The Subseasonal Experiment (SubX):

A multi-model subseasonal prediction experiment

Kathy Pegion*

George Mason University, Fairfax, VA, USA

Ben P. Kirtman

University of Miami, Rosenstiel School for Marine and Atmospheric Sciences, Miami, FL, USA

Dan C. Collins

NOAA/NCEP/Climate Prediction Center, College Park MD, USA

Emerson LaJoie

NOAA/NCEP/Climate Prediction Center and Innovim, Inc., College Park MD, USA

Robert Burgman

Florida International University, Miami, FL, USA

Ray Bell

University of Miami, Rosenstiel School for Marine and Atmospheric Sciences, Miami, FL, USA

Timothy DelSole

George Mason University and Center for Ocean-Land-Atmosphere Studies, Fairfax, VA

Dughong Min

University of Miami, Rosenstiel School for Marine and Atmospheric Sciences, Miami, FL, USA

Yuejian Zhu

NOAA/NCEP/Environmental Modeling Center, College Park, MD, USA

Wei Li

IMSG at NOAA/NCEP/Environmental Modeling Center, College Park, MD, USA

Eric Sinsky

IMSG at NOAA/NCEP/Environmental Modeling Center, College Park, MD, USA

Hong Guan

SRG at NOAA/NCEP/Environmental Modeling Center, College Park, MD, USA

Emily Becker

NOAA/NCEP/Climate Prediction Center, College Park MD, USA and Innovim, Inc., College Park MD, USA

Jon Gottschalck

NOAA/NCEP/Climate Prediction Center, College Park MD, USA

E. Joseph Metzger

Naval Research Laboratory, Oceanography Division, Stennis Space Center, MS, USA

Neil P Barton

Naval Research Laboratory, Marine Meteorology Division, Monterey, CA, USA

Deepthi Achuthavarier

Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, MD, USA and Universities Space Research Association, Columbia, MD, USA

Jelena Marshak

Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, MD, USA

Randal D. Koster

Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, MD, USA

Hai Lin

Recherche en prvision numrique atmosphrique, Environment and Climate Change Canada, Dorval, Quebec, Canada

Normand Gagnon

Canadian Meteorological Centre, Environment and Climate Change Canada, Dorval, Quebec, Canada

Michael Bell

International Research Institute for Climate and Society (IRI), Columbia University, Palisades, NY

Michael K. Tippett

Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY

Andrew W. Robertson

International Research Institute for Climate and Society (IRI), Columbia University, Palisades, NY

Shan Sun

University of Colorado Boulder, Cooperative Institute for Research in Environmental Sciences, Boulder, CO, USA and NOAA/OAR/ESRL/Global Systems Division, Boulder, CO, USA

Stanley G. Benjamin

NOAA/OAR/ESRL/Global Systems Division, Boulder, CO, USA

Benjamin W. Green

University of Colorado Boulder, Cooperative Institute for Research in Environmental Sciences, Boulder, CO, USA and NOAA/OAR/ESRL/Global Systems Division, Boulder, CO, USA

Rainer Bleck

University of Colorado Boulder, Cooperative Institute for Research in Environmental Sciences, Boulder, CO, USA and NOAA/OAR/ESRL/Global Systems Division, Boulder, CO, USA

Hye-Mi Kim

School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, NY, USA

**Corresponding author address:* Dept. of Atmospheric, Oceanic, and Earth Sciences, George Mason University, Fairfax, VA

E-mail: kpegion@gmu.edu

ABSTRACT

SubX is a multi-model subseasonal prediction experiment with both research and real-time components. Seven global models have produced seventeen years of retrospective (re-) forecasts and more than a year of weekly real-time forecasts. Both the re-forecasts and forecasts are archived at the Data Library of the International Research Institute for Climate and Society, Columbia University, for research on subseasonal predictability and predictions. The real-time forecasts started in July 2017 to provide guidance to the week 3-4 outlooks issued by the Climate Prediction Center at the NOAA National Centers for Environmental Prediction. Evaluation of SubX model biases demonstrates that model bias patterns are already established at week 1 and grow to week 4. Temperature and precipitation skill over the U.S. exists for week 3-4 predictions for specific regions and seasons. The SubX multi-model ensemble is more skillful than any individual model overall. Skill in simulating the Madden-Julian Oscillation and the North Atlantic Oscillation is also evaluated and found to be comparable to other subseasonal modeling systems. SubX is also able to make useful contributions to operational forecast guidance at the Climate Prediction Center.

1. Introduction

A well-known "gap" exists in our current prediction systems at the subseasonal (2-weeks to several months) timescale, as the memory of the atmospheric initial conditions is increasingly lost, while information in the slowly-evolving surface boundary conditions has had insufficient time to be felt (National Research Council (2010); Brunet et al. (2010); National Academies of Sciences, Engineering and Medicine (2017); Mariotti et al. (2018); Black et al. (2017)). Although there is evidence that predictability exists at this timescale in some regions and seasons (e.g. Pegion and Sardeshmukh (2011); DelSole et al. (2017); Li and Roberston (2015)), it is not clear whether the full potential of prediction skill has been realized. Additionally, many questions remain regarding our fundamental understanding of the physical processes giving rise to predictability, as well as how best to design, build, post-process, and verify a subseasonal prediction system.

Until recently, it has been difficult to assess the skill of subseasonal predictions. Re-forecast databases consisted of monthly or seasonal predictions that were not initialized frequently enough to capture the full range of subseasonal variability (e.g., NMME, DEMETER, CHFP, ENSEMBLES, APCC/CliPAS) (Kirtman et al. (2014); Palmer et al. (2004); Tompkins et al. (2017); Weisheimer and Reyes (2009); Wang et al. (2008)) or weather predictions that did not extend to long enough lead-times for subseasonal predictions (e.g. TIGGE, GEFS 2nd generation re-forecasts) (Swinbank et al. (2016); Hamill et al. (2013)). Initial efforts to produce subseasonal re-forecasts and evaluate skill focused primarily on the Madden-Julian Oscillation (MJO) and boreal summer intraseasonal oscillation (e.g., ISVHE, Neena et al. (2014) and NCEP-CFSv2 45-day re-forecasts, Saha et al. (2014); Wang et al. (2013)).

More recently, a focused community effort has developed to facilitate research on a broad range of subseasonal predictions and to understand current and potential capabilities for improving subseasonal skill. The World Weather Research Programme (WWRP)/World Climate Research Program (WCRP) Subseasonal to Seasonal (S2S) Prediction Project is an international project bringing together the weather and climate prediction communities to improve physical understanding and forecast skill for the S2S timescale (Robertson et al. (2015); Vitart et al. (2017)). A major contribution of this project is the development of a S2S forecast database consisting of operational forecasts (3 weeks behind real time), and re-forecasts, from 11 international global producing centers of long-range forecasts for S2S research purposes (Vitart et al. 2017). SubX contributes to the community S2S effort by providing a publicly available database of forecasts and re-forecasts. A unique contribution of SubX is that the real-time forecasts are made available *without delay* to support potential use in real-time applications. Additionally, the NOAA/Climate Program Office, Modeling Analysis and Predictions Program has developed an S2S Prediction Task Force consisting of researchers using the WWRP/WCRP S2S and SubX databases for research on subseasonal prediction and predictability (Mariotti et al. (2018) and Mariotti et al. (2018), manuscript submitted to EOS).

There is ever-increasing demand for predictions on these timescales, specifically predictions relevant for risk reduction and disaster preparedness, public health, energy, water management, agriculture, and marine fisheries (see White et al. (2017) for a review of S2S applications). In the U.S., the NOAA National Centers for Environmental Prediction (NCEP) Climate Prediction Center (CPC) was mandated to begin issuing week 3-4 outlooks for temperature and precipitation. Given that there are immediate needs for understanding predictability *and* making skillful operational predictions on these timescales, a research-to-operations (R2O) project provides the ideal testbed for quick progress in making subseasonal predictions while continuing research efforts that can lead to increased subseasonal prediction skill in the future.

SubX was launched to provide such a testbed. It follows in the footsteps of the North American Multi-model Ensemble (NMME), a R2O project focused on monthly and seasonal (1-month to 1-year) predictions (Kirtman et al. 2014). NMME contains a publicly available research archive of 36 years of re-forecast and forecast data, and has been providing real-time seasonal forecast guidance since 2011. Similarly, SubX brings together seven global models, following a specific protocol to make both re-forecasts and real-time forecasts on the subseasonal timescale. The collection of models consists of U.S. and Canadian operational models as well as research models. The inclusion of research models, another unique contribution of SubX, allows research groups to approach model improvements from a practical prediction perspective and to test those improvements in a real-time prediction framework. Given the timescale of interest, some models originate from the numerical weather prediction (NWP) community while others come from the seasonal prediction community, bringing together critical expertise from both communities to make progress on subseasonal prediction. The re-forecast and real-time forecast data are made publicly available to facilitate broad research and applications community use. Additionally, SubX forecasts are being provided each week to NCEP/CPC, and multi-model ensemble (MME) guidance is produced in support of their week 3-4 outlooks.

The purpose of this paper is to describe SubX, the available data (Section 2) and the evaluation of model biases and skill for operationally relevant variables (Section 3c,d). We also provide skill evaluation for some known sources of subseasonal predictability (Section 3e) and a description of how SubX contributes to the official NCEP/CPC week 3-4 outlooks (Section 4).

2. Protocol and Database

Each of the modeling groups participating in SubX agreed to follow a specific re-forecast and real-time forecast protocol. Given the demanding requirements of both re-forecasts and real-time forecasts, the protocol itself represents a compromise between the traditional operating modes of the NWP and seasonal prediction communities. For example, NWP groups are accustomed to running in real-time with frequent initializations, but producing shorter period re-forecast databases and only recently extending model runs to subseasonal timescales. In contrast, the seasonal prediction community typically produces large re-forecast datasets and extended range predictions, but not with weekly initializations.

While each modeling group was allowed to determine the details of their individual prediction system, (e.g., initialization, resolution, earth-system components, etc.), the SubX protocol required that each group adhere to a rigid scope of retrospective and real-time forecasts. The groups agreed to produce 17 years of re-forecasts out to a minimum of 32 days for the years 1999-2015. Initialization was required at least weekly, and a minimum of three ensemble members were required, although more were encouraged. Since the land-surface (e.g., soil moisture) is an important source of subseasonal predictability (Koster et al. (2010); Koster et al. (2011)), all models were required to include a land surface model and initialize both the atmosphere and land. The SubX project has also performed more than one year of real-time forecasts, beginning July 2017. During this demonstration period, forecasts were required to be made available to NCEP/CPC by 6pm every Wednesday. This requirement was relaxed to 6am Thursday partway through the real-time demonstration period. All data were provided on a uniform 1°x1° longitude-latitude grid as full fields to both NCEP/CPC for their internal use and the International Research Institute for Climate and Society Data Library (IRIDL) for public dissemination[1] (Kirtman et al. 2017).

a. Models

Seven modeling groups participate in SubX. These are:

- National Centers for Environmental Prediction (NCEP) Climate Forecast System, version 2 (NCEP-CFSv2);
- NCEP Environmental Modeling Center, Global Ensemble Forecast System (EMC-GEFS);
- Environmental and Climate Change Canada Global Ensemble Prediction System, Global Environmental Multi-scale Model (ECCC-GEM);
- National Aeronautics and Space Administration, Global Modeling and Assimilation Office, Goddard Earth Observing System (GMAO-GEOS);
- Naval Research Laboratory, Navy Earth System Model (NRL-NESM);
- National Center for Atmospheric Research Community Climate System Model, version 4 run at the University of Miami Rosenstiel School for Marine and Atmospheric Science (RSMAS-CCSM4);
- National Oceanic and Atmospheric Administration, Earth System Research Laboratory, Flow-Following Icosahedral Model (ESRL-FIM).

For additional details, see Table 1.

[1] http://iridl.ldeo.columbia.edu/SOURCES/.Models/.SubX/

All groups have provided re-forecasts for the 1999-2015 period with the exception of ECCC-GEM (1999-2014)[2] and most have provided additional re-forecasts to fill the gap between the end of the SubX re-forecast period and beginning of the real-time forecasts in July 2017. Five of the groups use fully coupled atmosphere-ocean-land-sea ice models (NCEP-CFSv2, GMAO-GEOS, NRL-NESM, RSMAS-CCSM4, ESRL-FIM), while two groups use models with atmosphere and land components forced with prescribed sea surface temperatures (EMC-GEFS, ECCC-GEM). In the EMC-GEFS forecast system, SSTs are specified by relaxing the SST analysis to a combination of climatological SST and bias-corrected SST from operational NCEP-CFSv2 forecasts. The longer the lead time, the more weighting given to the bias-corrected NCEP-CFSv2 forecast SST. In the ECCC-GEM forecast system, the SST anomaly averaged from the previous 30 days is persisted in the forecast. The sea-ice cover is adjusted in order to be consistent with the SST change. Most groups provide 4 ensemble members for the re-forecasts (NCEP-CFSv2, ECCC-GEM, GMAO-GEOS, NRL-NESM, ESRL-FIM) with some groups using lagged ensembles and others using their own ensemble generation systems to produce initial conditions. Some groups provide additional ensemble members in real-time (e.g. RSMAS-CCSM4, EMC-GEFS).

b. Description of Datasets

There is a demand for many S2S-relevant variables from the research community for evaluating a range of S2S phenomena. This demand together with daily output frequency, weekly initial conditions, seven models, and three or more ensemble members places extremely high demands on the data server, therefore a priority for fields to be distributed was defined. Ten fields were identified as critical to supporting NCEP/CPC operational products and were designated as Priority 1 variables. These variables include, geopotential height at 200 and 500 hPa, zonal and meridional

[2]ECCC-GEM runs their re-forecasts on the fly as part of their operational practice and will fill in 2015 at a later date

winds at 200 and 850 hPa, temperature at 2m, precipitation, surface temperature (SST + Land), and outgoing longwave radiation (see Table 2). This paper will focus on evaluation of the models using these Priority 1 variables. A second set of 21 additional fields have been identified as key variables for supporting S2S research, labelled Priority 2 variables (see Table 3). Both priority 1 and 2 variables are publicly available through the IRIDL.

