

Evaluation of MJO Predictive Skill in Multiphysics and Multimodel Global Ensembles

BENJAMIN W. GREEN AND SHAN SUN

*Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, and
NOAA/Earth System Research Laboratory/Global Systems Division, Boulder, Colorado*

RAINER BLECK

*NASA Goddard Institute for Space Studies, New York, New York, and Cooperative Institute for Research in
Environmental Sciences, University of Colorado Boulder, and NOAA/Earth System Research Laboratory/Global
Systems Division, Boulder, Colorado*

STANLEY G. BENJAMIN AND GEORG A. GRELL

NOAA/Earth System Research Laboratory/Global Systems Division, Boulder, Colorado

(Manuscript received 3 November 2016, in final form 24 February 2017)

ABSTRACT

Monthlong hindcasts of the Madden–Julian oscillation (MJO) from the atmospheric Flow-following Icosahedral Model coupled with an icosahedral-grid version of the Hybrid Coordinate Ocean Model (FIM-iHYCOM), and from the coupled Climate Forecast System, version 2 (CFSv2), are evaluated over the 12-yr period 1999–2010. Two sets of FIM-iHYCOM hindcasts are run to test the impact of using Grell–Freitas (FIM-CGF) versus simplified Arakawa–Schubert (FIM-SAS) deep convection parameterizations. Each hindcast set consists of four time-lagged ensemble members initialized weekly every 6 h from 1200 UTC Tuesday to 0600 UTC Wednesday.


The ensemble means of FIM-CGF, FIM-SAS, and CFSv2 produce skillful forecasts of a variant of the Real-time Multivariate MJO (RMM) index out to 19, 17, and 17 days, respectively; this is consistent with FIM-CGF having the lowest root-mean-square errors (RMSEs) for zonal winds at both 850 and 200 hPa. FIM-CGF and CFSv2 exhibit similar RMSEs in RMM, and their *multimodel* ensemble mean extends skillful RMM prediction out to 21 days. Conversely, adding FIM-SAS—with much higher RMSEs—to CFSv2 (as a multimodel ensemble) or FIM-CGF (as a *multiphysics* ensemble) yields either little benefit, or even a degradation, compared to the better single-model ensemble mean. This suggests that multiphysics/multimodel ensemble mean forecasts may only add value when the individual models possess similar skill and error. An atmosphere-only version of FIM-CGF loses skill after 11 days, highlighting the importance of ocean coupling. Further examination reveals some sensitivity in skill and error metrics to the choice of MJO index.

1. Introduction

The Madden–Julian oscillation [MJO; Madden and Julian (1971, 1972)] involves the coupling of a large-scale atmospheric baroclinic circulation to multiscale convective clusters and is the primary driver of intraseasonal (30–90 day) variability in the tropics

(Zhang 2005). The MJO impacts not just the tropics, but also the entire Earth system (Zhang 2013). Because it acts on relatively long time scales and has influences on global weather, a reasonable representation of the MJO in earth system models is central to intraseasonal prediction. Specifically, a better physical understanding of the MJO—along with improved simulation in numerical models—should help bridge the gap between weather forecasts [which have a predictability limit of ~ 2 weeks in the midlatitudes (Lorenz 1969)] and climate forecasts (interseasonal and longer time scales).

Forecasting the MJO has been a major challenge in weather and climate models alike. Because the MJO is

 Denotes content that is immediately available upon publication as open access.

Corresponding author: Dr. Benjamin W. Green, ben.green@noaa.gov

most commonly described as an eastward-propagating convective anomaly across the equatorial Indian Ocean through the Maritime Continent and into the western Pacific Ocean, it is not surprising that MJO simulations are highly sensitive to the representation of atmospheric convection (e.g., Liu et al. 2005; Zhou et al. 2012; Holloway et al. 2013; Boyle et al. 2015).

Another challenge is defining and identifying the MJO for forecasting and verification purposes. This is not straightforward because of the abovementioned circulation–convection coupling. All of the characteristics of this coupling (propagation speed, intensities of the wind and convective anomalies, and geographic location/extent) are influenced by interannual variability (including El Niño) and the seasonal cycle, and may even vary considerably between successive MJO events. Consequently, multiple algorithms have been proposed to quantify the strength and location of the MJO signal from both observations and numerical models.

The most widely used MJO identification algorithm is the Real-time Multivariate MJO (RMM) index originally proposed by Wheeler and Hendon (2004). Briefly, they used empirical orthogonal function (EOF) analysis to develop the RMM index (which is actually two orthogonal principal components: RMM1 and RMM2) based on three input fields: anomalous zonal winds at 850 and 200 hPa (U850 and U200, respectively) as well as anomalous outgoing longwave radiation (OLR). Anomalies are calculated by removing (i) the annual mean, (ii) the first three harmonics of the seasonal cycle, (iii) the preceding 120 days of anomaly fields, and (iv) the El Niño signal. These anomalies are then averaged at each longitude between 15°S and 15°N, divided by normalization factors, projected onto the observed EOFs, and then normalized again. There are many variants of the RMM, most of which omit step (iv) following Lin et al. (2008). Revisions to the RMM continue to be proposed (Liu et al. 2016); as described in section 2, this study also uses a slightly different method to calculate RMM.

There have been other MJO indices proposed that do not use the same three inputs as the RMM-based indices. Ventrice et al. (2013) followed the methodology of Lin et al. (2008) [i.e., like Wheeler and Hendon (2004) except retaining the El Niño signal] but replaced OLR with the 200-hPa velocity potential (VP200), naming their index the velocity potential MJO (VPM). While VP200 can serve as a proxy for deep convection (via large-scale upper-tropospheric divergence) like OLR, the latter field has some limitations not shared by the former (Ventrice et al. 2013). In addition to a modified RMM index, this study also uses a similarly modified version of the VPM (section 2) in order to examine the

sensitivity of MJO verification (via skill and error) to MJO index.

Another class of MJO indices aims to identify/quantify the MJO solely through its convective signal (typically, OLR). A major benefit of such OLR-only indices is their ability to much better capture the meridional propagation common during boreal summer (e.g., Kikuchi et al. 2012). Conversely, a drawback of these indices—including the all-season OLR-based MJO index and the filtered MJO OLR index, both of which were developed by Kiladis et al. (2014)—is that OLR is much noisier in time and space (and difficult for numerical models to properly represent) than wind fields. This manuscript focuses on RMM and VPM, both of which can be computed for all of the hindcasts examined here without loss of fidelity.

In both shorter-term weather prediction and longer-term climate prediction there has been an increasing utilization of *ensembles* of multiple deterministic model integrations to obtain better forecasts. One immediate advantage of ensembles is that they provide information (via spread) about forecast uncertainty that a single deterministic forecast cannot. Moreover, using the *ensemble mean* as a forecast itself has been shown in many cases to consistently beat the forecasts of the individual ensemble members (e.g., Toth and Kalnay 1997; Grimit and Mass 2002), including for the RMM index (Xiang et al. 2015).

Ensemble forecasting initially focused on using the same dynamic core with the same subgrid-scale parameterizations, but with each member having slightly perturbed initial conditions (e.g., Toth and Kalnay 1993; Tracton and Kalnay 1993). The idea behind perturbing initial conditions only was to account for uncertainty (and error) in the true initial state of the dynamical system. More recently, there has been a significant effort to construct ensembles that are not just varied in their initial conditions, but also in their physics schemes [hereafter *multiphysics* ensembles (MPEs)]¹ and even dynamic cores [hereafter *multimodel* ensembles (MMEs)]. MPEs and MMEs account for forward model error and have been shown to improve over single-model,

¹ There are two, not necessarily exclusive, approaches to varying model physics within an ensemble. One approach is for different members to have different set(s) of subgrid-scale physical parameterization schemes. The second approach is to use “stochastic physics” whereby a stochastic term—different for each ensemble member—is added to the model tendency terms. See Berner et al. (2011) and Bouttier et al. (2012) (and references therein) for detailed discussions of various stochastic physics methods. No stochastic physics are used in this study.

single-physics ensembles (SMSPEs) (e.g., [Evans et al. 2000](#); [Candille 2009](#)).

To the best of the authors' knowledge, only two published studies have demonstrated the benefit of MMEs for MJO forecasting. [Fu et al. \(2013\)](#) examined ~ 31 cases² of forecasts made weekly during November 2011–March 2012 from the 16-member ensemble of the real-time operational Climate Forecast System, version 2 (CFSv2), and a 10-member ensemble of their University of Hawaii coupled model (UH). For this very small sample size, they found skillful forecasts of RMM³ for the ensemble means of CFSv2 and UH out to 25 and 28 days, respectively. When they gave equal weights to the 16-member CFSv2 *ensemble mean* and to the 10-member UH *ensemble mean*, skill increased to 36 days. [Zhang et al. \(2013\)](#) showed that MMEs improved skill of RMM1 and RMM2 (evaluated separately) over a 21-yr hindcast period.

Many more studies have examined SMSPE forecast skill for the MJO. [Wang et al. \(2014\)](#) showed that the mean of 4-member CFSv2 ensemble hindcasts, performed daily in the period from 1999 to 2010, had RMM skill out to 20 days; in fact, a subset of their data is used in this study ([section 2](#)). [Vitart \(2014\)](#) showed that hindcasts made “on-the-fly” with the European Centre for Medium-Range Weather Forecasts (ECMWF) 5-member ensemble had RMM skill out to ~ 30 days covering the period 1995–2001. Similar results were found in a CFSv2-ECMWF comparison study by [Kim et al. \(2014\)](#). Hindcasts from the Australian Bureau of Meteorology coupled model exhibited RMM skill to 21 days ([Rashid et al. 2011](#)).

Some studies only focus on a subset of the year centered on boreal winter, during which the MJO has the greatest prediction skill (e.g., [Rashid et al. 2011](#)). A coupled climate model developed at the Geophysical Fluid Dynamics Laboratory had RMM skill extending to 27 days for 11 years of November–April hindcasts ([Xiang et al. 2015](#)). A comparison of eight different coupled-model hindcasts over an approximate two-decade period in the November–March months ([Neena et al. 2014](#)) found considerable variability in RMM skill (albeit with a different definition of “skill”

than mentioned above), with a version of the ECMWF model slightly different from that in [Vitart \(2014\)](#) having the best skill (out to 28 days). Finally, [Hamill and Kiladis \(2014\)](#) showed RMM skill just under 14 days in hindcasts from the atmosphere-only Global Ensemble Forecast System covering the December–February 1985–2012 period; the importance of ocean coupling for successful MJO forecasts has been demonstrated in several studies (e.g., [Woolnough et al. 2007](#); [Fu et al. 2013](#); [Seo et al. 2014](#); [Tseng et al. 2015](#); see also the review article by [DeMott et al. 2015](#)) and will be visited briefly herein.

This present study is unique in that it evaluates the performance of RMM/VPM forecasts over a long period (1999–2010)—without seasonal restriction—from two versions of the atmospheric Flow-following Icosahedral Model coupled with the icosahedral-grid version of the Hybrid Coordinate Ocean Model (FIM-iHYCOM; [Benjamin et al. 2017](#)), the CFSv2 hindcasts, and various novel combinations of MMEs and/or MPEs. The remainder of the manuscript is structured as follows. [Section 2](#) describes the coupled models used, hindcast postprocessing, and the datasets/methodologies required for evaluation of hindcast performance. Hindcast results are presented in [section 3](#). A discussion is provided in [section 4](#), followed by conclusions in [section 5](#).

2. Data and experimental methods

a. Hindcast models

1) FIM-iHYCOM

As described above, FIM-iHYCOM couples the hydrostatic atmospheric FIM ([Bleck et al. 2015](#)) to an icosahedral-grid version of HYCOM (cf. [Bleck 2002](#)) on a common horizontal mesh to eliminate interpolation of air–sea fluxes. All FIM-iHYCOM hindcasts presented here cover the period from January 1999 to December 2010 (to match the CFSv2 hindcast period, see below). The hindcasts were set up such that, at each week, a 4-member time-lagged ensemble was created by initializing FIM-iHYCOM at 1200 and 1800 UTC Tuesday and 0000 and 0600 UTC Wednesday.⁴ This gives 2500 individual ensemble runs, or 625 *cases* of 4-member ensemble (means) for each model. Initial

² Herein, the term “cases” is used to refer to the number of *independent sets of ensemble integrations*. For example, aggregated verification statistics from an ensemble (of arbitrary member size M) initialized weekly for $N = 625$ weeks would be considered to have $N = 625$ cases.

³ Unless otherwise stated explicitly, the temporal extent of forecast “skill” refers to the time at which the bivariate correlation between model-forecasted and observed RMM/VPM index falls below 0.5 (e.g., [Rashid et al. 2011](#)).

