Extratropical Prediction Skill of the Subseasonal-to-Seasonal (S2S) Prediction Models

Seok-Woo Son1, Hera Kim1, Kanghyun Song1, Sang-Wook Kim1, Patrick Martineau2, Yu-Kyung Hyun1, and Yoonjae Kim3

1School of Earth and Environmental Sciences, Seoul National University, Seoul, South Korea, 2Research Center for Advanced Science and Technology, The University of Tokyo, Tokyo, Japan, 3Earth System Research Division, National Institute of Meteorological Research, Jeju, South Korea

Abstract The deterministic prediction skill of the 10 operational models participating in the subseasonal-to-seasonal (S2S) prediction project is assessed for both the extratropical stratosphere and troposphere. Based on the mean squared skill score of 50- and 500-hPa geopotential height forecasts, the overall prediction skill is on average 16 days in the stratosphere and 9 days in the troposphere. The high-top models with a fully resolved stratosphere typically have a higher prediction skill than the low-top models. Among them, the European Centre for Medium-Range Weather Forecasts model shows the best performance in both hemispheres. The decomposition of model errors reveals that eddy errors are more important than zonal-mean errors in both the stratosphere and troposphere. While the errors in the stratosphere are dominated by planetary-scale eddies, those in the troposphere are equally influenced by planetary- and synoptic-scale eddies. This result indicates that subseasonal-to-seasonal prediction could be improved by better representing planetary-scale wave activities in the model.

Plain Language Summary How well do the operational models predict weather and climate beyond 2 weeks? To answer this question and to better understand the gap between short-term weather forecasts and long-term seasonal forecasts, the international research organizations recently launched the subseasonal-to-seasonal (S2S) prediction project with the participation of over 10 modeling centers across the world. As a first step to improve S2S prediction, this study assesses the prediction skill of S2S models. It is found that while most models reliably predict the stratospheric circulation for 2- to 4-week lead times, they do so only for lead times of about 1 to 2 weeks in the troposphere. Both the stratospheric and tropospheric errors are mainly explained by eddy errors. More specifically, the misrepresentation of planetary-scale waves, with length scales of a few thousand kilometers or larger, is the key factor that determines the prediction limit. This result suggests that S2S prediction could be improved by better representing large-scale wave activities in the model.

1. Introduction

The leading operational forecast centers, such as the European Centre for Medium-Range Weather Forecasts (ECMWF), are conducting extended forecasts ranging from weeks to months. Such forecasts aim to fill the gap between short-term weather forecasts and long-term seasonal forecasts. However, their overall performance and the factors determining the prediction limit are not fully understood.

To better understand the predictability on the subseasonal-to-seasonal (S2S) time scale, the World Weather Research Program and the World Climate Research Program jointly launched the S2S prediction project in 2015 (Vitart et al., 2015). One of the main goals of this project is to identify the prediction skill of the latest operational models and its sources. For the latter, it is known that S2S prediction is sensitive to various surface boundary conditions such as land surface processes, sea ice distribution, and ocean coupling (Saravanan & Chang, 2019; Chevallier et al., 2019). The skill is also sensitive to the atmospheric mean state not only in the troposphere but also in the stratosphere (Boville, 1984; Butler et al., 2019).

As a first step to improving S2S prediction, the present study assesses the overall prediction skill of the operational models that have participated in the S2S prediction project (Vitart et al., 2015) with a focus on the deterministic prediction skill in the extratropics. The skill is evaluated for both the extratropical stratosphere and troposphere, and compared across all S2S models as in Domeisen et al. (2019). However, unlike the previous studies that typically utilize anomaly correlation, the present study uses more quantitative evaluation...
Most importantly, the nature of prediction errors is quantified by decomposing the errors into their zonal-mean and eddy components. The eddy errors are also separated into planetary-scale and synoptic-scale errors, and eddy-amplitude and eddy-phase errors.

2. S2S models

A total of 10 S2S models, available in October of 2018, are analyzed. They are the operational models from the Australian Bureau of Meteorology (BoM), China Meteorological Administration (CMA), Institute of Atmospheric Sciences and Climate of the National Research Council (CNR-ISAC), Météo-France/Centre National de Recherches Météorologiques (CNRM), Environment and Climate Change Canada (ECCC), ECMWF, Japan Meteorological Agency, Korean Meteorological Administration, National Centers for Environmental Prediction (NCEP), and U.K. Met Office as listed in Table 1. The Hydrometeorological Centre of Russia model is excluded in this study because of technical errors in data archiving.

