Prospects for Improving Subseasonal Predictions

KATHY PEGION AND PRASHANT D. SARDESHMUKH

CIRES Climate Diagnostics Center, University of Colorado, and NOAA/Earth System Research Laboratory, Boulder, Colorado

(Manuscript received 30 December 2010, in final form 20 April 2011)

ABSTRACT

Extending atmospheric prediction skill beyond the predictability limit of about 10 days for daily weather rests on the hope that some time-averaged aspects of anomalous circulations remain predictable at longer forecast lead times, both because of the existence of natural low-frequency modes of atmospheric variability and coupling to the ocean with larger thermal inertia. In this paper the week-2 and week-3 forecast skill of two global coupled atmosphere–ocean models recently developed at NASA and NOAA is compared with that of much simpler linear inverse models (LIMs) based on the observed time-lag correlations of atmospheric circulation anomalies in the Northern Hemisphere and outgoing longwave radiation (OLR) anomalies in the tropics. The coupled models are found to beat the LIMs only slightly, and only if an ensemble prediction methodology is employed. To assess the potential for further skill improvement, a predictability analysis based on the relative magnitudes of forecast signal and forecast noise in the LIM framework is conducted. Estimating potential skill by such a method is argued to be superior to using the ensemble-mean and ensemble-spread information in the coupled model ensemble prediction system. The LIM-based predictability analysis yields relatively conservative estimates of the potential skill, and suggests that outside the tropics the average coupled model skill may already be close to the potential skill, although there may still be room for improvement in the tropical forecast skill.

1. Introduction

It has long been recognized that although daily weather is not predictable beyond about 10 days, some aspects of anomalous weekly and longer-term averages remain predictable because of slowly evolving boundary conditions and external forcing. Also, the geographical patterns of anomalous weekly and longer-term averages are distinct from those of daily weather, which suggests that they are associated with different and possibly predictable evolution mechanisms. The Pacific–North American (PNA) and North Atlantic Oscillation (NAO) patterns in the middle and high latitudes, and the patterns associated with the Madden–Julian oscillation (MJO) in the tropics, are good examples of such patterns. The MJO in particular, with its putative oscillation period of around 50 days, is thought to be a ''source of predictability'' not just for tropical but also extratropical variations through tropical–extratropical teleconnection mechanisms (e.g., Winkler et al. 2001; Waliser et al. 2003a,b; Newman et al. 2003; Liess et al. 2005; Pegion and Kirtman 2008; Fu et al. 2007, 2008; Dole 2008). It is also worth emphasizing that although El Niño–Southern Oscillation (ENSO) is primarily an interannual phenomenon, its rapid evolution over intervals as short as a few weeks can amount to another significant source of predictability even at subseasonal forecast ranges (Newman et al. 2003).

Partly motivated by the above considerations, there is growing interest in developing useful forecast information beyond 10 days. For example, the National Centers for Environmental Prediction Climate Prediction Center (NCEP/CPC) in the United States, and the Bureau of Meteorology (BoM) in Australia, currently make ''outlook'' type forecasts for extended forecast ranges. The NCEP/CPC global tropical hazards/benefits assessment (see online at http://www.cpc.ncep.noaa.gov/products/precip/CWlink/ghazards/index.php) for week 1 and week 2 provides forecasts of anomalous tropical temperature and precipitation. The U.S. hazards assessment product (see online at http://www.cpc.ncep.noaa.gov/products/expert_assessment/threats.shtml), also issued by NCEP/CPC, includes outlooks of potential hazards in the United States for the next 3–16 days. The Australian BoM produces a weekly tropical climate note

Corresponding author address: Kathy Pegion, NOAA/ESRL, 325 Broadway, R/PSD1, Boulder, CO 80305.
E-mail: kathy.pegion@noaa.gov

DOI: 10.1175/MWR-D-11-00004.1

© 2011 American Meteorological Society

(see online at http://www.bom.gov.au/climate/tropnote/tropnote.shtml) that includes a discussion of recent and projected intraseasonal patterns that are expected to impact northern Australia in the next 1–2 weeks. At present such outlook-style forecast products are based on a subjective combination of various statistical and dynamical methods, although there is momentum to make the process more objective using real-time general circulation model (GCM) forecasts (Gottshalck et al. 2008). Such an enterprise naturally presumes that useful forecasts at these longer lead times are routinely possible, which still remains to be demonstrated.

There are ongoing efforts to improve the accuracy of real-time GCM forecasts at subseasonal forecast ranges (National Research Council 2010). Although there have been clear demonstrations of nonzero forecast skill at the week-2 and week-3 ranges, the general skill of the forecasts on these subseasonal time scales remains low. To what extent this may be due to current forecast system deficiencies or the inherent unpredictability of the ocean–atmosphere system at these forecast ranges is an important question. Addressing this question is the principal focus of this paper.

Consistent with the predictability limit of about 10 days for daily weather, the nonlinear evolution of the atmosphere from an initial condition becomes strongly chaotic after 10 days. Only a relatively minor and arguably ''linear'' part of the evolution, associated with the dominant patterns of low-frequency variability, remains predictable. Primarily because of this, even very simple linear-regression-based statistical models [e.g., linear inverse models (LIMs)] provide important benchmarks for comprehensive GCMs at subseasonal forecast ranges. For example, Winkler et al. (2001) showed that a simple LIM of weekly averaged extratropical circulation and tropical diabatic heating anomalies was more skillful at week 3 in predicting upper-tropospheric streamfunction anomalies than the comprehensive GCM used in the Dynamical Extended Range Forecast (DERF) project at NCEP. Newman et al. (2003) showed further that a version of the NCEP Medium-Range Forecast (MRF) model that was used operationally in 1998 was less skillful at week 3 than Winkler et al.'s LIM in predicting 250-hPa streamfunction, and at week 2 in predicting tropical heating anomalies, in both winter and summer.

In this study, we compare the week-2 and week-3 forecast skill of two state-of-the-art coupled ocean–atmosphere general circulation models (CGCMs) with the corresponding skill of LIMs similar to those used by Winkler et al. (2001) and Newman et al. (2003). The CGCMs are described below in section 2, and the LIMs in section 3. These models differ from those in the previous studies in two important ways: 1) they are coupled atmosphere–ocean rather than atmosphere-only models; and 2) they are much newer, incorporating almost a full decade of model developments, and also improvements in data assimilation techniques for providing better forecast initial conditions. Our original motivation for this study was to assess whether using such improved models would lead one to reach a different conclusion vis-à-vis LIMs concerning week-2 and week-3 forecast skill, and if not, to what extent this may be due to the model skill already being near the maximum realizable or ''potential'' skill associated with intrinsic predictability limits at these forecast ranges.

The principal conclusion from our analysis is that the newer CGCMs can indeed beat simple LIMs at the week-2 and week-3 forecast ranges (section 4), but only if an ensemble forecasting methodology is adopted. A second, more surprising, conclusion is that further refinements to even an ensemble forecasting system will not necessarily lead to higher *average* extratropical forecast skill at these forecast ranges than is currently achievable (section 5), although there may still be room for improvement in the tropical forecast skill. This second conclusion is drawn from ratios of the forecast signal and forecast noise variances in a LIM-based predictability framework, which are argued to yield more conservative, but also more credible, estimates of potential skill than those derived directly from the ensemble-mean signal and ensemble-spread information in current ensemble prediction systems. It is shown that the latter approach yields unrealistically optimistic estimates of potential skill owing to both overestimation of the forecast signal and underestimation of the forecast noise in the ensemble modeling system considered here.

2. Dynamical models and hindcast datasets used

We assessed the week-2 and week-3 forecast skill of weekly averaged anomalous Northern Hemisphere 200- and 850-hPa streamfunction and tropical (30°S–30°N) outgoing longwave radiation (OLR) using output from the so-called reforecast (i.e., hindcast) runs (see Table 1) initialized in December, January, and February using two state-of-the-art CGCMs: 1) the NCEP Climate Forecast System version 1 (CFS; Saha et al. 2006; Wang et al. 2005), and 2) the National Aeronautics and Space Administration (NASA) Goddard Earth Observing System, version 5 (GEOS-5; Rienecker et al. 2008).

The NCEP/CFS consists of the Global Forecast System (GFS) atmospheric model at T62 horizontal resolution and 64 vertical levels coupled to the Geophysical Fluid Dynamics Laboratory Modular Ocean Model version 3 (MOM3; Pacanowski and Griffies 1998) with 40 vertical levels and $1° \times 1/3°$ horizontal resolution in the tropics (10°S–10°N) decreasing to $1° \times 1°$ resolution poleward of

TABLE 1. Summary of CGCM experiments.

Model	AGCM	OGCM	Hindcasts	Initialization
NASA GEOS-5	GEOS-5 $2° \times 2.5° \times 72$ vertical levels	MOM4 360×200 grid points $\times 50$ vertical levels	6-month hindcasts from 1980–2005	Daily 2100 UTC from replay runs
NCEP CFSv1	GFS T62L64	MOM3	9-month hindcasts from 1981–2005	15 times per month from NCEP R2 (atm) and GODAS (ocn)

30°. The atmosphere and ocean exchange surface flux and SST information once a day. The atmosphere in the reforecast runs was initialized using states from the NCEP–Department of Energy (DOE) Reanalysis-2 dataset (NCEP R2; Kanamitsu et al. 2002) and the ocean using states from the NCEP Global Ocean Data Assimilation System dataset (GODAS; Behringer and Xue 2004; Behringer 2007).

The NASA GEOS-5 coupled model consists of the GEOS-5 atmospheric model at $2° \times 2.5°$ longitude–latitude horizontal resolution and 72 vertical levels coupled to the GFDL/MOM version 4 (MOM4; Griffies et al. 2008) ocean model with 50 vertical levels and 360 and 200 grid points in the zonal and meridional directions, respectively. The atmosphere and ocean exchange fluxes and SST every 30 min. The atmosphere in the reforecast runs was initialized daily using the 2100 UTC atmospheric states from the NASA/Global Modeling and Assimilation Office Modern Era Reanalysis (MERRA; Rienecker et al. 2008) ''scout'' runs, which are coarse-resolution ($2° \times 2.5°$) runs of the MERRA system. The ocean initial states were taken from the so-called replay runs, which use the data assimilation system to produce initial ocean states that are physically consistent with the atmospheric initial conditions.

The CFS hindcasts were initialized on dates 1–3, 9–13, 19–23, and the last 2 days of each month for the years 1981–2005, and run for 9 months. The GEOS-5 hindcasts were initialized every day in 1979–2005 and run for 6 months. For both modeling systems, the daily forecast output was made available by the respective institutions. (The CFS data are also publicly available online at http://cfs.ncep.noaa.gov.) The two modeling systems and hindcast experiments are summarized in Table 1.

