AGU100 ADVANCING EARTH AND SPACE SCIENCE

JGR Atmospheres

RESEARCH ARTICLE
10.1029/2019JD031252

Special Section:
Bridging Weather and Climate: Subseasonal-to-Seasonal (S2S) Prediction

Key Points:
- Current dynamical models have skillful prediction of extratropical cyclone activity on subseasonal to seasonal time scales
- Skillful predictions are found over parts of the Pacific, North America, northern Europe, and East Asia
- Seasonal prediction skills are mainly due to ENSO, while subseasonal skills arise from ENSO and the polar vortex

Supporting Information:
- Supporting Information S1
- Figure S1

Correspondence to:
C. Zheng,
cheng.zheng.1@stonybrook.edu

Citation:
Zheng, C., Chang, E. K.-M., Kim, H., Zhang, M., & Wang, W. (2019). Subseasonal to seasonal prediction of wintertime northern hemisphere extratropical cyclone activity by S2S and NMME models. *Journal of Geophysical Research: Atmospheres*, *124*, 12,057–12,077. https://doi.org/10.1029/2019JD031252

Received 27 JUN 2019
Accepted 11 OCT 2019
Accepted article online 23 OCT 2019
Published online 29 NOV 2019

Subseasonal to Seasonal Prediction of Wintertime Northern Hemisphere Extratropical Cyclone Activity by S2S and NMME Models

Cheng Zheng[1], Edmund Kar-Man Chang[1], Hyemi Kim[1], Minghua Zhang[1], and Wanqiu Wang[2]

[1]School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, New York, USA, [2]NOAA/Climate Prediction Center, College Park, Maryland

Abstract In this study, the prediction of wintertime extratropical cyclone activity from subseasonal to seasonal timescale in current dynamical models' reforecasts is investigated. On seasonal time scales, the North American Multi-Model Ensemble (NMME) models show skillful predictions over the eastern North Pacific, North America and the western North Atlantic with at least 5 months lead. The prediction skill is highly related to El Nino-Southern Oscillation (ENSO), as using the ENSO-related SST pattern gives rise to prediction skill with very similar spatial pattern and amplitude. On subseasonal time scales, models in the Seasonal to Sub-seasonal Prediction (S2S) dataset have skillful predictions up to 4 weeks lead over regions from the eastern North Pacific to the western Atlantic, as well as northern Europe, the eastern Atlantic and East Asia. Generally, forecast skill improves with a larger ensemble size. The subseasonal prediction skill from the Pacific to the western Atlantic is related to ENSO, and that over eastern Atlantic, Europe and East Asia are associated with stratospheric polar vortex anomalies. Current models do not show much skill from the Madden-Julian Oscillation (MJO), as the MJO impact on extratropical cyclone activity is not well captured by the models. European Centre for Medium-Range Weather Forecasts (ECMWF) model has the highest single model subseasonal prediction skill. The prediction skill in the ECMWF model is higher than its estimated potential predictability, likely because the signal-to-noise ratio is too low in the model hindcasts.

1. Introduction

Extratropical cyclones can often give rise to high-impact weather such as heavy precipitation, strong wind, storm surge and severe snowfall. Thus, they have large impact on regional weather and climate, as well as the economy and human life. Predicting extratropical cyclone activity (ECA) from days to seasons and projecting the future changes of ECA are of great scientific interest and can be beneficial to the society.

Current dynamical models have skillful predictions of individual extratropical cyclones with a few days of lead time. As it is difficult to predict midlatitude instantaneous weather out to a few weeks (e.g. Zhang et al., 2019), the aggregate paths of extratropical cyclones, also referred to as extratropical storm tracks, can be a useful substitute for one to predict ECA on timescales longer than a few weeks. The ECA (or storm track activity), can be represented not only by cyclone tracks (e.g. Klein, 1957), but also by statistics on gridded atmospheric data, such as the variance in a frequency band of synoptic timescales (e.g. Blackmon, 1976; Chang et al., 2012; Lau, 1978). In this study, we will focus on wintertime Northern Hemisphere ECA based on variance statistics.

The observational, theoretical, and modeling aspects of the ECA have been extensively studied in the literature (see the review papers by Chang et al., 2002; Shaw et al., 2016). ECA varies on multiple time scales (Chang et al., 2002; Chang et al., 2013; Stockdale et al., 2010; Chang & Fu, 2002). On interannual time scales, the El Niño–Southern Oscillation (ENSO) has significant modulation on Northern Hemisphere ECA (Eichler & Higgins, 2006; Ma & Chang, 2017; Straus & Shukla, 1997; Zhang & Held, 1999). El Niño events drive an equatorward and downstream shift of the Pacific storm track and weakening of North America ECA, while La Niña events drive opposite changes. Recent studies showed that the quasi-biennial oscillation (QBO) can also modulate Northern Hemisphere ECA (Wang, Kim, & Chang, 2018) over both the Pacific and the Atlantic, but the impacts are mostly in the upper troposphere. The QBO can also modulate the MJO impact on Pacific ECA (Wang, Kim, Chang, & Son, 2018; see discussion below for MJO impact on ECA).

Predictability of seasonal ECA has been studied by Yang et al. (2015) using a Geophysical Fluid Dynamics Laboratory (GFDL) climate model. The leading predictable ECA pattern in Northern Hemisphere winter was found to be ENSO-related and is predictable up to 9 months in advance.

On subseasonal time scale, recent studies have shown that several phenomena, including ENSO, the polar vortex, the Madden Julian Oscillation (MJO), and the QBO, can provide sources for atmospheric predictability (e.g. Black et al., 2017; DelSole et al., 2017; Garfinkel et al., 2018; Tian et al., 2017; Xiang et al., 2019). The MJO causes a north-south shift of ECA over North America when the MJO associated tropical diabatic heating is in different locations (Zheng et al., 2018). The MJO also modifies ECA over North Pacific and North Atlantic (Deng & Jiang, 2011; Guo et al., 2017; Lee & Lim, 2012; Zheng et al., 2018). In addition, the polar vortex in the Northern Hemisphere stratosphere can also modulate the ECA (Kidston et al., 2015; Scaife et al., 2012). Stratospheric wind anomalies can modify the midlatitude jet through the "downward control" mechanism (Haynes et al., 1991). Stronger or weaker zonal flow generally leads to enhanced or suppressed ECA. Walter and Graf (2005) showed that stratosphere anomalous patterns modify ECA over the North Atlantic.

In this study, we will explore the prediction of ECA on seasonal time scale (up to 5 months) by using the North American Multi-Model Ensemble (NMME; Kirtman et al., 2014), and on subseasonal time scale (i.e., 1-4 weeks) by using the Seasonal to Sub-seasonal Prediction (S2S; Vitart et al., 2017) dataset. In section 2, we will introduce the datasets, as well as how ECA is defined and evaluated in the models. Seasonal and monthly prediction of ECA will be discussed by using NMME and statistical model in section 3. Subseasonal prediction skill of ECA in the S2S dataset will be shown in section 4. Discussions related to the source of subseasonal prediction skill and potential predictability will be provided in section 5. Conclusions and implications of this study will be discussed in section 6.

2. Data and Methods

2.1. Data

2.1.1. NMME and S2S models

Phase 2 of the NMME (Kirtman et al., 2014) comprises of seasonal forecasts and hindcasts made by 8 models, providing daily data for a number of atmospheric variables. Seven of the eight models (except for NCEP CFSv2) provide forecasts initialized at 00Z in the first day of each month. Because the GFDL models do not provide MSLP data (the variable we use to quantify storm track activity – see Section 2.2.1 below), while the NCAR CESM does not initialize the atmosphere with a current atmospheric analysis, in this study we examine hindcasts from four models – Environment Canada CanCM3 and CanCM4, NASA Goddard Space Flight Center (GSFC) GEOS5, and NCAR/University of Miami CCSM4.0. Each of these models provide a ten-member ensemble, with a common hindcast period of 1982 to 2012. Here, we focus on Northern Hemisphere (NH) winter (December to February; DJF) since storm tracks are most active in this season. Because there are some missing data for the 1998/99 winter season, this season is omitted from our analysis. Thus, we have examined hindcasts from 29 winter seasons or 87 winter months. Apart from daily MSLP data, we have also examined monthly mean SST hindcasts. Since NMME data are scheduled to be delivered by the modeling centers to the Climate Prediction Center (CPC) by the 8th of each month, the lead-zero forecasts (i.e. DJF or December forecasts made on December 1st) will only be available more than one week after the predicted period has begun. Thus lead 0 hindcast skills are examined mainly for reference and for the sake of completeness.

The S2S database consists of 11 coupled or uncoupled models including near-real-time ensemble forecasts and reforecasts up to 60 days. Since the S2S database is not produced by following an agreed upon protocol, there are significant differences among S2S models (e.g. resolution, forecast time range, frequency of initializing forecasts, ensemble sizes). Reforecasts of 6 models are used (see Table 1 and Text S1). MSLP data on a 1.5° by 1.5° horizontal resolution grid are available at 00 UTC at each forecast day from these 6 models. To investigate model performance in the upper level, 500-hPa geopotential height (Z500) data are also examined.

2.1.2. Reanalysis and other datasets

For verification, ECA is calculated by using the 6-hourly MSLP data on a 1.5° by 1.5° horizontal resolution grid from ECMWF Interim Re-Analysis (ERA-Interim, Dee et al., 2011). As only daily mean MSLP is

Table 1
Description of the 6 models in S2S dataset that are used in this study.

Model	Time range	Resolution	Reforecast frequency	Reforecast period	Reforecast Sizes	Ocean coupling	Sea ice coupling
CMA China Meteorological Administration	Day 0-60	T106 L40	daily	1994-2014	4	Yes	Yes
CNR-ISAC Institute of Atmospheric Sciences and Climate of the National Research Council (Model Version Date 2017/06/08)	Day 0-32	0.75x0.56 L54	every 5 days	1981-2010	5	No	No
CNRM Météo-France/Centre National de Recherche Meteorologiques	Day 0-61	T255 L91	4 times a month	1993-2014	15	Yes	Yes
ECCC Environment and Climate Change Canada Model Version: GEM Jan-2016	Day 0-32	0.45x0.45 L40	weekly	1995-2014	4	No	No
ECMWF European Centre for Medium-Range Weather Forecasts Model Version: CY43R3	Day L91	twice a week	1997-2016	11	Yes	Yes	
HMCR Hydrometeorological Centre of Russia	Day 0-61	1.1x1.4 L28	weekly	1985-2010	10	No	No

available from NMME models, daily mean MSLP of ERA-Interim is obtained by averaging 6-hourly MSLP data. 6-hourly geopotential height data on the same temporal and spatial resolution is also obtained from ERA-Interim. To be consistent with S2S dataset, only the data at 00 UTC in ERA-Interim is used to verify S2S models.

In addition, to evaluate whether the model prediction skill is associated with phenomena like ENSO, MJO or polar vortex, datasets which quantify these phenomena are also used in the study. We use the Nino 3.4 index to define the phase of ENSO. The Nino 3.4 index, which is obtained from National Oceanic and Atmospheric Administration (NOAA) Earth System Research Laboratory (ESRL) website, is calculated based Hadley Centre's sea ice and sea surface temperature (SST) data set (HadISST1; Rayner et al., 2003). For the MJO, we make use of the real-time multivariate (RMM, Wheeler & Hendon, 2004) index, which is the commonly used index for evaluating MJO forecasts (e.g. Kim et al., 2018). The RMM index, obtained from Bureau of Meteorology (BoM) website, is developed based on multivariate empirical orthogonal function (EOF) analysis of combined fields of outgoing longwave radiation (OLR), 850 and 200-hPa zonal wind anomalies. The Polar vortex index (PVI; similar to Xiang et al., 2019) is defined by averaging the zonal wind anomaly north of 60°N at 100-hPa in ERA-Interim.

