MJO Prediction Skill of the Subseasonal-to-Seasonal Prediction Models

YUNA LIM AND SEOK-WOO SON

School of Earth and Environmental Sciences, Seoul National University, Seoul, South Korea

DAEHYUN KIM

Department of Atmospheric Sciences, University of Washington, Seattle, Washington

(Manuscript received 12 August 2017, in final form 20 February 2018)

ABSTRACT

The Madden–Julian oscillation (MJO), the dominant mode of tropical intraseasonal variability, provides a major source of tropical and extratropical predictability on a subseasonal time scale. This study conducts a quantitative evaluation of the MJO prediction skill in state-of-the-art operational models, participating in the subseasonal-to-seasonal (S2S) prediction project. The relationship of MJO prediction skill with model biases in the mean moisture fields and in the longwave cloud–radiation feedbacks is also investigated.

The S2S models exhibit MJO prediction skill out to a range of 12 to 36 days. The MJO prediction skills in the S2S models are affected by both the MJO amplitude and phase errors, with the latter becoming more important at longer forecast lead times. Consistent with previous studies, MJO events with stronger initial MJO amplitude are typically better predicted. It is found that the sensitivity to the initial MJO phase varies notably from model to model.

In most models, a notable dry bias develops within a few days of forecast lead time in the deep tropics, especially across the Maritime Continent. The dry bias weakens the horizontal moisture gradient over the Indian Ocean and western Pacific, likely dampening the organization and propagation of the MJO. Most S2S models also underestimate the longwave cloud–radiation feedbacks in the tropics, which may affect the maintenance of the MJO convective envelope. The models with smaller bias in the mean horizontal moisture gradient and the longwave cloud–radiation feedbacks show higher MJO prediction skills, suggesting that improving those biases would enhance MJO prediction skill of the operational models.

1. Introduction

The Madden–Julian oscillation (MJO) is a planetary-scale, equatorially trapped convective disturbance that propagates eastward with a period of 30–60 days (e.g., Madden and Julian 1971, 1972; Zhang 2005). The MJO significantly modulates not only precipitation but also large-scale atmospheric circulation in the tropics (Zhang 2013). For example, MJO-related circulation anomalies affect the Indian and Australian monsoons, as well as the African monsoon (Yasunari 1979; Hendon and Liebmann 1990; Lavender and Matthews 2009). The MJO-related circulation anomalies also affect the genesis of tropical cyclones over all ocean basins (e.g., Hall et al. 2001).

The impact of the MJO is not limited to the tropics. The MJO’s influence is also evident in the extratropics. The upper-level divergence, induced by the MJO-related large-scale vertical motion, often excites the Rossby wave packet that propagates into the subtropical North Pacific, western North America, and North Atlantic region (Matthews et al. 2004; Lin et al. 2009; Seo and Son 2012). Through this teleconnection, the MJO significantly modulates surface weather and climate systems in East Asia, North America, and Europe (Jeong et al. 2005; Cassou 2008).

Given its wide influence, an accurate prediction of the MJO and its teleconnection is crucial for subseasonal predictions. In this regard, the MJO prediction skill in operational models has been extensively examined over the past decade. Among others, it has been reported that the MJO prediction skills of the National Centers for Environmental Prediction (NCEP) and the European Centre for Medium-Range Weather Forecasts (ECMWF) model are approximately 20 days (Wang et al. 2014; Kim et al. 2014) and 31 days (Vitart 2014), respectively. Rashid
et al. (2011) documented that the MJO prediction skill of the Australian Bureau of Meteorology (BoM) model is approximately 21 days. The Japan Meteorological Agency (JMA) and China Meteorological Administration (CMA) coupled models, respectively, showed limits of approximately 25 days (Neena et al. 2014) and 16 days (Liu et al. 2017). Overall, these studies suggest that the MJO prediction skill in recent operational models is approximately 16 to 31 days. A similar result was also found in the latest subseasonal-to-seasonal (S2S) models (Vitart 2017, hereafter V17).

Although MJO prediction skill is typically quantified by a single forecast lead day as indicated above (e.g., the forecast lead time at which point the skill metric is greater than a certain threshold value), it significantly varies from event to event depending on initial MJO amplitudes and phases (e.g., Lin et al. 2008; Rashid et al. 2011; Kim et al. 2014; Neena et al. 2014; Wang et al. 2014; Xiang et al. 2015). In general, MJO events with stronger initial MJO amplitudes are better predicted than the normal events (Rashid et al. 2011; Neena et al. 2014; Xiang et al. 2015). It has also been reported that MJO prediction skill could also be sensitive to initial MJO phase. For example, operational models often suffer from a drop of prediction skill when predicting MJO propagation across the Maritime Continent (e.g., Wang et al. 2014). However, this phenomenon, often referred to as the Maritime Continent MJO prediction barrier (Lin et al. 2008; Vitart and Molteni 2010; Wang et al. 2014), is not observed in several recent models (Kim et al. 2014; Neena et al. 2014; Xiang et al. 2015).

The primary goal of the present study is to assess the MJO prediction skill in a set of newly available operational models. Specifically, the reforecasts from the 10 operational models that participate in the S2S prediction project (Robertson et al. 2015; Vitart et al. 2017) are analyzed. While V17 examined the same set of models regarding their MJO prediction skill and reported that the MJO prediction skill of the S2S models ranges from 13 to 32 days, this study builds upon the V17 study and expands it in several aspects.

First of all, the present study evaluates MJO prediction skill using all available reforecasts. This is different from the approach of V17, who used the reforecast period that is common to all models (1999–2010). This approach allows us to best estimate each model’s MJO prediction skill. More than one skill metric is used to gain clear information about the errors. Second, we examine the sensitivity of MJO prediction skill to initial MJO amplitude and phase. Note that V17 did not investigate these sensitivities. Third, we propose a novel way to decompose the total prediction errors in one of the MJO skill scores into those that are related to MJO amplitude and phase, respectively, and examine their relative contributions to the total error. Last, motivated by the recent development in the theories for the MJO, the current study also attempts to better understand the nature of MJO prediction errors by exploring the relationship of MJO prediction skill with aspects of model biases.

There is a growing body of thought that considers the MJO as a “moisture mode” on an equatorial beta plane (Neelin and Yu 1994; Raymond 2001; Sobel and Maloney 2012, 2013; Adames and Kim 2016; Fuchs and Raymond 2017). Under this framework, which is based on the tight coupling between moisture and convection (e.g., Bretherton et al. 2004) and the smallness of buoyancy perturbations in the tropics (Charney 1963; Sobel et al. 2001), the evolution of large-scale, low-frequency convective anomalies associated with the MJO is explained by those of moisture anomalies. The moisture mode theory has provided a framework for studying and interpreting the column-integrated moisture budget of the MJO in observations and in models. The results of the budget studies collectively indicate that horizontal moisture advection, especially the advection of the mean moisture by the MJO perturbation wind, is the key process that moistens ahead (east) and dries behind (west) the region of enhanced moisture anomalies, controlling the MJO propagation (Kiranmayi and Maloney 2011; Andersen and Kuang 2012; Adames and Wallace 2015; Wang et al. 2017). These results suggest that, to accurately forecast the MJO, the operational models may need to represent a realistic horizontal distribution of the mean moisture. Gonzalez and Jiang (2017) showed that GCMs’ MJO simulation performances have a close relationship with their ability to represent accurately the basic-state moisture distribution over the Indo-Pacific warm pool. Kim (2017) examined the column-integrated moist static energy (MSE) budget of the MJO in the ECMWF reforecast dataset and showed that a dry bias in the mean moisture distribution caused a weakening of horizontal moisture advection, resulting in an early disruption of the MJO in the reforecasts. Motivated by these recent works that emphasize the role of the basic-state moisture distribution on the simulation of the MJO, we examined the relationship of the MJO prediction skills with biases in the mean moisture gradients in the S2S models.