It is important to note that each model has different configuration (Table 1; see also the S2S prediction website, https://confluence.ecmwf.int/display/S2S/Description). Among others, each model has a different vertical resolution and model top. The BoM, CMA, ECCC, and CNR-ISAC models (hereafter low-top models) have a relatively coarse vertical resolution and do not fully resolve the stratosphere, while the other six models (hereafter high-top models) have a well-resolved stratosphere with a model top above the stratopause (Domeisen et al., 2019).

Most models provide reforecast data at 1.5° × 1.5° resolution. The only exception is the BoM model which is archived at 2.5° × 2.5° resolution. For the comparison, BoM data are bilinearly interpolated to a 1.5° × 1.5° resolution. In the vertical, 10 pressure levels from 1,000 to 10 hPa are examined with a particular focus on the 50- and 500-hPa pressure levels.

All available reforecasts for the common reforecast period of 1999–2010 are analyzed. The two exceptions are the CMA and NCEP models. Due to storage issue, daily initializations of these two models are subsampled to six initializations per month (Lim et al., 2019). To best estimate model performance, the prediction skill is calculated with ensemble-mean forecasts. Note that the ensemble size varies substantially among the models, ranging from one (CNR-ISAC model) to 33 members (BoM model). This difference in ensemble size, along with the difference in reforecast frequency, does not allow a fair comparison between the models. As such, the model performance reported in this study should be understood only in a qualitative sense.

The S2S prediction skill is evaluated by comparing geopotential height forecasts to the reference data of the ECMWF Interim Reanalysis (ERA-Interim; Dee et al., 2011). A bias correction is first conducted (e.g., Choi et al., 2016). Although not shown, most models have systematic biases especially in the stratosphere. For instance, the BoM and CMA models have positive zonal-mean (geopotential height) biases in the stratosphere. In contrast, the CNR-ISAC model shows negative biases. The ECCC model has a rather small bias at all pressure levels for the first few weeks. While the NCEP model has positive biases similar to the CMA model, the CNRM, ECMWF, and U.K. Met Office models show negative biases in the stratosphere.

<table>
<thead>
<tr>
<th>Model</th>
<th>Resolution</th>
<th>Reforecast length</th>
<th>Reforecast frequency</th>
<th>Ensemble size</th>
<th>Initialization</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOM</td>
<td>T47 L17</td>
<td>D 1-62</td>
<td>6/month</td>
<td>33</td>
<td>ERA-40</td>
</tr>
<tr>
<td>CMA</td>
<td>T106 L40</td>
<td>D 0-59</td>
<td>6/month</td>
<td>4</td>
<td>NCEP-R1</td>
</tr>
<tr>
<td>CNR-ISAC</td>
<td>0.75° × 0.56° L54</td>
<td>D 0-32</td>
<td>Every 5 days</td>
<td>1</td>
<td>ERA-Interim</td>
</tr>
<tr>
<td>CNRM*</td>
<td>T255 L91</td>
<td>D 0-60</td>
<td>2/month</td>
<td>15</td>
<td>ERA-Interim</td>
</tr>
<tr>
<td>ECCC</td>
<td>0.45° × 0.45° L40</td>
<td>D 1-32</td>
<td>Weekly</td>
<td>4</td>
<td>ERA-Interim</td>
</tr>
<tr>
<td>ECMWF*</td>
<td>T639/319 L91</td>
<td>D 0-46</td>
<td>2/week</td>
<td>11</td>
<td>ERA-Interim</td>
</tr>
<tr>
<td>JMA*</td>
<td>T319 L60</td>
<td>D 1-34</td>
<td>3/month</td>
<td>5</td>
<td>JRA-55</td>
</tr>
<tr>
<td>KMA*</td>
<td>N216 L85</td>
<td>D 0-60</td>
<td>4/month</td>
<td>3</td>
<td>ERA-Interim</td>
</tr>
<tr>
<td>NCEP*</td>
<td>T126 L64</td>
<td>D 0-44</td>
<td>6/month</td>
<td>4</td>
<td>CFSR</td>
</tr>
<tr>
<td>UKMO*</td>
<td>N216 L85</td>
<td>D 0-60</td>
<td>4/month</td>
<td>4</td>
<td>ERA-Interim</td>
</tr>
</tbody>
</table>

Note. Common reforecast period of 1999–2010 is considered. The high-top models, which fully resolve the stratosphere, are denoted with asterisks.
These biases are removed by subtracting the climatology of the ensemble mean at each forecast day from the raw data.

