Subseasonal Forecasting with an Icosahedral, Vertically Quasi-Lagrangian Coupled Model. Part II: Probabilistic and Deterministic Forecast Skill

Shan Sun and Benjamin W. Green
University of Colorado Boulder, Cooperative Institute for Research in Environmental Sciences, NOAA/OAR/ESRL/Global Systems Division, Boulder, Colorado
Rainer Bleck
University of Colorado Boulder, Cooperative Institute for Research in Environmental Sciences, NOAA/OAR/ESRL/Global Systems Division, Boulder, Colorado, and NASA Goddard Institute for Space Studies, New York, New York
Stanley G. Benjamin
NOAA/OAR/ESRL/Global Systems Division, 325 Broadway, Boulder, Colorado

Submitted to *Monthly Weather Review* for publication as an article
January 4, 2018
Revised, April 2, 2018

Corresponding author: Shan Sun, shan.sun@noaa.gov

Abstract

Subseasonal forecast skill of the global hydrostatic atmospheric Flow-following Icosahedral Model FIM coupled to an icosahedral-grid version of the Hybrid Coordinate Ocean Model HYCOM is evaluated through 32-day predictions initialized weekly using a 4-member time-lagged ensemble over the 16-year period 1999-2014. Systematic biases in forecasts by the coupled system, referred to as FIM-iHYCOM, are described in a companion paper (Part I). This present study (Part II) assesses probabilistic and deterministic model skill for predictions of surface temperature, precipitation, and 500 hPa geopotential height in different seasons at different lead times ranging from 1 to 4 weeks. The coupled model appears to have reasonable agreement with reanalysis in terms of simulated weekly variability in sea surface temperatures, except in extratropical regions because the ocean model cannot explicitly resolve eddies there. This study also describes the ability of the model to simulate mid-latitude tropospheric blocking frequency, Madden-Julian oscillation patterns, and sudden stratospheric warming events – all of which have been shown to be relevant on subseasonal timescales. The metrics used here indicate that the subseasonal forecast skill of the model is comparable to that of several operational models, including the National Oceanic and Atmospheric Administration's (NOAA's) operational Climate Forecast System version 2 and the European Centre for Medium-Range Weather Forecasting model. Therefore, FIM-iHYCOM – as a participant in NOAA's Subseasonal Experiment – is expected to add value to multi-model ensemble forecasts produced through this effort.

1. Introduction

Considerable effort is presently invested in providing and improving subseasonal forecasts (~2 weeks to 2 months) because of the importance of this timescale in many sectors of society (Brunet et al. 2010; WMO 2015; NAS 2016). Several operational centers have been issuing subseasonal-to-seasonal (S2S) forecasts for more than a decade (Vitart 2004; Saha et al. 2014; Kim et al. 2014; Lin et al. 2016; Wheeler et al. 2017). Although subseasonal timescales fall outside the theoretical deterministic predictability limit of ~2 weeks in the mid-latitudes (Lorenz 1969), there is sufficient evidence that potential sources of subseasonal predictability are seen in relatively long-lived flow configurations in the tropics (Charney and Shukla 1981), such as the Madden-Julian Oscillation (MJO, Waliser et al. 2003), and tropospheric blocking at mid-latitudes (Matsueda 2011). Major changes in the wintertime upper-stratospheric circulation associated with sudden stratospheric warmings (SSWs) likewise are being investigated as potential precursors of persistent tropospheric circulation anomalies at high latitudes (Baldwin and Dunkerton 1999; Shaw and Perlwitz 2013). In addition, capitalizing on relatively slow-varying processes involving sea-ice extent, soil moisture (Beljaars et al. 1996), snow cover (Walland and Simmonds 1997) and the ocean state, has the potential for improving S2S predictions. In particular, ocean heat content fluctuations, especially when associated with the El Niño Southern Oscillation (ENSO), have a long-lasting impact not only on the tropics, but globally due to atmospheric teleconnections (Hoerling and Kumar 2002).

Studies have also shown that ensemble-based probabilistic prediction on the S2S timescale can exhibit skill relative to forecasts based on persistence (e.g., Palmer 2002; Zhu et al. 2014; Smith et al. 2015; Lin et al. 2016). There has been widespread and still increasing use of ensemble forecasting to improve forecast skill at all timescales. Two approaches for generating an ensemble prediction system based on dynamical forecast models have been used:

1. Use of a single model ensemble whose members are diversified by perturbed or lagged initial conditions. Some such systems also include stochastic perturbations of model physics (e.g., Kalnay and Dalcher 1987; Palmer and Tibaldi 1988; Straus and Shukla 2000; Vitart 2014);
2. Use of several models in a multi-model ensemble (e.g., Kirshnamurti et al. 2003; Palmer et al. 2004; Kirtman et al. 2014; Li and Robertson 2015; Vigaud et al. 2017).

This pragmatic shift to probabilistic forecasting to reflect the inherently chaotic nature of atmospheric behavior has effectively improved skill in the subseasonal-to-interannual prediction range. At these timescales, multi-model ensemble forecasts have been found to be better than those based on any one single model, as shown in, e.g., the Development of a European Multimodel Ensemble System for Seasonal to Interannual Prediction (DEMETER; Palmer et al. 2004) and North American Multimodel Ensemble (NMME; Kirtman et al. 2014). Nevertheless, forecasts are still susceptible to errors from deficiencies in the treatment of spatially unresolved physical processes and finite-difference approximations in the model equations. To maximize the gain from ensemble methods, approach 2 above requires that attention be paid to model diversity in order to assure sufficient spread among ensemble members and not just reduction in overall bias. This makes models with diverse subgrid-scale physical parameterizations and innovative numerics attractive in multi-model ensembles.

Nevertheless, continuing development of individual models still appears to offer the best chance for improving subseasonal prediction, mainly by improving representation of the earth-system processes listed above. The goals of this article are to evaluate the subseasonal forecast skill of a new coupled model – the atmospheric Flow-following Icosahedral Model (FIM; Bleck et al. 2015) coupled to an icosahedral grid version of the Hybrid Coordinate Ocean Model (HYCOM; Bleck 2002) – in a multi-year set of 32-day hindcasts, with emphasis on seasonal and geographic skill variations. Given that the coupled model, referred to here as FIM-iHYCOM, has been described in detail in Part I (Sun et al. 2018), we concentrate in section 2 of the present article (Part II) on the details of the model climatology and bias removal. Weekly

variability of sea surface temperature (SST) between reanalysis and FIM-iHYCOM is also compared. Skill measures in predicting a number of relevant variables and phenomena are presented in section 3, followed by a discussion in section 4.

2. Model climatology and SST variability

a. Model climatology and bias removal

As discussed in Part I, FIM-iHYCOM hindcasts were carried out over a 16-year period initialized weekly with 4 time-lagged ensemble members for a total of 835 weeks, yielding 3340 simulations. The model output is interpolated onto a 1°x1° horizontal grid, and averaged to either daily or weekly means for various lead times, depending on different applications shown later.

To increase the sample size for a model climatology, FIM-iHYCOM was also initialized daily at 0000 UTC from 1 January 1999 through 31 December 2014[1], a total of 5844 simulations. For the purpose of bias correction, at each latitude-longitude point and for each lead day independently, the resulting hindcasts were first averaged over the number of available years, i.e., 4 years for February 29, 16 years for the remaining 365 days. This yielded averaged fields, henceforth referred to as raw model climatology, with dimensions 366 (initialization days of the year) x 32 (forecast lead days) x 181 x 360 (latitude-longitude points). Then, for each lead day at each latitude-longitude point, 10 passes of a 25-day low-pass numerical filter described in the Appendix were applied across the 366-day dimension. The number of filter passes was chosen empirically to balance noise removal necessitated by limited sampling and retention of physical signals.

The resulting lead-dependent daily model climatology is used to perform bias correction before assessing skill in most of the results presented below, with the following exception: the results from probabilistic forecasts do not employ the bias correction method described above because, as noted in Section 3a, model bias is accounted for implicitly.

For the purpose of bias-correcting subseasonal forecasts from NOAA's Climate Forecast System (CFS) version 2 (CFSv2, Saha et al. 2014), climatologies for CFS Reanalysis (CFSR, Saha et al. 2010), and for lead-dependent CFSv2, calculated following Zhang and van den Dool (2012), were downloaded from the National Centers for Environmental Information (NCEI). Note that CFS climatology is calculated over the period 1982-2010 due to data/computational constraints. It should be noted that Saha et al. (2014) recommend using a "split climatology" (p. 2199) based upon data from 1999 onward for both precipitation and SST in the tropical Pacific; however, this is not a concern here because none of our precipitation results include data equatorward of 20°N.

Figure 1 shows one example of how the low-pass filter (red curve) removes sampling noise from the raw FIM-iHYCOM model climatology (blue curve) of 2-m temperature, hereafter T2m, at a particular geographic location at a lead time of 1 day. These climatologies can also be compared with the corresponding CFSR climatology (black curve); this curve is noticeably smoother than the FIM-iHYCOM filtered climatology. The difference between the red and black curves represents the bias correction that is applied to FIM-iHYCOM hindcasts.

b. SST variability in FIM-iHYCOM

Model overview articles, especially those dealing with subseasonal or longer timescales, often go beyond an assessment of systematic biases and examine the ability of the model to reproduce observed temporal variability (e.g., Pegion and Kirtman 2008; Saha et al. 2014). Typically, such variability analyses are from multi-year model integrations that are then filtered to isolate the desired timescale. Because such a long integration of FIM-iHYCOM does not yet exist, the previously-described 16-year hindcast dataset, along with the filtered model climatology, is used. To calculate variability in

[1] Note that there is some overlap, namely, 0000 UTC every Wednesday, between this hindcast and the above mentioned 4-member time-lagged ensemble initialized weekly.

FIM-iHYCOM, the general approach of Saha et al. (2014; their Figure 3) is followed: we compute anomalies *with respect to model climatology* and then compute the variability of these anomaly fields.

Because the focus of Parts I and II is on subseasonal prediction, variability was computed from weekly data, using the following procedure. First, the filtered lead-dependent model climatology was removed from each of the four ensemble members to give *daily* model anomalies. Then, for each ensemble member separately, the daily anomalies were averaged over weeklong periods to obtain *weekly model anomalies*. Next, for a given target season (e.g., December through February, hereafter DJF; see Part I for details on "target" season) *and* a given forecast lead week (1-4) for each ensemble member separately, the variance of the weekly model anomalies was calculated. As an example for DJF, there are ~12 weeks per season and 14 DJF seasons, so ~168 cases for each lead week were included in the variance calculation of each of the four ensemble members. Finally, for plotting purposes, the *mean of the four ensemble member variances* was taken.

The resulting SST variances for FIM-iHYCOM were compared with SST variances from CFSR. The methodology to compute weekly variances from CFSR is similar to that described above, except that there are no ensembles and no dependence on forecast lead time. Instead, CFSR variances were simply computed based on the 1999-2014 period for those Wednesday-Tuesday weeks whose midweek day (Saturday) fell in the desired target season.

Figure 2 shows, for DJF, the CFSR weekly SST variance along with the *difference* in variance between FIM-iHYCOM ("FIMr1.1" as in Part I) and CFSR for forecast lead weeks 1-4. Looking at Figure 2a and ignoring sea-ice areas (e.g., north of Japan and in Hudson Bay), the areas with the largest weekly SST variability are in (i) extratropical boundary and gyre-scale drift currents such as the Gulf Stream and Kuroshio extensions; (ii) the Antarctic Circumpolar Current; and (iii) the central and eastern equatorial Pacific Ocean. For FIM-iHYCOM lead week 1 (Figure 2b), the largest differences between simulated and observed SST variability are in the strong current regions just mentioned. The reduced SST variability in these regions is a consequence of iHYCOM's grid spacing of ~60 km, too coarse to adequately simulate eddying and meandering in extratropical current systems. (Even when observed SST is sampled at coarse resolution, as is done in CFSR, the effects of meandering and eddying on the larger-scale ocean state are still captured to some extent.) The reduced SST variability in these regions (blue shading) appears to be nearly identical between weeks 3 and 4 (Figures 2d,e). More interesting, though, is that by week 4 FIM-iHYCOM has higher SST variability than CFSR in the equatorial central Pacific, and also in the equatorial Indian Ocean. An investigation of the potential impact of increased SST variability in FIM-iHYCOM (relative to CFSR, which may serve as a proxy for observed conditions) in the Indian Ocean on the MJO is beyond the scope of the present article. Overall, given that iHYCOM in its present configuration cannot resolve extratropical eddies, the coupled model appears to be effective in representing weekly SST variability.

3. Skill of subseasonal hindcasts from FIM-iHYCOM and CFSv2

Buizza and Leutbecher (2015) confirm the forecast skill of time-averaged fields of temperature, wind and geopotential height to be significantly higher than that of time-averaged scores of instantaneous fields. (In their case, instantaneous fields had lower skill than 2-day-averaged fields, which had lower skill than 8-day-averaged fields.) Zhu et al. (2014) evaluate model forecast skill by linking the averaging time window to the lead time as an approach to seamless verification across different timescales. In this section, we have opted to measure forecast skill based on *weekly* averages of model results (except in section 3c, in which daily data are used) on a 1°x1° horizontal grid. Recall from Part I that the native resolution of FIM-iHYCOM is ~60 km, while that of CFSv2 is ~100 km.

a. Probabilistic skill of T2m and precipitation

In the NOAA-facilitated multi-model Subseasonal Experiment [SubX; NOAA (2017)], each model was required to contribute at least 4

ensemble members. As stated earlier, FIM-iHYCOM uses 4 time-lagged members; CFSv2 currently has 16 members per day (as of 1 April 2011) but its hindcasts have only 4 members per day (cf. Part I).

