PREDICTING NEAR-TERM CHANGES IN THE EARTH SYSTEM

A Large Ensemble of Initialized Decadal Prediction Simulations Using the Community Earth System Model

S. G. Yeager, G. Danabasoglu, N. A. Rosenbloom, W. Strand, S. C. Bates, G. A. Meehl, A. R. Karspeck, K. Lindsay, M. C. Long, H. Teng, and N. S. Lovenduski

A new community data resource offers unique capabilities for evaluating the potential for useful Earth system prediction on decadal time scales.

The field of near-term climate prediction has grown rapidly since the advent of the first studies, about a decade old now, showing that observation-based initialization of coupled general circulation model (CGCM) simulations of the last half-century can significantly enhance predictive capacity on time scales from a year to a decade or more in advance (Keenlyside et al. 2008; Pohlmann et al. 2009; Smith et al. 2007). Phase 5 of the Coupled Model Intercomparison Project (CMIP5) included, for the first time, a framework for testing the benefits of historical initialization, in addition to prescribed external forcings, for decadal climate outlooks that are sensitive to both initial conditions and forced boundary conditions (Meehl et al. 2009). This framework involves running fully coupled hindcast (i.e., retrospective forecast) ensembles initialized using observationally based state information at prescribed intervals over the historical period; these hindcasts are then verified against observations and compared to uninitialized, free-running simulations to determine both overall skill and the benefits of initialization. Analysis of CMIP5 decadal prediction (DP) experiments revealed a wide range in skill for different variables and for different prediction systems. While the potential for useful applications was demonstrated, the CMIP5 experience also highlighted a number of outstanding research questions that must be tackled in order to advance DP science into a more mature phase (Kirtman et al. 2013; Meehl et al. 2014). The prediction of near-term climate change has recently been recognized by the World Climate Research Programme (WCRP) as one of the grand challenges facing the international climate research community,

AFFILIATIONS: Yeager, Danabasoglu, Rosenbloom, Strand, Bates, Meehl, Karspeck, Lindsay, Long, and Teng—National Center for Atmospheric Research, Boulder, Colorado; Lovenduski—Department of Atmospheric and Oceanic Sciences, and Institute of Arctic and Alpine Research, University of Colorado Boulder, Boulder, Colorado
CORRESPONDING AUTHOR: Stephen Yeager, yeager@ucar.edu

The abstract for this article can be found in this issue, following the table of contents.
DOI:10.1175/BAMS-D-17-0098.1

A supplement to this article is available online (10.1175/BAMS-D-17-0098.2)

In final form 8 March 2018

and an extensive set of coordinated decadal prediction experiments slated for CMIP6 will help to further advance the frontiers of this still-nascent application of CGCMs (Boer et al. 2016).

The enormous computational cost of performing (and analyzing) DP experiments is a significant impediment to progress in near-term climate prediction, not least because it greatly restricts the community of active researchers. The tier-1 set of hindcasts required for basic participation in the Decadal Climate Prediction Project (DCPP) of CMIP6 calls for roughly 3,000 years of coupled model simulation (Boer et al. 2016). This figure is based on a set of 10-member ensembles initialized each year for the past 60 years and integrated forward for 5 years ($60 \times 10 \times 5 = 3{,}000$). The resource demand doubles if the hindcast length is extended to 10 years so that skill can be evaluated at decadal, as opposed to multiannual, lead times. The significant cost of such experiments makes it difficult, if not wholly unfeasible, to systematically evaluate the sensitivity to poorly constrained DP configuration choices such as the ensemble size, the method of ensemble generation, the annual start date, the number of start times, the initialization method, the number of initialized Earth system components (in addition to the ocean), and the component model resolution(s). Furthermore, the identification of skill enhancement due to initialization requires a complementary set of "uninitialized" (UI) historical simulations, which greatly adds to the expense of DP evaluation. Both the DP and UI ensembles should ostensibly be large enough so that the ensemble average operation effectively isolates the shared component of variance within the respective ensembles (Boer et al. 2013). The standard 10-member ensemble is probably insufficient for this purpose for many fields and regions of interest (Sienz et al. 2015), but is generally deemed adequate for pragmatic reasons.

A recently completed set of initialized prediction simulations using the Community Earth System Model (CESM) promises to be a valuable resource for evaluating decadal predictions of the Earth system in the large-ensemble limit. The CESM decadal prediction large ensemble (CESM-DPLE) is composed of 40 member ensembles initialized each 1 November between 1954 and 2015 (for a total of 62 start dates) and integrated for 122 months. What was originally a 10-member ensemble set was expanded by an additional 30 members in early 2017 thanks to a computational award granted by the Computational and Information Systems Laboratory (CISL) of the National Center for Atmospheric Research (NCAR). A unique aspect of the CESM-DPLE that, apart from ensemble size, makes it unprecedented in the field of near-term climate prediction is that the complementary, uninitialized historical simulations also compose a large ensemble set. The CESM Large Ensemble (CESM-LE; Kay et al. 2015) is a highly successful community project that has accumulated a 40-member ensemble of historical and projection simulations spanning 1920–2100. The CESM-DPLE was generated using the same code base, component model configurations, and historical and projected radiative forcings as in the CESM-LE. Together, CESM-DPLE and CESM-LE offer a powerful means of disentangling the impacts of external forcing versus initialization on hindcast skill and of ascertaining how ensemble size (of both the initialized and uninitialized simulation sets) influences DP assessment.

The CESM-DPLE is a rich, public dataset that will support a broad spectrum of scientific research related to Earth system prediction. It comprises roughly 600 TB of climate data archived at temporal frequencies ranging from 6 hourly to annual from each of the CESM component models (ocean, atmosphere, sea ice, and land). It includes ocean biogeochemistry fields (as does the CESM-LE), and thus it permits exploration of the predictability of fundamental components of the ocean biosphere and carbon cycle. Improved sampling of underlying climate probability distribution functions (PDFs) through the use of large ensembles provides more accurate measures of higher-order statistical moments in addition to the ensemble mean. Indeed, one of the scientific motivations for the CESM-DPLE project was to determine whether the CESM prediction system shows any evidence of predictable shifts in the likelihood of extreme climate phenomena—such as heat waves, cold spells, and floods—that inhabit the tails of climate PDFs. The large ensemble size will also facilitate process-oriented conditional subsampling of the ensemble in order to develop a deeper understanding of the critical mechanisms at play in near-term prediction. As noted above, the fact that CESM-DPLE represents the initialized counterpart to CESM-LE over the time period of roughly 1955–2025 opens up a host of possible lines of inquiry relating not only to near-term prediction skill and optimal DP system design, but more broadly addressing questions about the mechanisms and statistics of forced versus internal climate variability that might be elucidated by contrasting the two large ensembles.

This article is intended to document the CESM-DPLE experimental design, provide a broad overview of the prediction skill for a few key climate fields, and advertise some promising capabilities that merit more

in-depth examination in subsequent work. The hope is that this study will inspire a diverse community to examine the CESM-DPLE dataset, and that it will serve as a jumping-off point for more detailed and focused scrutiny of regional skill and associated mechanisms.

EXPERIMENTAL DESCRIPTION. The CESM-DPLE is based on CESM, version 1.1, using the same model and component configuration as that used in the CESM-LE (Kay et al. 2015). For completeness, some of the model details are repeated here. The atmosphere component is the Community Atmosphere Model, version 5 (CAM5; Hurrell et al. 2013), with a finite-volume dynamical core at nominal 1° horizontal resolution and 30 vertical levels. The ocean component is version 2 of the Parallel Ocean Program (POP) run at nominal 1° horizontal resolution and 60 vertical levels (Danabasoglu et al. 2012). The sea ice model is version 4 of the Los Alamos National Laboratory (LANL) Community Ice Code (CICE4; Hunke and Lipscomb 2008) and is run on the same horizontal grid as the ocean. The land model is version 4 of the Community Land Model (CLM4; Lawrence et al. 2011). The historical (through 2005) and projected (from 2006 onward) radiative forcings (including greenhouse and short-lived gases and aerosols) are identical to those used in CESM-LE. Following DCPP guidelines, historical volcanic aerosol forcings are applied in the DP experiments (Boer et al. 2016).

The CESM1.1 model includes the capability for simulating the global carbon cycle, wherein the land and ocean component models both include biogeochemistry modules that compute carbon exchanges with the atmosphere (Hurrell et al. 2013; Lindsay et al. 2014). On the land surface, CLM simulates gross primary productivity and routing into litter and soil carbon pools using prescribed vegetation distributions. The ocean model explicitly simulates seawater carbonate chemistry, includes a representation of the lower trophic levels of the marine ecosystem, and tracks several biogeochemical tracers including dissolved inorganic carbon, oxygen, and nutrients (Long et al. 2013, 2016; Moore et al. 2013). These features of the CESM-DPLE system provide relatively novel Earth system prediction capabilities, but as in the CESM-LE simulations (Kay et al. 2015), the ocean biogeochemistry and the simulated atmospheric CO_2 concentration are purely diagnostic (i.e., there is no feedback onto the simulated physical climate).

The CESM-DPLE is a collection of 2,480 independent historical integrations of the CESM model—40 distinct member simulations for each of 62 initialization dates (1 November from 1954 to 2015). As such, the computation was highly parallelizable. Initially, 10 members for each initialization date were completed on National Energy Research Scientific Computing Center (NERSC) machines. An additional 30 members (~19,000 simulation years) were completed within a 2-month span on the new Cheyenne supercomputer at the NCAR-Wyoming Supercomputer Center (NWSC). Initial conditions for the atmosphere and land models were obtained from a single member of the CESM-LE from which 1 November restart files were saved. While these initial conditions contain the effects of historical radiative forcings, they were not otherwise constrained by observations. Observations were introduced into CESM-DPLE primarily through the ocean and sea ice initial conditions, which were obtained from a coupled ocean–sea ice configuration of CESM1.1 forced at the surface with historical atmospheric state and flux fields (see "Initializing the ocean" sidebar for details). Such forced ocean–sea ice (FOSI) simulations using CESM have been shown to reproduce some key aspects of observed ocean and sea ice variability quite well despite the fact that there is no direct assimilation of either ocean or sea ice observations (Danabasoglu et al. 2016; Yeager and Danabasoglu 2014; Yeager et al. 2015). Thus, the observations contributing to the historical realism of CESM-DPLE are derived from the atmospheric reanalysis and flux products used to drive the FOSI simulation. Full-field (as opposed to anomaly) initialization is used for all model components; drift adjustment is generally required prior to analysis. Unless otherwise noted, the results shown below are based on anomaly analysis, with DP anomaly fields computed by removing a lead-time-dependent model climatology whose historical time span exactly matches that used for the verification data climatology (Boer et al. 2013; Kim et al. 2012).

The CESM-DPLE builds upon previous DP efforts at NCAR that made use of the Community Climate System Model, version 4 (CCSM4). The CCSM4 decadal prediction (CCSM4-DP) simulation set was submitted to the CMIP5 DP collection and has been analyzed in several publications (Karspeck et al. 2015; Meehl et al. 2016; Meehl and Teng 2012, 2014a,b; Yeager et al. 2012, 2015). Noteworthy setup differences that distinguish CESM-DPLE from CCSM4-DP (apart from ensemble size) include 1) the model code base (in particular, the use of CAM5 instead of CAM4), 2) the inclusion of ocean biogeochemistry (BGC), 3) the ensemble start date (1 November instead of 1 January), and 4) a new FOSI simulation with improved forcing for initializing the ocean

The ocean is the primary reservoir of memory in the climate system, and hence the historical ocean state is the single-most important consideration when it comes to initializing decadal climate predictions. The NCAR contribution to the CMIP5 decadal prediction collection (CCSM4-DP) used a FOSI simulation to obtain historical initial conditions for the ocean and sea ice components, with the FOSI surface fluxes computed with bulk formulas using atmospheric fields from the Coordinated Ocean-Ice Reference Experiment (CORE) forcing dataset. The CORE forcing protocol has been a widely used and extensively documented standard in the ocean modeling community (refer to the CORE-II virtual special issue of *Ocean Modelling* for related articles). Skillful reproduction of observed decadal changes in Labrador Sea hydrography, in particular, in the CORE-forced FOSI is believed to be a key aspect of the CCSM4-DP initialization that explains long lead-time prediction skill in the subpolar gyre of the North Atlantic (Yeager et al. 2012, 2015).

