable at 34 days in 46LCESM1 and only out to 28 days in 30LCESM1. This higher prediction skill for the NH polar vortex in 46LCESM1 is also found in other seasons (not shown). During the reforecast period between 1999 and 2015

there were 11 SSWs in DJF between 1999 and 2015 (Table 1). Tripathi et al. (2016) showed that all models examined were able to predict the 2012/2013 winter warming by all ensemble members when initialized 5 days before the central warming date of the event, and 50% to 100% of the ensemble members considered were able to predict an SSW when initialized 10 days before the observed event. In the CESM1 reforecast set, the initial day of the forecast is most of the time a Wednesday, hence forecasts closest to each SSW event are started either 2–6 days prior (see Table 1), and since the forecasts are ran weekly, initializations prior to that are 9–13 days before each SSW event, making the assessment of SSW predictability difficult to compare. Table 1 summarizes the SSWs examined in CESM1 and reforecast initialization dates relative to the central warming dates, which are defined as the dates on which the zonal mean zonal wind at 60°N and 10 hPa becomes easterly. Figure 8 summarizes the prediction skill of these SSWs for 30LCESM1 and

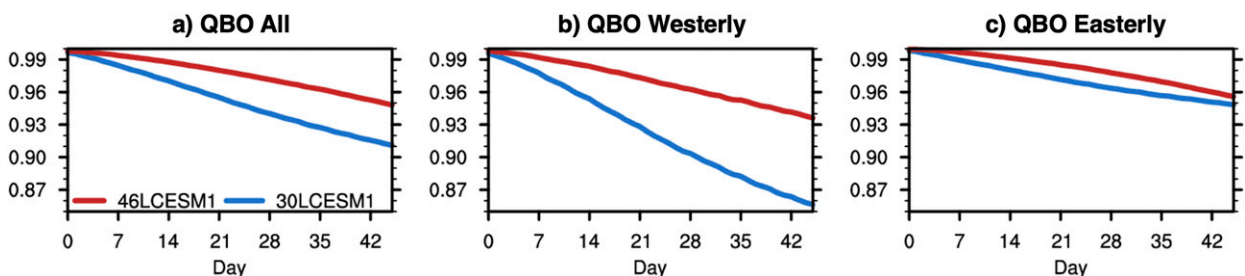


FIG. 6. Daily correlation coefficient of the quasi-biennial oscillation (QBO) between model and observations for 30LCESM1 and 46LCESM1 for (a) all dates, (b) QBO westerly dates, and (c) QBO easterly dates. The QBO is defined as the zonal mean zonal wind anomalies from climatology at 50 hPa averaged between 5°S and 5°N.

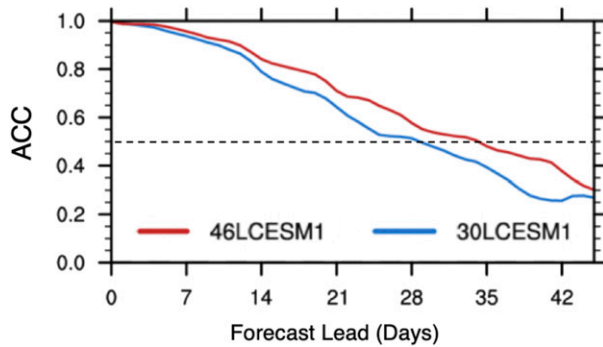


FIG. 7. Anomaly correlation coefficient (ACC) for DJF 10 hPa, 60°N zonal mean zonal wind for 30LCEM1 (blue) and 46LCEM1 (red). Dashed line represents the ACC value of 0.5.

46LCEM1. The time evolution of the ensemble mean zonal mean zonal wind at 60°N and 10 hPa for the individual events is shown in Fig. S3 and shows that they are in general similar between 30L and 46L CESM1. Figure 8 shows the number of individual ensemble members that predict a SSW within 3 days of the observed SSW. While the majority of ensemble members typically capture SSWs for initializations of 5 days or less before the central warming date, only 30% of the ensemble members are able to predict a SSW 9–13 days out. SSW prediction skill decreases further for initialization dates exceeding 15 days before the events, except for the 26 February 1999 and 22 February 2008 SSWs, in which 40%–50% of the ensemble members predict an SSW (Fig. S3). Predictability of SSWs in CESM1 is broadly similar to previous studies (Tripathi et al. 2016; Marshall and Scaife 2010). Figure 8 shows no clear evidence that 46LCEM1 predicts SSWs any better than 30LCEM1: in some instances an SSW event is predicted better by 30LCEM1 and in some by 46LCEM1. This suggests that the improved representation of the stratosphere in 46LCEM1 does not necessarily lead to better prediction of SSW events, which could partially explain why there is not clear difference in surface prediction skill between the two models.

TABLE 1. Central date of SSWs that occurred in DJF between 1999 and 2015 and initial date of three reforecast dates (R1, R2, and R3) that were initiated prior to the SSW. Reforecasts dates are shown in days relative to each SSW date.