3. Evaluation Metrics

The prediction skill is quantified by computing the mean squared error (MSE) and its skill score, the mean squared skill score (MSSS), for the bias-corrected geopotential height field. At forecast day \( \tau \), the area-weighted MSE is defined as below:

\[
MSE(\tau) = \frac{1}{N_f} \sum_{i=1}^{N_f} \text{MSE}_i(\tau),
\]

\[
\text{MSE}_i(\tau) = \frac{1}{A} \sum_{j=1}^{N_g} \left( Z_{M,i}(j, \tau) - Z_{O,i}(j, \tau) \right)^2 \cos \varphi_j,
\]

where \( Z_{M,i} \) is the bias-corrected geopotential height in the \( i \)th forecast, and \( Z_{O,i} \) is the geopotential in the reference data. The indices \( j \) and \( \tau \) denote the grid point and the forecast or reference day, respectively. The constants \( N_f \) and \( N_g \) indicate the number of reforecasts and the number of grid points from 30° to the pole, respectively (\( N_g = 9,600 \) for a 1.5° × 1.5° resolution). A similar property can be also computed for ERA-Interim:

\[
\text{MSE}_O(\tau) = \frac{1}{N_f} \sum_{i=1}^{N_f} \text{MSE}_O_i(\tau),
\]

\[
\text{MSE}_O_i(\tau) = \frac{1}{A} \sum_{j=1}^{N_g} \left( Z_{O,i}(j, \tau) - Z_{O}^{no}(j, \tau) \right)^2 \cos \varphi_j,
\]

where \( Z_{O}^{no} \) denotes the long-term climatology. Note that the \( \text{MSE}_O(\tau) \) is basically the area-mean variance of the observations, and only the same calendar days of each prediction are considered.

The MSE skill score is then quantified with the MSSS (Murphy, 1988a):

\[
\text{MSSS}(\tau) = 1 - \frac{\text{MSE}(\tau)}{\text{MSE}_O(\tau)}.
\]

The MSSS is 1 for a perfect forecast system, but in reality MSSS(0) is close to 1 and decreases with forecasts days (\( \tau \)). However, this is not always the case. As described in Table 1, four models are initialized with data sets other than ERA-Interim (https://confluence.ecmwf.int/display/S2S/Models). While the Japan Meteorological Agency model is initialized with JRA-55 reanalysis (Kobayashi et al., 2015), the CMA and NCEP models are initialized with NCEP-NCAR reanalysis (NCEP-R1; Kalnay et al., 1996) and NCEP-CFS reanalysis (CFSR; Saha et al., 2014). The BoM model uses the atmosphere and land initialization data that is halfway nudged to ERA-40 (Cottrill et al., 2013). Therefore, MSE(0) in these models is nonzero although close to zero. The low-top models also show nonzero MSE(0) in the stratosphere presumably due to their coarse vertical resolution.

The prediction skill is defined as the maximum lead time (in days) when the MSSS is positive and statistically significant at the 95% confidence level. The Welch’s \( t \) test is used for the significance test as explained in the Appendix A. A large sample size allows to carry a \( t \) test in this study. Note that this approach is slightly different from the previous studies that simply used MSSS = 0 (e.g., Goddard et al., 2013) or significant MSSS = 0 using a bootstrap test (e.g., Choi et al., 2016).

As described in Stan and Straus (2009), the MSE can be decomposed into the zonal-mean ([MSE]) and eddy components (MSE\(^*\)).

\[
\text{MSE}(\tau) = \text{MSE}(\tau) + \text{MSE}^*(\tau).
\]

By normalizing each term by the observed variance, the model errors can be expressed as
τ (τ) = \left( \frac{MSE (τ)}{MSE_0 (τ)} \right) + \left( \frac{MSE^* (τ)}{MSE_0 (τ)} \right)

\text{ER}^* (τ) = \left\{ \begin{array}{ll}
\text{ER}^* \text{amp} (τ) & \text{if } k \leq 3 \\
\text{ER}^* \text{phs} (τ) & \text{if } k > 4
\end{array} \right. 