3. Re-forecast Evaluation

a. Verification Datasets

Calculation of skill requires a verifying observational dataset. Where applicable, the datasets used correspond to those used by NCEP/CPC for verification of their forecasts. For 2m temperature over land, the CPC daily temperature dataset with horizontal resolution of 0.5°x0.5° is used[3]. This data is provided as a maximum and minimum daily temperature, thus the average daily temperature is calculated as the average of Tmax and Tmin (Fan and Van Den Dool 2008). For precipitation over land, the CPC Global Daily Precipitation dataset (0.5°x0.5°) is used (Xie et al. (2007); Chen et al. (2008)). Verification datasets are re-gridded to the coarser SubX model resolution of 1° x 1° prior to performing model evaluation.

We also evaluate the skill of two subseasonal phenomena that are known sources of S2S predictability - the Madden-Julian Oscillation (MJO) and the North Atlantic Oscillation (NAO). The MJO skill is evaluated using the real-time multivariate MJO index (RMM) (Wheeler and Hendon 2004). The observed index is calculated using the NCEP/NCAR Reanalysis (Kalnay et al. 1996) and NOAA Interpolated OLR (Liebmann and Smith 1996). The NAO is defined as the projection of the winter geopotential height at 500 hPa (Z500) onto the leading North Atlantic EOF spatial

[3]The original data can be found at ftp://ftp.cpc.ncep.noaa.gov/precip/PEOPLE/wd52ws/global_temp/

pattern of Z500 (0°-90°N, 93°W-47°E). The observed NAO index is calculated using 500 hPa geopotential height from NCEP/NCAR Reanalysis (Kalnay et al. 1996).

b. Multi-model Ensemble

Since the SubX models are initialized on different days, it is challenging to produce a MME (e.g. Vitart (2017)). In SubX, we choose to align the target dates of each model to produce a MME. Following nearly the same procedure used for NCEP/CPC real-time forecasts, Saturday is defined as the first day of a given week. All re-forecasts for all models that are produced during the prior week (previous Saturday through Thursday) are used to produce a MME forecast for weeks 1-4 individually, where week 1 is defined as the first Sat-Fri interval. Friday initializations are not included in an attempt to mimic real-time forecast procedures. In real-time, forecasts provided after Thurs 6am cannot be processed in time to be used by the forecasters. This procedure, which also involves forming averages of daily forecasts over the appropriate week, is repeated for weeks 2 through 4. Weeks 3 and 4 are then averaged together to produce week 3-4 forecasts. Using this procedure, a multi-model ensemble re-forecast, equally weighted by *model* can be produced by averaging the ensemble means of each of the models for their week 3-4 forecasts. We choose to equally weight by model when evaluating the re-forecasts in order to understand the contribution of each model to the MME. There are some potential drawbacks to the MME procedure. For example, some models will contribute older forecasts to the MME than others, depending on their initialization date. The extent to which decreased skill with longer lead time is balanced by increased ensemble size and model diversity in such an ensemble remains an open research question. Additionally, since the period over which forecasts are obtained is Sat-Thurs (a 6-day period, used to mimic the 6-day period of real-time forecast initializations described in Section

4) and some of the models initialize once every 7 days, there are times when a model will not be included in the MME, depending on how the re-forecast dates fall. This occurs with the ECCC-GEM and RSMAS-CCSM4 models. Finally, in rare cases, it is not possible to produce a week 3-4 forecast for the ECCC-GEM model since part of week 4 is not available due to the re-forecast initialization day and 32-day re-forecast length.

c. Model Biases

A forecast is typically initialized with an analysis in which observations have been assimilated, thereby constraining the analysis to represent the observed state as close as possible. As the forecast time increases, the model state on average moves from the observed climate towards a model-intrinsic climate, which is typically biased. Therefore, it is common practice in S2S predictions to estimate and remove the mean forecast bias using a set of re-forecasts (Smith et al. 1999). Additionally, the skill of forecasts at the S2S timescale is typically evaluated in terms of anomalies or differences from the mean climate, thus requiring a climatology based on re-forecasts. Both of these needs are met by determining the mean climate (i.e. climatology) as a function of lead time and initialization date. For seasonal predictions using monthly data, it is typical to calculate the model climatology as a multi-year average for each forecast start month and lead or target time (Tippett et al. 2018). However, calculation of the climatology is not trivial for subseasonal re-forecasts due to differences in initialization day and frequency among models. For example, some forecast models are initialized on the same Julian days every year while others are initialized on a day-of-the-week schedule, meaning that the Julian initialization dates shift from year to year. In the first case, the 17-year re-forecast period yields 17 model runs on some calendar dates and none on the rest. In the second case, only 2-3 model runs are available for each day of the year from which to determine the climatology. An additional challenge for the SubX project was that a climatology was needed to produce bias-corrected forecast anomalies in real-time for NCEP/CPC prior to the completion of the re-forecasts at some centers. The methodology described here was developed by the SubX Team to resolve these issues and is used for producing SubX real-time forecasts and model evaluation.

To compute the climatology, the first step is to calculate ensemble means for individual days of each forecast run. For most groups, lagged ensembles are produced using initialization dates from different hours of the same initialization day; these are averaged to yield ensemble means for the 24-h period spanning each forecast day. In the case of the NRL-NESM, which produces ensemble means over runs started on four consecutive days because ocean data assimilation is based on a 24-hour update cycle, the ensemble mean consists of a single member for each day. Next, for each day of the year (1-366), a multi-year average of the ensemble means is calculated. Depending on how model runs are scheduled, this may not produce a climatology for each day of the year for some models. Finally, a triangular smoothing window of 31 days (+/- 15 days) is applied in a periodic fashion such that late-December smoothing includes early January values and vice versa. This approach means that the forecast climatology can be computed from a partial re-forecast database and only re-forecasts with nearby initializations are required. Due to drift from the initial quasi-observed state to the models own internal mean state, the climatology for a given calendar day is expected to be different for different lead times. Therefore, the above procedure is performed for each lead time and each model individually. Removal of this climatology from the corresponding full fields produces anomalies and effectively performs a mean bias correction (Becker et al. 2014). Climatologies for the Priority 1 variables have been computed following this procedure and are available from the IRIDL.

Comparison of the model climatology with the observed climatology allows us to evaluate the model mean biases and their evolution at subseasonal timescales. While mean biases have been evaluated extensively at the monthly and seasonal timescales (e.g. Jin et al. (2008); Saha et al. (2014)), they have not been comprehensively evaluated in models at the subseasonal timescale, except in the context of the MJO (e.g. Agudelo et al. (2008); Hannah et al. (2015); Kim (2017); Lim et al. (2018); Janiga et al. (2018)). Two exceptions are Sun et al. (2018a) and Guan et al. (2018, manuscript submitted to WAF). These studies evaluate the mean biases in the ESRL-FIM and EMC-GEFS re-forecasts used in SubX, respectively. Evaluations of model biases are particularly important since there is evidence that model prediction errors are related to model mean bias errors (e.g. Lee et al. (2010); DelSole and Shukla (2010); Green et al. (2017)). The extent to which this is the case at subseasonal timescales is unknown. To evaluate the overall biases in the SubX system the average mean bias over all seven SubX models for week 1 (days 1-7) and week 4 (days 22-28) are calculated as model climatology minus observed climatology for 2m Temperature (Figure 1) and Precipitation (Figure 2), similar to Sun et al. (2018a). Observed climatology is calculated using the same methodology described above for the models with the verification datasets used by NCEP/CPC for temperature and precipitation (Section 3a). Model biases are already well established in both temperature and precipitation at week 1. On average, warm biases are evident in the central U.S. with the strongest biases $>1.5°$C during Jun-Jul-Aug (JJA). These warm biases are reduced by week 4 for re-forecasts initialized in Dec-Jan-Feb (DJF) and Sep-Oct-Nov (SON), but are increased for those initialized in Mar-Apr-May (MAM) and JJA. In DJF, cold biases are also present which increase to week 4, while re-forecasts initialized in SON show small changes from week 1 to week 4. For precipitation, a summer dry bias is evident in the central U.S. at week 1, which grows slightly to week 4. While model biases generally grow in amplitude from week 1 through week 4, increases in biases with lead days are smaller at longer leads and may be approaching saturation near the end of week 4. Overall, the SubX mean bias has a larger seasonal cycle than observed. The average bias over all models is generally smaller than any individual model biases in both temperature and precipitation (not shown).

d. Global and North America Skill Assessment

In this section, we evaluate the skill for the individual and multi-model combination of the SubX models using both deterministic and probabilistic skill measures. The skill assessment is performed for temperature and precipitation over land for global and North America domains. In most cases, the MME outperforms any individual model, one of the benefits of using a MME (Hagedorn et al. (2005); Weigel et al. (2008); Weisheimer and Reyes (2009); Kirtman et al. (2014); Becker et al. (2014); Becker and Van Den Dool (2016)).

1) Deterministic Skill

The deterministic skill of SubX re-forecasts is evaluated using the anomaly correlation coefficient (ACC) and root mean square error (RMSE). For temperature and precipitation, the results using both metrics are similar, therefore only the ACC is shown here. The ACC is calculated using the ensemble mean for each model.

Since the subseasonal timescale begins at week 2, we start by evaluating the DJF initialized re-forecasts with the ACC of global temperature and precipitation for week 2 (Figure 3). Most regions of the globe have ACC > 0.5 for 2m temperature at 2-weeks. For precipitation, there are substantially large regions with ACC >0.5, including the western U.S., east Asia, and Brazil.

Next, we evaluate week 3-4 skill over North America, the region and timescale relevant to NCEP/CPC outlooks. The week 3-4 MME ACC over North America is shown in Figure 4 for

2m Temperature and Figure 5 for precipitation for re-forecasts initialized over four seasons. Consistent with previous studies, winter skill is higher than summer skill for both temperature and precipitation (e.g. DelSole et al. (2017)). Temperature skill is positive for all seasons with regions of ACC >0.2 over most of North America with the exception of a few high latitude locations. Additionally, regions of skill >0.4 are also evident in each season. As expected, precipitation skill is lower than temperature, but there are substantial regions in each season for which the MME ACC >0.2. Figure 6 provides a comparison of the average ACC over North America for week 3-4 for the individual models and the MME. It is clear that although overall skill is low due to aggregation of low and high skill grid points, the MME exceeds the skill of any individual model in all seasons. It is also noted that there is no clear stratification in skill by model configuration (e.g. number of ensemble members, coupled vs. uncoupled, operational vs. research).

2) Probabilistic Skill

The SubX models are also evaluated using probabilistic skill scores, specifically, the ranked probability skill score (RPSS), for tercile categories of above, near, and below normal. Due to the small ensemble size of individual models, RPSS is calculated only for the full multi-model ensemble (typically 34 members). Figures 7 and 8 show the RPSS for week 3-4 North American 2m temperature and precipitation. Positive RPSS indicates skill better than a forecast of climatology, therefore any region with positive RPSS can be considered skillful. There are substantial regions and seasons of skill better than climatology for 2m temperature (Figure 7). For precipitation, skill is evident in spring and fall in the western and central U.S. (Figure 8).

3) Pattern Skill

The skill of SubX re-forecasts also can be assessed in terms of their pattern structure. A question of particular interest is whether the multi-model mean has significantly more skill than an individual model. This question can be addressed using the random walk test of DelSole and Tippett (2016), which is evaluated as follows. For each 2-week mean hindcast, the pattern correlation with respect to observations over U.S. and Canada is computed. The random walk score is a function of time that starts at zero and, for each hindcast, goes up one unit if the multi-model mean has a larger pattern correlation than the model being compared, otherwise the score goes down one unit. The score is tallied for each SubX model separately. Hypothetically, if the two hindcasts being compared are equally skillful, then the odds are 1:1 that the score will go up or down by one unit, in which case the average score should be zero and a 95% confidence interval is approximately $2\sqrt{N}$, where N is the number of independent verifications. To avoid verification periods that overlap with each other, only initial conditions separated by two or more weeks are considered. The resulting random walk scores for week 3-4 2m-temperature and precipitation are shown in figs. 9a-b. The scores for different seasons and years are concatenated. As seen in the figure, the score is positive for each model by the end of the period, indicating that the multi-model mean has larger pattern correlation more frequently than any single model. Moreover, the score is statistically significant at the 5% level in all cases except one, namely the CFSv2 hindcasts of 2m-temperature (although the score still is positive). These results demonstrate that the multi-model mean predicts the anomaly pattern for temperature and precipitation more skillfully, more frequently, than any individual model, and this frequency is statistically significant in almost all cases considered.

e. Sources of Subseasonal Predictability

A number of potential sources of predictability have been identified for the subseasonal timescales (National Research Council (2010); National Academies of Sciences, Engineering and Medicine (2017)). Correctly simulating the relevant processes and predicting their impacts is the key to successful subseasonal prediction; they should therefore be fully explored in subseasonal re-forecast databases. The available Priority 1 variables (Section 2b and Table 2) allow us to evaluate the skill of two of these predictability sources in the SubX models: the MJO and NAO.