The cloud–longwave radiation (CLW) feedback process has been suggested as the key process for the MJO’s maintenance (Kiranmayi and Maloney 2011; Andersen and Kuang 2012; Adames and Kim 2016). The increase of moisture and clouds during the active phase of the MJO reduces the amount of longwave cooling, causing
an anomalous longwave warming. This anomalous warming is balanced by upward motion, which moistens the column by vertical advection of moisture (Chikira 2014; Janiga and Zhang 2016; Wolding et al. 2016). Through this moistening, the increase of cloud amounts that is caused by the enhanced convection provides a favorable condition for further development of the anomalous convection. Kim et al. (2015) showed that GCMs with stronger CLW feedbacks tend to simulate more pronounced MJO variability. In the current study, such relationships between CLW feedback and MJO prediction skill are quantitatively evaluated using S2S models.

This paper is organized as follows. In section 2, observations, S2S model datasets, and the metrics of MJO forecast skill are introduced. The MJO prediction skill in the S2S models and its characteristics are then described in section 3. The possible cause(s) for the limited MJO prediction skill is discussed in section 4, followed by summary and conclusions in section 5.

2. Data and methods

a. Observations

Daily averaged upper (200-hPa) and lower (850-hPa) tropospheric zonal winds for the period of 1980–2013 are obtained from the ECMWF interim reanalysis data (ERA-Interim; Dee et al. 2011) for the model evaluation. The daily OLR data from the National Oceanic and Atmospheric Administration (NOAA; Liebmann and Smith 1996) and precipitation product from the Global Precipitation Climatology Project (GPCP; Huffman et al. 2001) are used to characterize the tropical convective activity for the periods of 1980–2013 and 1996–2013, respectively. The moisture distribution is further quantified with column-integrated water vapor data derived from the combined precipitable water products from the Special Sensor Microwave Imager (SSM/I; Wentz et al. 2012) and the Tropical Rainfall Measuring Mission Microwave Imager (TMI; Hou et al. 2001) over 1998–2013. We use all available observations to define the mean climatology. To reduce the uncertainty associated with different spatial resolution, all datasets are first interpolated into a common horizontal resolution of 2.5° before being used in the analyses.

b. S2S models

The S2S prediction project was launched jointly by the World Weather Research Programme (WWRP) and the World Climate Research Programme (WCRP), with the goal of better understanding and improving subseasonal-to-seasonal prediction (Vitart et al. 2017). As part of the project, an unprecedented amount of reforecast and near-real-time prediction datasets from operational weather/climate prediction centers worldwide are archived (Table 1). The participating centers include the BoM, 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, Hydro-meteorological Centre of Russia (HMCR), JMA, NCEP, and Met Office (UKMO). In the present study, only the long-term reforecast datasets are used.

Table 1 summarizes the reforecast datasets available at the S2S project data (Vitart et al. 2017) as of February 2017. All models provide extended reforecasts, which

<table>
<thead>
<tr>
<th>Modeling center</th>
<th>Model resolution</th>
<th>Reforecast period</th>
<th>Reforecast frequency</th>
<th>Total number of reforecasts</th>
<th>Reforecast length</th>
<th>Ensemble size</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoMa*</td>
<td>T47 L17</td>
<td>1981–2013</td>
<td>6 month⁻¹</td>
<td>990</td>
<td>Days 1–62</td>
<td>33</td>
</tr>
<tr>
<td>CMA*,a</td>
<td>T106 L40</td>
<td>1994–2013</td>
<td>6 month⁻¹</td>
<td>600</td>
<td>Days 1–60</td>
<td>4</td>
</tr>
<tr>
<td>CNR-ISAC</td>
<td>0.75° × 0.56° L54</td>
<td>1981–2010</td>
<td>Every 5 days</td>
<td>900</td>
<td>Days 1–31</td>
<td>1</td>
</tr>
<tr>
<td>CNRM*</td>
<td>T255 L91</td>
<td>1993–2013</td>
<td>2 month⁻¹</td>
<td>210</td>
<td>Days 1–61</td>
<td>15</td>
</tr>
<tr>
<td>ECCCa</td>
<td>0.45° × 0.45° L40</td>
<td>1995–2013</td>
<td>Weekly</td>
<td>418</td>
<td>Days 1–32</td>
<td>4</td>
</tr>
<tr>
<td>ECMWF*</td>
<td>T639/319 L91</td>
<td>1996–2013</td>
<td>2 week⁻¹</td>
<td>720</td>
<td>Days 1–46</td>
<td>11</td>
</tr>
<tr>
<td>HMCR</td>
<td>1.1° × 1.4° L28</td>
<td>1985–2010</td>
<td>Weekly</td>
<td>546</td>
<td>Days 1–61</td>
<td>10</td>
</tr>
<tr>
<td>JMA*</td>
<td>T319 L60</td>
<td>1981–2010</td>
<td>3 month⁻¹</td>
<td>450</td>
<td>Days 1–34</td>
<td>5</td>
</tr>
<tr>
<td>NCEP*,a</td>
<td>T126 L64</td>
<td>1999–2010</td>
<td>6 month⁻¹</td>
<td>360</td>
<td>Days 1–44</td>
<td>4</td>
</tr>
<tr>
<td>UKMO</td>
<td>N216 L85</td>
<td>1996–2009</td>
<td>4 month⁻¹</td>
<td>280</td>
<td>Days 1–60</td>
<td>3</td>
</tr>
</tbody>
</table>

*Models used in section 4.

b Subsampled to be compared with other models.
were initialized at least twice a month and integrated for up to 62 days, for at least 11 years. In this study, all available reforecast datasets are used, except for the CMA and NCEP models. Owing to storage issues, the daily reforecasts of these two models are subsampled by retaining only six initializations per month. Only the boreal winter, when the MJO is strong, is considered in this study. Specifically, the reforecasts initialized from November to March (NDJFM) are examined. Note that the reforecasts initialized in late November and December of 2013 are excluded in the analyses because the OLR observations are not available after January 2014. For instance, the reforecast initialized on 1 December 2013 cannot be compared with observations at day 31 (target date is 1 January 2014) and later.