2.2. Methods

2.2.1. Definition of ECA

To quantify ECA, a 24-h difference filter introduced by Wallace et al. (1988) is applied on the MSLP data,

$$pp = \overline{\{MSLP\,(\,t + 24hr\,) - MSLP\,(\,t\,)\}^2}, \tag{1}$$

pp, which is mean square of the 24-h difference of MSLP, is used to define ECA. In equation (1), the 24-h difference is calculated at each time step and on each grid point. Daily mean MSLP is used in the equation for NMME models, while MSLP at 00 UTC is used for S2S models. In many previous studies, this averaging is performed over a continuous time period. Here we will also examine seasonal, monthly, and weekly means. However, when verifying whether the prediction skill of ECA is associated with the MJO, we apply a similar approach as Guo et al. (2017) and Zheng et al. (2018), in which the averaging time period is not continuous. As shown by many previous studies (e.g., Chang et al., 2002; Wallace et al., 1988), the maxima from this 24-h difference filter lie over geographical locations where extratropical cyclones preferentially cross (see also Figure 1a). Variations in pp will serve as an indicator of variability of ECA. The winter climatology (DJF), and variability (standard deviation) on weekly, biweekly (week 3-4 prediction skill is evaluated in this study) and monthly time scale of ECA, are shown in Figure 1. The maxima in this metric clearly highlight the

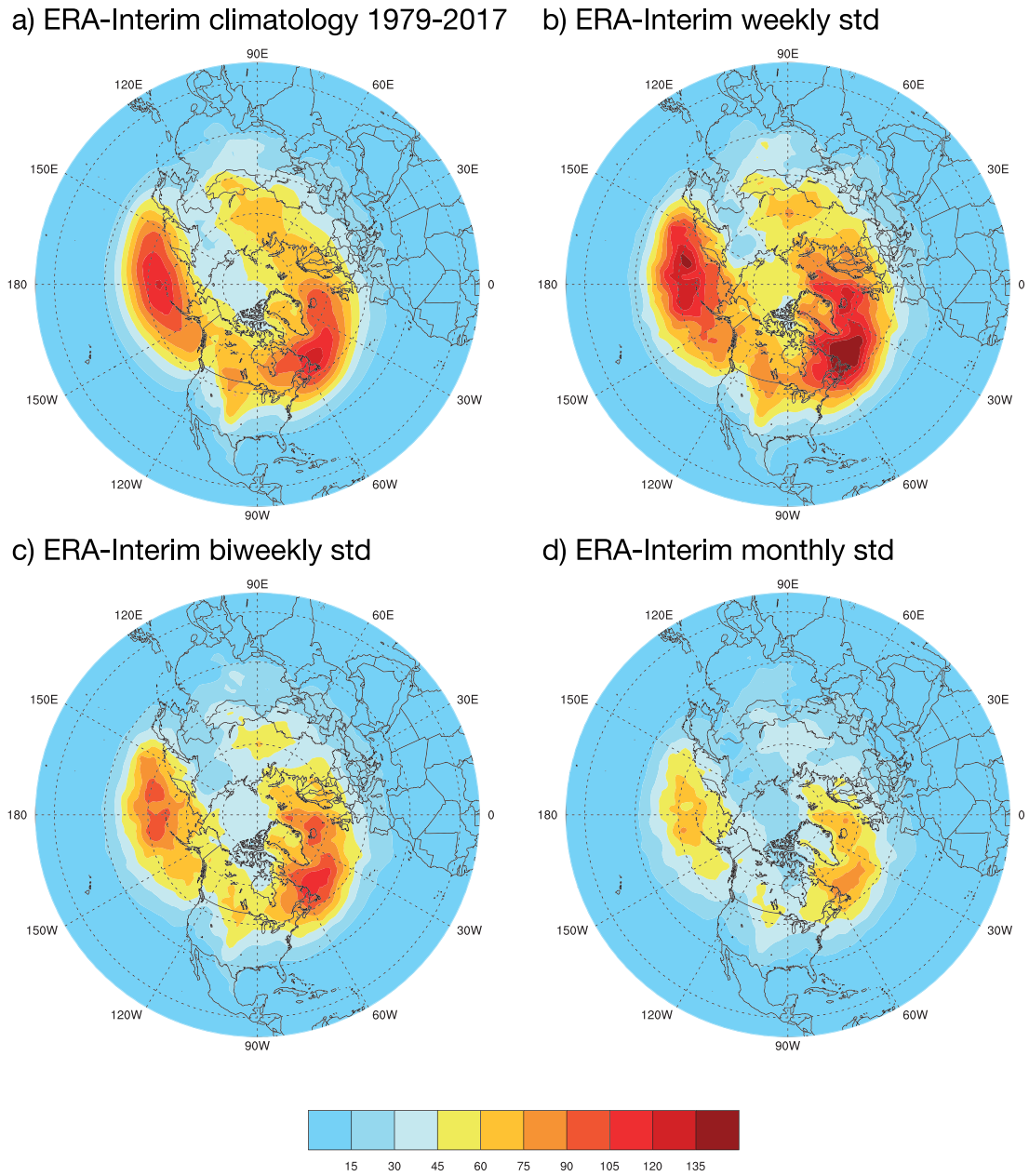

Figure 1. a) Climatology of the Northern Hemisphere extratropical cyclone activity (ECA) for 1979-2017 winters (December-February) based on all ERA-interim reanalysis 00z mean sea level pressure (MSLP) data. b) Standard deviation of 7-day running mean ECA. c) Standard deviation of 14-day running mean ECA. d) Standard deviation of monthly ECA. Units in hPa^2.

Pacific storm track spreading eastward across North America into the Atlantic and finally northeastward into northern Europe and Siberia (Figure 1a), which follows the direction of the mean flow. These are also the regions where storm track activity is most variable at all time scales (Figure 1b-d). Note that previous studies have shown that monthly and seasonal variability in this metric is well correlated with variability in precipitation and weather extremes (e.g. Chang et al., 2015; Yang et al., 2015; Ma & Chang, 2017).

2.2.2. Climatology and Anomalies of ECA in Model Forecasts

For sub-seasonal to seasonal range forecasts, the model bias can become dominant. For each individual S2S model and NMME model, during the reforecast period, the reforecasts are initialized at the same dates in the year. Thus, a model climatology, which depends on the model initialization time as well as forecast day (e.g. forecast day 1, forecast day 2 ...) is needed to correct the model bias. For an individual S2S model, all the

reforecast of ECA at each grid point can be written as, $pp_{model}(y, d_y, n, f_d)$, where y is year, d_y is initialization day, f_d is the forecast lead day. $n = 1,...,N$, where N is the total number of ensemble members in a single model. A proper model climatology, which depends on both reforecast initialization time and forecast day, can be obtained by averaging over all the years and all the ensemble members,

$$pp_{cli}(d_y, f_d) = \frac{\sum_y \sum_n pp_{model}(y, \ d_y, n, f_d)}{N \times Y}, \tag{2}$$

where Y is the total number of years. Then, model anomaly can be defined as the deviation from model climatology,

$$pp_{ano}(y, d_y, \mathrm{n}, f_d) = pp_{model}(y, d_y, n, f_d) - pp_{cli}(d_y, f_d). \tag{3}$$

The climatology and anomalies of ECA in NMME and Reanalysis can be defined by using a similar way, except that the Reanalysis climatology does not depend on forecast day and there is only one ensemble member when defining Reanalysis climatology.

2.2.3. Combining Different Models into a Multi-model Ensemble

Previous studies have shown that multi-model ensembles (MME) usually outperform single model ensembles (e.g. Becker et al., 2014; Hagedorn et al., 2005; Smith et al., 2013). It is rather straightforward to combine the NMME models into an MME as they are all initialized at the same time (first day of each month). However, as the S2S models reforecast are initialized on different dates, the S2S models cannot be directly combined into an MME. We have developed a way to combine the models with the following steps: (1) During the winter seasons in the overlapping period of the 6 S2S models (DJF from 1997/98 to 2009/10), for each day and each model, the gap between this day and the nearest reforecast that is initialized earlier than this day is defined as the lead of the reforecast (Figure S12) (2) For any day in this period, we select days that the lead of all the 6 models is less than or equal to 4 days (yellow shadings in Figure S12b). (3) If consecutive days satisfy the requirement, only the earliest day is selected (to make the lead smallest) and defined as Day0 (red boxes in Figure S12b). Then, the 6 models are combined into an MME for all the days which can be defined as Day0. Day0 is considered as the "forecast day 0" for the MME. Day 0-6, 7-13, 14-20 and 21-27 after Day0 are considered as week1, week2, week3 and week4 respectively in the following discussions. During DJF from 1997/98 to 2009/10, there are 156 days that can be considered as Day0 (about 12 cases per winter season, which on average is about once a week).

2.2.4. Prediction Skill, Potential Predictability and Model Evaluation

The prediction skill and potential predictability of NMME and S2S models are assessed primarily by using the anomaly correlation coefficient (ACC). The ACC measures the association between the anomalies of grid point forecast and analysis. Prediction skill, measures how well the ensemble mean (EM, here it can be either single model ensemble or MME) forecasts the observed (Reanalysis) value. When calculating multi-model ensemble mean, ensemble members from each model are weighted equally.

Potential predictability assesses how well one model's EM, which is based on $N-1$ members, predict the one member that is left out (e.g. Becker et al., 2014; Kumar et al., 2014). We first use each ensemble member as the "left out" member to calculate the EM ACC, then the average of the ACC of all members is considered as potential predictability. The anomaly correlation coefficient at each grid point and each forecast lead day can be written as,

$$ACC = \frac{\sum_y \sum_{d_y} pp_{ano}^{EM}(y, d_y) \ pp_{ano}^{obs}(y, d_y)}{\sqrt{\sum_y \sum_{d_y} \left[pp_{ano}^{EM}(y, d_y)\right]^2 \sum_y \sum_{d_y} \left[pp_{ano}^{obs}(y, d_y)\right]^2}} \tag{4}$$

where $pp_{ano}^{EM}(y, d_y)$ is the ensemble mean of model forecast anomalies, $pp_{ano}^{obs}(y, d_y)$ is the anomalies in Reanalysis. Assuming that the model simulates the correct physics of the real atmosphere, the potential predictability is an estimate of the upper bound of the prediction skill. As daily ECA is noisy, weekly (week 1, week 2, week 3, week 4) or biweekly (e.g. week 3-4) predictions are evaluated for the S2S models, while monthly and seasonal predictions are evaluated for NMME models.

To further evaluate S2S model predictions, we also apply the Heidke skill score (HSS), which is a common performance metric used by the CPC to evaluate extended-range probabilistic forecasts (Wilks, 2011). The HSS is a useful tool to assess the proportion of categories that is forecast correctly. For EM of single model or multi-model ensembles, each forecast is assigned to one of two (above normal or below normal) or three (top, middle, or bottom tercile) categories. The number of categories forecast correctly is designated as H. The expected number of categories forecast correctly from a random forecast, E, is one-third of the total number of forecasts, T, for three-category HSS, and is half of T for two-category HSS. Then HSS can be written as,

$$HSS = \frac{H-E}{T-E} \tag{5}$$

The HSS ranges in value from -1 (completely wrong set of forecasts) to 1 (perfect forecasts) for two-category HSS, and from -0.5 (completely wrong set of forecasts) to 1 (perfect forecasts) for three-category HSS. If HSS equals zero, it can be considered as the expected HSS of a climatological forecast if we define the climatological forecast as a random draw from two (or three) equiprobable forecast categories. Thus, if HSS values are above zero, it indicates that the forecasts are skillful.

3. Monthly and Seasonal Predictions of ECA in NMME Models

3.1. Seasonal Prediction Skill

As discussed above, we examined hindcasts made with 4 NMME models, each with 10 ensemble members. Overall, we do not find much difference in the prediction skill between individual model ensembles, with the multi-model ensemble mean (the average of all 40 members) exhibiting the highest anomaly correlation with observed variability. Hence, we will focus on the multi-model ensemble mean.

For lead zero (i.e., the hindcast for DJF pp using December 1 00Z initial conditions), the NMME hindcasts exhibit an extended band of significant ACC in the subtropics and mid-latitudes spreading from East Asia across the Pacific, North America, and the Atlantic (Figure 2a). The highest skill appears to be over Eastern Pacific and North America, with ACC reaching above 0.6 just off the west coast of North America, and over 0.5 over much of the continental US, western Canada, and Alaska. For one-month lead (Figure 2b), the main area with high ACC lies over the central part of North America and Alaska, with scattered regions of significant correlation over the Pacific and Atlantic. This pattern persists with leads of two months and longer (not shown).

These hindcast skills for leads of one month or more is likely related to the impacts of ENSO. In Figure 2c, the ACC between seasonal mean Nino-3.4 index and pp from ERA-Interim data is shown, showing strong negative correlation over continental U.S. and Canada, and moderately positive correlation over eastern Pacific and western Atlantic. Correlation between model predicted pp (zero-lag) and observed Nino-3.4 index is shown in Figure S17a, showing that the models can largely predict the ENSO related ECA variability. In fact, one can construct a simple statistical model for predicting pp using just the leading empirical orthogonal function (EOF) of equatorial Pacific SST (which is related to ENSO) as follows: The principal component (PC) of the leading EOF of SST is regressed with seasonal pp anomaly at each grid point, while leaving out data from a target season (leave-one-out-cross-validation). This model can then be used to hindcast the left out season by projecting the model predicted seasonal SST anomaly onto the leading EOF pattern. This projection coefficient is then multiplied by the regression coefficients to form the pp prediction. The skill of this simple statistical model is shown in Figure 2d, showing that using the leading EOF of predicted tropical SST alone can explain most of the NMME predicted skill for lead of one month. This is consistent with Yang et al. (2015) who showed that the leading predictable pattern of seasonal storm track variability is that associated with ENSO.