The eddy errors, \text{ER}^*, can be further decomposed into the eddy amplitude (\text{ER}^* \text{amp}) and phase errors (\text{ER}^* \text{phs}) for all wavenumbers. It is then possible to decompose the eddy errors into different wavenumbers.

\text{ER}^* (τ) = \text{ER}^* \text{amp} (τ) + \text{ER}^* \text{phs} (τ),

\text{ER}^* (τ) = \text{ER}^*_{k \leq 3} (τ) + \text{ER}^*_{k > 4} (τ).

See the Appendix B for mathematical representations. Here we note that this approach differs from the conventional error decomposition that represents the MSSS as a combination of anomaly correlation, conditional and unconditional errors (Murphy, 1988a). By comparing the model errors associated with the zonal-mean flow and those with eddies, we attempt to better quantify the sources of S2S prediction errors.

4. Results

This section first discusses the ECMWF model. Its prediction skill is assessed for 50- (Z50) and 500-hPa geopotential height (Z500) forecasts averaged from 30° to the pole in each hemisphere. As shown below, this model has the best performance among the 10 S2S models in both the stratosphere and troposphere for both hemispheres. The results of other S2S models are only briefly discussed later, and presented in the supporting information.

Figure 1 shows Z50 and Z500 MSSS time series as a function of forecast days for all reforecasts (annual, ANN) and those initialized in the boreal winter (December-January-February, DJF) and summer (June-July-August, JJA). The ECMWF model shows a comparable prediction skill in the two hemispheres with a much higher skill in the stratosphere (25 days; Figure 1a) than in the troposphere (12–13 days; Figure 1d). This result is consistent with the longer time scale of variability of the stratosphere (e.g., Baldwin et al., 2003).
The Northern Hemisphere (NH) prediction skill exhibits a pronounced seasonality (Figures 1b–1c and 1e–1f). In both the stratosphere and troposphere, DJF skills are much higher than JJA skills. For instance, Z50 skill in the NH winter is 28 days which is twice longer than Z50 skill in summer (Figures 1b and 1c). Although less dramatic, a similar result is also found in the troposphere (Figures 1e and 1f). This seasonality is likely caused by the persistence of polar vortex anomalies during weak or strong polar vortex state (e.g., Tripathi et al., 2015) which can influence the surface climate for over a month (e.g., Baldwin & Dunkerton, 2001).

Unlike in the NH, the prediction skill in the Southern Hemisphere (SH) shows a weak winter-summer difference. Even in the stratosphere, the winter (JJA) prediction skill (17 days) is almost identical to the summer (DJF) prediction skill (18 days). This is likely caused by a reduced austral midwinter polar vortex variability. As highlighted by Shiotani and Hirota (1985) and Baldwin et al. (2003), the SH polar vortex has a large variability in austral spring (e.g., October) and the associated downward coupling is also pronounced from austral spring to early summer (Byrne & Shepherd, 2018; Lim et al., 2018, 2019). In this regard, it is noteworthy that the SH stratospheric prediction skill from all reforecasts (25 days in Figure 1a) is larger than the one in DJF and JJA. This implies that the skill is much larger in spring than in other seasons.

The seasonality and the vertical structure of extratropical prediction skill are presented in Figure 2. For each month, the skill is computed by analyzing all reforecasts initialized during that month. It turns out that a large prediction skill in the stratosphere (larger than 20 days) is achieved when the polar vortex variability is largest, e.g., DJF in the NH but October-November-December (OND) in the SH. This is again due to the persistence of weak or strong polar vortex anomalies (see Figure 1 of Baldwin et al., 2003). A higher prediction skill in the stratosphere is also associated with a higher prediction skill in the troposphere (DJF in the NH and OND in the SH). This is consistent with Choi and Son (2019) who showed that the tropospheric prediction skill is sensitive to stratospheric initial conditions. Sigmond et al. (2013) and Tripathi et al. (2015) also showed that surface prediction skill is extended when models are initialized during weak or strong polar vortex state in comparison to normal condition.

