Lagged Ensembles, Forecast Configuration, and Seasonal Predictions

MINGYUE CHEN, WANQIU WANG, AND ARUN KUMAR
Climate Prediction Center, National Centers for Environmental Prediction, College Park, Maryland

(Manuscript received 2 July 2012, in final form 29 March 2013)

ABSTRACT

An analysis of lagged ensemble seasonal forecasts from the National Centers for Environmental Prediction (NCEP) Climate Forecast System, version 2 (CFSv2), is presented. The focus of the analysis is on the construction of lagged ensemble forecasts with increasing lead time (thus allowing use of larger ensemble sizes) and its influence on seasonal prediction skill. Predictions of seasonal means of sea surface temperature (SST), 200-hPa height ($z_{200}$), precipitation, and 2-m air temperature ($T_{2m}$) over land are analyzed. Measures of prediction skill include deterministic (anomaly correlation and mean square error) and probabilistic [rank probability skill score (RPSS)]. The results show that for a fixed lead time, and as one would expect, the skill of seasonal forecast improves as the ensemble size increases, while for a fixed ensemble size the forecast skill decreases as the lead time becomes longer. However, when a forecast is based on a lagged ensemble, there exists an optimal lagged ensemble time (OLET) when positive influence of increasing ensemble size and negative influence due to an increasing lead time result in a maximum in seasonal prediction skill. The OLET is shown to depend on the geographical location and variable. For precipitation and $T_{2m}$, OLET is relatively longer and skill gain is larger than that for SST and tropical $z_{200}$. OLET is also dependent on the skill measure with RPSS having the longest OLET. Results of this analysis will be useful in providing guidelines on the design and understanding relative merits for different configuration of seasonal prediction systems.

1. Introduction

Weather prediction enterprise at various operational centers is a well-coordinated effort, with forecast systems initiated at the same time (0000 UTC, 0600 UTC, etc.). This standardization in the timing of the forecast system has facilitated an easy exchange of forecast data among operational centers. The standardization in operational practices for weather prediction has also led to the development of coordinated databases, such as The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE; http://tigge.ecmwf.int), which have proved to be valuable resources for enabling science, connecting operational and research communities, and advancing weather prediction endeavors.

However, for operational long-range predictions, such as seasonal climate predictions, the current operational practices stand in a stark contrast to the coordination in the operational weather prediction community. Specifically, at some operational centers seasonal forecasts are done in a “burst” mode, and a large ensemble of seasonal forecasts is initiated on a particular calendar day. Examples of operational centers that run seasonal prediction systems in burst mode include the European Centre for Medium-Range Weather Forecasts (ECMWF) and the Australian Bureau of Meteorology (BoM), and the ensemble size of such prediction systems is typically 30–50.

In contrast with the seasonal forecasts run in the burst mode, at some other operational centers a much smaller set of ensembles is run on a daily basis, mimicking the setup of the weather prediction paradigm. One such example is at the National Centers for Environmental Prediction (NCEP), where an ensemble of four seasonal forecasts is initialized every day (Wang et al. 2010). In between the two extremes of the burst and the continuous mode, there are also operational seasonal prediction systems that initialize seasonal forecasts on a weekly basis, such as at the Met Office (UKMO).

A complementary characteristic of the seasonal prediction systems run in the burst versus continuous mode is the ensemble size, which tends to be larger for the
systems run in the burst mode (but the forecast initializations are typically farther apart) and is smaller for the seasonal prediction systems run in the continuous mode (but the forecasts are initialized much more often, e.g., on a daily basis). With a smaller ensemble size for seasonal prediction systems run in the continuous mode, construction of real-time forecasts relies on the lagged ensemble technique, wherein forecasts from initial conditions going back in time are pooled together to adequately sample the uncertainty associated with the seasonal means (Kumar and Hoerling 1995, 1997) to generate forecasts for various target seasons. One aspect of prediction systems run in the continuous mode is that lagged ensembles can be constructed with relative ease. A much more frequent initialization for the continuous mode systems also allows for analysis of the dependence of the seasonal forecast on the initial phase of certain modes of weather and climate variability, such as the Madden–Julian oscillation (Wang et al. 2011).

The lagged ensemble technique has two opposing factors that can influence prediction skill of seasonal means: while use of predictions from longer lead times increases the ensemble size and can improve prediction skill and reliability (Kumar and Hoerling 2000), inclusion of predictions from progressively longer lead times can also result in degradation in prediction skill (Weigel et al. 2008; Chen et al. 2010; Kumar et al. 2011). Because of two opposing factors, it is not clear what would be the influence of different choices of lagged ensemble on the prediction skill. For example, is the decay in skill for seasonal means with lead time sharp enough that it offsets the advantage due to the increase in the ensemble size? Or is the influence such that an increase in ensemble size and a longer lead forecast may result in an optimal lagged ensemble time (LET), for which prediction skill is maximized? If that is the case, what is the dependence of the optimal lagged ensemble on the season, variable, and geographical location?

In this paper, an analysis of a lagged seasonal prediction system based on the NCEP Climate Forecast System, version 2 (CFSv2) is carried out. The real-time forecast configuration for the CFSv2 has four seasonal forecasts every day, and real-time predictions are currently constructed based on initial conditions from the last 10 days, a number that has been chosen on an ad hoc basis (http://www.cpc.ncep.noaa.gov/products/CFSv2/CFSv2seasonal.shtml). Real-time CFSv2 predictions are supplemented by an extensive set of hindcasts spanning 1982–2010 and have a set of four integrations spaced every 5 days. This set of hindcasts (described further in section 2) is used to analyze the dependence of seasonal prediction skill on the lagged ensemble, as one includes forecasts from initial conditions from longer lead time while increasing the ensemble size based on which forecast is made. The focus of the analysis is variation in different measures of skill [e.g., anomaly correlation and rank probability skill score (RPSS)] with the lagged ensemble time (LET), and the results are discussed in section 3. A summary and a discussion about the configuration of seasonal forecast systems appear in section 4.