It should be noted that the reforecast period (third column in Table 1) and frequency (fourth column in Table 1) differ appreciably among the operational centers. For example, the ECMWF model was initialized 8 times per month for 18 years in the 2015 (for November–December) and 2016 (for January–March) version of real-time forecast, providing a total of 720 reforecasts. By contrast, the CNRM model was initialized only twice per month for 21 years, allowing only 210 reforecasts. Also, the size of ensemble differs substantially among the models (rightmost column in Table 1). For example, the CNR-ISAC model has only one ensemble member, whereas the BoM model uses 33 ensemble members.

Unless otherwise specified, all analyses are performed with ensemble-mean forecasts during NDJFM for all available years. While this approach does not allow a fair comparison between the models, it enables us to assess the best estimate of each model’s prediction skill. To reconcile this issue, the overall analyses are also repeated with only one or three ensemble members, as reported in the appendix. The sensitivity test is further performed for the common reforecast period of 1999–2009.

c. Evaluation metrics

The MJO skill assessment is often conducted with Real-time Multivariate MJO (RMM) indices (Wheeler and Hendon 2004, hereafter WH04). The RMM indices, that is, RMM1 and RMM2, are the principle components associated with the two leading empirical orthogonal functions (EOFs) of the combined field of equatorially averaged OLR, and zonal wind at 850 hPa (U850) and 200 hPa (U200). In the current study, the RMM indices of reforecasts are derived following the method described in Gottschalck et al. (2010) and V17, except for the following two differences. The leading pair of combined EOFs, which is required for projection, is obtained from ERA-Interim winds instead of precomputed combined EOFs from WH04. The zonal wind data of WH04 is obtained from NCEP–National Center for Atmospheric Research reanalysis data. In this study, the use of ERA-Interim winds for the leading pair of combined EOFs makes all the analyses more consistent because only one reanalysis dataset is used as a reference. Another difference is the way the climatology is defined. In V17, the climatology is computed in a cross-validated way by averaging the reforecasts initialized on the same day and month as the common period, except for the targeted year. For example, for the reforecast initialized on 1 January 1999, all the reforecasts initialized on 1 January 2000–10 are used to compute the climatology. This implies that the model climatology is based on 11 years and each reforecast initialized on the same day and month has a slightly different climatology. In the present study, the climatology is simply defined by averaging all the reforecasts initialized on the same day and month. For the ECMWF reforecast initialized on 1 January 1999, all the reforecasts initialized on 1 January 1996–2013 are used to define the climatology. In other words, the climatology is computed for the whole reforecast period (Table 1), and each reforecast initialized on the same day and month has the same climatology. These differences could introduce a subtle (but not major) difference in the MJO prediction skill compared to V17.

Using the RMM indices, the observed and forecasted MJO amplitudes, $A_O$ and $A_M$, and their covariability, $C_{OM}$, are defined as below.

$$A_O(t) = \sqrt{O_1^2(t) + O_2^2(t)}, \quad (1)$$
$$A_M(t, \tau) = \sqrt{M_1^2(t, \tau) + M_2^2(t, \tau)}, \quad \text{and} \quad (2)$$
$$C_{OM}(t, \tau) = O_1(t)M_1(t, \tau) + O_2(t)M_2(t, \tau), \quad (3)$$

where $O_1(t)$ and $O_2(t)$ are the observed RMM1 and RMM2 indices at time $t$, and $M_1(t, \tau)$ and $M_2(t, \tau)$ are the respective reforecasts with the lead time of $\tau$. Likewise, the observed and forecasted MJO phases in the RMM space, $\phi_O$ and $\phi_M$, are defined as below.

$$\phi_O(t) = \tan^{-1}\left[\frac{O_2(t)}{O_1(t)}\right], \quad \text{and} \quad (4)$$
$$\phi_M(t, \tau) = \tan^{-1}\left[\frac{M_2(t, \tau)}{M_1(t, \tau)}\right]. \quad (5)$$

Using these properties, several MJO skill metrics are calculated. One of the key metrics used in this study is
The so-called bivariate anomaly correlation coefficient (BCOR) (e.g., Rashid et al. 2011):

$$\text{BCOR}(\tau) = \frac{\sum_{i=1}^{N} C_{OM}(t, \tau)}{\sqrt{\sum_{i=1}^{N} A_{O}^2(t)} \sqrt{\sum_{i=1}^{N} A_{M}^2(t, \tau)}}$$  (6)

where $N$ is the number of reforecasts. Following previous studies, the MJO prediction skill of each model is determined as the forecast lead time when BCOR becomes lower than 0.5 (e.g., Rashid et al. 2011). Since the choice of this threshold value is somewhat arbitrary, another threshold value, such as 0.7, is also considered in the sensitivity test (e.g., Table 2).

To track the relative importance of MJO amplitude and phase errors to the total error, a bivariate-mean-squared error (BMSE) is defined in this study:

$$\text{BMSE}(\tau) = \frac{1}{N} \sum_{i=1}^{N} \left[ (M_1(t, \tau) - O_1(t))^2 + (M_2(t, \tau) - O_2(t))^2 \right] .$$  (7)

This metric is the same as the square of root-mean-squared error (RMSE) in the literature (e.g., Rashid et al. 2011). As RMSE $< \sqrt{2}$ is generally used to determine the prediction skill, BMSE $< 2.0$ is set as an upper limit of the reliable MJO prediction in this study. BMSE can be decomposed into its amplitude-error component, BMSEa, and phase-error component, BMSEP, as below.

$$\text{BMSE}(\tau) = \text{BMSEa}(\tau) + \text{BMSEP}(\tau) ,$$  (8)

where

$$\text{BMSEa}(\tau) = \frac{1}{N} \sum_{i=1}^{N} [A_M(t, \tau) - A_O(t)]^2 ,$$  and (9)

$$\text{BMSEP}(\tau) = \frac{1}{N} \sum_{i=1}^{N} 2A_M(t, \tau) A_O(t) \times \{ 1 - \cos[\phi_M(t, \tau) - \phi_O(t)] \} .$$  (10)

As shown below, the BMSEa is essentially the same as the mean-squared amplitude error [Eq. (13)]. Note that the BMSEP is not only determined by the MJO phase error but also weighted by the observed and forecasted MJO amplitudes. As the MJO amplitude decreases with forecast lead times, an increasing BMSEP, which is the case for all reforecasts, is primarily explained by the MJO phase error (not shown). Although this decomposition does not clearly separate MJO amplitude and phase errors, it still allows us qualitatively to attribute the total model errors into amplitude-dependent and phase-dependent ones.