1) The Madden-Julian Oscillation

The Madden-Julian Oscillation is the largest source of tropical variability on the subseasonal timescale. The MJO affects temperature and precipitation in the extratropics through various mechanisms, including the NAO (Cassou (2008); Lin et al. (2009)) and atmospheric rivers (e.g. Guan et al. (2012); Mundhenk et al. (2018)), among others (Zhang (2013); see Stan et al. (2017) for a review of MJO teleconnections). Given its impact, prediction of the MJO is considered a key component of a skillful subseasonal prediction system. Therefore, we evaluate its skill in SubX in terms of the bivariate ACC and RMSE for ensemble mean re-forecasts initialized Nov-Mar (Figure 10) (Rashid et al. (2010)). The skill of each model and the MME are calculated weekly and for weeks 3-4 combined, following the SubX MME ensemble methodology (Section 3b). Most SubX models have ACC >0.5 and RMSE < 1.4 out to week 3-4. This range of prediction skill is similar to the MJO skill of the WWRP/WCRP S2S models, with the exception of the ECMWF model which far exceeds the skill of any other S2S or SubX model (Vitart 2017). It is of interest that the two most skillful models have very different configurations. The GMAO-GEOS model is a fully coupled atmosphere-ocean-land-sea ice model that has contributed to the monthly and seasonal NMME. GMAO-GEOS contributes only 4 ensemble members in SubX. In contrast, the base model of EMC-GEFS (i.e. Global Forecast System) is a NWP atmosphere-land model forced with prescribed SST. The SubX version of GEFS takes into account the day-to-day SST variability from the bias-corrected operational NCEP-CFSv2 forecast and contributes 11 ensemble members to the SubX re-forecasts. The MME is more skillful than any individual model in both metrics.

2) THE NORTH ATLANTIC OSCILLATION

One of the key sources of extratropical subseasonal variability is the NAO, which has been linked to periods of extreme winter weather on subseasonal timescales in Eastern North America and Europe (e.g Hurrell et al. (2010)). Until recently, there was little evidence that the NAO could be skillfully predicted beyond weather timescales (e.g. Johansson (2007); Kim et al. (2012)); however, recent studies have found that the United Kingdom Met Office (UKMET) seasonal prediction system can produce skillful monthly predictions of the NAO up to 1-year due to high resolution in both the atmosphere (0.83° longitude by 0.55° latitude) and ocean (0.25° longitude-latitude) models, large-ensembles (>20 members), and long re-forecast periods ($\sim$ 40 years) (Scaife et al. (2014); Dunstone et al. (2016)). Given this newly found predictability of the NAO and its potential impacts on extreme weather at S2S timescales, we evaluate the skill of the NAO in the SubX models. Figure 11 shows the ensemble mean anomaly correlation (left) and RMSE (right) of the SubX models forecasting the NAO index averaged for weeks 1-4 individually and for weeks 3-4 combined using initialization dates during the northern hemisphere winter (Dec-Jan-Feb). The skill of each model and the MME are calculated following the SubX MME ensemble methodology (Section 3b). The most skillful models and the MME have ACC > 0.5 and RMSE < 1.4 to week 2. The MME has similar skill to the most skillful models in both metrics. However, the *week*

3-4 skill of the 34-member SubX MME is not as skillful as the *monthly* correlations found in the UKMET seasonal prediction system (Scaife et al. 2014).

4. Real-time Forecasts

SubX produces real-time forecasts each week and provides them to NCEP/CPC as dynamical guidance for their official week 3-4 temperature outlook and experimental week 3-4 precipitation outlook. These outlooks show regions of increased probability of above-normal or below-normal (i.e. two category) temperature and precipitation, and regions where the probabilities of above or below normal are equal (i.e. 50/50 chance of above or below normal). To illustrate, the official week 3-4 temperature and precipitation outlook produced on 6 July 2018 is shown in Figure 12. Recall that we evaluated the probabilistic skill of 3-category re-forecasts in Section 3. Ideally, we would be able to produce skillful forecasts that can differentiate between more than two categories. However, the two category probabilities are used for real-time forecasts because they are currently more skillful.

Forecast guidance products have been developed at NCEP/CPC using the SubX forecasts for 500hPa geopotential height, 2m temperature, and precipitation. For temperature and precipitation, MME bias corrected anomalies and probabilistic guidance products are shown in Figure 12 (left). The procedure for producing these guidance products is shown schematically in Figure 13. NCEP/CPC collects the weekly forecast data from each modeling group every Thursday by 6am ET, using the most recently initialized forecast runs available for each model from the prior Friday through Wednesday, with the latest initialization from 00 UTC Thursday provided by ECCC-GEM. Bias-corrected anomalies are calculated for each model and ensemble member using the re-forecast climatologies described in section 3c. From these anomalies, the week 3-4 multi-model mean anomalies are produced by averaging each ensemble member from each model, thus in the real-time forecasts each ensemble *member* is given equal weight in calculating the multi-model mean (Figure 12, upper left panels); recall that in Section 3b, multi-model results gave each *model* equal weight. Since some models produce additional ensemble members in real-time (Table 1), the SubX real-time forecasts have 78 ensemble members, while the MME re-forecasts described in Section 3 typically have 34 ensemble members. Each ensemble member is given equal weight in real-time forecast anomalies so that the multi-model anomaly forecasts are consistent with the multi-model probability forecasts. A preliminary analysis of multi-model ensemble anomaly correlations showed that multi-model anomalies that equally weighted ensemble *members* were more skillful than those that equally weighted *models* (not shown). This suggests that the ensemble mean anomalies of models with fewer ensemble members are less skillful, however individual ensemble members may be equally skillful. Determining the optimal weighting procedure is an active area of research. Probability guidance of above- and below-normal are then derived by counting the number of ensemble members from all model runs that exceed or do not exceed the individual model's climatological mean. The probabilistic map is produced for the 'above-only' category (cf. Figure 12) and probabilities of below-normal are inferred to be one minus the probability of above-normal.

Using guidance from SubX and other tools, NCEP/CPC forecasters produce the official maps for week 3-4 outlooks. These maps for July 6, 2018 temperature and precipitation show above- and below-normal areas consistent with the corresponding probabilities and anomalies from the SubX multi-model ensemble, demonstrating the use of SubX in the NCEP/CPC official outlooks (Figure 12).

5. Concluding Remarks

This paper introduces SubX to the S2S community. SubX is a multi-model R2O project in which seven models have produced a suite of historical re-forecasts and also provide weekly real-time forecasts. The re-forecast database has been completed and the real-time forecasts have been operating for over a year. Both real-time and re-forecasts are publicly available through the IRI Data Library. We wish to emphasize that the SubX database is complementary to the WWRP/WCRP S2S prediction project database. The inclusion of research and operational models and availability of both real-time and retrospective forecasts in SubX provides a unique contribution to community efforts in subseasonal predictability and prediction.

Here we have provided an initial assessment of subseasonal biases and skill for the SubX models as well as a demonstration of the SubX contribution to real-time operational predictions. There have been few evaluations of model biases for subseasonal timescales. We show that for the SubX models, bias patterns over the U.S. are already well established at week 1 and grow to week 4. Further research should evaluate the impact of these biases on prediction skill. The SubX MME demonstrates skill for week 3-4 predictions of temperature and precipitation in specific regions and seasons. This is confirmed using both probabilistic and deterministic skill metrics. On average, the MME is more skillful than individual models over North America. We also evaluated the skill of MJO and NAO predictions. MJO skill is comparable with most of the WWRP/WCRP S2S models. However, we have evaluated only a single metric. Future work should explore a broader range of MJO metrics. The NAO skill is also comparable to other modeling systems with the exception of the UKMET. Future work should explore the model configuration necessary to produce NAO skill consistent with the UKMET system. Finally, we have demonstrated that SubX can provide useful MME guidance to NCEP/CPC operational products in real-time. All seven modeling groups, including research models, have provided SubX forecasts each week on time throughout the real-time demonstration period. In addition to the results shown in this paper, many additional images showing model skill and biases are available on the SubX website [4].

The results shown in this paper have only scratched the surface of potential research on subseasonal predictability and prediction. With the availability of subseasonal re-forecast databases such as SubX and WWRP/WCRP S2S, it is now possible for the research community to extensively explore the full range of subseasonal predictability, and to develop methodologies for S2S post-processing including forecast calibration and multi-model ensembling (e.g. Vigaud et al. (2017a); Vigaud et al. (2017b)). The availability of real-time subseasonal forecasts in SubX also enables the development of real-time forecast demonstration prototypes for applications use in various socio-economic sectors. We encourage the community to utilize the SubX database to these ends.

Finally, we wish to highlight that the SubX database is also an ideal framework for testing model improvements for subseasonal predictions. For example, Sun et al. (2018, manuscript in preparation) have already undertaken an effort to test the impact of including more model levels to resolve the stratosphere following the SubX re-forecast protocol. This has made it possible to compare the results of their model improvements in a prediction framework and against the suite other SubX models. Colleagues at NRL are also testing the impact of better resolving the stratosphere in their model (N.Barton, personal communication). Additionally, Green et al. (2017) and Sun et al. (2018b) have used the SubX framework for testing the impact of a new subgrid-scale convection scheme. We encourage future model development efforts to utilize SubX as a framework for improving subseasonal predictions.

[4]http://cola.gmu.edu/kpegion/subx/

Acknowledgments. The SubX project is funded and was initiated by NOAAs Climate Program Offices Modeling, Analysis, Predictions, and Projections program (MAPP) in partnership with the NASA Modeling, Analysis, and Prediction program (MAP); the Office of Naval Research; and NOAAs NWS Office of Science and Technology Integration. Relevant NOAA award numbers are: NA16OAR4310149, NA16OAR4310151, NA16OAR4310150, NA16OAR4310143, NA16OAR4310141, NA16OAR4310146, NA16OAR4310145, NA16OAR4310148. S. Sun and B. W. Green are supported by funding from NOAA Award NA17OAR4320101. N. Barton and E.J. Metzger were funded by the Navy ESPC in the North-American Multi Model Ensemble project sponsored by the Office of Naval Research. Computer time for NRL-NESM was provided by the Department of Defense High Performance Computing Modernization Program. This is NRL contribution NRL/JA/7320-18-4121. Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, MD, USA and Universities Space Research Association, Columbia, MD, USA. CPC Precipitation and Temperature, NCEP/NCAR Reanalysis, and NOAA Interpolated OLR data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at https://www.esrl.noaa.gov/psd/. The Center for Ocean-Land-Atmosphere studies (COLA) provided extensive disk space for performing the model evaluations and also hosts the SubX Website. COLA support for SubX is provided by grants from NSF (1338427) and NASA (NNX14AM19G) and a Cooperative Agreement with NOAA (NA14OAR4310160).

References

Agudelo, P. A., C. D. Hoyos, P. J. Webster, and J. A. Curry, 2008: Application of a serial extended forecast experiment using the ECMWF model to interpret the predictive skill of tropical intraseasonal variability. *Climate Dyn.*, **32 (6)**, 855–872.

Becker, E., H. v. den Dool, and Q. Zhang, 2014: Predictability and Forecast Skill in NMME. *J. Climate*, **27 (15)**, 5891–5906.

Becker, E., and H. Van Den Dool, 2016: Probabilistic Seasonal Forecasts in the North American Multimodel Ensemble: A Baseline Skill Assessment. *J. Climate*, **29 (8)**, 3015–3026.

Black, J., N. C. Johnson, S. Baxter, S. B. Feldstein, D. S. Harnos, and M. L. L'Heureux, 2017: The Predictors and Forecast Skill of Northern Hemisphere Teleconnection Patterns for Lead Times of 3–4 Weeks. *Mon. Wea. Rev.*, **145 (7)**, 2855–2877.

Brunet, G., and Coauthors, 2010: Collaboration of the Weather and Climate Communities to Advance Subseasonal-to-Seasonal Prediction. *Bull. Amer. Meteor. Soc.*, **91 (10)**, 1397–1406.

Cassou, C., 2008: Intraseasonal interaction between the Madden–Julian Oscillation and the North Atlantic Oscillation. *Nature*, **455 (7212)**, 523–527.

Chen, M., W. Shi, P. Xie, V. B. S. Silva, V. E. Kousky, R. Wayne Higgins, and J. E. Janowiak, 2008: Assessing objective techniques for gauge-based analyses of global daily precipitation. *J. Geophys. Res.*, **113 (D4)**, D04 110–13.

DelSole, T., and J. Shukla, 2010: Model Fidelity versus Skill in Seasonal Forecasting. *J. Climate*, **23 (18)**, 4794–4806.