2. Data and analysis procedures

a. Data

The hindcasts analyzed in this study are from NCEP’s CFSv2. CFSv2 is a fully coupled dynamical prediction system. The atmospheric component of the CFSv2 is a 2007 version of the NCEP Global Forecast System (GFS), with a spectral truncation of 126 waves (T126) in the horizontal (equivalent to nearly a 100-km spatial resolution) and 64 layers in the vertical. The oceanic component is the Geophysical Fluid Dynamics Laboratory (GFDL) Modular Ocean Model, version 4 (MOM4). In CFSv2, configuration of the MOM4 has 40 levels in the vertical, a zonal resolution of ½°, and a meridional resolution of ¼° between 10°S and 10°N, gradually increasing through the tropics until becoming fixed at ½° poleward of 30°S and 30°N. CFSv2 forecasts are initialized from the NCEP Climate Forecast System Reanalysis (CFSR; Saha et al. 2010) atmospheric and ocean states. The CFSR is the latest version of the NCEP reanalysis effort and has the same modeling components as the CFSv2. One salient difference in the atmospheric component between CFSR and CFSv2 is that the atmospheric component of the CFSR has a much higher horizontal resolution of T382 (~38 km). The CFSR is a partially coupled data assimilation system in that, although the 6-hourly guess is based on a coupled forecast, the data assimilation for the ocean and atmosphere is performed separately.

For CFSv2 seasonal hindcasts, four forecast runs for nine target months were made every 5 days starting on 1 January, without considering 29 February in leap years. Each forecast run is for a full 9-month integration and allows us to analyze variations in the skill of seasonal mean forecasts from 0-day to 6-month lead times. In this study, hindcasts from 1982 to 2010 are analyzed.

b. Analysis procedures

We examine variations in prediction skill of seasonal means with lead time using the lagged ensembles forecast technique. An increase in LET allows construction of the lagged ensemble forecast from forecasts initialized from all intervening initial conditions and, therefore,
increases the size of the ensemble that goes in the forecast. To understand the skill variation with LET (and different ensemble size), we also study the skill variations for the same lead time but as a function of the ensemble size.

For the same target season, two kinds of ensemble seasonal mean forecasts are analyzed. In the first kind, ensemble forecast is constructed with varying ensemble size from the same initial time. For every 5 days there are four forecast members (as shown in Fig. 1). This allows us to analyze variation in prediction skill with ensemble size varying from one to four, but for the same lead time, and the traditional prediction skill die-off curve with forecast lead time is made.

The second kind of analysis forecast is the lagged ensemble, which includes forecast members from initial conditions from 0-day lead (i.e., the shortest lead time) going back in time to a specific LET (as shown schematically in Fig. 1). As LET increases, the ensemble size also increases. When four members for each initial time within the LET are included, for example, the lagged ensemble at 0-day LET consists of four members at 0-day-lead initial time, the ensemble size for 5-day LET consists of four members each at 0-day and 5-day lead (eight forecast members in total). The ensemble size at 10-day LET consists of four members each at 0-day, 5-day, and 10-day lead (12 members in total), and so on up to 200-day LET that consists of 164 forecast members in total. The LET, and the corresponding ensemble size used in the analysis, are listed in Table 1. From the lagged ensembles seasonal forecasts are thus constructed, and we analyze the prediction skill dependence on the LET.

We first examine how the seasonal prediction skill decreases with increasing forecast lead time for the forecasts with ensemble size varying from one to four. We then investigate the skill variation for the lagged ensemble forecasts to check if there is any improvement in prediction skill when more forecast members from longer lead times are included. After establishing that this is indeed the case, we next focus on the analysis to find 1) the OLET for which prediction skill maximizes and 2) the gain in skill by comparing skills from the lagged ensemble at OLET to that from 0-day lead time.

The Monte Carlo approach is used to estimate the significance level of the forecast skill. In our approach for the Monte Carlo test, skill after randomizing the forecasts is first computed, and this procedure is repeated 1000 times. Significance of “actual skill” is determined based on the fraction of times the actual skill exceeds the skill estimated after randomizing the forecasts. A 95% significance level means that actual skill was better than skill based on randomized forecasts in 95% of instances.

c. Variables analyzed

The variables analyzed include sea surface temperature (SST), precipitation, 2-m temperature (T2m) over land, and 200-hPa height (z200). The rationale for selecting these variables is the following.

1) SST is the most important forcing for the predictable component of the atmospheric interannual variability on seasonal time scale.

<table>
<thead>
<tr>
<th>LET</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>35</th>
<th>40</th>
<th>45</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lagged ensemble size (case 1)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Lagged ensemble size (case 2)</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td>16</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>Lagged ensemble size (case 3)</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>12</td>
<td>15</td>
<td>18</td>
<td>21</td>
<td>24</td>
<td>27</td>
<td>30</td>
</tr>
<tr>
<td>Lagged ensemble size (case 4)</td>
<td>4</td>
<td>8</td>
<td>12</td>
<td>16</td>
<td>20</td>
<td>24</td>
<td>28</td>
<td>32</td>
<td>36</td>
<td>40</td>
</tr>
</tbody>
</table>

The Monte Carlo approach is used to estimate the significance level of the forecast skill. In our approach for the Monte Carlo test, skill after randomizing the forecasts is first computed, and this procedure is repeated 1000 times. Significance of “actual skill” is determined based on the fraction of times the actual skill exceeds the skill estimated after randomizing the forecasts. A 95% significance level means that actual skill was better than skill based on randomized forecasts in 95% of instances.

<table>
<thead>
<tr>
<th>LET</th>
<th>50</th>
<th>55</th>
<th>60</th>
<th>65</th>
<th>70</th>
<th>75</th>
<th>80</th>
<th>85</th>
<th>90</th>
<th>95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lagged ensemble size (case 1)</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>Lagged ensemble size (case 2)</td>
<td>22</td>
<td>24</td>
<td>26</td>
<td>28</td>
<td>30</td>
<td>32</td>
<td>34</td>
<td>36</td>
<td>38</td>
<td>40</td>
</tr>
<tr>
<td>Lagged ensemble size (case 3)</td>
<td>33</td>
<td>36</td>
<td>39</td>
<td>42</td>
<td>45</td>
<td>48</td>
<td>51</td>
<td>54</td>
<td>57</td>
<td>60</td>
</tr>
<tr>
<td>Lagged ensemble size (case 4)</td>
<td>44</td>
<td>48</td>
<td>52</td>
<td>56</td>
<td>60</td>
<td>64</td>
<td>68</td>
<td>72</td>
<td>76</td>
<td>80</td>
</tr>
</tbody>
</table>
2) Precipitation and the associated tropical heating anomalies communicate the impact of SST on the atmosphere, thereby influencing global and regional hydrological cycles; high-frequency systems (e.g., the tropical cyclone variability) are closely associated with the predictability of seasonal mean large-scale circulation.