To understand the amplitude and phase errors more directly, other standard metrics of MJO amplitude and phase errors are also utilized. They include the mean amplitude error ($\overline{\text{AE}}$) and mean phase error ($\overline{\text{PE}}$):

$$\overline{\text{AE}}(\tau) = \frac{1}{N} \sum_{i=1}^{N} [A_M(t, \tau) - A_O(t)] ,$$  and (11)

$$\overline{\text{PE}}(\tau) = \frac{1}{N} \sum_{i=1}^{N} [\phi_M(t, \tau) - \phi_O(t)].$$  (12)

The variable $\overline{\text{PE}}$ typically ranges from $-\pi/2$ to $\pi/2$. In some cases, $\overline{\text{PE}}$ can rapidly increase with forecast lead times, and eventually jump from $\pi/2$ to $-\pi/2$ when the MJO amplitude is small (Rashid et al. 2011). To prevent

<table>
<thead>
<tr>
<th>Model</th>
<th>Phase 8–1</th>
<th>Phase 2–3</th>
<th>Phase 4–5</th>
<th>Phase 6–7</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoM</td>
<td>27 (15), 28</td>
<td>31 (16), 28</td>
<td>24 (15), 28</td>
<td>28 (14), 28</td>
</tr>
<tr>
<td>CMA</td>
<td>18 (12), 20</td>
<td>15 (9), 17</td>
<td>18 (11), 19</td>
<td>28 (14), 28</td>
</tr>
<tr>
<td>CNR-ISAC</td>
<td>15 (11), 13</td>
<td>14 (10), 13</td>
<td>17 (12), 14</td>
<td>15 (11), 13</td>
</tr>
<tr>
<td>CNRM</td>
<td>20 (14), 18</td>
<td>21 (10), 17</td>
<td>20 (13), 16</td>
<td>18 (12), 18</td>
</tr>
<tr>
<td>ECCC</td>
<td>17 (12), 19</td>
<td>19 (13), 20</td>
<td>16 (11), 22</td>
<td>15 (12), 15</td>
</tr>
<tr>
<td>ECMWF</td>
<td>36 (23), 40</td>
<td>38 (20), 40</td>
<td>40 (26), 41</td>
<td>36 (25), 39</td>
</tr>
<tr>
<td>HMCR</td>
<td>12 (6), 11</td>
<td>13 (5), 13</td>
<td>13 (7), 10</td>
<td>9 (5), 5</td>
</tr>
<tr>
<td>JMA</td>
<td>17 (12), 18</td>
<td>18 (12), 18</td>
<td>17 (11), 15</td>
<td>15 (11), 17</td>
</tr>
<tr>
<td>NCEP</td>
<td>24 (12), 17</td>
<td>30 (12), 18</td>
<td>18 (14), 19</td>
<td>31 (13), 15</td>
</tr>
<tr>
<td>UKMO</td>
<td>25 (16), 31</td>
<td>22 (14), 24</td>
<td>22 (17), 27</td>
<td>27 (15), 30</td>
</tr>
<tr>
<td>MMM</td>
<td>21.1 ± 7.0</td>
<td>22.1 ± 8.3</td>
<td>20.5 ± 7.5</td>
<td>21.1 ± 8.8</td>
</tr>
<tr>
<td></td>
<td>(13.3 ± 4.3)</td>
<td>(12.1 ± 4.1)</td>
<td>(13.7 ± 5.1)</td>
<td>(12.9 ± 5.0)</td>
</tr>
<tr>
<td></td>
<td>21.5 ± 8.9</td>
<td>20.8 ± 8.1</td>
<td>21.1 ± 9.0</td>
<td>19.7 ± 9.8</td>
</tr>
</tbody>
</table>

$C_{OM}(t, \tau)$ is the variable to determine the prediction skill of each model. The first number, followed by parentheses, denotes the BCOR skill. The number in parentheses is also the BCOR skill, but based on a correlation coefficient greater than 0.7 instead of 0.5. The second number indicates the BMSE skill. The multimodel mean (MMM) value and one standard deviation are also shown at the bottom.
such an artificial jump, $\overline{PE}$ is evaluated in units of degrees from $0^\circ$ to $360^\circ$, and subsequently converted into radian.

Both $\overline{AE}$ and $\overline{PE}$ characterize the mean errors, rather than the absolute errors of individual reforecasts. If the model errors are positive in some cases but negative in others, the mean errors would become negligible. To prevent such cancellation and to quantify model error better, the mean-squared amplitude error ($\overline{AE^2}$) and mean-squared phase error ($\overline{PE^2}$) are also considered in this study. They are computed as below.

$$\overline{AE^2}(\tau) = \frac{1}{N} \sum_{t=1}^{N} [A_M(t, \tau) - A_O(t)]^2,$$

and

$$\overline{PE^2}(\tau) = \frac{1}{N} \sum_{t=1}^{N} [\phi_M(t, \tau) - \phi_O(t)]^2.$$  \hspace{1cm} (13)

$$\overline{PE^2}(\tau) = \frac{1}{N} \sum_{t=1}^{N} [\phi_M(t, \tau) - \phi_O(t)]^2. \hspace{1cm} (14)$$

3. MJO prediction skill

a. BCOR skill

Figure 1a presents the MJO prediction skill of all models in terms of the BCOR metric. Reforecasts with
initial MJO amplitude smaller than 1 are excluded to ensure that the MJO signal is robust at least in the initial conditions. BCOR consistently decreases with forecast lead times and crosses the threshold of 0.5 in 12–36 days, which is consistent with the result of V17. Among the 10 S2S models, the ECMWF model shows a relatively high prediction skill of 36 days, with the slowest decrease of BCOR. By contrast, the HMCR model shows a relatively rapid decrease of BCOR in the first 2 weeks, which drops below 0.5 at the forecast lead time of 12 days. Except for these two models, others are clustered together in the first 2-week forecasts.

The overall MJO prediction skill, shown in Fig. 1a, is summarized in Table 2 (see the second column). The multimodel mean MJO prediction skill is approximately 3 weeks. Even with a higher threshold value (e.g., BCOR > 0.7), the MJO is qualitatively well predicted for up to approximately 2 weeks (the numbers in parentheses in Table 2). Quantitatively, the BoM, ECMWF, and UKMO models show relatively higher BCOR skills, with retaining BCOR greater than 0.5 (0.7) up to more than 25 (15) days of forecast lead time. As discussed later, these three models also exhibit relatively higher BMSE skills, ranging from 28 to 40 days (see the third numbers in the second column in Table 2).

It is worth mentioning that the above skill estimates, 12–36 days, are slightly different from those reported by V17, who showed that the MJO prediction skill of the S2S models ranges from 13 to 32 days. Each model’s performance in this result is also slightly different from those in V17’s result. For instance, the CNRM model is the second best in V17 (see their Fig. 1b) but only the fifth best in this study (Fig. 1a and Table 2), although the model rank is not meaningful in this study. This difference is likely caused by multiple factors: the different sampling strategy, the different reforecast periods used, and the slightly different ways of defining the model climatology. In regarding the sampling strategy, unlike V17, who used all reforecasts initialized, we exclude reforecasts that contain weak MJO [\(A_O(1) < 1\)] in their initial conditions. Also, we extend the initial months considered in V17 (December–March) by including November. The constraint of minimum MJO amplitude and the inclusion of November make quantitatively different results (not shown). As mentioned in the introduction, the present study considers the entire reforecast period available instead of focusing on the common reforecast period (1999–2010). As discussed later, quantitative results are somewhat sensitive to the choice of reforecast period (see appendix). Because of the difference in the reforecast period, the model climatology is also slightly different (see section 2c).