Prior to assessing the week-2 and week-3 forecast skill of these models, we consider in Fig. 1 their biases in tropical sea surface temperature (SST), OLR, and 850-hPa winds. These biases were estimated as the 6-month forecast biases relative to the Met Office's Hadley Center Sea Ice and SST (HadISST; Rayner et al. 2003), the National Oceanic and Atmospheric Administration (NOAA) Interpolated OLR (Liebmann and Smith 1996), and the NCEP–DOE Reanalysis-2 (Kanamitsu et al. 2002) fields, respectively. The GEOS-5 has a cold bias of as much as 5°C in the eastern Pacific cold tongue region and a warm bias of 1°C in the western Pacific, Indian, and southern Pacific Oceans. The CFS has generally much smaller SST biases, with a maximum bias of about 1°C confined to the far eastern Pacific. Both models have large positive OLR biases over the western Pacific, up to 60 W m^{-2} in the GEOS-5 and 30 W m^{-2} in the CFS, indicating that they both drift over 6 months toward a state of weak mean atmospheric convection. Associated with these large SST and OLR biases are also large near-surface wind biases, in both models. There are large (>5 m s^{-1}) easterly biases in the 850-hPa zonal winds in the GEOS-5 model throughout the subtropics and over the equatorial Pacific Ocean. The wind biases are somewhat weaker (~ 2.5 m s^{-1}) in the CFS model, and generally of opposite sign to the GEOS-5 biases except over the subtropical northern Pacific Ocean. The interrelationships between the SST, OLR, and 850-hPa wind biases of the two models are interesting and worthy of additional investigation, but will not be pursued further here.

Although the forecast biases at week 2 and week 3 are not as large as the 6-month biases shown in Fig. 1, they are nevertheless evident even at these shorter lead times (not shown) and were therefore removed prior to assessing the model forecast skill. This was done by removing a forecast-lead-dependent daily climatology from the model forecasts. The climatology was calculated as the average over all years for each day of the year at each forecast lead time. For example, forecasts initialized on 1 January were averaged over all years for days $1, 2, 3, \ldots,$ day N, thus providing a lead-dependent climatology for forecasts initialized on 1 January. It is important to note that although the biases were removed prior to assessing the forecast skill of the models, such a procedure does not eliminate all impacts of the biases, as will be evident in the forecast skill of the tropical OLR examined in section 4.

3. Linear inverse models

a. General description

LIMs are derived from the observed lag-covariance statistics of a dynamical system. They typically have only a few degrees of freedom, are extremely inexpensive to run, and provide challenging benchmarks for comprehensive GCMs

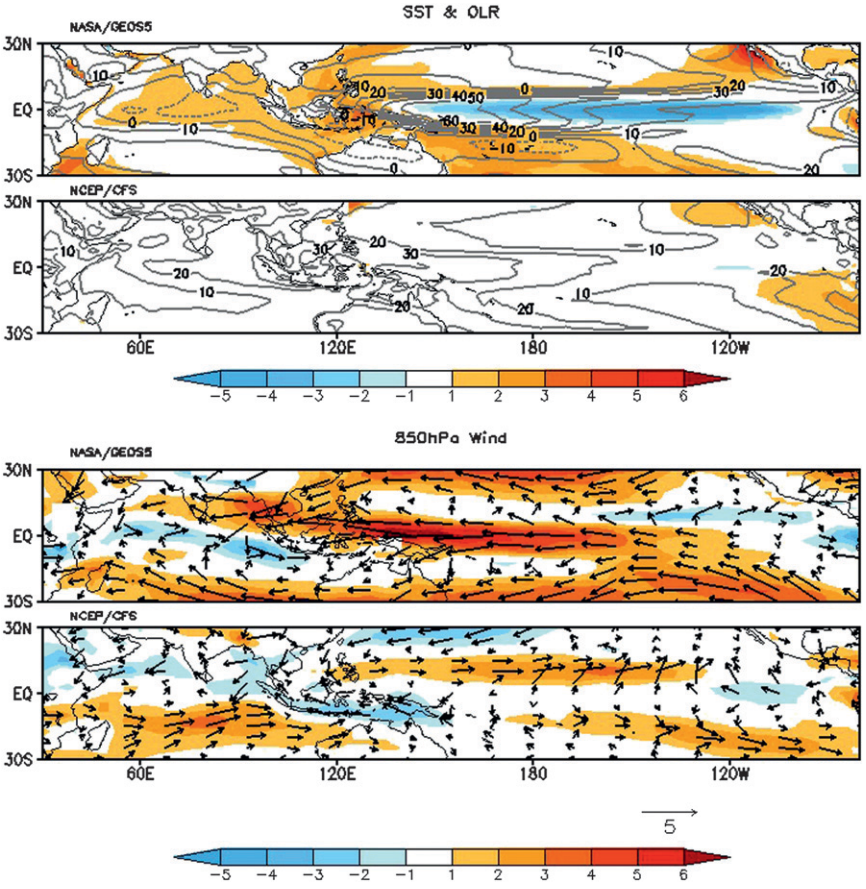

FIG. 1. (top) Bias of 6-month winter (DJF) forecasts for 1981–2005 (a) NASA GEOS-5 and (b) NCEP/CFS. Shading is SST bias (°C) relative to HadISST. Contours (gray) are OLR bias (W m^{-2}) relative to NOAA interpolated OLR. Contour interval is 1 W m^{-2} with the zero contour suppressed. (bottom) Arrows are bias in 850-hPa winds from NCEP R2. The shading indicates the bias in zonal wind speed.

as shown in many studies. Here, we provide a brief summary of linear inverse modeling and the specific model form implemented in this study. More detailed basic information may be found in Penland and Sardeshmukh (1995).

The fundamental assumption of linear inverse modeling is that the evolution of the dynamical system under consideration can be approximated as that of a linear stochastically forced system of the following form:

$$\frac{d\mathbf{x}}{dt} = \mathbf{B}\mathbf{x} + \mathbf{F}, \tag{1}$$

where $\mathbf{x}$ is the anomaly state vector, $\mathbf{B}$ is a constant matrix, and $\mathbf{F}$ is a stochastic noise vector. Any system of this type satisfies $\mathbf{C}(\tau) = \mathbf{G}(\tau)\mathbf{C}(0)$, with $\mathbf{G}(\tau) = \exp(\mathbf{B}\tau)$, where $\mathbf{C}(\tau)_{i,j} = \langle \mathbf{x}_i(t+\tau)\mathbf{x}_j(t) \rangle$ defines the covariance matrix of $\mathbf{x}$ for time lag τ. This relationship can be used to estimate $\mathbf{B}$ using observational estimates of $\mathbf{C}(0)$ and $\mathbf{C}(\tau_0)$ for some training lag τ_0. In such a system, any two states separated by a time interval τ are related as $\mathbf{x}(t+\tau) = \mathbf{G}(\tau)\mathbf{x}(t) + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a random error vector with covariance $\mathbf{E}(\tau) = \mathbf{C}(0) - \mathbf{G}(\tau)\mathbf{C}(0)\mathbf{G}^{\mathrm{T}}(\tau)$. In a forecasting context, $\mathbf{G}(\tau)\mathbf{x}(t)$ represents the "best" forecast, in a least squares sense, of $\mathbf{x}(t+\tau)$ given $\mathbf{x}(t)$, and $\mathbf{E}(\tau)$ represents its expected error covariance. Linear inverse modeling differs from multiple linear regression in that $\mathbf{G}(\tau_0)$ and $\mathbf{B}$ estimated using $\mathbf{C}(\tau_0)$ for some chosen lag τ_0 are then used to estimate $\mathbf{G}(\tau)$ for all other lags τ as $\mathbf{G}(\tau) = \exp(\mathbf{B}\tau)$. It is this property that enables the system to be approximated as a linear stochastically forced system of the form Eq. (1), in which all the predictable interactions among system components are encapsulated in the deterministic system feedback matrix $\mathbf{B}$. Given any $\mathbf{x}(t)$ one can then calculate $\mathbf{G}(\tau)\mathbf{x}(t)$ as the expected (or the "ensemble mean") forecast of $\mathbf{x}(t+\tau)$ for forecast lead time τ, and $\mathbf{E}(\tau)$ as the expected error covariance (or the "ensemble spread") of that forecast. Note that in the formulation described here,

$\mathbf{E}(\tau)$ depends only on the forecast lead time τ and not on the initial condition $\mathbf{x}(t)$. In other words, the expected forecast ensemble spread is assumed to be independent of the system state.

b. Implementation and data

In practice, it is most convenient to construct a LIM in a reduced EOF space of a system's variables. The LIM can range from simple to complicated, depending upon what variables are included in its state vector $\mathbf{x}$, at what vertical levels, and at what EOF truncation. To construct the simple benchmark LIMs used here, we first created a dataset of 7-day running mean anomalies during the winters of 1980/81–2004/05 of tropical OLR from the NOAA interpolated OLR dataset, and of the extratropical 200- and 850-hPa streamfunction from various reanalysis datasets (see Table 2). Then, following Newman et al. (2003) and Winkler et al. (2001), we defined the LIM state vector $\mathbf{x}$ as comprising the principal components (i.e., the amplitude coefficients) of the dominant 7 EOFs of tropical OLR and 30 EOFs of extratropical streamfunction of the 7-day running mean anomaly fields. Our state vector thus has $7 + 30 = 37$ components. We used the zero-lag and time-lag ($\tau_0 = 5$ days) covariance matrices of this state vector during DJF to construct our LIM in the 37-dimensional space by estimating $\mathbf{B} = (1/\tau_0)\ln[\mathbf{C}(\tau_0)\mathbf{C}^{-1}(0)]$. Seven different reanalysis datasets were used for the streamfunction data (see Table 2), yielding seven different LIMs. Forecasts made using each of these LIMs were initialized and validated against the corresponding reanalysis (for streamfunction) and the NOAA interpolated OLR (for OLR) dataset that was used to construct it. All forecasts were cross validated by withholding data for a year when constructing a LIM, making forecasts for that year, and then repeating the process for all years (often referred to as "jackknifing").

c. LIM forecast skill

We first present the forecast skill of the LIMs to demonstrate that despite their vast simplicity, they do provide tough benchmarks for CGCM forecast skill. The average forecast skill of the seven different LIMs, in terms of the average anomaly correlation coefficient (ACC) of predicted and observed anomalies over the domain, is shown in Table 3. The ACC scores are generally consistent across the seven LIMs, although we note in passing that the LIMs derived from the twentieth-century reanalysis (20CR) and NCEP/Climate Forecast System Reanalysis (CFSR) datasets tend to be relatively more and less skillful, respectively. To what extent this is indicative of deficiencies in those datasets is an interesting issue that we will not pursue here. With minor differences, the geographical variations of skill are also consistent for the seven LIMs (not shown). Figure 2 shows the ACC skill averaged over all the LIMs. Clearly, the LIMs have appreciable skill even at week 3. For all three variables, it is also interesting that the regions of relatively high week-2 and week-3 skill tend to be similar. For Ψ_{200}, these regions are in the subtropics, with local anomaly correlations >0.6 at week 2 and >0.5 at week 3. For Ψ_{850}, they are over North America and the tropical eastern Indian and western Pacific Oceans, where the local anomaly correlations are >0.4 for week 2 and >0.3 for week 3. In the tropics, there are two regions of relatively high skill: the Maritime Continent (>0.5 for week 2; >0.4 for week 3) and the central Pacific (>0.7 for week 2; >0.6 for week 3).

TABLE 2. Reanalyses and observations data used to construct and initialize LIMs as well as validate LIM and dynamical model forecasts. Note: National Center for Atmospheric Research (NCAR); 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40); Japanese 25-yr reanalysis (JRA-25).