DelSole, T., and M. K. Tippett, 2016: Forecast comparison based on random walks. *Mon. Wea. Rev.*, **144 (2)**, 615–626.

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 (10)**, 3499–3512.

Dunstone, N., D. Smith, A. Scaife, and L. Hermanson, 2016: Skilful predictions of the winter North Atlantic Oscillation one year ahead. *Nature*, **9**, 809–815.

Fan, Y., and H. Van Den Dool, 2008: A global monthly land surface air temperature analysis for 1948–present. *J. Geophys. Res.*, **113 (D1)**, D01 103–18.

Green, B. W., S. x, 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 (7)**, 2555–2574.

Guan, B., D. E. Waliser, N. P. Molotch, E. J. Fetzer, and P. J. Neiman, 2012: Does the Madden–Julian Oscillation Influence Wintertime Atmospheric Rivers and Snowpack in the Sierra Nevada? *Mon. Wea. Rev.*, **140 (2)**, 325–342.

Hagedorn, R., F. D. REYES, and T. N. Palmer, 2005: The rationale behind the success of multimodel ensembles in seasonal forecasting. *Tellus*, **57A**, 219–233.

Hamill, T. M., G. T. Bates, J. S. WHITAKER, D. R. Murray, M. Fiorino, T. J. Galarneau, Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Dataset. *Bull. Amer. Meteor. Soc.*, **94 (10)**, 1553–1565.

Hannah, W. M., E. D. Maloney, and M. S. Pritchard, 2015: Consequences of systematic model drift in DYNAMO MJO hindcasts with SP-CAM and CAM5. *J. Adv. Modeling and Earth Systems.*, **7 (3)**, 1051–1074.

Hogan, T., and Coauthors, 2014: The Navy Global Environmental Model. *Oceanography*, **27 (3)**, 116–125.

Hurrell, J. W., Y. Kushnir, G. Ottersen, and M. Visbeck, 2010: An overview of the North Atlantic Oscillation. *The North Atlantic Oscillation: Climatic Significance and Environmental Impact*, American Geophysical Union, Washington, D. C., 1–35.

Infanti, J. M., and B. P. Kirtman, 2016: Prediction and predictability of land and atmosphere initialized CCSM4 climate forecasts over North America. *J. Geophys. Res.*, **121 (21)**, 12,690–12,701.

Janiga, M. A., C. J Schreck III, J. A. Ridout, M. Flatau, N. P. Barton, E. J. Metzger, and C. A. Reynolds, 2018: Subseasonal Forecasts of Convectively Coupled Equatorial Waves and the MJO: Activity and Predictive Skill. *Mon. Wea. Rev.*, **146 (8)**, 2337–2360.

Jin, E. K., and Coauthors, 2008: Current status of ENSO prediction skill in coupled ocean–atmosphere models. *Climate Dyn.*, **31 (6)**, 647–664.

Johansson, Å., 2007: Prediction Skill of the NAO and PNA from Daily to Seasonal Time Scales. *J. Climate*, **20 (10)**, 1957–1975.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, **77**, 437–472.

Kim, H.-M., 2017: The impact of the mean moisture bias on the key physics of MJO propagation in the ECMWF reforecast. *J. Geophys. Res.*, **122 (15)**, 7772–7784.

Kim, H.-M., P. J. Webster, and J. A. Curry, 2012: Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter. *Climate Dyn.*, **39 (12)**, 2957–2973.

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 (4)**, 585–601.

Kirtman, B. P., and Coauthors, 2017: The subseasonal experiment (subx). IRI Data Library, doi: 10.7916/d8pg249h.

Koster, R. D., M. J. Suarez, A. Ducharne, M. Stieglitz, and P. Kumar, 2007: A catchment-based approach to modeling land surface processes in a general circulation model: 1. Model structure. *J. Geophys. Res.*, 1–14.

Koster, R. D., and Coauthors, 2010: Contribution of land surface initialization to subseasonal forecast skill: First results from a multi-model experiment. *Geophys. Res. Lett.*, **37 (2)**, L02 402–18.

Koster, R. D., and Coauthors, 2011: The Second Phase of the Global Land–Atmosphere Coupling Experiment: Soil Moisture Contributions to Subseasonal Forecast Skill. *J. Hydrometeor.*, **12 (5)**, 805–822.

Lee, J.-Y., and Coauthors, 2010: How are seasonal prediction skills related to models' performance on mean state and annual cycle? *Climate Dyn.*, **35 (2-3)**, 267–283.

Li, S., and A. W. Roberston, 2015: Evaluation of submonthly precipitation forecast skill from global ensemble prediction systems. *Mon. Wea. Rev.*, **143**, 2871–2889.

Liebmann, B., and C. Smith, 1996: Description of a complete (interpolated) outgoing longwave radiation dataset. *Bull. Amer. Meteor. Soc.*, **77**, 1275–1277.

Lim, Y., S.-W. Son, and D. Kim, 2018: MJO Prediction Skill of the Subseasonal-to-Seasonal Prediction Models. *J. Climate*, **31 (10)**, 4075–4094.

Lin, H., G. Brunet, and J. Derome, 2009: An Observed Connection between the North Atlantic Oscillation and the Madden–Julian Oscillation. *J. Climate*, **22 (2)**, 364–380.

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 (12)**, 4867–4883.

Mariotti, A., P. M. Ruti, and M. Rixen, 2018: Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. *npj Climate and Atmospheric Science*, 1–4.

Metzger, E. J., and Coauthors, 2014: US Navy Operational Global Ocean and Arctic Ice Prediction Systems. *Oceanography*, **27 (3)**, 32–43.

Molod, A., L. Takacs, M. J. Suarez, J. Bacmeister, I.-S. Song, and A. Eichmann, 2012: The Geos-5 Atmospheric General Circulation Model: Mean Climate and Development From Merra to Fortuna . Tech. Rep. TM–2012-104606, NASA.

Mundhenk, B. D., E. A. Barnes, E. D. Maloney, and C. F. Baggett, 2018: Skillful empirical subseasonal prediction of landfalling atmospheric river activity using the Madden–Julian oscillation and quasi-biennial oscillation. *npj Climate and Atmospheric Science*, 1–7.

National Academies of Sciences, Engineering and Medicine, 2017: Next Generation Earth System Prediction: Strategies for Subseasonal to Seasonal Forecasts. Tech. rep., The National Academies Press, Washington, DC.

National Research Council, 2010: *Assessment of Intraseasonal to Interannual Climate Prediction and Predictability*. National Academies Press, Washington, D.C.

Neena, J. M., J. Y. Lee, D. Waliser, and B. Wang, 2014: Predictability of the Madden–Julian Oscillation in the Intraseasonal Variability Hindcast Experiment (ISVHE)*. *J. Climate*, **27**, 4531–4543.

Palmer, T. N., A. Alessandri, U. A. B. o. the, and 2004, 2004: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bull. Amer. Meteor. Soc.*, **85**, 853–872.

Pegion, K., and P. D. Sardeshmukh, 2011: Prospects for Improving Subseasonal Predictions. *Mon. Wea. Rev.*, **139 (11)**, 3648–3666.

Rashid, H. A., H. H. Hendon, M. C. Wheeler, and O. Alves, 2010: Prediction of the Madden–Julian oscillation with the POAMA dynamical prediction system. *Climate Dyn.*, **36 (3-4)**, 649–661.

Reichle, R., and Q. Liu, 2014: Observation-Corrected Precipitation Estimates in GEOS-5. Tech. Rep. TM–2014-104606, NASA.

Rienecker, M. M., and Coauthors, 2008: The GEOS-5 Data Assimilation System— Documentation of Versions 5.0.1, 5.1.0, and 5.2.0. Tech. Rep. TM–2008–104606, NASA.

Robertson, A. W., A. Kumar, M. Peña, and F. Vitart, 2015: Improving and Promoting Subseasonal to Seasonal Prediction. *Bull. Amer. Meteor. Soc.*, **96 (3)**, ES49–ES53.

Saha, S., and Coauthors, 2014: The NCEP Climate Forecast System Version 2. *J. Climate*, **27 (6)**, 2185–2208.

Scaife, A. A., A. Arribas, and E. Blockley, 2014: Skillful long-range prediction of European and North American winters. *Geophys. Res. Lett.*, **41**, 2514–2519.

Smith, T. M., R. L. J. o. Climate, and 1999, 1999: GCM systematic error correction and specification of the seasonal mean Pacific–North America region atmosphere from global SSTs. *J. Climate*, **12 (1)**, 273–288.

Stan, C., D. M. Straus, J. S. Frederiksen, H. Lin, E. D. Maloney, and C. Schumacher, 2017: Review of Tropical-Extratropical Teleconnections on Intraseasonal Time Scales. *Rev. Geophys.*, **55 (4)**, 902–937.

Sun, S., R. Bleck, S. G. Benjamin, B. W. Green, and G. A. Grell, 2018a: Subseasonal Forecasting with an Icosahedral, Vertically Quasi-Lagrangian Coupled Model. Part I: Model Overview and Evaluation of Systematic Errors. *Mon. Wea. Rev.*, **146 (5)**, 1601–1617.

Sun, S., B. W. Green, R. Bleck, and S. G. Benjamin, 2018b: Subseasonal Forecasting with an Icosahedral, Vertically Quasi-Lagrangian Coupled Model. Part II: Probabilistic and Deterministic Forecast Skill. *Mon. Wea. Rev.*, **146 (5)**, 1619–1639.

Swinbank, R., and Coauthors, 2016: The TIGGE Project and Its Achievements. *Bull. Amer. Meteor. Soc.*, **97**, 49–67.

Tippett, M. K., L. Trenary, T. DelSole, K. Pegion, and M. L. L'Heureux, 2018: Sources of Bias in the Monthly CFSv2 Forecast Climatology. *J. Appl. Meteor. Climatol.*, **57 (5)**, 1111–1122.

Tompkins, A. M., and Coauthors, 2017: The Climate-System Historical Forecast Project: Providing Open Access to Seasonal Forecast Ensembles from Centers around the Globe. *Bull. Amer. Meteor. Soc.*, **98 (11)**, 2293–2301.

Vigaud, N., A. W. Robertson, and M. K. Tippett, 2017a: Multimodel Ensembling of Subseasonal Precipitation Forecasts over North America. *Mon. Wea. Rev.*, **145 (10)**, 3913–3928.

Vigaud, N., A. W. Robertson, M. K. Tippett, and N. Acharya, 2017b: Subseasonal Predictability of Boreal Summer Monsoon Rainfall from Ensemble Forecasts. *Frontiers in Environmental Science*, **5**, 2197–19.

Vitart, F., 2017: Madden-Julian Oscillation prediction and teleconnections in the S2S database. *Quart. J. Roy. Meteor. Soc.*, **143 (706)**, 2210–2220.

Vitart, F., C. Ardilouze, and A. Bonet, 2017: The Subseasonal to Seasonal (S2S) Prediction Project Database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173.

Wang, B., and Coauthors, 2008: Advance and prospectus of seasonal prediction: assessment of the APCC/CliPAS 14-model ensemble retrospective seasonal prediction (1980–2004). *Climate Dyn.*, **33 (1)**, 93–117.

Wang, W., M.-P. Hung, S. J. Weaver, A. Kumar, and X. Fu, 2013: MJO prediction in the NCEP Climate Forecast System version 2. *Climate Dyn.*, **42 (9-10)**, 2509–2520.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2008: Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quart. J. Roy. Meteor. Soc.*, **134 (630)**, 241–260.

Weisheimer, A., and F. D. Reyes, 2009: ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions—Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophys. Res. Lett.*, **36**, L21 711.

Wheeler, M. C., and H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, **132 (8)**, 1917–1932.

White, C. J., and Coauthors, 2017: Potential applications of subseasonal-to-seasonal (s2s) predictions. *Meteor. Appl.*, **24 (3)**, 315–325.

Xie, P., M. Chen, S. Yang, A. Yatagai, T. Hayasaka, Y. Fukushima, and C. Liu, 2007: A Gauge-Based Analysis of Daily Precipitation over East Asia. *J. Hydrometeor.*, **8 (3)**, 607–626.

Zhang, C., 2013: Madden–Julian oscillation: Bridging weather and climate. *Bull. Amer. Meteor. Soc.*, **94**, 1849–1870.

Zhou, X., Y. Zhu, D. Hou, and D. Kleist, 2016: A comparison of perturbations from an ensemble transform and an ensemble Kalman filter for the NCEP Global Ensemble Forecast System. *Wea. Forecasting*, **31**, 2057–2074.

Zhou, X., Y. Zhu, D. Hou, Y. Luo, and J. P. and, 2017: Performance of the new NCEP Global Ensemble Forecast System in a parallel experiment. *Wea. Forecasting*, **32**, 1989–2004.

Zhu, Y., X. Zhou, W. Li, and D. Hou, 2018: Towards the Improvement of Sub-Seasonal Prediction in the NCEP Global Ensemble Forecast System (GEFS). *J. Geophys. Res.*, **123**, 6732–6745.

LIST OF TABLES

TABLE 1. Summary of models participating in SubX, A=atmosphere, O=Ocean, I=sea ice, and L=land. Numbers in the ensemble members column apply to re-forecasts and real-time forecasts unless indicated by brackets [] which indicate a different number of ensemble members used in real-time forecasts than those used in the re-forecasts.