Despite high skill in the Atlantic sector, CCSM4-DP hindcasts exhibited a strong initialization shock in the tropical Pacific that resulted in spurious El Niño (La Niña) conditions in early lead years in ensembles initialized between roughly 1955 and 1974 (1985 and 2005) (Karspeck et al. 2015; Teng et al. 2017). Inspection of the CORE-forced FOSI simulation used for initializing CCSM4-DP revealed a spurious weakening trend in the large-scale, zonal SST gradient along the equatorial Pacific compared to the stable SST gradient over the late twentieth century in the observed record (Fig. SB1a). Other experimental FOSI simulations forced with NOAA Twentieth Century Reanalysis, version 2 (20CRv2; Compo et al. 2011), and adjusted Japanese 55-year Reanalysis Project (JRA55-do; Tsujino et al. 2018.) fields showed much less of a ΔSST trend than the CORE-forced FOSI and, hence, were more consistent with observations in the tropical Pacific. The negative ΔSST trend in the CORE-forced FOSI appears to be related to a pronounced slackening of

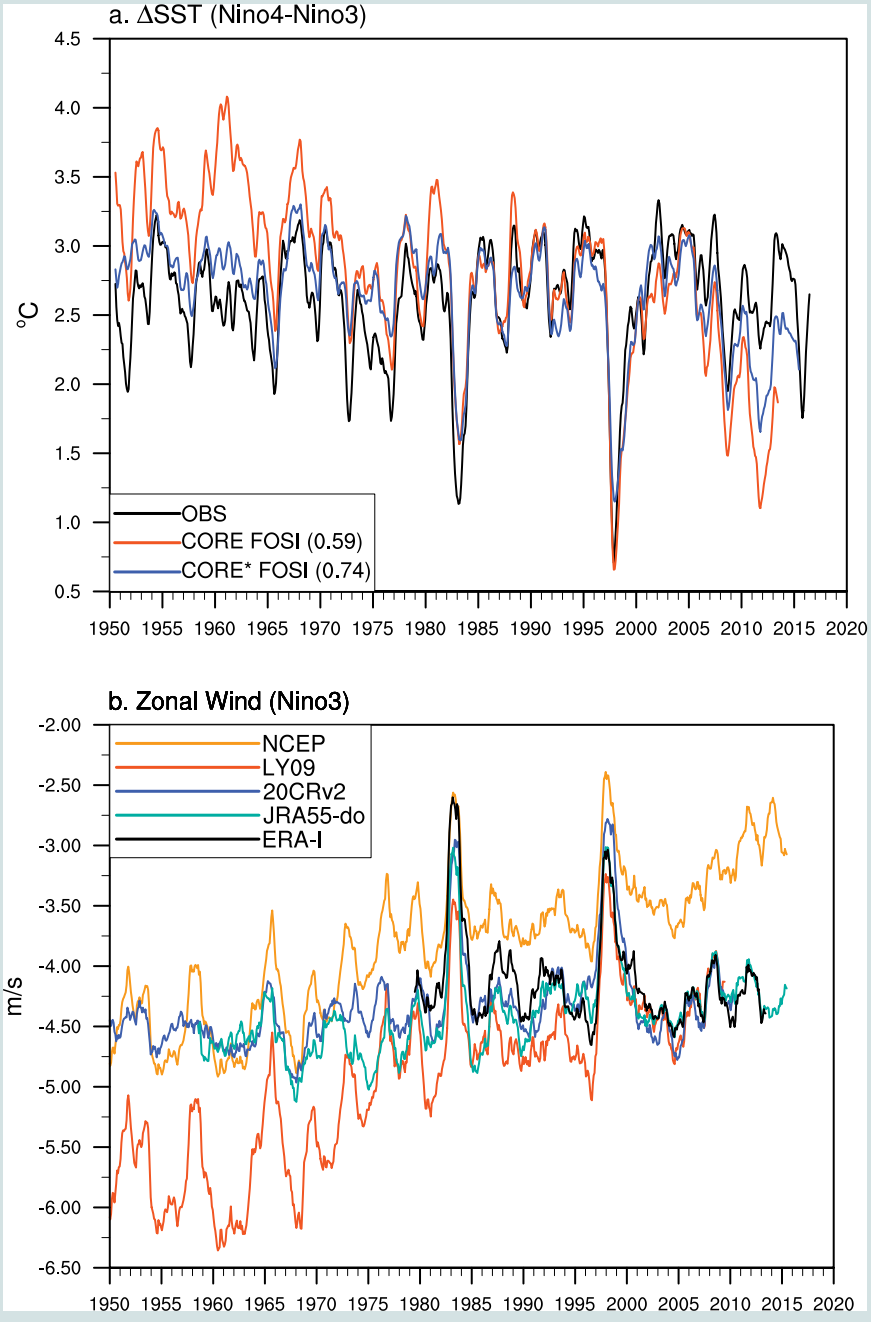

Fig. SB1. Monthly time series of (a) the large-scale zonal SST gradient in the equatorial Pacific, quantified as the difference between western (Niño-4) and eastern (Niño-3) regional averages and (b) 10-m zonal wind speed averaged over the eastern equatorial Pacific (Niño-3). The ΔSST curves are from observations (OBS; Hurrell et al. 2008) and the FOSI simulations used to initialize CCSM4-DP (CORE) and CESM-DPLE (CORE*), with the correlation with observations given in the legend in (a). The zonal wind curves are from a variety of raw or adjusted atmospheric reanalysis products: NCEP (Kalnay et al. 1996), CORE-adjusted NCEP (LY09), 20CRv2 (Compo et al. 2011), adjusted JRA55-do (Tsujino et al. 2018), and ERA-I (Dee et al. 2011).

the equatorial trade winds in the Pacific in the NCEP–NCAR reanalysis that is used as the base dataset for CORE (Fig. SB1b). The time-independent adjustment to NCEP–NCAR reanalysis winds that are part of the CORE forcing protocol (Large and Yeager 2009, hereafter LY09) correct the weak trade wind bias in that dataset in the modern satellite era, but the adjustment also results in trade winds that are probably too strong prior to 1975, and it amplifies the weakening trend inherent in the NCEP–NCAR reanalysis (Fig. SB1b). The fact that LY09 winds in the equatorial Pacific appear to be an outlier compared to more recent atmospheric reanalysis products such as 20CRv2 and JRA55-do, as well as the European Centre for Medium-Range Weather Forecasts (ECMWF) interim reanalysis (ERA-Interim, hereafter ERA-I; Dee et al. 2011), would appear to explain the poor SST simulation in that region in the CORE-forced FOSI.

To eliminate the spurious ΔSST trend, which presumably gives rise to spurious Bjerknes feedback effects upon coupling that contribute to initialization shock behavior, a new FOSI was developed with forcing coming primarily from CORE data streams but with a nonstandard, blended wind field. In particular, CORE winds were used everywhere except in the tropical band (30°S–30°N), where either 20CRv2 winds (spanning 1948–2010) or JRA55-do winds (to extend the simulation through 2015) were used. This new CORE*-forced FOSI successfully eliminated the spurious ΔSST trend in the Pacific (Fig. SB1a) while retaining desirable aspects of standard CORE forcing elsewhere. The CESM-DPLE used the CORE*-forced FOSI for initializing the ocean and sea ice components, and this change is believed to explain the dramatic reduction in tropical Pacific initialization shock as well as much of the large, global-scale skill improvements in a variety of fields compared to CCSM4-DP (see the supplemental material). Similar DP sensitivity to ocean initial states generated using erroneous tropical Pacific wind forcing has been noted in the Max Planck Institute decadal prediction system (Pohlmann et al. 2016).

and sea ice components. Significant skill improvements in CESM-DPLE compared to CCSM4-DP, discussed in the supplemental material (https://doi.org/10.1175/10.1175/BAMS-D-17-0098.2), are believed to derive primarily from setup difference 4 above, which largely eliminated a spurious trend in the east–west sea surface temperature (SST) gradient in the tropical Pacific (see the sidebar). A summary of the experimental setup of CESM-DPLE (and noteworthy changes from the setup used for CCSM4-DP) is provided in Table 1.

HINDCAST EVALUATION METHODS. The full-field initialization necessitates a drift adjustment procedure prior to hindcast verification against observations. This is accomplished by transforming the raw DP output into anomalies relative to the climatological forecast for each lead time: $Y'_{j\tau} = Y_{j\tau} - \overline{Y}_{\tau}$, where $Y_{j\tau}$ represents the ensemble-average forecast from start year j at lead time τ and $\overline{Y}_{\tau}$ represents an additional average over start years for a given τ. When verifying against sparse observational datasets, care is taken to compute $\overline{Y}_{\tau}$ using only $j\tau$ pairs for which there exist corresponding observations (Doblas-Reyes et al. 2013; Kim et al. 2012).

Hindcast verification in this paper follows the framework outlined in Goddard et al. (2013). Verification metrics include the anomaly correlation coefficient (ACC) and the mean-square skill score ($\text{MSSS} = 1 - \text{MSE}_{\text{DP}}/\text{MSE}_{\text{ref}}$) computed against standard observational benchmarks (see below). The MSSS quantifies the change in mean-square error between the DP ensemble and observations (MSE_{DP}) and the MSE of a reference hindcast (MSE_{ref}). Reference predictions considered herein include persistence and UI simulations. For comparison with DP hindcasts of N-year-average anomalies, the persistence hindcast is computed as the most recent N-year-average anomaly that had been observed at the time of DP initialization. Annual means are defined over January–December, and the persistence hindcast includes the November and December observations from the year of initialization. The UI hindcast is simply the N-year-average anomaly for a particular time period computed from the set of uninitialized historical simulations. As noted above, anomalies are defined relative to identically sampled climatologies that are computed separately for each distinct data stream. We focus here on decadal-time-scale predictions of low-frequency variability, and therefore we consider multiyear annual- and seasonal-mean predictions for a subset of lead times [e.g., lead years 1–5 (LY1–5)].

The nonparametric block bootstrap technique outlined in Goddard et al. (2013) is used to assess the statistical significance of hindcast skill scores. To test whether a score (e.g., ACC) or score difference (e.g., ΔACC) is significantly different from zero, a bootstrapped distribution of (4,000) scores is computed at each spatial location by resampling (with replacement) the hindcast ensembles across both the time and member dimensions. To account for temporal autocorrelation, the resampling in time maintains continuity in 5-yr blocks (although the results are not strongly sensitive to this choice of block length). The

bootstrapped PDF of the scores reflects the uncertainty of the test statistic associated with the limited ensemble size and temporal sampling and can be used to derive p values. For example, a positive (negative) score with a bootstrapped distribution showing only 100 scores below (above) zero would have a p value of 100/4,000 = 0.025. Scores of either sign with low p values are of interest insofar as they indicate areas of notable success (or failure) of the prediction system. Scores that are not locally significant (i.e., p value > 0.1) are denoted in map plots with a slash mark (/). Fisher's z transformation is applied to ACC scores prior to the determination of p values. Evaluation of hindcast skill in terms of gridded skill score maps demands consideration of the field significance of the results (Ventura et al. 2004; Wilks 2006, 2016). In our skill-map plots, we assess field significance by controlling the false discovery rate (FDR) following Wilks (2016), assuming moderate to strong spatial correlation ($\alpha_{global} = 0.1$; $\alpha_{FDR} = 2\alpha_{global}$). An advantage of the FDR method is that global field significance is implied by the existence of local p values that, when sorted, fall below a ramping threshold. These particularly low p values are denoted in map plots with a dot (·).

The observational benchmarks used for hindcast evaluation in this study are as follows: the Met Office EN4.2.1 gridded ocean temperature product for the upper-ocean heat content (Good et al. 2013), the Extended Reconstructed Sea Surface Temperature, version 5 (ERSSTv5), dataset from the National Oceanic and Atmospheric Administration (NOAA) (Huang et al. 2017), the University of East Anglia Climatic Research Unit time series, version 4.00 (CRU-TS4.0), land surface temperature and precipitation dataset (Harris et al. 2014), and estimates of ocean net primary productivity (NPP) generated from the Moderate Resolution Imaging Spectroradiometer (MODIS; 2003–15) using the Vertically Generalized Production Model (Behrenfeld and Falkowski 1997).

Table 1. Overview of the experimental setups of the two initialized decadal prediction experiments run to date at NCAR. Boldface font in the right column highlights noteworthy changes from the earlier (CCSM4-DP) setup. Note that the UI is a complementary simulation set that employs the same external radiative forcings as the DP set. Refer to the sidebar for a detailed description of the modified CORE forcing used to generate ocean and sea ice initial conditions for CESM-DPLE.