	SSW date	R1	R2	R3
1	26 Feb1999	-16	-9	-2
2	11 Feb2001	-18	-11	-4
3	30 Dec 2001	-18	-11	-4
4	18 Jan 2003	-17	-10	-3
5	5 Jan 2004	-19	-12	-5
6	21 Jan 2006	-17	-10	-3
7	24 Feb 2007	-17	-10	-3
8	22 Feb 2008	-16	-9	-2
9	24 Jan 2009	-17	-10	-3
10	9 Feb 2010	-20	-13	-6
11	6 Jan 2013	-18	-11	-4

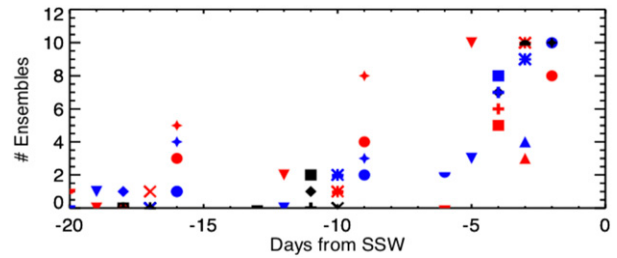


FIG. 8. Summary of the prediction skill of SSWs for different forecast initiations (individual SSW events are shown in Fig. S3). Symbols indicate the number of ensemble members that predicted an SSW event within 3 days from the actual observed event. Different symbols are used to distinguish between the different SSWs events and have no particular meaning. Blue (red) symbols are for 30LCEM1 (46LCEM1). When the number of ensemble members predicting SSW is the same in both models, the symbols are black.

d. NAO and MJO

The NAO and the MJO are two modes of variability that have been shown to be affected by stratosphere–troposphere coupling. We examine these in 30LCEM1 and 46LCEM1 to determine if the improved representation of the stratosphere lead to changes in prediction skill of the NAO and the MJO. The NAO is a key driver of winter climate and associated extremes over Europe and North America (Hurrell 1995; Scaife et al. 2008). Until recently, long-range forecast systems showed low levels of skill in predicting fluctuations in the NAO at seasonal lead times (e.g., Johansson 2007); however, recent seasonal forecasting systems have demonstrated significant skill in predicting the NAO when initialized a month before the beginning of winter (Scaife et al. 2014a,b; Riddle et al. 2013; Stockdale et al. 2015). Pegion et al. (2019) showed that the individual SubX models, as well as their multimodel mean, exhibit ACC > 0.5 for the NAO through the week 2 forecasts; however, the prediction skill drops off quickly after that. Figure 9 shows the ACC for the NAO for 30LCEM1 and 46LCEM1. The NAO index was obtained by first calculating EOF analysis of ERA-Interim monthly (December–February) sea level pressure anomalies over the Atlantic sector (20°–90°N, 90°W–60°E) and treating the leading EOF pattern as the NAO. The NAO index was then calculated by projecting the SLP anomaly in the reanalysis and model onto the leading EOF. Figure 9 shows that CESM1 has high NAO prediction skill in week 1 and week 2 (ACC of ~0.95 and 0.75, respectively). The NAO prediction skill is ~0.5 for week 3, and 0.4 for week 4, exceeding the NAO prediction skill for that time frame by other SubX models (not shown, see Pegion et al. 2019 for figure). Although CESM1 exhibits high NAO skill, Fig. 9 shows clearly that there is no significant difference between the overall NAO prediction skill between the 30LCEM1 and 46LCEM1.

The MJO is the dominant mode of the tropical intraseasonal variability and considered as the leading source of subseasonal predictability in the tropics as well as in extratropics (e.g., Stan et al. 2017). The MJO prediction skills, such as the anomaly

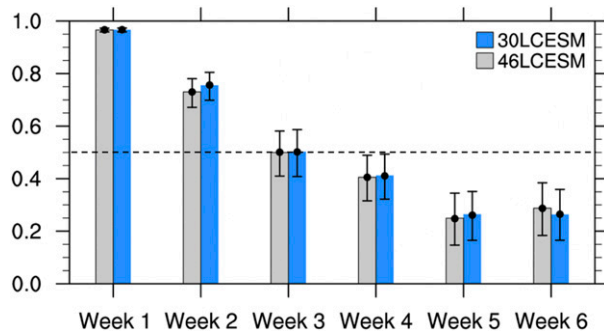


FIG. 9. ACC of the weekly mean NAO at increasing lead times during NDJFM for 30LCESM (blue bars) and 46LCESM (gray bars). Error bars indicate the 5% and 95% statistical levels based on bootstrapping. NAO is calculated by projecting the 1000-hPa geopotential height anomaly onto the NDJFM monthly leading EOF pattern over the North Atlantic region (20° – 90° N, 90° W– 60° E).