⁴ The Modeling, Analysis, Predictions and Projections Program, part of the Climate Program Office of NOAA, has organized the upcoming Subseasonal Experiment (“SubX,” [NOAA 2016](#)). Initializing the FIM-iHYCOM hindcasts weekly around each Wednesday follows the preliminary SubX protocol.

TABLE 1. Configurations of each of the four hindcasts examined in this study. For a homogeneous comparison (see text), all four hindcasts used runs initialized every Tuesday at 1200 and 1800 UTC and every Wednesday at 0000 and 0600 UTC (to create a four-member time-lagged ensemble once per week) over the common 12-yr period 1999–2010. Moreover, only output data within 768 h (32.0 days) after the initial 0000 UTC Wednesday were considered. Note that FIM-AGF does not include an ocean model: monthly sea surface temperatures from the Hadley Centre were linearly interpolated to every day to provide a boundary condition to the atmosphere. Here, the symbols σ , θ , ρ , and p denote sigma/terrain, isentropic, isopycnic, and pressure, respectively.

		FIM-AGF	FIM-CGF	FIM-SAS	CFSv2
Atmospheric model	Dynamic core	FIM	FIM	FIM	GFS
	Horizontal grid (resolution, structure)	~60-km G7, icosahedral	~60-km G7, icosahedral	~60-km G7, icosahedral	~100 km, T126 spectral
	Vertical grid (No. of layers, structure)	64 layers, hybrid σ - θ	64 layers, hybrid σ - θ	64 layers, hybrid σ - θ	64 layers, hybrid σ - p
	Deep convective scheme	Revised GF	Revised GF	SAS (2015 GFS)	SAS (Saha et al. 2010)
	All other physics	2015 GFS	2015 GFS	2015 GFS	Saha et al. (2014)
Ocean model	Dynamic core	None	iHYCOM	iHYCOM	MOM4
	Horizontal grid (resolution, structure)	—	~60-km G7, icosahedral	~60-km G7, icosahedral	Variable (Saha et al. 2010, 1031–1032)
	Vertical grid (No. of layers, structure)	—	32 layers, hybrid σ - ρ	32 layers, hybrid σ - ρ	40 layers, stretched height

conditions for FIM-iHYCOM were provided directly by CFS reanalysis (CFSR; Saha et al. 2010). No direct data assimilation using FIM was attempted. FIM has 64 vertical layers; iHYCOM has 32 vertical layers, 2 of which are in the upper 5 m to better capture the diurnal oscillation in sea surface temperature. The hindcasts were run at a horizontal grid spacing of ~60 km, with corresponding time steps of 90 s for the dynamic core and 180 s for physics and coupling. Table 1 summarizes all of the hindcast experiments used in this study.

To test the sensitivity to parameterization of deep convection, two parallel FIM-iHYCOM hindcasts were run: one using a revised version of the Grell–Freitas (GF) scheme [Grell and Freitas (2014); hindcast set hereafter as FIM-CGF], and another with a version of the simplified Arakawa–Schubert (SAS; Han and Pan 2011) scheme employed in the Global Forecast System (GFS) physics suite [hindcast set hereafter as FIM-SAS; see also section 4 of Bleck et al. (2015)]. Otherwise, FIM-CGF and FIM-SAS are identical (including the fact that both parameterize shallow convection with a version of SAS).

As mentioned above, several studies have shown that the MJO is much better represented/forecasted in fully coupled atmosphere–ocean models than in atmosphere-only models forced by “offline” sea surface temperatures. To test this, the FIM-CGF configuration described above was run over the hindcast period but with iHYCOM turned off. This atmosphere-only run—FIM-AGF—was forced by observed monthly sea surface temperatures (from the

Hadley Centre) linearly interpolated to daily values. In section 3, it will be shown that FIM-AGF performs much worse than any of the coupled-model hindcasts; thus, FIM-AGF will not be considered in any detailed analysis.

Output variables archived on the native icosahedral grid were transformed to a $2.5^\circ \times 2.5^\circ$ horizontal grid (approximately 5 times the effective grid spacing of the native grid near the equator). For three-dimensional variables, there was also a transformation from the native adaptive vertical coordinate to isobaric levels. Each ensemble member was integrated forward at least 32.5 days (780 h)—thus, the 1200 UTC Tuesday (earliest) member ends exactly 32.0 days (768 h) after 0000 UTC Wednesday.

2) CFSv2

Hindcasts from CFSv2 (Saha et al. 2014) were also examined. CFSv2, like FIM-iHYCOM, couples the atmosphere with the ocean. The hydrostatic atmospheric model in CFSv2 is run at T126 (~100-km resolution) with 64 hybrid vertical layers and uses a SAS-based deep convection parameterization (but a slightly different version than that used by FIM-SAS). The ocean is simulated using version 4 of the Modular Ocean Model, on a different horizontal grid finer than the atmospheric component of CFSv2. This is unlike FIM-iHYCOM, which is unique in that the horizontal grids for the atmosphere and ocean are perfectly matched. CFSR analyses (Saha et al. 2010) were used for all fields at the initial time.

Of all the CFSv2 hindcasts from January 1999 to December 2010 (initialized four times daily), only a subset (1/7) were downloaded to match the initialization times (four weekly) of FIM-iHYCOM described above. Moreover, although CFSv2 integrates out to at least 45 days, only the first 780 h of hindcasts were downloaded to match FIM-iHYCOM. Both CFSR and CFSv2 data were downloaded from the National Centers for Environmental Information (<https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/climate-forecast-system-version2-cfsv2>) on the native grid and then converted to a $2.5^\circ \times 2.5^\circ$ grid.

b. Postprocessing and MJO metrics

This study uses slightly modified versions of both the RMM and VPM indices. Recall that the RMM index requires inputs of anomalous U850, U200, and OLR. VPM is similar to RMM except anomalous VP200 is used instead of OLR. Both indices require daily anomaly fields on a $2.5^\circ \times 2.5^\circ$ horizontal grid in the tropical band from 15°S – 15°N (Wheeler and Hendon 2004; Ventrice et al. 2013).

Both RMM and VPM require the removal of a climatology to obtain anomaly fields. Ideally, this climatology would come from the model itself [i.e., a “model climatology,” Gottschalck et al. (2010)]. However, only CFSv2 has an available model climatology; there is no such climatology for FIM-iHYCOM because the simulations were only initialized on a weekly basis and thus the sample size for a given calendar day is too small. Therefore, the approach of Gottschalck et al. (2010) was followed: removing the climatology based on the years 1979–2001 of the National Centers for Environmental Prediction–National Center for Atmospheric Research reanalysis (Kalnay et al. 1996). Using this reanalysis (which is relatively independent of both CFSv2 and FIM-iHYCOM) to remove climatology facilitates a fairer comparison between the two models than if CFSR had been used. Daily satellite-derived analyses from NOAA were used for OLR (Liebmann and Smith 1996). The sensitivity to climatology source (model hindcast versus reanalysis) was tested in CFSv2 (not shown): the former yielded small but noticeable improvements in RMM/VPM skill/error at lead times beyond ~ 12 days. However, it is possible that the FIM-iHYCOM simulations—which have not been tuned for subseasonal and longer-range forecasts—could see more substantial improvements in RMM/VPM forecasts if anomalies were calculated from (currently nonexistent) model climatologies rather than from reanalyses.

Another modification to the RMM/VPM calculations that is implemented here regards the removal of the preceding 120-day mean (which is done to remove

interannual variability). Because the FIM-iHYCOM hindcasts were only initialized on a weekly basis, there are no FIM-iHYCOM analyses for 6 out of 7 days (i.e., as much as 85% of the data needed to calculate 120-day means are missing). Therefore, the interannual component was retained for FIM-iHYCOM, CFSv2, and the reanalysis fields—again facilitating a fair comparison. It should be noted that Neena et al. (2014; p. 4534) state they “have verified that removing or retaining the interannual variability in the anomalies does not qualitatively affect the predictability estimates” in their study. The reanalysis fields used for verification (1999–2010) are obtained from the same data sources used to calculate and remove climatology.

3. Results

Most of the results presented below consist of verification of MJO index (RMM or VPM) against observations in terms of bivariate correlation, root-mean-square error (RMSE), ensemble spread, and climatologies of MJO index amplitude and phase (approximate geographic location). Because these indices are constructed from combinations of multiple fields, the time evolutions of the spatial RMSEs for the three variables that contribute at least 40% to the leading pair of EOFs—U850 and U200 for RMM; VP200 for VPM (Ventrice et al. 2013)—are also shown.

a. Skill, error, and spread

1) SINGLE-MODEL ENSEMBLE MEANS AND SPREAD

A common benchmark to evaluate MJO forecast skill is the lead time at which the bivariate correlation coefficient [Eq. (1) of Lin et al. (2008)] of the RMM (or VPM) index falls below 0.5. Bivariate RMSE [Eq. (2) of Lin et al. (2008)]—a measure of forecast error—will be $\sqrt{2}$ for a climatological forecast of RMM/VPM index (first two principal components both zero); this value serves as another benchmark for evaluating model performance in forecasting RMM/VPM.

Figure 1 shows bivariate correlations and RMSEs for the SMSPE means of FIM-CGF, FIM-AGF, FIM-SAS, and CFSv2 for both RMM and VPM; ensemble spreads [calculated using the right-hand side of Eq. (15) in Fortin et al. (2014)] are also shown in Figs. 1c and 1d. The correlations are lower and RMSEs are higher (viz., worse) for the individual ensemble members aggregated by initialization time (i.e., all 0600 UTC Wednesday members) than for the four-member ensemble means; this is a common result (e.g., Xiang et al. 2015) and thus not shown here.

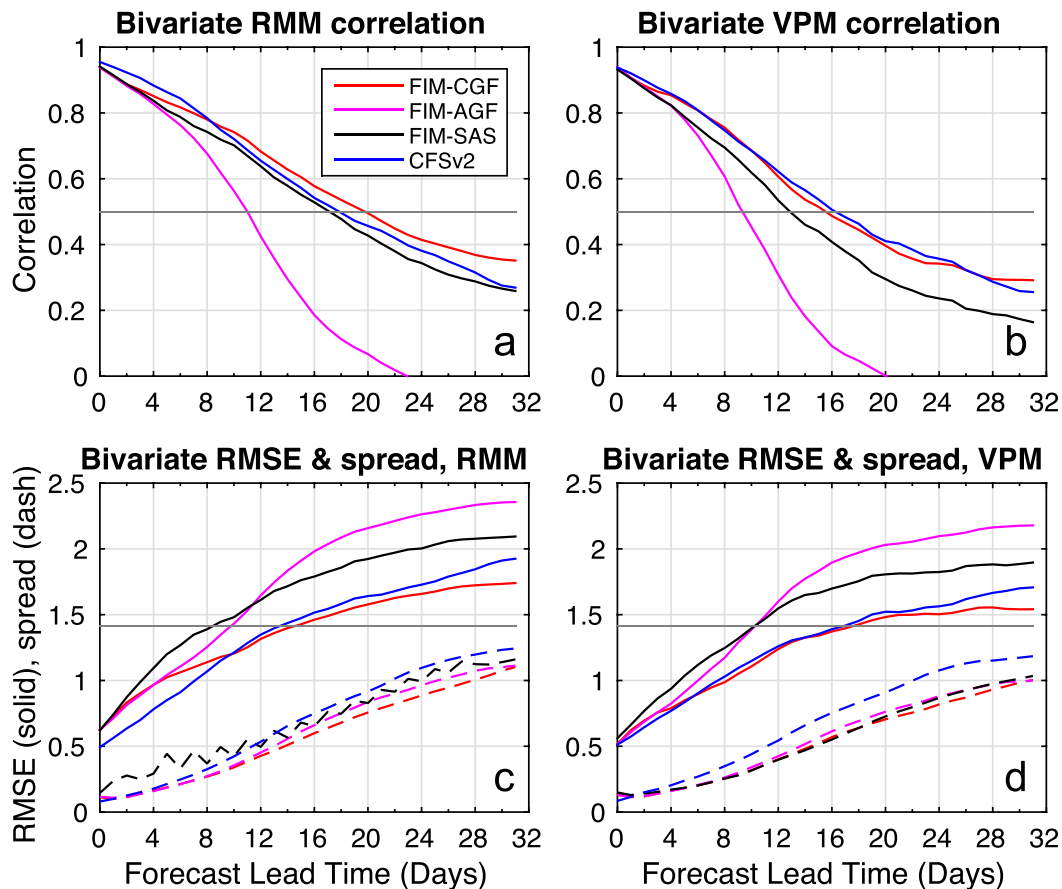


FIG. 1. Performance of four-member single-model ensemble hindcasts as functions of forecast lead time. (a) Bivariate correlation for the RMM index from each model's ensemble mean. (b) As in (a), but for the VPM index. (c),(d) As in (a),(b), but for ensemble mean RMSE (solid) and ensemble spread (dashed). As shown in the legend in (a), the red, magenta, black, and blue curves in all panels represent the FIM-CGF, FIM-AGF, FIM-SAS, and CFSv2 hindcasts, respectively. In (a),(b), bivariate correlations above the gray line at 0.5 are considered to be skillful; similarly, in (c),(d), RMSE below the gray line at $\sqrt{2}$ indicates model performance better than a climatological forecast.