	CCSM4-DP	**CESM-DPLE**
Model	CCSM4	**CESM1.1**
atm	CAM4 (FV 1°, 26 levels)	**CAM5** (FV 1°, **30 levels**)
ocn	POP2 (1°, 60 levels)	POP2 (1°, 60 levels) **with BGC**
ice	CICE4 (1°)	CICE4 (1°)
lnd	CLM4	CLM4
UI ensemble	6-member CCSM4 twentieth-century ensemble (Meehl et al. 2012)	**40-member CESM twentieth-century Large Ensemble** (Kay et al. 2015)
Forcing		
through 2005	CMIP5 historical	CMIP5 historical
from 2006 onward	CMIP5 representative concentration pathway (RCP) 4.5	CMIP5 **RCP 8.5**
Initialization		
method	Full field	Full field
atm	UI	UI
ocn	CORE-forced FOSI	**CORE*-forced FOSI**
ice	CORE-forced FOSI	**CORE*-forced FOSI**
lnd	UI	UI
Ensembles		
Ensemble size	10	**40**
Start dates	Annual; 1 Jan 1955–2014 (N = 60)	Annual; **1 Nov 1954–2015 (N = 62)**
Ensemble generation	Variable Jan start days and round-off perturbation of atm initial conditions	Round-off perturbation of atm initial conditions
Simulation length	120 months	122 months

HINDCAST SKILL IN THE SURFACE OCEAN. We first assess skill at predicting upper-ocean heat content, particularly Atlantic Ocean heat content, because it is widely believed to be the foundation of skillful decadal predictions of surface climate (Yeager and Robson 2017). Pentadal variations in heat content of the upper 295 m (T295) are well predicted in broad tropical and extratropical regions within each ocean basin at short lead times (Fig. 1a). At longer lead times, there is a loss of skill in the tropical Pacific, tropical Indian, and eastern extratropical Pacific Oceans, while scores remain high elsewhere (Figs. 1b,c). The ACC in CESM-DPLE is considerably greater than persistence in most regions exhibiting positive skill (Fig. 1d), and this skill improvement tends to increase with lead time (Figs. 1e,f). The regions of significantly high skill and skill improvement are also generally found to be field significant.

Much of the skill improvement over persistence in pentadal hindcasts can be attributed to external forcing, and hence the comparison with uninitialized (but externally forced) historical simulations is a higher bar for gauging success. ACC score differences between CESM-DPLE and CESM-LE reveal the impact of initialization on predictions of T295 (Figs. 1g–i). The subpolar North Atlantic (SPNA) stands out as a region where the skill improvement associated with initialization is largest and most lasting, with ΔACC exceeding 0.4 at all lead times considered (Figs. 1g–i). There are indications that the high SPNA skill extends northward toward the Arctic, with large improvements over persistence and modest but significant improvements over uninitialized simulations at short lead times in the Nordic seas (NS). The high skill for SPNA heat content in initialized hindcasts is generally understood to derive from realistic initialization of, and limited prediction of, Atlantic thermohaline circulation anomalies (Yeager and Robson 2017), and there is mounting evidence that upper-ocean anomalies in the SPNA propagate across the Iceland–Scotland Ridge to provide a source for NS predictability (Årthun et al. 2017; Årthun and Eldevik 2016; Yeager et al. 2015).

Other regions of significantly enhanced skill for heat content include the eastern subtropical North Atlantic off the coast of Africa, the subtropical South Atlantic, the north- and southeastern subtropical Pacific, the west and south Indian Ocean to the east of

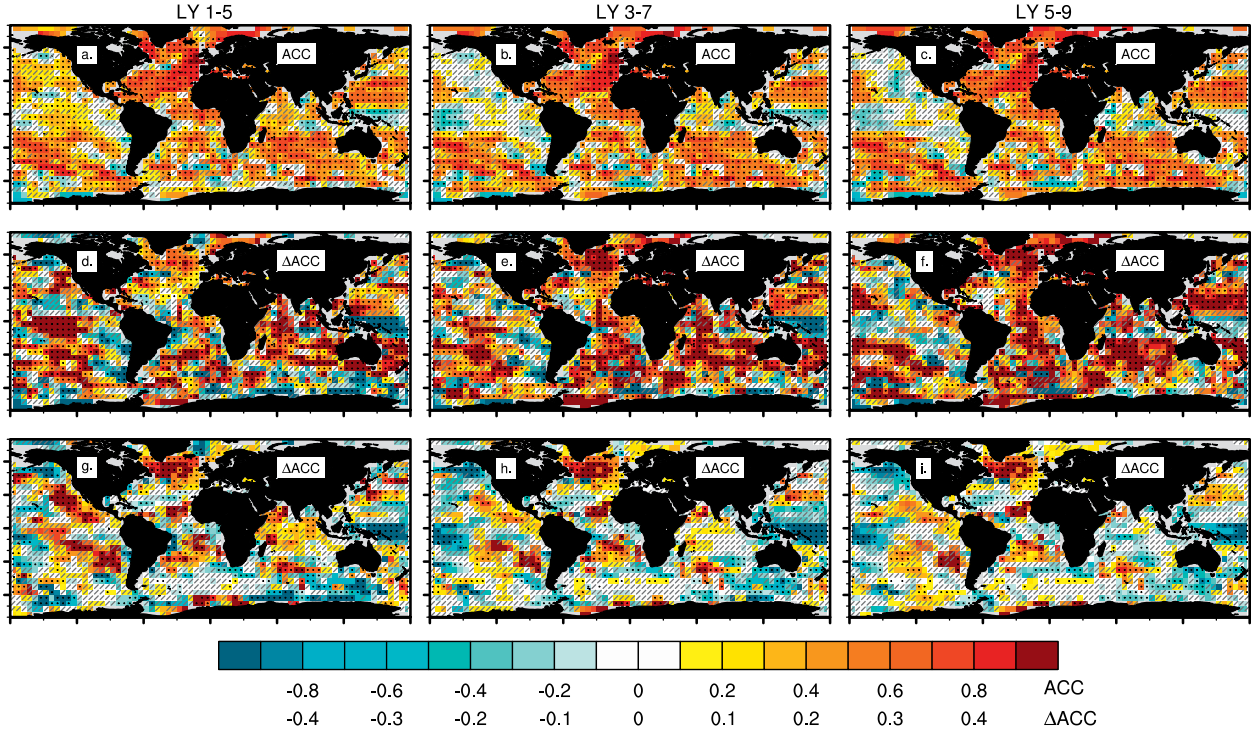

Fig. 1. (a)–(c) ACC of annual upper-ocean heat content above 295 m (T295) from CESM-DPLE relative to Met Office EN4.2.1 data (Good et al. 2013) for lead times of 1–5, 3–7, and 5–9 years, respectively. ACC skill score differences (d)–(f) between CESM-DPLE and persistence and (g)–(i) between CESM-DPLE and CESM-LE. All fields were mapped onto a 5° × 5° grid prior to analysis. The scale used for (d)–(i) is half that used for (a)–(c). The absence (presence) of a gray slash indicates scores that are (are not) significant at the 10% level ($\alpha = 0.1$); stippling further indicates points whose p values pass an FDR test for global (70°S–70°N) field significance ($\alpha_{global} = 0.1$).

Madagascar, and the northwestern Pacific (Figs. 1g–i). The mechanisms underpinning high T295 skill in CESM-DPLE in these regions, and the potential significance of this skill for other surface and subsurface fields of interest, should be examined in future, more regionally focused, studies. However, we will note that these regions exhibit an interesting correspondence to known mode water regions (Hanawa and Talley 2001), suggesting that the skill improvements may be associated with realistic initialization of, and subsequent subduction and interior advection of, mode water anomalies. There are also regions of significantly degraded T295 skill in the western tropical and north-central Pacific and the Southern Ocean (SO). The reasons for negative skill differences remain unclear at this time, but we speculate that they are related to initialization shocks that result in spurious tropical air–sea interaction at short lead times, particularly in the eastern Pacific, that can have large far-field impacts (see the sidebar and more extended discussion below).

Trend bias correction is a potential source of skill in initialized hindcasts that is derived simply from initializing closer to the observed long-term trend, and it can result in enhanced skill scores even without any improvement in the simulation of internal climate variability mechanisms. To check whether trend bias correction is a significant factor, we have redone Fig. 1 using detrended data (Fig. ES1). Almost all of the regions of improved T295 skill discussed above are recovered, and even enhanced, after detrending, implying that trend bias correction is not the dominant source of ACC skill for this field.

In some regions, long-lasting skill at predicting upper-ocean heat content finds surface expression in terms of high skill scores for annual-mean SSTs that can be clearly associated with initialization (Fig. 2). As for T295, CESM-DPLE exhibits widespread, significant skill at predicting pentadal SST variations out to decadal lead times (Figs. 2a–c). The northeast Pacific and the Pacific–Atlantic sectors of the SO stand out as regions of low predictability with skill that degrades with lead time (Figs. 2a–c). In contrast, the central and southeastern tropical Pacific shows a marked increase in skill with lead time. We will return to this phenomenon below. While there is widespread improvement over persistence (Figs. 2d–f), the skill comparison with CESM-LE reveals that external forcing accounts for a large fraction of the global

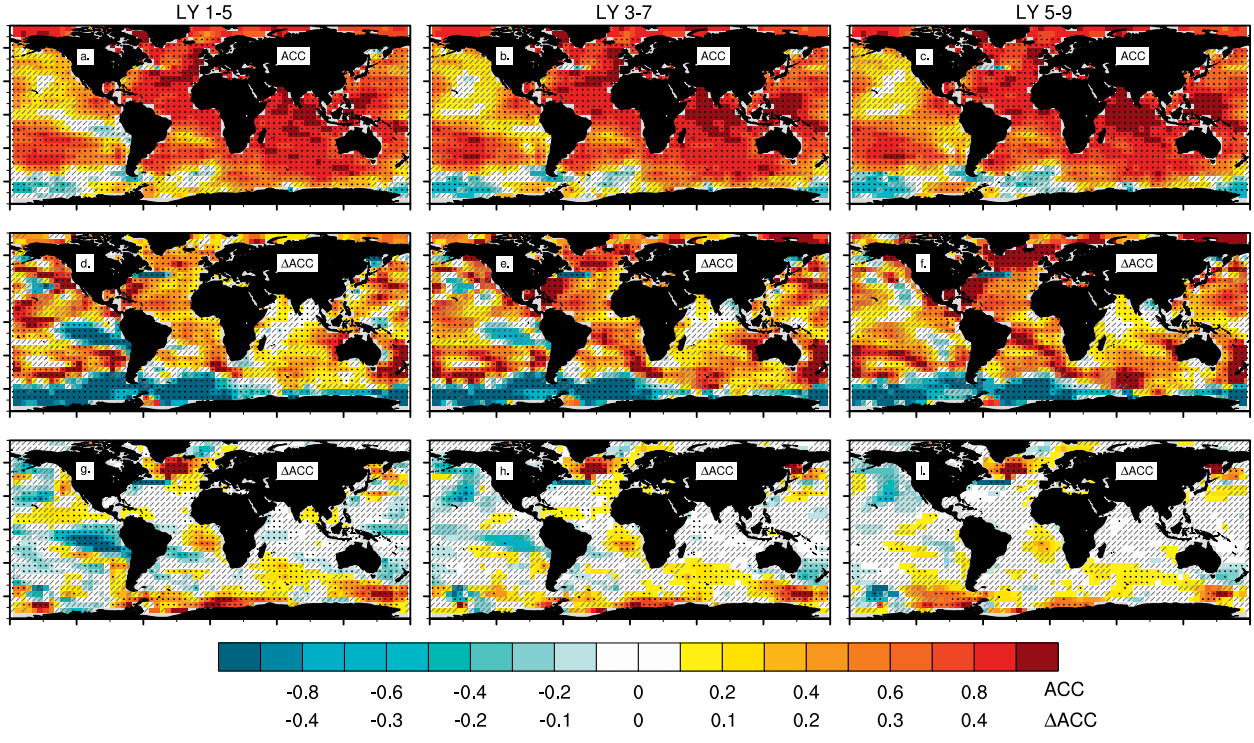

Fig. 2. (a)–(c) ACC of annual SST from CESM-DPLE relative to ERSSTv5 observations (Huang et al. 2017) for lead times of 1–5, 3–7, and 5–9 years, respectively. ACC skill score differences (d)–(f) between CESM-DPLE and persistence and (g)–(i) between CESM-DPLE and CESM-LE. All fields were mapped onto a 5° × 5° grid prior to analysis. The scale used for (d)–(i) is half that used for (a)–(c). The absence (presence) of a gray slash indicates scores that are (are not) significant at the 10% level (α = 0.1); stippling further indicates points whose *p* values pass an FDR test for global (70°S–70°N) field significance (α_{global} = 0.1).

SST skill in CESM-DPLE (Figs. 2g–i). In particular, the very high ACC scores in the Indian and tropical western Pacific Oceans in CESM-DPLE are no better than in the uninitialized ensemble, consistent with the known dominance of the externally forced trend on SST variance in this region (Han et al. 2010).

There are, however, several regions showing locally and field significant SST skill improvement over uninitialized simulations that appear to be geographically related to the T295 improvements noted above. The North Atlantic (and in particular the SPNA) again stands out as a region where initialization confers large benefits (Figs. 2g–i). For LY3–7, the skill improvement over uninitialized simulations (Fig. 2h) bears a strong resemblance to the canonical pattern of Atlantic multidecadal variability (AMV; Sutton and Hodson 2005), with heavy loading in the SPNA and an extension into the tropical North Atlantic through the eastern subtropics. This horseshoe pattern of SST skill improvement, and the extension of skill improvement into the Nordic seas, reflects the underlying T295 improvements (Fig. 1h). There is reason to suspect that some of the baseline SST skill in the SPNA in the uninitialized ensemble is obtained through incorrect physical mechanisms. Yeager et al. (2012) point out that UI ensembles simulate late-twentieth-century SPNA warming despite Atlantic meridional overturning circulation (AMOC) weakening, presumably through anomalous surface fluxes, whereas FOSI and initialized DP simulations rely on AMOC-related advective heat convergence. Thus, the improved mechanistic fidelity associated with initialization (positive SPNA SST trend due to a strengthening AMOC) is not necessarily reflected in Fig. 2.