With only 4 ensemble members, traditional "counting" methods to construct probabilistic forecasts (e.g., how many of the ensemble members exceed a certain threshold X) are of limited use. Fortunately, extended logistic regression [ELR; see Wilks (2009) for further details] can be used to construct a continuous range (bounded by [0,1]) of forecast probabilities. This technique was adopted by Vigaud et al. (2017) to use the ensemble mean forecast (and observed climatological terciles) as input into an ELR to create probabilistic forecasts of weekly precipitation on subseasonal timescales. Vigaud et al. (2017) used ELR-based probabilistic forecasts to look at both Ranked Probability Skill Score (RPSS) and reliability diagrams for three S2S prediction systems, including the European Centre for Medium-Range Weather Forecasting (ECMWF) Ensemble Prediction System (EPS) and CFSv2. We evaluate FIM-iHYCOM's probabilistic skill for weekly averaged T2m, and also for weekly accumulated precipitation. To do this, we follow a very similar methodology as that used by Vigaud et al. (2017). In short, this involves evaluating weekly forecasts *initialized* in either January, February, March (JFM) or July, August, September (JAS) over the period 1999-2010: this period was chosen in order to facilitate a direct comparison with Vigaud et al. (2017). FIM-iHYCOM results are compared with those from CFSv2 in Figures 3-6. Moreover, the reanalyses used to build the ELR were CFSR for T2m, and the Global Precipitation Climatology Project [GPCP, Huffman et al. (2001)] for precipitation. We chose GPCP because (i) it incorporates both satellite measurements and ground observations and (ii) it was used by Vigaud et al. (2017) in their ELR.

It should be noted that bias correction in the manner used throughout much of this article was not employed for probabilistic forecasts, i.e., the raw ensemble mean forecasts were used as input to the ELR. But by relating biased forecasts to reanalyses, the training of the regression model implicitly accounts for model bias. Given the fact that Vigaud et al. (2017) did not appear to apply bias correction *a priori* to the ensemble mean forecasts, we did not test the impact of *a priori* bias correction on the training of the ELR and the resultant ELR-computed forecast probabilities.

In this section, we compare the ability of FIM-iHYCOM and CFSv2 to predict below-normal, near-normal, and above-normal conditions for T2m and precipitation.

1) RPSS

The RPSS is one metric to assess probabilistic skill; as described in pp. 299-302 of Wilks (2006; note that "SS_{RPS}" in his Equation (7.49) is the same as RPSS), RPSS is useful for probabilistic forecasts of multi-category (three or more) events (e.g., below, near, or above average temperature and precipitation). Positive values of RPSS indicate forecasts better than a climatological prediction (in this case, assigning a 1/3 probability to each of the three categories). Figure 3 shows RPSS from FIM-iHYCOM and CFSv2 forecasts of T2m for JFM starts as well as JAS starts. Looking at JFM (top two rows of Figure 3), both models have similar spatial distributions of RPSS for all four forecast lead weeks, but FIM-iHYCOM has more areas of non-negative RPSS than CFSv2 for weeks 3 and 4. For JAS starts (bottom two rows of Figure 3), again FIM-iHYCOM has higher RPSS values than CFSv2 for weeks 3 and 4. These results suggest that FIM-iHYCOM provides better probabilistic forecasts than CFSv2 of T2m at subseasonal timescales over the United States.

Figure 4 follows Figure 3, respectively, but for forecasts of precipitation. RPSS for precipitation decreases much faster as a function of lead week than T2m, and the skill of FIM-iHYCOM is comparable to that of CFSv2. The much higher precipitation RPSS for CFSv2 week 1 in Figure 4 – compared with Figs. 5 and 6 of Vigaud et al. (2017) – is a consequence of the different days used to define week 1 [days 1-7 here; days 2-8 in Vigaud et al. (2017)]. By weeks 3 and 4, neither FIM-iHYCOM nor CFSv2 exhibits any cohesive areas of positive RPSS outside the tropics. Overall, the results are comparable to those shown in Figs. 5-6 of Vigaud et al. (2017) for the ECMWF EPS.

2) RELIABILITY DIAGRAMS

Reliability diagrams (e.g., Wilks 2006) provide a useful visualization of a model's probabilistic performance, because they show the probability of observing an event (or of a variable exceeding a certain threshold) given a forecast probability of that same event (in our case, obtained from the ELR). On a reliability diagram, a perfect probabilistic forecasting system will have all points falling on the straight line $y = x$.

Figure 5 shows reliability diagrams (but without the distribution of forecast probabilities) for below, near, and above normal T2m for FIM-iHYCOM forecasts initialized in JFM, for all North American land points between 20°N and 50°N [following Fig. 3a-c of Vigaud et al. (2017)]. In week 1, T2m forecasts are all slightly underconfident (p. 288-289 of Wilks 2006) but are overconfident by week 4. Not surprisingly, the near-normal category is the hardest to predict (e.g., van den Dool and Toth 1991; Kharin and Zwiers 2003): after the first two weeks, there is no resolution – regardless of forecast probability, observed frequency is near 1/3. For the below- and above-normal T2m categories, however, there is more reliability and resolution through week 3. In week 4 the curves show losses in reliability and resolution, and the forecast probabilities (not shown) are concentrated on 1/3. For all three categories, as lead time increases, the forecasts become less sharp (a smaller range of forecast probabilities is issued and there is a tendency to forecast climatology). Overall, we find that FIM-iHYCOM can contribute to real-time prediction of T2m through at least 3 weeks of lead time.

Figure 6 is very similar to the top row of Fig. 3 in Vigaud et al. (2017). Here, reliability diagrams for precipitation based on FIM-iHYCOM forecasts initialized in JFM are shown aggregated over the same area as in Figure 5. The slopes of the lines in these reliability diagrams all indicate an overconfident forecast by week 2. Consistent with the RPSS results shown earlier, precipitation is more difficult to forecast than T2m. Again, the near-normal category is hardest to predict: after the first week, there is no resolution. For the below- and above-normal precipitation categories, however, there is some reliability and resolution through week 2; in weeks 3 and 4 there is no resolution, and the forecast probabilities (not shown) are concentrated on 1/3.

b. Deterministic verification of selected fields

1) ANOMALY CORRELATION COEFFICIENTS

Numerous variations on the theme of deterministic forecast skill assessment have appeared in the literature over the years. In the field of subseasonal prediction in particular, a "best" measure has yet to emerge. One commonly used skill metric for subseasonal verification so far (e.g., Saha et al. 2014; Zhu et al. 2014; Li and Robertson 2015; Lin et al. 2016) is the Anomaly Correlation Coefficient (ACC), which we adopt here to quantify the skill of deterministic predictions.

We define weekly intervals the same way as in Part I (cf. Fig. 1 of that article), namely days 1-7, 8-14, 15-21, 22-28 for weeks 1 to 4, respectively. When we categorize *target* weeks by month, the day in the middle of the target week determines the target month (see Section 2b of Part I for a detailed description).

The ACC is calculated from

$$ACC = \frac{\sum_{k=1}^{K}\sum_{i=1}^{N}(f_{i,k}-F_i)(a_{i,k}-A_i)}{\sqrt{\sum_{k=1}^{K}\sum_{i=1}^{N}(f_{i,k}-F_i)^2 \sum_{k=1}^{K}\sum_{i=1}^{N}(a_{i,k}-A_i)^2}} \quad (1)$$

Here, N extends over all data points spanning the desired range of latitudes and longitudes; K spans all hindcasts available. The $f_{i,k}$ are the ensemble-averaged model forecasts, and $a_{i,k}$ are the corresponding (re)analysis values, both weighted by the cosine of latitude to account for meridian convergence in the latitude-longitude grid. F_i, A_i are the forecast and (re)analysis climatologies at each latitude-longitude gridpoint i.

We verify ensemble mean forecasts of T2m and 500 hPa geopotential height (H500) against CFSR. Precipitation is verified against GPCP during the period of 1999 to 2014.

Figure 7 shows the geographic distributions of the ACC over North America for forecasts of T2m from both FIM-iHYCOM and CFSv2. As in Part I and in section 2b here, the results are composited based on *target* season: 14 for DJF and 16 for JJA. There are 196 cases in DJF and 224 cases in JJA, respectively, which

implies that an ACC value of ~0.1 is statistically significantly greater than zero (at the 95% confidence level) based on a Student's *t* test. There is clearly a substantial contrast in T2m skill between ocean and land (T2m over the ocean is strongly influenced by slowly-evolving sea-surface temperatures), which is why Figures 9 and 11 (see below) only consider land points for T2m. Despite an overall rapid decrease in T2m ACC over land after week 2, there is some statistically significant skill in the southeastern United States beyond week 2. For a given target season (DJF or JJA), FIM-iHYCOM has comparable ACCs with CFSv2. Over land, DJF generally has higher ACCs than JJA for a given lead week; the opposite is the case for points over water. The last row of Figure 7 shows the ACCs of 2-week (14-day) forecasts for weeks 3 and 4 combined. This is along the line of seamless verification in which the temporal averaging window increases as lead time increases (e.g., Zhu et al. 2014). There are regions in which the ACCs of the combined weeks 3-4 forecasts of T2m are higher (although not necessarily statistically significantly higher) than the ACCs of the week 3 forecast. This lends further support to the notion that for subseasonal prediction, useful information could be extracted from a combined weeks 3-4 forecast that might not be evident when considering weeks 3 and 4 separately.

The skill of precipitation is generally quite low beyond week 1 (e.g., Zhu et al. 2014; Li and Robertson 2015). Recent studies have shown that forecast skill at weeks 3 and 4 may be higher than previously thought, via use of more sophisticated data analysis techniques (DelSole et al. 2017) and/or by targeting periods with known sources of predictability (Vigaud et al. 2017). Figure 8 follows Figure 7, but for precipitation. Although ACCs drop rapidly after week 1, there is some skill along the west coast of the United States at weeks 2 and 3, especially in DJF. As with T2m, there are some areas (mainly in the subtropics) in which the combined weeks 3-4 ACCs are higher than the week 3 ACCs for precipitation (bottom row of Figure 8). Overall, FIM-iHYCOM and CFSv2 have similar precipitation ACCs, despite the fact that they employ very different convective parameterizations [modified Grell and Freitas (2014) for FIM-iHYCOM, Simplified Arakawa-Schubert (e.g., Han and Pan 2011) for CFSv2]. In general, the ACCs from both FIM-iHYCOM and CFSv2 are comparable to the Predictive Ocean-Atmosphere Model for Australia [POAMA; cf. Zhu et al. (2014), Wheeler et al. (2017)], and CFSv2 and EPS in Li and Robertson (2015).

Figure 9 compares the ACCs at different lead times of weekly averages in DJF and JJA for T2m (aggregating all land points in the northern hemisphere), as well as for precipitation and H500 (both aggregated over all points from 20°N to 80°N), from both FIM-iHYCOM and CFSv2. In addition to weekly averages, the 2-week average of weeks 3 and 4 is shown. The ACCs from both models are very similar and, not surprisingly, decrease with increasing lead time. Aggregated ACCs remain above zero through 4 weeks, and are mostly higher in DJF than in JJA. Consistent with Figures 7 and 8, the ACCs of the combined weeks 3-4 average forecast is close to that of the week 3 ACCs and higher than the week 4 ACCs.

Figure 10 shows the geographic distributions of H500 ACCs for both FIM-iHYCOM and CFSv2, similar to Figures 7 and 8. Consistent with Figure 9, H500 ACCs through week 3 are higher in DJF than in JJA. Also consistent with Figure 9 is the fact that FIM-iHYCOM and CFSv2 have very similar northern hemisphere ACC magnitudes. A higher level of skill is seen in the North Pacific for all lead weeks in both models, especially in DJF.

In summary, the ACCs of T2m, precipitation, and H500 in FIM-iHYCOM are comparable to those from CFSv2. They are also comparable to those found in other subseasonal prediction systems, including the Canadian Global Ensemble Prediction System (GEPS) shown by Lin et al. (2016) when evaluated using the definitions of weekly averages and target months in their study (not shown), although the spatial patterns differ. These skills are in general higher in the winter season than in summer, as found in other studies (e.g., Kirtman et al. 2014; Zhu et al. 2014; Lin et al. 2016; DelSole et al. 2017), possibly due to the relative dominance in winter of well-resolved synoptic-scale processes such as baroclinic instability as opposed to mesoscale and convection-scale processes that are poorly resolved and represented.

2) SPREAD-ERROR RELATIONSHIP

Two other metrics commonly used to assess model skill are root-mean-square error (RMSE) and ensemble spread. In ensemble prediction, it is important that the RMSE of the ensemble mean is of comparable magnitude to the ensemble spread of the field in question, as they are, for example, in Fig. 1 of Fortin et al. (2014). In general, an ensemble spread that is substantially smaller than (greater than) the RMSE of the ensemble mean indicates an ensemble that is underdispersive (overdispersive). In either case, the ensemble probability distribution poorly represents the true probability distribution; the interested reader is referred to the statistically-based discussion of Fortin et al. (2014). We calculated the spread using the right-hand-side of Eq. (15) of Fortin et al. (2014), reproduced here for convenience:

$$RMSE \approx \sqrt{\left(\frac{R+1}{R}\right)\frac{1}{T}\sum_{t=1}^{T} s_t^2}\sqrt{\left(\frac{R+1}{R}\right)}\left(\overline{s_t^2}\right)^{1/2} \quad (2)$$

Here, R is the ensemble size (4 for the case of FIM-iHYCOM), T is the number of cases (hindcast weeks for FIM-iHYCOM), and s_t^2 is the ensemble variance for case (week) t. The term under the radical symbol on the far right-hand-side of (2) accounts for ensemble size (without this, spread would be less than error simply due to the small number of ensemble members). The results are shown in Figure 11 for T2m, precipitation, and H500 (all three fields composited over the same regions used for Figure 9) as a function of lead time for both FIM-iHYCOM and CFSv2. Overall, the two models are comparable in terms of their RMSE and spread. For T2m and H500, the models are under-dispersive (spread < RMSE) even after accounting for the small ensemble size; however, for precipitation, CFSv2 (and, to a lesser extent, FIM-iHYCOM) have some instances in which spread exceeds error, likely owing to the highly non-Gaussian nature of precipitation. The underdispersive (overconfident) nature of the T2m results in Figure 11a is consistent with the shape of the reliability diagrams (Figure 5) for weeks 3 and 4.

c. Prediction of various subseasonally-relevant phenomena

1) BLOCKING FREQUENCY

Many extreme weather events such as heat waves and flooding are found to be associated with episodes during which the normal mid-latitude zonal flow is temporarily blocked by a meridionally aligned cyclone-anticyclone couplet often referred to as modon (Flierl et al. 1980). Early studies (e.g., Miyakoda et al. 1983) have suggested that there is some predictability for blocking at lead times up to one month; more recent work (e.g., Matsueda 2011) has shown examples of blocking impacting weather on multi-week timescales. Thus, it is important to examine model's ability to simulate blocking at subseasonal timescales. Furthermore, reproducing realistic blocking frequency is considered to be necessary, though not sufficient, for skillful forecasts of blocking. In light of the many shapes and forms of midlatitude blocks, which a single blocking index may not be able to account for satisfactorily, we document their frequency and geographic distribution by relying on two different blocking indices.