3.2. Monthly Prediction Skill

For monthly hindcasts, the multi-model ensemble mean displays significant ACC with observed pp variability over nearly everywhere in the mid latitudes (Figure 3a), with highest correlation over northeastern Asia and around the Greenwich Meridion. Over continental US and Canada, the ACC is mostly 0.4 or higher. For one-month lead (Figure 3b), the highest correlation is found over North America and its adjacent oceans,

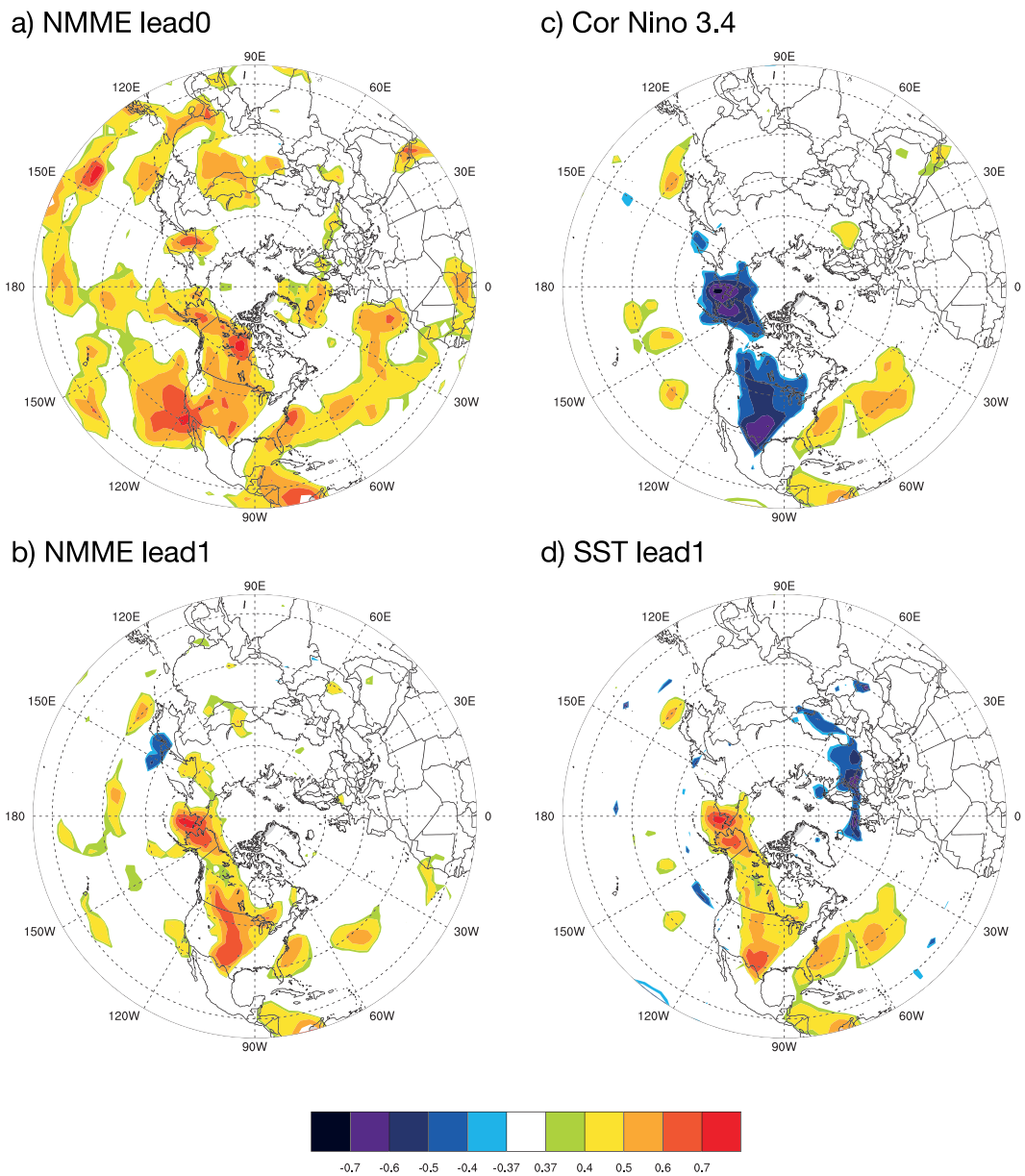

Figure 2. a) Prediction skill of seasonal (DJF) extratropical cyclone activity of NMME ensemble mean at lead zero. b) The same as a) but for one month lead hindcast. c) Anomaly correlation between seasonal mean Nino 3.4 index and seasonal mean extratropical cyclone activity in Reanalysis. d) Prediction skill of extratropical cyclone activity from a simple statistical model by using the leading EOF of equatorial SST (see main text for details). For 29 seasons, a correlation of 0.37 is significant at the 95% level.

similar to lead-1 seasonal hindcasts. Month-to-month correlation between pp and Nino-3.4 is shown in Figure 3c (see also Figure S17b for correlation between model predicted pp and Nino-3.4), while the hindcasts made by a simple leave-one-season-out cross validation regression model based on the leading PC of model hindcast monthly equatorial Pacific SST is shown in Figure 3d. It is clear that the one-month lead NMME hindcasts are only marginally better than that given by this simple statistical model. Similar anomaly correlation patterns but with slightly lower correlations for both NMME and SST regression are found for longer leads (not shown), out to a lead time of 5 months (the longest-lead hindcast cycle examined in this study). Hence, Figure 2 and 3 together suggest that at least for the NMME models, the main source of predictability of storm track variability with leads of one month or longer is related to ENSO.

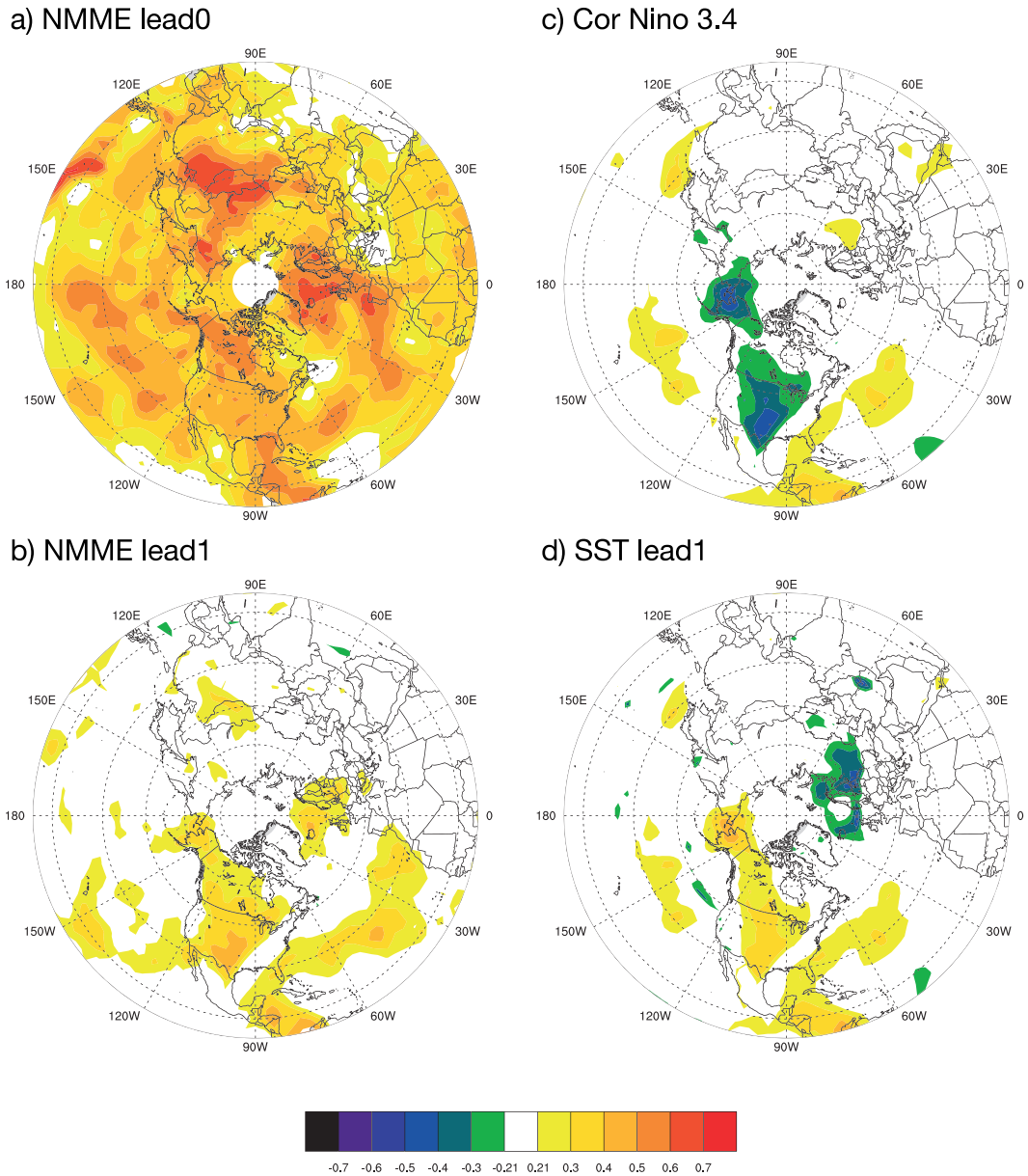

Figure 3. The same as Figure 2, but for monthly forecast. For 87 months, a correlation of 0.21 is significant at the 95% level.

Figure 3a suggests that during the first month, there is substantial storm track prediction skill that appears to be not directly related to ENSO. The predictability in week 1-2 (weather forecast time scale) and week 3-4 (subseasonal time scale) both can contribute to the prediction skill in the first month. The subseasonal predictability of storm track variability will be the focus of the remainder of this study.

4. Sub-seasonal Predictions of ECA in S2S Models

4.1. Temporal Evolution of Prediction Skill in MME

For subseasonal prediction, we first investigate the MME prediction skill of weekly (week 1 to 4) ECA (Figure 4). Over most regions in the midlatitudes, the prediction skill in week 1 is above 0.6 and can reach as high as 0.8. The prediction skill decreases from week 1 to week 4, with the ACC reaching only about 0.3 to 0.4 in regions where the prediction skill is highest in week 4. In regions over East Asia, central and