3) T2m is a variable strongly influenced by small-scale variations in regional orography, but it is also closely related to large-scale circulation.

4) The 200-hPa heights depict global atmospheric teleconnection patterns due to interannual variability related to tropical Pacific SST variability associated with ENSO.

T2m and precipitation are also key variables of societal importance (O’Lenic et al. 2008; Peng et al. 2012). Moreover, z200 is a well-documented variable in the assessment of seasonal climate variability and predictability (Kumar and Hoerling 1998; Kumar et al. 2003; Peng and Kumar 2005; Lau et al. 2008), and over the extratropical latitudes, because of the equivalent barotropic nature of low-frequency climate anomalies, it is also associated with the surface climate variability (Trenberth et al. 1998).

d. Forecast skill measurements

 Measures of forecast skill used in the study include anomaly correlation coefficient (ACC) and mean square error (MSE) for the ensemble mean deterministic forecasts and the RPSS for probabilistic forecasts. These skill measures are summarized below.

1) The ACC verifies the spatial pattern (or time series) agreement between the forecast and observation. ACC is defined as

\[ \text{ACC} = \frac{\text{cov}(F, O)}{\sigma_F \sigma_O}, \]  

where \( \text{cov}(F, O) \) is the covariance between ensemble mean forecast and observed anomalies and \( \sigma_F \) and \( \sigma_O \) are the standard deviations of forecast and observed anomalies, respectively, over a specific time period and/or a spatial domain. ACC quantifies phase similarity between forecast and observation map or a time series. The ACC score ranges from -1.0 to 1.0. Irrespective of the amplitude, if the forecast has a perfect phasing with observations, the score of ACC equals 1.0.

2) The MSE is defined as

\[ \text{MSE} = \sigma_F^2 + \sigma_O^2 - 2 \times \text{cov}(F, O). \]  

MSE, by definition, is always greater than or equal to zero. If the forecast is perfect, the MSE equals zero. For the forecasts for which there is no phase coherency with the observations, that is, the covariance term in Eq. (2) is zero, then the MSE has its maximum value, which equals the sum of forecast and observed variance. For seasonal predictions, MSE can also be used to infer the lower bound for the unpredictable signal in the observed seasonal mean anomalies (Kumar et al. 2007).

3) The RPSS verifies probabilistic forecasts that take into account the information about the spread of individual members within the ensembles. There are several versions of RPSS. In this study, we rely on a commonly used definition of RPSS, as described in Kumar et al. (2001) and Wilks (1995). The calculation of RPSS includes three steps. First, the rank probability score (RPS) for a probabilistic forecast for \( n \) equiprobable forecast categories is calculated as the squared difference between the cumulative probabilities of each forecast and observation pair:

\[ \text{RPS} = \sum_{m=1}^{n} (Y_m - O_m)^2, \]  

where \( Y_m \) and \( O_m \) are the cumulative probabilities for the \( m \)th forecast category for each pair of the forecast members and the observations, respectively, and are defined as

\[ Y_m = \sum_{j=1}^{m} y_j, \quad O_m = \sum_{j=1}^{m} o_j. \]  

In the above expression, \( y_j \) is the probabilistic forecast for the event to fall in category \( j \), and \( o_j = 1 \) if the observation is in category \( j \) and \( o_j = 0 \) if the observation falls into a category \( i \neq j \). In this analysis we divide the probability density function of seasonal means into three equally probable categories—below normal, normal, above normal categories—and, therefore, \( n = 3 \), as is the traditional practice for making seasonal forecasts (Peng et al. 2012). The forecast probabilities for each category are obtained by a simple counting procedure whereby the percentage of forecast members in a category bin equals the forecast probability for that particular bin.

The second step in computing RPSS involves computing the rank probability score for the forecast using climatological probabilities (RPScl). Finally, the RPSS is defined as the ratio of RPS for ensemble forecasts to that for the climatology as the forecast:

\[ \text{RPSS} = 1.0 - \frac{\langle \text{RPS} \rangle}{\langle \text{RPS}_{cl} \rangle}, \]
where \(\langle \ldots \rangle\) denotes the average over all the forecast–observation pairs in the sample. RPSS, therefore, is a measure of the percent change in the RPS of forecast over the RPS based on climatological probabilities as the forecast. Positive values of RPSS indicate that the forecast has better skill than using climatological probabilities.

In the analysis, all skill measures are calculated for 3-month-running seasonal mean anomalies for the target season against observations of SST from Reynolds et al. (2002), precipitation from the Climate Prediction Center (CPC) Climate Anomaly Monitoring System–Outgoing Longwave Radiation Precipitation Index (CAMS-OPI) (Janowiak and Xie 1999), T2m from CPC Global Historical Climatology Network (GHCN) monthly land surface air temperature analysis (Fan and van den Dool 2008), and z200 from CFSR (Saha et al. 2010). The climatology for both the hindcast and observation is based on the time period from 1982 to 2010.

For the sake of brevity, and to illustrate the concept of OLET, all skill comparisons for different lead times and LETs are shown for global averages first, and then in more detail with spatial distributions. Since there are large spatial variations in standard deviation for the selected variables, especially for z200 between tropics and extratropics, we calculated the global averages with consideration of areal weight for each grid. To calculate the global average of ACC and RPSS, we first calculate the global average of the covariance, standard deviation, RPS of forecast, and RPS of climatology, and then obtained the global average for the skill measure.