There are some interesting differences between RMM and VPM. For example, although all SMSPEs have more skill (via correlation) in RMM than VPM (Figs. 1a vs 1b), there is also more error in RMM than VPM (Figs. 1c vs 1d). This result by itself is important, because one cannot make *unqualified* claims about *MJO* predictive skill or predictability; specifically, stating that “model X has skillful prediction of (the RMM index) out to Y days” is preferred over stating that “model X has skillful *MJO* prediction out to Y days.” Accordingly, Fig. 2 shows the temporal extent to which forecasts of (modified) RMM and VPM from each of the four SMSPE means are skillful (correlation threshold of 0.5), and have errors less than those expected by climatology. Overall, FIM-CGF and CFSv2 are comparable in their ability to simulate RMM and VPM (although FIM-CGF has noticeably higher RMM correlations than CFSv2

after ~ 10 days), while FIM-SAS is worse—especially in terms of RMSE. Unsurprisingly, FIM-AGF has substantially lower skill and higher RMSE than any of the coupled models; thus, FIM-AGF will not be considered in any further analysis. Finally, all SMSPEs are underdispersive: the spreads are lower than the RMSEs.

2) MULTIMODEL, MULTIPHYSICS ENSEMBLE MEANS AND SPREAD

One unique aspect of this study is the evaluation of MPE and MME hindcasts of *MJO* indices over a large dataset spanning 12 years. Bivariate correlations, RMSEs, and ensemble spreads of these MPE/MME hindcasts are shown in Fig. 3. Specifically, there are five ensemble means: “CFSv2 + FIM-CGF” and “CFSv2 + FIM-SAS” are eight-member MMEs; “FIM-CGF + FIM-SAS” is an eight-member MPE. “All 3 equal” and

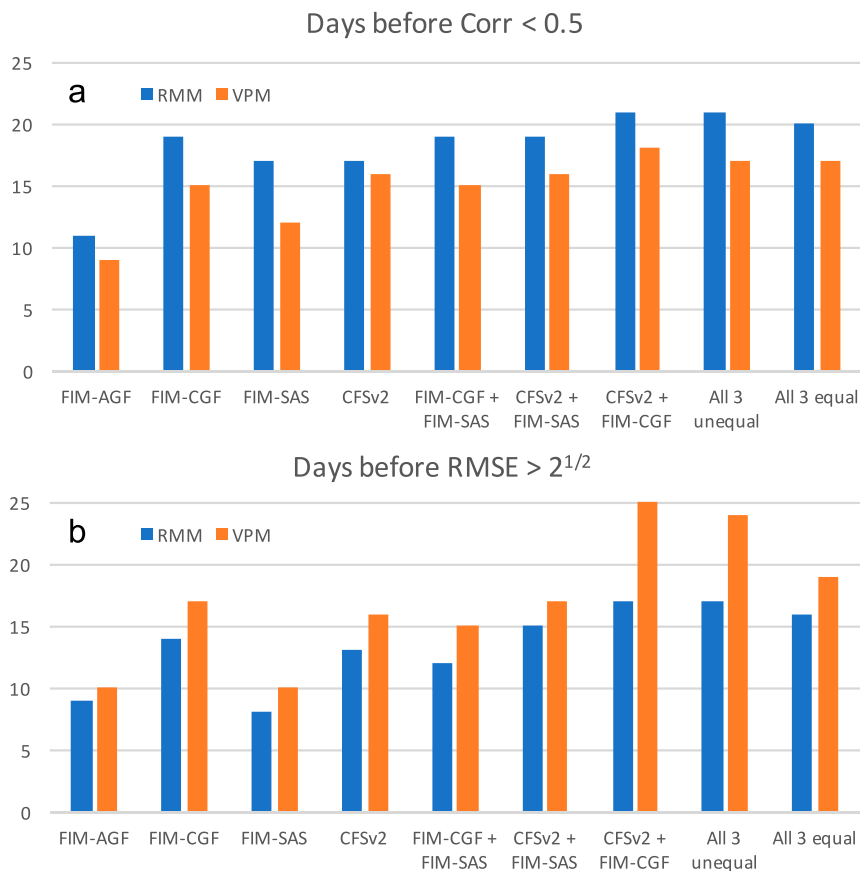


FIG. 2. Summary of MJO performance from the various single-model, multiphysics, and multimodel ensemble hindcast experiments, as measured by the average lead time in days before (a) the bivariate correlation becomes less than 0.5 and (b) RMSE becomes greater than $\sqrt{2}$ for both the RMM (blue) and VPM (orange) indices.

“All 3 unequal” are both 12-member MMEs of FIM-CGF, FIM-SAS, and CFSv2; the difference is in the weighting given to each individual model: All 3 equal assigns a weight of 1/3 to each SMSPE mean, whereas All 3 unequal assigns weights of 0.4 to the ensemble means of FIM-CGF and CFSv2 but a weight of 0.2 to the ensemble mean of FIM-SAS. The addition of a 12-member ensemble with unequal weights is motivated by the finding that FIM-SAS performs worse than either FIM-CGF or CFSv2, and thus including FIM-SAS in an MME may degrade performance.⁵ Indeed, that is exactly the case: while difficult to see graphically, the 8-member CFSv2 + FIM-CGF has higher correlations

and lower RMSEs than either 12-member ensemble (for both RMM and VPM); see also Fig. 2. The negative impact of adding FIM-SAS to the 8-member CFSv2 + FIM-CGF is further illustrated by comparing the two 12-member ensembles: All 3 equal performs worse than All 3 unequal, which performs worse than CFSv2 + FIM-CGF (which can be thought of as a 12-member ensemble but with zero weight applied to the four FIM-SAS members).

Nevertheless, MPEs and MMEs can add skill and reduce error over their SMSPE counterparts, as evidenced by comparing Fig. 3 with Fig. 1, and looking at Fig. 2. Also, while still underdispersive, the MPEs/MMEs have a better spread/RMSE relationship (closer to unity) than the SMSPEs (cf. Figs. 3c,d and 1c,d). One caveat, though, is that ensemble size has not yet been accounted for—an issue that will now be addressed.

Figure 4 focuses only on the eight-member CFSv2 + FIM-CGF MME; Figs. 5 and 6 are similar but for CFSv2 + FIM-SAS and FIM-CGF + FIM-SAS,

⁵The specific weights were chosen somewhat arbitrarily, but with two objectives in mind: first, equally weight the two similar “good” models (FIM-CGF and CFSv2); second, the “poor” model (FIM-SAS) should have less weight than either of the good models alone.

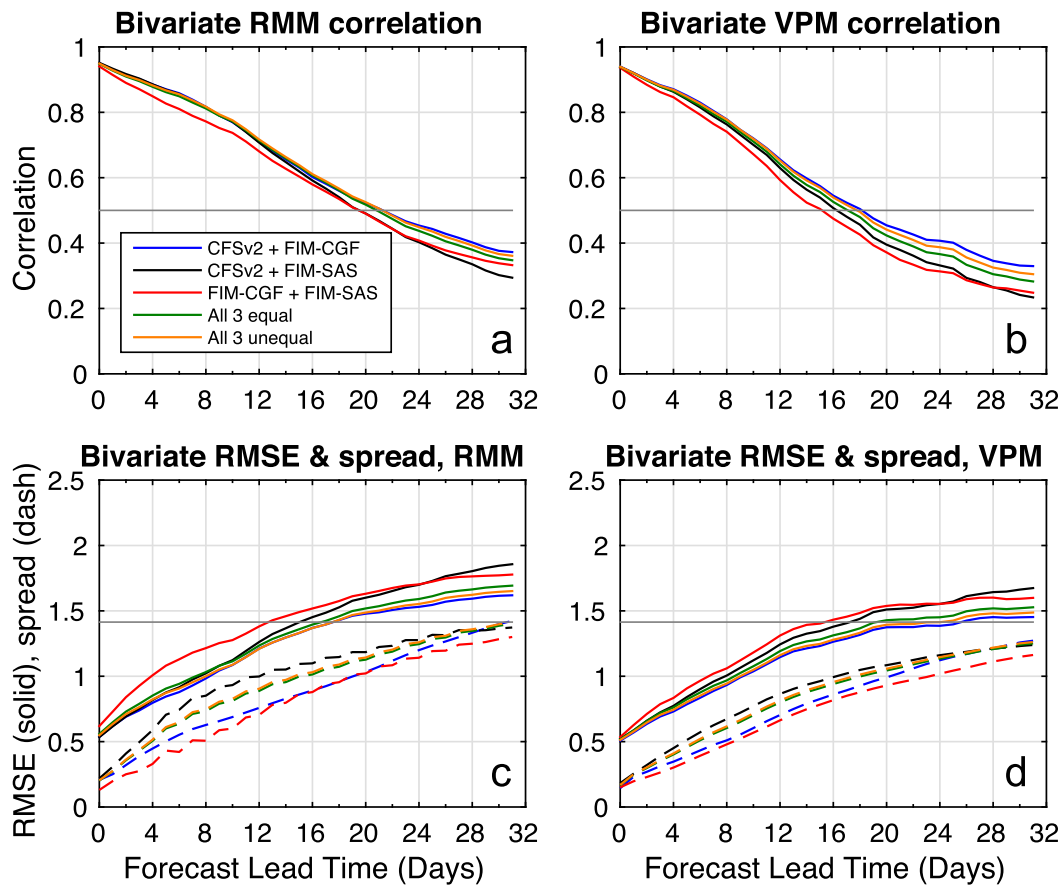


FIG. 3. As in Fig. 1, but for the 8-member multimodel CFSv2 + FIM-CGF ensemble (blue); 8-member multimodel CFSv2 + FIM-SAS ensemble (black); 8-member multi-physics FIM-CGF + FIM-SAS ensemble (red); 12-member multimodel, equally weighted CFSv2 + FIM-CGF + FIM-SAS ensemble (green); and 12-member multimodel, unequally-weighted (see text) CFSv2 + FIM-CGF + FIM-SAS ensemble (orange).

respectively. In all three figures, the green lines represent the eight-member ensemble mean (or spread), and are identical to the correspondingly-labeled eight-member ensembles in Fig. 3. Moreover, the red and blue lines represent the two component four-member SMSPEs (means and spreads), corresponding to the similarly labeled lines in Fig. 1. But what makes Figs. 4–6 unique are the gray lines, which correspond to the 8 choose 4 = 70 possible combinations of MPEs (or MMEs) with a size of 4 (viz., the same size as the SMSPEs). Note that two of these 70 combinations are actually the SMSPEs themselves; the other 68 combinations are truly MPEs (or MMEs).

As already mentioned, the best MME combines CFSv2 with FIM-CGF. In terms of bivariate correlation, RMSE, and ensemble spread, the benefit of the eight-member CFSv2 + FIM-CGF over the two (component) four-member SMSPEs is clear for both RMM and VPM (Fig. 4). In fact, the bivariate correlations of this MME are higher than both component

SMSPEs for all lead times. But this benefit cannot be explained solely by the increase in ensemble size [a result also found by Evans et al. (2000) and Candille (2009)]; nearly all of the four-member MMEs have higher correlations and lower RMSEs than the two (component) four-member SMSPEs. Moreover, the four-member multimodel spread is always higher than the corresponding SMSPE spreads. It is conjectured that the similar, but perhaps complementary, skills and errors of CFSv2 and FIM-CGF has a “synergistic” impact on their combined MME, allowing for better forecasts than the two component SMSPEs.