Model	Components	Ensemble Members	Length (Days)	Years	Reference(s)
NCEP-CFSv2	A,O,I,L	4	45	1999-2016	Saha et al. (2014)
EMC-GEFS	A,L	11 [21]	35	1999-2016	Zhou et al. (2016); Zhou et al. (2017); Zhu et al. (2018)
ECCC-GEM	A,L	4 [20]	32	1999-2014	Lin et al. (2016)
GMAO-GEOS	A,O,I,L	4	45	1999-2015	Koster et al. (2007); Molod et al. (2012); Reichle and Liu (2014); Rienecker et al. (2008)
NRL-NESM	A,O,I,L	4	45	1999-2016	Hogan et al. (2014); Metzger et al. (2014)
RSMAS-CCSM4	A,O,I,L	3 [9]	45	1999-2016	Infanti and Kirtman (2016)
ESRL-FIM	A,O,I,L	4	32	1999-2016	Sun et al. (2018a); Sun et al. (2018b)

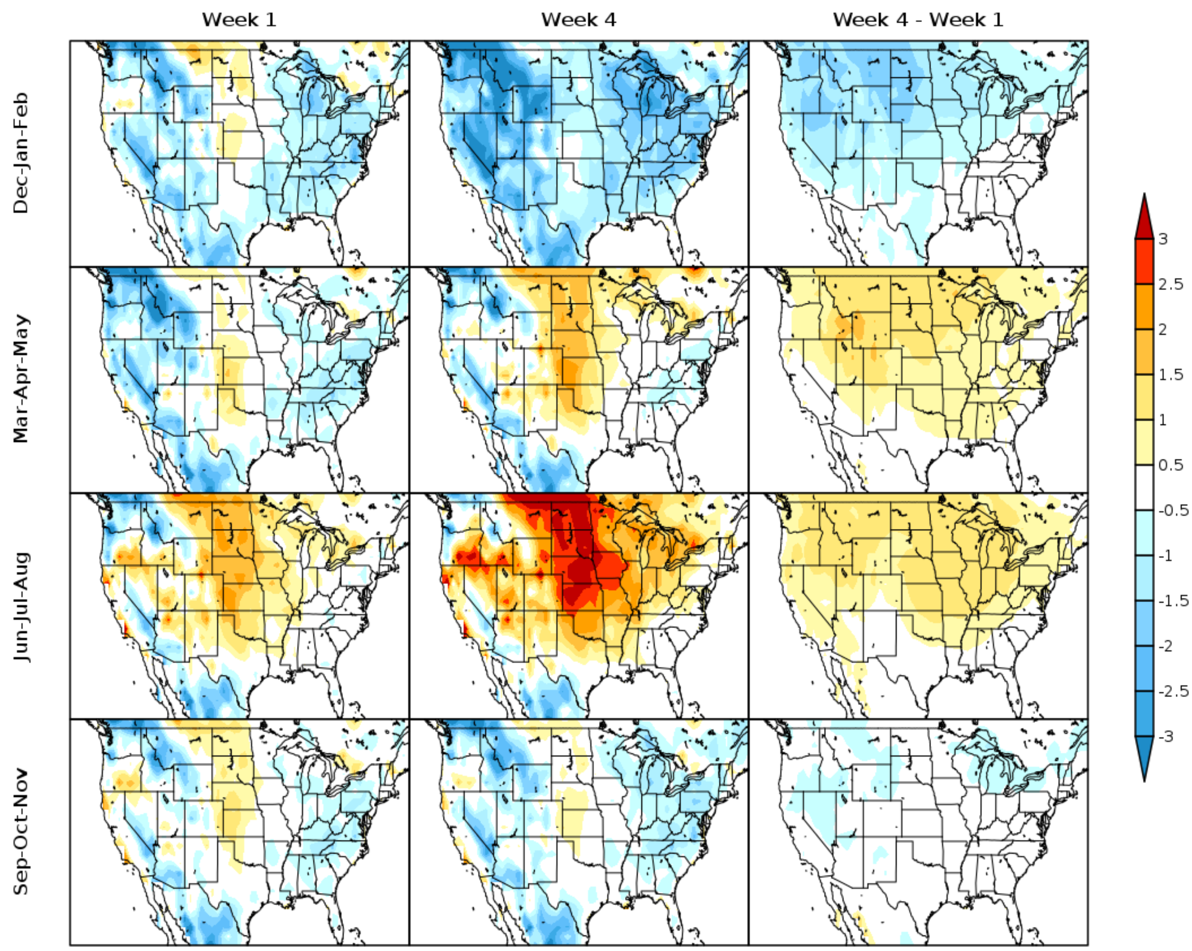
TABLE 2. Priority 1 variables: fields required to support Climate Prediction Center operational products

Variable	Level	Unit	Frequency
Geopotential Height	500 hPa	m	Average of instantaneous values at 0,6,12, 18 UTC
Geopotential Height	200 hPa	m	Average of instantaneous values at 0,6,12, 18 UTC
Zonal Velocity	850 hPa	ms^{-1}	Average of instantaneous values at 0,6,12, 18 UTC
Zonal Velocity	200 hPa	ms^{-1}	Average of instantaneous values at 0,6,12, 18 UTC
Meridional Velocity	850 hPa	ms^{-1}	Average of instantaneous values at 0,6,12, 18 UTC
Meridional Velocity	200 hPa	ms^{-1}	Average of instantaneous values at 0,6,12, 18 UTC
Temperature	2m	K	Daily Average (0-23:59 UTC)
Precipitation Flux	Surface	$kgm^{-2}s^{-1}$	Accumulated every 24 hours
Surface Temperature (SST+Land)	Surface	K	Daily Average
Outgoing Longwave Radiation	top of atmosphere	Wm^{-2}	Accumulated every 24 hours

TABLE 3. Priority 2 variables: fields needed to support evaluation of many S2S phenomena for research purposes

Variable	Level	Unit	Frequency
Specific Humidity	850 hPa	1	Daily Average (0-23:59 UTC)
Vertical Velocity	500 hPa	Pas^{-1}	Average of instantaneous values at 0,6,12, 18 UTC
Zonal Velocity	100 hPa	ms^{-1}	Average of instantaneous values at 0,6,12, 18 UTC
Meridional Velocity	100 hPa	ms^{-1}	Average of instantaneous values at 0,6,12, 18 UTC
Zonal Wind	10m	m^{-1}	Average of instantaneous values at 0,6,12, 18 UTC
Meridional Wind	10m	ms^{-1}	Average of instantaneous values at 0,6,12, 18 UTC
Daily Maximum Temperature	2m	K	24hr instantaneous
Daily Minimum Temperature	2m	K	24hr instantaneous
Latent Heat Flux	sfc	Wm^{-2}	Accumulated every 24 hours
Sensible Heat Flux	sfc	Wm^{-2}	Accumulated every 24 hours
Zonal wind stress	sfc	Nm^{-2}	Daily Average (0-23:59UTC)
Meridional wind stress	sfc	Nm^{-2}	Daily Average (0-23:59UTC)
Mean pressure	sea level	Pa	Average of instantaneous values at 0,6,12, 18 UTC
Snow water equivalent	N/A	kgm^{-2}	Accumulated every 24 hours
Net Radiation	sfc	Wm^{-2}	Accumulated every 24 hours
Snow Density	N/A	kgm^{-2}	Daily Average (0-23:59UTC)
Snow Cover	N/A	percent	Daily Average (0-23:59UTC)
Vertically integrated soil moisture	N/A	kgm^{-2}	Daily Average
Sea ice concentration	N/A	%	Daily Average (0-23:59UTC)
Convective Available Potential Energy	N/A	Jkg^{-1}	Daily Average

LIST OF FIGURES



785 FIG. 1. Multi-model biases for 2m temperature (°C) for week 1 (left), week 4 (middle), and week 4 minus week 1 (right) for re-forecasts initialized in Dec-Jan-Feb (top row), Mar-Apr-May (second row), Jun-Jul-Aug (third row), and Sep-Oct-Nov (bottom row). Biases are calculated as model minus verification.

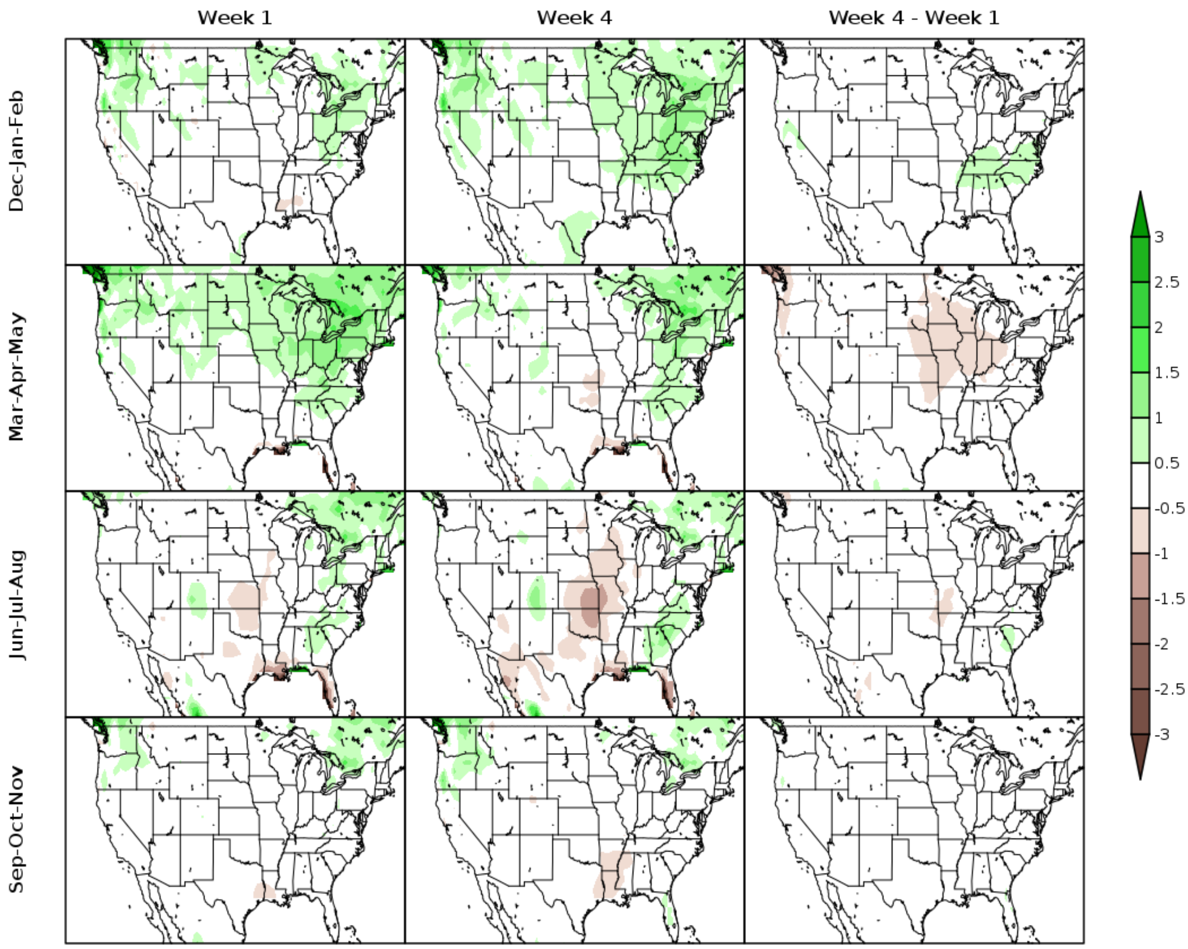


FIG. 2. Multi-model biases for precipitation (mm day-1) for week 1 (left), week 4 (middle), and week 4 minus week 1(right) for re-forecasts initialized in Dec-Jan-Feb (top row), Mar-Apr-May (second row), Jun-Jul-Aug (third row), and Sep-Oct-Nov (bottom row). Biases are calculated as model minus verification.