3. Results

a. Prediction skill of ACC

In this section, we analyze variation of spatially averaged ACC prediction skill for seasonal means as a function of LET. To understand how the ACC changes with increasing LET, we begin with a discussion of ACC skill variation with ensemble size in forecasts. Solid curves in Fig. 2 show the ACC variations as a function of lead time (x axis) and ensemble size varying from one to four (different colored curves). For a fixed ensemble size, each curve is the traditional die-off curve with forecast lead time. The ACC is averaged over the global region (60°S–60°N), and the prediction skill is calculated for all 3-month-running target seasons between 1982 and 2010 for lead times from 0- to 90-day lead. Each solid colored curve (from orange to brown) in Fig. 2 corresponds to the prediction skill with ensemble size increasing from one to four. Note that all the calculations...
in the analysis are based on lead times up to 6 months. Since the skill variation is almost flat after 90-day lead, the results are shown only up to 90-day lead. As one would expect, four prior variables analyzed in this study have different levels of prediction skills because of differences in inherent predictability. The details of physical reasons for differences in the level of prediction skill for different variables will be discussed later in the context of spatial variations in skill.

As seen in Fig. 2, for a constant ensemble size the ACC for all variables is largest at the 0-day lead and declines thereafter because the contribution of initial conditions on the prediction fades away as the forecast lead time increases. T2m over land loses the initial memory the fastest among the four variables, while decay in skill of z200 is the slowest because of its high predictability over the tropical regions, which can be attributed to slowly evolving SST conditions. Relatively slower decay of precipitation prediction skill is also mainly due to the inclusion of tropical regions both over land and ocean, where the interannual variations are also associated with the slow variation of SST.

For a fix lead time, the ACC increases as more ensemble members are included in the forecast. This is consistent with the well-known relationship identified in previous studies (e.g., Kumar and Hoerling 2000) between ACC and ensemble size. The rate of increase in skill, however, decreases with increasing ensemble size, for example, increase in ACC from one member to two is much larger than that from three members to four. Previous studies (Kumar and Hoerling 2000; Kumar et al. 2001) have documented that an ensemble size of 10–20 may be sufficient to ensure average skill close to the expected value of skill, and further increase in ensemble size has a diminishing return in further improving the ACC. Current analysis, based on the maximum of four members of CFSv2 hindcast data, reveals a similar tendency.

Gain in prediction skill with increasing ensemble size is also dependent on the average skill. Figure 3 summarizes ACC gain from one-member to four-member ensemble forecasts for different ranges of average ACC skills (x axis). The ACC gains are the largest at intermediate values of ACC skill and are small at both low and high values of ACC skill for all four variables. This is because the increase in skill with ensemble is a function of signal-to-noise ratio (which in turn is related to the expected value of various skill measures). For example, Kumar and Hoerling (2000) demonstrated that for increasing ensemble size, the largest gain in skill occurs at...
moderate signal-to-noise ratios. When the signal-to-noise ratio is high (and the expected value of prediction skill is high), an increase in ensemble cannot lead to further improvements in skill. Similarly when the signal-to-noise ratio is low (and the expected value of prediction skill is also low), a large increase in ensemble is also of no further advantage. This tendency for small gains in skill at the low and high ends of ACC and larger gains at
moderate levels of ACC leads to a Gaussian-shaped curve, as in Fig. 3, for gain in skill with the ensemble size. To further understand the ACC variations with lead times for different ensemble sizes, the denominator and numerator terms in Eq. (1) are shown separately in Fig. 2 with dashed curves. For the convenience of the comparison, both terms are normalized by the variance of observation (the right-side axis). There are several points to note.

1) The forecast standard deviation goes down with increasing ensemble size. This is expected as uncorrelated variability in the forecasts from different initial conditions gets averaged by the processes of ensemble averaging.

2) The standard deviation of forecast for a fixed ensemble size shows only a slight reduction as lead time increases (note that the standard deviation of observation is constant for target seasons), indicating that the standard deviation of seasonal mean forecasts quickly converges to the model’s climatological value (Peng et al. 2009); departure in the normalized standard deviation from 100 for an ensemble size of one is indicative of bias in the model’s simulation in replicating the amplitude of seasonal variability.

3) Because of the near invariance with lead time of the forecast standard deviation, the rate in reduction of the ACC with lead time is mainly determined by the rate of reduction in covariance between forecast and observation with lead time. For a given lead time, it can be shown theoretically that the covariance between forecast and observation is independent of the ensemble size since the covariance includes only the predictable part that is common between forecast and observations (blue dashed curves in Fig. 2). The gain of ACC with an increase in the ensemble size is therefore due to the reduction of the standard deviation of ensemble forecast when more members are included in the forecast.

The above results show that, as expected, the seasonal mean prediction skill (as measured by the ACC) increases with increasing size of ensemble but decreases with an increase in the lead time. As discussed in section 1, these two opposing factors lead to following questions: 1) Will there be an advantage by including more members from longer lead times (i.e., using the approach of lagged ensemble) on forecast skill? and 2) If the answer is yes, then at what lagged ensemble time will the gain in skill because of increasing ensemble size (but with longer forecast leads) be eventually offset by the decay in forecast skill by the loss of useful information due to increasing lead time? These questions are addressed in the remainder of the paper.

Figure 4 shows the variations in ACC as a function of LET for the ensemble size increasing from one member for each initial time to four members (orange to brown curves). The approach of lagged ensemble allows larger and larger total size of ensemble with increasing LET as all forecasts between the longest lead (i.e., the LET) to shortest lead are used. We can see that there is indeed an OLET at which the ACC reaches its maximum. The ACC increases from its value at 0-day LET to a maximum (marked with a cross sign) as the LET increases, and more members from longer leads are included in the construction of ensemble mean forecast. The ACC then gradually decreases as the LET increases even further. For example, for the lagged ensemble with one (four) member at each initial time, the maximal gain of ACC for the seasonal mean SST is 8% (3%) and the corresponding OLET is 10 days (5 days). The variation in ACC for lagged ensemble and existence of OLET is an indication that the lagged ensemble approach does indeed lead to improvements in skill because of the positive influence of an increase in the ensemble size, which eventually is offset by forecasts with longer lead time being included in the lagged ensemble. Further, variation in ACC for lagged ensemble no longer follows the natural tendency of skill die-off curves that have a monotonic decrease with lead time (e.g., Fig. 2). A point to note is that OLET becomes shorter, and the magnitude of the ACC gain relative to its 0-day-lead value decreases when more members from each initial time