The first index is the widely used Tibaldi and Molteni (1990) blocking frequency index, TM index hereafter, which is essentially based on the reversal of the meridional gradient of H500 at mid latitudes. In addition to identifying so-called "instantaneous" blocks (which in this case are blocks identified by the TM index for a single day of the daily-averaged H500 field), the original TM index definition also has an option to add a longevity threshold to blocking identification. Including a temporal threshold (for this article, a minimum of four days) means that only *persistent* blocks are identified. The challenge to the forecast model is not merely to predict blocking events *per se* but to predict the *long-lived* ones correctly, as the latter often lead to extreme weather.

The second blocking index, developed by Pelly and Hoskins (2003), PH index hereafter, is arguably a more physical alternative to the TM index. It is based on a reversal of the meridional potential temperature gradient on a tropopause-level potential vorticity (PV) isosurface within a latitude range centered on the longitude-

dependent mean storm track. The latter refinement gives the PH index some advantage over the TM index whose meridional interval search is independent of longitude. The PH index emphasizes temperature gradient reversals near the tropopause, that is, at the edge of the stratospheric surf zone (McIntyre and Palmer 1984), compared to those evident in the mid-troposphere detected by the TM Index.

Figure 12 shows the blocking frequency from FIM-iHYCOM and CFSv2 at selected forecast lead times from day 7 to 28. It is based on the full 16-year forecast period (no seasonal restriction) of the weekly-initialized 4-member ensembles, where each member is treated an independent forecast – no ensemble means are considered. The solid curves in all panels are based on the TM index. No temporal threshold is used in Figures 12a,b, whereas a 4-day temporal threshold to capture blocking "episodes" is used in Figures 12c,d. Due to the noisy nature of the derived blocking frequency which is unavoidable given the limited sample size, all curves are longitudinally smoothed using 15 passes of the 9-point low-pass filter described in the Appendix. (This number of passes was found to remove wave components of roughly 20° of longitude and shorter.) As a proxy for observed conditions, the blocking frequencies from each model's initial conditions (day 0) are also shown in the top panel[2].

From the perspective of the TM index, FIM-iHYCOM and CFSv2 show similar blocking frequency at all lead weeks – specifically, a frequent Euro-Atlantic block and a weaker Pacific block. The slight decline in blocking frequency with lead time seen in the FIM-iHYCOM results appears to be related to excessive deepening of troughs over the high-latitude ocean basins (not shown) which has the effect of lowering the probability of H500 gradient reversals. This decline trend is largely removed by the bias correction as shown in the gray curves. Given the fact that different time periods are used here, it is not surprising that the blocking frequency shown in Figures 12a,b differs from Jung et al. (2012) for the ECMWF model and Hamill and Kiladis (2014) for the NOAA GEFS model, both also being based on the TM index. Applying the 4-day duration filter lowers the TM blocking frequency by about 30%, as shown in Figures 12c,d.

The dashed lines in Figure 12c show the frequency of 4-day blocking episodes in FIM-iHYCOM as measured by the PH index. As discussed by Pelly and Hoskins (2003), the southward displacement of the storm tracks over the western and central Pacific limits the ability of many blocks identified by the TM index to actually interfere with the westerly flow in that region. The PH index corrects for that and, in the process, shifts the Pacific blocking maximum eastward relative to the one generated by the TM index. This shift, a definite improvement from the synoptic meteorology perspective, is quite noticeable in Figure 12c. Our results confirm the finding of Pelly and Hoskins (2003) that their tropopause-based index captures more blocking events than does the TM index, especially over Europe.

Due to the use of an isentropic vertical coordinate (including a prognostic equation for layer thickness, the denominator in the PV expression) the PV field generated by FIM contains details not resolved in models employing a conventional fixed vertical grid. For this reason, we refrain from comparing PH-index based blocking statistics between FIM-iHYCOM and CFSv2, focusing instead on a comparison of the two blocking indices in FIM-iHYCOM forecasts.

2) MADDEN-JULIAN OSCILLATION

As discussed in Part I, the Madden-Julian Oscillation (MJO) is seen as important for subseasonal timescales because it is responsible for most of the 30-90 day tropical variability (Zhang 2005) and impacts the entire Earth system (Zhang 2013); therefore, a good representation of the MJO in FIM-iHYCOM is necessary but not sufficient to provide reasonably skillful subseasonal forecasts. Green et al. (2017) provide a detailed analysis of the overall ability of an earlier version of FIM-iHYCOM to simulate two

[2] Recall from Part I that the initial conditions for both FIM-iHYCOM and CFSv2 come from CFSR; thus, the small differences in the black curves in Figure 12 (between the two models) are a consequence of interpolating to the FIM-iHYCOM native icosahedral grid before all data are interpolated to 1°x1°.

different MJO indices (and the corresponding input fields). The key differences between the hindcasts in Green et al. (2017) and those shown here are detailed in Section 2a of Part I.

Figure 13 compares the CFSv2 results [over the period 1999-2010 as in Green et al. (2017)] with the full 16-year FIM-iHYCOM hindcast period[3] in terms of ability to predict a variant of the Real-time Multivariate MJO (RMM) index (cf. Wheeler and Hendon 2004; Green et al. 2017), and a similar variant of the Velocity Potential MJO (VPM) index (cf. Ventrice et al. 2013; Green et al. 2017). Both indices require daily (i.e., not weekly) data. Figure 13 follows the methodology of Green et al. (2017) with one major exception: in Green et al. (2017), reanalysis climatology was used whereas here, model climatologies (for both FIM-iHYCOM and CFSv2) were used (i.e., bias correction described in section 2a above was applied). It should be noted here that the verifications for RMM and VPM shown in Figure 13 are based on the NCEP/NCAR reanalysis (Kalnay et al. 1996) rather than CFSR. The present hindcast configuration of FIM-iHYCOM exhibits RMM skill [often defined as exceeding a threshold of 0.5 for the bivariate correlation (e.g., Rashid et al. 2011)] out to ~19 days, essentially the same as that of the CFSv2 hindcast (Figure 13); VPM skill is ~16 days for FIM-iHYCOM and ~18 days for CFSv2. The RMM skill of CFSv2 is consistent with what is shown in Wang et al. (2014). Overall, the performance of the present hindcast in terms of RMM and VPM is comparable to the results shown in Green et al. (2017) for their earlier version of FIM-iHYCOM, not just for bivariate correlation but also for RMSE (both models are comparable; note also that RMSE < $\sqrt{2}$ represents errors less than those of a climatological forecast) and spread (both models are underdispersive). One notable exception is that the current FIM-iHYCOM hindcast for RMM performs better (but not necessarily statistically significantly so) than that used by Green et al. (2017) in terms of both higher correlations and lower RMSEs during the first ~8 days; interestingly, this cannot be explained by the inclusion of bias correction in the present hindcast (not shown).

3) SUDDEN STRATOSPHERIC WARMING

Upward propagation of zonal wavenumber-1 or -2 planetary waves into the upper polar stratosphere, which typically occurs during periods of weak high-latitude stratospheric westerlies (Charney and Drazin 1961; Dickinson 1968; Schoeberl 1978) but on rare occasions also takes place at times when the polar vortex is fully developed, is the generally accepted cause of occasional wintertime circulation changes in the upper stratosphere and mesosphere. Accompanied by very large temperature anomalies (Scherhag 1952) which are attributed to adiabatic compression in descending air near where a low-latitude, low-PV streamer gains ground against the high-PV polar vortex, these events have historically been referred to as sudden stratospheric warmings (SSWs). The commonly used criterion for so-called major warmings is that the zonally averaged 10 hPa zonal wind at 60°N ($\bar{u}_{60}$) changes from westerly to easterly (e.g., Tripathi et al. 2016), a result of the intrusion of low-PV air into the polar region.

There is evidence (e.g., Shaw and Perlwitz 2013) that an SSW, while initiated by a vertically propagating planetary wave, renders the upper stratosphere more reflective to subsequent upward-propagating waves, thereby possibly affecting the tropospheric circulation in the days following the warming. Hence, SSWs are deemed relevant for subseasonal prediction.

We quantitatively assess SSW prediction skill in terms of the ACC (defined as Eq. (1) above) of predicted $\bar{u}_{60}$ for various lead times. Results from FIM-iHYCOM simulations initialized in the months October through March during the 16-year hindcast period are shown in Figure 14. The model skill is comparable to the ECMWF EPS forecast skill shown for the year 2011 in Fig. 10 of Vitart (2014), which is added in Figure 14 for easier comparison.

Other models (Fig. 4 in Tripathi et al. 2016) show skill in predicting SSW for a key case

[3] The impact of including FIM-iHYCOM hindcasts from 2011-2014 is deemed negligible (not shown).

in January 2013 up to 15 days in advance. Figure 15 illustrates the occurrence and predictive skill of SSW for two representative boreal winters (2008/2009 and 2012/2013) in FIM-iHYCOM simulations in terms of the maximum temperature on the 10 hPa surface (T_{max}) inside the polar cap north of 60°N, as well as $\bar{u}_{60}$. The axes in the diagrams are model initialization time and forecast lead time (0-32 days). Lines of equal forecast verification time (dot-dashed in the figure) slope from upper left to lower right. In order to arrive at coherent contour plots despite the large data gap (1 week) on the abscissa, additional data points were created by linear interpolation in the oblique direction marked by the dot-dashed lines. Successfully predicted SSW events are depicted in this reference frame by $\bar{u}_{60}$ and T_{max} isopleths that are aligned with the dot-dashed lines.

Figure 15 shows that FIM-iHYCOM succeeds in predicting the SSWs in the two selected years up to three weeks in advance. However, the gradual loss of "slope" of the $\bar{u}_{60}$ isopleths illustrates a general tendency of FIM-iHYCOM to be late in predicting flow reversal at lead times beyond 10 days.

4. Discussion

The main goal of this study is to validate the new FIM-iHYCOM coupled model and generate a baseline for a systematic assessment of prediction skill with a focus on subseasonal timescales, as an attempt to bridge the skill gap between weather forecasting and seasonal prediction. This is the second of two articles on subseasonal prediction with the new coupled modeling system, focusing on the model's probabilistic and deterministic prediction skill. Part I provided a detailed description of the coupled model along with an evaluation of systematic seasonal biases.

A comprehensive hindcast dataset at 60-km horizontal resolution was constructed by running a 4-member time-lagged ensemble on a weekly basis for the 16-year period 1999-2014. The main focus is on the verification of 2-m temperature, precipitation and 500-hPa geopotential height, but we also attempt to shed light on the coupled model's ability to predict specific processes deemed relevant for longer-range prediction, namely, blocking, the Madden-Julian Oscillation, and sudden stratospheric warming events. There is also a cursory comparison between model-simulated variability and that of observations. In subject areas where we are able to compare FIM-iHYCOM to CFSv2 (and other models), we find that the skill is comparable.

We analyzed weekly mean forecasts on a 1°x1° horizontal grid after removing biases extracted from the model's own climatology, except in the case of probabilistic skill where model bias is accounted for implicitly. Despite the existing model biases shown in Part I, the competitive probabilistic and deterministic skills of FIM-iHYCOM for T2m, precipitation and H500, as well as the ability of the model to simulate various phenomena (blocking, the MJO, and SSWs), appear to be similar to those of the operational models of CFSv2, EPS, GEPS, and POAMA mentioned earlier. In addition, we compared the blocking frequency simulated by FIM-iHYCOM using two different indices, and confirmed that the PH index is meteorologically more relevant than the TM index at least for the Pacific blocks.

Our work is based on only 4 ensemble members, and no stochastic forcing is applied during model integration. Despite these limitations, FIM-iHYCOM shows promising skill in many aspects of the metrics used here and suggests a likely positive contribution to NOAA's multi-model Subseasonal Experiment (SubX).

Due to the nonstandard spatial discretization of the model equations on the sphere and in the vertical direction, and the inclusion of a different scale-aware convection scheme [namely, a variant of Grell and Freitas (2014)], the primary purpose of FIM-iHYCOM will be to enrich genetic diversity in multi-model ensembles that are increasingly being used in predicting atmospheric circulation anomalies and associated weather phenomena on timescales beyond 2 weeks. To become a candidate for inclusion in multi-model ensembles, an individual model must be demonstrated to be state-of-the-art in terms of its ability to simulate atmospheric flow patterns and weather-regime

transitions relevant to subseasonal weather prediction; this is the primary goal of this article.

In our future work, we will use FIM-iHYCOM, as well as coupled models based on NOAA's Next-Generation Global Prediction System (NGGPS), to conduct more detailed, process-based studies of some of the subseasonally relevant phenomena that were given brief attention in this article. We will also investigate subseasonal as well as longer-term phenomena not yet covered in this article, such as tropical cyclone frequency, the stratospheric quasi-biennial oscillation, and El Niño-Southern Oscillation.

The dataset generated during this study has been made available to the research community through NOAA's SubX project. Combined with other existing subseasonal datasets, SubX offers the potential to improve our understanding of the mechanisms that are crucial for subseasonal prediction. The multi-model dataset also serves as a benchmark against which future coupled models will be compared, including models incorporating NGGPS, which is based on finite-volume numerics on a cubed-sphere grid (e.g., Putman and Lin 2007).