Many previous studies have shown that MJO prediction skill is dependent on the initial MJO amplitude. In general, models are more skillful when the initial MJO amplitude is stronger (Lin et al. 2008; Rashid et al. 2011; Kim et al. 2014; Neena et al. 2014; Xiang et al. 2015). This relationship is evaluated using the S2S models in Fig. 2. Following Kim et al. (2016), for each model, all reforecasts are grouped into three categories: strong [\(A_O(1) \geq 1.5\)], moderate [0.7 \(\leq A_O(1) < 1.5\)], and weak [\(A_O(1) < 0.7\)] MJO amplitude in the initial conditions. Note that Fig. 2 is the only case where we analyze the reforecasts with initial MJO amplitudes weaker than 1. On average, the S2S models show 21.7 \(\pm\) 7.2, 19.2 \(\pm\) 7.6, and 15.8 \(\pm\) 7.5 days of prediction skills for initially strong, moderate, and weak MJO events, respectively (Fig. 2), confirming the previous findings (Rashid et al. 2011; Neena et al. 2014; Xiang et al. 2015). It is noteworthy that the CNR-ISAC and UKMO models do not show strong sensitivity of their MJO prediction skill to initial MJO amplitude.

Figure 2 also reveals a large difference in the BCORs among the three groups, especially at early periods in the reforecast. That is, the BCORs of the initially weak MJO events are considerably lower than those of the initially moderate and strong MJO events. For the initially moderate and strong MJO events, the BCORs also tend to decay rather slowly with forecast lead times than that of the initially weak MJO events. Because of these differences, the sensitivity of MJO prediction skill to initial MJO amplitude becomes larger if a higher threshold value of BCOR is used. For example, when MJO prediction skill is evaluated with BCOR > 0.7, the S2S models show MJO prediction skills of 14.3 \(\pm\) 4.4, 9.7 \(\pm\) 4.8, and 7.5 \(\pm\) 6.3 days for the initially strong, moderate, and weak MJO events, respectively. For weak MJO events, more than half of the models (i.e., BoM, CMA, CNR-ISAC, CNRM, HMCR, and NCEP models) either exhibit only a 1-day prediction skill or none with BCOR > 0.8. This result again confirms that models predict the MJO better when initial MJO amplitude is strong.

Several studies have documented that MJO prediction skill is sensitive to the initial MJO phase. In particular, models show relatively lower skill when initial MJO phase is 2 or 3, indicating that the models have difficulty in representing the MJO’s propagation across the Maritime Continent (Vitart and Molteni 2010; Wang et al. 2014). This relatively poor MJO prediction skill with initial MJO phase 2–3 is often referred to as the Maritime Continent prediction barrier. However, more recent studies have reported that in the latest operational models, the Maritime Continent prediction barrier is not as pronounced as in the old models.
Therefore, it is of great interest to see whether the Maritime Continent prediction barrier is present in the S2S models.

(Kim et al. 2014; Neena et al. 2014; Xiang et al. 2015). Figure 3 illustrates the sensitivity of MJO prediction skill to initial MJO phase. Most models show some sensitivity to initial MJO phase, though the sensitivity differs among the models. In other words, the S2S
models show no systematic sensitivity of their MJO prediction skill to initial MJO phase. These results also suggest that the Maritime Continent prediction barrier is not a common symptom of operational models. The only model that shows a hint of the Maritime Continent prediction barrier is the NCEP model (Fig. 3i), the descendent of the model used in Wang et al. (2014). In terms of BCOR skill, MJO prediction skills of the NCEP model are 30, 18, 31, and 23 days for initial MJO phases 8–1, 2–3, 4–5, and 6–7, respectively (Table 2). On the other hand, the ECMWF model shows the highest prediction skill in initial MJO phase 2–3 (38, 40, 36, and 15 days, respectively).
31 days of the MJO prediction skills in each initial MJO phase group).

b. BMSE skill

Figure 1b presents the BMSE time series of each model. Not surprisingly, BMSE steadily increases with forecast lead times and crosses the threshold value (2.0, gray dotted line in Fig. 1b) on forecast lead times of about 11–40 days. When averaged across all models, the MJO prediction skill evaluated with BMSE < 2.0 is approximately 3 weeks (see the third numbers in the second column of Table 2). Among the 10 S2S models, the BoM, ECMWF, and UKMO models show relatively higher BMSE skills, as for the BCOR skills (Table 2).

The advantage of the BMSE metric over the BCOR metric is that it can be decomposed into the amplitude-error-dependent component, BMSEa, and the phase-error-dependent component, BMSEp, as described in section 2c. In Fig. 4, the time evolutions of BMSEa (red) and BMSEp (blue) are displayed for each model. Overall, BMSEp is larger than BMSEa during the 40-day period considered, suggesting that MJO phase error dominates the total error. Although BMSEa is initially larger than BMSEp in some models (i.e., ECCC, HMCR, and JMA models), the latter becomes dominant after 2 weeks (Figs. 4e,g,h). In the CNR-ISAC, CNRM, and NCEP models, BMSEp is even twice as large as BMSEa (Figs. 4c,d,i). Here, one notable exception is the ECMWF model (Fig. 4f), which shows BMSEp that is comparable to BMSEa during the 40-day period. Note that the results presented in Fig. 4 are not significantly affected by initial MJO amplitude or phase (not shown). These results suggest that BMSEp plays a more important role than BMSEa in the growth of BMSE and MJO prediction skill. The results also imply that the improved MJO prediction in the S2S models can be effectively achieved by reducing the source(s) of MJO phase errors. As discussed later, such errors are partly associated with the model mean biases in moisture distribution.

c. Amplitude and phase errors

Figures 1c–f display $\Delta AE$, $\Delta PE$, $\Delta AE^2$, and $\Delta PE^2$ obtained using all available reforecasts for each model. The variables $\Delta AE$ and $\Delta PE$ show negative values for all models throughout the 40-day period considered, indicating that most S2S models underestimate MJO amplitude and phase speed (Figs. 1c,d). Their temporal evolutions of $\Delta AE$s differ, although all models have negative $\Delta AE$s (Fig. 1c). They can be largely grouped into four categories in terms of the evolution of $\Delta AE$. The group of the CNR-ISAC, CNRM, and NCEP models has almost negligible $\Delta AE$ throughout the forecast. In another group of three models, that is, the ECCC, HMCR, and JMA models, $\Delta AE$ grows rapidly with a minus sign during the first week and remains at approximately −0.7 afterward. The rapid development of $\Delta AE$ is quite systematic especially in the latter two models. The third group, consisting of the CMA, ECMWF, and UKMO models, shows an almost linear increase of negative $\Delta AE$ with forecast lead times. The last group is the BoM model. This model does not belong to any of the above three categories. Its $\Delta AE$ is substantially high at forecast day 1 (about −0.30) as in the HMCR model. However, its temporal evolution is markedly different from any of the other models. The $\Delta AE$ of the BoM model decreases with forecast lead times until day 10 and then increases afterward, consistent with Fig. 4c of Rashid et al. (2011). This nonlinear evolution is mainly due to the positive and negative errors in the reforecasts canceling each other out, as it does not appear in $\Delta AE^2$ (Fig. 1e).