Dataset	Reference
NCEP–NCAR reanalysis (NCEP R1)	Kalnay et al. (1996)
NCEP–DOE Reanalysis-2 (NCEP R2)	Kanamitsu et al. (2002)
ERA-40	Uppala et al. (2005)
NOAA 20CR	Compo et al. (2011)
NASA MERRA	Rienecker et al. (2008)
JRA-25	Onogi et al. (2007)
NCEP CFSR	Saha et al. (2010)
NOAA interpolated OLR	Liebmann and Smith (1996)

4. Comparison of LIM and coupled model skill

a. "Single member" hindcasts

Figure 3 compares the skill of the week 3 CGCM hindcasts of upper- and lower-tropospheric streamfunction anomalies with the corresponding skill of the LIM derived from the NASA MERRA streamfunction reanalysis and NOAA interpolated OLR. At week 2, the GEOS-5 streamfunction skill is higher than the LIM skill (not shown), but as Fig. 3 makes clear, at week 3 the difference in skill is statistically insignificant except at 850 hPa over the Sudan and the Arabian Sea, where the GEOS-5 ACC skill is about 0.3 higher than the LIM skill. The CFS streamfunction skill is very similar to the LIM skill at both week 2 (not shown) and week 3. Figure 4 compares the CGCM and LIM skill for tropical OLR at weeks 2 and 3. The GEOS-5 skill is higher than the LIM skill at week 2, but at week 3 the skill of both models is similar to the LIM skill. The results in both Figs. 3 and 4 highlight the fact that even these state-of-the-art coupled models are unable to substantially outperform a simple LIM at week 3, and confirm that such

TABLE 3. ACC skill of seven LIMs averaged over the Northern Hemisphere for Ψ_{200}, Ψ_{850}, and the tropics for OLR. The skill of each LIM is determined relative to the data used to construct it.

		Ψ_{200}		Ψ_{850}		OLR	
LIM	No. of years	Week 2	Week 3	Week 2	Week 3	Week 2	Week 3
NOAA 20CR	25	0.47	0.38	0.34	0.27	0.25	0.20
NASA MERRA	29	0.42	0.33	0.29	0.22	0.25	0.19
JRA-25	29	0.45	0.36	0.32	0.25	0.23	0.20
ERA-40	21	0.43	0.35	0.31	0.26	0.24	0.20
NCEP R1	29	0.42	0.33	0.30	0.24	0.25	0.20
NCEP R2	29	0.44	0.34	0.30	0.23	0.26	0.20
NCEP CFSR	29	0.35	0.25	0.29	0.22	0.26	0.22

LIMs continue to provide challenging benchmarks for comprehensive dynamical modeling systems.

Very similar results to those in Figs. 3 and 4 are obtained when the skill of the two CGCMs is compared with that of any of the seven LIMs. This is efficiently depicted using Taylor diagrams (Taylor 2001) in Fig. 5. In each panel, the colored dots represent the streamfunction and OLR forecasts made using the 7 LIMs, and their comparison with the corresponding CGCM forecasts, represented by the ''reference'' dot in the bottom-right portion of the diagram. The radial distance of the colored dots from the origin is a measure of a LIM's root-mean-square (RMS) forecast amplitude normalized by that of the reference forecasts, the angular distance from the horizontal axis is a measure of the average anomaly pattern correlation of the LIM and reference forecasts, and the distance from the colored dot to the reference dot is a measure of the normalized RMS difference between the LIM and reference forecast fields. The black reference dot in the top (bottom) panel represents GEOS-5 (CFS) forecasts. The individual colored dots show the results for each variable (upper- and lower-tropospheric streamfunction, and tropical OLR) for each LIM. The tight clustering of like-colored dots in all four panels shows that very similar results to those in Figs. 3 and 4 are obtained when comparing the CGCMs with any one of the seven LIMs.

b. Ensemble-mean forecasts

The comparisons of the CGCM forecasts with the LIM forecasts above, as well as in Winkler et al. (2001) and Newman et al. (2003), are of deterministic single-member ensemble-mean CGCM forecasts with LIM forecasts. As mentioned earlier, the LIM forecasts are ''expected value'' forecasts, and as such may also be interpreted as ''infinite member'' ensemble-mean forecasts. It would therefore be more appropriate to compare the skill of the LIM forecasts with that of ensemble-mean CGCM forecasts. Unfortunately, the CGCM forecasts were not generated as part of an ensemble forecasting experiment but as single forecasts, and so such ensembles were not available. However, the fact that the GEOS-5 forecasts were initialized from each day in 1980–2005 gave us the opportunity to construct crude 7-member lagged-average forecast ensembles (Dalcher et al. 1988; Hoffman and Kalnay 1983), using 7 forecasts initialized on 7 consecutive days. Specifically, the 7-member ensemble for week 2 was constructed using the day $8 + n$ through $14 + n$ forecasts, and that for week 3 using the day $15 + n$ through $21 + n$ forecasts, initialized on day $-n$ with n ranging from 0 to 6. Such a procedure ensures that there is no ''cheating'': the week-2 (week-3) ensemble consists of only week-2 (week-3) and *older* forecasts. This ensemble construction technique is obviously crude, and its skill is certainly reduced by the inclusion of the older forecasts. Still, the ensemble means of such forecast ensembles have been used in both synoptic and subseasonal predictions and shown to be more skillful than single forecasts (Dalcher et al. 1988; Hoffman and Kalnay 1983; Tracton et al. 1989; Brankovic et al. 1990).

The difference in skill of the 7-member GEOS-5 ensemble-mean forecasts and LIM forecasts for 200- and 850-hPa streamfunction is shown in Fig. 6 for weeks 2 and 3. It is clear that using the ensemble mean from even such a crude ensemble improves the GEOS-5 skill with respect to the LIM skill, as is also evident from comparing the right-hand panels in Fig. 6 with the top panels in Fig. 3.

c. Summary of skill

Figure 7 provides a comprehensive summary of the week-2 and week-3 forecast skill of the LIM, CFS, and GEOS-5 models, as well as that of the GEOS-5 ensemble, for each of the three variables (Ψ_{200}, Ψ_{850}, and OLR) in the format of Taylor diagrams. The skill is measured with respect to the NASA MERRA reanalysis for the streamfunction and the NOAA interpolated OLR for the OLR anomaly fields represented by the reference dot in the bottom-right portion of the diagrams. A particularly attractive feature of Taylor diagrams is that nondimensionalizing all fields by the standard deviation of the corresponding reference field allows results for different variables such as streamfunction and OLR to be shown on the same plot. If the forecast skill were perfect, all the plotted symbols in

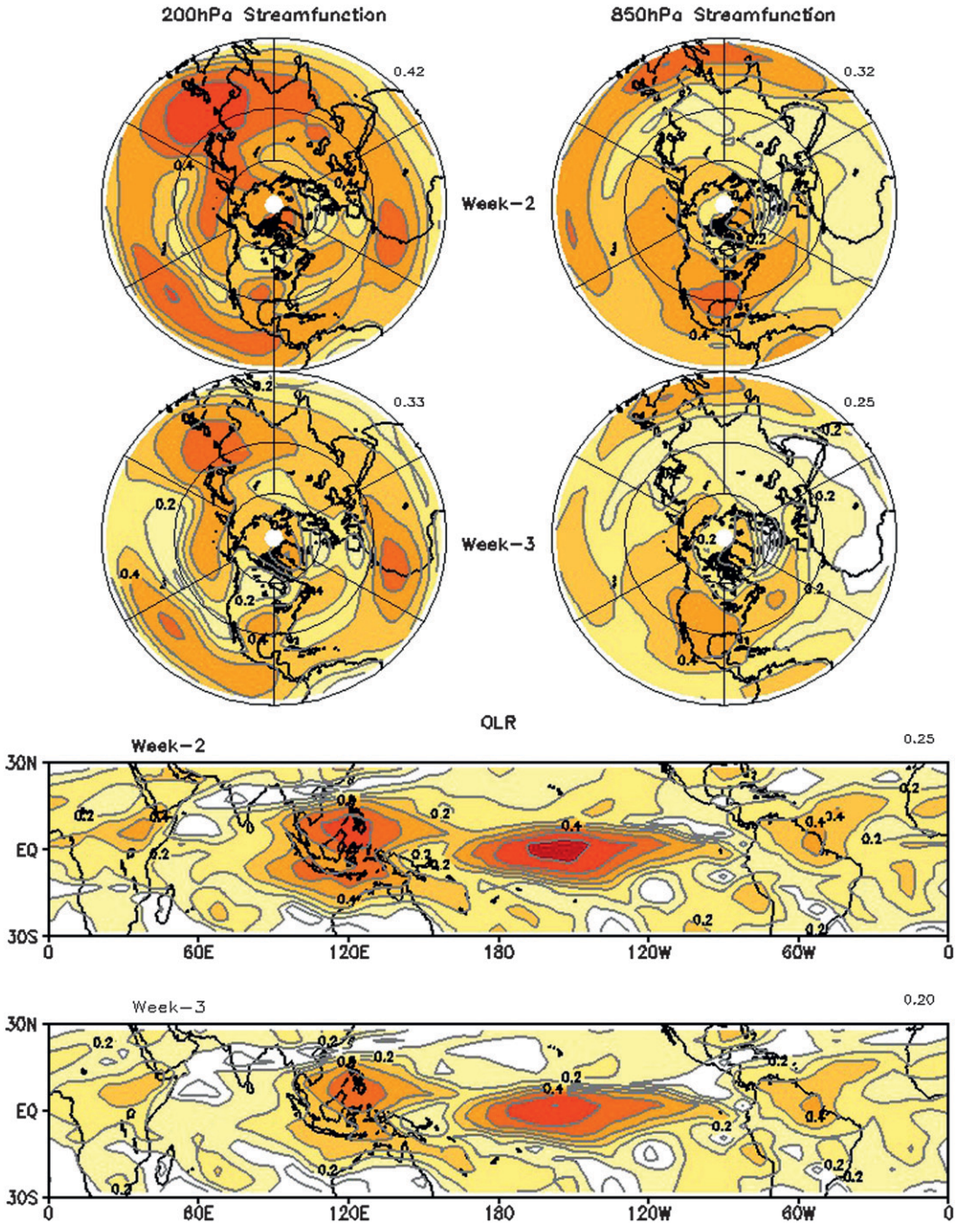

FIG. 2. Average anomaly correlation skill of winter (DJF) LIM forecasts made using seven LIMs derived from seven reanalyses (see Table 2). The average anomaly correlation over the domain is indicated for the top right. Each LIM is initialized by and validated against the reanalysis used to construct it. All forecasts are cross validated. The contour interval is 0.1.