The main characteristics of the ECMWF model skills, i.e., a higher prediction skill in the stratosphere than in the troposphere and a pronounced winter-summer seasonality in the NH, are robustly found in all high-top models (Tables 2 and 3). Although not shown, essentially the same results are also found when the anomaly correlation coefficient is computed (Domeisen et al., 2019). There are a few noticeable differences among the models in Tables 2 and 3. They are discussed later.

To better understand the nature of the model prediction errors in the NH, the vertical structure of the MSSS is examined in Figures 3a–3c. It is evident that the prediction skill is almost constant in the troposphere but sharply increases in the stratosphere (Figure 3a). This is particularly true in DJF (Figure 3b). In JJA, the skill only slightly increases with height as the prediction errors rapidly increase in the stratosphere (Figure 3c). Figures 3d–3f further illustrate the normalized MSE, i.e., the ER in section 3. By definition, the sum of the first two rows in Figure 3 is equal to one. The ER rapidly increases in the summer stratosphere (Figure 3f) in contrast to a rather slow increase in the winter stratosphere (Figure 3e). This result, however, does not imply a large absolute error in the summer stratosphere. It is instead caused by a weak observed internal variability. Since the ER is defined as the MSE divided by the observed variance, it becomes large when the observed variance is quite small. As shown in Figures 4a–4c, the MSE itself is smaller in the summer stratosphere than in the winter stratosphere. But, the observed variance in the summer stratosphere is even much smaller than the one in the winter stratosphere (Figures 4j–4l). This causes a relatively larger ER in...
Table 2
Prediction Skills in Days or in Pentads (in Parentheses) in the Northern Extratropics (30°–90°N)

<table>
<thead>
<tr>
<th></th>
<th>ANN</th>
<th>DJF</th>
<th>JJA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50 hPa</td>
<td>500 hPa</td>
<td>50 hPa</td>
</tr>
<tr>
<td>BoM</td>
<td>11 (2)</td>
<td>9 (2)</td>
<td>15 (3)</td>
</tr>
<tr>
<td>CMA</td>
<td>11 (2)</td>
<td>6 (2)</td>
<td>12 (3)</td>
</tr>
<tr>
<td>CNRM-ISAC</td>
<td>11 (2)</td>
<td>7 (2)</td>
<td>13 (3)</td>
</tr>
<tr>
<td>CNRM*</td>
<td>19 (3)</td>
<td>9 (2)</td>
<td>22 (4)</td>
</tr>
<tr>
<td>ECMWF*</td>
<td>18 (4)</td>
<td>10 (2)</td>
<td>22 (4)</td>
</tr>
<tr>
<td>ECMWF</td>
<td>25 (5)</td>
<td>13 (3)</td>
<td>28 (6)</td>
</tr>
<tr>
<td>JMA*</td>
<td>19 (4)</td>
<td>11 (3)</td>
<td>21 (4)</td>
</tr>
<tr>
<td>JMA</td>
<td>16 (3)</td>
<td>9 (2)</td>
<td>20 (4)</td>
</tr>
<tr>
<td>KMA*</td>
<td>16 (3)</td>
<td>10 (2)</td>
<td>21 (4)</td>
</tr>
<tr>
<td>KMA</td>
<td>16 (3)</td>
<td>9 (2)</td>
<td>19 (4)</td>
</tr>
<tr>
<td>ECCC</td>
<td>21.60 ± 9.30 ± 19.20 ± 10.40 ± 9.10 ± 8.20 ± 4.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMM</td>
<td>18.50 ± 10.17 ± 21.83 ± 15.50 ± 11.50 ± 8.83 ± 3.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-Top</td>
<td>3.20</td>
<td>1.46</td>
<td>2.91</td>
</tr>
</tbody>
</table>

Note: The values at 50 and 500 hPa are computed for all, DJF, and JJA initializations. The multimodel mean (MMM) and intermodel standard deviation for all 10 models and a subset including the six high-top models are shown at the bottom.