<table>
<thead>
<tr>
<th>Variables</th>
<th>Members each initial time in lagged ensemble</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC OLET (day)</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>ACC gain (%)</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>MSE OLET (day)</td>
<td>20</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>MSE reduction (%)</td>
<td>22</td>
<td>12</td>
<td>8</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>RPSS OLET (day)</td>
<td>64</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>RPSS gain (%)</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td></td>
</tr>
</tbody>
</table>

Unauthenticated | Downloaded 05/29/22 03:14 PM UTC
are included in the lagged ensemble. Extrapolating this observation further, this implies that as larger and larger numbers of forecasts are initiated from the same time, use of lagged ensemble in improving forecast skill will not be a useful approach; for seasonal forecast systems run in the continuous mode, gain in skill using lagged ensemble will only accrue when the size of ensemble run from each initial time is small.

To glean insight into how the ACC for lagged ensemble changes with LET, the ACC of ensemble with four members from each initial time is decomposed into its numerator and denominator terms [in Eq. (1)] and is shown in Fig. 5. The values displayed are normalized by their corresponding values at 0-day lead. Let us consider the ACC variation over the following three periods.

1) From 0-day lead to the OLET (marked with a cross sign), the ACC increases to its maximum because of a faster decrease of the forecast standard deviation term (red curve) than that of the covariance term (blue curve).

2) From the OLET to the LET, the forecast standard deviation (red curve) crosses the covariance (blue curve) again (after 0-day lead). This crossover point is 15/135/110/135 (days) for SST/z200/precipitation/T2m (however, as the skills for LETs shorter than 90 days are displayed for the sake of clarity, the crossover point for SST can only be seen in Fig. 5a). At the crossover point, the ACC decreases during this period from its maximum back to its 0-day lead value, and the gain of the ACC from the inclusion of more ensemble members in the forecast is offset completely by the decline of ACC because of inclusion of forecasts from longer leads.

3) Beyond the crossover point, decrease in the covariance term becomes dominant and the ACC is less than its 0-day lead value, indicating that the loss of useful information from longer lead forecasts is larger than that which could be offset by the improvement in skill from inclusion of more ensemble members.

All four variables analyzed in the study show similar characteristics in the variations of ACC with lagged ensemble time but with different OLETs and different
magnitudes of maximum gain in the ACC relative to its 0-day lead. In general, OLET is longer for precipitation and T2m, that is, more members from longer lead times continue to increase skill, than for SST and z200 (see the second and third rows of Table 2). With the use of four members at each initial time lagged ensemble, OLETs are 20, 15, 10, and 5 days for precipitation, T2m, z200, and SST, respectively. The percentages of ACC gain relative to its 0-day lead value are higher for precipitation and T2m than for SST and z200 and are 10%, 10%, 6%, and 3% for precipitation, T2m, z200, and SST, respectively, for the case of lagged ensemble with four members from each initial time. Table 2 lists the OLET and percentages of ACC gain relative to its 0-day lead value for all variables and for ensemble size increasing from one to four at each initial time. A relatively larger ACC gain and a longer OLET for z200 compared to that for SST is due to the inclusion of the extratropical regions where the ACC is low for z200. We will confirm this through the spatial variations in OLET discussed next.

Figure 6 shows spatial distribution of seasonal mean prediction skill ACC for variables at 0-day lead with four ensemble members from each initial time for December–February (DJF) and June–August (JJA). Because of the slow variation of SST (higher persistence) and remote influence of ENSO on the other oceans (Alexander et al. 2002; Deser et al. 2010; Chen et al. 2012), high ACC values (>0.7) for SST cover many oceanic regions, for example, from the tropical central to eastern Pacific, part of the North Pacific, the northern tropical Atlantic, and the tropical eastern and western Indian Ocean, with the largest skill in the equatorial eastern Pacific, where the interannual variability associated with ENSO is largest. The regions of the most skillful forecasts are confined to the equatorial eastern Pacific in DJF than that in JJA. On the other hand, low ACCs are over the regions where SST variability is governed mainly by the higher frequency, and unpredictable atmospheric variability, such as the northwest Pacific in DJF and the North Atlantic and southern Indian Ocean in both seasons (Chen et al. 2012). The SST has the largest spatial coverage of the significant skills, except for a region over the southern Indian Ocean (JJA).
It is well known that the primary source of predictability for seasonal mean atmospheric variability is the slowly varying boundary forcings, SST in particular, and further, this influence has geographic preference (Shukla and Wallace 1983; Brankovic et al. 1994; Kumar and Hoerling 1998; Anderson et al. 1999; Kumar et al. 2003; Phelps et al. 2004; Peng et al. 2009). Thus, relatively high ACCs for the atmospheric variables are mainly confined to the tropical latitudes, where the influence of interannual variability of SST dominates. In response to the interannual variability of tropical SST anomalies, z200 ACC shows the largest values over tropical latitudes and a general decrease beyond the extratropical latitudes. The zonal feature in the z200 skill basically comes from the tropical dynamics. Briefly, the interannual variability of tropical heights is highly constrained by interannual SST variability related to ENSO (see, e.g., Kumar et al. 2003). Although the ENSO SST variability, and associated changes in precipitation and diabetic heating, are mostly localized in the equatorial eastern Pacific, since tropical atmosphere cannot sustain horizontal gradients, the influence of localized heat source anomalies is dispersed throughout the tropical belt via dynamical mechanisms on a fast time scale, for example, involving eastward propagating atmospheric Kelvin waves. Near zonally symmetric influence of ENSO SST variability on z200 seasonal mean heights has been documented in numerous studies (e.g., Trenberth et al. 1998; Kumar and Hoerling 1995) and is the fundamental reason for zonally symmetric features that is prevalent in various skill measures for z200 in tropics. The spatial distribution of z200 skill is similar in DJF and JJA seasons with significant values over the tropical to subtropical areas.