Acknowledgments. This project was supported by NOAA OAR funding for week 3-4 forecast improvement and the Earth System Prediction Capability program. Co-authors Sun, Green, and Bleck are also supported by funding from NOAA OAR Award NA17OAR4320101. We thank Dr. Nicolas Vigaud for his helpful suggestions on building the extended logistic regression model, as well as Drs. John Brown and George Kiladis for providing internal reviews. Dr. Kathy Pegion also gave useful feedback. Three anonymous reviewers provided extensive feedback on an earlier version of this work. The authors acknowledge the NOAA Research and Development High Performance Computing Program for providing computing and data storage resources that have been instrumental in generating the results reported in this article (URL: http://rdhpcs.noaa.gov), as well as the Texas Advanced Computing Center (TACC) at The University of Texas at Austin (URL: http://www.tacc.utexas.edu). CFSv2 and CFSR data were obtained from the National Centers for Environmental Information.

APPENDIX
Low-Pass Filter Details

A description of the procedure for generating a model-specific climatology would not be complete without documentation of the method used for eliminating short-term fluctuations from the raw model output. Here we provide details of the low-pass filters employed in this work.

We use filters developed by Fleck and Fryer (1953) which offer advantages over both the use of running averages (which suffer from serious "ringing" throughout the frequency range) and the use of a finite number of Fourier components (which tends to suppress seasonal peaks). Fleck-Fryer filters have flawless damping characteristics at wave numbers outside the low-pass window, a consequence of setting to zero as many derivatives as possible of the filter transfer function at the high wavenumber end. The number of derivatives that can be specified increases with the number of filter weights.

Starting with the center weight, the 13 independent weights for the 25-point time filter used in generating the model climatology are listed in the middle column of Table A1. (Note that filters must be symmetric with respect to the center weight to avoid phase shifts in the processed data.) The 5 independent weights of the 9-point longitudinal filter used in processing blocking statistics are listed in the right column of Table A1. The resulting transfer functions for these two filters used here are shown in Figure A1.

References

Baldwin, M. P., and T. J. Dunkerton, 1999: Propagation of the Arctic oscillation from the stratosphere to the troposphere. *J. Geophys. Res.,* **104,** 30937-30946, doi:10.1029/1999JD900445.

Beljaars, A. C. M., P. Viterbo, M. J. Miller, and A. K. Betts, 1996: The anomalous rainfall over the United States during July 1993: Sensitivity to land surface parameterization and soil moisture anomalies. *Mon. Wea. Rev.,* **124,** 362-383, doi:10.1175/1520-0493(1996(124<0362:TAROTU>2.0.CO;2.

Brunet, G. N., and Coauthors, 2010: Collaboration of the weather and climate communities to advance subseasonal-to-seasonal prediction. *Bull. Amer. Meteor. Soc.,* **91,** 1397-1406, doi:10.1175/2010BAMS3013.1.

Buizza, R., and M. Leutbecher, 2015: The forecast skill horizon. *Quart. J. Roy. Meteor. Soc.,* **141,** 3366-3382, doi:10.1002/qj.2619.

Charney, J. G., and P. G. Drazin, 1961: Propagation of planetary-scale disturbances from the lower into the upper atmosphere. *J. Geophys. Res.,* **66,** 83-109, doi:10.1029/JZ066i001p00083.

Charney, J. G., and J. Shukla, 1981: Predictability of monsoons. Monsoon Dynamics, J. Lighthill and R. P. Pearce, Eds., Cambridge Univ. Press, New York, 99.

DelSole, T., L. Trenary, M. K. Tippett, and K. Pegion, 2017: Predictability of week-3-4 average temperature and precipitation over the contiguous United States. *J. Climate,* **30,** 3499-3512, doi:10.1175/JCLI-D-16-0567.1.

Dickinson, R. E., 1968: Planetary Rossby waves propagating vertically through weak westerly wind wave guides. *J. Atmos. Sci.,* **25,** 984-1002, doi:10.1175/1520-0469(1968)025<0984:PRWPVT>2.0.CO;2.

Fleck, J. T., and W. D. Fryer, 1953: An exploration of numerical filtering techniques. Cornell Aero Lab., Inc., Rept. No. XA-869-P-1, 79 pp.

Flierl, G. R., V. D. Larichev, J. C. McWilliams and G. M. Reznik, 1980: The dynamics of baroclinic and barotropic solitary eddies. Dyn. Atmosph. Oceans,1-41.

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.

Green, B. W., S. Sun, R. Bleck, S. G. Benjamin, and G. A. Grell, 2017: Evaluation of MJO predictive skill in multiphysics and multimodel global ensembles. *Mon. Wea. Rev.,* **145,** 2555-2574, doi:10.1175/MWR-D-16-0419.1.

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.

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.

Hoerling, M. P., and A. Kumar, 2002: Atmospheric response patterns associated with tropical forcing. *J. Climate,* **15,** 2184-2203, doi:10.1175/1520-0442(2002)015<2184:ARPAWT>2.0.CO;2.

Huffman, G. J., R. F. Adler, M. M. Morrissey, D. T. Bolvin, S. Curtis, R. Joyce, B. McGavock, and J. Susskind, 2001: Global precipitation at one-degree daily resolution from multisatellite observations. *J. Hydrometeor.,* **2,** 36-50, doi:10.1175/1525-7541(2001)002<0036:GPAODD>2.0.CO;2.

Jung, T., and Coauthors, 2012: High-resolution global climate simulations with the ECMWF model in Project Athena: Experimental design, model climate, and seasonal forecast skill. *J. Climate,* **25,**

3155-3172, doi:10.1175/JCLI-D-11-00265.1.

Kalnay, E., and A. Dalcher, 1987: Forecasting forecast skill. *Mon. Wea. Rev.,* **115,** 349-356, doi:10.1175/1520-0493(1987)115<0349:FFS>2.0.CO;2.

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.

Kharin, V. V., and F. W. Zwiers, 2003: Improved seasonal probability forecasts. *J. Climate,* **16,** 1684-1701, doi:10.1175/1520-0442(2003)016<1684:ISPF>2.0.CO;2.

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.

Kirtman, B. P., and Coauthors, 2014: The North American Multimodel Ensemble: Phase-1 seasonal-to-interannual prediction; Phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.,* **95,** 585-601, doi:10.1175/BAMS-D-12-00050.1.

Li, S., and A. W. Robertson, 2015: Evaluation of submonthly precipitation forecast skill from global ensemble prediction systems. *Mon. Wea. Rev.,* **143,** 2871-2889, doi:10.1175/MWR-D-14-00277.1.

Lin, H., N. Gagnon, S. Beauregard, R. Muncaster, M. Markovic, B. Denis, and M. Charron, 2016: GEPS-based monthly prediction at the Canadian Meteorological Centre. *Mon. Wea. Rev.,* **144,** 4867-4883, doi:10.1175/MWR-D-16-0138.1.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus,* **21,** 289-307, doi:10.1111/j.2153-3490.1969.tb0044.x.

Matsueda, M., 2011: Predictability of Euro-Russian blocking in summer of 2010. *Geophys. Res. Lett.,* **38,** L06801, doi:10.1029/2010GL046557.

McIntyre, M. E., and T. N. Palmer, 1984: The 'surf zone' in the stratosphere. *J. Atm. Terr. Phys.,* **46,** 825-849, doi:10.1016/0021-9169(84)90063-1.

Miyakoda, K., T. Gordon, R. Caverly, W. Stern, J. Sirutis, and W. Bourke, 1983: Simulation of a blocking event in January 1977. *Mon. Wea. Rev.,* **141,** 846-869, doi:10.1175/1520-0493(1983)111<0846:SOABEI>2.0.CO;2.

NAS, 2016: Next Generation Earth System Prediction: Strategies for Subseasonal to Seasonal Forecasts. National Academy of Sciences, Engineering, and Medicine, The National Academies Press, Washington, D.C. Available online at https://www.nap.edu/catalog/21873/next-generation-earth-system-prediction-strategies-for-subseasonal-to-seasonal

NOAA, 2017: S2S Prediction Task Force: Subseasonal to Seasonal (2016-2019). Accessed 04 December 2017. [Available online at https://cpo.noaa.gov/Meet-the-Divisions/Earth-System-Science-and-Modeling/MAPP/MAPP-Task-Forces/S2S-Prediction-Task-Force.]

Palmer, T. N., and S. Tibaldi, 1988: On the prediction of forecast skill. *Mon. Wea. Rev.,* **116,** 2453-2480, doi:10.1175/1520-0493(1988)116<2453:OTPOFS>2.0.CO;2.

Palmer, T. N., 2002: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quart. J. Roy. Meteor. Soc.,* **128,** 747-774, doi:10.1256/0035900021643593.

Palmer, T. N., and Coauthors, 2004: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bull. Amer. Meteor. Soc.,* **85,** 853-872, doi:10.1175/BAMS-85-6-853.

Pegion, K., and B. P. Kirtman, 2008: The impact of air-sea interactions on the simulation of tropical intraseasonal variability. *J. Climate,* **21,** 6616-6635, doi:10.1175/2008JCLI2180.1.

Pelly, J. L., and B. J. Hoskins, 2003: A new perspective on blocking. *J. Atmos. Sci.,* **60,** 743-755, doi:10.1175/1520-0469(2003)060<0743:ANPOB>2.0.CO;2.

Putman, W. M., and S.-J. Lin, 2007: Finite-volume transport on various cubed-sphere grids. *J. Comp. Phys.,* **227,** 55-78, doi: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. *Clim. Dyn.,* **36,** 649-661, doi: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.

Saha, S., and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate,* **27,** 2185-2208, doi:10.1175/JCLI-D-12-00823.1.

Scherhag, R., 1952: Die explosionsartigen stratosphärenerwärmungen des spätwinters 1951/1952. Berichte Deutsch. Wetterdienst. US-Zone, 38, 51-63.

Schoeberl, M. R., 1978: Stratospheric warmings: observations and theory. *Rev. Geophys. Space Phys.,* **16,** 521-538, doi:10.1029/RG016i004p00521.

Shaw, T. A. and J. Perlwitz, 2013: The life cycle of northern hemisphere downward wave coupling between the stratosphere and troposphere. *J. Climate,* **26,** 1745-1763, doi:10.1175/JCLI-D-12-00251.1.

Smith, L. A., H. Du, E. B. Suckling, and F. Niehörster, 2015: Probabilistic skill in ensemble seasonal forecasts. *Quart. J. Roy. Meteor. Soc.,* **141,** 1085-1100, doi:10.1002/qj.2403.

Straus, D. M., and J. Shukla, 2000: Distinguishing between the SST-forced variability and internal variability in mid latitudes: Analysis of observations and GCM simulations. *Quart. J. Roy. Meteor. Soc.,* **126,** 2323-2350, doi:10.1002/qj.49712656716.

Sun, S., R. Bleck, S. G. Benjamin, B. W. Green, and G. A. Grell, 2018: Subseasonal forecasting with an icosahedral, vertically quasi-Lagrangian coupled model. Part I: Model overview and evaluation of systematic errors. *Mon. Wea. Rev.,* in review.

Tibaldi, S., and F. Molteni, 1990: On the operational predictability of blocking. *Tellus,* **42A,** 343-365, doi:10.1034/j.1600-0870.1990.t01-2-00003.x.

Tripathi, O. P., and Coauthors, 2016: Examining the predictability of the stratospheric sudden warming of January 2013 using multiple NWP systems. *Mon. Wea. Rev.,* **144,** 1935-1960, doi:10.1175/MWR-D-15-0010.1.

van den Dool, H. M., and Toth, Z., 1991: Why do forecasts for "Near Normal" often fail? *Wea. Forecasting,* **6,** 76-85, doi:10.1175/1520-0434(1991)006<0076:WDFFNO>2.0.CO;2.

Ventrice, 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.

Vigaud, N., A. W. Robertson, and M. K. Tippett, 2017: Multimodel ensembling of subseasonal precipitation forecasts over North America. *Mon. Wea. Rev.,* **145,** 3913-3928, doi:10.1175/MWR-D-17-0092.1.

Vitart, F., 2004: Monthly forecasting at ECMWF. *Mon. Wea. Rev.,* **132,** 2761-2779, doi:10.1175/MWR2826.1.

Vitart, F., 2014: Evolution of ECMWF sub-seasonal forecast skill scores. *Quart. J. Roy. Meteor. Soc.,* **140,** 1889-1899, doi:10.1002/qj.2256.

Waliser, D. E., K. M. Lau, W. Stern, and C. Jones, 2003: Potential predictability of the Madden-Julian oscillation. *Bull. Amer. Meteor. Soc.,* **84,** 33-50, doi:10.1175/BAMS-84-1-33.

Walland, D. J., and I. Simmonds, 1997: Modelled atmospheric response to changes in northern hemisphere snow cover. *Clim. Dyn.,* **13,** 25-34, doi:10.1007/s003820050150.

Wang, W., M.-P. Hung, S. J. Weaver, A. Kumar, and X. Fu, 2014: MJO prediction in the NCEP Climate Forecast System version 2. *Clim. Dyn.,* **42,** 2509-2520, doi: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.

Wheeler, M. C., H. Zhu, A. H. Sobel, D. Hudson, and F. Vitart, 2017: Seamless precipitation prediction skill comparison between two global models. *Quart. J. Roy. Meteor. Soc.,* **143,** 374-383, doi:10.1002/qj.2928.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2d ed. Academic Press, 627 pp.

Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorol. Appl.,* **16,** 361-368, doi:10.1002/met.134.

WMO, 2015: Seamless prediction of the earth system: From minutes to months. Tech. Rep. 1156, World Meteorological Organization. Available online at https://library.wmo.int/pmb_ged/wmo_1156_en.pdf.

Zhang, C., 2005: Madden-Julian oscillation. *Rev. Geophys.,* **43,** RG2003, doi:10.1029/2004RG000158.

Zhang, C., 2013: Madden-Julian oscillation: Bridging weather and climate. *Bull. Amer. Meteor. Soc.,* **94,** 1849-1870, doi:10.1175/BAMS-D-12-00026.1.

Zhang, Q., and H. van den Dool, 2012: Relative merit of model improvement versus availability of retrospective forecasts: The case of Climate Forecast System MJO prediction. *Wea. Forecasting,* **27,** 1045-1051, doi:10.1175/WAF-D-11-00133.1.