When the noticeably worse FIM-SAS is combined with CFSv2 (Fig. 5) into an MME, the synergistic effect is not as evident. In terms of RMM, the eight-member MME beats both four-member component SMSPEs with higher correlation and lower RMSE. But for VPM, adding FIM-SAS to CFSv2 yields no improvement, and many of the four-member MMEs perform worse than CFSv2 alone. Only the spread is reliably increased by

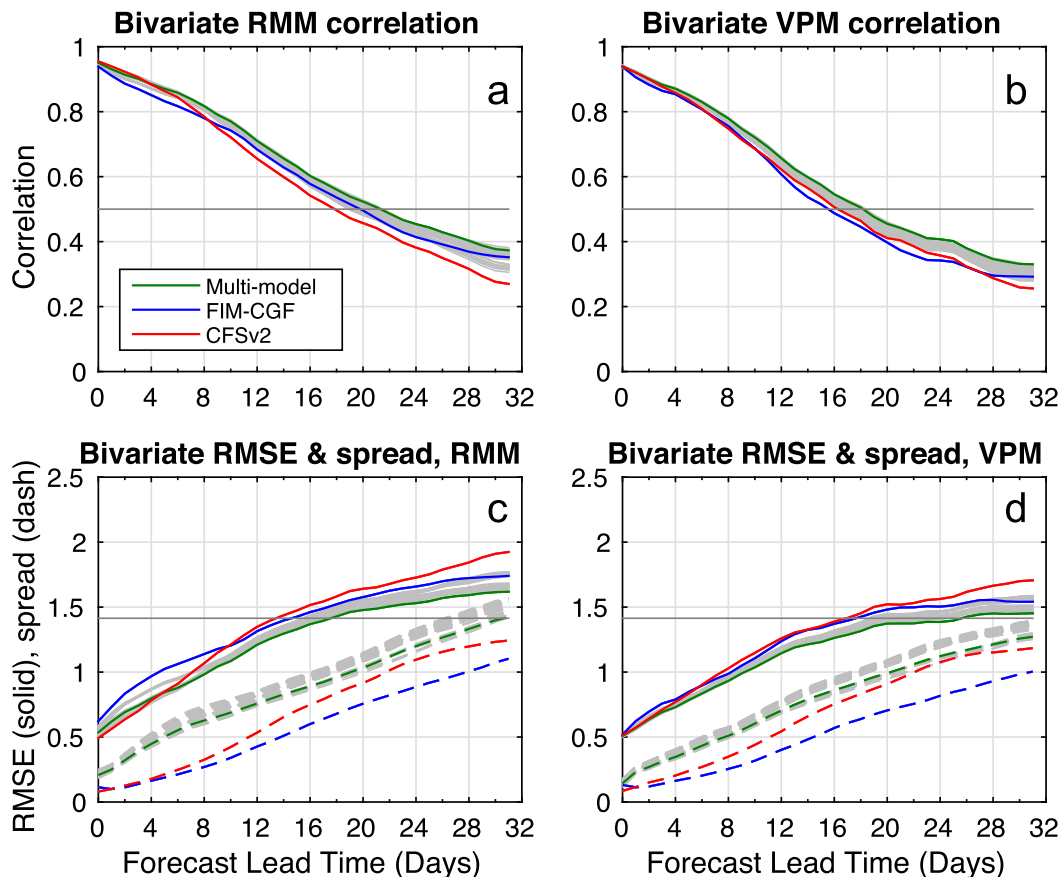


FIG. 4. As in Figs. 1 and 3, but the green curve represents the full eight-member multimodel CFSv2 + FIM-CGF ensemble (identical to the blue curves in Fig. 3), the blue curve represents the four-member single-model FIM-CGF ensemble (identical to the red curves in Fig. 1), and the red curve represents the four-member single-model CFSv2 ensemble (identical to the blue curves in Fig. 1). Additionally, the gray curves represent the 80 combinations of 4-member, multimodel CFSv2 + FIM-CGF ensembles.

the MMEs, which is no surprise given the large differences between the two component SMSPEs.

Finally, the *multiphysics* FIM-CGF + FIM-SAS (Fig. 6) show more clearly the negative influence of FIM-SAS: for both RMM and VPM, the eight-member MPE mean has lower skill and higher RMSE than the four-member FIM-CGF. The reasons for the MPE being worse than one of the component SMSPEs alone are not clear, but it is conjectured that the cause is a combination of (i) FIM-SAS being much worse than FIM-CGF and (ii) the lack of model (physics) diversity—here, the dynamic core is identical and only the deep convection parameterization is changed—compared with the *multimodel* CFSv2 + FIM-SAS.

So far, hindcast performance has only been evaluated in terms of bivariate correlation, RMSE, and spread for RMM and VPM. These metrics do not give any sense of the model climatology (viz., how the simulated

distributions of RMM/VPM amplitude and phase compare with observations). This issue will be addressed in section 3b. Based on the above results, operational centers looking to use ensembles to improve MJO (index) forecasts should consider using an MME approach in which the individual models perform comparably.

b. Observed and model climatologies of amplitude and phase for RMM and VPM

1) RMM AND VPM AMPLITUDE

By design, both RMM and VPM have a long-term average amplitude of unity; however, there may be considerable deviation from that climatological value at any given time. This is illustrated in Figs. 7a and 7b for the distributions of observed RMM and VPM amplitude, respectively. The observed distributions are calculated using every day from January 1999 to December

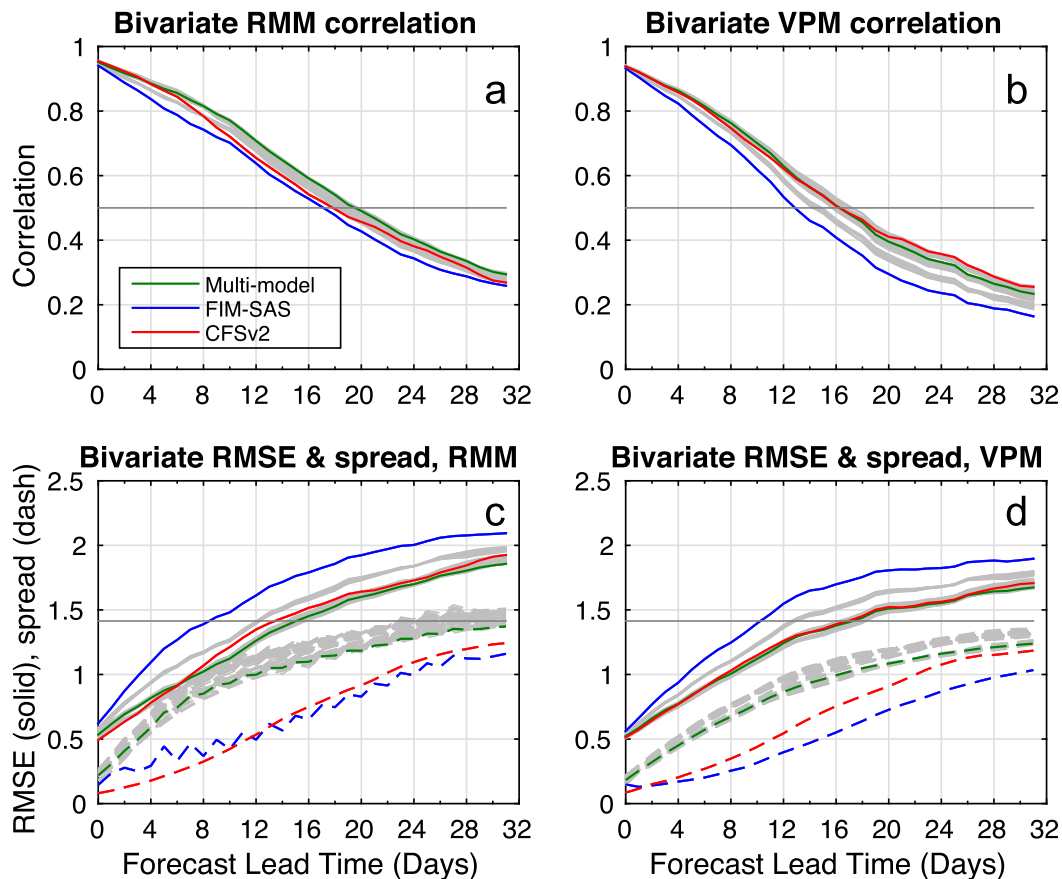


FIG. 5. As in Fig. 4, but for the full eight-member multimodel CFSv2 + FIM-SAS ensemble (green), the four-member single-model FIM-SAS ensemble (blue), and the four-member single-model CFSv2 ensemble (red).

2010, and plotted as a function of “forecast lead time” (by simply using the same values for all 32 days, as evidenced by the vertical strips of color) for ease of comparison with the models (e.g., Figs. 7c–h, see below). While both RMM and VPM climatological amplitudes appear to be centered close to unity, the former has a heavier tail toward high amplitude whereas the latter has a heavier tail toward low amplitude.

A perfect model—when averaged over a sufficiently large number of forecast cases, as could be argued here—would have a distribution of MJO index amplitude independent of forecast lead time and equal to that of observations (i.e., the model would faithfully represent the observed climatology of the index without any “drift” toward an incorrect model climatology). Therefore, changes in the model climatology (as a function of forecast lead time) would indicate an imperfect model—likely due to errors in the governing equations and physical parameterizations rather than errors in the initial conditions, given the long time scales of these hindcasts.

Looking at Figs. 7c–h, it is clear from an ensemble-mean perspective that none of the three coupled SMSPE hindcasts (CFSv2, Figs. 7c,d; FIM-CGF, Figs. 7e,f; and FIM-SAS, Figs. 7g,h) sustain a consistent model climatology similar to observations. Perhaps the best model in this regard is CFSv2: except for a possible weak bias in RMM during the last 2 weeks of the forecast, the distributions of both MJO indices appear very similar visually to their observational counterparts. Both FIM-iHYCOM hindcasts show, to varying degrees, a significant strengthening trend in RMM and VPM amplitudes during the first week of the forecast, followed by a weakening trend in the last two weeks of the forecast (to the point where FIM-CGF has an obvious weak bias for both indices). The FIM-iHYCOM hindcasts—especially FIM-SAS—have broader amplitude distributions than observations. FIM-SAS also has a persistent strong bias throughout the forecast. It is conjectured that FIM-iHYCOM’s lack of cycled data assimilation (initial conditions are from CFSR), combined with use of different convective

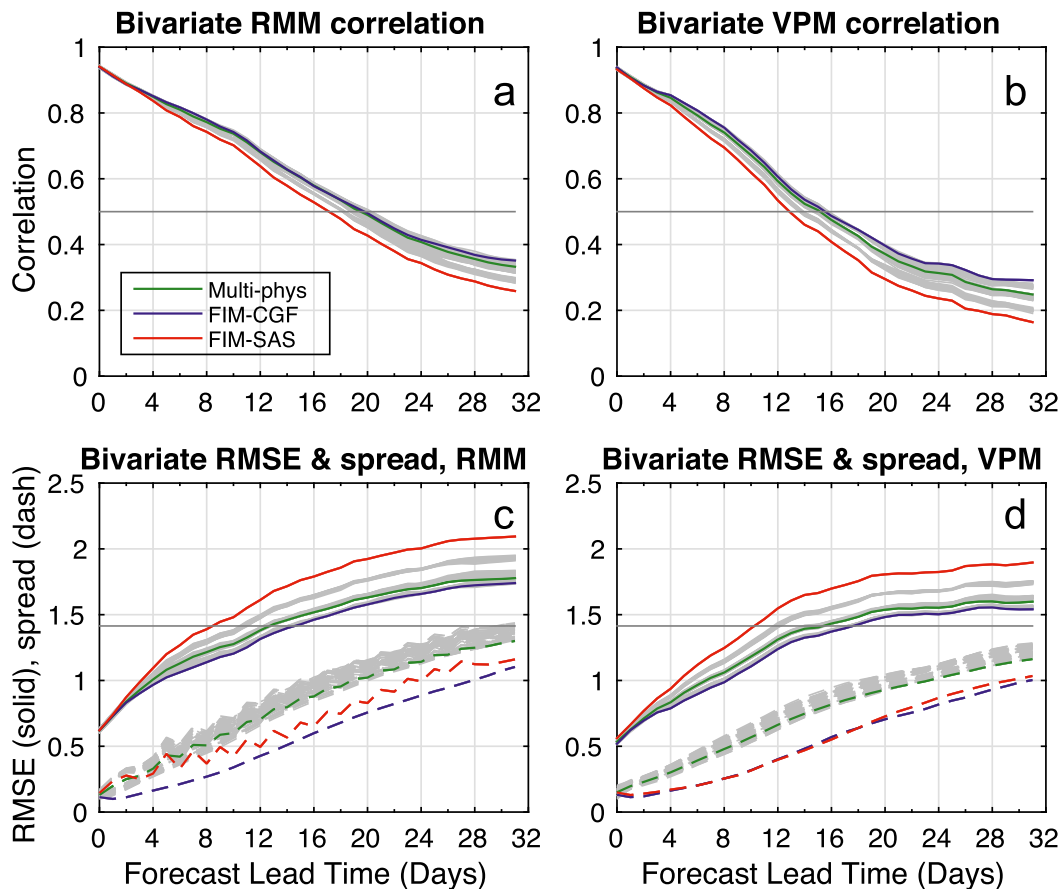


FIG. 6. As in Figs. 4 and 5, but for the full eight-member multiphysics FIM-CGF + FIM-SAS ensemble (green), the four-member single model FIM-CGF ensemble (blue), and the four-member single-model FIM-SAS ensemble (red).

schemes, yields an adjustment period (better associated with model error than initial condition error) to a different convective balance in the FIM-iHYCOM forecasts. Specifically, CFSR analyses are based on a version of the SAS deep convection scheme not used in any of the FIM-iHYCOM hindcasts. Thus, the initial adjustment in FIM-iHYCOM, especially for FIM-CGF, toward a different convective balance is evident in the distribution shifts (as a function of time) in Figs. 7e–h.