bivariate correlation coefficient and the RMSE, are usually defined by comparing the predicted and observed real-time multivariate MJO indices (RMM; [Wheeler and Hendon 2004](#)). With these measures, CESM1 is categorized in a high skill group, where the MJO prediction skill is about 27 days defined by the day when ACC reaches 0.5 ([Kim et al. 2019](#)). Each of 10 ensemble members from 30LCESM1 and 46LCESM1 were combined in [Kim et al. \(2019\)](#) due to the very similar characteristics and skill in MJO forecasts in the two models. [Figure 10](#) compares the ACC and RMSE of ensemble mean RMM indices in each model as a function of forecast lead days during boreal winter verified against the RMM indices obtained by ERAI. Details about the datasets and verification methods are described in [Kim et al. \(2019\)](#). Here, we select MJO samples when the amplitude of the RMM indices exceeds 1.0 at the initialized date that ranges from October to March (total 262 MJO samples). Although the ACC (RMSE) in 30LCESM1 is slightly higher (lower) than 46LCESM1 up to about 5 weeks, the difference of skill in the two models are negligible. We have also examined the MJO prediction skill during QBO westerly and during QBO easterly conditions separately. We found the MJO skill to be similar in both QBO phases in the 30LCESM1 and 46LCESM1.

4. Model biases

Although CESM1 has overall good predictive skill on sub-seasonal time scales, we compare here the biases in temperature, wind, and precipitation between an AMIP simulation performed with CAM5 extended until the end of year 2015 and S2S reforecasts for the 30LCESM1 to begin to understand how inadequacies in model physics might be contributing to decreasing predictive skill. At the subseasonal time-scale model biases could also be caused by an imbalance between the ocean state and atmospheric state, and errors in the ocean initial state. Surface temperature and precipitation biases for 46LCAM AMIP simulation and 46LCESM1 reforecasts are indistinguishable from those for 30LCAM AMIP and 30LCESM1 reforecasts, respectively, and hence are not shown. The top two rows of [Fig. 11a](#) show

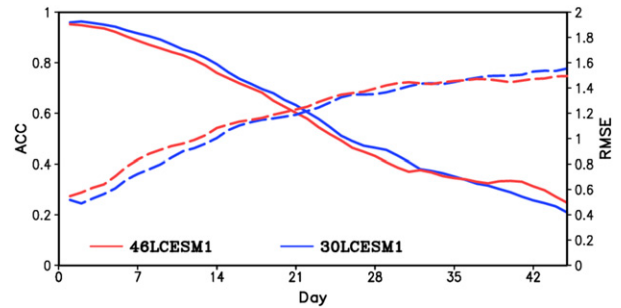


FIG. 10. MJO prediction skill for 30LCESM1 (blue) and 46LCESM1 (red) hindcasts initialized with strong MJO events during boreal winter (ONDJFM). The prediction skill is evaluated based on the anomaly correlation coefficient (ACC; solid lines) and root-mean-squared error (RMSE; dashed lines) between the model and observation RMM indices.

that CAM5 has a primarily cold bias over most of the land areas except for central Eurasia, central United States, and southeast South America, and southwest Australia in DJF. In JJA, CAM5 has a warm bias over the majority of Eurasia (up to 3° C), and throughout most of the United States (up to 5° C). There are smaller biases, mostly cold, in South America, Africa, and Australia in JJA. [Figures 11b,c](#) show the biases calculated from S2S hindcasts for lead times of 1 week and 4 weeks, respectively. The patterns of temperature biases for both DJF and JJA are very similar between the S2S hindcasts and AMIP runs indicating that these come primarily from the representation of atmospheric physics in CAM5. The magnitude of biases in the S2S hindcasts grows with time in the majority of areas as indicated by differences between week 4 and week 1 ([Fig. 11d](#)), but the pattern of biases is established in the first week of simulation.

Bottom two rows in [Fig. 11](#) show precipitation biases in a fashion similar to the temperature biases discussed above. In DJF, the AMIP simulation shows primarily wet biases with largest excess precipitation in southern Africa, majority of Australia, the northwest United States and near the west coast of South America. Dry biases are found in north-central South Africa and near the southeast corner of the United States. The pattern of biases is nearly perfectly reproduced in the S2S hindcasts at lead times of week 1 and week 4 ([Figs. 11b,c](#)) with the amplitude of biases at week 4 reaching the values of biases in AMIP simulations. In JJA, wet biases in CAM5 AMIP simulations are primarily in central Africa, and along the south and eastern coast lines of Asia, as well as in the northwestern United States. Dry biases extend through in the north–south direction through the central United States, and in Northern South America. Similarly to DJF, precipitation biases in the S2S hindcasts exist in the same location by week 1. Magnitudes of precipitation biases are comparable in magnitude to the biases in AMIP simulations already at week 1 and do not grow much more especially in DJF.