Zhu, H., M. C. Wheeler, A. H. Sobel, and D. Hudson, 2014: Seamless precipitation prediction skill in the tropics and extratropics from a global model. *Mon. Wea. Rev.,* **142,** 1556-1569, doi:10.1175/MWR-D-13-00222.1.

List of figures

FIG. 1. T2m climatology at 40°N, 90°W for FIM-iHYCOM for day 1 lead time. Green curve: raw climatology (simple average over the 16-year period); red curve: smoothed climatology after 10 passes of 25-day filter. CFSR climatology provided by NCEI is shown for comparison (black curve).

FIG. 2. (a) Variance of weekly-averaged SST (K^2) for target season DJF from CFSR. (b)-(e) FIM-iHYCOM variance minus CFSR variance for lead weeks 1-4, respectively.

FIG. 3. RPSS for T2m forecasts, verified against CFSR, initialized in JFM (top two rows) and JAS (bottom two rows) for FIM-iHYCOM (rows 1 and 3) and CFSv2 (rows 2 and 4) over a region encompassing the conterminous United States. Lead weeks 1 to 4 are shown from left to right.

FIG. 4. As in Fig. 3 but for precipitation forecasts verified against GPCP. Note that the color bar matches that in Figs. 5 and 6 of Vigaud et al. (2017).

FIG. 5. Reliability diagrams for T2m forecasts (verified against CFSR) from FIM-iHYCOM, restricted to North American land points between 20°N and 50°N and initialized in JFM. Lead weeks 1 through 4 shown in different colors. Left to right: below-normal, near-normal, and above-normal categories.

FIG. 6. As in Fig. 5, but for precipitation forecasts verified against GPCP.

FIG. 7. Maps of T2m ACCs for DJF (left two columns) and JJA (right two columns) from FIM-iHYCOM (columns 1 and 3) and CFSv2 (columns 2 and 4) at lead times of (top four rows) 1 to 4 weeks. Bottom row shows the ACCs as computed from the average of weeks 3 and 4. Values of ACC $\geq$ ~0.1 are significantly different from zero (at 95% confidence).

FIG. 8. As in Fig. 7, but for precipitation.

FIG. 9. ACCs for FIM-iHYCOM (blue) and CFSv2 (red) forecasts as a function of forecast lead week (1-4) for target seasons DJF (left) and JJA (right). Top: T2m over land points in the northern hemisphere. Middle: precipitation between 20°N and 80°N. Bottom: H500 between 20°N and 80°N. Dashed lines show the ACCs computed from the average of weeks 3 and 4.

FIG. 10. Similar to Figs. 7 and 8, but for H500.

FIG. 11. Similar to Fig. 9, but for RMSE (solid) and spread (dashed). Units of RMSE and spread for T2m, precipitation, and H500 are K, mm dy^{-1}, and gpm, respectively. The geographic areas over which RMSE and spread are computed match those of Fig. 9.

FIG. 12. Northern Hemisphere blocking frequency as a function of longitude from 16 years (1999-2014) of ensemble forecasts (each member treated as an independent sample) for lead times of 7, 14, 21, and 28 days. (a) Solid lines: TM index extracted from weekly-sampled 4-member FIM-iHYCOM hindcasts with no temporal threshold. Gray lines: TM index based on bias-corrected H500; colored lines: TM index based on H500 without bias correction. (b) As in (a), but for CFSv2. (c) and (d) are similar to (a) and (b), but with a temporal threshold of 4 days. PH index (dashed lines) added in (c). "Lead Day 0" – initial conditions – added in (a) and (b) as proxy for reanalysis.

FIG. 13. Model performance as a function of lead time for FIM-iHYCOM (blue) and CFSv2 (red) ensemble mean forecasts of the RMM index (left) and VPM index (right) [as in Green et al. (2017)]. Top: Bivariate correlation (gray line = 0.5). Bottom: Bivariate root-mean-square error (RMSE, solid; gray line = $\sqrt{2}$) and 4-member ensemble spread (dashed).

FIG. 14. ACCs of $\bar{u}_{60}$ predictions from FIM-iHYCOM for months October - March in years 1999/2000 to 2013/2014, plotted against forecast lead time (days). Circles show ACC values for 2011 reproduced from Fig. 10 of Vitart (2014).

FIG. 15. Illustration of FIM-iHYCOM predictions for boreal winters 2008/2009 (left) and 2012/2013 (right). Top: maximum 10 hPa

temperature in polar cap north of 60°N (°C). Bottom: $\bar{u}_{60}$ (m s^{-1}). Abscissa: model initialization time. Ordinate: forecast lead time (0-32 days). Slanted dot-dashed lines are lines of equal model verification time.

FIG. A1. Transfer functions for two low-pass filters used in this study. Red curves show the effect of applying the original filter (shown in blue) multiple times, as indicated. Abscissa: wave number in units of inverse data intervals.

List of tables

Table A1. Filter weights (starting with the central weight and progressing outwards) of the symmetric 25-point temporal and 9-point spatial filters discussed in the Appendix.

Point (from center)	25-point filter	9-point filter
1	1.611802578E-01	2.7343750E-01
2	1.487817764E-01	2.1875000E-01
3	1.168999672E-01	1.0937500E-01
4	7.793331146E-02	3.1250000E-02
5	4.383748770E-02	3.9062500E-03
6	2.062940598E-02	
7	8.022546768E-03	
8	2.533435822E-03	
9	6.333589554E-04	
10	1.206398010E-04	
11	1.645088196E-05	
12	1.430511475E-06	
13	5.960464478E-08	

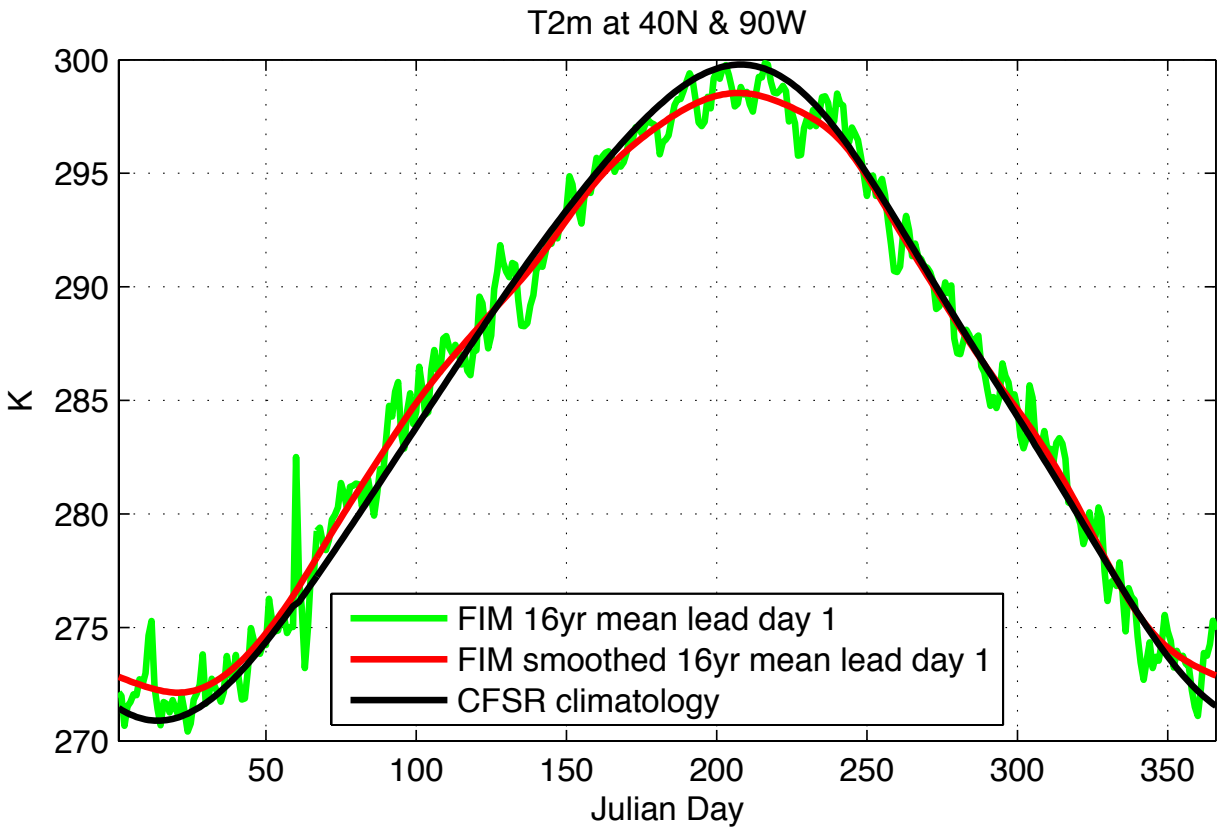

FIG. 1. T2m climatology at 40°N, 90°W for FIM-iHYCOM for day 1 lead time. Green curve: raw climatology (simple average over the 16-year period); red curve: smoothed climatology after 10 passes of 25-day filter. CFSR climatology provided by NCEI is shown for comparison (black curve).

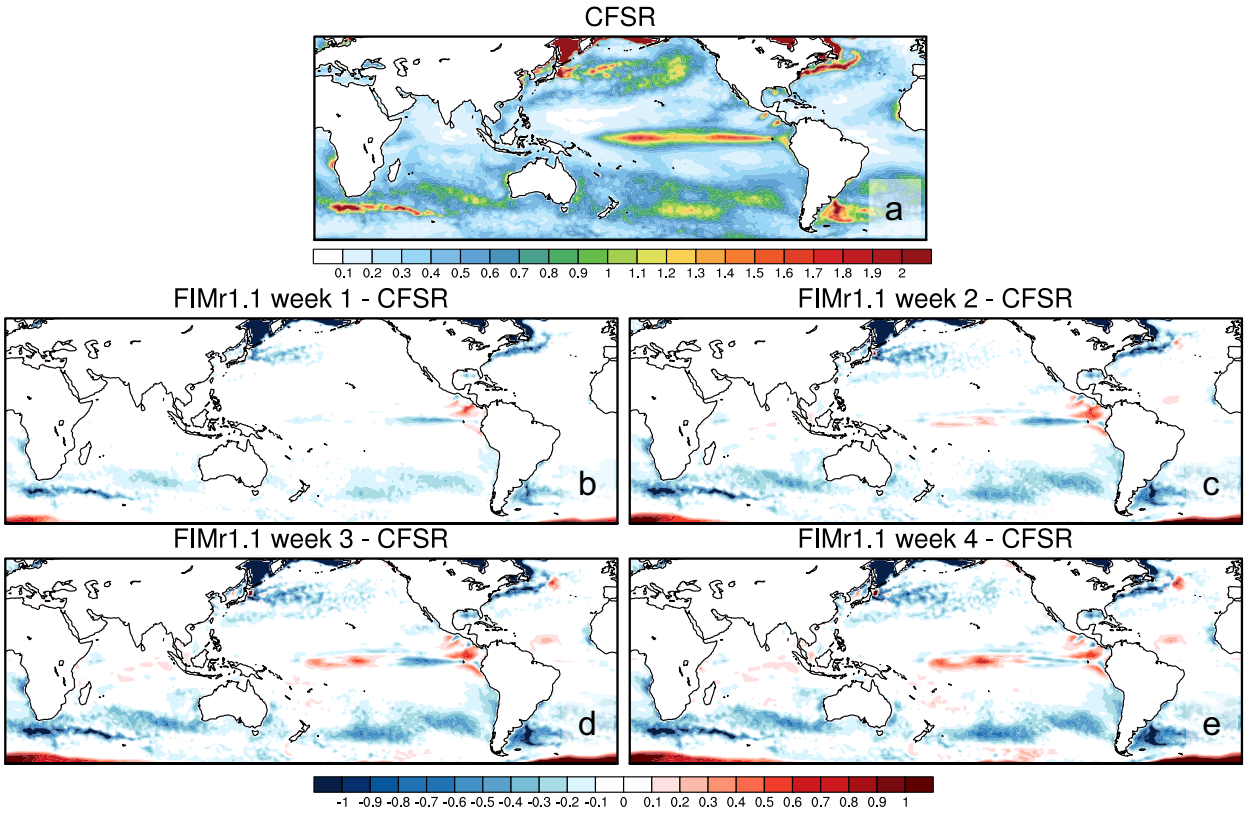

FIG. 2. (a) Variance of weekly-averaged SST (K^2) for target season DJF from CFSR. (b)-(e) FIM-iHYCOM variance minus CFSR variance for lead weeks 1-4, respectively.