For precipitation, a region with relatively high skill (and statistically significant skill) is confined in the tropical central–eastern Pacific, which is collocated with the region of higher SST prediction skill and larger amplitude of SST interannual variability associated with ENSO. The precipitation ACC, however, is very low over land (<0.3), except over a few tropical land locations that also have been identified to be the regions influenced by SST variability in the tropical eastern Pacific–related ENSO. For T2m over land, the ACC skill is quite low. ACC with value above 0.4 appears only over limited regions, such as northern Canada (DJF),

![Fig. 8. Spatial distribution of OLET at which the ACC reaches its maximum for the ensemble with four members from each initial time for (left) DJF and (right) JJA and for (top) SST, (upper middle) z200, (lower middle) precipitation, and (bottom) T2m. The areas with ACC less than 0.1 at 0-day lead are masked out.](image-url)
eastern Canada (JJA), northern Mexico, the northern part of South America, southern Asia, northern Africa, and the southeast and northwest of Australia.

In summary, four variables analyzed in this study have different levels of prediction skills on the seasonal time scale. As is well known in long-range predictions, precipitation is the hardest variable to predict. SST has the largest skill in the equatorial eastern Pacific because of variability associated with ENSO; z200, because of its strong dependence on ENSO SST variability, has high prediction skill in the tropics, which decreases monotonically toward extratropical latitudes. The spatial dependence in skill of these variables for long-range prediction is a function of signal (mainly due to ENSO SST) and noise due to atmospheric internal variability. These features for seasonal predictions have been extensively studied. For example, because of the inherently fast varying, unpredictable nature of atmospheric variability, the seasonal prediction skill of precipitation and T2m over land is relatively low compared with SST. The spatial distribution of OLET and corresponding gain in ACC is next discussed.

Figure 7 shows the spatial distribution of gain in ACC for lagged ensembles, and Fig. 8 shows the corresponding lagged ensemble when it occurs. The ACC gain is calculated as the percent of increase from its value at 0-day lead and is masked out for those regions with ACC less than 0.1. Distinct regional variations are observed for the ACC gain and OLET. For SST, which has high predictability over large oceanic regions, the ACC gain is less than 5% and is distributed uniformly over all ocean basins except some small regions, such as the eastern tropical and southern Indian Ocean, southern central Pacific, and North Pacific. The corresponding OLET over most oceans is shorter than 5 days, that is, there is no further ACC gain by adding more forecast members from longer leads. However, for the regions where ACC gain is higher than 5%, OLET can be up to 15 days.

For z200, the ACC gain from lagged ensemble is less than 5% over tropical areas for both seasons (the second
row in Fig. 7) where the ACC with 0-day lead is very high to begin with (the second row in Fig. 6). The percentage gain of ACC is larger over extratropics where the ACC for 0-day lead itself is relatively low. The corresponding OLET is less than 5 days over the tropics but about 10 days over most extratropical areas.

The ACC shows relatively large gain (>10%) for precipitation over most of regions except for the central to eastern tropical Pacific, where the high ACC is located (see the third row in Figs. 6 and 7). OLET is generally less than 15 days over most areas of tropics, and the distribution pattern is also relatively uniform. However, OLET over extratropical regions is generally larger than 15 days, although the spatial pattern is fairly noisy. T2m over land shows more than 10% ACC gain over most regions, except for high-latitude North America, central Eurasia, and some parts of southern Africa. Correspondingly, OLET is generally larger over the regions with large ACC gain. Over some regions, such as the southwest United States, northeast South America, and tropical Africa, OLET can be up to 65 days, which indicates the ACC can have a marginal gain by adding forecast members from very long lead times. On the other hand, the actual gain in the ACC is already close to zero beyond about 2 months (not shown).

The general characteristics of relative increase in ACC follows the known relationship between variations in ACC with ensemble size, wherein a larger (smaller) gain in skill occurs when average ACC is smaller (larger) (Kumar and Hoerling 2000; Kumar et al. 2001). Following this, larger relative gain in the ACC with lagged ensemble is realized over the regions where the skill is smaller to begin with and requires larger OLET with larger ensemble sizes. Regions where skill for 0-day-lead forecast is large (e.g., tropical regions for SST and z200) have little gain from an increase in ensemble size, and the OLET is relatively small.

To summarize the seasonal characteristics of the analysis, Fig. 9 shows the average of OLET, ACC at 0-day lead, and at the OLET for precipitation and SST (the two most crucial variables in the seasonal forecast) over two subregions (the Indian Ocean and land area and the central–eastern Pacific Ocean) for four different seasons: DJF, March–May (MAM), JJA, and September–November (SON). The general features for the OLET are the same as discussed thus far—OLET for SST is shorter than for precipitation; because of lower prediction skill for both SST and precipitation over the Indian Ocean compared to the Pacific Ocean, OLET is longer over the Indian Ocean. The figure also shows a gain in skill at OLET compared to that at 0-day lead for precipitation and SST over the Indian Ocean and land areas, but there is only a marginal skill improvement at OLET for the central–eastern Pacific Ocean areas where ACC at 0-day lead is already high. These features can be seen from the spatial distribution in Fig. 7.

FIG. 10. As in Fig. 5, but for the MSE and its related terms: the ensemble forecast variance (red solid curves), the observation variance (red dashed curves), and the covariance between ensemble forecast and observation (blue solid curves). The values shown are normalized by the observation variance.
Another deterministic skill measure used for assessing the performance of lagged ensemble forecast is the MSE. MSE quantifies the squared difference between the ensemble mean forecast against the observation. In general, and without showing the results, for a fixed ensemble size, the MSE of the ensemble mean forecast for all four variables is smallest at 0-day lead and increases thereafter as the forecast lead time increases. Precipitation shows the largest MSE among the four variables while SST has the fastest increase as lead time becomes longer. As the change of forecast variance slows down as lead time gets longer (see Fig. 2), the increase in MSE with lead time is mainly due to the decrease of the covariance between forecast and observation.