Our findings on the similarity of biases in climate-length simulations and in shorter forecasts are in agreement with previous similar studies. For example, [Xie et al. \(2012\)](#) examined systematic errors in moist-process related fields and

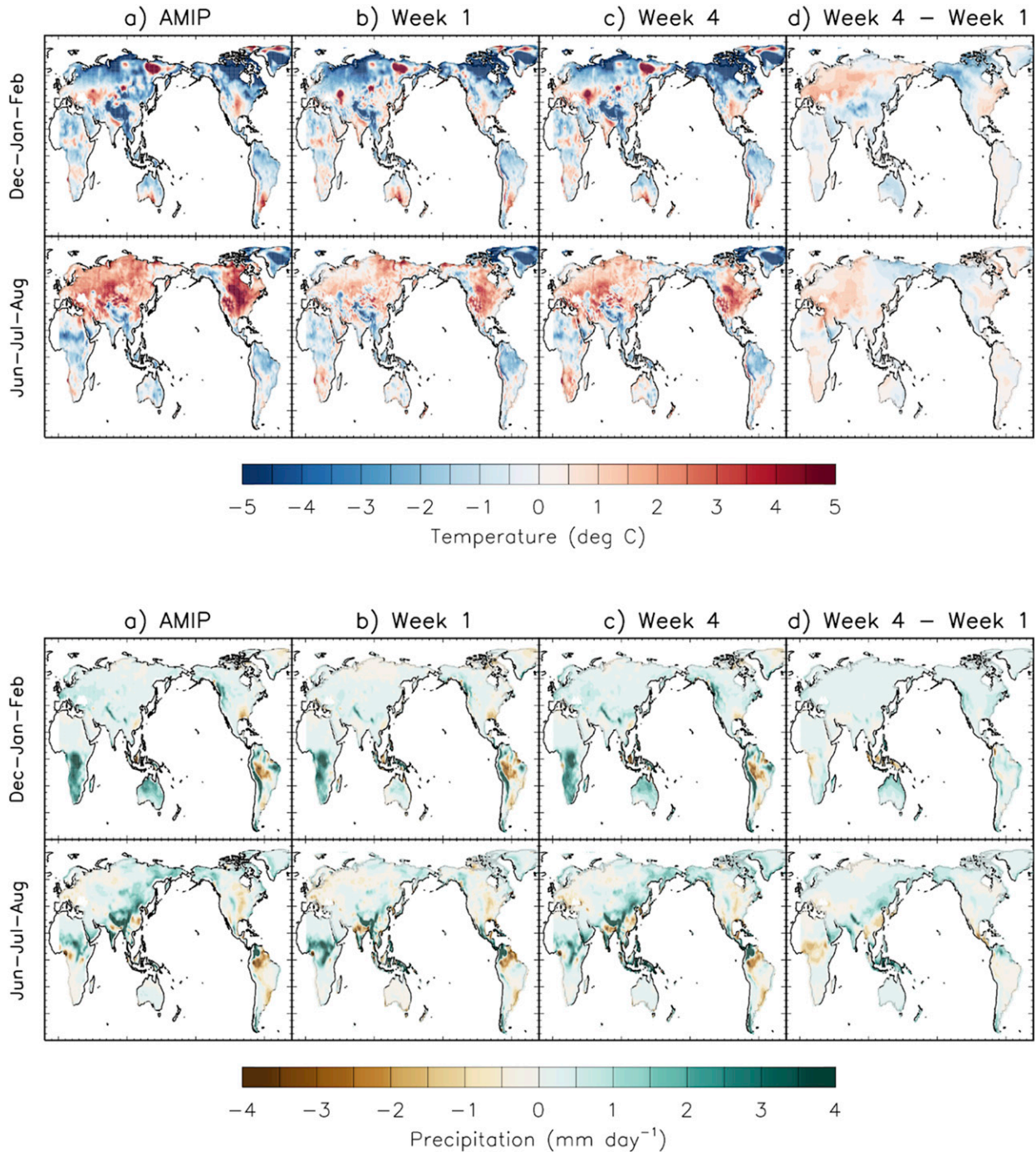


FIG. 11. (top rows) Temperature and (bottom rows) precipitation biases (model minus observations) calculated from (a) AMIP simulations with 30LCAM5, (b) 30LCESM1 S2S hindcasts week 1, (c) 30LCESM1 S2S hindcasts for week 4, and (d) S2S hindcasts week 4–week 1. Biases are shown for winter (DJF) in the first and third row and summer (JJA) in the second and fourth row. Observed data for 2-m temperature over land is the CPC daily temperature dataset with horizontal resolution of $0.5^\circ \times 0.5^\circ$. For precipitation over land, the CPC Global Daily Precipitation dataset ($0.5^\circ \times 0.5^\circ$) is used (Xie et al. 2007; Chen et al. 2008). Verification datasets are regridded to the coarser SubX resolution. Comparisons are for the time period between 1999 and 2015.

surface temperature and has shown that a strong correlation exists between hindcasts errors and climate errors in CAM4 and CAM5. Some of the errors, for example in tropospheric temperature, water vapor, and precipitation develop quite fast and were apparent in just the day 2 hindcasts, suggesting that these errors are likely a result of model parameterization errors. [Ma et al. \(2014\)](#) examined the correspondence between short- and long-term systematic errors in five general circulation models. Their analysis showed that most systematic errors of precipitation, clouds, and radiation processes in the long-term runs were present by day 4 in the short-term hindcasts in all models. In addition, the amplitude of the biases was comparable to that of long-term biases within a few days, again suggesting that it is the fast-physics that contribute most to the biases.