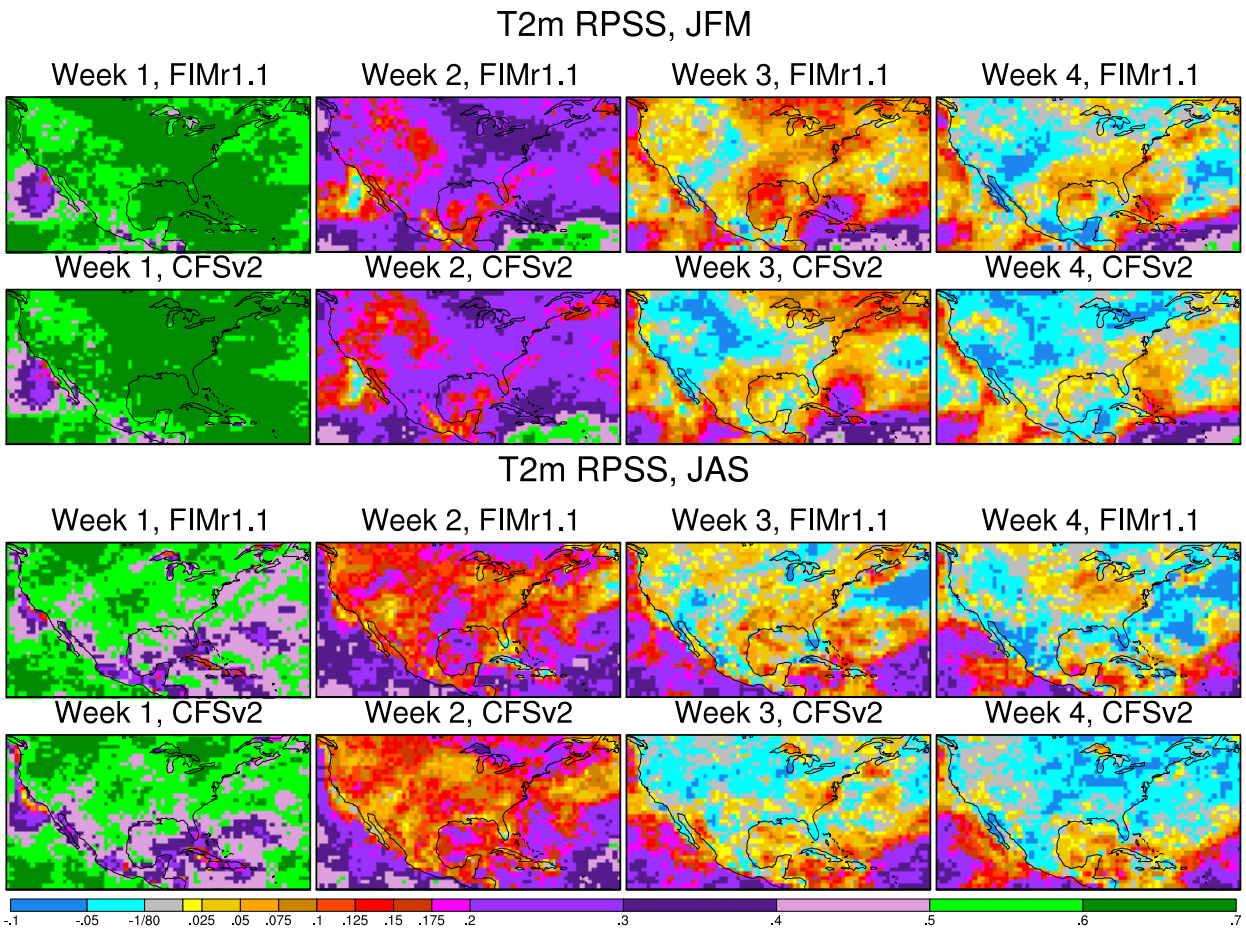

FIG. 3. RPSS for T2m forecasts, verified against CFSR, initialized in JFM (top two rows) and JAS (bottom two rows) for FIM-iHYCOM (rows 1 and 3) and CFSv2 (rows 2 and 4) over a region encompassing the conterminous United States. Lead weeks 1 to 4 are shown from left to right.

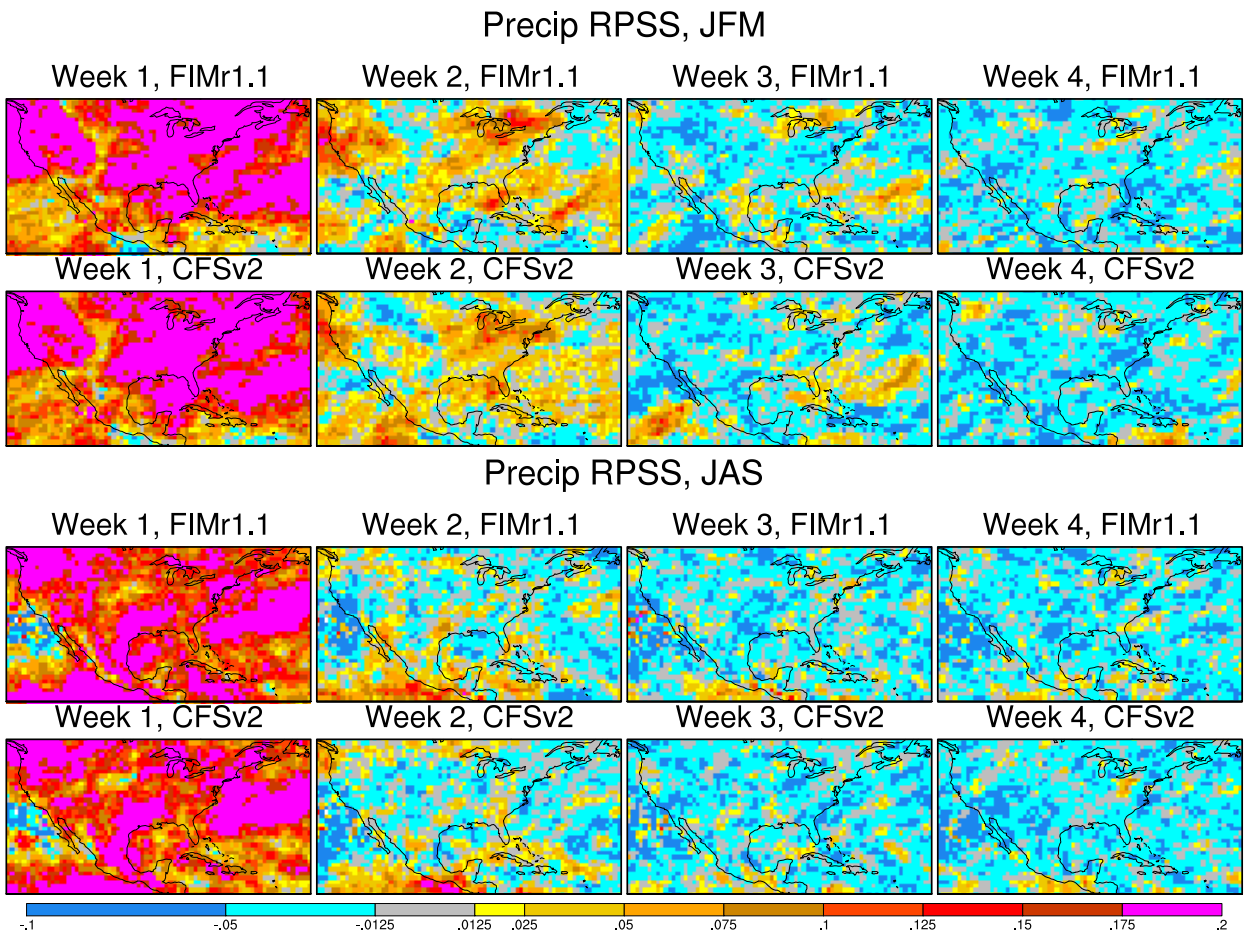

FIG. 4. As in Fig. 3 but for precipitation forecasts verified against GPCP. Note that the color bar matches that in Figs. 5 and 6 of Vigaud et al. (2017).

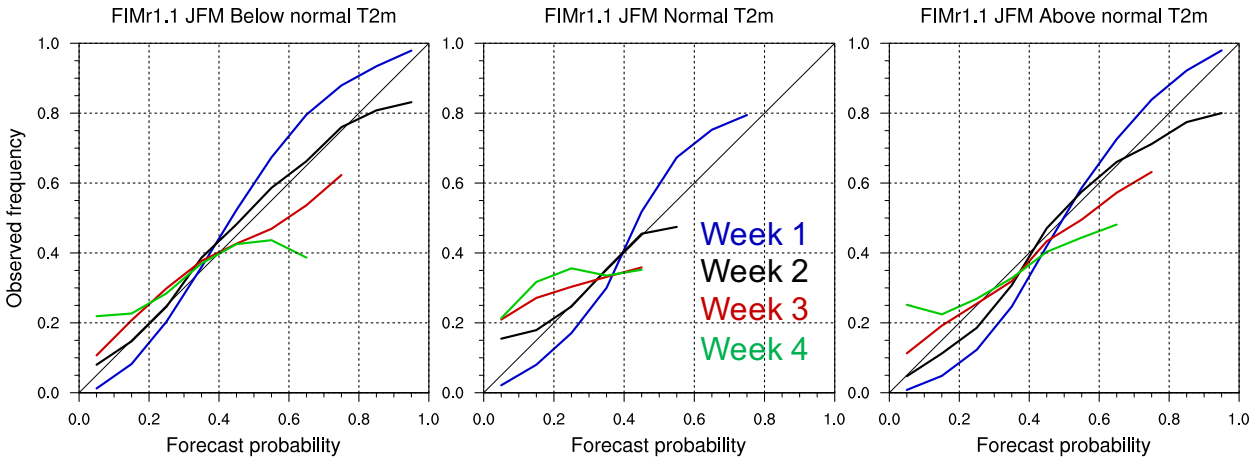

FIG. 5. Reliability diagrams for T2m forecasts (verified against CFSR) from FIM-iHYCOM, restricted to North American land points between 20°N and 50°N and initialized in JFM. Lead weeks 1 through 4 shown in different colors. Left to right: below-normal, near-normal, and above-normal categories.

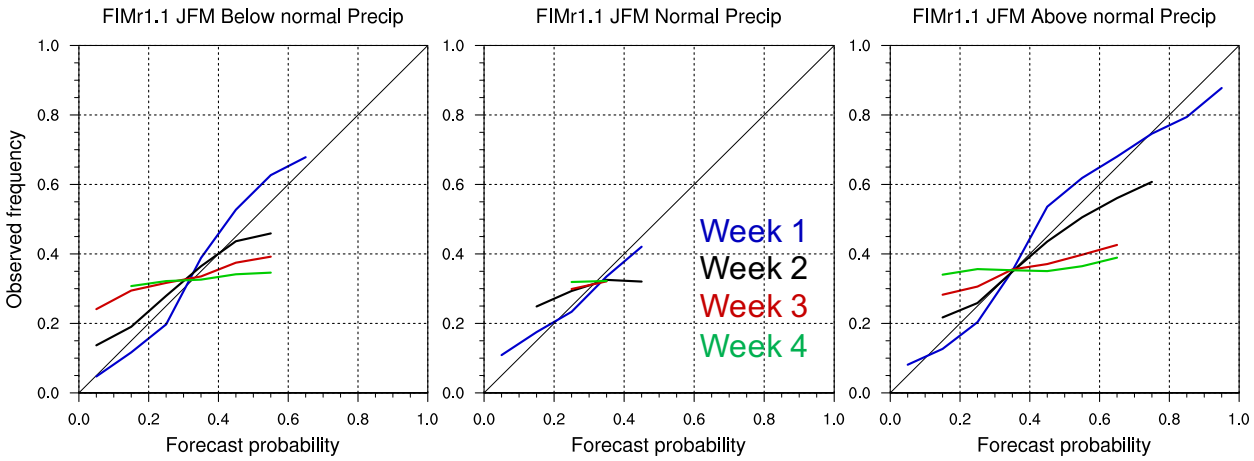

FIG. 6. As in Fig. 5, but for precipitation forecasts verified aga

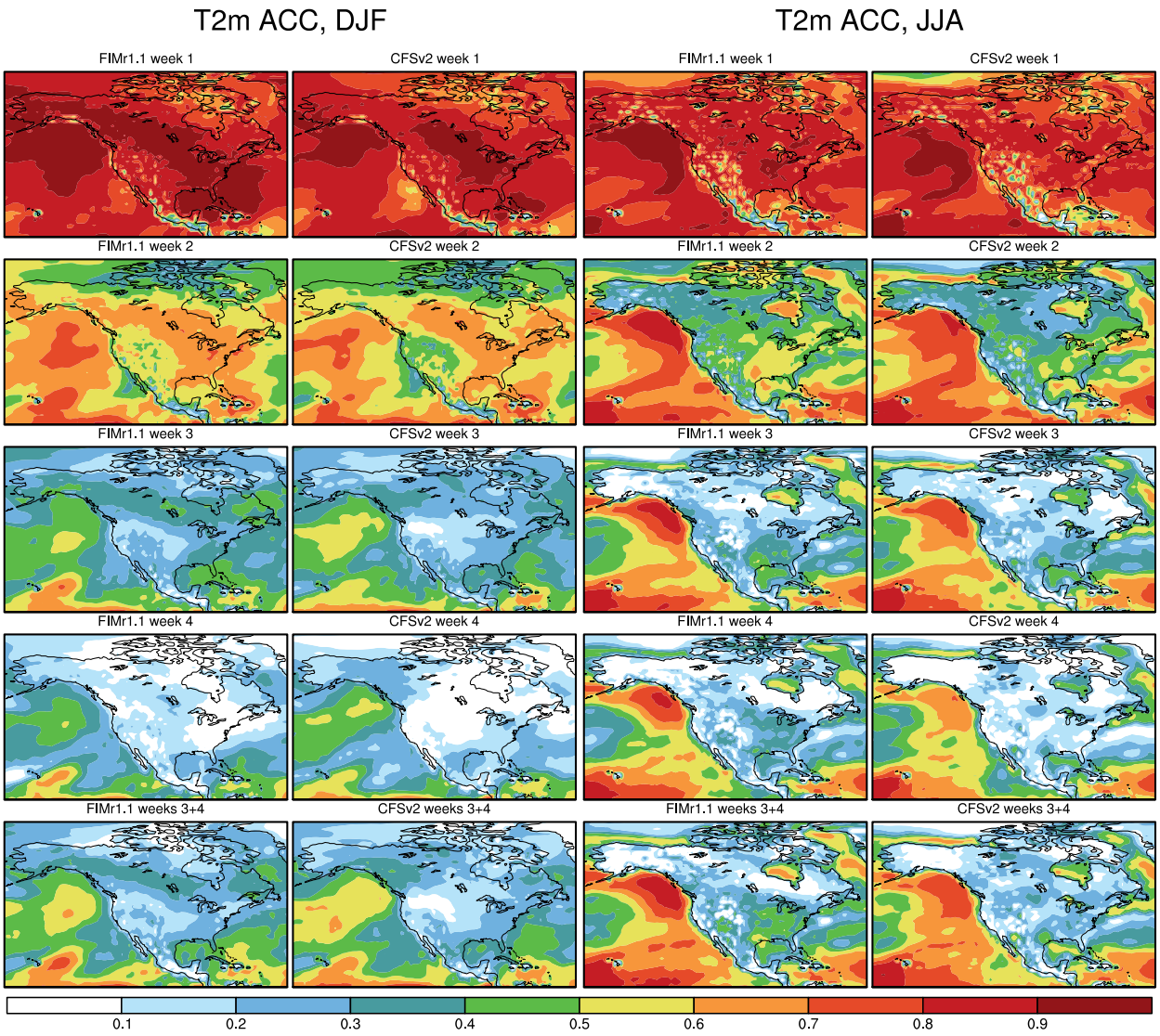

FIG. 7. Maps of T2m ACCs for DJF (left two columns) and JJA (right two columns) from FIM-iHYCOM (columns 1 and 3) and CFSv2 (columns 2 and 4) at lead times of (top four rows) 1 to 4 weeks. Bottom row shows the ACCs as computed from the average of weeks 3 and 4. Values of ACC $\geq$ ~0.1 are significantly different from zero (at 95% confidence).