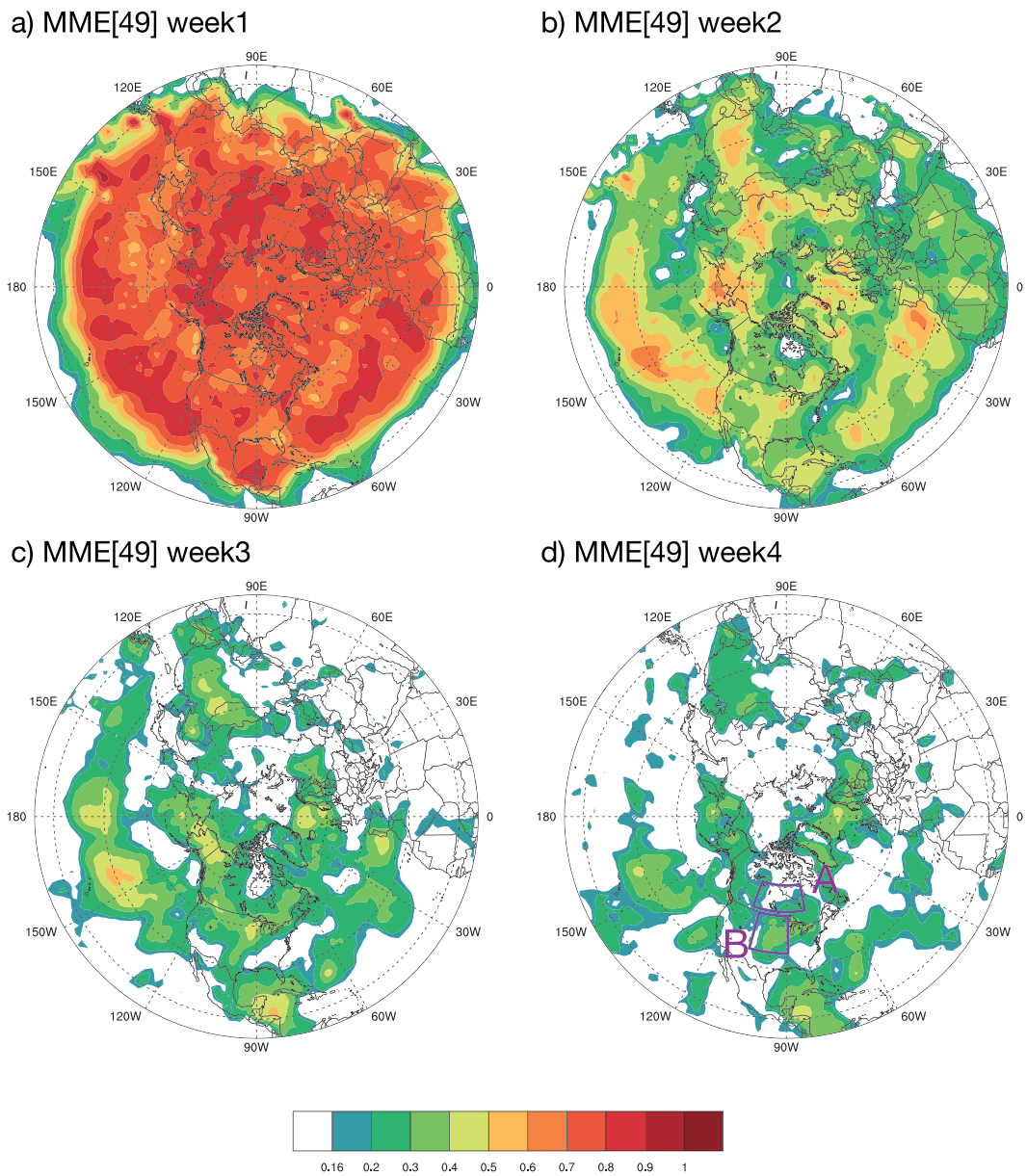

Figure 4. a)-d) Prediction skill (anomaly correlation coefficient, ACC) of multi-model ensemble (MME, 49 ensemble members) of extratropical cyclone activity for week 1 to 4 respectively. The region A (50.25°N-60.75°N, 110.25°W-78.75°W) and region B (35.25°N-48.75°N, 105.75°W-89.25°W) are plotted in d). See section 5 for definition of region A and B. For the 156 cases that are investigated here, a correlation of 0.16 is significant at 95% level. Note that the average interval between each case is about a week. In addition, over most of the regions, autocorrelation with 1-week lag of weekly ECA is not significant at 95% level.

eastern North Pacific, Bering Sea and Alaska, central parts of North America, Gulf of Mexico and North Atlantic, as well as Scandinavia and Norwegian Sea, the MME has relatively high ACC scores compared to other regions. As sub-seasonal forecast is one of the main focuses of the study, we will mainly investigate the week 3-4 (following CPC week 3-4 outlook and many other studies) performance of the S2S models.

4.2. Prediction Skill of Individual Models and MME in Weeks 3-4

The prediction skill of individual models is shown in Figure 5a-5f. CNRM, ECMWF and HMCR have better prediction skill than CMA, CNR-ISAC and ECCC (also see Figure S1 and Table S1). This does not necessarily mean CMA, CNR-ISAC and ECCC are worse than the other three models, as CMA, CNR-ISAC and ECCC only have 4 or 5 members while CNRM, ECMWF and HMCR have at least 10 members. Usually larger

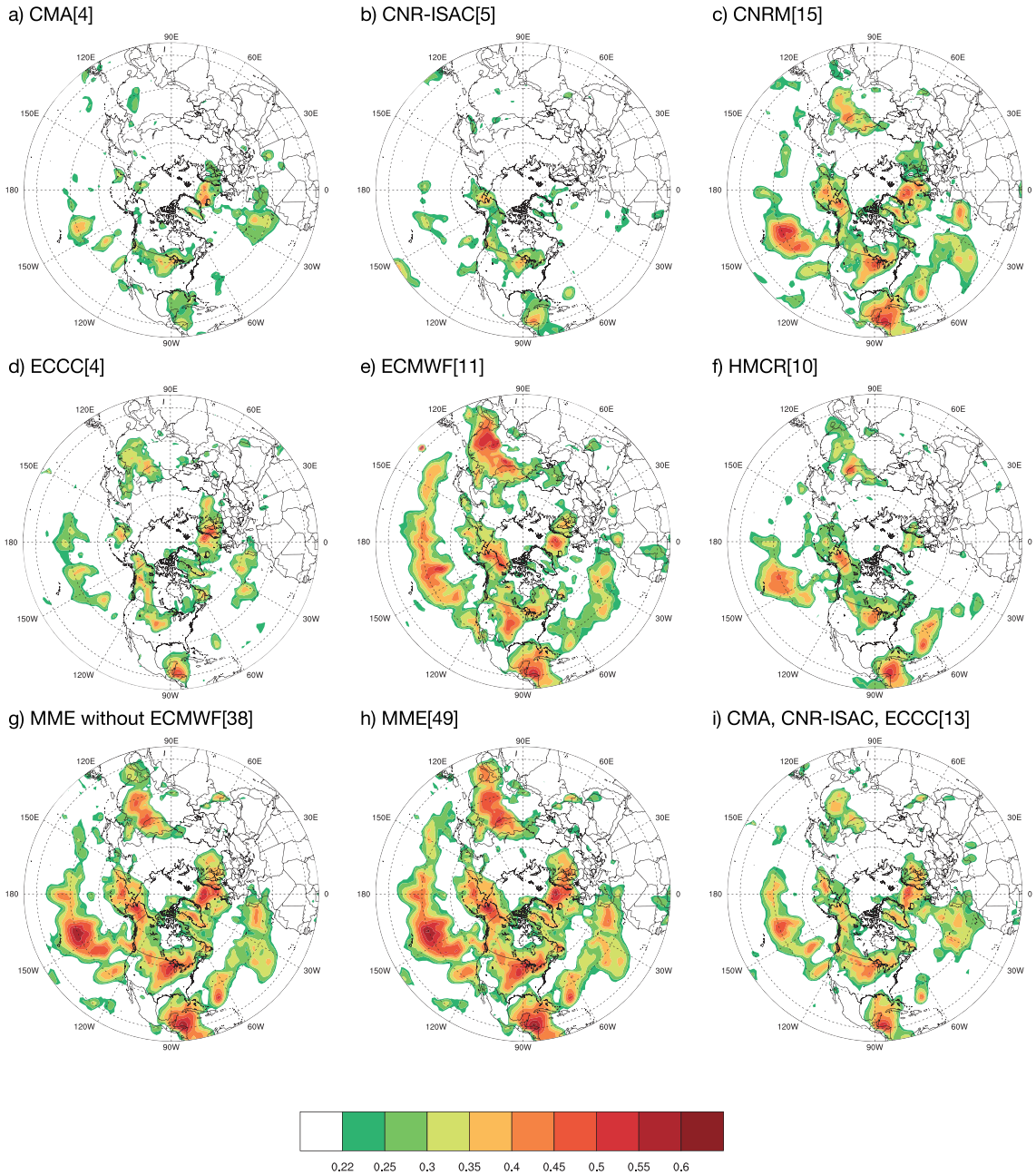

Figure 5. a)-f) Prediction skill (anomaly correlation coefficient, ACC) of week 3-4 extratropical cyclone activity for CMA (4 members), CNR-ISAC (5 members), CNRM (15 members), ECCC (4 members), ECMWF (11 members) and HMCR (10 members) respectively. g) The same as a) but combining CMA, CNR-ISAC, CNRM, ECCC and HMCR into a 38-member multi-model ensemble (MME). h) The same as g) but combining all the 6 models into a 49-member MME. i) The same as g) but combining CMA, CNR-ISAC and ECCC 13-member multi-model ensemble MME. For the 156 cases that are investigated here, a correlation of 0.22 is significant at 95% level. Note the over most of the regions, autocorrelation with 2-week lag of biweekly ECA is not significant at 95% level. As the average interval between each case is about a week, the estimated degree of freedom is 78 (half of 156).

ensemble size gives rise to higher prediction skill (e.g. Scaife et al., 2014; also see Figure S1). Though the EM prediction skill of ECCC is worse than CNRM and HMCR, the prediction skill of one single member of ECCC is very similar to CNRM and outperforms HMCR (Figure S1). Among the 6 models, ECMWF has the best prediction skill. If we combine CMA, CNR-ISAC and ECCC into a 13-member MME (Figure 5i), the skill is comparable to CNRM and HMCR. This shows that a larger ensemble size is beneficial to improve prediction skill, and that combining different models may also be beneficial as it may cancel part of the model biases. In all 6 models, the high prediction skill regions are consistent with the regions

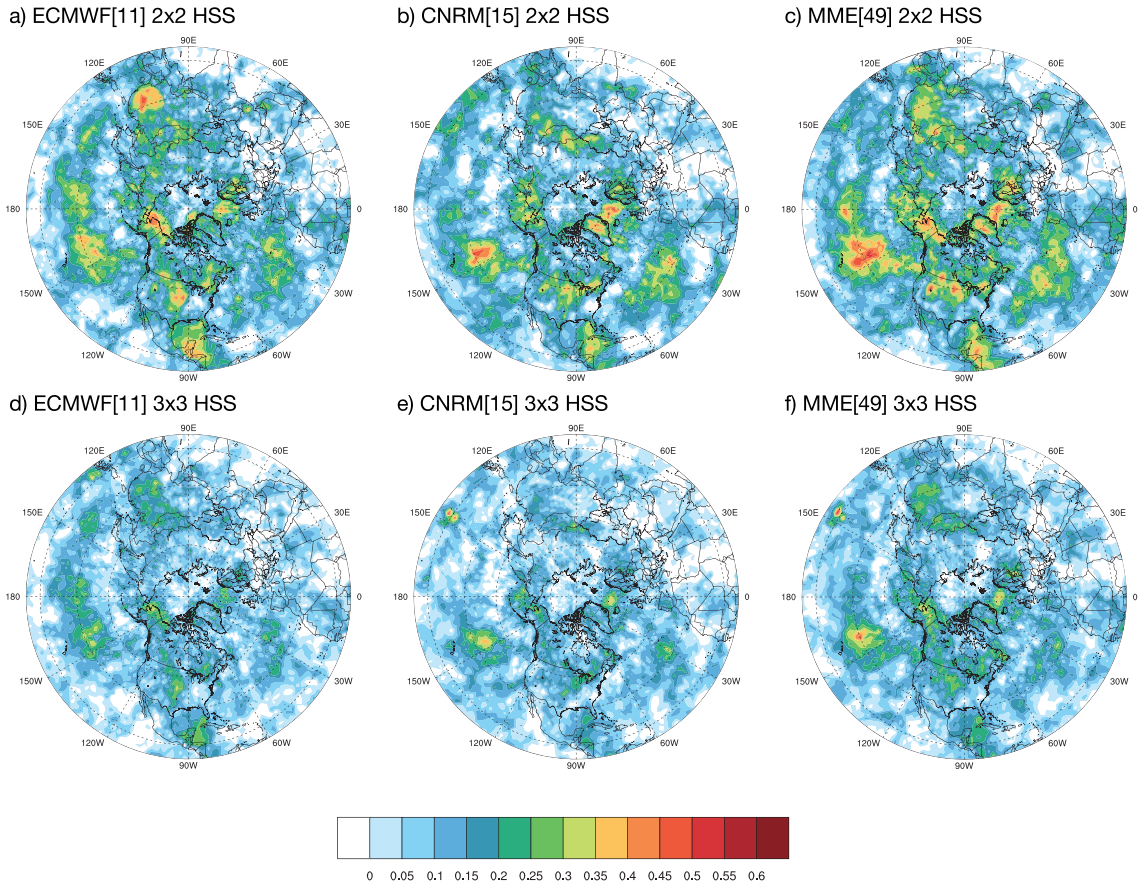

Figure 6. a) Heidke Skill Score (HSS) of ECMWF model for 2-category forecast of extratropical cyclone activity. b) The same as a) but for CNRM model. c) The same as a) but for multi-model ensemble mean (MME). d)-f) The same as a)-c), but for 3-category forecast.

discussed in section 4.1, which indicates that the dynamics giving rise to the prediction skill in different models is similar.

Though ECMWF performs best among single models, combining the other 5 models into an MME (Figure 5g) still outperforms ECMWF model overall, though ECMWF still has better skill over East Asia and western Pacific. This suggest that though the other models do not have as good skill as ECMWF individually, they still provide useful information for prediction. One potential reason why MME without ECWMF has better skill is also due to that the MME without ECMWF has larger ensemble size (38) than ECMWF (11). Another possible reason is that cancellation of individual model biases in MME may also improve the prediction skill. When all the models are combined into a 49-member MME, not surprisingly it has the best skill (Figure 5h, Figure S1 and Table S1). In regions over East Asia, central and eastern North Pacific, Bering Sea and Alaska, North America, Gulf of Mexico and North Atlantic, as well as Scandinavia and Norwegian Sea, the prediction of week 3-4 ECA can reach as high as 0.5. Potential sources of predictability over these regions will be discussed in Section 5. As ECMWF and CNRM has the best prediction skill and largest ensemble sizes among the 6 models, these two models will be investigated in detail in the following sections.