5. Summary and future directions

We have used here the Coupled Earth System Model (CESM1), typically used for studies of climate change and variability to study subseasonal prediction. We have performed two sets of S2S reforecasts with CESM1, one with the default 30-level and a 46-level version with an improved representation of the stratosphere, in particular with a better representation of SSWs and the QBO. Each hindcasts set followed the SubX protocol with weekly initializations from 1999 to 2015 and consisted of 10 ensemble members in each reforecast.

We have shown that surface temperature and precipitation prediction skill of CESM1 on subseasonal time scales (weeks 3–4) is comparable to the skill of operational S2S models despite the relatively simple initialization techniques and lack of data assimilation. Furthermore, the addition of CESM1 to the ensemble of SubX models raises the skill of the multimodel ensemble.

We hypothesized that an inclusion of a better resolved stratosphere would lead to improved subseasonal prediction skill. The 46-level CESM1 shows a higher prediction skill than the 30-level CESM1 of stratospheric variables, including polar vortex winds and QBO. However, there is no clear improvement from using the higher-top model on the time scales of 3–4 weeks at the surface. The predictability of SSWs, the NAO, and the MJO is also nearly the same in both versions of the model. In addition [Kim et al. \(2019\)](#) showed that there is no difference between the prediction skill of the MJO between QBOE and QBOW phases in these two model versions. This could be in part due to the fact that the QBO even in the 46-level CESM1 does not reach sufficiently down into the lower stratosphere, and hence the observed influences of the QBO on the MJO are not reproduced, or the mean state of the MJO in CESM1 is not reproduced well enough to begin with, and hence the influence of the QBO is not noticeable. Consistent with our study, [Domeisen et al. \(2019a\)](#) found that high-top models have higher subseasonal skill in the stratosphere compared to low-top models. [Domeisen et al. \(2019b\)](#) also found that despite the significant surface impact of the stratosphere, enhanced predictability of surface temperature linked to weak and strong vortex events was difficult to demonstrate.

The lack of improved surface prediction with the 46LCESM1 as compared to 30LCESM1 does not mean that the stratosphere is not important to surface prediction. As was demonstrated here and in [Kim et al. \(2019\)](#), the mean stratospheric state between weeks 3–4 of reforecasts is quite similar to the initial condition in both versions of CESM1, and anomaly correlation coefficients of stratospheric variables in both models are high. This is due to the slow-varying nature of the stratosphere. Hence although the stratospheric state is predicted a little better in the 46-level version of CESM1, the initial state of the stratosphere persists through the subseasonal time scale in both versions of the model. As a result, the benefit of a better resolved stratosphere does not translate to improvement in surface prediction skill. This is unlikely the case on seasonal and longer time scales, as the stratospheric state will depart from observations much faster in the low-top model. In other words, since the stratosphere is slowly varying, especially the QBO, which has a time scale of ~ 28 months, the initial observed state of the stratosphere persists throughout the duration of the subseasonal forecast window, regardless of the quality of the representation of the stratosphere. This is likely the reason for the similarity in skill between the two versions of CESM1. Last, in this work we showed that week 1 biases of initialized hindcasts are the same as in multidecadal simulations with prescribed sea surface temperatures, pointing to the fact that fast-physics parameterization contributes most to the biases of the system.

Simulations with CCSM4 carried out as part of the SubX project and the simulations carried out here with CESM1 demonstrate the utility of a community climate model as a useful research framework for subseasonal prediction that could be used by a broader community. The caveat being that subseasonal hindcasts are very computationally expensive, and the infrastructure of running the system is not trivial with CESM1 and not easily transferable to individual users, and not extendable to present day (due to initial condition issues for what is now an outdated land model). However, at the time of publication of this manuscript, S2S hindcasts are being carried out with the latest version of CESM, CESM2 ([Danabasoglu et al. 2020](#)), with the goal of providing the community a research framework for subseasonal prediction research including relevant datasets and infrastructure for performing additional simulations that could help to understand sources of predictability.

Acknowledgments. This work was supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement 1852977 and by NOAA's Climate Program Office (CPO) Modeling, Analysis, Predictions and Projections (MAPP). Portions of this study were supported by the Regional and Global Model Analysis (RGMA) component of the Earth and Environmental System Modeling Program of the U.S. Department of Energy's Office of Biological and Environmental Research (BER) via National Science Foundation IA 1844590, and the NOAA Climate Program Office Climate Variability and Predictability Program. We thank Judith Perlwitz for useful discussions in the initial stages of this work. HK was

supported by NSF Grant AGS-1652289 and KMA R&D Program Grant KMI2018-03110. KP was supported by NOAA/OWAQ SPC-000940 and NOAA/MAPP NA16OAR4310146.