Fig. 7 would gravitate to the reference point. This is obviously not the case, although consistent with the results already presented, the symbols are generally closer to the reference point for week-2 than for week-3 forecasts, and the symbols for the GEOS-5 ensemble forecasts are closer than those for the GEOS-5 single-member and LIM forecasts, especially in week 3. Still, the dominant impression from Fig. 7 is that at week 2 the domain-averaged anomaly correlation skill of the models is at best 0.6 for Ψ_{200}, 0.5 for Ψ_{850}, and 0.35 for OLR, and drops below 0.5

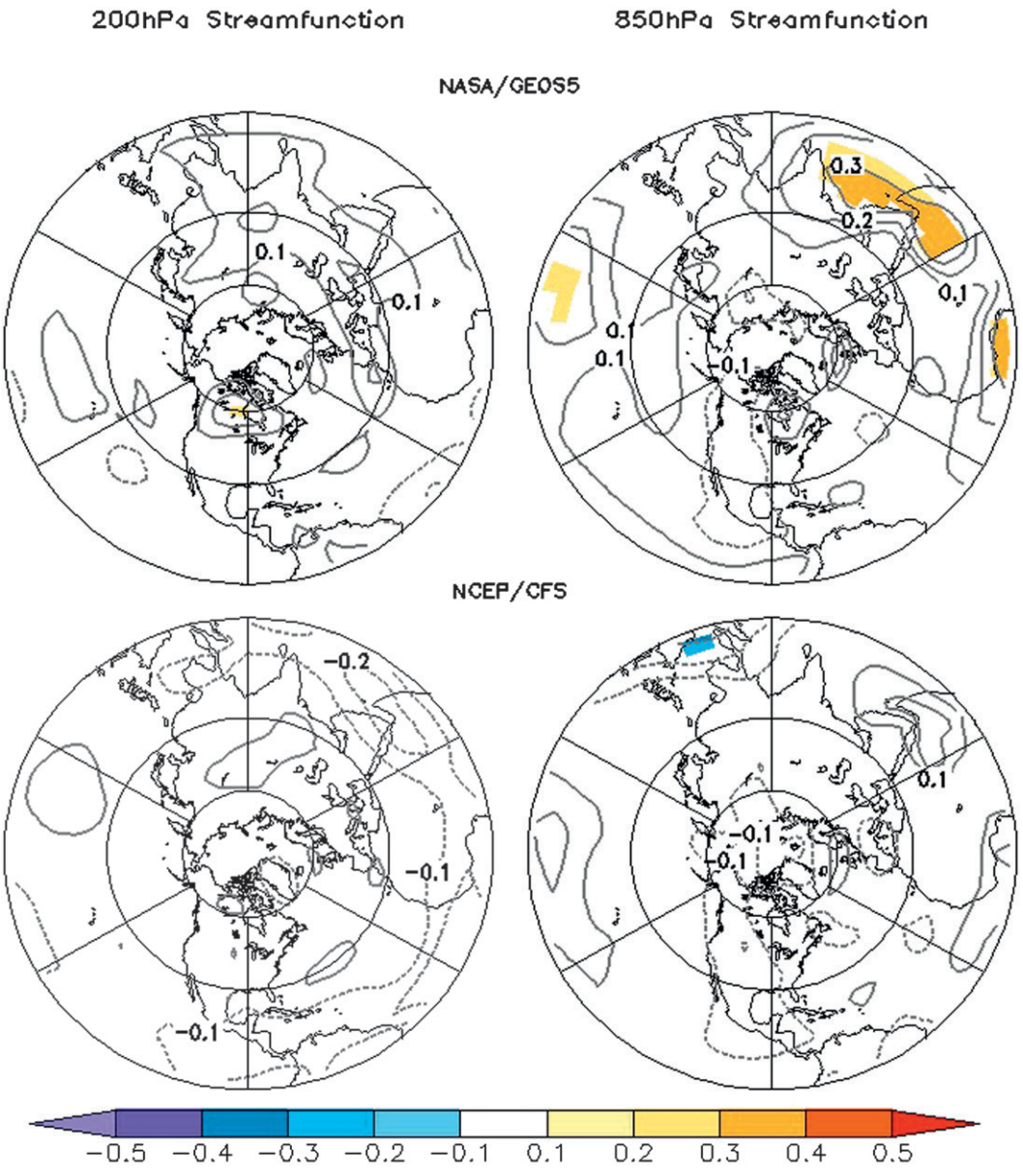

FIG. 3. Difference between local anomaly correlation skill of (top) NASA GEOS-5 and LIM and (bottom) NCEP/CFS and LIM for week-3 forecasts of (left) Ψ_{200} and (right) Ψ_{850}. Differences significant at the 95% level based on a two-tailed test of a normal distribution are shaded.

for all of these variables at week 3. The results of this study thus provide further confirmation that ''useful'' skill (which one may arbitrarily define as anomaly correlations of greater than, say, 0.5) is hard to come by for these extended forecast ranges. This raises a fundamental question concerning the future direction of subseasonal prediction efforts. Is it possible to increase subseasonal forecast skill to useful levels, or is the current modest skill already near the maximum realizable or ''potential'' skill for these forecast ranges? We will address this issue through a predictability analysis in the next section.

Before ending this section, we note one aspect of the LIM forecasts in Fig. 7 that was not apparent in the previous figures. This is that they are of generally weaker amplitude than the CGCM forecasts, and of even weaker amplitude than the observed reference anomaly fields. We stress that this by itself does *not* indicate a deficiency in the LIM forecasts. Recall that

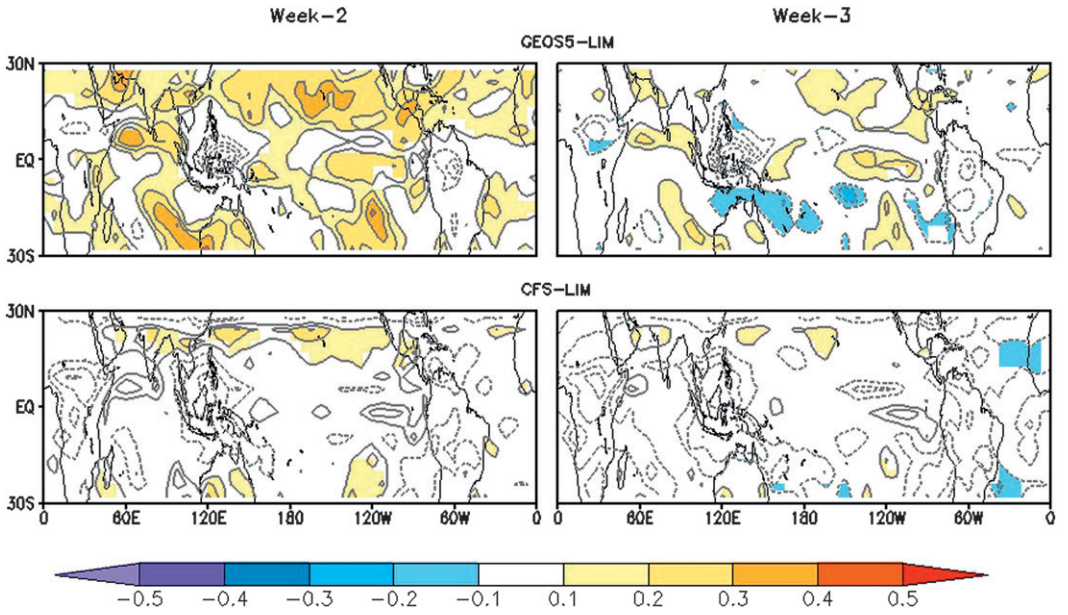

FIG. 4. Difference between anomaly correlation skill of (top) NASA GEOS-5 and LIM and (bottom) NCEP/CFS and LIM for (left) week-2 and (right) week-3 forecasts of OLR. Differences significant at the 95% level based on a two-tailed test of a normal distribution are shaded.

the total anomaly variance at any forecast lead time is the sum of the forecast signal variance and forecast error variance, and the LIM forecast amplitude in Fig. 7 is merely a depiction of the signal standard deviation as a fraction of the total standard deviation, which has to be less than 1. A similar statement can be made concerning the amplitudes of the GEOS-5 ensemble-mean forecasts, which although larger than those of the LIM forecasts, are also weaker than the amplitudes of the observed anomaly fields. It is noted that the standardized root-mean-square error (RMSE) of the forecasts relative to MERRA and NOAA/OLR is also shown in Fig. 7. While the LIM amplitude is much less than GEOS-5, its RMSE is similar to GEOS-5 for both weeks 2 and 3, indicating that the amplitude errors do not have a significant impact on the ability of the LIM to make skillful forecasts. The issue of whether the LIM's forecast amplitudes are more realistic than the ensemble-mean GEOS-5 forecast amplitudes also has an important bearing on estimates of potential skill, as shown in the next section.

5. A predictability analysis

The predictability of a chaotic system may be defined in many different ways. Traditionally, the deterministic predictability limit of weather has been defined as the forecast range at which small errors in the initial conditions grow to be large enough that they are no longer distinguishable from the noise of the system. (Lorenz 1965). The assumption is that the predictable signal comes from the initial conditions and diminishes over time. For seasonal predictions, predictability metrics have been focused on the signal-to-noise ratio (e.g., Compo and Sardeshmukh 2004; Sardeshmukh et al. 2000; Straus et al. 2003), with the assumption that the predictable signal primarily comes from slowly varying boundary conditions. For predictions on intermediate time scales (e.g., intraseasonal, subseasonal), both of these methods have been used to estimate predictability (e.g., Waliser et al. 2003a,b; Pegion and Kirtman 2008; Fu et al. 2007, 2008). For both metrics, the predictability is a measure of whether the signal is large enough to be distinguished from the internal variability or "noise," regardless of the "source" of the signal. Here we define predictability based on the single-to-noise ratio, then translate that into a commonly used metric for assessing prediction skill, so that potential and realized skill can be easily compared. Our definition of predictability is simply the expected anomaly correlation skill of infinite-member ensemble-mean forecasts using a perfect model. We will also refer to this as the potential or maximum realizable skill. As discussed by Sardeshmukh et al. (2000) and Compo and Sardeshmukh (2004), this expected anomaly correlation is a simple function of the forecast signal-to-noise ratio S, defined as

$$S = \frac{\|\text{ensemble mean anomaly}\|}{\|\text{ensemble spread}\|}, \qquad (2)$$

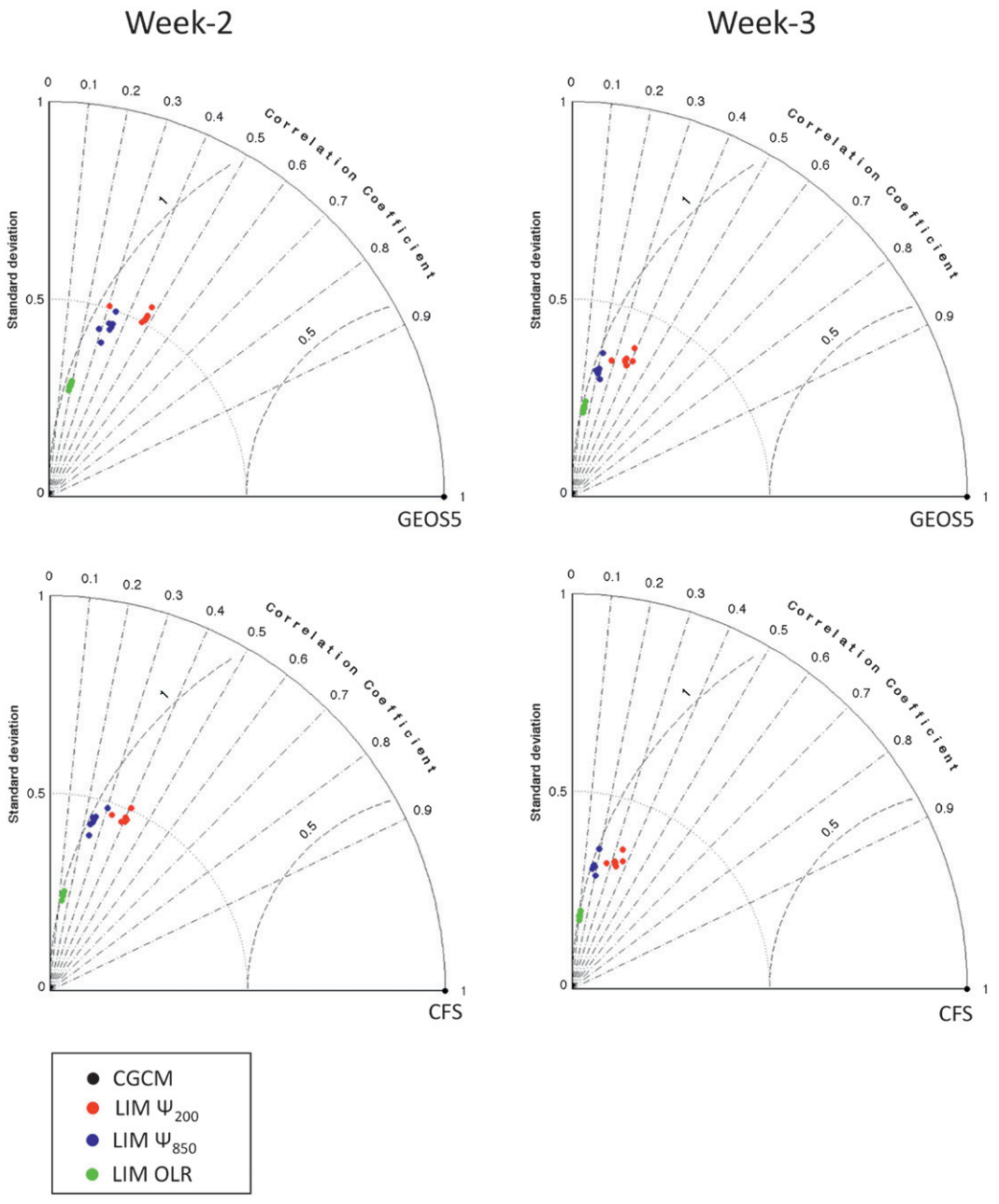

FIG. 5. Taylor diagrams comparing (top) LIM and GEOS-5 forecasts and (bottom) LIM and CFS forecasts for (left) week 2 and (right) week 3. Results are shown for forecasts of Ψ_{200} (red), Ψ_{850} (blue), and OLR (green). The seven different LIMs are shown as filled circles. The reference CGCM forecasts are indicated by the black circles on the horizontal axis.