Focusing on the absolute magnitude of the errors, the mean-squared amplitude errors are illustrated in Fig. 1e. Unlike $\Delta AE$, $\Delta AE^2$ quasi-linearly increases with forecast lead times. Furthermore, its intermodel spread is smaller than that in $\Delta AE$, especially at later forecast lead times. As such, the models cannot be simply divided into the four groups. Some models, such as the JMA and HMCR models, show a relatively rapid increase at early forecast lead times, whereas others show a quasi-linear increase. Although not distinctive, the ECMWF model exhibits the minimum $\Delta AE^2$ until forecast day 25.

Figure 1d presents the time series of $\Delta PE$ for each model. Unlike $\Delta AE$, which shows a large intermodel spread even at day 1 with respect to the maximum spread around day 20–30, $\Delta PE$ shows a relatively small intermodel spread, especially during the first 5 days. $\Delta PE$s in the CNR-ISAC and UKMO models are near zero, whereas the maximum $\Delta PE$s of the CNRM and NCEP models are about −3 m/16. None of the S2S models show $\Delta PE$s smaller than −π/4 (about 1 phase in the RMM space) throughout the forecast.

Figure 1f illustrates the $\Delta PE^2$, which shows a steady increase with forecast lead times. Unlike $\Delta PE$, the square of the phase error tends to increase linearly. Although the intermodel spread of $\Delta PE$ becomes small at forecast day 40, the spread of $\Delta PE^2$ increases continuously with forecast lead times. In contrast to $\Delta AE^2$, $\Delta PE^2$ does not saturate with forecast lead times, suggesting that the overall MJO prediction skill could be more sensitive to the MJO phase error than to the amplitude error at longer forecast lead times. Note that the ECMWF model has the smallest $\Delta PE^2$ throughout the forecast, whereas the HMCR model has a large $\Delta PE^2$ during the first 2 weeks. These two models have the best and worst BCOR and BMSE skills (Table 2; see also Figs. 1a,b).
d. Relationship between error and skill metrics

Figure 5 summarizes relationships among the MJO prediction skill scores in the 2-week (squares) and 4-week (circles) forecasts. Here, the 2-week and 4-week forecasts are defined by averaging values over forecast lead days 8–14 and 22–28, respectively. Not surprisingly, the BCOR and BMSE values are highly correlated with each other (Fig. 5a). The correlation coefficient between the two in the 2-week forecast \( r_2 \) is \(-0.97\). This one-to-one
relationship indicates that the BMSE metric is comparable to the BCOR metric. Their ratio, which crosses BCOR = 0.5 (dashed horizontal line) and BMSE = 2.0 (dashed vertical line), further suggests that the BMSE skill can replace the BCOR skill. In fact, as summarized in Table 2 (see the second column), the BMSE skills are quantitatively similar to the BCOR skills in most models [approximately 21 days in multimodel mean (MMM) prediction skills]. Although these two skill metrics tend to diverge at longer lead times, they are still significantly correlated at the 4-week forecast ($r_4 = -0.8$).

The two representative mean-error metrics, that is, $\overline{AE}$ and $\overline{PE}$, are not closely related to each other (Fig. 5b). This result is again due to cancellations between large positive and negative errors. The variables $\overline{AE^2}$ and $\overline{PE^2}$ are significantly correlated, even in the 4-week forecast. Their correlation coefficient across all 10 models is 0.91 and 0.87 at the 2-week and 4-week forecast, respectively, suggesting that the amplitude and phase errors are inherently related to each other.

The relationship of $\overline{AE^2}$ and $\overline{PE^2}$ with the BCOR is evaluated in Figs. 5d and 5e, respectively. As anticipated from Figs. 5a and 5c, both $\overline{AE^2}$ and $\overline{PE^2}$ are highly correlated with BCOR, with a correlation coefficient greater than −0.94 in the 2-week forecast and −0.87 in the 4-week forecast. Between them, $\overline{PE^2}$ shows a higher correlation with BCOR than $\overline{AE^2}$. Their correlation coefficient remains −0.98 even in the 4-week forecast. This result again suggests that the MJO prediction skill in the S2S models is more sensitive to the phase error than the amplitude error, especially at longer forecast lead times (see also Fig. 4).

4. Mean-state biases and their impact on the MJO forecast

As mentioned in the introduction, studies of column-integrated moisture or moist static energy have shown
that the CLW feedbacks and horizontal distribution of the mean-state moisture are key to the MJO maintenance and propagation of the observed MJO. In this section, we investigate the relationship between the MJO prediction skills in the S2S models and model biases in the basic-state moisture distribution and CLW feedbacks. Among the 10 S2S models, only seven models provide column-integrated water vapor as indicated in Table 1. Therefore, all analyses below are performed with these seven models.

Figure 6a shows the NDJFM mean column-integrated water vapor (CWV) from the satellite observations described in section 2a. The observed CWV distribution exhibits a distinct maximum in the Indo-Pacific warm pool region, featuring large-scale zonal and meridional gradients of CWV across the Indo-Pacific warm pool region. The CWV maximum over the warm pool area is typically underestimated in the S2S models (Figs. 6b–h).

Each model’s CWV, averaged over the first 30-day forecast, exhibits a significant dry bias around the Maritime Continent. In the subtropics (poleward of 15°S/N), most models show either dry biases that are weaker than equatorial dry biases (e.g., CMA, ECMWF, JMA, and NCEP models) or wet biases (e.g., BoM, CNRM, and ECCC models). This pattern of moisture biases suggests that both zonal and meridional gradients of the background moisture are underestimated in most models.

If all other conditions (e.g., large-scale circulations) are equal, a weaker horizontal moisture gradient would dampen horizontal moisture advection associated with the MJO. Because it is the horizontal moisture advection that dominates the moistening and drying tendencies to the east and west of the enhanced MJO convections, respectively, thereby pushing the MJO convection anomalies to move eastward, a weaker horizontal advection would slow down the eastward propagation of the MJO.
This line of consideration leads us to hypothesize that the models with a smaller moisture gradient bias may have a better MJO prediction skill.

The above hypothesis is qualitatively tested by examining the relationship between MJO prediction skill and moisture gradient bias among the S2S models (Fig. 7). In Fig. 7, the reforecasts initialized in initial MJO phases 2–3 (Figs. 7a,c) and 6–7 (Figs. 7b,d) with a minimum amplitude of 1.0 are used. For the reforecasts with initial MJO phase 2–3, a scalar metric of zonal moisture gradient is computed by taking the difference between the area-averaged CWV over the equatorial Maritime Continent (10°S–10°N, 100°–120°E) and that of the eastern Indian Ocean (10°S–10°N, 60°–80°E). The meridional moisture gradient is defined by the difference between the area-averaged CWV over the equatorial (5°S–5°N, 60°–120°E) and subtropical (25°–20°S, 60°–120°E, and 15°–20°N, 60°–120°E) regions. For the reforecasts with initial MJO phase 6–7, slightly different domains are used when computing the zonal and meridional moisture gradients. Specifically, the zonal gradient is defined by taking the difference between the central Pacific (10°S–10°N, 170°–190°E) and the Maritime Continent (10°S–10°N, 110°–150°E), whereas the meridional gradient is quantified by the difference between the equatorial (10°–5°S, 110°–170°E) and subtropical (35°–30°S, 110°–170°E, and 10°–15°N, 110°–170°E) regions.