For the lagged ensemble forecasts, Fig. 10 shows the three terms on the right side of Eq. (2): the variance of ensemble forecast (red curve), the variance of observation (red dashed line), and two times of covariance between forecast and observation (blue curve). The results are shown for the lagged ensemble with four members from each initial time. Lagged ensemble MSE (black curve) itself is the sum of two red curves minus the blue curve. Since for a fixed target the variance of observation is constant, the change of MSE is due to change in the other two terms. For SST, z200, and T2m, MSE decreases from its 0-day-lead value to its minimum at the OLET because of a faster decrease of the forecast variance than the reduction in the covariance term. Beyond the OLET, MSE starts to increase when the decrease of forecast variance becomes slower than that for the covariance term. OLET becomes shorter, and the corresponding minimum MSE value is also reduced when more members from each initial time are included in the lagged ensemble. The OLET of MSE is generally longer than that of ACC (see Table 2).

For the precipitation, the MSE shows monotonic decrease even up to 6 months LET (the longest LET available for seasonal mean forecast in CFSv2 hindcast) because of a faster decrease of the forecast variance than the covariance for all LETs. In fact, after certain lead days (see 45-day LET in Fig. 10c), MSE almost stays constant because of a comparable rate of decrease in the forecast variance and the covariance.

All four variables show similar characteristics in the variations of MSE with the LET. The OLET is longer for precipitation (even beyond the lead time analyzed).
and T2m than for SST and z200. For the lagged ensemble with four members from each initial time, OLET is 45, 20, and 10 (days) for T2m, z200, and SST, respectively (see Table 2). The percentage of MSE reduction relative to its 0-day-lead value is also larger in precipitation and T2m than that for the other two variables. The percentages of MSE reduction relative to its 0-day-lead value for four variables are listed in Table 2.

Figure 11 shows spatial distribution of MSE of the four-member ensemble at 0-day-lead time for DJF and JJA, and at grid point MSE is normalized by the observed variance. MSE is less than 100% of its observed variance over most of the oceans for SST, less than 60% over tropics but higher over extratropics for z200, more than 140% over most regions for precipitation, and around 100% for T2m.

For the lagged ensembles that include more members from longer leads, the MSE decreases from 0-day lead to OLET at which the MSE reduction by an increase in ensemble size is offset by the decay in the forecast skill due to longer lead. Figure 12 shows the spatial distribution of maximum MSE reduction from lagged ensembles, and Fig. 13 shows the OLET at which the MSE reaches its minimal value. The MSE change is expressed as the percent of its value at 0-day lead. As can be seen, MSE reduction and OLET for lagged ensembles show different regional variations for four variables. It is interesting to note that the similarity in OLET between ACC and MSE (Figs. 8 and 13), and the spatial correspondence between ACC gain and MSE reduction (Figs. 7 and 12), is quite good. For SST, which has high prediction skill over wide oceanic regions, its MSE reduction is less than 20% over most areas, except for most of the Indian Ocean (DJF) and the northern Indian Ocean and west part of the tropical Pacific (JJA). The corresponding OLET is less than 15 days over most areas but somewhat longer over the tropical Indian Ocean to west part of the tropical Pacific.

For z200, the MSE reduction from lagged ensemble is also less than 20% over tropical regions where the MSE itself is small (see Fig. 11). The reduction of MSE is relatively large over extratropics where the MSE is high to begin with. OLET is generally shorter over the tropical areas than over the extratropical regions.

For precipitation, MSE shows large reduction over wide areas, and the corresponding OLET is beyond the...
lead time analyzed (90 days), particularly over the North Pacific, the Atlantic, northern Europe and Asia regions, and Southern Hemisphere extratropical regions. However, for the central to eastern tropical Pacific where the MSE values are low, MSE shows a small reduction and the corresponding OLET also drops to about 15 days. For T2m, evident MSE reduction is observed over most of the land areas. OLETs are relatively long. For those regions marked by dark red colors in Fig. 13, OLET is longer than 2 months.

c. Prediction skill of RPSS

In addition to skill of ensemble mean as a single deterministic forecast, we also analyzed the variation in skill for probabilistic forecasts based on the RPSS. Because of the sample requirement in RPSS, the following forecast analysis is only shown for the lagged ensembles with four members from each initial time.

Figure 14 shows spatial distribution of RPSS at 0-day lead for different variables in DJF and JJA. As described in section 2c, the positive values of RPSS indicate the forecast has better skill than assigning climatological probabilities as the forecast. For better visualization, RPSSs less than $-0.2$ are masked out in Figs. 14, 15, and 16 and are not included in the calculation of global average (Fig. 17).

As can be seen, RPSS for SST forecast skill is greater than zero over most oceanic regions, except for extratropical southern oceans. It reaches above 0.4 over the central tropical Pacific. The RPSS is even higher for z200 over the tropics but becomes negative over most of the extratropics. The feature of negative value for RPSS for model-based forecasts with small ensembles was noted by Buizza and Palmer (1998), Richardson (2001), and Kumar et al. (2001) and was attributed to the intrinsic unreliability of small ensembles leading to inconsistencies in the formulation of the RPSS (Weigel et al. 2007). Both precipitation and T2m only show positive RPSS over limited tropical regions. Since different skill measures implicitly map onto each other [although the mapping could be nonlinear (Weigel et al. 2008; Kumar 2007; Peng et al. 2012)], it explains the spatial similarity in the patterns displayed in Fig. 14 (RPSS) and Fig. 6 (ACC).

As the LET gets longer and more members are included in the lagged ensembles, all four variables have a gain in RPSS almost everywhere over the globe (Fig. 15), especially over the regions with low RPSS at
0-day lead. The RPSS gains of precipitation, T2m, and extratropical z200 are most prominent, and the corresponding OLET at which the RPSS reaches its maximum is also relatively longer (see Fig. 16). For those regions with OLET longer than 65 days in Fig. 16 (dark red color), the gains of RPSS are very limited as it reaches a saturation beyond 2 months lead, as shown in Fig. 17, which summarizes the variation of globally averaged RPSS with lagged ensemble from 0-day LET to 90-day LET averaged for all seasons together. RPSS of all four variables increases as LET gets longer and more members are included. After the OLET (cross mark), RPSS decreases for SST and z200 but seems to asymptote for precipitation and T2m (the reason for which is not immediately apparent). OLET is longer and the relative gain in RPSS to its 0-day-lead value is larger for precipitation and T2m than for SST and z200 (see Table 2). Further, OLET for RPSS is generally longer than that of ACC and MSE.