Other regions showing noteworthy SST skill improvement (Figs. 2g–i) include the eastern subtropical South Atlantic; the south Indian Ocean off Madagascar; the SO, particularly the Bellingshausen Sea to the west of the Antarctic Peninsula; the northwest Pacific region to the east of Japan and the Kamchatka Peninsula; and the southeast Pacific off the coast of Chile. As for the SPNA, the enhanced SST skill in most of these regions appears to be related to collocated improvements in T295 skill and is resilient to detrending (Fig. ES2), suggesting that these SST improvements are not simply artifacts of trend bias correction. The SO SST skill improvement over CESM-LE derives in part from improved representation of nontrend variability (Figs. ES2g–i), but the relation with T295 skill improvement is not as clear as in other regions (cf. Figs. 1g–i). While there are significant improvements over uninitialized simulations throughout much of the SO, it is still a region of low overall skill that, apart from the Indian Ocean sector, does not show improvement over persistence (even when detrended; Fig. ES2). Further work is needed to clarify the nature of the SO response in CESM-DPLE.

The improvement in SST skill in the equatorial Pacific with lead time (Fig. 2) merits further discussion. As discussed in the sidebar (and the supplemental material), unbalanced initial conditions in the equatorial Pacific are hypothesized to give rise to spurious El Niño (La Niña) conditions that degrade the skill of initialized hindcasts at short lead times (Pohlmann et al. 2016; Teng et al. 2017). While the initialization shock in CESM-DPLE is much reduced compared to its predecessor (CCSM4-DP), there is still a significant improvement in SST skill in many tropical and extratropical regions as lead time increases (Fig. 3), indicative of short-term adjustments to initialization that tend to degrade skill, particularly in the El Niño region. The expanding blue (negative ΔACC) regions in Fig. 3 are expected, and we speculate that the growth of the orange regions (positive ΔACC) is dynamically linked to the improved SST in the (south)eastern tropical Pacific. ACC in the western tropical Pacific, for instance, is seen to improve significantly with lead time (Fig. 3), resulting in improved comparisons with reference forecasts (Figs. 2 and ES2). The early-lead-time skill degradation in this region (Fig. 2g) may be related to the combination of spurious variability in the eastern tropical Pacific and model bias that extends ENSO activity too far to the west (Van Oldenborgh et al. 2012).

PREDICTING SURFACE TEMPERATURE AND PRECIPITATION. A key outstanding challenge in DP research is to ascertain the extent to which multiyear skill in predicting ocean heat content, SST, and sea ice extent (Yeager et al. 2015) might translate into useful predictions of surface climate over land. Skill maps of ACC for annual or seasonal surface air temperature (SAT) over land show near-ubiquitous, high, and significant skill that generally outperforms persistence (Fig. ES4). Much of this skill derives from the strong externally forced trend in SAT, and so it is difficult to detect ACC improvement over uninitialized simulations unless a linear trend is first removed (Fig. ES5). However, the MSSS (using CESM-LE as the reference forecast) does reveal skill improvement even in the presence of strong forced SAT trends (Fig. 4). Initialization results in (field) significant improvements in pentadal predictions of annual SAT over western Europe, Greenland, the Mediterranean, northern and southern Africa, Arabia, South Asia, northeastern

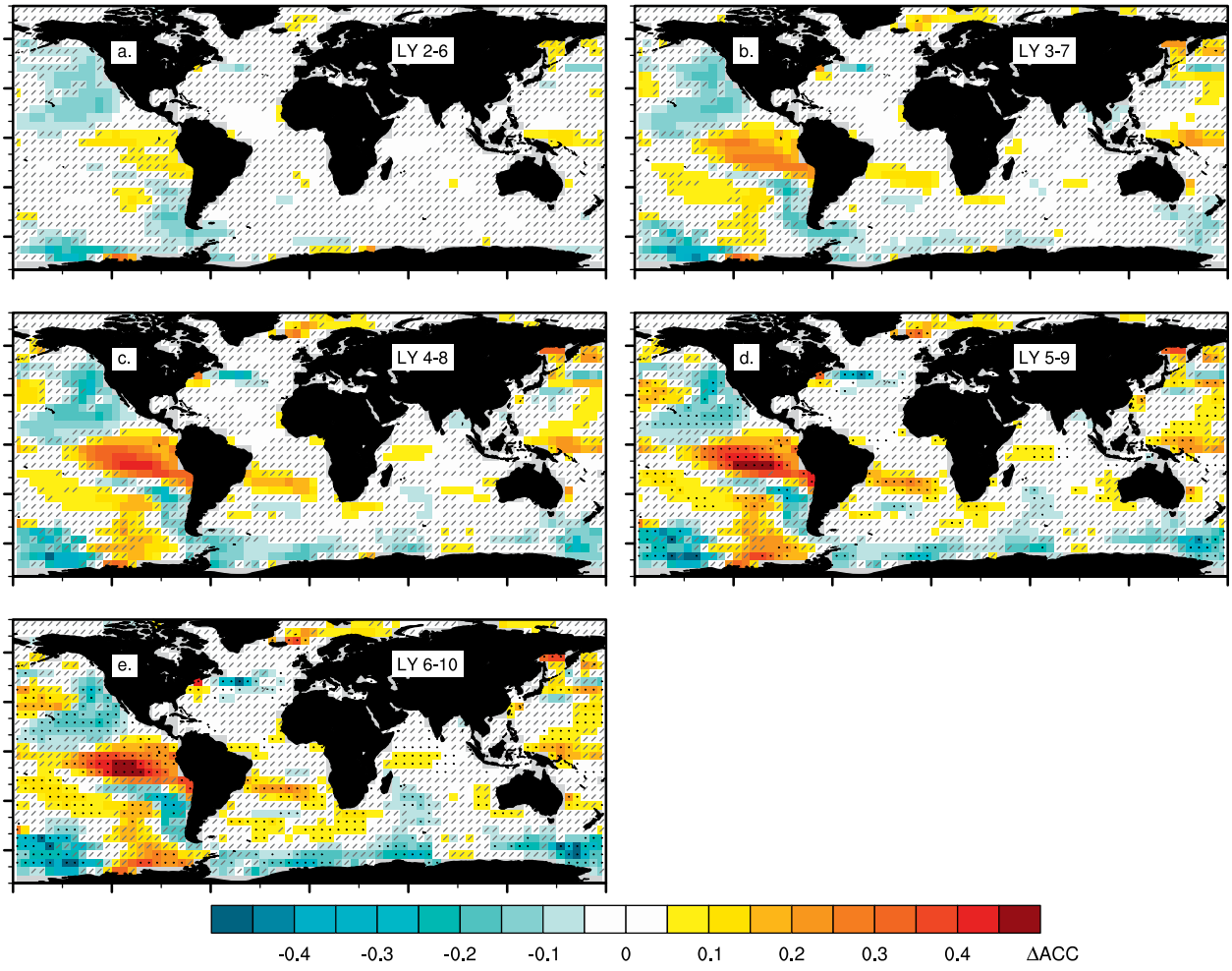

Fig. 3. ΔACC for annual SST from CESM-DPLE relative to ERSSTv5 observations (Huang et al. 2017) for lead years (a) 2–6, (b) 3–7, (c) 4–8, (d) 5–9, and (e) 6–10. In each panel, ΔACC is computed relative to the score for LY1–5 (see Fig. 2a). All fields were mapped onto a 5° × 5° grid prior to analysis. The absence (presence) of a gray slash indicates scores that are (are not) significant at the 10% level (α = 0.1); stippling further indicates points whose *p* values pass an FDR test for global (70°S–70°N) field significance (α_{global} = 0.1).

Eurasia, China, and the southwestern United States. The enhanced annual skill appears to be related to widespread error reduction during boreal summer [June–August (JJA); Figs. 4g–i], as boreal winter [December–February (DJF)] improvements are confined primarily to northeast Africa and Arabia (Figs. 4d–f).

The notable increase in SAT MSSS with lead time in many regions is likely related to the broad patterns of improved SST forcing of the atmosphere noted earlier (Fig. 3). Trend bias correction may be contributing to some of the SAT skill in Europe and Asia given that scores there are lower when the trend is removed (Fig. ES6). Even after detrending, however, significantly positive MSSS (Fig. ES6) and ΔACC (Fig. ES5) scores are found over large swaths of the greater Mediterranean and central and eastern Asia regions, implying significant improvements in the representation of nontrend variability associated with initialization. These results appear consistent with a recent study that concluded that skillful multiyear prediction of boreal summer temperature over northeast Asia derives from a global atmospheric teleconnection pattern modulated by low-frequency North Atlantic SST variability (Monerie et al. 2017).

CESM-DPLE shows promising prospects for useful decadal predictions of hydroclimate over land, and this represents a significant advance over the CMIP5-era CCSM4-DP system. Figure 5 shows the ACC skill map for boreal summer [June–September (JAS)] land precipitation. Locally and field significant positive ACC scores are found in western Europe, central and northeastern Eurasia, the African Sahel, parts of south and eastern Africa, Alaska, the northeastern and northwestern continental

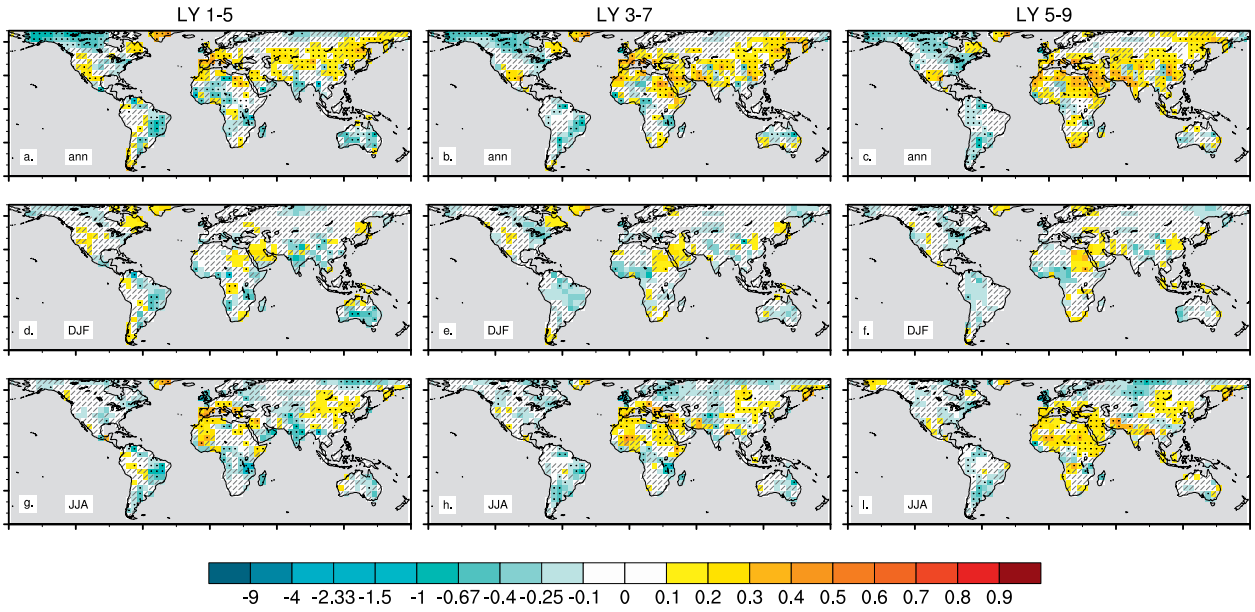

Fig. 4. MSSS of SAT from the CESM-DPLE using CESM-LE as the reference forecast for lead times of 1–5, 3–7, and 5–9 years. MSE is computed relative to CRU-TSv4.00 data (Harris et al. 2014). Rows show the scores for (a)–(c) annual, (d)–(f) boreal winter (DJF), and (g)–(i) boreal summer (JJA) means. The absence (presence) of a gray slash indicates scores that are (are not) significant at the 10% level (α = 0.1); stippling further indicates points whose *p* values pass an FDR test for global (70°S–70°N) field significance (α_{global} = 0.1). The nonlinear color bar reflects symmetric changes in the MSE ratio.