from which the expected anomaly correlation skill of n-member ensemble-mean forecasts for lead τ can be derived as

$$\rho_n(\tau) = \frac{S^2(\tau)}{\left\{[S^2(\tau)+1]\left[S^2(\tau)+\frac{1}{n}\right]\right\}^{1/2}}. \quad (3)$$

a. Comparison of expected skill estimated using LIM and GEOS-5 ensemble forecasts

For the GEOS-5 ensemble, the signal-to-noise ratio S was calculated explicitly as the ratio of the RMS ensemble mean and ensemble spread values for each variable at each grid point. The potential skill was then calculated by specifying these S values and $n = 7$ in Eq. (3). For the

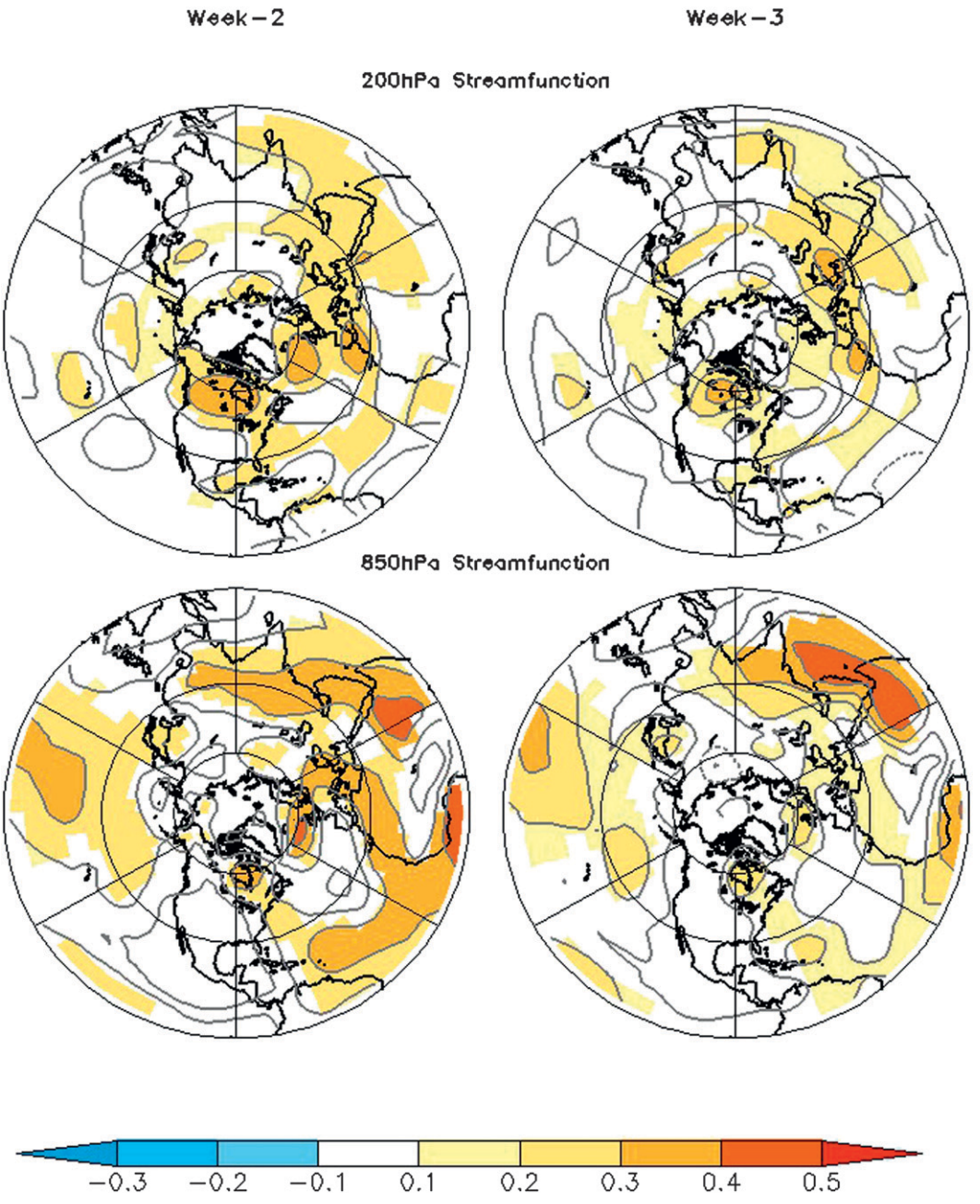

FIG. 6. Difference between anomaly correlation skill of NASA GEOS-5 seven-member ensemble-mean forecasts and LIM forecasts for (left) week 2 and (right) week 3 of (top) Ψ_{200} and (bottom) Ψ_{850}. Differences significant at the 95% level based on a two-tailed test of a normal distribution are shaded.

LIM, the total anomaly covariance matrix of $\mathbf{x}$ for any lead time τ, which is identical to $\mathbf{C}(0)$ in a statistically stationary system, can be partitioned into the forecast signal covariance matrix $\mathbf{G}(\tau)\mathbf{C}(0)\mathbf{G}^{\mathrm{T}}(\tau)$ and the forecast noise (or expected error) covariance matrix $\mathbf{E}(\tau) = \mathbf{C}(0) - \mathbf{G}(\tau)\mathbf{C}(0)\mathbf{G}^{\mathrm{T}}(\tau)$. These matrices were estimated and transformed to grid space, and the forecast signal variance F and noise variance E for each variable at each grid point were identified with the corresponding diagonal elements as

$$F(\tau) = \mathrm{diag}[\mathbf{G}(\tau)\mathbf{C}(0)\mathbf{G}^{\mathrm{T}}(\tau)] \tag{4}$$

and

$$E(\tau) = \mathrm{diag}[\mathbf{C}(0) - \mathbf{G}(\tau)\mathbf{C}(0)\mathbf{G}^{\mathrm{T}}(\tau)], \tag{5}$$

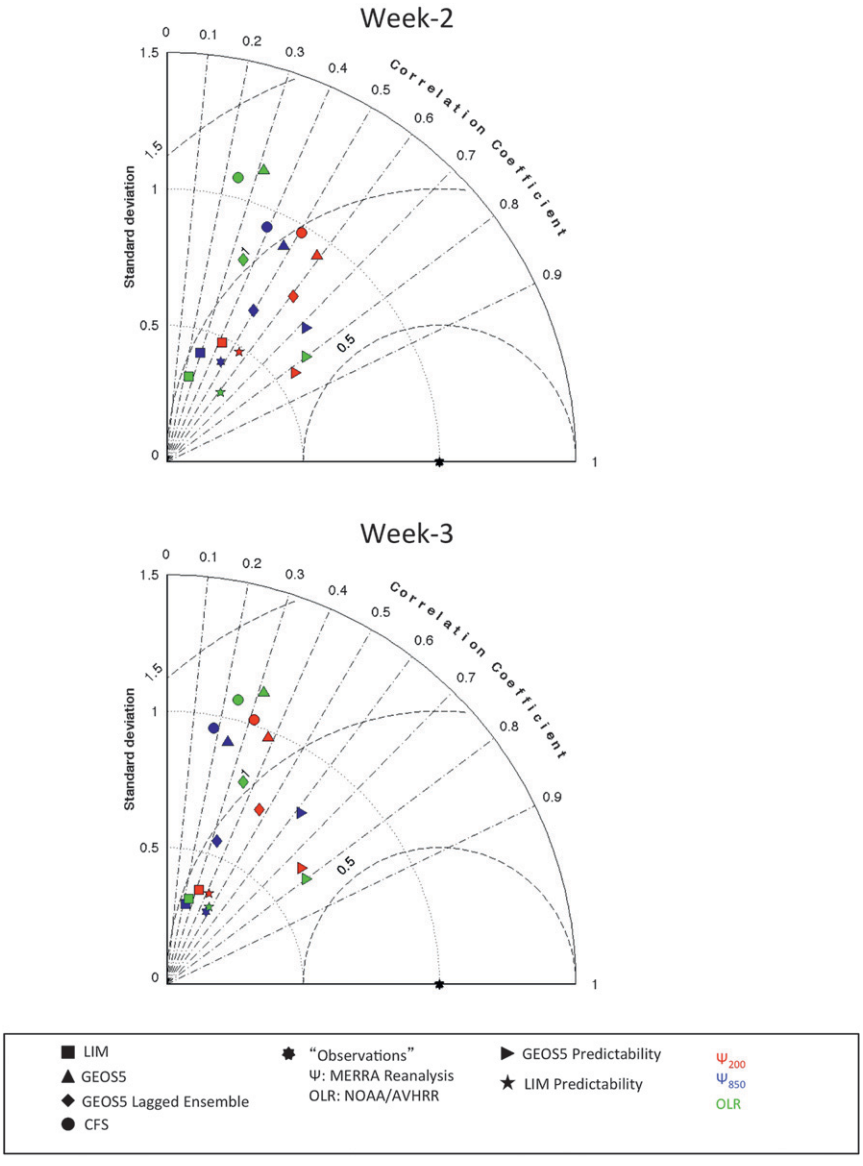

FIG. 7. Taylor diagram summarizing skill of LIM (squares), GEOS-5 (upward pointing triangle), and CFS (circles) for Ψ_{200} (red), Ψ_{850} (blue), and OLR (green). Lagged average ensemble skill is indicated as diamonds. Predictability estimates are indicated by stars (LIM based) and right-pointing triangles (GEOS-5 based). The black asterisk denotes the observations that are from the NASA MERRA reanalysis for streamfunction and the NOAA interpolated OLR.

from which the signal-to-noise ratio S was calculated as the square root of F/E. The expected anomaly correlation skill, assuming a perfect model and infinite-member ensemble, was then calculated by specifying this S and $n = \infty$ in Eq. (3).