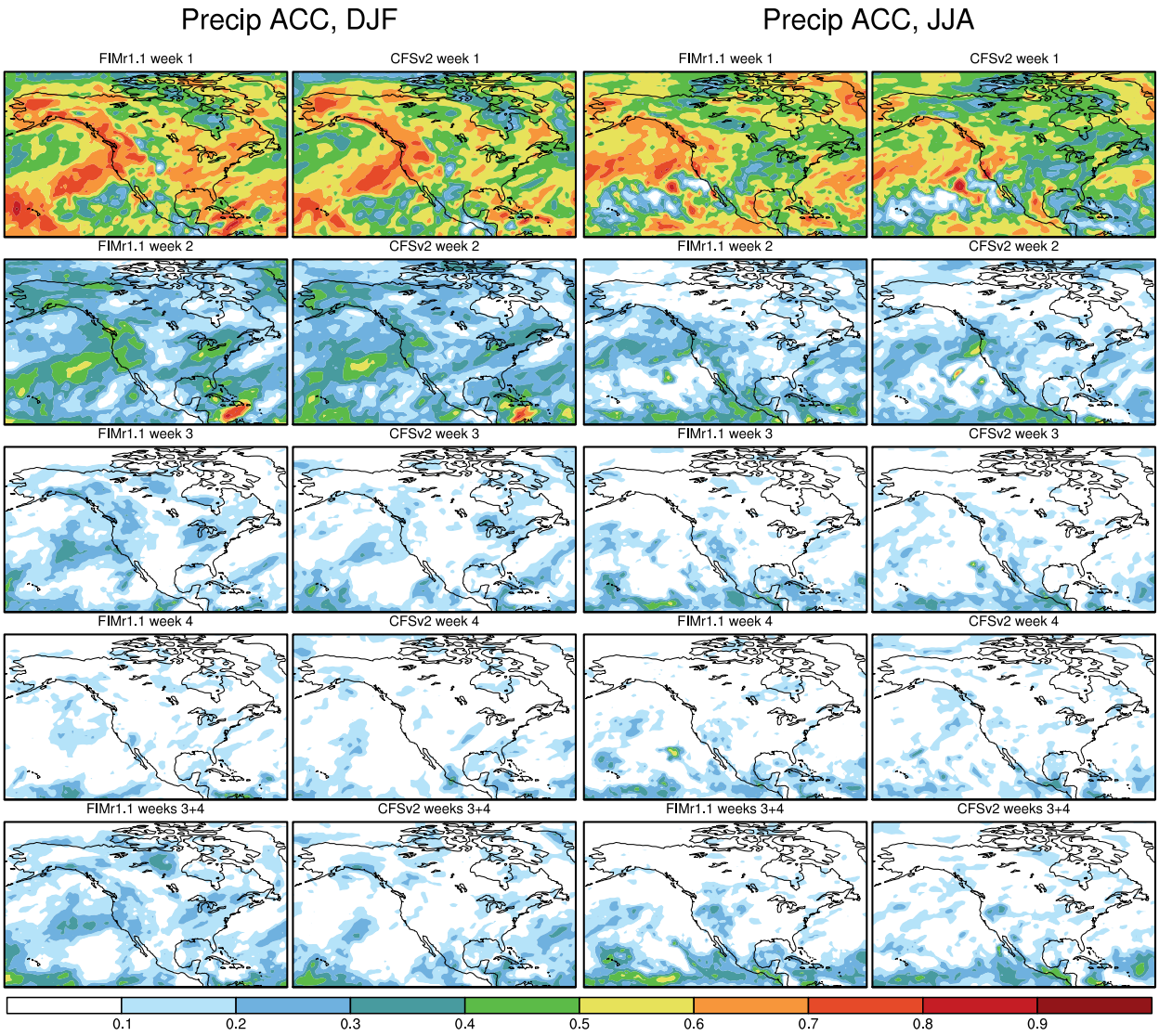

FIG. 8. As in Fig. 7, but for precipitation.

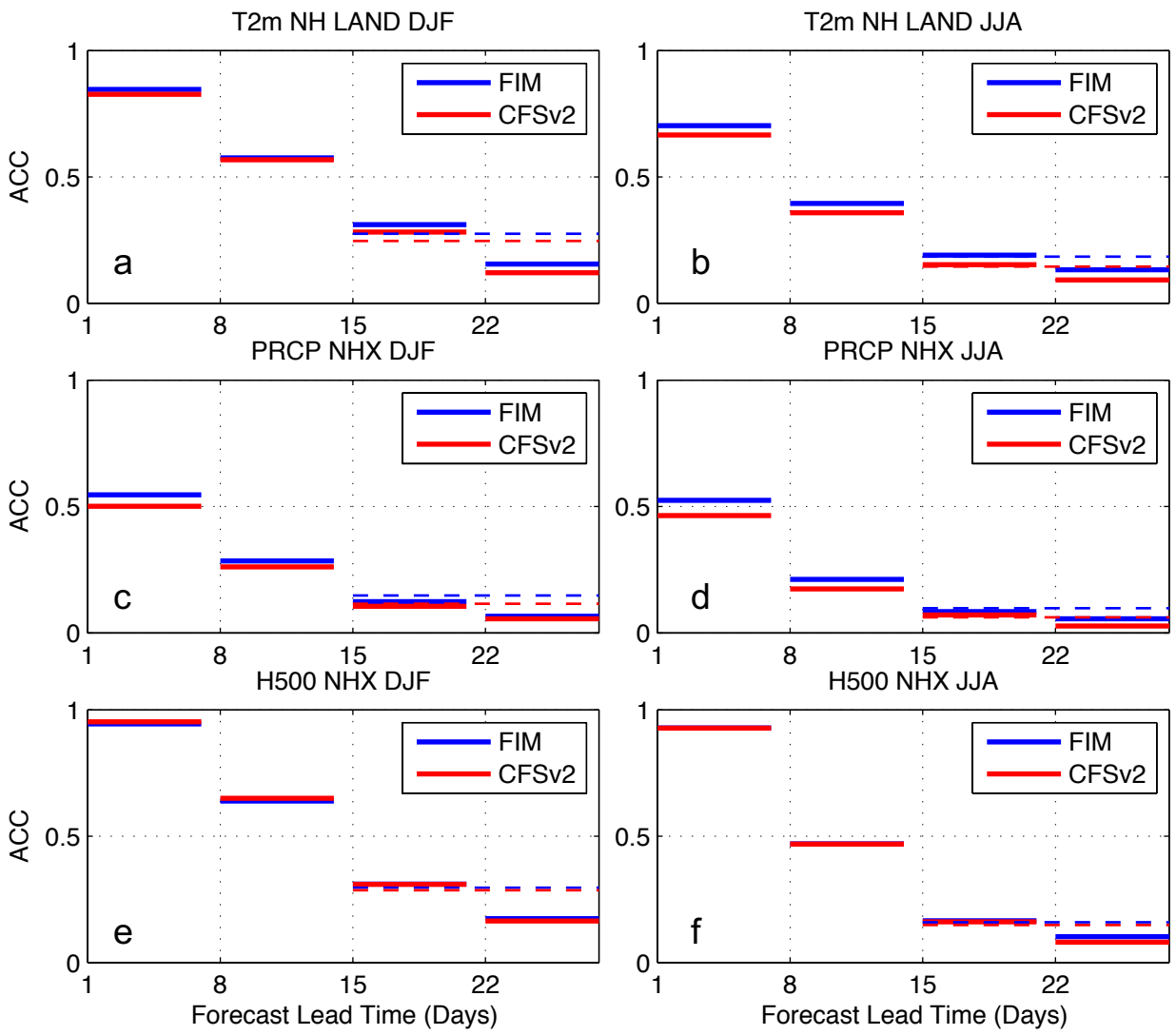

FIG. 9. ACCs for FIM-iHYCOM (blue) and CFSv2 (red) forecasts as a function of forecast lead week (1-4) for target seasons DJF (left) and JJA (right). Top: T2m over land points in the northern hemisphere. Middle: precipitation between 20°N and 80°N. Bottom: H500 between 20°N and 80°N. Dashed lines show the ACCs computed from the average of weeks 3 and 4.

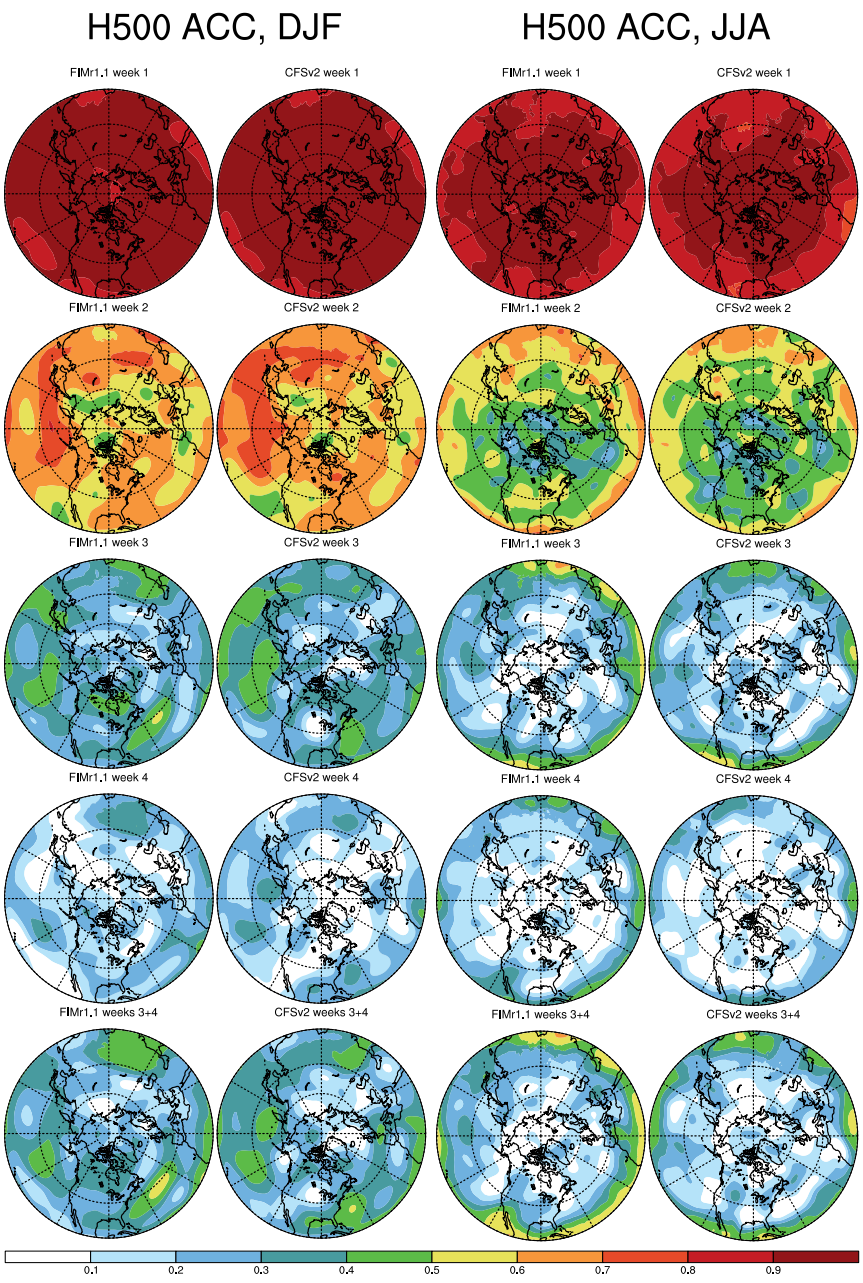

FIG. 10. Similar to Figs. 7 and 8, but for H500.