As shown earlier, using MMEs has the potential to yield better RMM and VPM forecasts (as measured by bivariate correlation and RMSE) than the individual component models. But how do the MMEs fare in terms of the climatology of index amplitude? Perhaps not surprisingly, it turns out that the MPE/MMEs simply yield patterns that tend to smooth out the features of the individual component models (i.e., the effects of averaging are obvious) and are thus not shown here. So, while MMEs can be beneficial for improving RMM/VPM forecast skill (Figs. 3 and 4), they should not be

expected to yield a model climatology better than the individual component models.

2) RMM AND VPM PHASE

A phase (angle) that describes the instantaneous “position” or “location” of the MJO signal can be derived from the first two principal components of an index like RMM or VPM. Because of the nature of the EOFs, these phases are not equal in geographic extent: for example, RMM phases 8 and 1 (1/4 of RMM phases) cover over half the globe (the entire Western Hemisphere, plus Africa), whereas RMM phases 4 and 5—also 1/4 of RMM phases—cover the Maritime Continent (which spans only a small fraction of longitudes). Moreover, Ventrice et al. (2013; p. 4204) argue that the geographic locations of the VPM phases “should nearly match those for the RMM” phases; this greatly simplifies the analyses of RMM and VPM phase distributions.

The observed frequencies of RMM and VPM phases over the 12-yr period are shown in Figs. 8a and 8b.

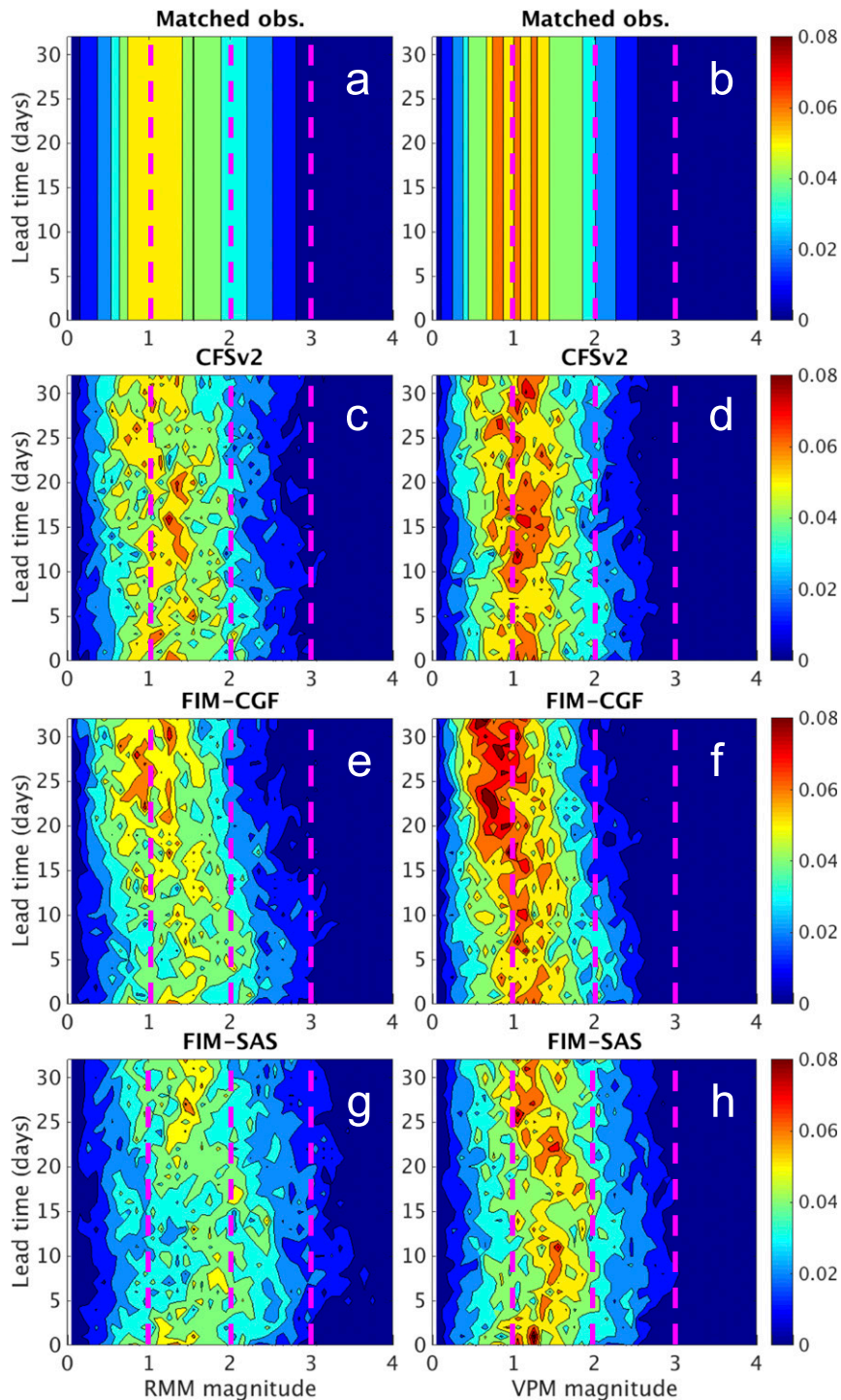


FIG. 7. Frequency plots of MJO magnitude (abscissa) as measured by the (left) RMM index and (right) VPM index, as functions of forecast lead time on the ordinate. (a),(b) Observed frequencies; as described in the text, the observed distributions are independent of (lead) time. (c),(d) Frequency plots for the 4-member CFSv2 ensemble mean (aggregated over 625 cases) as a function of forecast lead time. (e),(f) As in (c),(d), but for the four-member FIM-CGF ensemble mean. (g),(h) As in (e),(f) but for the four-member FIM-SAS ensemble mean. The pink dashed lines denote RMM/VPM magnitudes of 1, 2, and 3.

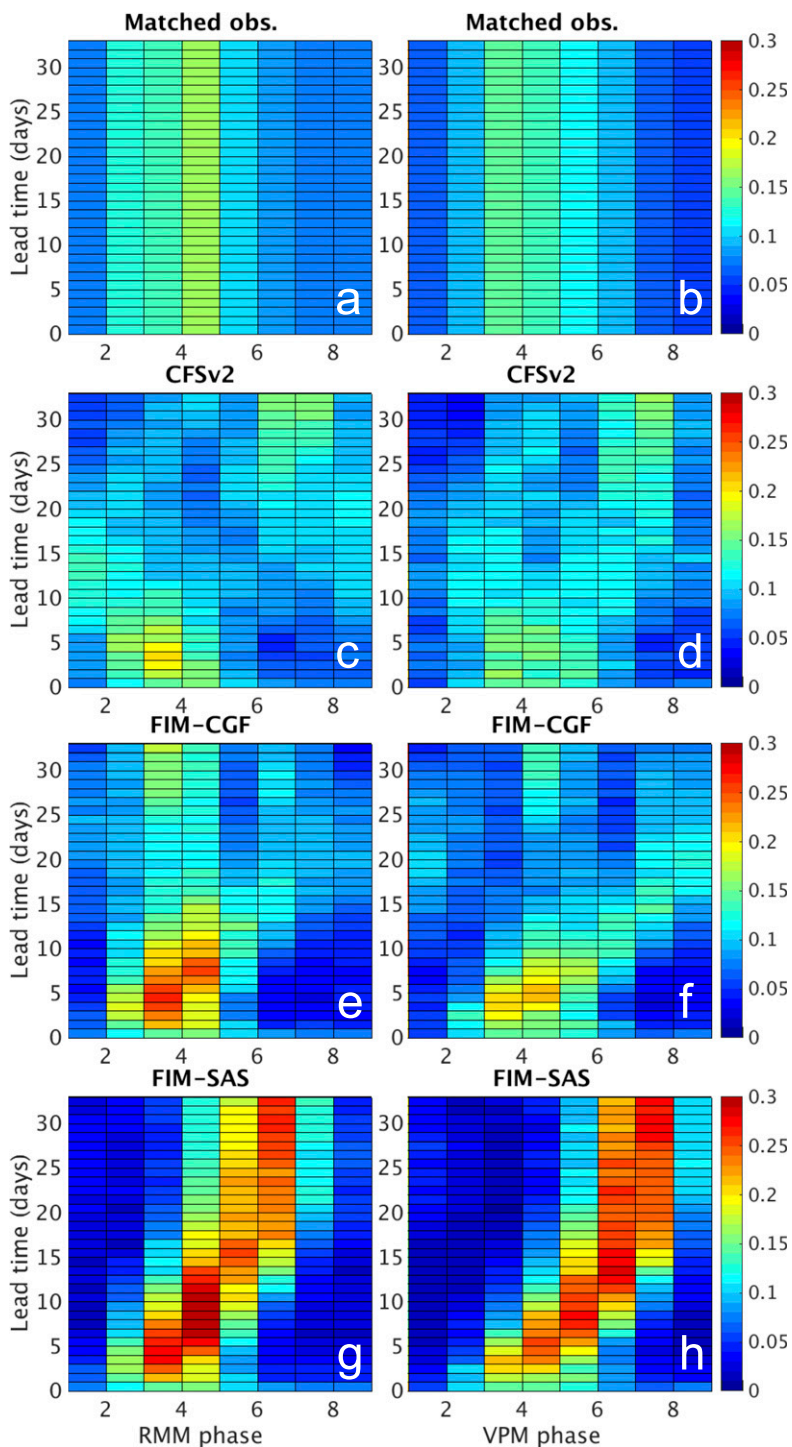


FIG. 8. As in Fig. 7, but for frequency distribution of the MJO phase. As described in the text, all cases where (left) RMM and (right) VPM magnitude falls below a threshold of 0.75 are placed into a “phase 0.” Although the temporal evolution of the frequency of phase 0 cases is not shown in this figure, phase 0 cases are included in the calculation of relative frequency. Thus, while phases 1–8 do not graphically sum to 1 (for a given lead time, i.e., horizontal row), the relative frequencies of phases 1–8 are still evident.

Clearly, the preferred phases are 2–4 for RMM and 3–5 for VPM (i.e., over the Indian Ocean and Maritime Continent). While both RMM and VPM traditionally only have eight phases (as defined by an octant in 2D Cartesian space), small changes in either principal component (i.e., RMM1 or RMM2) can yield major changes in the calculated phase when the amplitude is near zero. To avoid this problem, one can bin (into a “phase 0”) cases in which the index amplitude falls below a certain threshold. Often, this threshold is unity (e.g., [Ventrice et al. 2013](#)); here, however, the threshold is set to 0.75 to capture more marginally strong MJO cases in the weak-biased FIM-CGF ([Figs. 7e,f](#)). As expected by using the lower threshold, fewer than 50% of observational cases fall into phase 0: approximately 15%–20% for RMM and ~25% for VPM (not shown). The observational frequencies here are similar to those shown in the top panel of [Fig. 6 of Ventrice et al. \(2013\)](#): in their study, phase 5 is most favored for both RMM and VPM.⁶

Similar to the distributions of MJO index amplitude as a function of forecast lead time, the model hindcasts have phase distributions that are in poor agreement with observations (cf. [Figs. 8a,b and 8c–h](#)). All models have varying degrees of phase shift with forecast lead time. CFSv2 has a preference for phases 6 and 7 (for both indices) during weeks 3–4. FIM-CGF looks better than CFSv2 in terms of phase distribution, with a sustained preference for phases 3–4 for RMM (phase 4 for VPM), although in FIM-CGF both indices do show a secondary peak toward the end of the forecast (phase 6 for RMM, phases 7–8 for VPM). And as with the amplitude distribution, the phase distribution is worst for FIM-SAS: the preferred phases of RMM and VPM settle at 5–6 and 6–7, respectively, beyond week 2. Clearly, none of these models can sustain a climatology that agrees with observations, instead adjusting to their own, different, internal climatology. Finally, like with amplitude, the phase distributions for the MPE/MMEs just smooth out the features of the individual SMSPEs and do not yield patterns closer to observations (not shown).

c. RMSEs for hindcasts of U850, U200, and VP200

Understanding that RMM and VPM are constructed from a combination of multiple fields, it is worthwhile to place the index results shown above in the context of the

raw fields. A thorough analysis of how the hindcasts simulate the raw fields (i.e., stratified by index phase/amplitude, seasonal cycle, etc.) is beyond the scope of this paper. Instead, maps of RMSE in the tropical band (15°S–15°N) aggregated over all 625 cases as a function of forecast lead time are shown for all three coupled SMSPE mean hindcasts, for the variables contributing at least 40% to the leading EOF pair of either RMM or VPM ([Ventrice et al. 2013](#)): U850, U200, and VP200 ([Figs. 9, 10, and 11](#), respectively). FIM-SAS exhibits the highest RMSEs (especially in U200 and VP200)—not just in the areas of active MJO convection, but throughout the tropics—which is consistent with the poor performance of this SMSPE in MJO index forecasts and underscores the importance of realistically simulating the entire tropical band in order to get better forecasts of RMM/VPM. In contrast, FIM-CGF has the lowest RMSEs in each of the three fields, especially in weeks 3 and 4; CFSv2 generally falls in between the two FIM-iHYCOM hindcasts.