Figure 8 compares the potential skill of week 3 Ψ_{200} and Ψ_{850} forecasts estimated using the LIM and the GEOS-5 ensemble. The GEOS-5-based estimates are substantially higher than the LIM-based estimates, by about 0.3–0.4, for both variables. The average potential skill over the Northern Hemisphere is also depicted in the Taylor diagrams in Fig. 7 for the LIM (stars) and the GEOS-5 ensemble (right-pointing triangles), in terms of both the estimated signal-to-noise ratios (along the radial coordinate) and the corresponding expected anomaly correlations (along the angular coordinate). The LIM-based potential skill is relatively low (<0.5) for week 3 and only slightly higher for week 2. On the

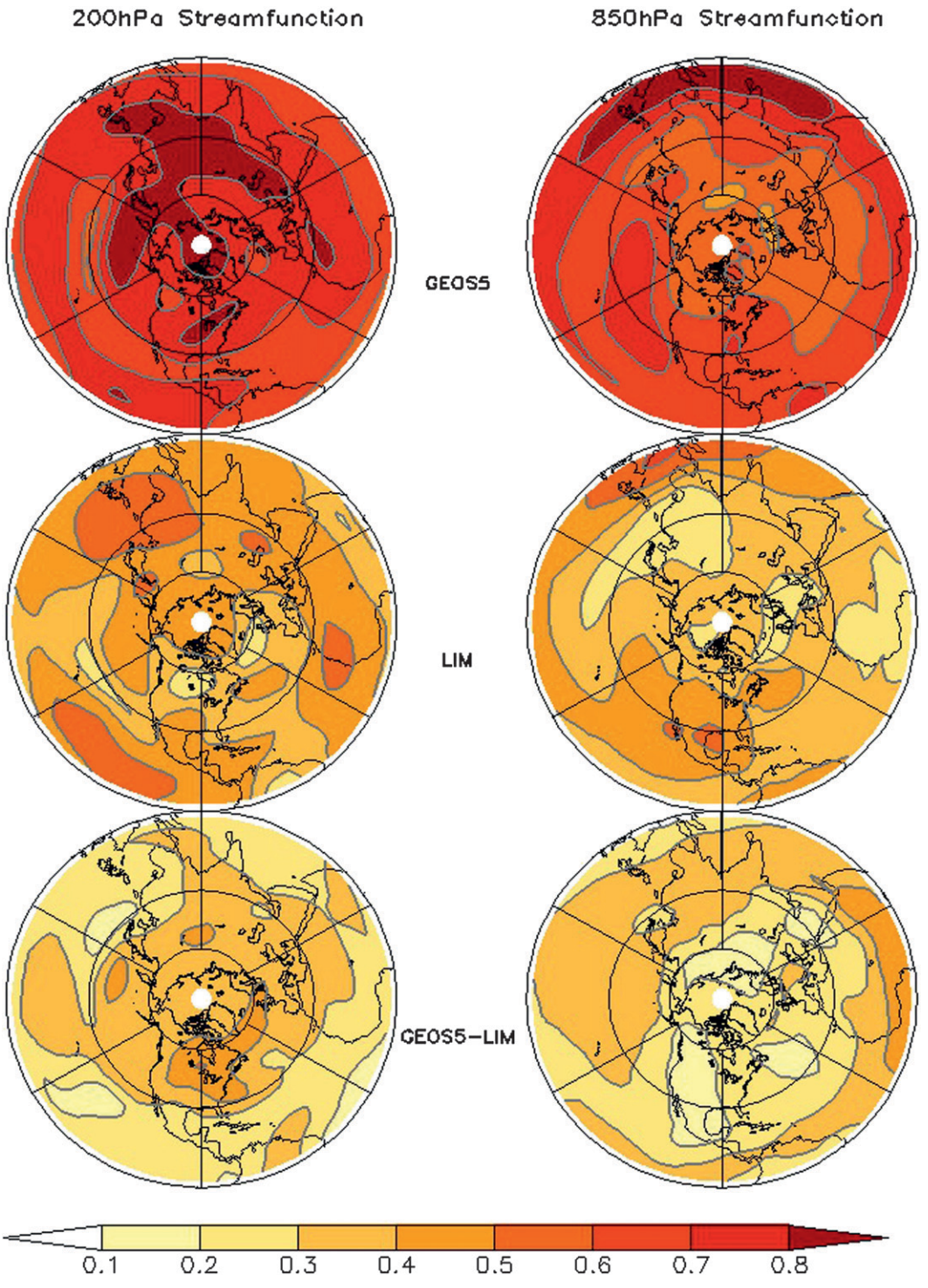

FIG. 8. Potential anomaly correlation skill of week 3 (left) Ψ_{200} and (right) Ψ_{850} forecasts estimated using the (top) GEOS-5 seven-member ensemble and (middle) LIM. (bottom) The difference between GEOS-5-based and LIM-based potential skill estimates is shown.

other hand, the GEOS-5-based potential skill, for all three variables, is >0.7 for week 2 and >0.6 for week 3. The GEOS-5-based estimates thus suggest a potential for substantial improvement even in *average* subseasonal forecast skill to levels that would generally be considered ''useful,'' whereas the LIM-based estimates do not suggest such a potential for improvement. Clarifying which of these two estimates might be more credible has important implications for subseasonal prediction efforts.

Figure 9 provides one pathway to such a clarification. It compares, for week-3 forecasts of 850-hPa streamfunction, the forecast signal variances (top panels), forecast noise variances (middle panels), and forecast error variances (bottom panels) of the LIM and GEOS-5 ensemble forecasts. The noise variances, as noted previously, were estimated using Eq. (5) for the LIM and the spread of the 7-member ensemble for the GEOS-5. The forecast error variances in the bottom panels were determined as variances of the errors of the LIM forecasts and the ensemble-mean GEOS-5 forecasts with respect to the NASA MERRA reanalysis. Figure 9 shows that the GEOS-5 signal variances are much larger, and the noise variances are much smaller, that the corresponding LIM variances. This yields much larger estimates of forecast signal-to-noise ratios using the GEOS-5 ensemble system, and consequently much higher estimates of potential skill in Fig. 8.

Ensemble prediction systems are required to be ''reliable'' in that the verifying analysis should, ideally, always fall within the forecast ensemble. A necessary condition for this is that the forecast ensemble variance (i.e., the noise variance) should match the forecast error variance. Comparing the middle and bottom panels of Fig. 9, it is clear that the GEOS-5 ensemble is seriously deficient in this regard, whereas the LIM is not. This suggests that the GEOS-5 noise variances are unrealistically weak, which contributes to inflating the GEOS-5-based potential skill estimates in Fig. 8. It is also noted that the forecast errors for the LIM and GEOS-5 are not dramatically different, further emphasizing that despite the smaller forecast amplitude by the LIM compared to the GEOS-5 ensemble, its prediction skill is not negatively impacted.

Our lagged average GEOS-5 ensemble is obviously a crude ensemble, so it is not surprising that its spread does not properly represent the forecast error variance. Nonetheless, it is well recognized that the underestimation of ensemble spread is a widespread problem even in sophisticated ensemble prediction systems, which leads to overoptimistic estimates of potential skill. The LIM noise variance in Fig. 9 is much more consistent with the forecast error variance, and to that extent leads to more credible estimates of potential skill in Fig. 8. Note that this consistency in Fig. 9 is not enforced by design, as it would have been if the LIM dynamical operator **B** had been determined using a training lag $\tau_0 = 3$ weeks instead of $\tau_0 = 5$ days.

One can provide evidence that the GEOS-5 ensemble not only underestimates the noise variance in Fig. 9, but also overestimates the signal variance, contributing to a further inflation of the GEOS-5-based potential skill estimates in Fig. 8. Figure 10 compares the time-lag autocorrelation of Ψ_{850} in the GEOS-5 and LIM forecasts with that in the NASA MERRA reanalysis, at time lags of 2 and 3 weeks. Specifically, at each grid point we compute the correlations $\rho_f(\tau)$ of the predicted week-2 and week-3 anomalies with the day-0 anomalies, and compare them with the correlations $\rho_{\text{obs}}(\tau)$ of the corresponding observed week-2 and week-3 anomalies with the day-0 anomalies, using the NASA MERRA reanalysis for this purpose. The figure shows the difference $\rho_f(\tau) - \rho_{\text{obs}}(\tau)$ at $\tau = 2$ weeks and $\tau = 3$ weeks. It suggests that the GEOS-5 forecasts are generally more persistent than the MERRA reanalysis. Note that $\rho_f(\tau)$ in the GEOS-5 forecasts was calculated using single-member GEOS-5 forecasts, rather than the ensemble-mean GEOS-5 forecasts, so the problem in Fig. 10 is not associated with the construction of the GEOS-5 ensemble. This overpersistence of the GEOS-5 forecasts is associated with spuriously large GEOS-5 forecast amplitudes at weeks 2 and 3, which leads to inflated potential skill estimates for weeks 2 and 3 in Fig. 8. The autocorrelations $\rho_f(\tau)$ of the LIM forecasts are more similar to the $\rho_{\text{obs}}(\tau)$ of the MERRA reanalysis, with the exception of subtropical Africa and Asia, where they are smaller than $\rho_{\text{obs}}(\tau)$ indicating a weaker forecast signal in these regions. It is noted that the GEOS-5 is not overpersistent in a region spanning from Africa into parts of Asia and eastern Europe. However, the GEOS-5 ensemble underestimates the noise in this region. Perhaps this is one region where a better ensemble design could lead to both higher estimates of predictability as well as higher realized skill.

To summarize, the GEOS-5 ensemble underestimates the noise and overestimates the signal, leading to unrealistically high potential skill estimates for streamfunction at weeks 2 and 3. Although the underestimation of the noise may be related to the crude construction of our ensemble, the overestimation of the signal clearly arises from the model being too persistent. The LIM noise estimates are more credible because they are more consistent with the forecast error variance. The fact that the LIM lag autocorrelations are also more consistent with those in MERRA reanalysis suggests that the LIM signal estimates are also more credible. This leads us to conclude