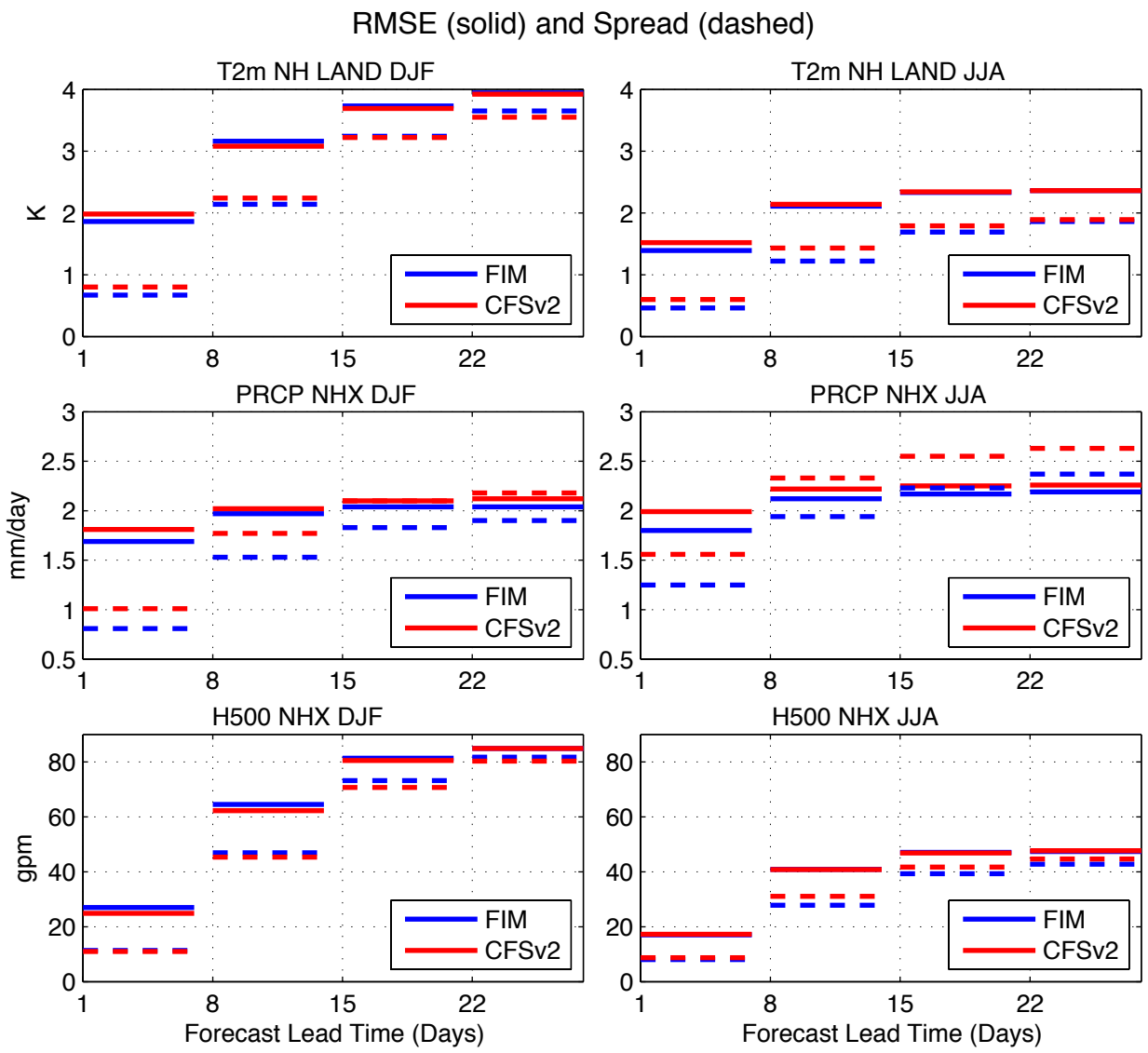

FIG. 11. Similar to Fig. 9, but for RMSE (solid) and spread (dashed). Units of RMSE and spread for T2m, precipitation, and H500 are K, mm dy^{-1}, and gpm, respectively. The geographic areas over which RMSE and spread are computed match those of Fig. 9.

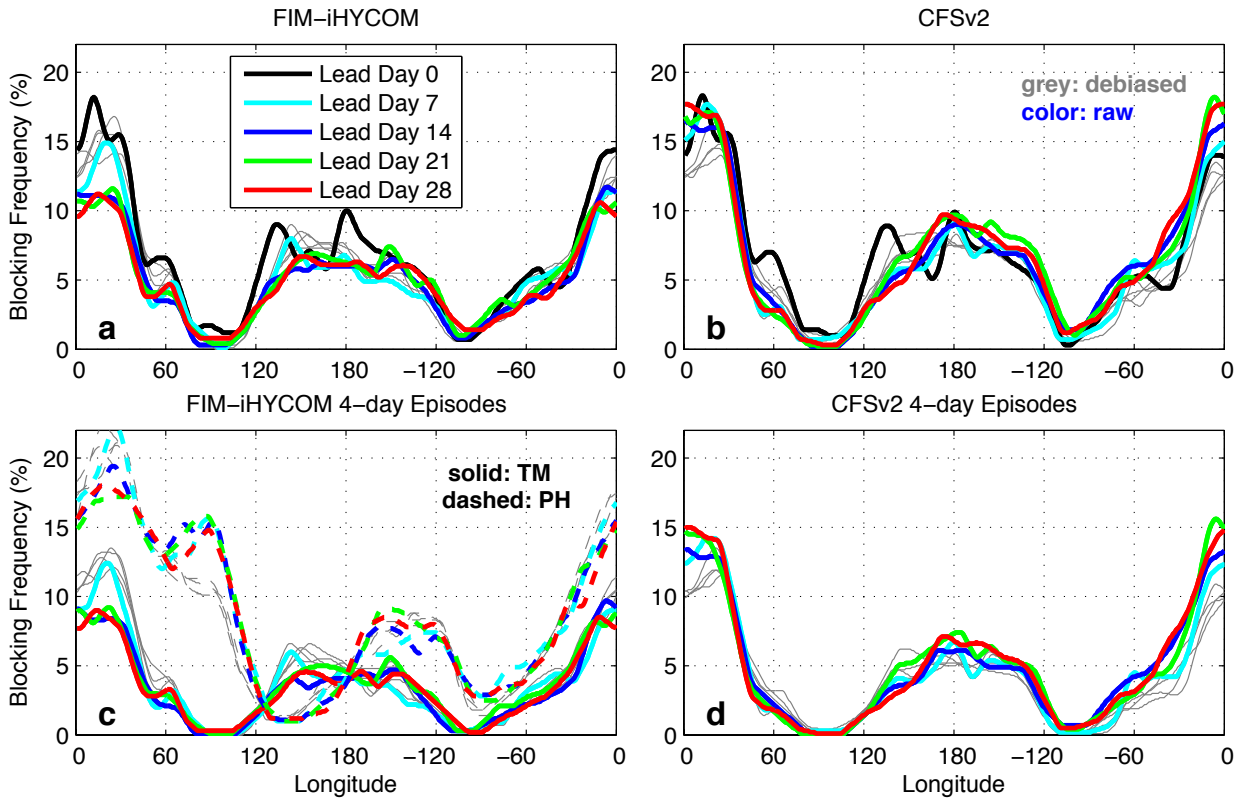

FIG. 12. Northern Hemisphere blocking frequency as a function of longitude from 16 years (1999-2014) of ensemble forecasts (each member treated as an independent sample) for lead times of 7, 14, 21, and 28 days. (a) Solid lines: TM index extracted from weekly-sampled 4-member FIM-iHYCOM hindcasts with no temporal threshold. Gray lines: TM index based on bias-corrected H500; colored lines: TM index based on H500 without bias correction. (b) As in (a), but for CFSv2. (c) and (d) are similar to (a) and (b), but with a temporal threshold of 4 days. PH index (dashed lines) added in (c). "Lead Day 0" – initial conditions – added in (a) and (b) as proxy for reanalysis.

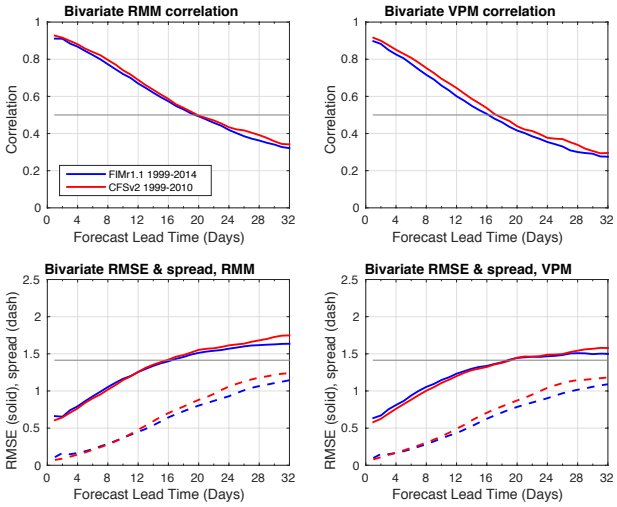

FIG. 13. Model performance as a function of lead time for FIM-iHYCOM (blue) and CFSv2 (red) ensemble mean forecasts of the RMM index (left) and VPM index (right) [as in Green et al. (2017)]. Top: Bivariate correlation (gray line = 0.5). Bottom: Bivariate root-mean-square error (RMSE, solid; gray line = $\sqrt{2}$) and 4-member ensemble spread (dashed).

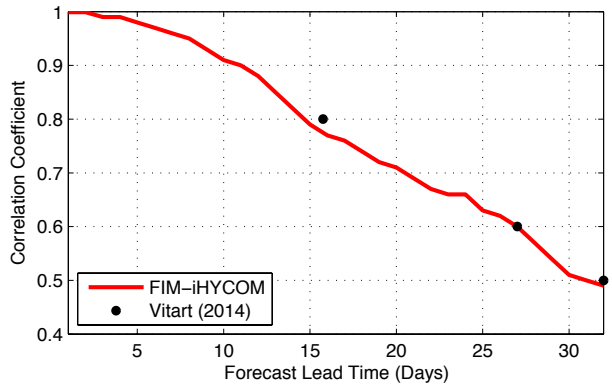

FIG. 14. ACCs of $\bar{u}_{60}$ predictions from FIM-iHYCOM for months October - March in years 1999/2000 to 2013/2014, plotted against forecast lead time (days). Circles show ACC values for 2011 reproduced from Fig. 10 of Vitart (2014)

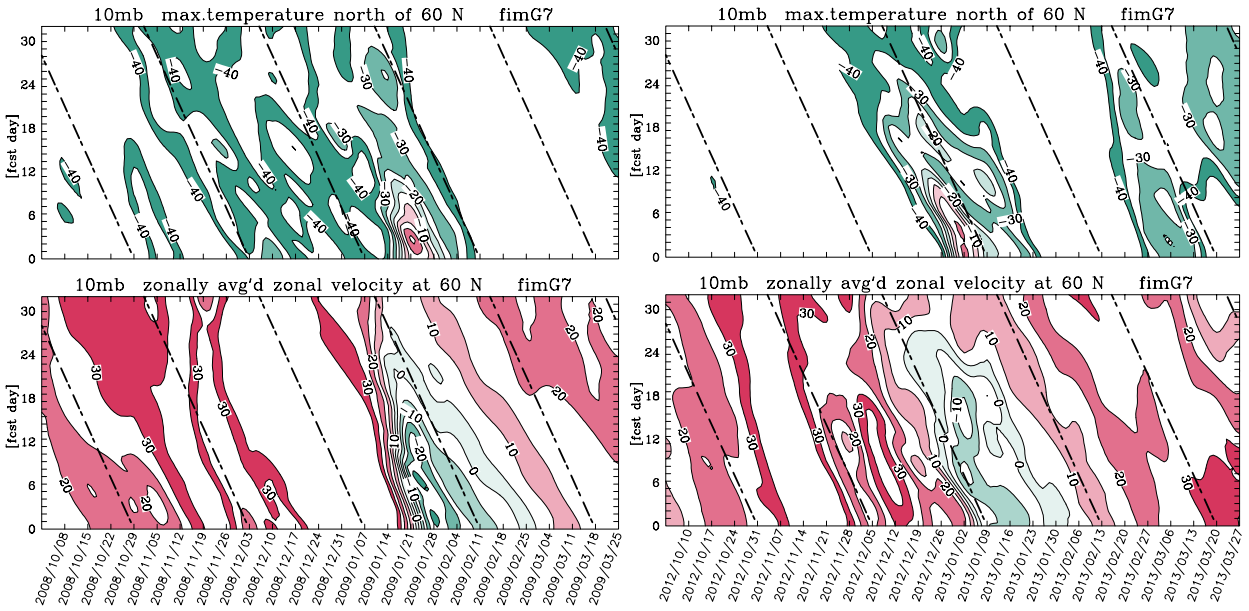

FIG. 15. Illustration of FIM-iHYCOM predictions for boreal winters 2008/2009 (left) and 2012/2013 (right). Top: maximum 10 hPa temperature in polar cap north of 60°N (°C). Bottom: $\bar{u}_{60}$ (m s^{-1}). Abscissa: model initialization time. Ordinate: forecast lead time (0-32 days). Slanted dot-dashed lines are lines of equal model verification time.

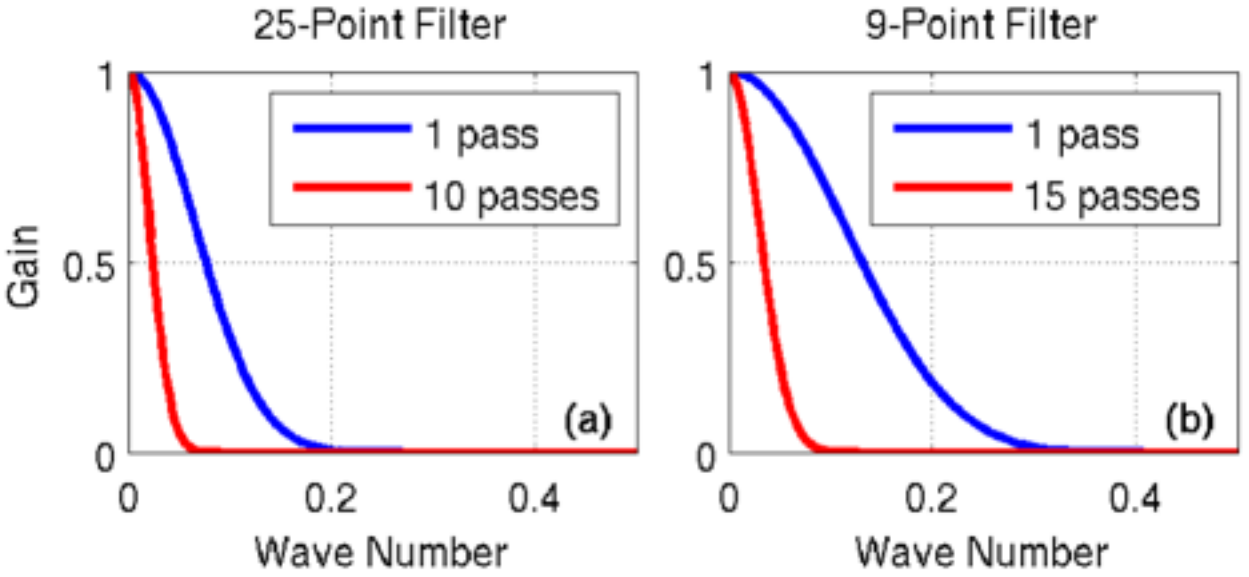

FIG. A1. Transfer functions for two low-pass filters used in this study. Red curves show the effect of applying the original filter (shown in blue) multiple times, as indicated. Abscissa: wave number in units of inverse data intervals