[Figure 9](#) shows that all three coupled-model hindcasts have the largest U850 RMSEs over the Indian Ocean and far western Pacific, with a surprising local RMSE minimum over the Maritime Continent. CFSv2 actually has U850 RMSEs that are as large as those in FIM-SAS and substantially higher than FIM-CGF. This could explain why FIM-CGF has noticeably higher bivariate correlations (after ~10 days) than CFSv2 for RMM ([Fig. 1a](#)) but not VPM ([Fig. 1b](#)): U850 contributes over 40% to RMM but only ~25% to VPM. Thus, larger U850 errors would impact RMM more than VPM. For U200, which has similar fractional contributions to both RMM and VPM as U850, the lowest RMSEs are again in FIM-CGF ([Fig. 10](#)), although CFSv2 is closer to FIM-CGF than to FIM-SAS in terms of geographic distribution but roughly halfway between the two FIM-iHYCOM hindcasts in terms of overall error magnitude. The highest U200 RMSEs in all three models are not over the areas of active MJO convection, but rather over the oceans in the Western Hemisphere. The VP200 RMSEs ([Fig. 11](#)) show similar geographic distributions for CFSv2 and FIM-CGF, with the latter SMSPE having somewhat smaller magnitudes. FIM-SAS has substantially higher VP200 RMSEs throughout the tropics, with the highest errors over equatorial Africa.

4. Discussion

This study, which evaluates the performance of two global models (and combinations thereof) in terms of their ability to forecast and represent MJO indices, raises some important issues that should be considered in future research regarding MJO simulation in global models.

⁶ Differences in the methodologies used here and by [Ventrice et al. \(2013\)](#) are responsible for the slight discrepancies in preferred MJO phases; namely, time range and frequency of observational data, different thresholds for strong versus weak MJOs, and the impact of retaining/removing the preceding 120-day mean.

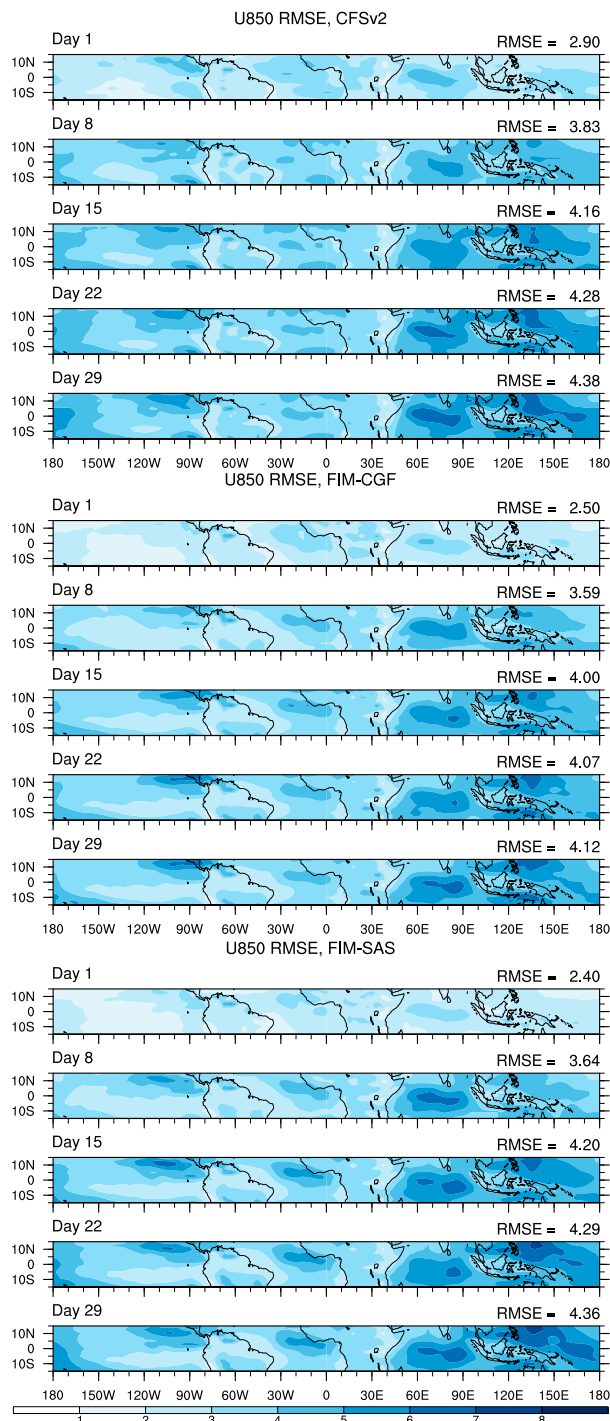


FIG. 9. RMSEs (verified against reanalysis) of U850 (ms^{-1}) aggregated over all 625 cases for each of the three coupled SMSPE means [(top) CFSv2, (middle) FIM-CGF, (bottom) FIM-SAS] as a function of forecast lead time (days 1, 8, 15, 22, and 29 from top to bottom for each SMSPE), evaluated at each grid point in the tropical band from 15°S to 15°N. Also shown in each panel is the RMSE averaged over all grid points in the tropical band.

The first issue is that of fully coupling the atmospheric model to an ocean model. The importance of air–sea interaction and ocean coupling on the MJO has already been documented extensively (e.g., Woolnough et al. 2007; Fu et al. 2013; Seo et al. 2014; Tseng et al. 2015). Such a result was replicated here (Figs. 1 and 2): specifically, turning off ocean coupling resulted in a loss of 8 days of forecast skill for the RMM index (6 days for VPM), and RMSE exceeding what would be expected from climatology 5 days earlier for RMM (7 days earlier for VPM). Moreover, noticeable differences between the coupled and uncoupled runs begin to appear at lead times of ~ 3 –5 days, which is clearly within the realm of numerical weather prediction models. Thus, medium-range forecasts made from atmosphere-only models (such as the GFS, which runs out to 16 days) will have much more difficulty forecasting the MJO (index) and any associated extratropical interactions—as evidenced by the ~ 14 -day RMM skill in a version of the GFS ensemble [Fig. 13a of Hamill and Kiladis (2014)]. This would suggest that global atmospheric weather models should be coupled with an ocean model from the start, if forecasting to extended (intraseasonal) lead times is an eventual goal. It should be noted that coupling iHYCOM to FIM only increases computational expense by $\sim 20\%$.

The second issue raised by this study is that different methods to quantitatively identify the MJO—here, modified versions of the RMM and VPM indices—can have substantially different skill and error scores within the same model. This is most evident in Figs. 1–6, where models (including MMEs) generally have higher skill—but also higher RMSE—for RMM than for VPM. Therefore, care was taken in this study—and should be taken in future studies—to avoid stating the ability of a model to predict *the* MJO (rather than *a specific MJO metric/index* like RMM or VPM) with skill out to so many days. Also, while some studies (Lin et al. 2008; Rashid et al. 2011; Wang et al. 2014) have found that bivariate correlation drops to 0.5 at around the same lead time as RMSE becomes $\sqrt{2}$, Xiang et al. (2015, their Fig. 2) show that it takes ~ 4 –6 days longer for correlation to reach 0.5 than for RMSE to exceed $\sqrt{2}$; this discrepancy is also evident in our results for RMM (Fig. 2).

Purely from the perspective of bivariate correlations of RMM/VPM, determining if one model has statistically significantly higher skill than another model is not straightforward. The bivariate correlation formula given by Eq. (1) of Lin et al. (2008) is a special case of the anomaly correlation coefficient (ACC) [e.g., Eq. (1) of Jones et al. (2004, 2015)] in which only *two* points (RMM1, RMM2) are considered rather than the *thousands* of

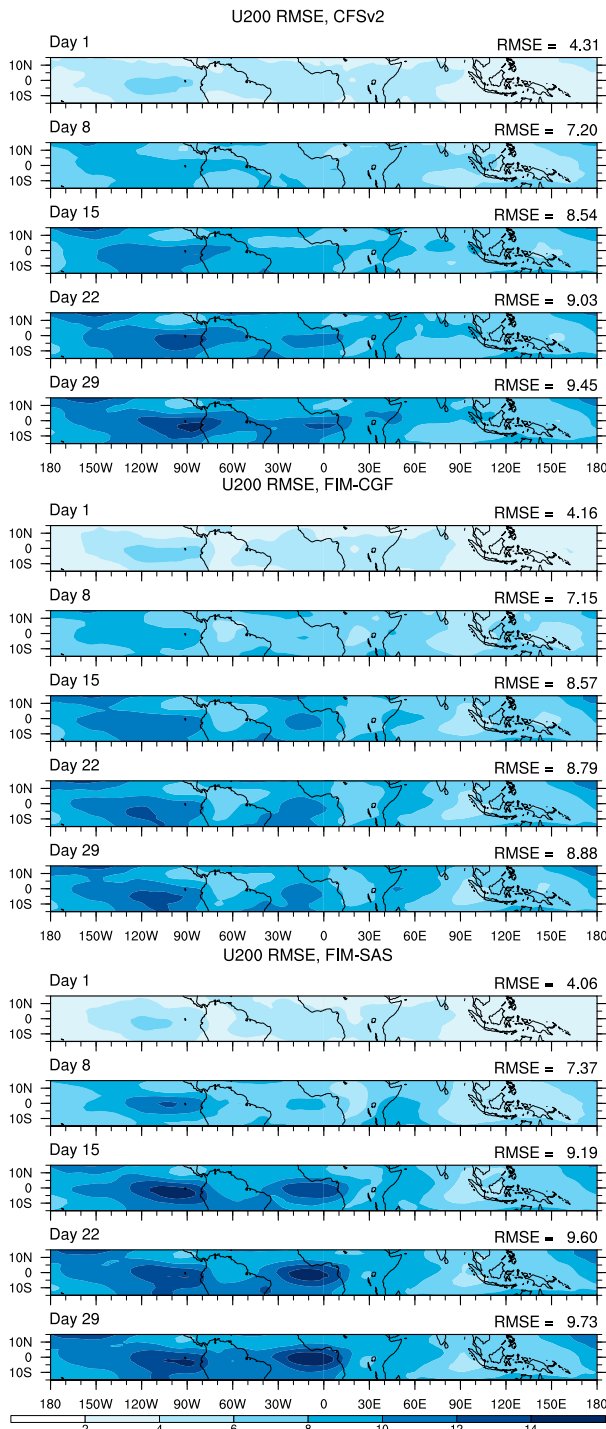


FIG. 10. As in Fig. 9, but for U200 ($m s^{-1}$).