Data availability statement. CESM1 and SubX hindcasts are freely available from the IRI data library: <https://iridl.ideo.columbia.edu/SOURCES/.Models/.SubX/>. CPC precipitation and temperature, NCEP–NCAR Reanalysis, and NOAA interpolated OLR data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, from their website at www.esrl.noaa.gov/psd/. ERAI can be downloaded from <https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim>.

REFERENCES

- Baldwin, M. P., and T. J. Dunkerton, 2001: Stratospheric harbingers of anomalous weather regimes. *Science*, **294**, 581–584, <https://doi.org/10.1126/science.1063315>.
- , D. B. Stephenson, D. W. J. Thompson, T. J. Dunkerton, A. J. Charlton, and A. O'Neill, 2003: Stratospheric memory and skill of extended-range weather forecasts. *Science*, **301**, 636–640, <https://doi.org/10.1126/science.1087143>.
- , and Coauthors, 2018: 100 years of progress in understanding the stratosphere and mesosphere. *A Century of Progress in Atmospheric and Related Sciences: Celebrating the American Meteorological Society Centennial*, *Meteor. Monogr.*, No. 59, 27.1–27.62, <https://doi.org/10.1175/AMSMONOGRAPH5-D-19-0003.1>.
- Butler, A., and Coauthors, 2019: Sub-seasonal predictability and the stratosphere. *Sub-seasonal to Seasonal Prediction*, A. W. Robertson and F. Vitart, Eds., Elsevier, 585 pp.
- Charlton, A. J., and L. M. Polvani, 2007: A new look at stratospheric sudden warmings. Part I: Climatology and modeling benchmarks. *J. Climate*, **20**, 449–469, <https://doi.org/10.1175/JCLI3996.1>.
- Chen, M., W. Shi, P. Xie, V. B. S. Silva, V. E. Kousky, R. W. Higgins, and J. E. Janowiak, 2008: Assessing objective techniques for gauge-based analyses of global daily precipitation. *J. Geophys. Res.*, **113**, D04110, <https://doi.org/10.1029/2007JD009132>.
- Danabasoglu, G., S. C. Bates, B. P. Briegleb, S. R. Jayne, M. Jochum, W. G. Large, S. Peacock, and S. G. Yeager, 2012: The CCSM4 ocean component. *J. Climate*, **25**, 1361–1389, <https://doi.org/10.1175/JCLI-D-11-00091.1>.
- , and Coauthors, 2020: The Community Earth System Model version 2 (CESM2). *J. Adv. Model. Earth Syst.*, **12**, e2019MS001916, <https://doi.org/10.1029/2019MS001916>.
- Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, <https://doi.org/10.1002/qj.828>.
- DelSole, T., and M. K. Tippett, 2014: Comparing forecast skill. *Mon. Wea. Rev.*, **142**, 4658–4678, <https://doi.org/10.1175/MWR-D-14-00045.1>.
- Dennis, J. M., and Coauthors, 2012: CAM-SE: A scalable spectral element dynamical core for the community atmosphere model. *Int. J. High Perform. Comput. Appl.*, **26**, 74–89, <https://doi.org/10.1177/1094342011428142>.
- Domeisen, D. I. V., and Coauthors, 2019a: The role of the stratosphere in subseasonal to seasonal prediction. Part I: Predictability of the stratosphere. *J. Geophys. Res. Atmos.*, **124**, e2019JD030923, <https://doi.org/10.1029/2019JD030920>.
- , and Coauthors, 2019b: The role of the stratosphere in sub-seasonal to seasonal prediction. Part II: Predictability arising from stratosphere–troposphere coupling. *J. Geophys. Res. Atmos.*, **124**, e2019JD030923, <https://doi.org/10.1029/2019JD030923>.
- Griffies, S. M., and Coauthors, 2014: An assessment of global and regional sea level for years 1993–2007 in a suite of interannual CORE-II simulations. *Ocean Modell.*, **78**, 35–89, <https://doi.org/10.1016/j.ocemod.2014.03.004>.
- , and Coauthors, 2016: OMIP contribution to CMIP6: Experimental and diagnostic protocol for the physical component of the ocean model intercomparison project. *Geosci. Model Dev.*, **9**, 3231–3296, <https://doi.org/10.5194/gmd-9-3231-2016>.
- Hitchcock, P., and P. Haynes, 2016: Stratospheric control of planetary waves. *Geophys. Res. Lett.*, **43**, 11 884–11 892, <https://doi.org/10.1002/2016GL071372>.
- Hurrell, J. W., 1995: Decadal trends in the North Atlantic Oscillation: Regional temperatures and precipitation. *Science*, **269**, 676–679, <https://doi.org/10.1126/science.269.5224.676>.
- , J. J. Hack, D. Shea, J. M. Caron, and J. Rosinski, 2008: A new sea surface temperature and sea ice boundary dataset for the Community Atmosphere Model. *J. Climate*, **21**, 5145–5153, <https://doi.org/10.1175/2008JCLI2292.1>.
- , 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>.
- Johansson, Å., 2007: Prediction skill of the NAO and PNA from daily to seasonal time scales. *J. Climate*, **20**, 1957–1975, <https://doi.org/10.1175/JCLI4072.1>.
- Jones, C., D. E. Waliser, J. K. E. Schemm, and W. K. M. Lau, 2000: Prediction skill of the Madden and Julian Oscillation in dynamical extended range forecasts. *Climate Dyn.*, **16**, 273–289, <https://doi.org/10.1007/s003820050327>.
- Kay, J. E., and Coauthors, 2015: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Amer. Meteor. Soc.*, **96**, 1333–1349, <https://doi.org/10.1175/BAMS-D-13-00255.1>.
- Kim, H., J. H. Richter, and Z. Martin, 2019: Insignificant QBO–MJO prediction skill relationship in the SubX and S2S sub-seasonal reforecasts. *J. Geophys. Res. Atmos.*, **124**, 12 655–12 666, <https://doi.org/10.1029/2019JD031416>.
- Kirtman, B. P., and Coauthors, 2014: The North American multi-model ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, <https://doi.org/10.1175/BAMS-D-12-00050.1>.
- Lawrence, D. M., and Coauthors, 2011: Parameterization improvements and functional and structural advances in version 4 of the Community Land Model. *J. Adv. Model. Earth Syst.*, **3**, M03001, <https://doi.org/10.1029/2011MS00045>.
- Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539, <https://doi.org/10.1016/j.jcp.2007.02.014>.
- Lim, Y., and Coauthors, 2019: Influence of the QBO on MJO prediction skill in the subseasonal-to-seasonal prediction models. *Climate Dyn.*, **53**, 1681–1695, <https://doi.org/10.1007/s00382-019-04719-y>.
- Ma, H., and Coauthors, 2014: On the correspondence between mean forecast errors and climate errors in CMIP5 models. *J. Climate*, **27**, 1781–1798, <https://doi.org/10.1175/JCLI-D-13-00474.1>.
- Magnusson, L., J. Nycander, and E. Kallen, 2009: Flow-dependent versus flow-independent initial perturbations for ensemble