It is evident from Fig. 7 that the BCOR skills are closely related to the horizontal moisture gradient biases. The ECMWF model, which shows the highest prediction skill of MJO phase 2–3 (40-day BCOR skill; Table 2), has the smallest moisture gradient biases, whereas the JMA model, with a prediction skill of only 17 days (Table 2), shows the largest moisture gradient biases. Note that the meridional moisture gradients have a higher correlation with MJO prediction skill than zonal moisture gradients (cf. correlation coefficients in Fig. 7a and those in Fig. 7c). This result partly explains the different MJO prediction skills among the BoM, CMA, CNRM, and ECCC models, which have similar zonal moisture gradient biases. As shown in Figs. 7b,d, essentially the same result is found for MJO phase 6–7. One difference from MJO phase 2–3 is that the BCOR skills are more closely related to the zonal moisture gradient biases than the meridional moisture gradient biases (cf. Figs. 7b,d). These results presented in Fig. 7 are only weakly sensitive to the domain used to calculate the zonal and meridional moisture gradients.

The above analyses are repeated for the 4-week forecasts (not shown). As anticipated, the linear relationship becomes weaker. For instance, the correlation of BCOR skills to the zonal moisture gradient biases is lowered from 0.76 in the 2-week forecasts (Fig. 7a) to 0.57 in the 4-week forecasts. None of the 4-week correlations are statistically significant. It is, however, important to note that a qualitative similar relationship still holds even in the 4-week forecasts. It is worthwhile to note that the correlation coefficients become lower when the ECMWF model is excluded (see gray lines in Figs. 7 and 9). This suggests that a larger number of models would be useful to better understand the factors that influence MJO prediction skills in operational models.

Next, we examine the relationship between MJO prediction skill and model biases in the CLW feedbacks. The strength of the observed CLW feedbacks is presented in Fig. 8a. Here, the CLW feedbacks are quantified by regressing OLR anomalies against precipitation anomalies (both in units of W m$^{-2}$) and then multiplying $-1$ to the resulting regression coefficients (Lin and Mapes 2004). To isolate the MJO-related feedbacks, both OLR and precipitation anomalies are obtained by subtracting the daily climatology and the mean of the 120-day segment that ends on the day of interest. The resulting values (Fig. 8a) indicate the ratio of anomalous longwave heating rate to anomalous condensational heating rate in the column.

Figures 8b–h show that the S2S models have a wide spread in their representation of the CLW feedbacks, possibly suggesting an important role of physical parameterizations (e.g., cloud microphysics, radiation). Although the bias patterns differ substantially among the models, most models exhibit negative biases in the CLW feedbacks over the Indo-Pacific warm pool region (15°S–15°N, 60°E–180°). An exception is the ECMWF model, which shows somewhat stronger CLW feedbacks over the Indian Ocean than their observational counterparts.

Figure 9 shows that MJO prediction skill is tightly linked to the CLW feedback bias. Here the CLW feedback bias is averaged over the Indo-Pacific warm pool region by only considering the oceanic grid points. Consistent with Fig. 7, only the 2-week forecasts, initialized in MJO phase 2–3 or 6–7, are considered. Although not shown, the overall results are not sensitive to the MJO phase. When all MJO events are considered, the correlation coefficient between the MJO prediction skills and CLW feedback biases becomes slightly larger (0.85 for BCOR in Fig. 9a).

The above results suggest that the MJO prediction skills in the S2S models are closely related to model biases in the mean-state moisture distribution and in the CLW feedbacks. Here it should be stated that the moisture gradient and CLW biases are not
independent to each other. They are physically related through precipitation biases in the Maritime Continent. In fact, the correlation of the zonal moisture gradient biases to the CLW biases is 0.53 for MJO phase 2–3 and 0.78 for MJO phase 6–7. This suggests that the moisture gradient and CLW biases mutually, not independently, influence the MJO prediction errors.

**Fig. 7.** Relationship between the model mean biases in moisture gradient and the BCOR skills in the 2-week forecast: (a),(b) the zonal moisture gradient biases vs BCORs for the reforecasts initialized in MJO phase 2–3 and MJO phase 6–7, and (c),(d) as in (a),(b), but for the meridional moisture gradient biases. See the text for the definition of zonal and meridional moisture gradients. The correlation coef., $r^2$, that is statistically significant at the 95% confidence level, is denoted by an asterisk. The regression line is also added. The gray $r^2$ and the gray regression line indicate the analysis result without the ECMWF model.
5. Summary and discussion

This study assesses the MJO prediction skill of the state-of-the-art operational models in the S2S prediction project. The long-term reforecast datasets from 10 operational centers that have been collected and archived for the S2S project are used to quantify the overall MJO prediction skill and its intermodel spread. The relationship of MJO prediction skill with model biases is also examined.

All S2S models underestimate the MJO’s amplitude and phase speed within the 40-day period considered in the current study. This result indicates that the models tend to underestimate the MJO organization and its eastward propagation. In terms of the BCOR skill being greater than 0.5, the S2S models show useful prediction skill up to forecast lead times of 12–36 days. This skill estimate is comparable to those obtained in other recent studies (Rashid et al. 2011; Neena et al. 2014; Kim et al. 2014; Vitart 2014; Wang et al. 2014; Liu et al. 2017; V17). A quantitatively similar result is also found when the BMSE is applied. For both metrics, the multimodel mean MJO prediction skill is approximately 21 days with a pronounced intermodel spread. Among the 10 S2S models, the BoM, ECMWF, and UKMO models exhibit relatively higher performances. Their BCOR/BMSE skills are 27/28, 36/40, and 25/31 days, respectively. These skill scores are considerably higher than the skill scores of their antecedent models. For instance, the BoM, NCEP, and ECMWF models examined in this study show 4- to 6-day longer BCOR skill than their earlier versions (Rashid et al. 2011; Vitart 2014; Wang et al. 2014).

MJO prediction skill shows a distinct sensitivity to initial MJO amplitude. In general, a higher MJO prediction skill is obtained in the reforecasts with stronger initial MJO amplitude. This dependency becomes clearer...
when a higher threshold is applied to the BCOR skill (e.g., 0.7 instead of 0.5). While the MJO prediction skill in the S2S models depends on initial MJO phase, the sensitivity of MJO prediction skill to initial MJO phase differs substantially among the models. This result suggests that the so-called Maritime Continent MJO prediction barrier (Lin et al. 2008; Vitart and Molteni 2010; Wang et al. 2014) is not a common symptom in the latest operational models.