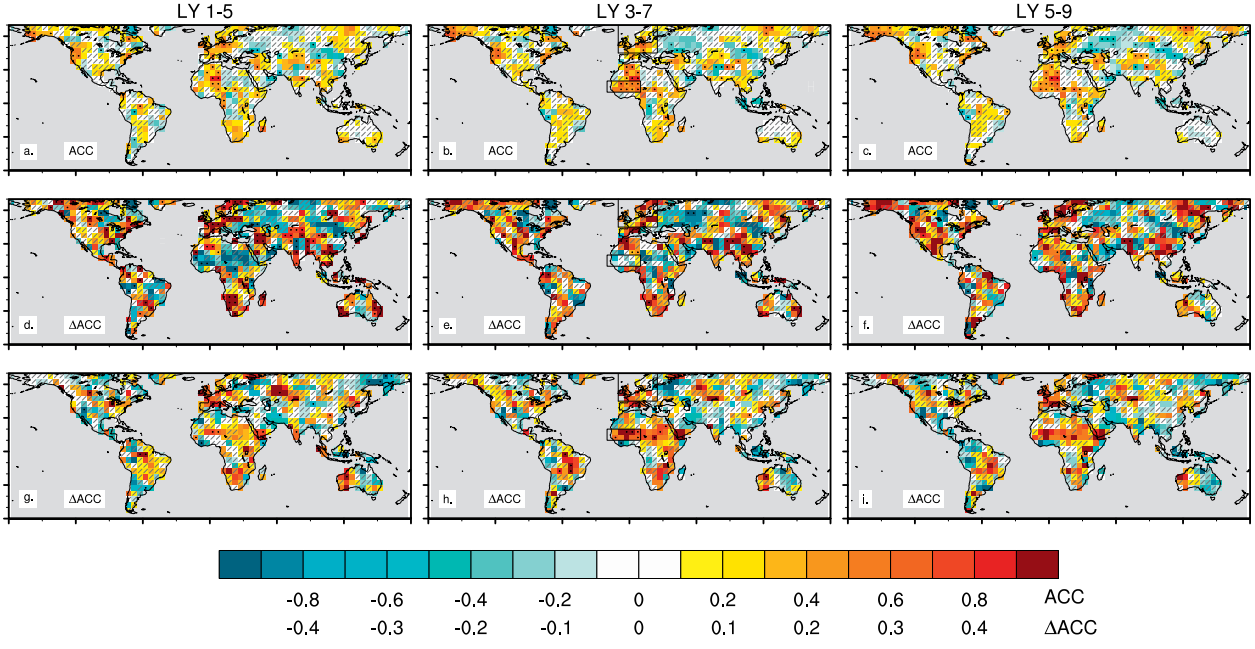

Fig. 5. (a)–(c) ACC of boreal summer (JAS) land surface precipitation (PREC) from the CESM-DPLE relative to CRU-TSv4.00 observations (Harris et al. 2014) for lead times of 1–5, 3–7, and 5–9 years. ACC skill score differences (d)–(f) between CESM-DPLE and persistence and (g)–(i) between CESM-DPLE and CESM-LE. All fields were mapped onto a 5° × 5° grid prior to analysis. The scale used for (d)–(i) is half that used for (a)–(c). The absence (presence) of a gray slash indicates scores that are (are not) significant at the 10% level (α = 0.1); stippling further indicates points whose *p* values pass an FDR test for global (70°S–70°N) field significance (α_{global} = 0.1).

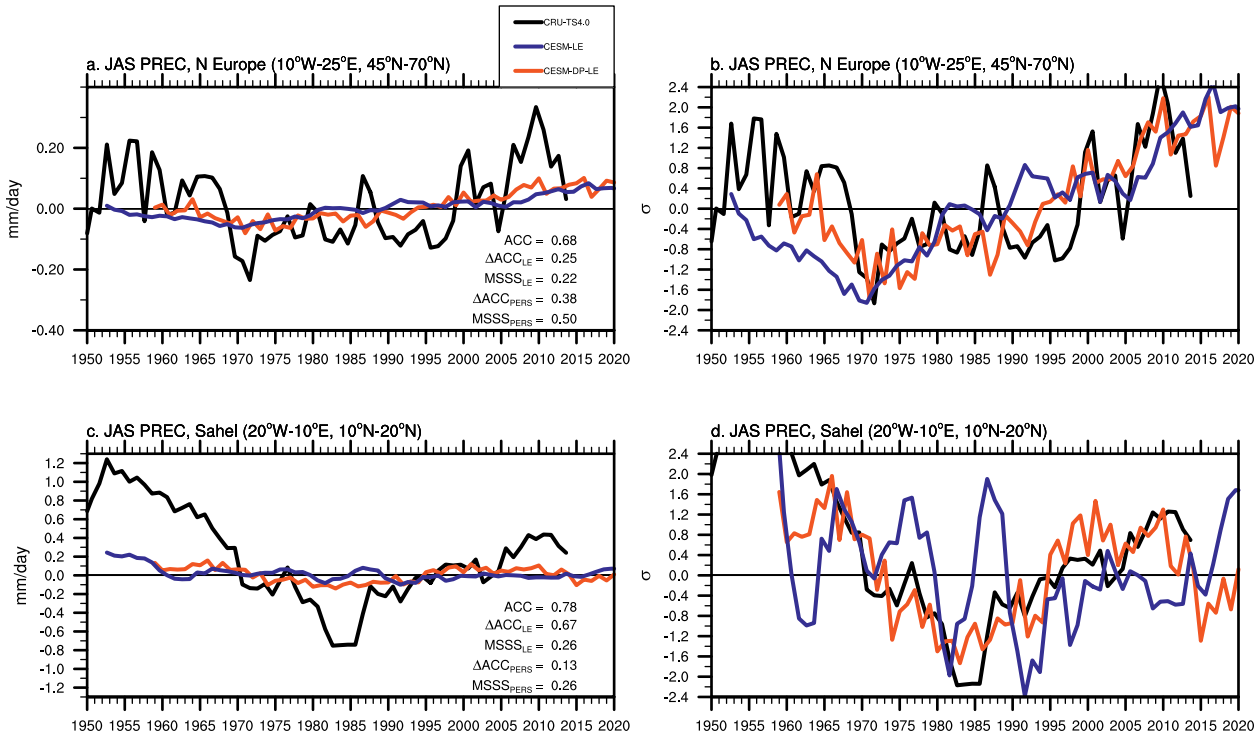

Fig. 6. Regional average boreal summer (JAS) precipitation for (a),(b) northern Europe (45°–70°N, 10°W–25°E) and (c),(d) the West African Sahel (10°–20°N, 20°W–10°E). (left) Raw time series (mm day^{-1}) and (right) normalized time series. The CESM-DPLE time series (red) is the ensemble mean over LY3–7; the CESM-LE ensemble-mean (blue) and observed (black) time series have been smoothed with a running 5-yr-mean filter. The regions are shown in Fig. 5. Skill scores are listed in (a) and (c).

United States, and northern Quebec (Figs. 5a–c). The number of field significant positive scores clearly outnumbers the negative scores. The skill variations with lead time are noticeable but small compared to the spatial variations in skill. Regions of high skill that also exhibit significant skill improvement over CESM-LE include western Europe, the Sahel, southeast Africa, and northwestern North America. Field significant improvements over uninitialized simulations are scattered at this spatial resolution, but there are noteworthy concentrations of positive ΔACC in the Sahel, Brazil, northwestern North America, western Europe, central Eurasia, southern Africa, and western Australia (Figs. 5g–i). However, not all of these regions show coherent improvements over persistence (Figs. 5d–f). Again, we find that the noted skill improvements over CESM-LE are resilient to detrending (Fig. ES7).

Numerous studies have linked the multidecadal variability of NATL SST (often referred to as Atlantic multidecadal variability) to seasonal climate fluctuations over Europe, Africa, Asia, and the Americas. It therefore seems likely that the CESM-DPLE surface climate skill in these regions stems in large part from (enhanced) skill at predicting NATL SST (Figs. 2 and ES2). However, the precise origins of skill over land in CESM-DPLE remain under investigation. Low-frequency warming (cooling) in the NATL has been associated with enhanced (suppressed) summer precipitation across Europe (Sutton and Dong 2012; Sutton and Hodson 2005), and it has recently been shown that NATL SST-driven changes in the supply of water vapor support skillful seasonal predictions of summertime convective rainfall over northern Europe (Dunstone et al. 2018). Averaging over the same region as Sutton and Dong (2012) and Dunstone et al. (2018), our system suggests that decadal predictions of summer precipitation over northern Europe may be viable (Figs. 6a,b). While the ensemble-mean signal is clearly very weak in both CESM-DPLE and CESM-LE (Fig. 6a), the ACC (which is insensitive to magnitude) for pentadal anomalies is 0.68 with a ΔACC of 0.25 over CESM-LE. This skill improvement might be even larger if it were possible to extend hindcast start dates further back in time in order to sample more of the positive phase of the AMV and European summer precipitation in the 1950s (Fig. 6b).

The striking skill improvement over the Sahel is also in line with our current understanding of AMV impacts over Africa (Mohino et al. 2016; Sutton and

Hodson 2005; Wang et al. 2012; Zhang and Delworth 2006). On multiyear time scales, a warmer NATL is associated with a northward shift in the intertropical convergence zone (ITCZ) and enhanced moisture supply for the West African monsoon (Green et al. 2017; Sheen et al. 2017). CESM-DPLE generates an ACC of 0.78 for regionally averaged summer (JAS) precipitation over the West African Sahel for lead years 3–7 (Figs. 6c,d), slightly better than persistence. The corresponding correlation from CESM-LE is only 0.11, and the high CESM-DPLE correlation is slightly higher than that obtained from a 52-member multimodel mean of CMIP5 DP hindcasts analyzed over the same region and lead interval (Martin and Thorncroft 2014). The skill for regionally averaged JAS precipitation over the Sahel increases considerably from LY1–5 to LY3–7 and then slowly diminishes (Fig. 7a). We speculate that this lead-time dependence is related to the changes in SST skill discussed earlier (Fig. 3), which result in notably higher ACC scores in the subtropical Atlantic when nontrend variability is isolated (Fig. ES3).

Previous works have identified the relative SST index (RSI; the difference between subtropical North Atlantic SST and global tropical SST) as a good predictor of Sahel summertime precipitation as well as model potential for skillful Sahel prediction (Giannini et al. 2013; Martin and Thorncroft 2014). CESM-DPLE does indeed show higher RSI skill than any of the reference forecasts considered herein—much higher than the previous CCSM4-DP system that was characterized by a large initialization shock (Fig. 7b; see also the supplemental material). While the ability (or lack thereof) to predict RSI offers some explanation for the differences in Sahel precipitation skill between the different forecast systems (Fig. 7a), it does not really explain why CESM-DPLE skill peaks at LY3–7. Curiously, the high CESM-DPLE skill at predicting SPNA SST noted earlier is no better than in the old CCSM4-DP system (Fig. 7c), implying that

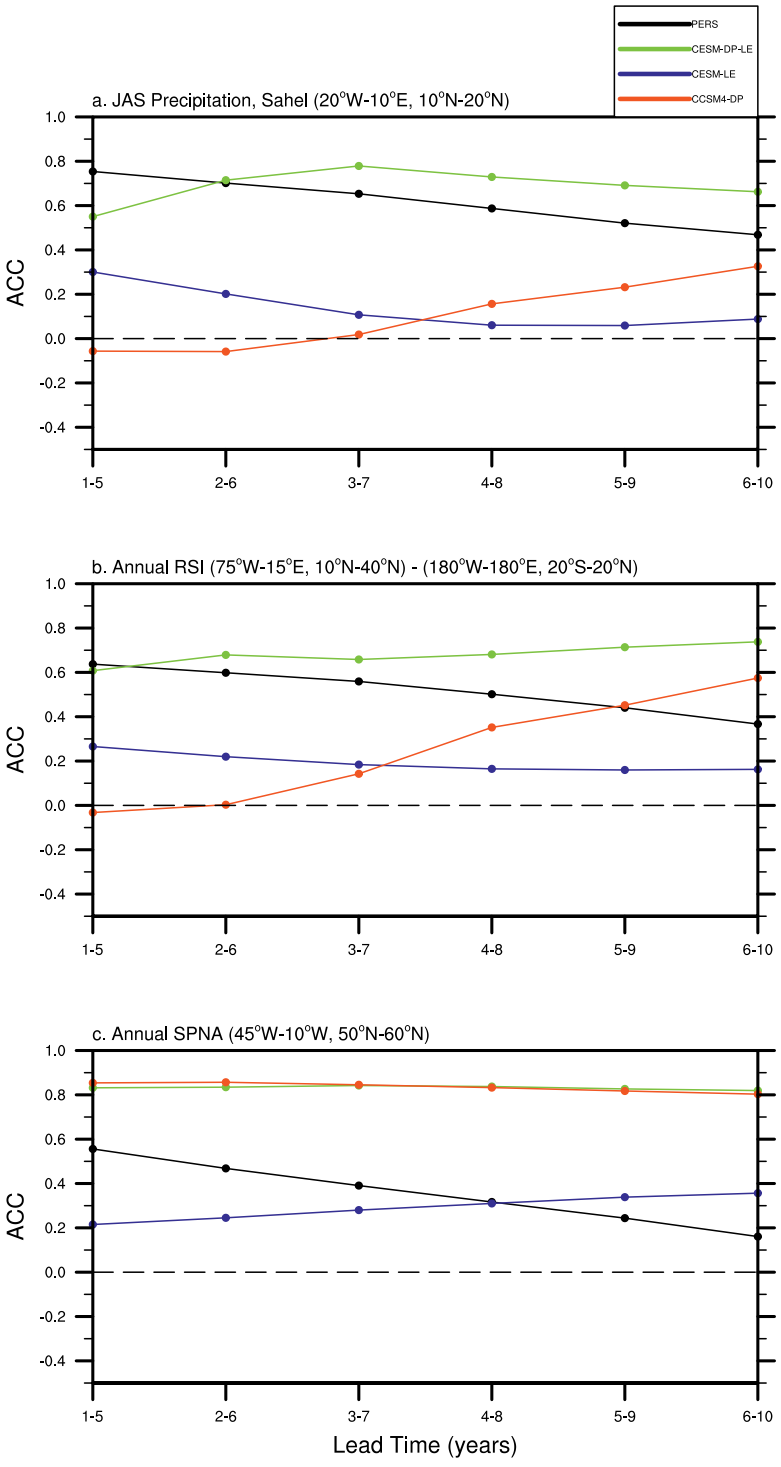

Fig. 7. ACC as a function of lead time for (a) boreal summer (JAS) precipitation averaged over the Sahel (10°–20°N, 20°W–10°E), (b) the annual RSI (the difference between subtropical North Atlantic SST averaged over 10°–40°N, 75°W–15°E and global tropical SST averaged over 20°S–20°N), and (c) annual SST in the SPNA (50°–60°N, 45°–10°W).