- prediction. *Tellus*, **61A**, 194–209, <https://doi.org/10.1111/j.1600-0870.2008.00385.x>.
- Mariotti, A., and Coauthors, 2020: Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bull. Amer. Meteor. Soc.*, **101**, E608–E625, <https://doi.org/10.1175/BAMS-D-18-0326.1>.
- Marshall, A. G., and A. A. Scaife, 2010: Improved predictability of stratospheric sudden warming events in an atmospheric general circulation model with enhanced stratospheric resolution. *J. Geophys. Res.*, **115**, D16114, <https://doi.org/10.1029/2009JD012643>.
- McIntyre, M. E., 1982: How well do we understand sudden stratospheric warmings? *J. Meteor. Soc. Japan*, **60**, 37–65.
- Murphy, A. H., and E. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.*, **117**, 572–582, [https://doi.org/10.1175/1520-0493\(1989\)117<C0572:SSACCI>2.0.CO;2](https://doi.org/10.1175/1520-0493(1989)117<C0572:SSACCI>2.0.CO;2).
- NAS, 2016: *Next Generation Earth System Prediction: Strategies for Subseasonal to Seasonal Forecasts*. The National Academies Press, 350 pp., <https://doi.org/10.17226/21873>.
- Neale, R. B., and Coauthors, 2012: Description of the NCAR Community Atmosphere Model (CAM 5.0). NCAR Tech. Note TN-486, 274 pp., www.cesm.ucar.edu/models/cesm1.0/cam/docs/description/cam5_desc.pdf.
- Nie, Y., A. A. Scaife, H.-L. Ren, R. E. Comer, M. B. Andrews, P. Davis, and N. Martin, 2019: Stratospheric initial conditions provide seasonal predictability of the North Atlantic and Arctic Oscillations. *Environ. Res. Lett.*, **14**, 034006, <https://doi.org/10.1088/1748-9326/ab0385>.
- Nishimoto, E., and S. Yoden, 2017: Influence of the stratospheric quasi-biennial oscillation on the Madden–Julian oscillation during austral summer. *J. Atmos. Sci.*, **74**, 1105–1125, <https://doi.org/10.1175/JAS-D-16-0205.1>.
- Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment. *Bull. Amer. Meteor. Soc.*, **100**, 2043–2060, <https://doi.org/10.1175/BAMS-D-18-0270.1>.
- Richter, J. H., F. Sassi, and R. R. Garcia, 2010: Toward a physically based gravity wave source parameterization in a general circulation model. *J. Atmos. Sci.*, **67**, 136–156, <https://doi.org/10.1175/2009JAS3112.1>.
- , A. Solomon, and J. T. Bacmeister, 2014: Effects of vertical resolution and nonorographic gravity wave drag on the simulated climate in the Community Atmosphere Model, version 5. *J. Adv. Model. Earth Syst.*, **6**, 357–383, <https://doi.org/10.1002/2013MS000303>.
- , C. Deser, and L. Sun, 2015: Effects of stratospheric variability on El Niño teleconnections. *Environ. Res. Lett.*, **10**, 124021, <https://doi.org/10.1088/1748-9326/10/12/124021>.
- Riddle, E. E., A. H. Butler, J. C. Furtado, J. L. Cohen, and K. Kumar, 2013: CFSv2 ensemble prediction of the wintertime Arctic Oscillation. *Climate Dyn.*, **41**, 1099–1116, <https://doi.org/10.1007/s00382-013-1850-5>.
- Robertson, A. W., A. Kumar, M. Peña, and F. Vitart, 2015: Improving and promoting subseasonal to seasonal prediction. *Bull. Amer. Meteor. Soc.*, **96**, ES49–ES53, <https://doi.org/10.1175/BAMS-D-14-00139.1>.
- Scaife, A. A., C. K. Folland, L. V. Alexander, A. Moberg, and J. D. Knight, 2008: European climate extremes and the North Atlantic Oscillation. *J. Climate*, **21**, 72–83, <https://doi.org/10.1175/2007JCLI1631.1>.