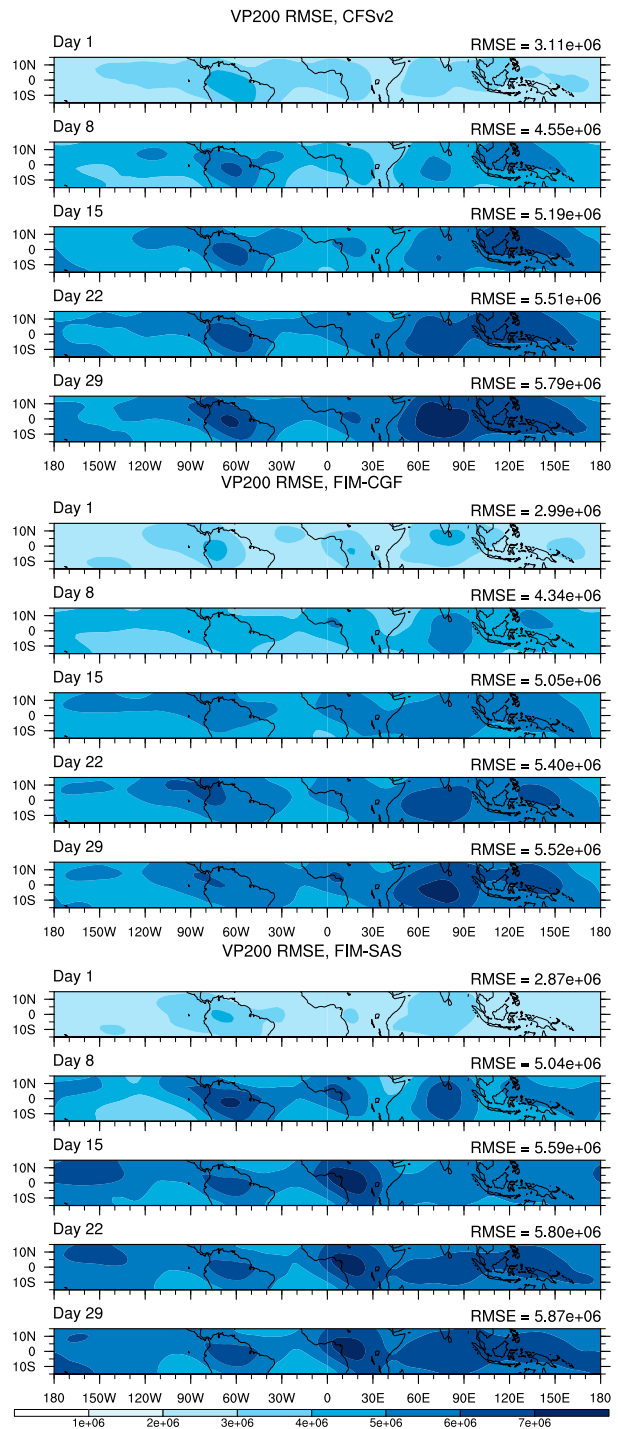


FIG. 11. As in Fig. 10, but for VP200 ($m^2 s^{-1}$).

grid points typically used to calculate ACC over a large area. Equations (2) and (3) of Jones et al. (2004, 2015) outline a strategy to determine whether two ACCs have a statistically significant difference. In their Eq. (3), the variance (for each forecast lead time) of the

Fisher-transformed ACC [their Eq. (2)] falls in the denominator of their test statistic Z . Thus, increased variance—but all else equal (mean and sample size)—decreases Z and the likelihood that the two correlations will have statistically significant differences. As expected,

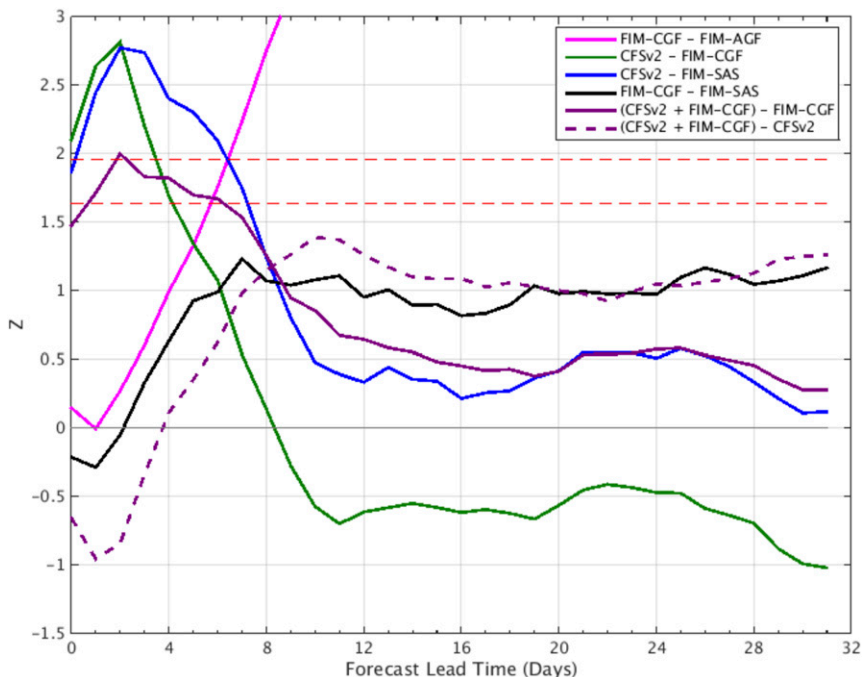


FIG. 12. Test statistic Z as a function of forecast lead time following Eqs. (2) and (3) of Jones et al. (2004, 2015) for differences in bivariate RMM correlation between different SMSPEs and the CFSv2 + FIM-CGF MME. For each curve labeled in the legend, a positive Z means that the first model labeled has a higher correlation than the second model labeled; the red dashed lines of $Z = 1.64$ and $Z = 1.96$ correspond to significance levels of 90% and 95%, respectively.

only having two points in an ACC yields extremely large variance, even for short lead times (not shown). Figure 12 shows Z for the differences in bivariate RMM correlation between select hindcast ensemble means. The first model has statistically significantly higher bivariate RMM correlation at 90% and 95% confidence when Z exceeds 1.64 and 1.96, respectively. Unsurprisingly, FIM-CGF has statistically significantly higher skill than the uncoupled FIM-AGF after ~6 days (magenta curve in Fig. 12). Otherwise, there are only statistically significant differences in the first week, with CFSv2 beating FIM-SAS and FIM-CGF. These differences—and the CFSv2 + FIM-CGF MME improving upon the FIM-CGF SMSPE at the 90% level—are likely related to the adjustment period of the FIM simulations, which do not have their own cycling data assimilation. There are never statistically significant differences in VPM correlation amongst the coupled models (not shown).

It is also interesting to note that while FIM-CGF [which uses a variant of the Grell and Freitas (2014) deep convection scheme] performs similarly in terms of RMM/VPM to CFSv2 (which uses the SAS deep convection scheme), FIM-SAS (which uses a slightly different variant of SAS than CFSv2) performs worse than

the other two models. Furthermore, RMSEs of the raw fields that go into RMM/VPM calculations (Figs. 9–11) reveal that FIM-CGF has smaller errors than both FIM-SAS and CFSv2. Therefore, in this study, GF is better suited than SAS deep convection for MJO representation and forecasting (using the RMM and VPM indices), although obviously similar tests are needed within other global models.

Finally, using MPEs or MMEs has the potential to improve skill and reduce error in RMM/VPM forecasts compared to every one of the individual component models. Here, an MME that combined CFSv2 with FIM-CGF gave higher bivariate correlations (more skill), lower RMSEs, and more spread than either CFSv2 or FIM-CGF alone—even when accounting for ensemble size (Fig. 4). This “synergistic” impact of an MME has been demonstrated for the RMM index (Fu et al. 2013; Zhang et al. 2013) and in studies of other fields/phenomena (e.g., Evans et al. 2000; Candille 2009). However, it was also shown here that adding FIM-SAS to CFSv2, FIM-CGF, or the CFSv2 + FIM-CGF MME yielded minimal to no improvement, or even degraded forecast performance. The most likely explanation is that the noticeably worse FIM-SAS SMSPE forecasts (cf. SMSPE forecasts from both FIM-CGF and CFSv2) simply

acted as a hindrance when combined with FIM-CGF and/or CFSv2. A second possibility that cannot be eliminated is that FIM-SAS, when combined with CFSv2 (atmospheric physics parameterizations very similar; different dynamical cores for both the atmosphere and ocean) or FIM-CGF (same dynamical core and physics parameterizations except for deep convection) does not provide enough ensemble diversity. This is particularly the case for the FIM-CGF + FIM-SAS MPE, in which the eight-member mean always underperforms the four-member FIM-CGF mean. Most of the issues raised in this discussion require further investigation.

5. Conclusions

This study is the first to evaluate extensively—over a common 12-yr period with over 600 forecast cases—ensembles of multiple global atmosphere–ocean coupled models in their individual, *and combined*, ability to predict and represent common MJO indices. Specifically, hindcasts from the coupled atmosphere–ocean FIM-iHYCOM modeling system were compared with those from CFSv2. In agreement with numerous other studies, it was found that a fully coupled atmosphere–ocean modeling system much better represents and predicts two MJO indices (RMM and VPM) than an atmosphere-only model; namely, coupling iHYCOM to FIM extended the skillful prediction of RMM and VPM by 8 and 6 days, respectively (Fig. 2).

It is interesting to note that RMM has higher RMSE than VPM, but also higher correlations (Figs. 1–3). One would expect a priori that VPM would have higher correlations than RMM, because VPM only accounts for the large-scale circulation whereas RMM includes the much-less-predictable OLR (i.e., convection). Further investigation of this surprising result is beyond the scope of this paper, but further underscores that the differences in RMM and VPM—particularly when it comes to model evaluation—serve as a caution signal that no single index will best represent the broadband, multiscale nature of the MJO.

Two versions of FIM-iHYCOM (FIM-CGF and FIM-SAS) were run to test the impact of deep convective parameterization. FIM-CGF had lower RMSEs and higher bivariate correlations—for both RMM and VPM—than did FIM-SAS; however, FIM-CGF performed comparably with CFSv2 (which used a slightly different version of SAS) in terms of VPM, and ~2 days better in terms of RMM (Figs. 1 and 2). But in terms of the raw fields used to calculate RMM and VPM, FIM-CGF consistently had the lowest RMSEs.

When FIM-CGF and CFSv2 were combined into a multimodel ensemble (MME), there were improvements over *both* component models in skill (bivariate correlation), RMSE, and ensemble spread (Figs. 3 and 4)—even when accounting for the increase in ensemble size. It is conjectured that the CFSv2 + FIM-CGF MME beats both CFSv2 and FIM-CGF single-model, single-physics ensembles because the two component models are of similar skill and have sufficient model diversity (different deep convective parameterizations and different dynamic cores for both atmosphere and ocean). But when FIM-SAS was added to either FIM-CGF, CFSv2, or CFSv2 + FIM-CGF, model performance did not improve—and in some cases, actually degraded. This is most likely due to the worse performance of FIM-SAS by itself, although it is possible that a relative lack of model diversity provided by FIM-SAS (same dynamic core as FIM-CGF, and similar convective parameterization as CFSv2) could also be contributing (Zhang et al. 2013).

The distributions of RMM and VPM amplitude and phase were evaluated as functions of forecast lead time for the models and compared with the observed climatological distributions. For CFSv2, the amplitude distributions for both indices were fairly steady with forecast lead time and close to observations. Both FIM-iHYCOM runs exhibited changes/adjustments in RMM/VPM amplitude with forecast lead time (likely because FIM-iHYCOM was initialized by CFSR, rather than its own model-consistent analysis), and featured notable biases. None of the models had a steady distribution of RMM/VPM phase that matched observations; instead, there was a preference to phases east of those favored by climatology with increasing lead time. The distributions of RMM/VPM amplitude and phase in the MMEs essentially represented smoothing/averaging of the individual models, with no obvious improvement toward observed climatology (not shown).

Future work will leverage the finding that the combined CFSv2 + FIM-CGF MME provides better forecasts of both RMM and VPM than either single model alone. Results from this study will be combined with those from a 30-km FIM-CGF hindcast and used as a benchmark against which future coupled models will be compared. Specifically, this includes experimental versions of NOAA's Next-Generation Global Prediction System using the Finite Volume on a Cubed Sphere (FV3; e.g., Putman and Lin 2007) atmospheric dynamical core. A version of FV3 was used in the coupled-model MJO study of Xiang et al. (2015), but with initial conditions nudged from the current spectral GFS. NOAA will develop a cycling data assimilation system for FV3 as part of its scheduled operational implementation.

This will provide FV3-native analyses (initial conditions) and eliminate the problem of adjusting to a different convective balance caused by using analyses from another modeling system [as was the case for FIM-iHYCOM here, and in Xiang et al. (2015)].

Acknowledgments. The lead author was supported in part by a Visiting Postdoctoral Fellowship through the Cooperative Institute for Research in Environmental Sciences at the University of Colorado Boulder. This project was also supported by NOAA/OAR funding for week 3–4 forecast improvement and the Earth System Prediction Capability program. We thank George Kiladis for providing an internal review of this manuscript, and Juliana Dias for helpful comments. The authors acknowledge the NOAA/Research and Development High Performance Computing Program for providing computing and storage resources that have contributed to the research results reported within this paper (<http://rdhpcs.noaa.gov>). We also thank Paul Roundy and two anonymous reviewers for their excellent constructive comments that improved this manuscript. Matthew Wheeler provided observed RMM data (with the interannual component retained).