4.3. Week 3-4 HSS

As an alternative way to evaluate the model performance in predicting ECA in week 3-4, the HSS of ECMWF, CNRM and MME are shown in Figure 6 (area average HSS north of 10N° is shown in Table S1). The spatial structure of the HSS is very similar to the prediction skill of ECA (Figure 5e, 5c and 5h). The high HSS regions are over central Pacific, Bering Sea and Alaska, central North America, Gulf of Mexico, and some regions over Atlantic. Overall, the MME has higher HSS than ECMWF and CNRM, though in some

regions, individual model may outperform the MME. Again, this suggests that combining single models to MME usually improves the model forecast. Overall, current dynamical models exhibit some skill in predicting ECA in week 3-4 over most mid latitude regions.

5. Discussions

5.1. Source of Subseasonal Prediction Skill

5.1.1. MJO

As discussed in section 1, the MJO has significant impact on ECA in sub-seasonal time scale. In addition, current dynamical models have skillful prediction of the MJO out to 3-4 weeks (e.g. Kim et al., 2014; Kim et al., 2018; Lim et al., 2018; Vitart, 2017; Xiang et al., 2015). If the model can capture both the MJO and the MJO extratropical impact well enough, the MJO can be a source of predictability of ECA in week 3 to 4.

Here, following Zheng et al. (2018), we make MJO lag composites of ECA over regions A and B (see Figure 4 d) in ERA-Interim, ECMWF and CNRM, with respect to the 8 MJO phases and 28 lag days. Region A is the same (see Text S2) as region A in Zheng et al. (2018), where the signal-to-noise ratio of MJO impact on ECA is largest. Region B is over an area where both ECMWF and CNRM have good prediction skill. As there are only 156 cases when we combine the models into MME, if we separate these cases into 8 MJO phases, the lag composites will be noisy as there are only a limited number of cases. Therefore, we use all the reforecasts available from ECMWF (500 cases) and CNRM (252 cases) in DJF to make the model composites (forecast day number 0 is then the day when the reforecast is initialized instead of Day0 for MME). In Reanalysis composites, the MJO phase is defined as the phase at lag day 0, while in model (ECMWF and CNRM) composites, the MJO phase is defined as the phase at forecast day 0. As daily ECA is noisy, we perform a 7-day running mean on the ECA data prior to making MJO composites.

In region A, the MJO has significant impact on ECA in Reanalysis composites. The impact is significant in week 3 (lag day 17 for week 3 as a 7-day running mean is performed) after phases 6 to 8 (Figure 7a or 7e). The slope structure (e.g. significant impact in phase 8-2 in week 1 and subsequent significant impact in phase 6-8 in week 3) in Reanalysis composite is related to the propagation of the MJO. Both ECMWF (Figure 7c) and CNRM (Figure 7g) are able to capture some features of the MJO impact. But in the ECMWF hindcasts, starting from week 2, the amplitude of the MJO impact is weak compared to Reanalysis. In CNRM hindcasts, though the amplitude of MJO impact is similar to Reanalysis, the strong signal is not in the correct phase or lag (e.g. the negative signal is strong in phase 6 but weak in phase 8 during week 3-4 in Reanalysis data, but in CNRM hindcasts the negative signal is weak in phase 6 and strong over all the lags in phase 8). In both ECMWF and CNRM hindcasts, the slope structures in MJO composite are “flatter” compared with that in Reanalysis. Previous studies have shown that the propagation speed of the MJO has slow bias for these two models (see Figure 1 in Lim et al., 2018). As shown in Zheng and Chang (2019), if the MJO propagates slower or cannot propagate into specific phases, then the MJO extratropical impact will keep the same sign for a longer time. This may be the reason why the MJO composites in the models have flatter slopes. Single model member composites also show similar problems as ensemble mean composites (see Text S3). Therefore, neither the ECMWF nor the CNRM model are able to fully capture the MJO impact on ECA over region A, where in Reanalysis it is the region where the MJO impact is found to be most significant.

In region B, Zheng et al. (2018) found that the MJO does not have significant impact on ECA. Consistently, the MJO composite is weak and there is almost no significant signal after week 3 in Reanalysis data (Figure 7b). In the composites of ECMWF and CNRM model hindcasts (Figure 7d and 7h; single member composite in Figure S4 and S5), the ECA signal in region B also do not show strong relationship with the MJO phases. Thus, in region B, where the model has the highest prediction skill over North America (and one of the regions where the prediction skill is highest over land), the high prediction skill is not directly related to the MJO.

5.1.2. ENSO

Is the high prediction skill in week 3-4 forecast coming from ENSO, which is similar with monthly and seasonal forecast (section 3)? The absolute value of ACC between ERA-Interim ECA during week 3-4 in MME cases and DJF average Nino 3.4 index anomaly (also known as Oceanic Nino index (ONI)) is shown in Figure 8a. This can be regarded as using seasonal average Nino 3.4 index to hindcast week 3-4 ECA in the MME cases and can be directly compared with all the panels in Figure 4. Though using ONI index to

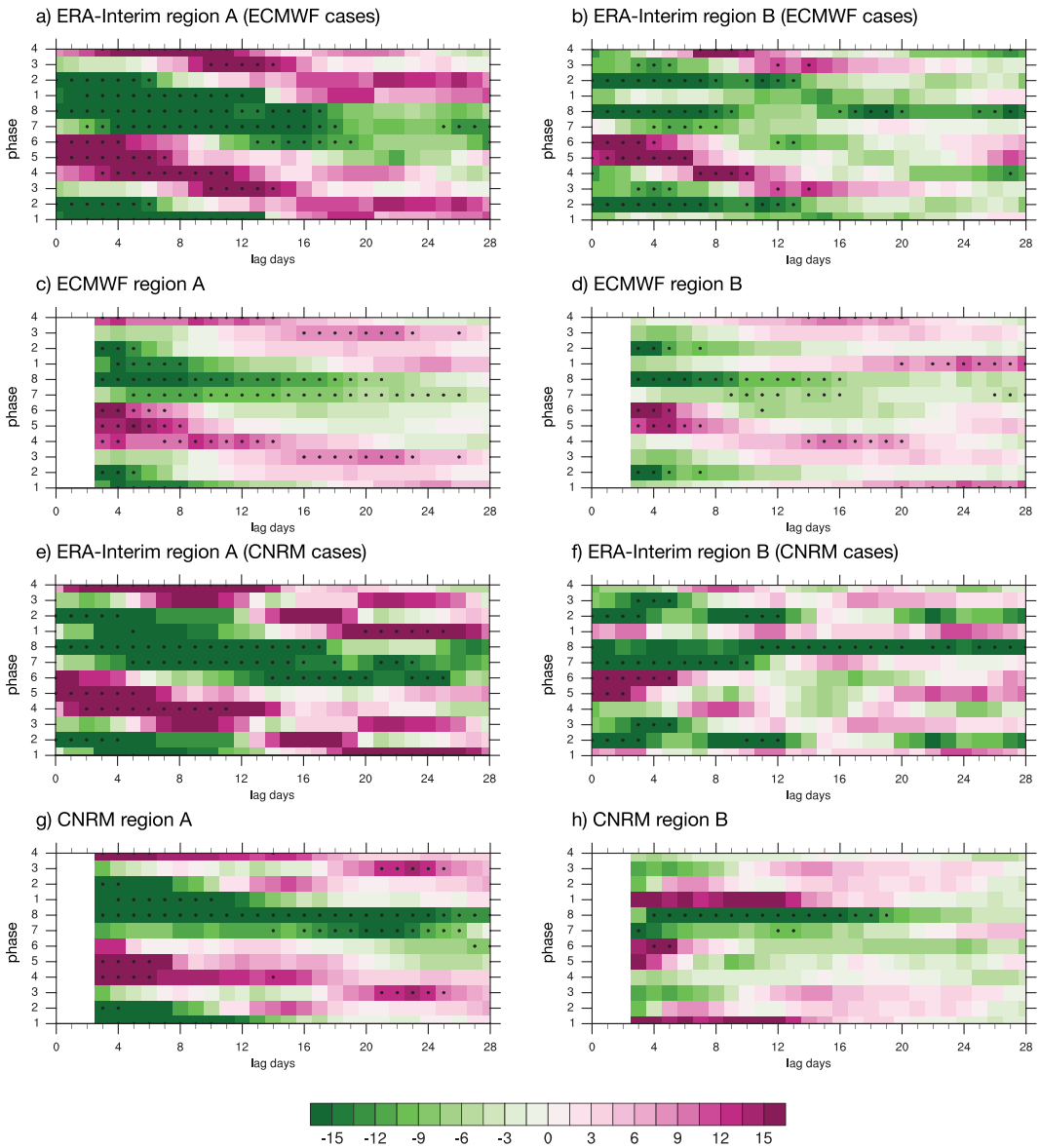

Figure 7. a) Lag composite of extratropical cyclone activity anomaly over region A with respect to 8 MJO phases with and lag day 0-28 for ERA-Interim in all ECMWF cases. A 7-day running mean is performed on the extratropical cyclone activity prior to making the composite. The dotted boxes are statistically significant at 95% from the results of a Monte Carlo test (see Zheng et al., 2018 for details of the method). b) The same as a) but for region B. c) The same as a) but for ECMWF model ensemble mean instead of ERA-Interim. Lag day 0 is the initialization day for the ECMWF reforecast. d) The same as c) but for region B. e)-f) The same as a)-b) but for all CNRM cases instead of ECMWF cases. g)-h) The same as c)-d) but for CNRM model instead of ECMWF model.

hindcast the week 3-4 ECA does not have as good skill as MME (Figure 5h), it is still comparable to single model ensembles (Figure 5a-5f). The important point to note is that the spatial pattern in Figure 8a is very similar to the panels in Figure 5 over the Pacific, North America and the western Atlantic. Many of the regions where dynamical models have high prediction skill (central Pacific, Alaska, central US, and Atlantic) correspond to regions where the ACC is high by using ONI to hindcast. Thus at least part of the prediction skill in these regions likely originates from ENSO.

How well do the models capture this signal? Similar ACC maps as Figure 8a but using individual members of ECMWF and CNRM are produced (Figure S6 and S7). Most of the model members capture the correlation between ONI and ECA, but the correlation is not as strong as that in Reanalysis, especially for ECMWF

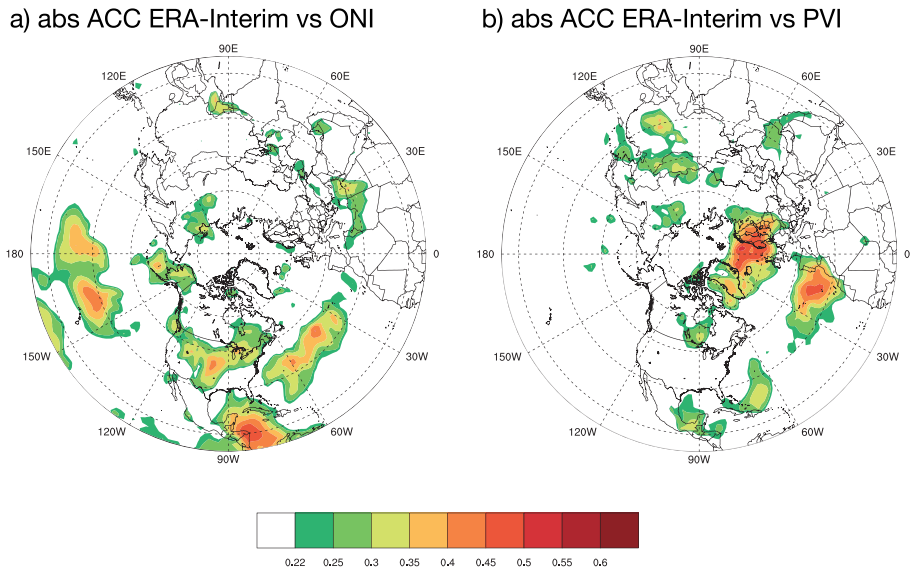

Figure 8. a) Absolute value of anomaly correlation coefficient (ACC) between ERA-Interim week 3-4 extratropical cyclone activity and winter season mean (December to February) Nino 3.4 index (also known as ONI index) in all MME cases. b) Similar to a), but for ACC between ERA-Interim week 3-4 extratropical cyclone activity and week 2-3 PVI index. Similar to Figure 5, for the 156 cases that are investigated here, a correlation of 0.22 is significant at 95% level.

reforecasts. Model EM ECA also has similar correlation pattern with ONI (Figure S13). The correlation between ECA over region B in week 3-4 MME cases and ONI index are shown in Table 2. Region B is where the models have relatively high prediction skill, and the correlation between ONI and ECA in Reanalysis is near 0.4. Though the ensemble mean of both ECMWF and CNRM model have higher correlation with ONI, however, only 1 of 11 ECMWF members and 2 of 15 CNRM members have higher correlation than the Reanalysis. As most of the members have lower correlation than the Reanalysis, the signal-to-noise ratio of ENSO-related signal in model hindcasts is lower than that in the Reanalysis in this region.