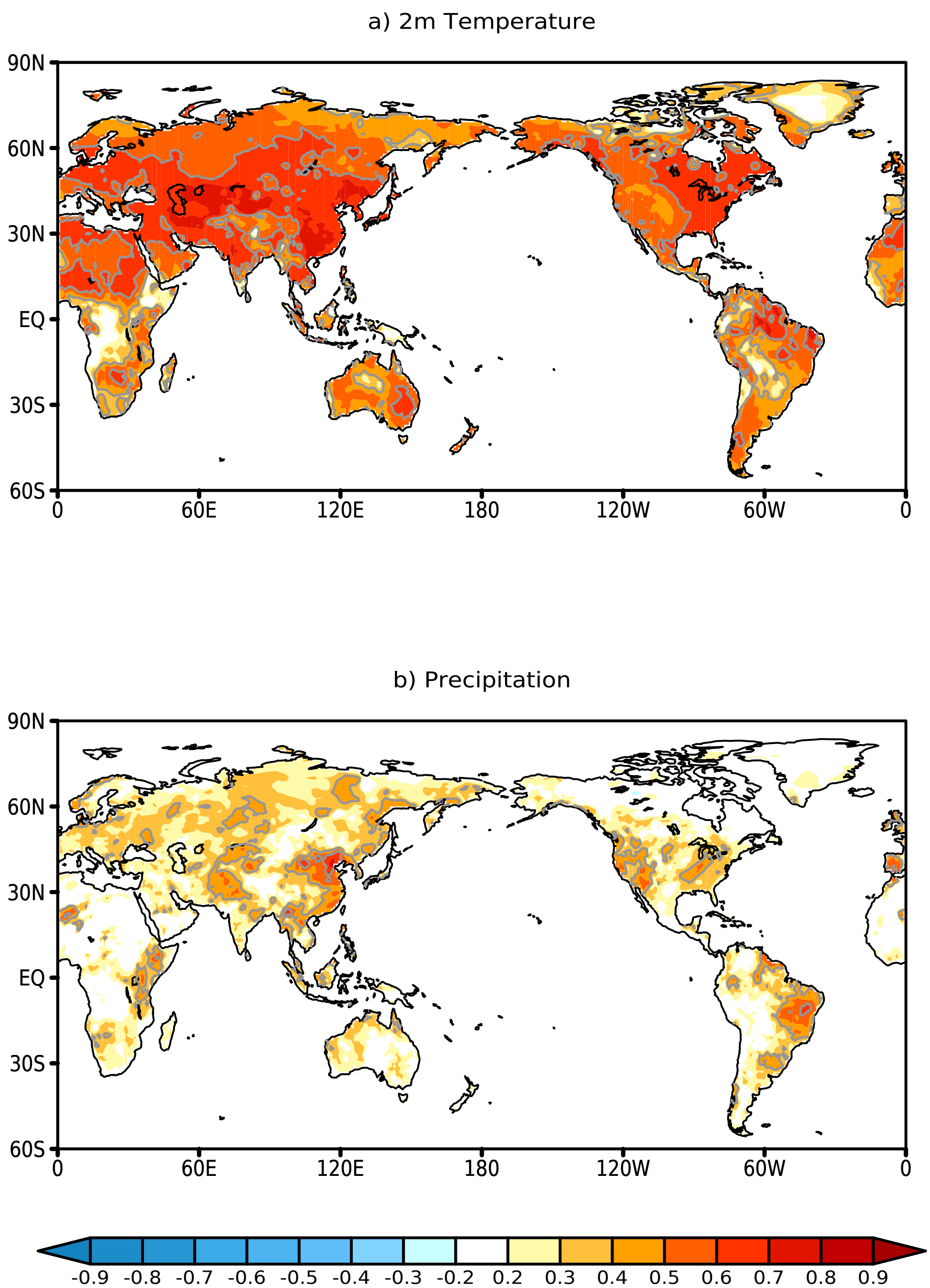


FIG. 3. Multi-model Ensemble ACC for week-2 (a) 2m temperature and (b) precipitation. ACC is calculated over re-forecasts with initial conditions for from Dec-Jan-Feb. Gray contour lines are drawn for ACC of 0.4 and 0.6.

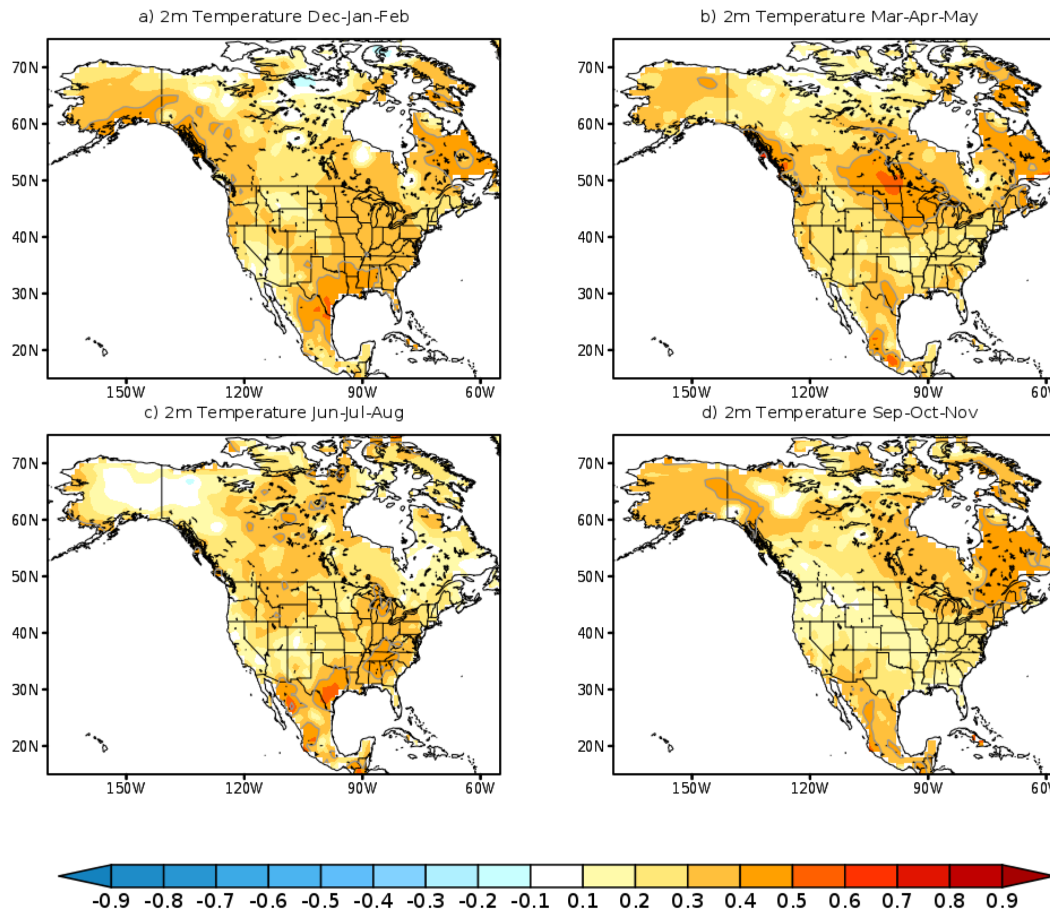


FIG. 4. Multi-model Ensemble ACC for week 3-4 2m temperature over North America. ACC is calculated over re-forecasts with initial conditions for (a) Dec-Jan-Feb, (b) Mar-Apr-May, (c) Jun-Jul-Aug, and (d) Sep-Oct-Nov.

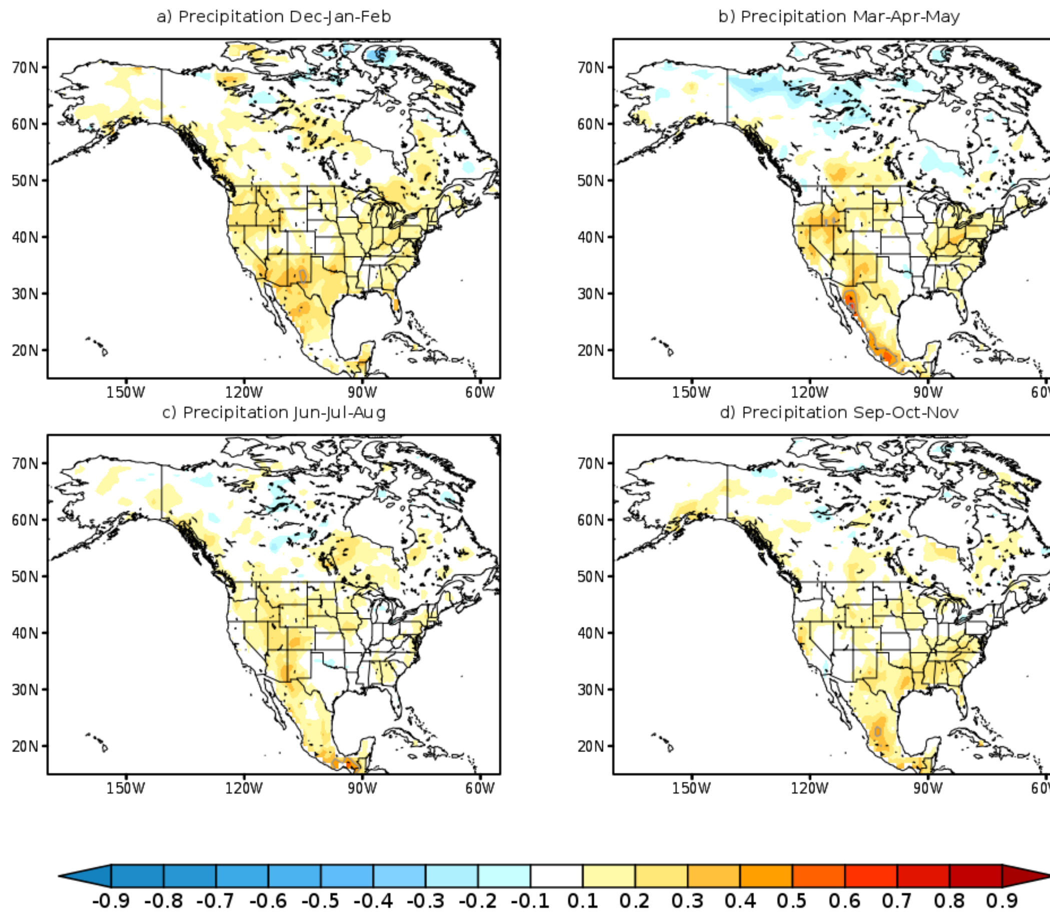


FIG. 5. Multi-model Ensemble ACC for week 3-4 precipitation over North America. ACC is calculated over re-forecasts with initial conditions for (a) Dec-Jan-Feb, (b) Mar-Apr-May, (c) Jun-Jul-Aug, and (d) Sep-Oct-Nov.

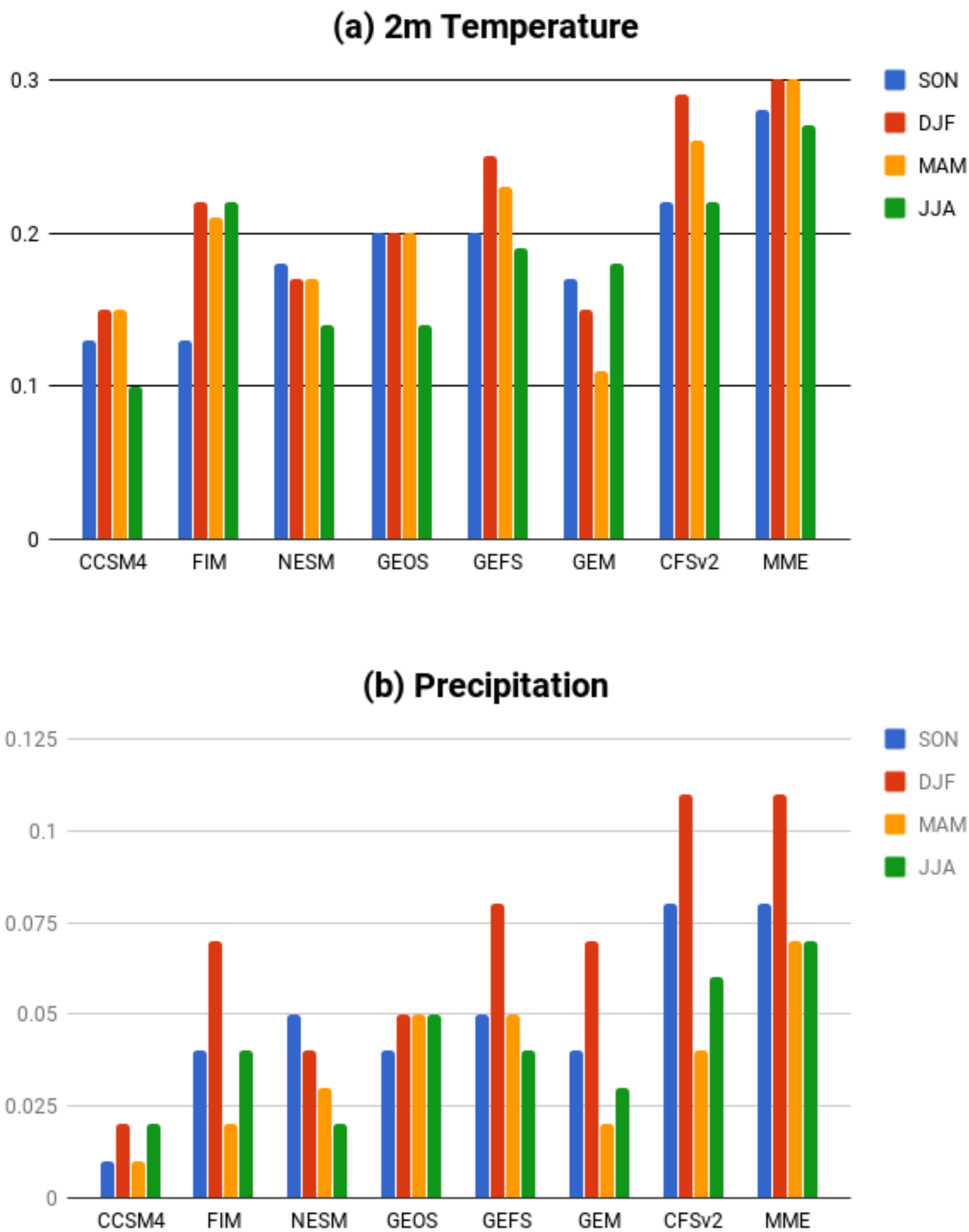


FIG. 6. Average week 3-4 ACC for (a) 2m temperature and (b) precipitation over North American domain shown in Figures 3 and 4 [15°N-75°N; 170°W-55°W]. ACC is calculated over re-forecasts with initializations