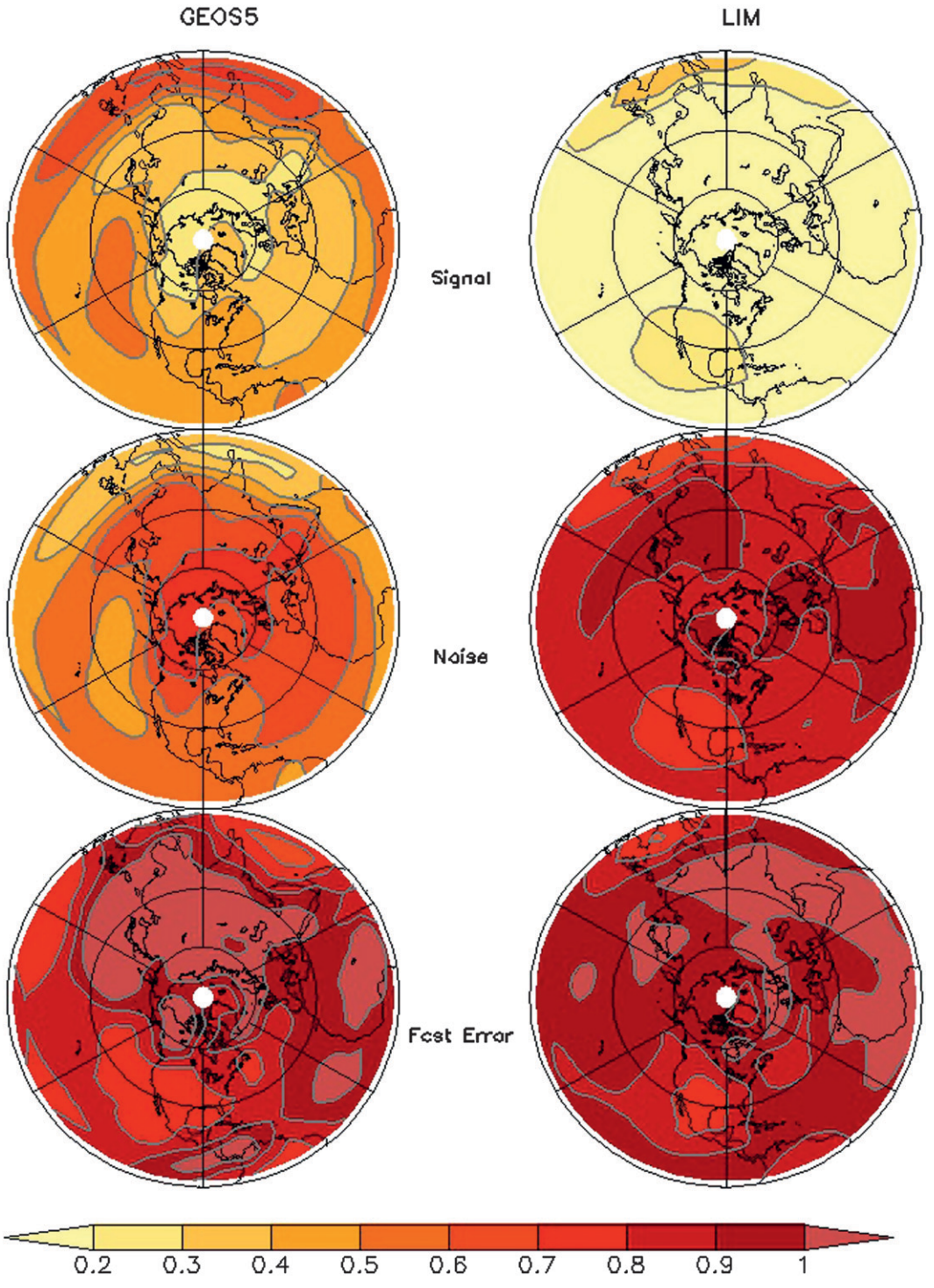

FIG. 9. (top) Forecast signal variance, (middle) forecast noise variance, and (bottom) forecast error variance of week-3 forecasts of Ψ_{850}, estimated from the (left) GEOS-5 seven-member ensemble and (right) LIM. Values are normalized by the observed variance determined from the NASA MERRA reanalyses.

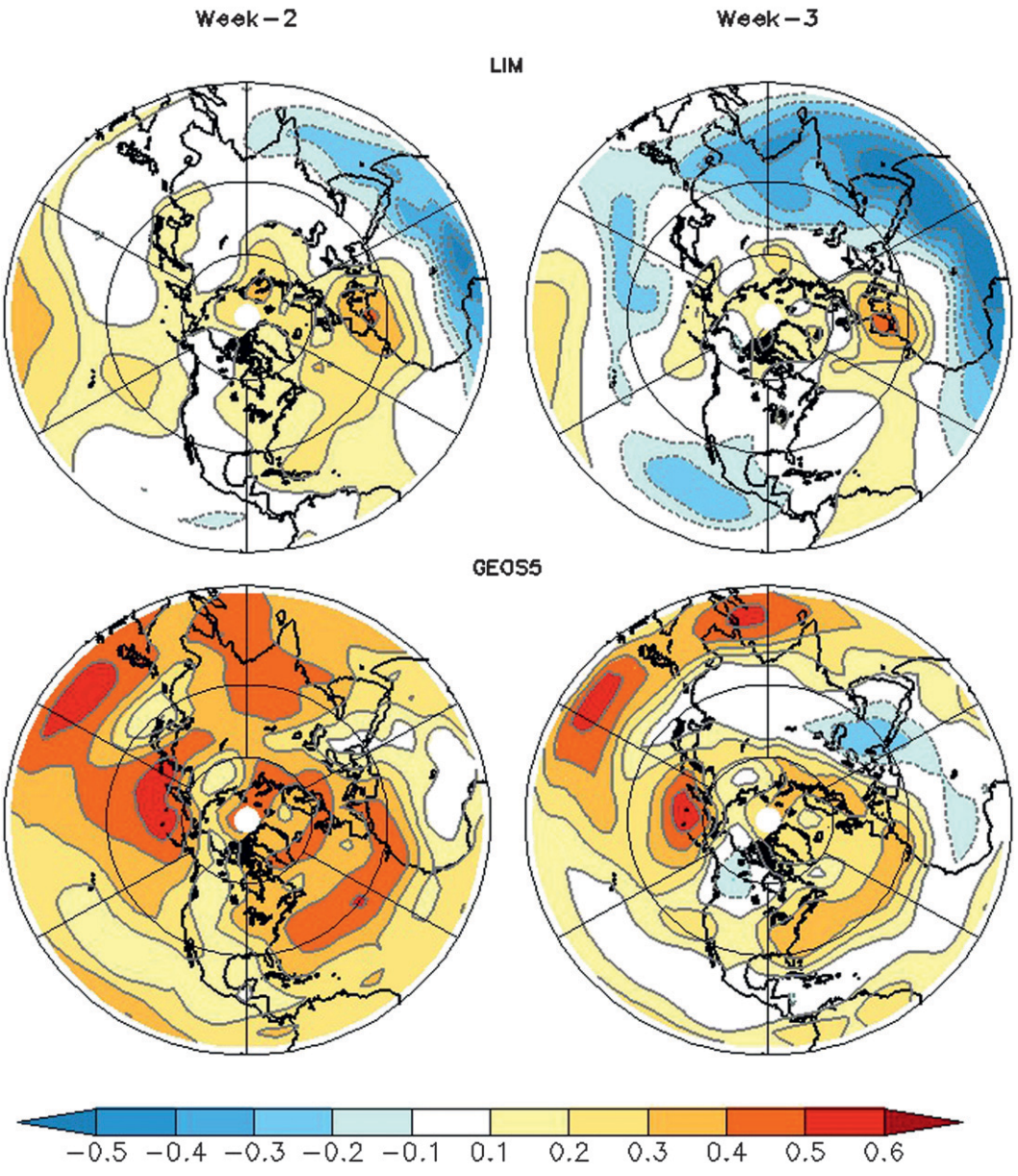

FIG. 10. (top) Difference between the lag autocorrelation of Ψ_{850} in the LIM forecasts and the NASA MERRA reanalyses at (left) week 2 and (right) week 3. (bottom) Difference between the lag autocorrelation of Ψ_{850} in the GEOS-5 forecasts and the NASA MERRA reanalyses at (left) week 2 and (right) week 3.

that the LIM-based potential skill estimates for weeks 2 and 3 in Fig. 8 are more realistic than those obtained using the GEOS-5 ensemble.

b. How much skill is left to be realized?

The potential for subseasonal forecast skill improvement is the difference between the potential and actual skill. We use the more credible LIM-based estimates of potential skill to make this assessment in Fig. 11. The figure shows the difference between these potential skill estimates and the actual skill of the ensemble-mean GEOS-5 forecasts at weeks 2 and 3. For the Northern Hemisphere streamfunction, there is apparently very little skill left to be realized on these time scales. This is both a testament to the skill of the GEOS-5 model as well as a disappointing realization for improving predictions on these time scales in the extratropics. In the tropics, on the other hand, there is apparently still scope for substantial skill improvement (~0.2–0.4), although less so in the equatorial zone (10°N–10°S).

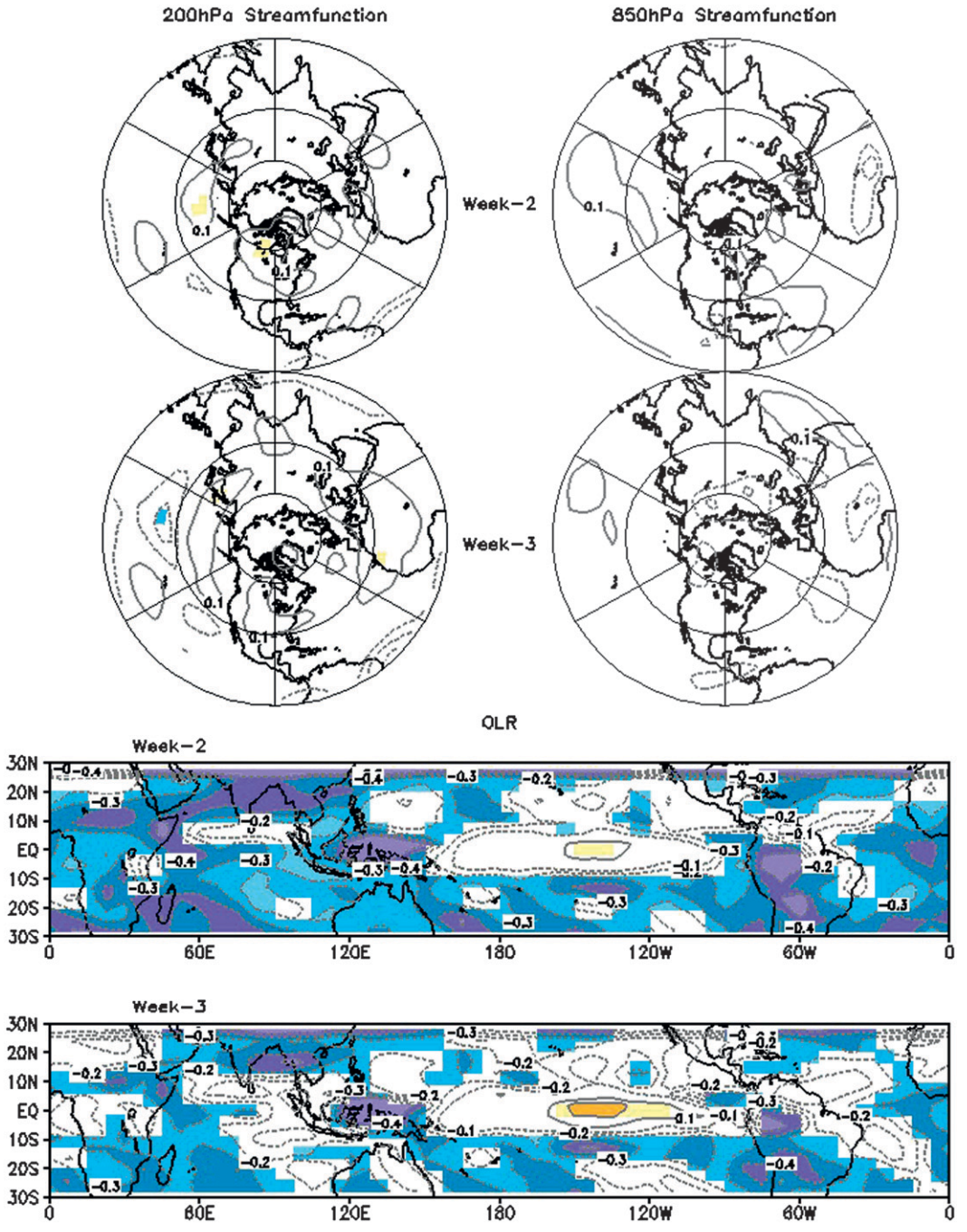

FIG. 11. Difference between the actual anomaly correlation skill of the GEOS-5 ensemble-mean forecasts and the LIM-based potential anomaly correlation skill. Results are shown for (left) Ψ_{200} and (right) Ψ_{850} and OLR for (top) week-2 and (bottom) week-3 forecasts. Differences significant at the 95% level based on a one-tailed test of a normal distribution are shaded.