SPNA SST skill alone does not guarantee skillful Sahel precipitation even though it is likely an important ingredient (Dunstone et al. 2011).

The apparent success of CESM-DPLE in skillfully hindcasting the relevant SST drivers of Sahel precipitation suggests that the forecast of drought conditions through 2020 (which contrasts sharply with the CESM-LE near-term projection of above-normal precipitation) should be taken into consideration by relevant stakeholders (Fig. 6d). We note that this forecast is not inconsistent with a number of recent studies that anticipate a shift toward a cooler NATL (negative AMV) as a result of a weakening Atlantic thermohaline circulation (Hermanson et al. 2014; Robson et al. 2014, 2016; Yeager et al. 2015).

NEW CAPABILITIES. *Large ensemble.* The large ensemble size of CESM-DPLE, in conjunction with that of the CESM-LE, permits unprecedented exploration of the sensitivity of DP skill assessment to the level of noise reduction achieved through ensemble averaging (Boer et al. 2013). Figure 8 shows how the precipitation skill over the northern Europe and Sahel regions discussed above (Fig. 6) varies as a function of ensemble size. With 40-member ensembles, the initialized skill score of $r = 0.68$ for summer precipitation over Europe can be confidently distinguished from the UI ensemble that yields $r > 0.4$ for this region (Fig. 8a). While there is a discernible benefit from initialization, the external forcing is clearly contributing quite a lot to the overall skill in this region. Even with a 40-member UI ensemble, there is considerable uncertainty in the skill associated with external forcing (the 90% confidence interval for the CESM-LE correlation spans about 0.3 for a 40-member ensemble and this increases to about 0.7 for a 5-member ensemble). Cleanly distinguishing initialized from uninitialized simulation skill in this region clearly requires large ensembles for both simulation sets (probably 20+ members). For the Sahel region, the 90% confidence interval for a 10-member UI set spans about 0.8 (Fig. 8b). In that region, however, the benefits of initialization are clearly evident even with a 10-member ensemble despite the UI uncertainty. To confidently (at the 90% level) beat persistence in the Sahel, however, probably requires an initialized ensemble of 30 or more. In short, the

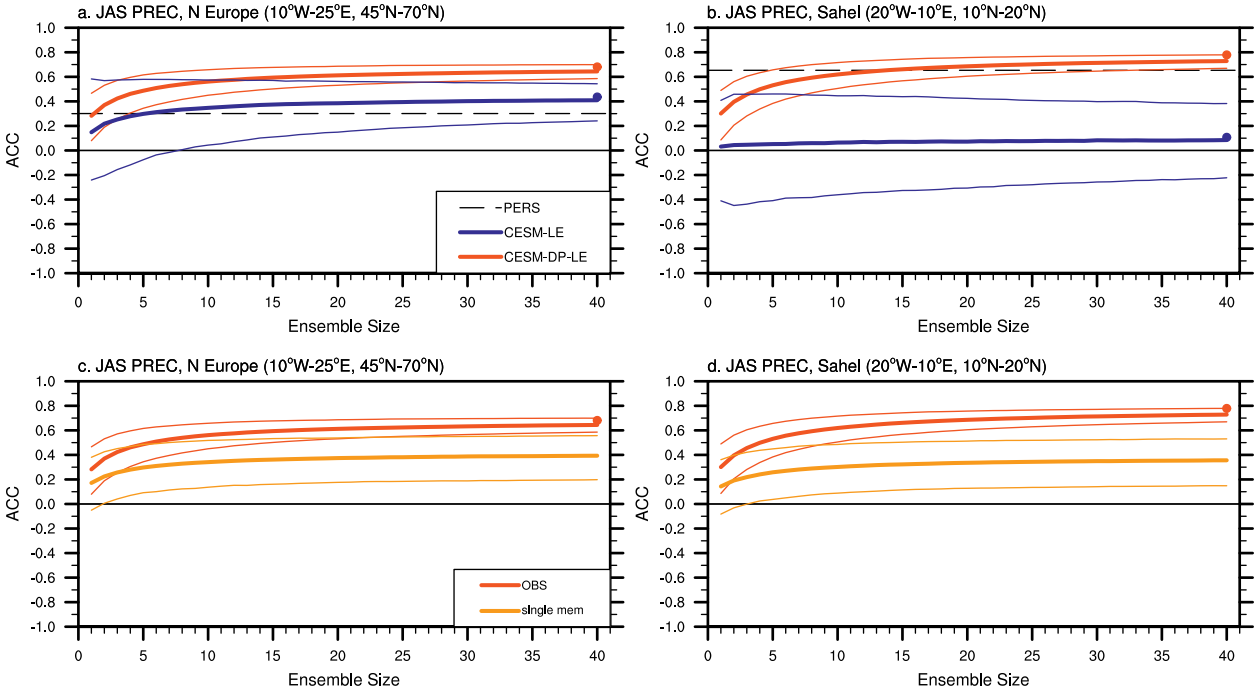

Fig. 8. ACC skill for boreal summer (JAS) land precipitation relative to the CRU-TS4.0 dataset (Harris et al. 2014) for pentadal anomalies over (a),(c) northern Europe (45°–70°N, 10°W–25°E) and (b),(d) the Sahel (10°–20°N, 20°W–10°E). CESM-DPLE data are from LY3–7, as shown in Fig. 6. The skill dependence on ensemble size is computed using a bootstrapped resampling (with replacement) of ensemble members from the 40-member pools of CESM-LE (blue) and CESM-DPLE (red). The thick lines are the median, and the thin lines are the 5th and 95th percentiles, of a score distribution of size 10,000. The black dashed lines show the skill of the persistence forecast, and the filled dots give the ACC score for the unique 40-member ensemble mean from each system. In (c) and (d), the orange lines show the ACC distribution for CESM-DPLE ensemble-mean predictions of random single-member time series drawn from the CESM-DPLE pool.

confidence intervals for predictions of precipitation over land are large, particularly for the baseline skill associated with external forcing, which is very ill-defined for UI ensembles of size 10 or less, which are not uncommon in the DP literature. The implication is that robust assessment of DP skill enhancement associated with initialization for fields such as precipitation may require much larger ensembles than current protocols recommend (Boer et al. 2016).

The increase of skill with ensemble size is in line with recent studies suggesting that unrealistically small signal-to-noise ratios in current DP systems can be overcome through the noise-dampening effect of large ensembles (Dunstone et al. 2016; Eade et al. 2014; Scaife et al. 2014). With respect to predictions of regional precipitation over land in the two areas highlighted above, CESM-DPLE does appear to exhibit a signal-to-noise paradox similar to other DP systems insofar as significantly higher correlation scores are obtained when verifying against observations than when verifying against single-member model "truth" (Figs. 8c,d). The ratio of predictable components (RPC; computed as the ratio of median correlations: r_{obs}/r_{model}) is 1.64 (2.05) for 40 member predictions of summer precipitation over northern Europe (Sahel). This suggests that CESM-DPLE predictions of precipitation over land are underconfident, characterized by unrealistically low signal to noise, and that the real-world predictability may be higher than what is implied by the model ensemble spread (Eade et al. 2014).

The large ensemble size will also facilitate exploration of the predictability of higher moments of the PDFs of climate fields, including the extreme tails that correspond to the most impactful and costly climate phenomena. Previous efforts to quantify the skill of predicting climate extremes at decadal lead times have employed relatively small ensembles (≈10) that may suffer from sampling issues (Eade et al. 2012). Examination of probabilistic skill metrics that test the quality of the ensemble spread such as the reliability (the ability of the system to realistically partition forecast probability across different forecast categories) and discrimination (the ability of the system to distinguish between observed events and nonevents) will be an important next step that will be greatly facilitated by the large ensemble.

Ocean biogeochemistry. The inclusion of ocean BGC fields in CESM-DPLE is an exciting new capability that will facilitate a wide array of research into the predictability of ocean biogeochemistry, with potentially important implications for environmental managers and policy makers. Reliable near-term predictions of marine NPP (the net rate of photosynthetic carbon fixation by phytoplankton in the surface ocean) are of particular relevance for fisheries management. To that end, Fig. 9 illustrates the potential predictability of NPP integrated over the upper 150 m of the water column in CESM-DPLE. In Fig. 9, NPP anomalies from CESM-DPLE (and from the reference forecasts) are verified against NPP anomalies from the CORE*-forced FOSI simulation used to initialize the DP simulations, rather than against observed anomalies. Therefore, the skill represents the potential for actual predictability in the limit of perfect knowledge of the initial state. Regions of very high ACC that are clearly distinguishable from both persistence and externally forced forecast skill are found in each of the world's oceans out to decadal lead times. Similar to the upper-ocean heat content and SST (Figs. 1 and 2), the Atlantic stands out as a basin with particularly long-lasting skill that derives from initialization (Figs. 9c,f,i). There are also indications of enhanced skill in western boundary current regions such as the Gulf Stream, Kuroshio, Agulhas, and East Australian Current. In several of the eastern boundary upwelling systems, such as the Canary, Benguela, and Humboldt Current regions, potential predictability is high and shows significant improvement with initialization. More extensive analysis is needed to determine the underlying mechanisms of this potential skill, and whether similar skill is seen in other elements of the ocean biosphere.

The limited coverage of BGC observations in space and time presents new challenges when it comes to assessing model fidelity and prediction skill. Satellite-derived estimates of marine NPP over the global ocean can be used for prediction skill assessment (Seferian et al. 2014), but verification is then limited to the period for which satellite ocean color observations exist. Figures 10a and 10b show the annual NPP ACC scores for the CORE*-forced FOSI and the CESM-DPLE LY1 hindcast, respectively. Here, skill is computed relative to MODIS-estimated annual NPP anomalies at each location over 2003–15. High skill is found in the central equatorial and western subtropical Pacific, central subtropical Indian, and subtropical Atlantic Oceans.