Table 2
The prediction skill over region B for ECMWF and CNRM model are shown in the first line. The absolute value of anomaly correlation c oefficient between ONI index and extratropical cyclone activity over region B in ERA-Interim, ensemble mean of ECMWF and CNRM model, as well as individual ensemble member are shown in the lines below. Individual members that have higher anomaly correlation between ONI and extratropical cyclone activity than that in the Reanalysis are highlighted.

	ECMWF	CNRM	ERA-Interim
Prediction Skill	0.568	0.470	
ONI vs Reanalysis			0.396
ONI vs ensemble mean	0.614	0.644	
ONI vs ensemble #1	0.261	0.292	
ONI vs ensemble #2	0.330	0.391	
ONI vs ensemble #3	0.208	0.263	
ONI vs ensemble #4	0.310	0.369	
ONI vs ensemble #5	0.269	0.292	
ONI vs ensemble #6	0.330	0.275	
ONI vs ensemble #7	0.329	***0.432***	
ONI vs ensemble #8	***0.414***	0.355	
ONI vs ensemble #9	0.292	***0.418***	
ONI vs ensemble #10	0.270	0.335	
ONI vs ensemble #11	0.331	0.269	
ONI vs ensemble #12		0.322	
ONI vs ensemble #13		0.169	
ONI vs ensemble #14		0.382	
ONI vs ensemble #15		0.385	

5.1.3. Polar Vortex

As discussed before, polar stratospheric anomalies can modulate ECA, especially over the Atlantic. Here we explore whether storm track predictability is related to the polar vortex. The absolute value of the ACC between the PVI in weeks 2-3 and ECA in weeks 3-4 for MME cases in ERA-Interim is shown in Figure 8b. The correlation is relatively high over the eastern North Atlantic, Scandinavia and Norwegian Sea, as well as over East Asia. The spatial pattern is similar to the forecast skill (Figure 5) from the Atlantic to the entire Eurasia, with similar amplitude over the Atlantic. This suggests that the prediction skill in these regions likely originates from the anomalies in the stratosphere. Similar correlation can also be found in the model hindcasts (Figure S14 and S15 for individual members of ECMWF and CNRM, Figure S16 for model EM), and the models also have good prediction skill of PVI in week 2-3 (Table S2). Note that for East Asia, though the correlation between PVI in week 2-3 and ECA is not high, averaging PVI over a longer period (Figure S8; day -28 to day 14, which is four weeks before forecast initialization date to week 2) increases the correlation over East Asia compared to Figure 8b, while the correlation over the Atlantic decreases. This suggests that the prediction skill over East Asia may be related to low frequency variability of the polar vortex. Detailed study of the stratospheric impact on East Asia

ECA is needed to fully explain the processes that give rise to the predictability.

5.1.4. QBO

The seasonal mean (DJF) zonal mean zonal wind at the equator at 30-hPa (QBO index, following Hamilton, 1984, Marshall & Scaife, 2009, and many others) is obtained from NOAA ESRL website. We make use of the index to examine the impact of the QBO on ECA on subseasonal time scale. Note the at least for the first few weeks, models can capture the QBO well (see Table S3; Garfinkel et al., 2018; Lim et al., 2019). The ACC between seasonal mean QBO and ECA in week 3-4 for MME case in ERA-Interim is not high (Figure S9). Though Wang, Kim, and Chang (2018) showed that the QBO significantly modulates ECA over both the Pacific and the Atlantic, here only relatively low correlation is found over the Atlantic and correlation over the Pacific is even lower. One possible reason is that the QBO impact on the extratropical storm track found by Wang, Kim, and Chang (2018) is mostly in the upper troposphere, while we are focusing on the ECA near the surface. This suggests that, at least in the 156 MME cases we analyzed, the QBO does not play an important role in subseasonal prediction of surface ECA.

5.2. Potential Predictability in ECMWF and CNRM

The potential predictability of storm track activity of ECMWF and CNRM are investigated because of the relatively high prediction skill of these two models. The potential predictability of CNRM (Figure 9b) over most regions is higher than its prediction skill (Figure 9a), although in some regions (e.g. eastern Pacific, northeast North America) the prediction skill is slightly higher than the potential predictability. This indicates that there is potential to improve the model prediction skill to reach the potential predictability if we assume the models are perfect forecast systems. However, the potential predictability of ECA in ECMWF hindcasts (Figure 9e) is lower than its prediction skill over many regions (Figure 9d). This is counterintuitive as it suggests that the ECMWF model is better at predicting the real atmosphere than its own ensemble members.

Previous studies have found that when predicting the North Atlantic Oscillation (NAO), for some models the potential predictability is lower than the prediction skill (e.g. Eade et al., 2014; Scaife et al., 2014). One proposed reason is that the signal-to-noise ratio is too low in the model. If the predictable component in the model could be well correlated with the predictable component in the real world, but the signal-to-noise (amplitude of predictable component versus model total variability) ratio is lower in the model than in the real atmosphere, then the potential predictability can be lower than the prediction skill as the low signal-to-noise ratio makes it harder for the model to predict its own ensemble members.

We use the root mean square (RMS) of the EM anomaly σ_{EM} to estimate the predictable component. If we have a large enough ensemble and a perfect model, the EM will be the same as the predictable component. As discussed in Eade et al. (2014), using σ_{EM} to represent the predictable component is an overestimation, since the amplitude of the EM will likely decrease if the ensemble size increases. The estimated signal-to-noise ratio of the model, η_{model}, can be defined as $\eta_{model} = \frac{\sigma_{EM}}{\sigma_{model}}$, where σ_{model} is model variability (RMS of model total anomaly). η_{model} of ECMWF and CNRM is shown in Figure 9f and 9c. As potential predictability can be considered as the ratio between model predictable component amplitude and model variability amplitude, η_{model} is expected to approach potential predictability when there is a very large ensemble. Since σ_{EM} is an overestimation of the model predictable component, η_{model} can be considered as the upper limit of potential predictability. The spatial structure of η_{model} (Figure 9f and Figure 9c) and potential predictability (Figure 9e and 9b) are very similar, which shows that η_{model} is a good indicator of model potential predictability.

Though η_{model} may only be slightly higher in CNRM (Figure 9c) than in ECMWF (Figure 9f), this may not be a fair comparison as η_{model} depends on ensemble size (since σ_{EM} depends on ensemble sizes, see discussion before). If we use the first 11 ensemble members in CNRM to calculate η_{model} (Figure S10c), it is clear that η_{model} in CNRM is indeed higher than that in ECMWF model. This can explain why CNRM potential predictability is higher than ECMWF. As η_{model} is a good indicator of model potential predictability, this suggests that the ECMWF potential predictability lower than prediction skill could be due to too small η_{model}, or that the signal-to-noise ratio is too small. The small signal-to-noise ratio is consistent with our previous discussions regarding the MJO and ENSO related signals found in the hindcasts, as individual ensemble

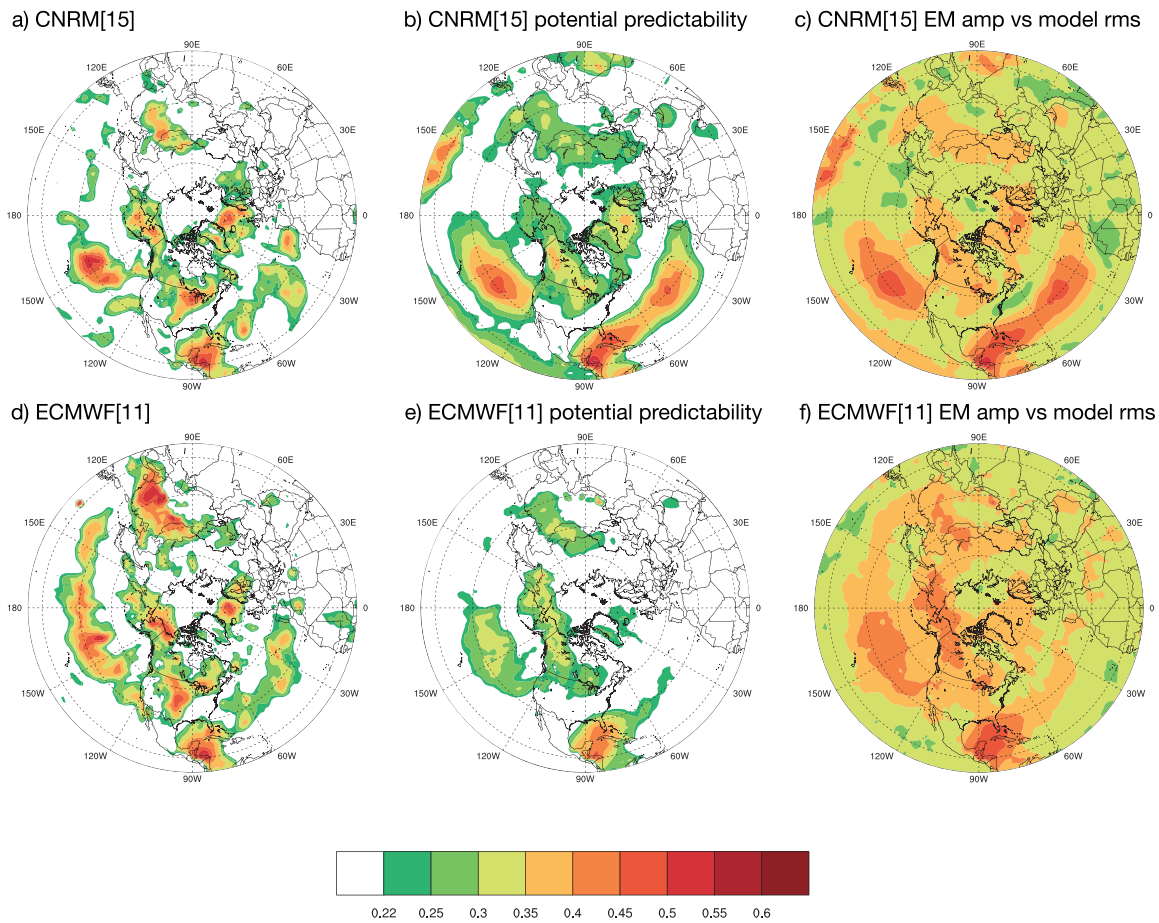

Figure 9. a) The same as Figure 5c). b) Potential predictability of week 3-4 extratropical cyclone activity for CNRM model. c) The estimated signal-to-noise ratio of CNRM (η_{model}), which is the ratio between amplitude of ensemble mean (root mean square of ensemble mean of each cases) and model total variability (root mean square of total model anomaly of each cases). d)-f) The same as a)-c) but for ECMWF model.

member has lower correlation with ENSO index (Figure S6 and S7) then that for Reanalysis (Figure 8a; also see Table 2). This low signal-to-noise problem is not limited to ECA but is also the case for Z500 (see Text S4 and Figure S11).

What causes the small signal-to-noise ratio in ECMWF? It can be due to too weak model predictable component (too small signal) or too large model variability (too large noise). We first investigate whether model variability σ_{model} is larger than Reanalysis variability σ_{obs} (RMS of Reanalysis anomaly). The ratio between model variability and Reanalysis variability can be written as $\chi_{noise} = \frac{\sigma_{model}}{\sigma_{obs}}$. χ_{noise} of ECMWF and CNRM is shown in Figure 10a and 10b. Though over the ocean χ_{noise} is around 1, over most regions over land the ratio is above 1. CNRM has larger ratio than ECMWF. This means that both ECMWF and CNRM has larger ECA variability (noise) than Reanalysis, and CNRM has larger variability than ECMWF.