The conclusion drawn from Fig. 11 of sharply different potentials for improvement in extratropical versus tropical forecast skill might appear paradoxical. How can there be a potential for substantial skill improvement in the tropics, but not in the extratropics, given that the tropics are an important source of predictability for the extratropics on subseasonal scales as demonstrated in many studies (e.g., Winkler et al. 2001; Newman et al. 2003)? One answer to this question is that there is no paradox if the most important tropical ''sources of

extratropical predictability'' lie in regions where the potential for skill improvement is low, such as the equatorial zone in Fig. 11. This issue can be objectively addressed through a singular vector (SV) analysis of the LIM's evolution operator $\mathbf{G}(\tau)$. The dominant SVs represent those tropical diabatic heating (or OLR) anomaly forcing patterns that maximize the extratropical streamfunction response at time τ, and thus objectively identify the dominant tropical sources of extratropical predictability. Winkler et al. (2001) performed just such a singular vector analysis, and their dominant singular vector of $\mathbf{G}(\tau = 14$ days) (shown in their Fig. 13) is entirely consistent with our proposed resolution of the apparent paradox in Fig. 11. Specifically, that singular vector has largest tropical heating amplitude in the equatorial zone, which in Fig. 11 is also the region where the potential for improvement in OLR forecast skill is lowest. We plan to investigate this issue further in a future study.

6. Summary and conclusions

In this paper we compared the subseasonal forecast skill of two state-of-the-art coupled ocean–atmosphere general circulation models (CGCMs) developed at NCEP and NASA with the skill of very simple linear inverse models (LIMs) derived from the observed lag-covariance statistics of 7-day running mean extratropical streamfunction (Ψ_{200} and Ψ_{850}) and tropical OLR anomaly fields. We found that LIMs remain a challenging benchmark for single-member CGCM forecasts to beat at week 3 in the extratropics and at both weeks 2 and 3 in the tropics. We also found that the skill of ensemble-mean CGCM forecasts from even a crude seven-member lagged average forecast ensemble is generally higher than the LIM skill in the extratropics, at both week 2 and week 3, indicating that an ensemble forecasting system is needed for CGCMs to outperform LIMs. The significantly better performance of our crude CGCM forecast ensemble suggests that the ensemble construction method need not be sophisticated, although it is not clear how a more sophisticated ensemble would impact the forecast skill.

For all variables and models, we found that the anomaly correlation skill at the week-2 and week-3 forecast ranges remains low (<0.5), which limits the utility of such forecasts. To determine if more useful forecasts are possible, we estimated the potential anomaly correlation skill at these forecast ranges using the CGCMs and LIMs under a ''perfect model'' assumption, and argued that the LIM-based potential skill estimates were more accurate. We showed that the CGCM-based potential estimates were unrealistically high as a result of both overestimation of the forecast signal and underestimation of the noise, whereas both the signal and noise amplitudes estimated in the LIM-framework were more realistic.

To assess the potential for forecast skill improvement at these ranges using CGCMs, we compared the actual week-2 and week-3 forecast skill of the CGCMs with the LIM-based potential skill estimates and found that there is apparently little skill left to be realized *on average* for weekly averaged Northern Hemisphere streamfunction and equatorial OLR anomalies. Taken at face value, our results indicate that there is little potential for ''useful'' skill (arbitrarily defined as >0.5) at subseasonal time scales and the ensemble-mean forecasts from at least one crudely constructed CGCM ensemble is already able to realize this skill. This is disappointing news for improving subseasonal prediction skill in the extratropics, particularly in light of the fact that this refers to the skill of spatially smooth streamfunction anomaly fields and not of sensible weather elements such as near-surface temperature and precipitation, which is likely even smaller.

In closing, we emphasize that the results presented in this paper are for *average* forecast skill at weeks 2 and 3. They suggest that forecasts at these ranges are not, and are not likely to become, particularly skillful *in general*. They do not, however, preclude the existence of a small but significant fraction of forecast cases in which there is a potential for relatively high skill. Such relatively skillful cases can be identified a priori as cases in which the forecast signal is relatively large and/or the forecast noise is relatively small. Newman et al. (2003) showed how such relatively skillful cases could be identified using a LIM, as cases in which the initial condition has a relatively large projection on the dominant singular vectors of the LIM's evolution operator $\mathbf{G}(\tau)$. Given the necessity of generating expensive CGCM forecast ensembles to maximize forecast skill at extended forecast ranges, a strategy of restricting the generation of such ensembles only to potentially useful forecast cases using Newman et al's (2003) methodology might prove attractive. This is a topic of current research.

Acknowledgments. We thank S. Schubert and Y. Chang from NASA/GMAO for providing us with the GEOS-5 reforecast data. Constructive comments by two anonymous reviewers helped to improve this manuscript. This work was partly supported by the National Aeronautics and Space Administration Grant ESE-NN-H-04-Z-YS-008-N.

REFERENCES

Behringer, D. W., 2007: The Global Ocean Data Assimilation System (GODAS) at NCEP. Preprints, *11th Symp. on Integrated Observing and Assimilation Systems for Atmosphere, Oceans, and Land Surface,* San Antonio, TX, Amer. Meteor. Soc., 3.3.

[Available online at http://ams.confex.com/ams/pdfpapers/119541.pdf.]

——, and Y. Xue, 2004: Evaluation of the global ocean data assimilation system at NCEP: The Pacific Ocean. Preprints, *Eighth Symp. on Integrated Observing and Assimilation System for Atmosphere, Ocean, and Land Surface,* Seattle, WA, Amer. Meteor. Soc., 2.3. [Available online at http://ams.confex.com/ams/pdfpapers/70720.pdf.]

Brankovic, C., T. N. Palmer, F. Molteni, S. Tibaldi, and U. Cubasch, 1990: Extended-range predictions with ECMWF models: Time-lagged ensemble forecasting. *Quart. J. Roy. Meteor. Soc.,* **116,** 867–912.

Compo, G. P., and P. D. Sardeshmukh, 2004: Storm track predictability on seasonal and decadal scales. *J. Climate,* **17,** 3701–3720.

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

Dalcher, A., E. Kalnay, and R. N. Hoffman, 1988: Medium-range lagged average forecasts. *Mon. Wea. Rev.,* **116,** 402–416.

Dole, R. M., 2008: Linking weather and climate. *Synoptic-Dynamic Meteorology and Weather Analysis and Forecasting: A Tribute to Fred Sanders,* L. Bosart and H. Bluestein, Eds., Amer. Meteor. Soc., 297–348.

Fu, X., B. Wang, D. E. Waliser, and L. Tao, 2007: Impact of atmosphere–ocean coupling on the predictability of monsoon intraseasonal oscillations. *J. Atmos. Sci.,* **64,** 157–174.

——, B. Yang, Q. Bao, and B. Wang, 2008: Sea surface temperature feedback extends the predictability of the tropical intraseasonal oscillation. *Mon. Wea. Rev.,* **136,** 577–597.

Gottshalck, J., and Coauthors, 2008: Madden–Julian Oscillation forecasting at operational modelling centres. *CLIVAR Exchanges,* No. 47, International CLIVAR Project Office, Southampton, United Kingdom, 18–19.

Griffies, S. M., M. J. Harrison, R. C. Pacanowski, and A. Rosati, 2008: A technical guide to MOM 4. NOAA/Geophysical Fluid Dynamics Laboratory Ocean Group Tech. Rep. 5, 291 pp.

Hoffman, R. N., and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo Forecasting. *Tellus,* **35A,** 100–118.

Kalnay, and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–470.

Kanamitsu, M., and Coauthors, 2002: NCEP–DOE AMIP-II Reanalysis (R-2). *Bull. Amer. Meteor. Soc.,* **83,** 1631–1643.

Liebmann, B., and C. A. Smith, 1996: Description of a complete (interpolated) outgoing longwave radiation dataset. *Bull. Amer. Meteor. Soc.,* **77,** 1275–1277.

Liess, S., D. E. Waliser, and S. D. Schubert, 2005: Predictability studies of the intraseasonal oscillation with the ECHAM5 GCM. *J. Atmos. Sci.,* **62,** 3320–3336.

Lorenz, E. N., 1965: A study of predictability of a 28-variable atmospheric model. *Tellus,* **17,** 321–333.

National Research Council, 2010: *Assessment of Intraseasonal to Interannual Climate Prediction and Predictability*. National Academies Press, 192 pp.

Newman, M., P. D. Sardeshmukh, C. R. Winkler, and J. S. Whitaker, 2003: A study of subseasonal predictability. *Mon. Wea. Rev.,* **131,** 1715–1732.

Onogi, K., and Coauthors, 2007: The JRA-25 Reanalysis. *J. Meteor. Soc. Japan,* **85,** 369–432.

Pacanowski, R. C., and S. M. Griffies, 1998: The MOM3 manual. NOAA/Geophysical Fluid Dynamics Laboratory Ocean Group Tech. Rep. 4, 680 pp.

Pegion, K., and B. P. Kirtman, 2008: The impact of air–sea interactions on the predictability of the tropical intraseasonal oscillation. *J. Climate,* **21,** 5870–5886.

Penland, C., and P. D. Sardeshmukh, 1995: The optimal growth of tropical sea surface temperature anomalies. *J. Climate,* **8,** 1999–2024.

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

Rienecker, M. M., and Coauthors, 2008: The GEOS-5 Data Assimilation System—Documentation of versions 5.0.1, 5.1.0, and 5.2.0. Technical Report Series on Global Modeling and Data Assimilation, Vol. 27, NASA/TM-2008-104606, NASA, 118 pp.

Saha, S., and Coauthors, 2006: The NCEP Climate Forecast System. *J. Climate,* **19,** 3483–3517.

——, and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.,* **91,** 1015–1057.

Sardeshmukh, P. D., G. P. Compo, and C. Penland, 2000: Changes of probability associated with El Niño. *J. Climate,* **13,** 4268–4286.

Straus, D., J. Shukla, D. Paolino, S. Schubert, M. Suarez, P. Pegion, and A. Kumar, 2003: Predictability of the seasonal mean atmospheric circulation during autumn, winter, and spring. *J. Climate,* **16,** 3629–3649.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.,* **106,** 7183–7192.

Tracton, M. S., K. Mo, and W. Chen, 1989: Dynamical Extended Range Forecasting (DERF) at the National Meteorological Center. *Mon. Wea. Rev.,* **117,** 1604–1635.

Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. Soc.,* **131,** 2961–3012, doi:10.1256/qj.04.176.

Waliser, D. E., K. M. Lau, W. Stern, and C. Jones, 2003a: Potential predictability of the Madden–Julian oscillation. *Bull. Amer. Meteor. Soc.,* **84,** 33–50.

——, W. Stern, S. Schubert, and K. M. Lau, 2003b: Dynamic predictability of intraseasonal variability associated with the Asian summer monsoon. *Quart. J. Roy. Meteor. Soc.,* **129,** 2897–2925.

Wang, W., S. Saha, H. Pan, S. Nadiga, and G. White, 2005: Simulation of ENSO in the New NCEP Coupled Forecast System. *Mon. Wea. Rev.,* **133,** 1574–1593.

Winkler, C. R., M. Newman, and P. D. Sardeshmukh, 2001: A linear model of wintertime low-frequency variability. Part I: Formulation and forecast skill. *J. Climate,* **14,** 4474–4494.