The Canary Current region (12°–22°N, 10°–25°W; Chavez and Messié 2009) is a hotspot of productivity and a critical area for African fisheries (Food and Agriculture Organization 2009), and so skillful decadal prediction in this region would be very beneficial for African resource management. In this region, CESM-DPLE exhibits both potential skill (verified against CORE* over 1955–2015; Fig. 9a) at various lead times and actual skill (verified against MODIS-based

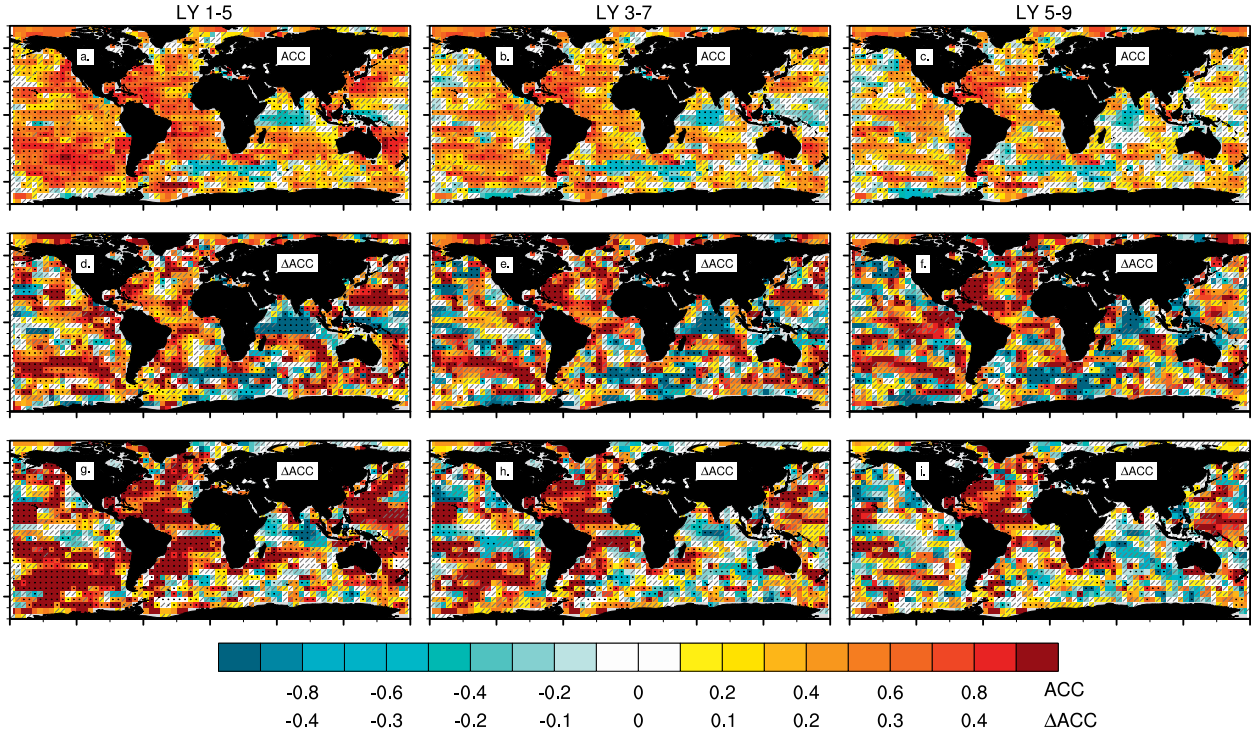

Fig. 9. (a)–(c) ACC of annual NPP from the CESM-DPLE relative to the CORE*-forced FOSI simulation used for initialization for lead times of 1–5, 3–7, and 5–9 years. ACC skill score differences between (d)–(f) between CESM-DPLE and persistence and (g)–(i) between CESM-DPLE and CESM-LE. All fields were mapped onto a 5° × 5° grid prior to analysis. The scale used for (d)–(i) is half that used for (a)–(c). The absence (presence) of a gray slash indicates scores that are (are not) significant at the 10% level (α = 0.1); stippling further indicates points whose *p* values pass an FDR test for global (70°S–70°N) field significance (α_{global} = 0.1).

observational estimates over 2003–15; Figs. 10b,c) at LY1. Furthermore, the CORE*-forced FOSI simulation shows good skill at reproducing observed NPP variability in this region (Figs. 10a,c,d), giving confidence that the high ACC scores obtained when using the full multidecadal time series of CESM-DPLE (Fig. 9) are a reliable indication of realizable skill. While NPP estimated from satellite-derived chlorophyll has many caveats, among them chlorophyll biases in the Canary Current due to Saharan dust (Chavez and Messié 2009), these results nevertheless suggest that skillful prediction of NPP in the Canary Current region may be possible.

SUMMARY. The CESM-DPLE is a new "big data" resource for the community that will permit advancements in the science of decadal prediction of the Earth system that could not be achieved through a small-scale, sole-investigator approach. The large number of hindcast start dates and ensemble members, paired with the equally large ensemble of uninitialized historical simulations that compose the CESM Large Ensemble (Kay et al. 2015), offer unprecedented statistical power for disentangling the intrinsic versus extrinsic sources of skill, exploring signal-to-noise characteristics, and studying climate extrema. The preliminary assessment of the dataset provided herein shows that CESM-DPLE exhibits quite promising levels of skill for many different fields, across a broad range of forecast lead times up to decadal scales, both in the ocean and over land. Significant skill improvement over the earlier CMIP5-era prediction set that used CCSM4 recommends its use both for new investigations and for reevaluation of earlier conclusions that were based on the less skillful CCSM4-DP. In particular, the combined analysis of CESM-DPLE and CESM-LE is revealing significant and potentially useful skill at predicting low-frequency variations in hydroclimate over land, such as over Europe and Africa, that appears to highlight the role of the ocean in modulating decadal climate variations. Large ensembles are needed to draw robust conclusions about the role of initialization in predictions of noisy atmospheric fields and to realize skill given what appear to be unrealistically small signal-to-noise ratios in the model. The inclusion of prognostic ocean biogeochemistry in CESM-DPLE opens up new prospects for predictability research that move beyond the

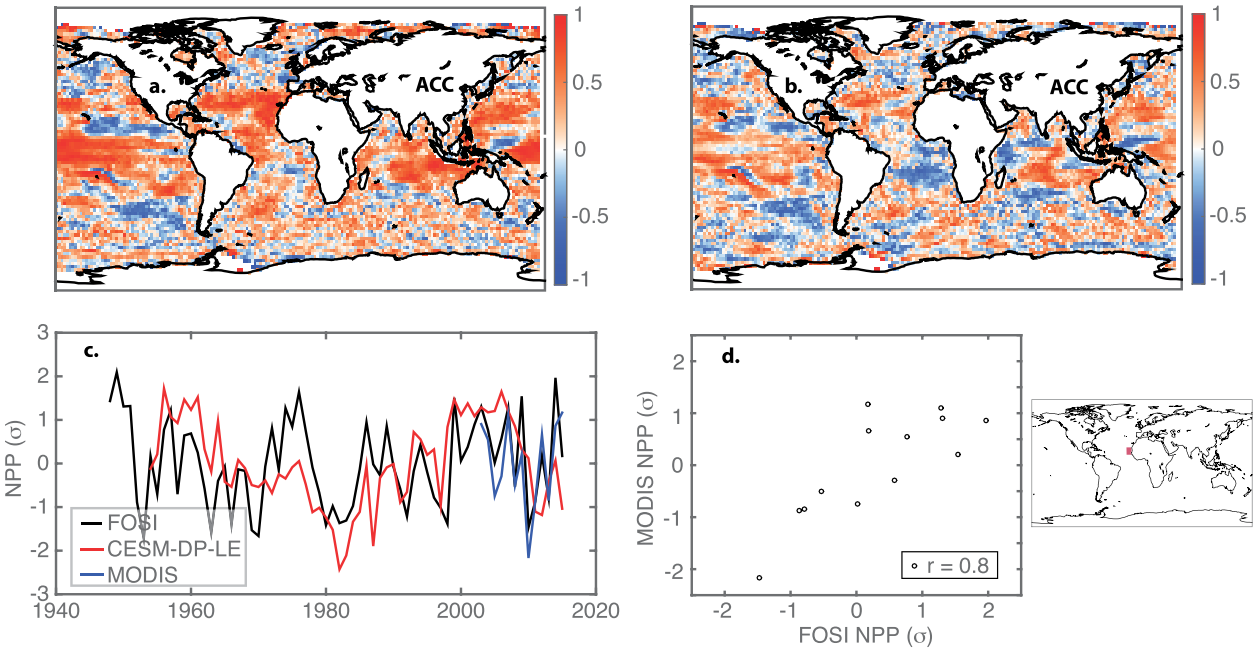

Fig. 10. (a) ACC of annual NPP from the CORE*-forced FOSI simulation, relative to NPP anomalies estimated from the MODIS-based observational product over 2003–15. (b) As in (a), but for CESM-DPLE LY1 relative to MODIS. (c) Temporal evolution of standardized NPP anomalies in the Canary Current region (pink region on map inset) from the CORE*-forced FOSI simulation (black), CESM-DPLE LY1 (red), and MODIS (blue). (d) CORE*-forced FOSI vs MODIS standardized, annual NPP anomalies in the Canary Current region over 2003–15. NPP was gridded onto a regular 2° grid prior to analysis.

physical climate system. A preliminary analysis of NPP suggests that skillful multiyear predictions of ocean biogeochemistry relevant to fisheries management (e.g., in the Canary Current region) are possible.

The output from CESM-DPLE (as well as from the CORE* simulation used to initialize the ocean and sea ice components) is available as raw, single-variable time series files. A web page (www.cesm.ucar.edu/projects/community-projects/DPLE) provides specifics about the simulations, links to the data, a publication list, and additional overview diagnostics for select fields. The companion CESM-LE simulation set has similar web documentation (www.cesm.ucar.edu/projects/community-projects/LENS), including links to output and relevant publications.

ACKNOWLEDGMENTS. This work was supported by the National Oceanic and Atmospheric Administration (NOAA) Climate Program Office under Climate Variability and Predictability Program Grants NA09OAR4310163 and NA13OAR4310138, by the National Science Foundation (NSF) Collaborative Research EaSM2 Grant OCE-1243015, and by the NSF through its sponsorship of the National Center for Atmospheric Research. GS, SB, NR, and HT are supported by the Regional and Global Climate Modeling Program (RGCM) of the U.S. Department of Energy's, Office of Science (BER), Cooperative Agreement DE-FC02-97ER62402. NL is supported by the NSF (OCE-1558225 and NSF-1752724). The CESM-DPLE was generated using computational resources provided by the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract DE-AC02-05CH11231, as well as by an Accelerated Scientific Discovery grant for Cheyenne (https://doi.org/10.5065/D6RX99HX) that was awarded by NCAR's Computational and Information Systems Laboratory. We thank Oregon State University for the satellite-derived NPP data (www.science.oregonstate.edu/ocean.productivity/).

REFERENCES

Årthun, M., and T. Eldevik, 2016: On anomalous ocean heat transport toward the Arctic and associated climate predictability. *J. Climate*, **29**, 689–704, https://doi.org/10.1175/JCLI-D-15-0448.1.

——, ——, E. Viste, H. Drange, T. Furevik, H. L. Johnson, and N. S. Keenlyside, 2017: Skillful prediction of northern climate provided by the ocean. *Nat. Commun.*, **8**, 15875, https://doi.org/10.1038/ncomms15875.

Behrenfeld, M. J., and P. G. Falkowski, 1997: Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnol. Oceanogr.*, **42**, 1–20, https://doi.org/10.4319/lo.1997.42.1.0001.

Boer, G. J., V. V. Kharin, and W. J. Merryfield, 2013: Decadal predictability and forecast skill. *Climate Dyn.*, **41**, 1817–1833, https://doi.org/10.1007/s00382-013-1705-0.

——, and Coauthors, 2016: The Decadal Climate Prediction Project (DCPP) contribution to CMIP6. *Geosci. Model Dev.*, **9**, 3751–3777, https://doi.org/10.5194/gmd-9-3751-2016.

Chavez, F. P., and M. Messié, 2009: A comparison of eastern boundary upwelling ecosystems. *Prog. Oceanogr.*, **83**, 80–96, https://doi.org/10.1016/j.pocean.2009.07.032.

Compo, G. P., and Coauthors, 2011: The Twentieth Century Reanalysis Project. *Quart. J. Roy. Meteor. Soc.*, **137**, 1–28, https://doi.org/10.1002/qj.776.

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, 2016: North Atlantic simulations in Coordinated Ocean-ice Reference Experiments phase II (CORE-II). Part II: Inter-annual to decadal variability. *Ocean Modell.*, **97**, 65–90, https://doi.org/10.1016/j.ocemod.2015.11.007.

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.

Doblas-Reyes, F. J., and Coauthors, 2013: Initialized near-term regional climate change prediction. *Nat. Commun.*, **4**, 1715, https://doi.org/10.1038/ncomms2704.

Dunstone, N. J., D. M. Smith, and R. Eade, 2011: Multi-year predictability of the tropical Atlantic atmosphere driven by the high latitude North Atlantic Ocean. *Geophys. Res. Lett.*, **38**, L14701, https://doi.org/10.1029/2011GL047949.

——, ——, A. Scaife, L. Hermanson, R. Eade, N. Robinson, M. Andrews, and J. Knight, 2016: Skilful predictions of the winter North Atlantic Oscillation one year ahead. *Nat. Geosci.*, **9**, 809–814, https://doi.org/10.1038/ngeo2824.

——, and Coauthors, 2018: Skilful seasonal predictions of summer European rainfall. *Geophys. Res. Lett.*, **45**, 3246–3254, https://doi.org/10.1002/2017GL076337.

Eade, R., E. Hamilton, D. M. Smith, R. J. Graham, and A. A. Scaife, 2012: Forecasting the number of extreme daily events out to a decade ahead. *J. Geophys. Res.*, **117**, D21110, https://doi.org/10.1029/2012JD018015.

——, D. Smith, A. Scaife, E. Wallace, N. Dunstone, L. Hermanson, and N. Robinson, 2014: Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophys. Res. Lett.*, **41**, 5620–5628, https://doi.org/10.1002/2014GL061146.