REFERENCES

- Benjamin, S., S. Sun, G. Grell, B. Green, R. Bleck, and H. Li, 2017: Improved subseasonal prediction with advanced coupled models including the 30km FIM-HYCOM coupled model. *European Geosciences Union General Assembly 2017*, Vienna, Austria, EGU, EGU2017–11097. [Available online at <http://meetingorganizer.copernicus.org/EGU2017/EGU2017-11097.pdf>.]
- Berner, J., S.-Y. Ha, J. P. Hacker, A. Fournier, and C. Snyder, 2011: Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. *Mon. Wea. Rev.*, **139**, 1972–1995, doi:10.1175/2010MWR3595.1.
- Bleck, R., 2002: An oceanic general circulation model framed in hybrid isopycnic-Cartesian coordinates. *Ocean Modell.*, **4**, 55–88, doi:10.1016/S1463-5003(01)00012-9.
- , and Coauthors, 2015: A vertically flow-following icosahedral grid model for medium-range and seasonal prediction. Part I: Model description. *Mon. Wea. Rev.*, **143**, 2386–2403, doi:10.1175/MWR-D-14-00300.1.
- Bouttier, F., B. Vié, O. Nuisser, and L. Raynaud, 2012: Impact of stochastic physics in a convection-permitting ensemble. *Mon. Wea. Rev.*, **140**, 3706–3721, doi:10.1175/MWR-D-12-00031.1.
- Boyle, J. S., S. A. Klein, D. D. Lucas, H.-Y. Ma, J. Tannahill, and S. Xie, 2015: The parametric sensitivity of CAM5's MJO. *J. Geophys. Res. Atmos.*, **120**, 1424–1444, doi:10.1002/2014JD022507.
- Candille, G., 2009: The multiensemble approach: The NAEFS example. *Mon. Wea. Rev.*, **137**, 1655–1665, doi:10.1175/2008MWR2682.1.
- DeMott, C. A., N. P. Klingaman, and S. J. Woolnough, 2015: Atmosphere–ocean coupled processes in the Madden–Julian oscillation. *Rev. Geophys.*, **53**, 1099–1154, doi:10.1002/2014RG000478.
- Evans, R. E., M. S. J. Harrison, R. J. Graham, and K. R. Mylne, 2000: Joint medium-range ensembles from the Met. Office and ECMWF systems. *Mon. Wea. Rev.*, **128**, 3104–3127, doi:10.1175/1520-0493(2000)128<3104:JMREFT>2.0.CO;2.
- Fortin, V., M. Abaza, F. Anctil, and R. Turcotte, 2014: Why should ensemble spread match the RMSE of the ensemble mean? *J. Hydrometeor.*, **15**, 1708–1713, doi:10.1175/JHM-D-14-0008.1.
- Fu, X., J.-Y. Lee, P.-C. Hsu, H. Taniguchi, B. Wang, W. Wang, and S. Weaver, 2013: Multi-model MJO forecasting during DYNAMO/CINDY period. *Climate Dyn.*, **41**, 1067–1081, doi:10.1007/s00382-013-1859-9.
- Gottschalck, J., and Coauthors, 2010: A framework for assessing operational Madden–Julian oscillation forecasts: A CLIVAR MJO working group project. *Bull. Amer. Meteor. Soc.*, **91**, 1247–1258, doi:10.1175/2010BAMS2816.1.
- Grell, G. A., and S. R. Freitas, 2014: A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling. *Atmos. Chem. Phys.*, **14**, 5233–5250, doi:10.5194/acp-14-5233-2014.
- Grimit, E. P., and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting*, **17**, 192–205, doi:10.1175/1520-0434(2002)017<0192:IROAMS>2.0.CO;2.
- Hamill, T. M., and G. N. Kiladis, 2014: Skill of the MJO and Northern Hemisphere blocking in GEFS medium-range reforecasts. *Mon. Wea. Rev.*, **142**, 868–885, doi:10.1175/MWR-D-13-00199.1.
- Han, J., and H.-L. Pan, 2011: Revision of convection and vertical diffusion schemes in the NCEP Global Forecast System. *Wea. Forecasting*, **26**, 520–533, doi:10.1175/WAF-D-10-05038.1.
- Holloway, C. E., S. J. Woolnough, and G. M. S. Lister, 2013: The effects of explicit versus parameterized convection on the MJO in a large-domain high-resolution tropical case study. Part I: Characterization of large-scale organization and propagation. *J. Atmos. Sci.*, **70**, 1342–1369, doi:10.1175/JAS-D-12-0227.1.
- Jones, C., D. E. Waliser, K. M. Lau, and W. Stern, 2004: The Madden–Julian Oscillation and its impact on Northern Hemisphere winter predictability. *Mon. Wea. Rev.*, **132**, 1462–1471, doi:10.1175/1520-0493(2004)132<1462:TMOAJI>2.0.CO;2.
- , A. Hazra, and L. M. V. Carvalho, 2015: The Madden–Julian oscillation and boreal winter forecast skill: An analysis of NCEP CFSv2 reforecasts. *J. Climate*, **28**, 6297–6307, doi:10.1175/JCLI-D-15-0149.1.
- Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**, 437–471, doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.
- Kikuchi, K., B. Wang, and Y. Kajikawa, 2012: Bimodal representation of the tropical intraseasonal oscillation. *Climate Dyn.*, **38**, 1989–2000, doi:10.1007/s00382-011-1159-1.
- Kiladis, G. N., J. Dias, K. H. Straub, M. C. Wheeler, S. N. Tulich, K. Kikuchi, K. M. Weickmann, and M. J. Ventrice, 2014: A comparison of OLR and circulation-based indices for tracking the MJO. *Mon. Wea. Rev.*, **142**, 1697–1715, doi:10.1175/MWR-D-13-00301.1.
- Kim, H.-M., P. J. Webster, V. E. Toma, and D. Kim, 2014: Predictability and prediction skill of the MJO in two operational forecasting systems. *J. Climate*, **27**, 5364–5378, doi:10.1175/JCLI-D-13-00480.1.
- Liebmann, B., and C. A. Smith, 1996: Description of a complete (interpolated) outgoing longwave radiation dataset. *Bull. Amer. Meteor. Soc.*, **77**, 1275–1277.

- Lin, H., G. Brunet, and J. Derome, 2008: Forecast skill of the Madden–Julian Oscillation in two Canadian atmospheric models. *Mon. Wea. Rev.*, **136**, 4130–4149, doi:[10.1175/2008MWR2459.1](https://doi.org/10.1175/2008MWR2459.1).
- Liu, P., B. Wang, K. R. Sperber, T. Li, and G. A. Meehl, 2005: MJO in the NCAR CAM2 with the Tiedtke convective scheme. *J. Climate*, **18**, 3007–3020, doi:[10.1175/JCLI3458.1](https://doi.org/10.1175/JCLI3458.1).
- , Q. Zhang, C. Zhang, Y. Zhu, M. Khairoutdinov, H.-M. Kim, C. Schumacher, and M. Zhang, 2016: A revised real-time multivariate MJO index. *Mon. Wea. Rev.*, **144**, 627–642, doi:[10.1175/MWR-D-15-0237.1](https://doi.org/10.1175/MWR-D-15-0237.1).
- Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307, doi:[10.3402/tellusa.v21i3.10086](https://doi.org/10.3402/tellusa.v21i3.10086).
- Madden, R. A., and P. R. Julian, 1971: Detection of a 40–50 day oscillation in the zonal wind in the tropical Pacific. *J. Atmos. Sci.*, **28**, 702–708, doi:[10.1175/1520-0469\(1971\)028<0702:DOADOI>2.0.CO;2](https://doi.org/10.1175/1520-0469(1971)028<0702:DOADOI>2.0.CO;2).
- , and —, 1972: Description of global-scale circulation cells in the tropics with a 40–50 day period. *J. Atmos. Sci.*, **29**, 1109–1123, doi:[10.1175/1520-0469\(1972\)029<1109:DOGSCC>2.0.CO;2](https://doi.org/10.1175/1520-0469(1972)029<1109:DOGSCC>2.0.CO;2).
- Neena, J. M., J. Y. Lee, D. Waliser, B. Wang, and X. Jiang, 2014: Predictability of the Madden–Julian oscillation in the Intra-seasonal Variability Hindcast Experiment (ISVHE). *J. Climate*, **27**, 4531–4543, doi:[10.1175/JCLI-D-13-00624.1](https://doi.org/10.1175/JCLI-D-13-00624.1).
- NOAA, 2016: S2S Prediction Task Force. NOAA, accessed 31 January 2017. [Available online at <http://cpo.noaa.gov/ClimateDivisions/EarthSystemScienceandModeling/ModelingAnalysisPredictionsandProjections/MAPPTaskForces/S2SPredictionTaskForce.aspx>].
- Putman, W. M., and S.-J. Lin, 2007: Finite-volume transport on various cubed-sphere grids. *J. Comput. Phys.*, **227**, 55–78, doi:[10.1016/j.jcp.2007.07.022](https://doi.org/10.1016/j.jcp.2007.07.022).
- Rashid, H. A., H. H. Hendon, M. C. Wheeler, and O. Alves, 2011: Prediction of the Madden–Julian oscillation with the POAMA dynamical prediction system. *Climate Dyn.*, **36**, 649–661, doi:[10.1007/s00382-010-0754-x](https://doi.org/10.1007/s00382-010-0754-x).
- Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91**, 1015–1057, doi:[10.1175/2010BAMS3001.1](https://doi.org/10.1175/2010BAMS3001.1).
- , and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, doi:[10.1175/JCLI-D-12-00823.1](https://doi.org/10.1175/JCLI-D-12-00823.1).
- Seo, H., A. C. Subramanian, A. J. Miller, and N. R. Cavanaugh, 2014: Coupled impacts of the diurnal cycle of sea surface temperature on the Madden–Julian oscillation. *J. Climate*, **27**, 8422–8443, doi:[10.1175/JCLI-D-14-00141.1](https://doi.org/10.1175/JCLI-D-14-00141.1).
- Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, doi:[10.1175/1520-0477\(1993\)074<2317:EFANTG>2.0.CO;2](https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2).
- , and —, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297–3319, doi:[10.1175/1520-0493\(1997\)125<3297:EFANAT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2).
- Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting*, **8**, 379–398, doi:[10.1175/1520-0434\(1993\)008<0379:OEPATN>2.0.CO;2](https://doi.org/10.1175/1520-0434(1993)008<0379:OEPATN>2.0.CO;2).
- Tseng, W.-L., B.-J. Tsuang, N. S. Keenlyside, H.-H. Hsu, and C.-Y. Tu, 2015: Resolving the upper-ocean warm layer improves the simulation of the Madden–Julian oscillation. *Climate Dyn.*, **44**, 1487–1503, doi:[10.1007/s00382-014-2315-1](https://doi.org/10.1007/s00382-014-2315-1).
- Ventrone, M. J., M. C. Wheeler, H. H. Hendon, C. J. Schreck III, C. D. Thorncroft, and G. N. Kiladis, 2013: A modified multivariate Madden–Julian oscillation index using velocity potential. *Mon. Wea. Rev.*, **141**, 4197–4210, doi:[10.1175/MWR-D-12-00327.1](https://doi.org/10.1175/MWR-D-12-00327.1).
- Vitart, F., 2014: Evolution of ECMWF sub-seasonal forecast skill scores. *Quart. J. Roy. Meteor. Soc.*, **140**, 1889–1899, doi:[10.1002/qj.2256](https://doi.org/10.1002/qj.2256).
- Wang, W., M.-P. Hung, S. J. Weaver, A. Kumar, and X. Fu, 2014: MJO prediction in the NCEP Climate Forecast System version 2. *Climate Dyn.*, **42**, 2509–2520, doi:[10.1007/s00382-013-1806-9](https://doi.org/10.1007/s00382-013-1806-9).
- Wheeler, M. C., and H. H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, **132**, 1917–1932, doi:[10.1175/1520-0493\(2004\)132<1917:AARMMI>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<1917:AARMMI>2.0.CO;2).
- Woolnough, S. J., F. Vitart, and M. A. Balmaseda, 2007: The role of the ocean in the Madden–Julian oscillation: Implications for MJO prediction. *Quart. J. Roy. Meteor. Soc.*, **133**, 117–128, doi:[10.1002/qj.4](https://doi.org/10.1002/qj.4).
- Xiang, B., M. Zhao, X. Jiang, S.-J. Lin, T. Li, X. Fu, and G. Vecchi, 2015: The 3–4 week MJO prediction skill in a GFDL coupled model. *J. Climate*, **28**, 5351–5364, doi:[10.1175/JCLI-D-15-0102.1](https://doi.org/10.1175/JCLI-D-15-0102.1).
- Zhang, C., 2005: Madden–Julian oscillation. *Rev. Geophys.*, **43**, RG2003, doi:[10.1029/2004RG000158](https://doi.org/10.1029/2004RG000158).
- , 2013: Madden–Julian oscillation: Bridging weather and climate. *Bull. Amer. Meteor. Soc.*, **94**, 1849–1870, doi:[10.1175/BAMS-D-12-00026.1](https://doi.org/10.1175/BAMS-D-12-00026.1).
- , J. Gottschalck, E. D. Maloney, M. W. Moncrieff, F. Vitart, D. E. Waliser, B. Wang, and M. C. Wheeler, 2013: Cracking the MJO nut. *Geophys. Res. Lett.*, **40**, 1223–1230, doi:[10.1002/grl.50244](https://doi.org/10.1002/grl.50244).
- Zhou, L., R. B. Neale, M. Jochum, and R. Murtugudde, 2012: Improved Madden–Julian oscillations with improved physics: The impact of modified convection parameterizations. *J. Climate*, **25**, 1116–1136, doi:[10.1175/2011JCLI4059.1](https://doi.org/10.1175/2011JCLI4059.1).