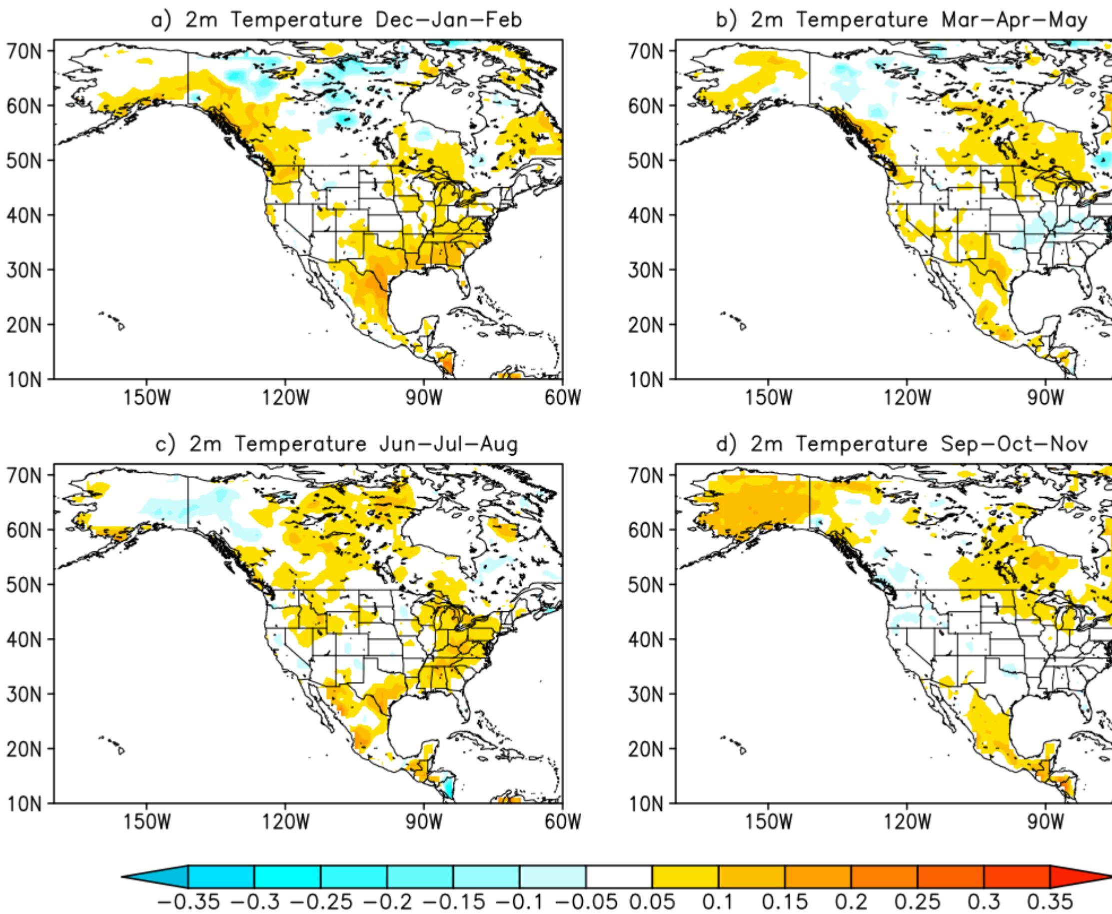


FIG. 7. Multi-model RPSS for week 3-4 2m temperature. RPSS is calculated over re-forecasts initialized in (a) Dec-Jan-Feb, (b) Mar-Apr-May, (c) Jun-Jul-Aug, and (d) Sep-Oct-Nov.

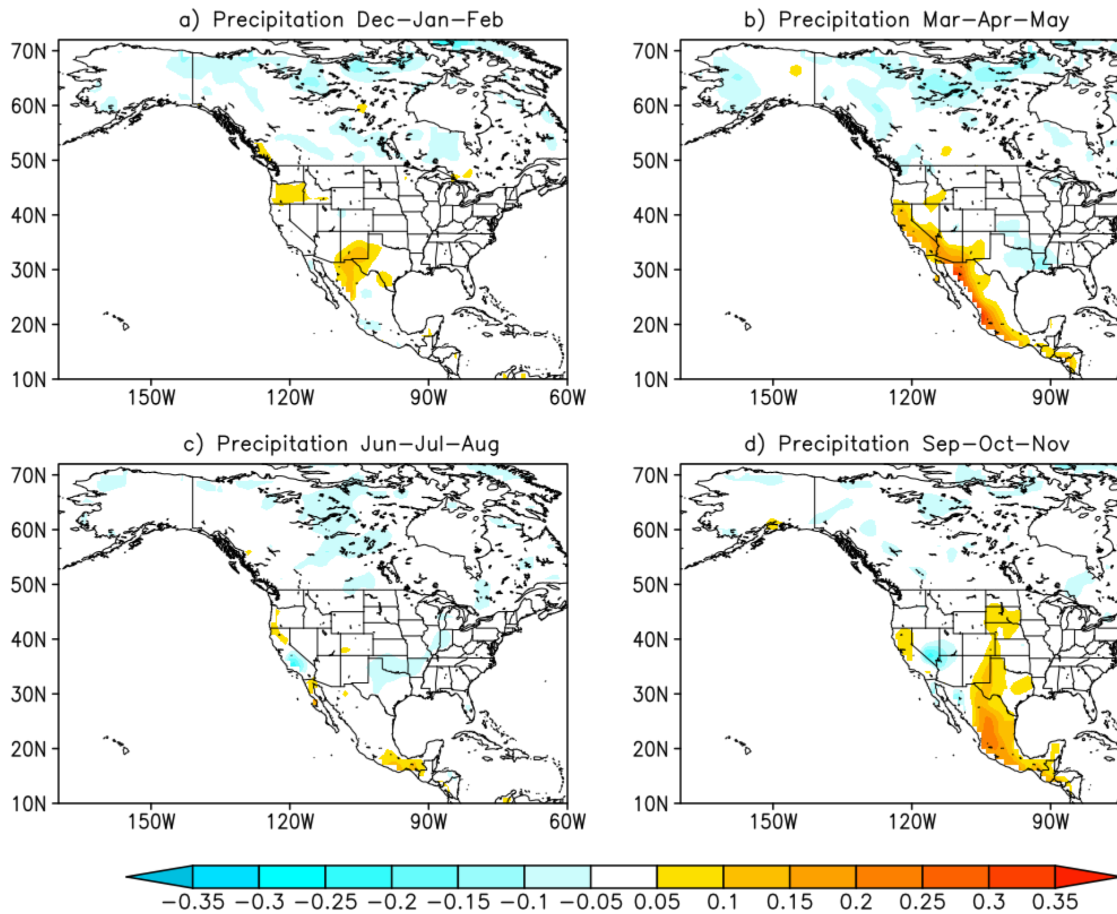


FIG. 8. Multi-model RPSS for week 3-4 precipitation. RPSS is calculated over re-forecasts initialized in (a) Dec-Jan-Feb, (b) Mar-Apr-May, (c) Jun-Jul-Aug, and (d) Sep-Oct-Nov initialized forecasts.

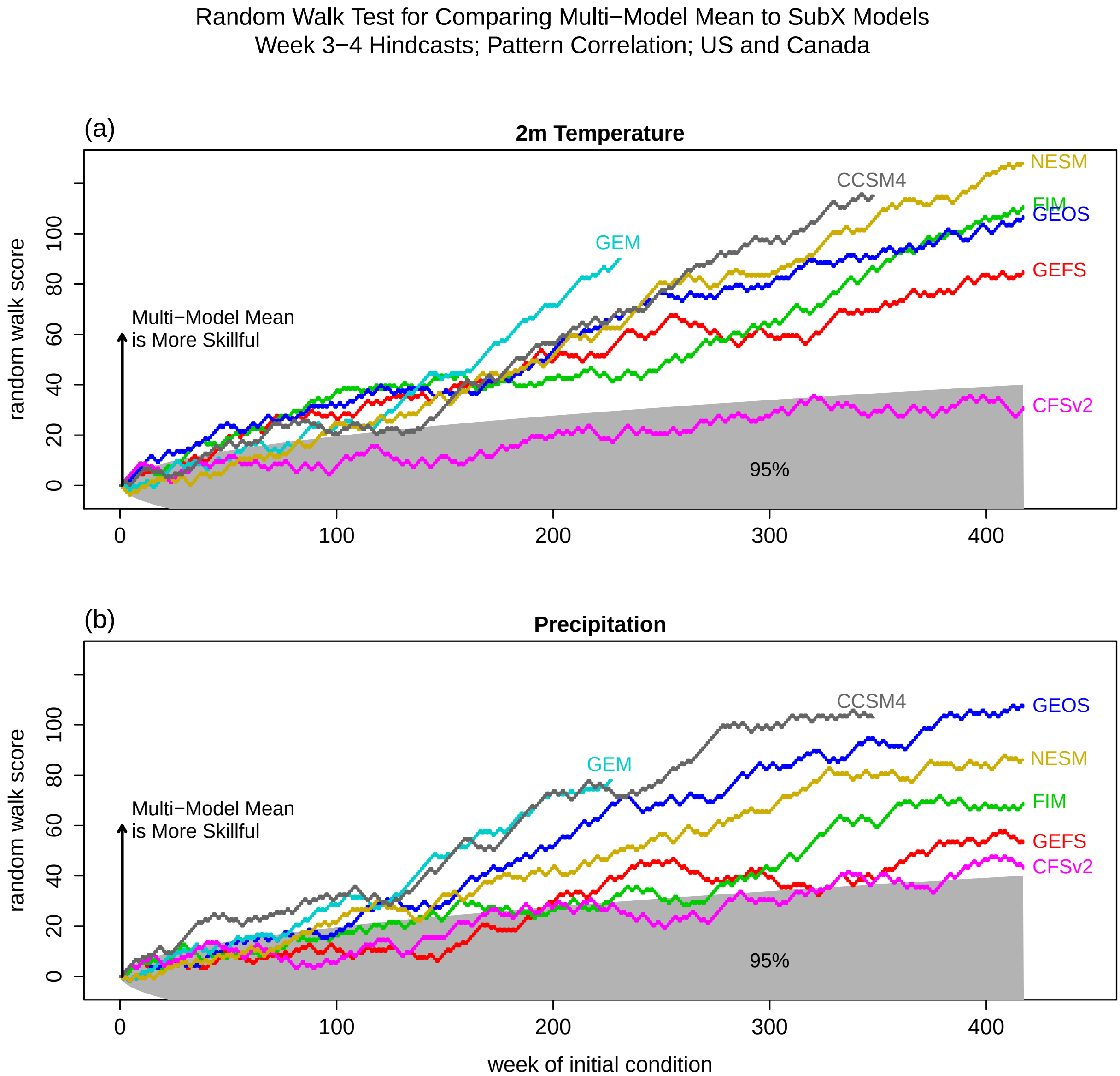


FIG. 9. Results of performing the random walk test (as described in the text) for comparing the multi-model mean to individual model hindcasts of week 3-4 temperature (a) and precipitation (b). The scores available for each model are strung together. Some models (e.g., GEM, CCSM4) do not have hindcasts for each verifying 2-week period because of the timing of their initial conditions. The x-axis refers to the week of the initial condition, but the corresponding date may differ across models because of verification gaps. The shaded region indicates the 95% probability range in which the random walk would lie if a given model were equally as skillful as the multi-model mean. In particular, a random walk that goes above the shaded region indicates that the multi-model mean has a higher pattern correlation more frequently (at the 5% level) than the model being compared.