Table 3
Same as Table 2 but for the Southern Extratropics (30°–90°S)

<table>
<thead>
<tr>
<th></th>
<th>ANN</th>
<th>DJF</th>
<th>JJA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50 hPa</td>
<td>500 hPa</td>
<td>50 hPa</td>
</tr>
<tr>
<td>BoM</td>
<td>10 (2)</td>
<td>9 (2)</td>
<td>7 (1)</td>
</tr>
<tr>
<td>CMA</td>
<td>12 (3)</td>
<td>6 (2)</td>
<td>12 (3)</td>
</tr>
<tr>
<td>CNRM-ISAC</td>
<td>11 (3)</td>
<td>7 (2)</td>
<td>9 (2)</td>
</tr>
<tr>
<td>CNRM*</td>
<td>20 (4)</td>
<td>10 (2)</td>
<td>16 (3)</td>
</tr>
<tr>
<td>ECMWF*</td>
<td>17 (4)</td>
<td>10 (2)</td>
<td>16 (3)</td>
</tr>
<tr>
<td>ECMWF</td>
<td>25 (6)</td>
<td>12 (3)</td>
<td>18 (4)</td>
</tr>
<tr>
<td>JMA*</td>
<td>16 (3)</td>
<td>9 (2)</td>
<td>15 (3)</td>
</tr>
<tr>
<td>JMA</td>
<td>15 (3)</td>
<td>9 (2)</td>
<td>14 (3)</td>
</tr>
<tr>
<td>KMA*</td>
<td>15 (3)</td>
<td>9 (2)</td>
<td>14 (3)</td>
</tr>
<tr>
<td>KMA</td>
<td>15 (3)</td>
<td>9 (2)</td>
<td>14 (3)</td>
</tr>
<tr>
<td>ECCC</td>
<td>15.60 ± 9.00 ± 13.50 ± 8.60 ± 12.40 ± 8.90 ± 4.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMM</td>
<td>17.67 ± 9.67 ± 15.17 ± 9.17 ± 13.83 ± 9.50 ± 3.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-Top</td>
<td>3.73</td>
<td>1.11</td>
<td>1.46</td>
</tr>
</tbody>
</table>

The same analyses are repeated for the SH (Figure 5). The overall evolution of the MSSS (and ER) is quite similar to its counterpart in the NH (e.g., compare NH DJF in Figure 3 to SH JJA in Figure 5). One key difference is a pronounced planetary-scale ER* in the SH stratosphere (Figure 5o) which is much larger than the one in the NH (Figure 3n). This causes a relatively poor prediction skill in the SH-winter stratosphere compared to the NH-winter stratosphere (e.g., Figure 1).

The above analyses are conducted for all other models as summarized in Tables 2 and 3 (see also Figures S1 and S2 for the pressure-time evolution of the MSSS in each model). It turns out that the ECMWF model has the best skill in both the stratosphere and troposphere for all seasons. Not surprisingly, the high-top models show a higher prediction skill than the low-top models. On average, the high-top models have a deterministic prediction skill of 18.50 days in the NH stratosphere and 10.17 days in the troposphere. In comparison, they are 17.67 days and 9.67 days in the SH. An exception is the ECCC model. Although low-top model, this model shows an excellent performance that is equivalent to other high-top models. This may suggest that Z50 prediction skill is not very sensitive to upper stratospheric processes.

It is important to note from Tables 2 and 3 that models with a higher Z50 skill generally have a higher Z500 skill. For all reforecasts (ANN), the correlation between the skills at these two reference levels across all models is 0.88 in the NH and 0.83 in the SH. If only the six high-top models are considered, the correlation becomes 0.87 in the NH and 0.98 in the SH (Figures 6a and 6b). However, this result is based on a small sample and does not necessarily indicate a stratosphere-troposphere connection. In fact, a strong linear relationship appears also in the summer hemisphere when the stratosphere-troposphere coupling is negligible. This result suggests that the overall prediction skills in the troposphere and the stratosphere are likely determined by common factors such as planetary-scale eddy errors.
Figures 6c–6h present the sources of prediction errors at 15–28 forecast days (third to fourth weeks). The square root of ER* components are shown. By definition, the distance from the origin is equal to the total ER*. The relative importance of [ER] versus ER*, planetary- versus synoptic-scale ER*, and ER*phs versus ER*amp are quantified for Z50 and Z500. Most models, except the two low-top models (BoM and CNR...
ISAC models), show quantitatively similar results to the ECMWF model. For instance, Figures 6c and 6d confirm that the eddy errors are more important than the zonal-mean errors in both the stratosphere and troposphere. While ER* is dominated by planetary-scale eddies in the stratosphere, ER* in the troposphere is equally influenced by planetary- and synoptic-scale eddies. The contributions of eddy phase and amplitude errors are also comparable between the two levels.