The relative importance of MJO amplitude and phase errors in the extended MJO forecast is examined by decomposing the BMSE into the amplitude-error-dependent component (BMSEA), and the phase-error-dependent component (BMSEP). While both BMSEA and BMSEP contribute to the total error (BMSE), BMSEP becomes increasingly more important with forecast lead times. Consistent with this finding, the BCOR and BMSE skill estimates show a more tight relationship with PE than AE. Quantitatively, PE explains approximately 98% of the intermodel spread of BCOR skills in the 4-week forecast, whereas AE explains approximately 87% of it.

To elucidate the MJO prediction skill in the S2S models and its intermodel spread, model biases in the mean moisture distribution and in the CLW feedbacks and their relationships with the BCOR skill estimates are explored. Most models exhibit pronounced dry biases around the Maritime Continent, and underestimate the CLW feedbacks. The former results in a weak horizontal moisture gradient over the Indian Ocean and the western Pacific, which would weaken MJO organization and propagation. These results show that the model with a larger moisture gradient bias has lower BCOR. Likewise, the model with a larger bias in CLW feedbacks has relatively low BCOR. These results suggest that MJO prediction skill could be improved by correcting these model biases. In this regard, the process-based studies focusing on the model biases are warranted. Modeling studies with varying cloud parameterizations or radiation schemes could also be useful.

It should be emphasized that the intermodel comparison reported in this study needs to be taken only qualitatively. We are not aiming to rank model performance, but instead are attempting to best estimate each model’s MJO prediction skill. In fact, a one-to-one comparison is unfair because each model has a different reforecast frequency, period, and ensemble
size. For instance, the MJO prediction skill is highly sensitive to the ensemble size (e.g., Neena et al. 2014). As shown in the appendix, when the analysis is repeated with only one ensemble member, the multimodel mean BCOR skill decreases from 21 to 17 days (Table A1). The intermodel spread also decreases from 7 to 4 days, indicating that the intermodel spread of MJO prediction skill reported in this study partly results from the different ensemble sizes of the models. For a more fair comparison, each model’s reforecasts should be resampled at the same initialization dates with the same ensemble size over the same analysis period. However, this is not the purpose of the present study.

Acknowledgments. Y. Lim and S.-W. Son were supported by the Korea Meteorological Administration Research and Development Program under Grant KMIPA 2015-2100. D. Kim was supported by the startup grant from the University of Washington.

APPENDIX

Sensitivity Test to the Ensemble Size and Analysis Period

The sensitivity of the MJO prediction skill to the ensemble size and the analysis period is briefly tested in this section. As summarized in Table 1, the S2S models have a wide range of ensemble sizes (see also the first column of Table A1). For example, the BoM and ECMWF models, which show the best BCOR skills, have the largest and the third largest ensemble members (i.e., 33 and 11), whereas the CNR-ISAC model has only one ensemble member. As the prediction skill generally increases when more ensemble members are used (e.g., Neena et al. 2014), the difference in the MJO prediction skill between these models could be partly caused by the different ensemble size.

Table A1 summarizes the impact of the different ensemble sizes in the MJO forecast. For comparison with the standard metric that is based on all available ensemble members (ALL; second column), the three ensemble members (EN3; third column) or only one ensemble member (EN1; fourth column) are used. This sampling corresponds to the ensemble size of the UKMO model and that of the CNR-ISAC model. All models show decreased MJO prediction skills by reducing ensemble size. For instance, the BCOR/BMSE skills of the ECMWF model become 31/34 days in EN3 and 26/24 days in EN1. These values are substantially lower than those in ALL, that is, 36/40 days. The multimodel mean BCOR/BMSE skills also decrease from 21.1/21.5 days in ALL to 17.1/14.8 days in EN1. Such a decrease in the MJO prediction skill, especially in the models with many ensemble members, also reduces the intermodel spread in the BCOR/BMSE skills (i.e., 7.0/8.9 days in ALL but 4.1/4.7 days in EN1).

Each model also has different reforecast periods (Table 1). To test its impact on the MJO prediction skill, all analyses are repeated for the common period of 1999–2009. It turns out that the overall results are not sensitive to the analysis period. Although some models (i.e., BoM, CMA, ECC, ECMWF, NCEP, and UKMO models) show a 1- to 3-day decrease in the BCOR skill, others show either no change or a 1- to 2-day increase in BCOR skill. When the BMSE is considered, most models, except the CNRM and CNR-ISAC models, show essentially no change in the MJO prediction skill (see the fifth column in Table A1).

<table>
<thead>
<tr>
<th>Model</th>
<th>ALL</th>
<th>EN3</th>
<th>EN1</th>
<th>1999–2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoM (33)</td>
<td>27 (15, 28)</td>
<td>23 (13), 20</td>
<td>18 (10), 13</td>
<td>24 (13, 28)</td>
</tr>
<tr>
<td>CMA (4)</td>
<td>18 (12), 20</td>
<td>18 (12), 17</td>
<td>16 (11), 14</td>
<td>17 (11), 20</td>
</tr>
<tr>
<td>CNR-ISAC (1)</td>
<td>15 (11), 13</td>
<td>15 (11), 13</td>
<td>17 (12), 14</td>
<td>17 (12), 14</td>
</tr>
<tr>
<td>CNRM (15)</td>
<td>20 (14), 18</td>
<td>19 (13), 16</td>
<td>17 (11), 13</td>
<td>21 (15), 19</td>
</tr>
<tr>
<td>ECC (4)</td>
<td>17 (12), 19</td>
<td>16 (12), 18</td>
<td>14 (9), 12</td>
<td>16 (11), 19</td>
</tr>
<tr>
<td>ECMWF (11)</td>
<td>36 (23), 40</td>
<td>31 (20), 34</td>
<td>26 (16), 24</td>
<td>34 (20), 40</td>
</tr>
<tr>
<td>HMCR (10)</td>
<td>12 (6), 11</td>
<td>11 (5), 10</td>
<td>10, (5), 7</td>
<td>12 (5), 11</td>
</tr>
<tr>
<td>JMA (5)</td>
<td>17 (12), 18</td>
<td>17 (12), 17</td>
<td>15 (11), 16</td>
<td>17 (12), 18</td>
</tr>
<tr>
<td>NCEP (4)</td>
<td>24 (12), 17</td>
<td>21 (12), 16</td>
<td>19 (12), 14</td>
<td>21 (13), 17</td>
</tr>
<tr>
<td>UKMO (3)</td>
<td>25 (16), 31</td>
<td>25 (16), 31</td>
<td>21 (15), 22</td>
<td>23 (16), 31</td>
</tr>
<tr>
<td>MMM</td>
<td>21.1 ± 7.0 (13.3 ± 4.3), 21.5 ± 8.9</td>
<td>20.1 ± 5.8 (12.8 ± 4.0), 19.9 ± 7.7</td>
<td>17.1 ± 4.1 (11.1 ± 2.9), 14.8 ± 4.7</td>
<td>20.2 ± 5.7 (12.8 ± 3.7), 21.7 ± 8.3</td>
</tr>
</tbody>
</table>
REFERENCES