It is not easy to investigate whether the predictable component (signal) in the models is too weak compared to Reanalysis, since the predictable component in Reanalysis is unknown. Here, following Eade et al. (2014), we use "ratio of predictable components" (RPC) to evaluate the predictable signal in the model. RPC can be written as,

$$RPC = \frac{PC_{obs}}{PC_{model}} \geq \frac{r}{\sqrt{\sigma_{EM}2/\sigma model^2}} = \frac{r}{\eta_{model}} \tag{6}$$

or,

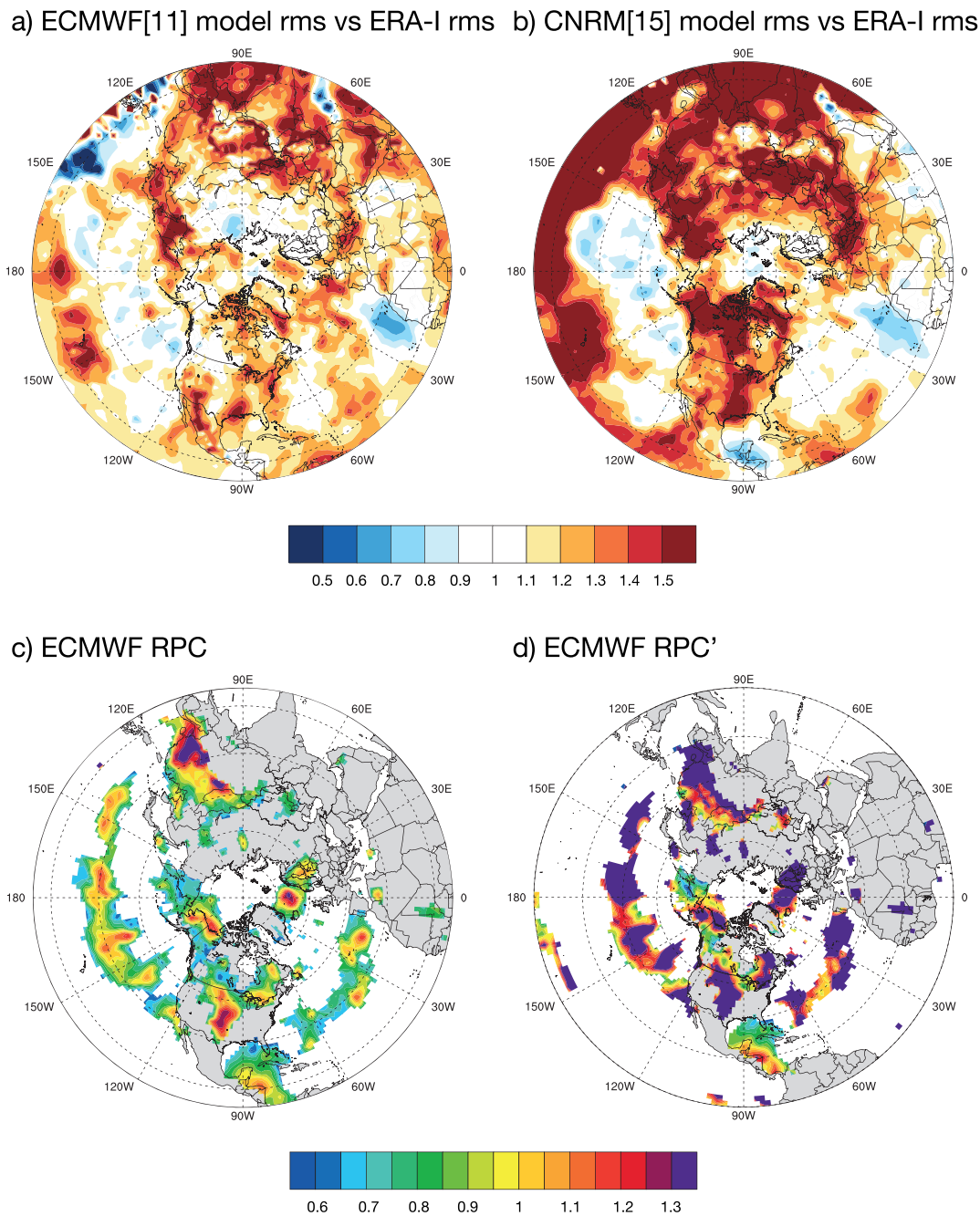

Figure 10. a) Ratio between root means square of ECMWF model total anomaly and root mean square of ERA-Interim total anomaly of extratropical cyclone activity. b) The same as a) but for CNRM model instead of ECMWF model. c) The RPC (equation (6) and (7)) value of ECMWF model. d) An alternative estimation of the ratio between Reanalysis predictable component and ECMWF model predictable component (RPC'), see details in equation (8). Only regions where ACC is above 0.22 (statistically significant at 95%) are shown in c) and d).

$$RPC = \frac{PC_{obs}}{PC_{model}} \geq \frac{r\,\sigma_{model}}{\sigma_{EM}} \tag{7}$$

where PC_{obs} is predictable component in Reanalysis, PC_{model} is predictable component in the model, and r is prediction skill (ACC) of the model. From equation (7), the predictable component in the Reanalysis is estimated from the fraction of the variance that can be explained by model forecasts, diagnosed from the ACC between Reanalysis and EM of model reforecast, as r^2 reflects the proportion of the Reanalysis anomaly

accounted for by the model anomaly (also see Text S5). This is likely an underestimation of predictable component, since larger ensembles or future improvements to the models can increase the correlations. The predictable component in the models is estimated by EM of the model. This is an overestimation since the variance of the ensemble mean would be reduced if there is a larger ensemble. So, the right-hand-side of equation (6) and (7) is the lower bound of RPC.

The right-hand-side of equation (6) can be estimated (Figure 10c) by the prediction skill (Figure 9d for ECMWF) divided by η_{model} (Figure 9f for ECMWF). We will not consider regions where the ACC is low – as the estimated predictable component in Reanalysis is proportional to ACC in our estimation, it is hard to tell whether small amplitude of estimated predictable component in Reanalysis is due to small amplitude of predictable component in reality or due to that the model is unable to fully capture the predictable component. Thus, only regions where ACC is above 0.22 (which is statistically significant at 95%) are shown in Figure 10c. In regions where the ACC is high in ECMWF (Pacific, East Asia, Alaska, Gulf of Mexico and North America), the ACC could reach above 0.45. However, in most of these regions (except Gulf of Mexico), η_{model} is smaller than 0.45. The RPC value is larger than 1 in these regions (see Figure 10c). Note that our estimation is the lower bound of RPC, which means the actual RPC could be larger than 1 over more regions. Therefore, over the regions where the prediction skill is high, ECMWF likely has smaller amplitude of predictable component than that in Reanalysis.

The RPC value in Figure 10c is the lower bound of the ratio between predictable component in Reanalysis and model. The following equation provides an alternative way to estimate the predictable component in Reanalysis and in model:

$$RPC' = \frac{PC_{obs}}{PC_{model}} \approx \frac{r\sigma_{obs}}{r_{potential}\sigma_{model}}, \tag{8}$$

where $r_{potential}$ is the potential predictability in the model (e.g. Figure 9e). In this equation, the amplitude of the Reanalysis predictable component is estimated by the prediction skill multiplied by Reanalysis variability, while the model predictable component is estimated by the model potential predictability times the model variability. The value of this ratio for ECMWF is shown in Figure 10d. Over most of the regions where the prediction skill is high, the predictable component in the Reanalysis is considerably larger than that in ECMWF. This provides further support to our hypothesis that the signal is too weak in ECWMF forecasts, and is likely one important reason why ECMWF potential predictability (Figure 9e) is lower than ECMWF prediction skill (Figure 9d).

6. Conclusions

In this study, we have evaluated the prediction of ECA on monthly to seasonal time scales using NMME and on subseasonal time scales using S2S dataset. For monthly and seasonal forecast, the high prediction skill lies in central part of North America and Alaska, with scattered regions of significant correlation over the Pacific and Atlantic after one month lead. On subseasonal time scales, high prediction skill is found over East Asia, central and eastern North Pacific, central part of North America, Gulf of Mexico and western Caribbean Sea, central North Atlantic, as well as Scandinavia and Norwegian Sea. ECMWF model has the best prediction skill among the 6 models we evaluated, though combining the other 5 models will lead to higher prediction skill. The prediction skill highly depends on the ensemble size. Models with 5 ensemble members or fewer do not show high prediction skill, but the ensemble which combines these models together into a larger ensemble performs much better. The reason why the MME outperforms each individual model is probably due to larger ensemble size, as well as cancelation of errors when combining different models together. The prediction of ECA is also evaluated by using the HSS, which shows that ECMWF, CNRM and MME have skill over most midlatitude regions.

The monthly and seasonal prediction skill mostly comes from ENSO after one month. For subseasonal prediction, various predictability sources that may give rise to the predication skill are examined. Though the MJO is found to have significant impact on ECA in subseasonal time scale, the benefit to the prediction skill in the S2S model is limited, as prediction skill is relatively low in the regions where the MJO have high impact. For ECMWF model, the MJO extratropical impact is too weak compared to Reanalysis. The CNRM model cannot capture the extratropical impact in the correct phase and lag. There are various

reasons why the models cannot capture the MJO extratropical response correctly. Most of the models have too weak MJO amplitude (e.g. Lim et al., 2018; Vitart et al., 2017), which may lead to too weak extratropical response. The propagation of the MJO, which may not be correctly captured by the models, highly modulates the timing, sign and duration of the MJO extratropical response (Zheng & Chang, 2019). In addition, biases of the midlatitude mean state in the models could be important in week 3-4 prediction since they can modify the extratropical response of the MJO in the midlatitudes (e.g. Henderson et al., 2017). For the Pacific, North America and Atlantic, the prediction skill is highly related to ENSO on subseasonal time scale, which is similar to that on seasonal time scale. These results are consistent with those of Johnson et al. (2014) who showed that week 3-4 surface temperature prediction skill over North America is largely due to ENSO and the trend, with little contributions from the MJO. For regions over Atlantic to Eurasia, the prediction skill is likely associated with stratospheric anomalies. A more detailed study of the prediction of ECA over Eurasia, especial over East Asia, is needed to better explain how the stratosphere is contributing to the prediction skill in these regions. The QBO shows only marginal contribution to the prediction skill mainly over the Atlantic in the subseasonal time scale.

Potential predictability is a common way to estimate the upper limit of what we can predict if we assume the model is a perfect forecast system. Here, we show that the potential predictability of ECA estimated from the ECMWF hindcasts is lower than the prediction skill. The reason is that the signal-to-noise ratio (the predictable part versus total variability) is too small in ECMWF model. Further analyses suggest that the noise over land areas is too large, and the amplitude of the predictable signals in the model may be too small. These suggest the need to improve the model in order to improved prediction in the future.

Acknowledgments
This research has been mainly conducted as part of the NOAA MAPP S2S Prediction Task Force and supported by NOAA Grant NA16OAR4310070. Analyses of NMME data were supported by DOE grant DE-SC0014050. The ERA-Interim Reanalysis, NMME model data and S2S data are cited in the reference list. The RMM index data is available at http://www.bom.gov.au/climate/mjo/. QBO and Nino 3.4 index are available at NOAA ESRL website. The authors would also like to thank three anonymous reviewers for comments that have helped to clarify and improve this paper.