- , and Coauthors, 2014a: Predictability of the quasi-biennial oscillation and its Northern winter teleconnection on seasonal to decadal timescales. *Geophys. Res. Lett.*, **41**, 1752–1758, <https://doi.org/10.1002/2013GL059160>.
- , and Coauthors, 2014b: Skillful long-range prediction of European and North American winters. *Geophys. Res. Lett.*, **41**, 2514–2519, <https://doi.org/10.1002/2014GL059637>.
- Son, S. W., Y. Lim, C. Yoo, H. Hendon, and J. Kim, 2017: Stratospheric control of the Madden–Julian oscillation. *J. Climate*, **30**, 1909–1922, <https://doi.org/10.1175/JCLI-D-16-0620.1>.
- Stan, C., D. M. Straus, J. S. Frederiksen, H. Lin, E. D. Maloney, and C. Schumacher, 2017: Review of tropical-extratropical teleconnections on intraseasonal time scales. *Rev. Geophys.*, **55**, 902–937, <https://doi.org/10.1002/2016RG000538>.
- Stockdale, T. N., F. Molenti, and L. Ferranti, 2015: Atmospheric initial conditions and the predictability of the Arctic Oscillation. *Geophys. Res. Lett.*, **42**, 1173–1179, <https://doi.org/10.1002/2014GL062681>.
- Tripathi, O. P., and Coauthors, 2016: Examining the predictability of the stratospheric sudden warming of January 2013 using multiple NWP systems. *Mon. Wea. Rev.*, **144**, 1935–1960, <https://doi.org/10.1175/MWR-D-15-0010.1>.
- Tsujino, H., and Coauthors, 2018: JRA-55 based surface dataset for driving ocean–sea-ice models (JRA55-do). *Ocean Modell.*, **130**, 79–139, <https://doi.org/10.1016/j.ocemod.2018.07.002>.
- , and Coauthors, 2020: Evaluation of global ocean–sea-ice model simulations based on the experimental protocols of the Ocean Model Intercomparison Project phase 2 (OMIP-2). *Geosci. Model Dev.*, **13**, 3643–3708, <https://doi.org/10.5194/gmd-13-3643-2020>.
- Vitart, F., and A. W. Robertson, 2018: The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *npj Climate Atmos. Sci.*, **1**, 3, <https://doi.org/10.1038/S41612-018-0013-0>.
- Wheeler, M. C., and H. H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, **132**, 1917–1932, [https://doi.org/10.1175/1520-0493\(2004\)132<1917:AARMMI>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<1917:AARMMI>2.0.CO;2).
- White, C. J., and Coauthors, 2017: Potential applications of Subseasonal-to-Seasonal (S2S) predictions. *Meteor. Appl.*, **24**, 315–325, <https://doi.org/10.1002/met.1654>.
- Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.
- Xie, P., M. Chen, S. Yang, A. Yatagai, T. Hayasaka, Y. Fukushima, and C. Liu, 2007: A gauge-based analysis of daily precipitation over East Asia. *J. Hydrometeorol.*, **8**, 607–626, <https://doi.org/10.1175/JHM583.1>.
- Xie, S., H. Ma, J. S. Boyle, S. A. Klein, and Y. Zhang, 2012: On the correspondence between short- and long-time-scale systematic errors in CAM4/CAM5 for the Year of Tropical Convection. *J. Climate*, **25**, 7937–7955, <https://doi.org/10.1175/JCLI-D-12-00134.1>.
- Yeager, S. G., and Coauthors, 2018: Predicting near-term changes in the Earth system: A large ensemble of initialized decadal prediction simulations using the Community Earth System Model. *Bull. Amer. Meteor. Soc.*, **99**, 1867–1886, <https://doi.org/10.1175/BAMS-D-17-0098.1>.
- Yoo, C., and S.-W. Son, 2016: Modulation of the boreal wintertime Madden-Julian Oscillation by the stratospheric quasi-biennial oscillation. *Geophys. Res. Lett.*, **43**, 1392–1398, <https://doi.org/10.1002/2016GL067762>.