It is important to note that the relatively small zonal-mean errors in the stratosphere are achieved in part by the bias correction which, as addressed in section 2, reduces the zonal-mean errors in the stratosphere. Figure 7 presents the MSSSs and ERs of the ECMWF model, as in Figures 3a–3l, but without bias correction. The total ER (Figures 7d–7f) substantially increases in the stratosphere, and this can be largely attributed to the zonal-mean errors (Figures 7g–7i). This is particularly true in the summer stratosphere (Figure 7i). The bias correction, however, has a minimal impact on the tropospheric prediction errors. Although not shown, similar results are found in several other models including all low-top models.

5. Discussion

This study evaluates the overall prediction skill of S2S models. Based on the MSSS metric, S2S models have average prediction skills of approximately 16.2 days in the NH stratosphere and 9.3 days in the NH troposphere. These skills are approximately 15.6 days and 9.0 days in the SH. These values increase if only high-top models, which fully resolve the stratosphere, are considered. In the NH, the overall prediction skills of high-top models are approximately 18.5 days in the stratosphere and 10.2 days in the troposphere. In the SH, they are approximately 17.7 and 9.7 days.

The NH stratospheric prediction skill exhibits a pronounced seasonality with the highest skill in winter and the lowest skill in summer. The enhanced winter prediction skill is likely associated with an enhanced

Figure 4. (a–c) Total MSE, (d–f) its zonal mean and (g–i) eddy components of ECMWF NH reforecasts, compared to (j–l) the observed variance.

4. Conclusion

This study has evaluated the overall prediction skill of S2S models and has found that the NH stratospheric prediction skill exhibits a pronounced seasonality with the highest skill in winter and the lowest skill in summer. The enhanced winter prediction skill is likely associated with an enhanced
persistence of stratospheric polar vortex anomalies under weak or strong vortex state (Baldwin et al., 2003). Consistent with previous studies, a higher prediction skill in the stratosphere is associated with a higher prediction skill in the troposphere, especially during the NH winter. In the SH, the large stratospheric prediction skill is found in OND when the stratospheric variability is large. It is again accompanied by an improvement of tropospheric prediction in October and November. Such synchronization in the

Figure 5. Same as Figure 3 but for SH reforecasts.
Figure 6. Multimodel comparison of (a, b) prediction skills, (c, d) zonal-mean and eddy ERs, (e, f) planetary-scale and small-scale ERs, and (g, h) eddy-amplitude and eddy-phase ERs averaged over 15–28 forecast days for all NH and SH reforecasts. The values at 50 and 500 hPa are denoted with circles and squares with filled markers for ECMWF model. In (a) and (b), the linear regressions and correlation coefficients for a subset of six high-top models are indicated.
stratospheric and tropospheric skills may result in part from the downward dynamical coupling (Baldwin et al., 2003; Seviour et al., 2014).

These results are based on daily prediction skills. However, in the literature, S2S prediction skills have been often quantified with pentad or weakly data (e.g., Tripathi et al., 2015). To test the sensitivity of our results to the temporal resolution, all analyses are repeated with pentad means (see Tables 2 and 3). Not surprisingly, the prediction skills based on pentad data are similar to the daily prediction skills. For instance, in the NH stratosphere, the ECMWF model skill in winter (6 pentads; 25–29 days) is much higher than that in summer (3 pentads). Such a difference does not appear in the SH stratosphere (4 pentads for both seasons). In the troposphere, the prediction skill is of 3 pentads regardless of the hemisphere or season.