References

Becker, E., den Dool, H. V., & Zhang, Q. (2014). Predictability and forecast skill in NMME. *Journal of Climate*, *27*(15), 5891–5906. https://doi.org/10.1175/JCLI-D-13-00597.1

Black, J., Johnson, N. C., Baxter, S., Feldstein, S. B., Harnos, D. S., & L'Heureux, M. L. (2017). The predictors and forecast skill of Northern Hemisphere teleconnection patterns for lead times of 3–4 weeks. *Monthly Weather Review*, *145*(7), 2855–2877. https://doi.org/10.1175/MWR-D-16-0394.1

Blackmon, M. L. (1976). A climatological spectral study of the 500 mb geopotential height of the Northern Hemisphere. *Journal of the Atmospheric Sciences*, *33*(8), 1607–1623. https://doi.org/10.1175/1520-0469(1976)033<1607:ACSSOT>2.0.CO;2

Chang, E. K., & Fu, Y. (2002). Interdecadal variations in Northern Hemisphere winter storm track intensity. *Journal of Climate*, *15*(6), 642–658. https://doi.org/10.1175/1520-0442(2002)015<0642:IVINHW>2.0.CO;2

Chang, E. K. M., Guo, Y., & Xia, X.. (2012). CMIP5 multimodel ensemble projection of storm track change under global warming. *Journal of Geophysical Research Atmospheres*, *117*(D23), D23118. https:doi:org/10.1029/2012jd018578

Chang, E. K., Guo, Y., Xia, X., & Zheng, M. (2013). Storm-track activity in IPCC AR4/CMIP3 model simulations. *Journal of Climate*, *26*(1), 246–260. https://doi.org/10.1175/JCLI-D-11-00707.1

Chang, E. K. M., Zheng, C., Lanigan, P., Yau, A. M. W., & Neelin, J. D. (2015). Significant modulation of variability and projected change in California winter precipitation by extratropical cyclone activity. *Geophysical Research Letters*, *42*, 5983–5991. https://doi.org/10.1002/2015gl064424

Chang, E. K., Lee, S., & Swanson, K. L. (2002). Storm track dynamics. *Journal of Climate*, *15*(16), 2163–2183. https://doi.org/10.1175/1520-0442(2002)015<02163:STD>2.0.CO;2

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

DelSole, T., Trenary, L., Tippett, M. K., & Pegion, K. (2017). Predictability of week-3–4 average temperature and precipitation over the contiguous United States. *Journal of Climate*, *30*(10), 3499–3512. https://doi.org/10.1175/JCLI-D-16-0567.1

Deng, Y., & Jiang, T. (2011). Intraseasonal modulation of the North Pacific storm track by tropical convection in boreal winter. *Journal of Climate*, *24*(4), 1122–1137. https://doi.org/10.1175/2010JCLI3676.1

Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L., & Robinson, N. (2014). Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophysical Research Letters*, *41*, 5620–5628. https://doi.org/10.1002/2014GL061146

Eichler, T., & Higgins, W. (2006). Climatology and ENSO-related variability of North American extratropical cyclone activity. *Journal of Climate*, *19*(10), 2076–2093. https://doi.org/10.1175/JCLI3725.1

Garfinkel, C. I., Schwartz, C., Domeisen, D. I., Son, S. W., Butler, A. H., & White, I. P. (2018). Extratropical Atmospheric Predictability From the Quasi-Biennial Oscillation in Subseasonal Forecast Models. *Journal of Geophysical Research: Atmospheres*, *123*, 7855–7866. https://doi.org/10.1029/2018JD028724

Guo, Y., Shinoda, T., Lin, J., & Chang, E. K. (2017). Variations of Northern Hemisphere storm track and extratropical cyclone activity associated with the Madden–Julian oscillation. *Journal of Climate*, *30*(13), 4799–4818. https://doi.org/10.1175/JCLI-D-16-0513.1

Hagedorn, R., DOBLAS-REYES, F. J., & Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting–I. Basic concept. *Tellus A*, *57*(3), 219–233. https://doi.org/10.1111/j.1600-0870.2005.00103.x

Haynes, P. H., McIntyre, M. E., Shepherd, T. G., Marks, C. J., & Shine, K. P. (1991). On the « Downward Control » of Extratropical Diabatic Circulations by Eddy-Induced Mean Zonal Forces. *Journal of the Atmospheric Sciences*, *48*(4), 651–678. https://doi.org/10.1175/1520-0469(1991)0482.0.co;2

Hamilton, K. (1984). Mean wind evolution through the quasi-biennial cycle in the tropical lower stratosphere. *Journal of the Atmospheric Sciences*, *41*(13), 2113–2125. https://doi.org/10.1175/1520-0469(1984)041<2113:MWETTQ>2.0.CO;2

Henderson, S. A., Maloney, E. D., & Son, S. W. (2017). Madden–Julian oscillation Pacific teleconnections: The impact of the basic state and MJO representation in general circulation models. *Journal of Climate*, *30*(12), 4567–4587. https://doi.org/10.1175/JCLI-D-16-0789.1

Johnson, N. C., Collins, D. C., Feldstein, S. B., L'Heureux, M. L., & Riddle, E. E. (2014). Skillful wintertime North American temperature forecasts out to 4 weeks based on the state of ENSO and the MJO. *Weather and Forecasting*, *29*(1), 23–38. https://doi.org/10.1175/WAF-D-13-00102.1

Kidston, J., Scaife, A. A., Hardiman, S. C., Mitchell, D. M., Butchart, N., Baldwin, M. P., & Gray, L. J. (2015). Stratospheric influence on tropospheric jet streams, storm tracks and surface weather. *Nature Geoscience*, *8*(6), 433–440. https://doi.org/10.1038/ngeo2424

Kim, H., Vitart, F., & Waliser, D. E. (2018). Prediction of the Madden–Julian Oscillation: A Review. *Journal of Climate*, *31*(23), 9425–9443. https://doi.org/10.1175/JCLI-D-18-0210.1

Kim, H. M., Webster, P. J., Toma, V. E., & Kim, D. (2014). Predictability and prediction skill of the MJO in two operational forecasting systems. *Journal of Climate*, *27*(14), 5364–5378. https://doi.org/10.1175/JCLI-D-13-00480.1

Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L. III, Paolino, D. A., Zhang, Q., et al. (2014). The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bulletin of the American Meteorological Society*, *95*(4), 585–601. https://doi.org/10.1175/BAMS-D-12-00050.1

Klein, W. H. (1957). Principal tracks and mean frequencies of cyclones and anticyclones in the Northern Hemisphere. Research Paper 40, U.S. Weather Bureau, 60 pp.

Kumar, A., Peng, P., & Chen, M. (2014). Is there a relationship between potential and actual skill? *Monthly Weather Review*, *142*(6), 2220–2227. https://doi.org/10.1175/MWR-D-13-00287.1

Lau, N. C. (1978). On the three-dimensional structure of the observed transient eddy statistics of the Northern Hemisphere wintertime circulation. *Journal of the Atmospheric Sciences*, *35*(10), 1900–1923. https://doi.org/10.1175/1520-0469(1978)035<1900:OTTDSO>2.0.CO;2

Lee, Y. Y., & Lim, G. H. (2012). Dependency of the North Pacific winter storm tracks on the zonal distribution of MJO convection. *Journal of Geophysical Research*, *117*, A01207. https://doi.org/10.1029/2011JA017246

Lim, Y., Son, S. W., & Kim, D. (2018). MJO prediction skill of the subseasonal-to-seasonal prediction models. *Journal of Climate*, *31*(10), 4075–4094. https://doi.org/10.1175/JCLI-D-17-0545.1

Lim, Y., Son, S. W., Marshall, A. G., Hendon, H. H., & Seo, K. H. (2019). Influence of theQBO on MJO prediction skill in the subseasonal-to-seasonal prediction models. *Climate Dynamics*, *53*(3-4), 1681–1695. https://doi.org/10.1007/s00382-019-04719-y

Ma, C.-G., & Chang, E. K. M. (2017). Impacts of Storm-Track Variations on Wintertime Extreme Weather Events over the Continental United States. *Journal of Climate*, *30*(12), 4601–4624. https://dio.org/10.1175/jcli-d-16-0560.1

Marshall, A. G., & Scaife, A. A. (2009). Impact of the QBO on surface winter climate. *Journal of Geophysical Research*, *114*, D18110. https://doi.org/10.1029/2009JD011737

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., et al. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research*, *108*(D14), 4407. https://doi.org/10.1029/2002JD002670

Scaife, A. A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R. T., Dunstone, N., et al. (2014). Skillful long-range prediction of European and North American winters. *Geophysical Research Letters*, *41*, 2514–2519. https://doi.org/10.1002/2014GL059637

Scaife, A. A., Spangehl, T., Fereday, D. R., Cubasch, U., Langematz, U., Akiyoshi, H., et al. (2012). Climate change projections and stratosphere–troposphere interaction. *Climate Dynamics*, *38*(9-10), 2089–2097. https://doi.org/10.1007/s00382-011-1080-7

Shaw, T. A., Baldwin, M., Barnes, E. A., Caballero, R., Garfinkel, C. I., Hwang, Y.-T., ... Voigt, A. (2016). Storm track processes and the opposing influences of climate change. *Nature Geoscience*, *9*(9), 656–664. https://doi.org/10.1038/ngeo2783

Smith, D. M., Scaife, A. A., Boer, G. J., Caian, M., Doblas-Reyes, F. J., Guemas, V., et al. (2013). Real-time multi-model decadal climate predictions. *Climate Dynamics*, *41*(11-12), 2875–2888. https://doi.org/10.1007/s00382-012-1600-0

Stockdale, T. N., Alves, O., Boer, G., Deque, M., Ding, Y., Kumar, A., ... Yun, W. T. (2010). Understanding and Predicting Seasonal-to-Interannual Climate Variability - The Producer Perspective. *Procedia Environmental Sciences*, *1*, 55–80. https://doi.org/10.1016/j.proenv.2010.09.006

Straus, D. M., & Shukla, J. (1997). Variations of midlatitude transient dynamics associated with ENSO. *Journal of the Atmospheric Sciences*, *54*(7), 777–790. https://doi.org/10.1175/1520-0469(1997)054<0777:VOMTDA>2.0.CO;2

Tian, D., Wood, E. F., & Yuan, X. (2017). CFSv2-based sub-seasonal precipitation and temperature forecast skill over the contiguous United States. *Hydrology and Earth System Sciences*, *21*(3), 1477–1490. https://doi.org/10.5194/hess-21-1477-2017

Vitart, F. (2017). Madden—Julian Oscillation prediction and teleconnections in the S2S database. *Quarterly Journal of the Royal Meteorological Society*, *143*, 2210–2220. https://doi.org/10.1002/qj.3079

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., et al. (2017). The subseasonal to seasonal (S2S) prediction project database. *Bulletin of the American Meteorological Society*, *98*(1), 163–173. https://doi.org/10.1175/BAMS-D-16-0017.1

Wallace, J. M., Lim, G. H., & Blackmon, M. L. (1988). Relationship between cyclone tracks, anticyclone tracks and baroclinic waveguides. *Journal of the Atmospheric Sciences*, *45*(3), 439–462. https://doi.org/10.1175/1520-0469(1988)045<0439:RBCTAT>2.0.CO;2

Walter, K., & Graf, H. F. (2005). The North Atlantic variability structure, storm tracks, and precipitation depending on the polar vortex strength. *Atmospheric Chemistry and Physics*, *5*(1), 239–248. https://doi.org/10.5194/acp-5-239-2005

Wang, J., Kim, H. M., & Chang, E. K. (2018). Interannual modulation of Northern Hemisphere winter storm tracks by the QBO. *Geophysical Research Letters*, *45*, 2786–2794. https://doi.org/10.1002/2017GL076929

Wang, J., Kim, H. M., Chang, E. K., & Son, S. W. (2018). Modulation of the MJO and North Pacific storm track relationship by the QBO. *Journal of Geophysical Research: Atmospheres*, *123*, 3976–3992. https://doi.org/10.1029/2017JD027977

Wheeler, M. C., & Hendon, H. H. (2004). An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Monthly Weather Review*, *132*(8), 1917–1932. https://doi.org/10.1175/1520-0493(2004)132<1917:AARMMI>2.0.CO;2

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*, (p. 676). Elsevier. New York, NY: Academic Press.

Xiang, B., Lin, S.-J., Zhao, M., Johnson, N. C., Yang, X., & Jiang, X. (2019). Subseasonal week 3–5 surface air temperature prediction during boreal wintertime in a GFDL model. *Geophysical Research Letters*, *46*, 416–425. https://doi.org/10.1029/2018GL081314

Xiang, B., Zhao, M., Jiang, X., Lin, S. J., Li, T., Fu, X., & Vecchi, G. (2015). The 3–4-week MJO prediction skill in a GFDL coupled model. *Journal of Climate*, *28*(13), 5351–5364. https://doi.org/10.1175/JCLI-D-15-0102.1

Yang, X., Vecchi, G. A., Gudgel, R. G., Delworth, T. L., Zhang, S., Rosati, A., ... Balaji, V. (2015). Seasonal Predictability of Extratropical Storm Tracks in GFDL's High-Resolution Climate Prediction Model. *Journal of Climate*, *28*(9), 3592–3611. https://doi.org/10.1175/jcli-d-14-00517.1

Zhang, F., Sun, Y. Q., Magnusson, L., Buizza, R., Lin, S. J., Chen, J. H., & Emanuel, K. (2019). What is the predictability limit of midlatitude weather? *Journal of the Atmospheric Sciences*, *76*(4), 1077–1091. https://doi.org/10.1175/JAS-D-18-0269.1

Zhang, Y., & Held, I. M. (1999). A linear stochastic model of a GCM's midlatitude storm tracks. *Journal of the Atmospheric Sciences*, *56*(19), 3416–3435. https://doi.org/10.1175/1520-0469(1999)056<3416:ALSMOA>2.0.CO;2

Zheng, C., & Chang, E. K. M. (2019). The role of MJO propagation, lifetime, and intensity on modulating the temporal evolution of the MJO extratropical response. *Journal of Geophysical Research: Atmospheres*, *124*, 5352–5378. https://doi.org/10.1029/2019JD030258

Zheng, C., Chang, E. K. M., Kim, H., Zhang, M., & Wang, W. (2018). Impacts of the Madden–Julian Oscillation on storm-track activity, surface air temperature, and precipitation over North America. *Journal of Climate*, *31*(15), 6113–6134. https://doi.org/10.1175/JCLI-D-17-0534.1