4. Summary and discussion

In this paper, skill variations of lagged ensemble forecasts with the lagged ensemble time (LET) were analyzed. The analysis was based on the NCEP seasonal prediction system CFSv2 from 1982 to 2010. The motivation was to better understand the influence of different choices for the lagged ensemble on the prediction skill.

For the ensembles that include forecast members from the same initial time, the results conform to prior expectations, in that seasonal forecast skills of ACC, MSE, and RPSS become better as ensemble size increases and are degraded as the lead time gets longer. The improvement of prediction skill (or gain in ACC and reduction in MSE) slows down with increasing ensemble size, for example, ACC increase (MSE decrease) from one member to two is much larger than that from three members to four members. For increasing lead time, ACC increase (MSE decrease) is due to the decrease of forecast variance. Further, gain in prediction skill with increasing ensemble size is dependent on the inherent predictability of a given variable. The ACC gains are the largest at intermediate values of ACC skill for all four variables for the reasons consistent with the result of Kumar and Hoerling (2000).

For the lagged ensembles that include forecast members from initial conditions at 0-day lead going back in
time to the LET, that is, forecast members from longer lead times, there is an optimal LET (OLET) at which the ACC (MSE) reaches its maximal (minimal) values. For lead time increasing from 0-day lead to the OLET, the decrease of forecast variance because of the inclusion of more members exceeds the decrease of covariance between forecast and observation, which for lagged ensemble results in a gain in ACC (reduction in MSE). As the LET gets longer, the change in forecast variance is gradually offset by the change in covariance between forecast and observation. The OLET and the amount of skill gain relative to its 0-day-lead values show dependence on location, variable, and the ensembles size. The OLET is longer and the gain in skill relative to its 0-day-lead values is more evident for the regions with low prediction skill, for variables with lower predictability (such as precipitation and T2m), and for the ensembles when less members are included from each initial time.

For the lagged ensemble probabilistic forecast, all four variables analyzed have also OLET at which the RPSS reaches its maximum value. Similar to ACC and MSE, OLET and skill gain vary with locations and variables. OLET is relatively longer and the RPSS gain is relatively larger for precipitation and T2m than for SST and z200. Over most regions, RPSS of precipitation and T2m increase all the way to the longest LET analyzed, but the tendency for an increase almost stops beyond around 2 months lead. The OLET for a given variable varies for different skill measures. In general, OLET of RPSS is longer than that of ACC and MSE. For example, for seasonal mean SST and four members each initial time lagged ensemble, OLET is 5, 10, and 20 (days) for ACC, MSE, and RPSS, respectively (also see Table 2).

We should point out that the rate of increase (decrease) in skill with lagged ensemble is a function of intrinsic signal-to-noise for different variables and need not have a consistency across different variables. For example, Kumar and Hoerling (2000) show that for increasing ensemble size, largest increase in skill occurs at moderate signal-to-noise ratio. When signal-to-noise ratio is high (and the expected value of prediction skill is high), increase in ensemble does not lead to much improvement in skill. Similarly when signal-to-noise ratio is low (and the expected value of prediction skill is also low), a large increase in ensemble is of no advantage. So for SST that has high prediction skill, increase in ensemble

![Fig. 15. As in Fig. 7, but for the gain in RPSS. The areas with RPSS less than −0.2 at 0-day lead are masked out.](image-url)
size based on lagged ensemble may not lead to a large gain in skill, while for atmospheric variables in extratropics, where signal-to-noise ratio is low, the skill associated with the tropical Pacific SST signal only emerges above the noise for larger ensembles. This is the fundamental reason that even though, for long-range predictions, the source of skill in prediction in atmospheric variables resides in skillful prediction of SSTs,

![Diagram](image)

**FIG. 16.** As in Fig. 8, but for the OLET in RPSS. The areas with RPSS less than −0.2 at 0-day lead are masked out.

![Diagram](image)

**FIG. 17.** As in Fig. 4, but for RPSS for the lagged ensemble with four members from each initial time.
the characteristics of the behavior of skill with lagged ensemble do not have to follow the same shape as that for SST itself.

Results from this study clearly indicate that when a smaller number of forecast members for seasonal climate prediction are available (as generally is the case for systems run in a continuous mode), the prediction skill can be improved by using lagged ensemble technique. However, the analysis also highlights the complexity in the choice of optimal lagged ensembles that can depend on the variable, geographical region, and the skill measure under consideration, and for this reason the choice of LET needs to be application specific. For the seasonal prediction systems run in the burst mode, the variation in skill with lead time analyzed, that is, results associated with the die-off curves (e.g., Fig. 2), also provide relevant information on sacrifice in skill with the lead time depending on the degree of separation of initial conditions from the target forecast season.

The current format of operational practices for seasonal prediction systems follows divergent strategies with some prediction systems run in the burst mode (with a large ensemble size) while others are run in a continuous mode (with a smaller ensemble size). What is the best strategy, however, is unclear, and the present analysis may provide a general guidance on future choices and a framework for addressing the issue of lagged ensemble in designing configuration of seasonal forecast systems at operational centers. A lagged ensemble, however, is not the only facet that should be part of the discussion about the choice of best configuration for seasonal forecast systems, as issues related to strategies for the generation of initial perturbations will also play a role. The answers will also depend on question such as how often a seasonal prediction system should run and how often the information (and with what lead time) needs to be provided to the users for informing their decision making process.

It would be desirable to include a similar analysis with other forecast systems; however, implementation of the intercomparison study would require exchanging a fair amount of hindcast data from different seasonal forecast systems that use a similar forecast configuration as NCEP and is beyond the scope of this study. We hope to carry out this analysis as part of the datasets available from the U.S. National Multi-Model Ensemble (NMME; http://www.cpc.ncep.noaa.gov/products/NMME/).

Acknowledgments. We thank Drs. Peitao Peng and Bhaskar Jha and three reviewers for their comments that led to improvements in the final version of the manuscript.

REFERENCES