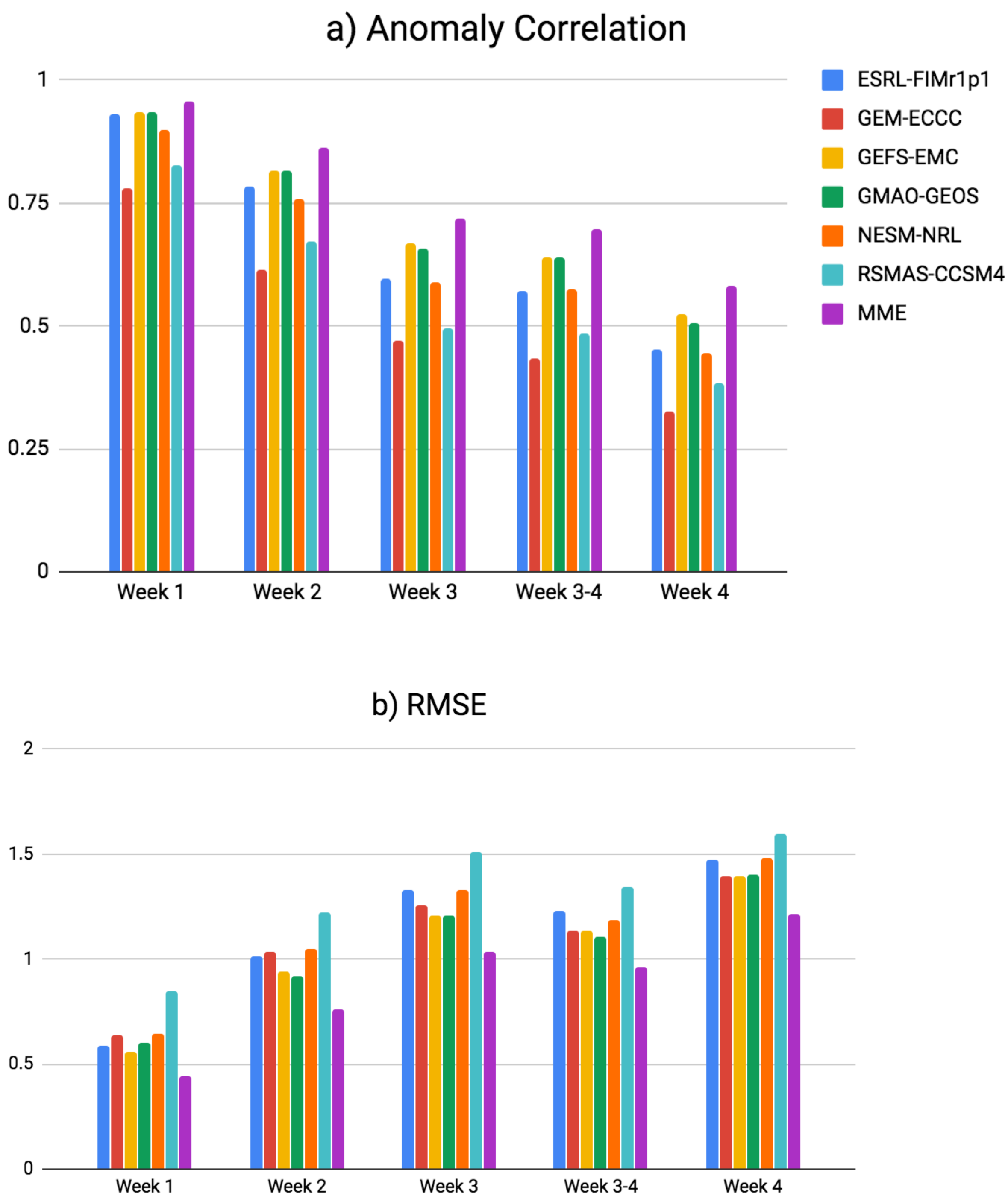


FIG. 10. RMM index skill in terms of ACC (a) and RMSE (b) for Nov-Mar initialized re-forecasts.

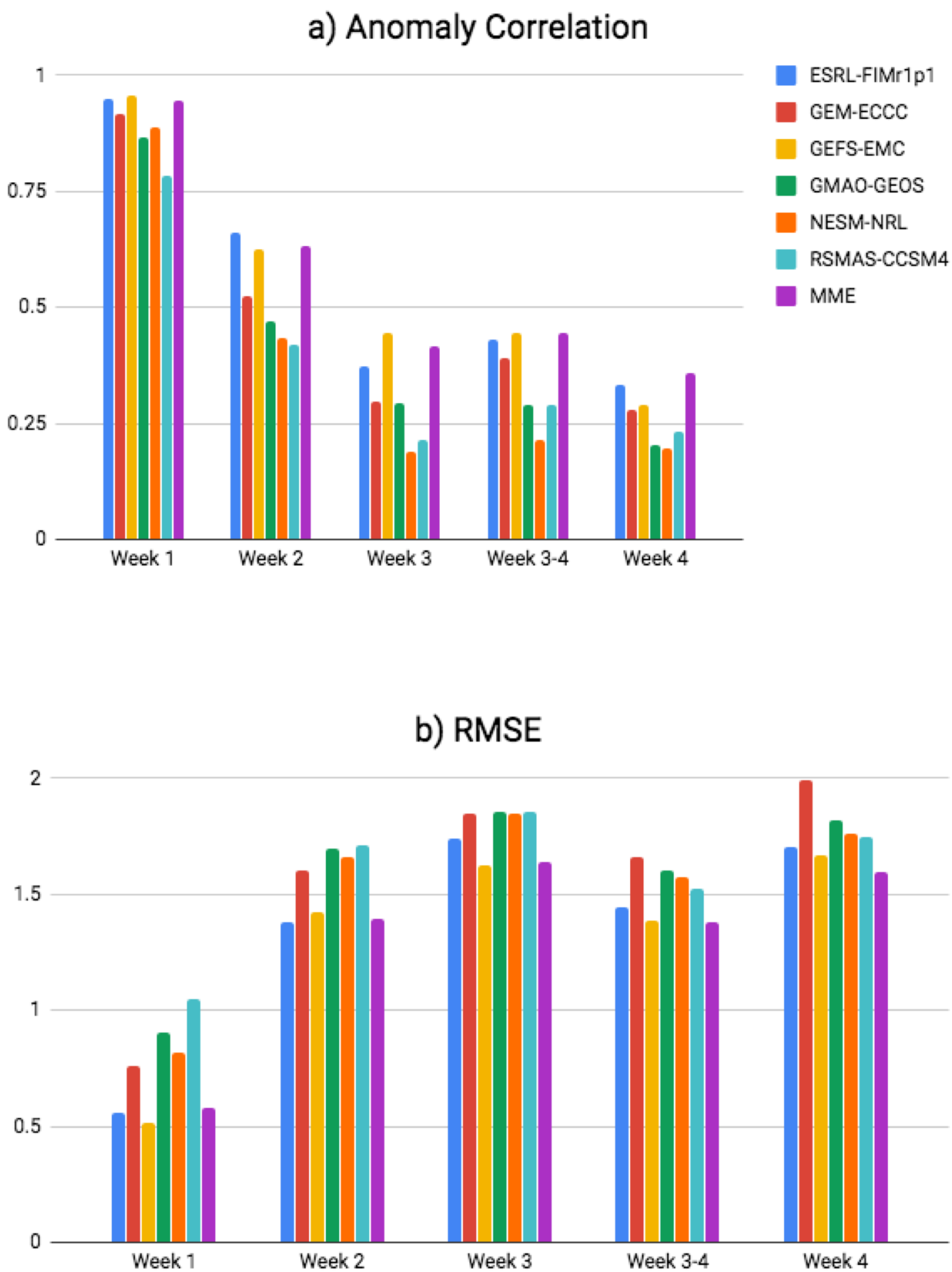


FIG. 11. NAO skill ACC (left) and RMSE (right) for Dec-Feb initialized re-forecasts.

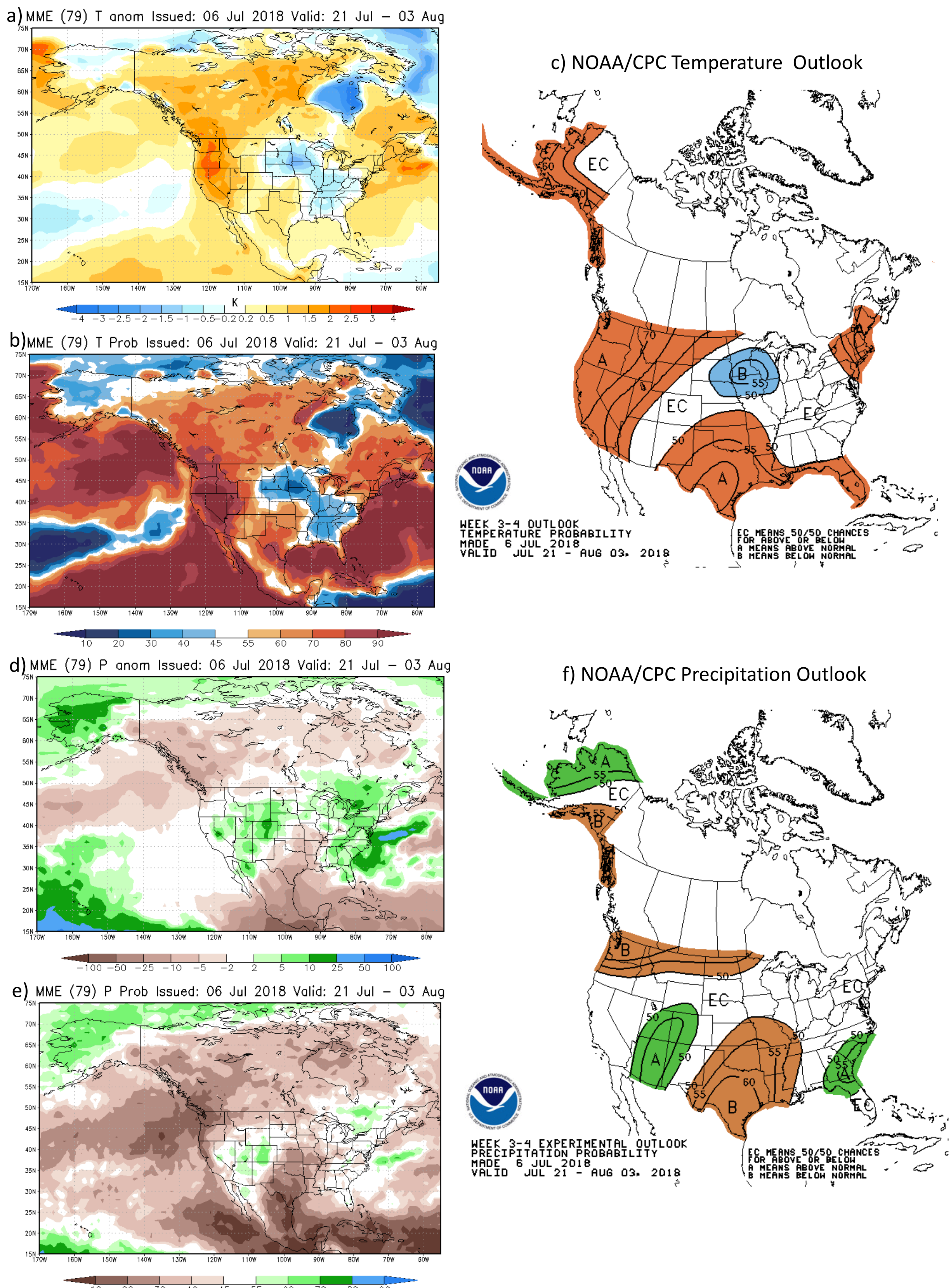


FIG. 12. SubX real-time multi-model anomaly and probability guidance for (a,b) temperature and (d,e) precipitation and corresponding CPC official week 3-4 outlook products for (c) temperature and (f) precipitation.

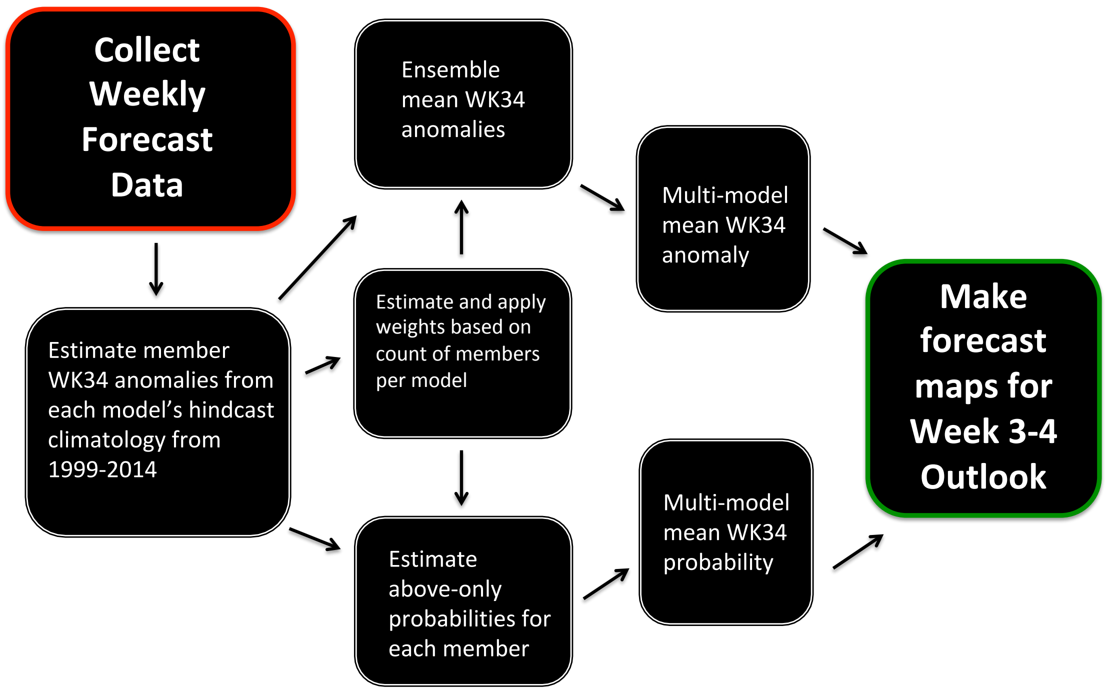


FIG. 13. Schematic diagram of the CPC procedure for processing SubX model data each week and producing anomaly and probabilistic maps for week 3-4 outlook guidance.