Food and Agriculture Organization, 2009: Protection of the Canary Current Large Marine Ecosystem (CCLME). FAO/Global Environment Facility Project Doc., 65 pp., www.fao.org/3/a-bo678e.pdf.

Giannini, A., S. Salack, T. Lodoun, A. Ali, A. T. Gaye, and O. Ndiaye, 2013: A unifying view of climate change in the Sahel linking intra-seasonal, interannual and longer time scales. *Environ. Res. Lett.*, **8**, 024010, https://doi.org/10.1088/1748-9326/8/2/024010.

Goddard, L., and Coauthors, 2013: A verification framework for interannual-to-decadal predictions experiments. *Climate Dyn.*, **40**, 245–272, https://doi.org/10.1007/s00382-012-1481-2.

Good, S. A., M. J. Martin, and N. A. Rayner, 2013: EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *J. Geophys. Res. Oceans*, **118**, 6704–6716, https://doi.org/10.1002/2013JC009067.

Green, B., J. Marshall, and A. Donohoe, 2017: Twentieth century correlations between extratropical SST variability and ITCZ shifts: AMO, PDO, and ITCZ variability. *Geophys. Res. Lett.*, **44**, 9039–9047, https://doi.org/10.1002/2017GL075044.

Han, W., and Coauthors, 2010: Patterns of Indian Ocean sea-level change in a warming climate. *Nat. Geosci.*, **3**, 546–550, https://doi.org/10.1038/ngeo901.

Hanawa, K., and L. Talley, 2001: Mode waters. *Ocean Circulation and Climate: Observing and Modelling the Global Ocean*, G. Siedler, J. Church, and J. Gould, Eds., International Geophysics Series, Vol. 77, Academic Press, 373–386.

Harris, I., P. D. Jones, T. J. Osborn, and D. H. Lister, 2014: Updated high-resolution grids of monthly climatic observations—The CRU TS3.10 dataset. *Int. J. Climatol.*, **34**, 623–642, https://doi.org/10.1002/joc.3711.

Hermanson, L., R. Eade, N. H. Robinson, N. J. Dunstone, M. B. Andrews, J. R. Knight, A. A. Scaife, and D. M. Smith, 2014: Forecast cooling of the Atlantic subpolar gyre and associated impacts. *Geophys. Res. Lett.*, **41**, 5167–5174, https://doi.org/10.1002/2014GL060420.

Huang, B., and Coauthors, 2017: Extended Reconstructed Sea Surface Temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *J. Climate*, **30**, 8179–8205, https://doi.org/10.1175/JCLI-D-16-0836.1.

Hunke, E. C., and W. H. Lipscomb, 2008: CICE: The Los Alamos sea ice model, documentation and software, version 4.0. Los Alamos National Laboratory Tech. Rep. LA-CC-06-012, 76 pp.

Hurrell, J. W., 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.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**, 437–471, https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Karspeck, A., S. Yeager, G. Danabasoglu, and H. Teng, 2015: An evaluation of experimental decadal predictions using CCSM4. *Climate Dyn.*, **44**, 907–923, https://doi.org/10.1007/s00382-014-2212-7.

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.

Keenlyside, N. S., M. Latif, J. Jungclaus, L. Kornblueh, and E. Roeckner, 2008: Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature*, **453**, 84–88, https://doi.org/10.1038/nature06921.

Kim, H.-M., P. J. Webster, and J. A. Curry, 2012: Evaluation of short-term climate change prediction in multimodel CMIP5 decadal hindcasts. *Geophys. Res. Lett.*, **39**, L10701, https://doi.org/10.1029/2012GL051644.

Kirtman, B., and Coauthors, 2013: Near-term climate change: Projections and predictability. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 953–1028.

Large, W. G., and S. G. Yeager, 2009: The global climatology of an interannually varying air–sea flux data set. *Climate Dyn.*, **33**, 341–364, https://doi.org/10.1007/s00382-008-0441-3.

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/10.1029/2011MS000453.

Lindsay, K., and Coauthors, 2014: Preindustrial-control and twentieth-century carbon cycle experiments with the Earth System Model CESM1(BGC). *J. Climate*, **27**, 8981–9005, https://doi.org/10.1175/JCLI-D-12-00565.1.

Long, M. C., K. Lindsay, S. Peacock, J. K. Moore, and S. C. Doney, 2013: Twentieth-century oceanic carbon uptake and storage in CESM1(BGC). *J. Climate*, **26**, 6775–6800, https://doi.org/10.1175/JCLI-D-12-00184.1.

Long, M. C., C. Deutsch, and T. Ito, 2016: Finding forced trends in oceanic oxygen. *Global Biogeochem. Cycles*, **30**, 381–397, https://doi.org/10.1002/2015GB005310.

Martin, E. R., and C. Thorncroft, 2014: Sahel rainfall in multimodel CMIP5 decadal hindcasts. *Geophys. Res. Lett.*, **41**, 2169–2175, https://doi.org/10.1002/2014GL059338.

Meehl, G. A., and H. Teng, 2012: Case studies for initialized decadal hindcasts and predictions for the Pacific region. *Geophys. Res. Lett.*, **39**, L22705, https://doi.org/10.1029/2012GL053423.

——, and ——, 2014a: CMIP5 multi-model hindcasts for the mid-1970s shift and early 2000s hiatus and predictions for 2016–2035. *Geophys. Res. Lett.*, **41**, 1711–1716, https://doi.org/10.1002/2014GL059256.

——, and ——, 2014b: Regional precipitation simulations for the mid-1970s shift and early-2000s hiatus: Regional precipitation. *Geophys. Res. Lett.*, **41**, 7658–7665, https://doi.org/10.1002/2014GL061778.

——, and Coauthors, 2009: Decadal prediction: Can it be skillful? *Bull. Amer. Meteor. Soc.*, **90**, 1467–1486, https://doi.org/10.1175/2009BAMS2778.1.

——, and Coauthors, 2012: Climate system response to external forcings and climate change projections in CCSM4. *J. Climate*, **25**, 3661–3683, https://doi.org/10.1175/JCLI-D-11-00240.1.

——, and Coauthors, 2014: Decadal climate prediction: An update from the trenches. *Bull. Amer. Meteor. Soc.*, **95**, 243–267, https://doi.org/10.1175/BAMS-D-12-00241.1.

——, A. Hu, and H. Teng, 2016: Initialized decadal prediction for transition to positive phase of the interdecadal Pacific oscillation. *Nat. Commun.*, **7**, 11718, https://doi.org/10.1038/ncomms11718.

Mohino, E., N. Keenlyside, and H. Pohlmann, 2016: Decadal prediction of Sahel rainfall: Where does the skill (or lack thereof) come from? *Climate Dyn.*, **47**, 3593–3612, https://doi.org/10.1007/s00382-016-3416-9.

Monerie, P.-A., J. Robson, B. Dong, and N. Dunstone, 2017: A role of the Atlantic Ocean in predicting summer surface air temperature over North East Asia? *Climate Dyn.*, https://doi.org/10.1007/s00382-017-3935-z, in press.

Moore, J. K., K. Lindsay, S. C. Doney, M. C. Long, and K. Misumi, 2013: Marine ecosystem dynamics and biogeochemical cycling in the Community Earth System Model [CESM1(BGC)]: Comparison of the 1990s with the 2090s under the RCP4.5 and RCP8.5 scenarios. *J. Climate*, **26**, 9291–9312, https://doi.org/10.1175/JCLI-D-12-00566.1.

Pohlmann, H., J. H. Jungclaus, A. Köhl, D. Stammer, and J. Marotzke, 2009: Initializing decadal climate predictions with the GECCO oceanic synthesis: Effects on the North Atlantic. *J. Climate*, **22**, 3926–3938, https://doi.org/10.1175/2009JCLI2535.1.

——, J. Kröger, R. J. Greatbatch, and W. A. Müller, 2016: Initialization shock in decadal hindcasts due to errors in wind stress over the tropical Pacific. *Climate Dyn.*, **49**, 2685–2693, https://doi.org/10.1007/s00382-016-3486-8.

Robson, J., D. Hodson, E. Hawkins, and R. Sutton, 2014: Atlantic overturning in decline? *Nat. Geosci.*, **7**, 2–3, https://doi.org/10.1038/ngeo2050.

——, P. Ortega, and R. Sutton, 2016: A reversal of climatic trends in the North Atlantic since 2005. *Nat. Geosci.*, **9**, 513–517, https://doi.org/10.1038/ngeo2727.

Scaife, A. A., and Coauthors, 2014: Skillful long-range prediction of European and North American winters. *Geophys. Res. Lett.*, **41**, 2514–2519, https://doi.org/10.1002/2014GL059637.

Seferian, R., L. Bopp, M. Gehlen, D. Swingedouw, J. Mignot, E. Guilyardi, and J. Servonnat, 2014: Multiyear predictability of tropical marine productivity. *Proc. Natl. Acad. Sci. USA*, **111**, 11 646–11 651, https://doi.org/10.1073/pnas.1315855111.

Sheen, K. L., D. M. Smith, N. J. Dunstone, R. Eade, D. P. Rowell, and M. Vellinga, 2017: Skilful prediction of Sahel summer rainfall on inter-annual and multi-year timescales. *Nat. Commun.*, **8**, 14966, https://doi.org/10.1038/ncomms14966.

Sienz, F., W. A. Müller, and H. Pohlmann, 2015: Ensemble size impact on the decadal predictive skill assessment. *Meteor. Z.*, **25**, 645–655, https://doi.org/10.1127/metz/2016/0670.

Smith, D. M., S. Cusack, A. W. Colman, C. K. Folland, G. R. Harris, and J. M. Murphy, 2007: Improved surface temperature prediction for the coming decade from a global climate model. *Science*, **317**, 796–799, https://doi.org/10.1126/science.1139540.

Sutton, R. T., and D. L. R. Hodson, 2005: Atlantic Ocean forcing of North American and European summer climate. *Science*, **309**, 115–118, https://doi.org/10.1126/science.1109496.

——, and B. Dong, 2012: Atlantic Ocean influence on a shift in European climate in the 1990s. *Nat. Geosci.*, **5**, 788–792, https://doi.org/10.1038/ngeo1595.

Teng, H., G. A. Meehl, G. Branstator, S. Yeager, and A. Karspeck, 2017: Initialization shock in CCSM4 decadal prediction experiments. *CLIVAR Exchanges*, No. 72, International CLIVAR Project Office, Southampton, United Kingdom, 41–46.

Tsujino, H., and Coauthors, 2018: JRA-55 based surface dataset for driving ocean–sea-ice models (JRA55-do). *Ocean Modell.*, in press.

Van Oldenborgh, G. J., F. J. Doblas-Reyes, B. Wouters, and W. Hazeleger, 2012: Decadal prediction skill in a multi-model ensemble. *Climate Dyn.*, **38**, 1263–1280, https://doi.org/10.1007/s00382-012-1313-4.

Ventura, V., C. J. Paciorek, and J. S. Risbey, 2004: Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data. *J. Climate*, **17**, 4343–4356, https://doi.org/10.1175/3199.1.

Wang, C., S. Dong, A. T. Evan, G. R. Foltz, and S.-K. Lee, 2012: Multidecadal covariability of North Atlantic sea surface temperature, African dust, Sahel rainfall, and Atlantic hurricanes. *J. Climate*, **25**, 5404–5415, https://doi.org/10.1175/JCLI-D-11-00413.1.

Wilks, D. S., 2006: On "field significance" and the false discovery rate. *J. Appl. Meteor. Climatol.*, **45**, 1181–1189, https://doi.org/10.1175/JAM2404.1.

——, 2016: "The stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. *Bull. Amer. Meteor. Soc.*, **97**, 2263–2273, https://doi.org/10.1175/BAMS-D-15-00267.1.

Yeager, S., and G. Danabasoglu, 2014: The origins of late-twentieth-century variations in the large-scale North Atlantic circulation. *J. Climate*, **27**, 3222–3247, https://doi.org/10.1175/JCLI-D-13-00125.1.

——, and J. I. Robson, 2017: Recent progress in understanding and predicting Atlantic decadal climate variability. *Curr. Climate Change Rep.*, **3**, 112–127, https://doi.org/10.1007/s40641-017-0064-z.

——, A. Karspeck, G. Danabasoglu, J. Tribbia, and H. Teng, 2012: A decadal prediction case study: Late twentieth-century North Atlantic ocean heat content. *J. Climate*, **25**, 5173–5189, https://doi.org/10.1175/JCLI-D-11-00595.1.

——, ——, and ——, 2015: Predicted slowdown in the rate of Atlantic sea ice loss. *Geophys. Res. Lett.*, **42**, 10 704–10 713, https://doi.org/10.1002/2015GL065364.

Zhang, R., and T. L. Delworth, 2006: Impact of Atlantic multidecadal oscillations on India/Sahel rainfall and Atlantic hurricanes. *Geophys. Res. Lett.*, **33**, L17712, https://doi.org/10.1029/2006GL026267.