The decomposition of errors further reveals that S2S prediction is primarily limited by the eddy errors rather than the zonal-mean errors in both the stratosphere and the troposphere. While the stratospheric errors are dominated by planetary-scale eddies, the tropospheric errors are influenced by both planetary- and synoptic-scale eddies. This result indicates that S2S prediction could be improved by better representing planetary-scale wave activities. Note that since wave activities are sensitive to the mean flow, they are indirectly related to the zonal-mean errors (see linear correlations between 50-hPa prediction errors in Figures 6c–6f). Thus, zonal-mean errors should also be addressed to improve S2S predictions.

Among the 10 S2S models examined in the present study, the ECMWF model shows the best performance in both hemispheres for all seasons (Tables 2 and 3). It is unclear why this model outperforms other high-top models. Under idealized conditions, the ensemble errors could be quantified as below (Murphy, 1988b):

$$E_M(\tau) = \frac{M + 1}{2M} E_1(\tau),$$

where $E_M$, $E_1$, $M$ denote the ensemble-mean prediction error, the single-member error, and the ensemble error, respectively.
The zonal wavenumber, $s$, is located within the 5% tail. Here the number of degrees of freedom, $\nu$, is estimated with

$$\nu(\tau) \approx (N_f-1) \frac{s_M^2(\tau) + s_O^2(\tau)}{s_M^2(\tau) + s_O^2(\tau)}.$$  

where $s_M^2(\tau)$ and $s_O^2(\tau)$ denote the variances of $MSE(\tau)$ and $MSE_O(\tau)$, respectively. Note that $\nu$ value becomes very small as the MSE difference (e.g., $MSE(\tau) - MSE_O(\tau)$) approaches zero and $s_M^2(\tau)$ rapidly increases.

The $t$ value is computed for all reforecasts. The significance of the MSSS is then determined when the $t$ value is located within the 5% tail. Here the number of degrees of freedom, $\nu$, is estimated with

$$\nu(\tau) = \frac{s_M^2(\tau) + s_O^2(\tau)}{s_M^2(\tau) + s_O^2(\tau)}.$$  

Appendix B.: Error Decomposition
At a given latitude and pressure level (e.g., 30° and 500 hPa), the geopotential height $z(l, \tau)$ can be expanded into a discrete Fourier series.

$$z(l, \tau) = [z(\tau)] + \sum_{k=-K}^{K} a(k, \tau) \cos(k 2\pi l/N) + \sum_{k=-K}^{K} b(k, \tau) \sin(k 2\pi l/N),$$

$$= [z(\tau)] + \sum_{k=-K}^{K} A(k, \tau) \cos(k 2\pi l/N - \Psi(k, \tau)),$$

where $l$ denotes longitude and $N$ is the number of grid points in the zonal direction ($N = 240$ for a 1.5° resolution). The zonal-mean value, amplitude, and phase for each zonal wavenumber $k$ are represented as below.

$$[z(\tau)] = \frac{1}{N} \sum_{l=1}^{N} z(l, \tau).$$

$$A(k, \tau) = \sqrt{a^2(k, \tau) + b^2(k, \tau)}.$$

$$\Psi(k, \tau) = \arctan \left( \frac{b(k, \tau)}{a(k, \tau)} \right).$$
The mean square error at a given latitude band, $mse(\tau)$, is then defined as in Stan and Straus (2009):

$$mse(\tau) = [z_M(l, \tau) - z_O(l, \tau)]^2 = [mse](\tau) + mse^*(\tau),$$

$$mse^*(\tau) = mse\_amp(\tau) + mse\_phs(\tau),$$

where

$$[mse](\tau) = ([z_M(\tau)] - [z_O(\tau)])^2,$$

$$mse\_amp(\tau) = \frac{1}{2} \sum_{k=1}^{K} (A_M(k, \tau) - A_O(k, \tau))^2,$$

$$mse\_phs(\tau) = \sum_{k=1}^{K} A_M(k, \tau) A_O(k, \tau) \{1 - \cos(\Psi_M(k, \tau) - \Psi_O(k, \tau))\}.$$

The planetary- and synoptic-scale errors are defined by integrating zonal wavenumber 1 to 3 and 4 to higher, respectively. Here it is important to note that the phase error, $mse\_phs(\tau)$, is weighted by the amplitude error. Note that $mse$ differs from MSE in section 3. The latter is an area-weighted latitudinal integration of the former. Note also that Lim et al. (2018) took a similar approach to evaluate S2S prediction of the Madden-Julian oscillation.

References


