Subseasonal Prediction of Wintertime Northern Hemisphere Extratropical Cyclone Activity by SubX and S2S Models

CHENG ZHENG,[a] EDMUND KAR-MAN CHANG,[b] HYEMI KIM,[b] MINGHUA ZHANG,[b] AND WANQIU WANG[c]

[a] *Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York*
[b] *School of Marine and Atmospheric Sciences, Stony Brook University, State University of New York, Stony Brook, New York*
[c] *NOAA/Climate Prediction Center, College Park, Maryland*

(Manuscript received 4 September 2020, in final form 16 November 2020)

ABSTRACT: The prediction of wintertime extratropical cyclone activity (ECA) on subseasonal time scales by models participating in the Subseasonal Experiment (SubX) and the Seasonal to Subseasonal Prediction (S2S) is assessed. Consistent with a previous study that investigated the S2S models, the SubX models have skillful predictions of ECA over regions from central North Pacific across North America to western North Atlantic, as well as East Asia and northern and southern part of eastern North Atlantic at 3–4 weeks lead time. SubX provides daily mean data, while S2S provides instantaneous data at 0000 UTC each day. This leads to different variance of ECA. Different S2S and SubX models have different reforecast initialization times and reforecast time periods. These factors can all lead to differences in prediction skill. To fairly compare the prediction skill between different models, we develop a novel way to evaluate the prediction of individual model across the two ensembles by comparing every model to the Climate Forecast System, version 2 (CFSv2), as CFSv2 has 6-hourly output and forecasts initialized every day. Among the S2S and SubX models, the European Centre for Medium-Range Weather Forecasts model exhibits the best prediction skill, followed by CFSv2. Our results also suggest that while the prediction skill is sensitive to forecast lead time, including forecasts up to 4 days old into the ensemble may still be useful for weeks 3–4 predictions of ECA.

KEYWORDS: Ensembles; Forecast verification/skill; Hindcasts; Intraseasonal variability

1. Introduction

Extratropical cyclones have large impacts on regional weather and climate. They also have significant societal impacts, as these cyclones can bring heavy precipitation, strong winds, storm surge, and heavy snowfall, especially in wintertime. Therefore, accurate predictions of extratropical cyclone activity (ECA) can help to secure life and property against disastrous events, and provide useful information for decision makers in transportation, water security, agriculture and energy. To take multiple extratropical cyclones into account on weekly to seasonal time scales, the aggregate paths of extratropical cyclones, also referred to as extratropical storm tracks, are often used to represent ECA. Previous works have extensively studied observational, theoretical, and modeling aspects of ECA [see the review papers by Chang et al. (2002) and Shaw et al. (2016)]. Different phenomena can modulate ECA on various time scales (Chang et al. 2002; Chang et al. 2013; Stockdale et al. 2010). El Niño–Southern Oscillation (ENSO) significantly modulates Northern Hemisphere (NH) ECA on interannual time scales (Straus and Shukla 1997; Zhang and Held 1999; Eichler and Higgins 2006; Ma and Chang 2017). During El Niño years, an equatorward and eastward shift of boreal winter ECA is found over the Pacific, and ECA over North America weakens. The Madden–Julian oscillation (MJO) has significant impact on ECA over the North Pacific, the North Atlantic, and North America (Zheng et al. 2018; Deng and Jiang 2011; Lee and Lim 2012; Guo et al. 2017) via the MJO-induced Rossby waves that propagate into the midlatitudes. The quasi-biennial oscillation (QBO) gives rise to variability of NH ECA as well, especially in the upper troposphere (Wang et al. 2018a). Note that QBO also modulates the MJO impact on ECA over the Pacific (Wang et al. 2018b). The polar vortex in the NH stratosphere also has significant influence on ECA (Kidston et al. 2015; Scaife et al. 2012), especially over the North Atlantic (e.g., Walter and Graf 2005). Through the "downward control" mechanism (Haynes et al. 1991), the midlatitude jet and the North Atlantic Oscillation (NAO) are modulated by stratospheric wind anomalies, resulting in enhanced or suppressed ECA due to stronger or weaker zonal flow. The phenomenon mentioned above can be potential predictors for ECA on subseasonal time scales.

On the weather time scale, the track and intensity of a single extratropical cyclone can be skillfully predicted with a few days of lead time (e.g., Froude et al. 2007a,b; Froude 2010). For longer time scales (more than 2 weeks), due to the chaotic nature of the atmosphere, a single extratropical cyclone is not expected to be well predicted (Froude et al. 2007a,b; Froude 2010). Thus, for subseasonal prediction in this study, we will focus on ECA, which represents the aggregated influences (e.g., pressure, wind, eddy kinetic energy) of multiple extratropical cyclones. There have been two ways to represent ECA, the first uses cyclone tracks (e.g., Klein 1957), and the second uses statistics on gridded atmospheric data, for example, using variance of meridional wind or mean sea level pressure (MSLP) in a frequency band covering the synoptic time scales (e.g., Blackmon 1976; Lau 1978; Chang and Fu 2002). As passages of extratropical cyclones close to a given

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/WAF-D-20-0157.s1.

Corresponding author: Cheng Zheng, czheng@ldeo.columbia.edu

DOI: 10.1175/WAF-D-20-0157.1

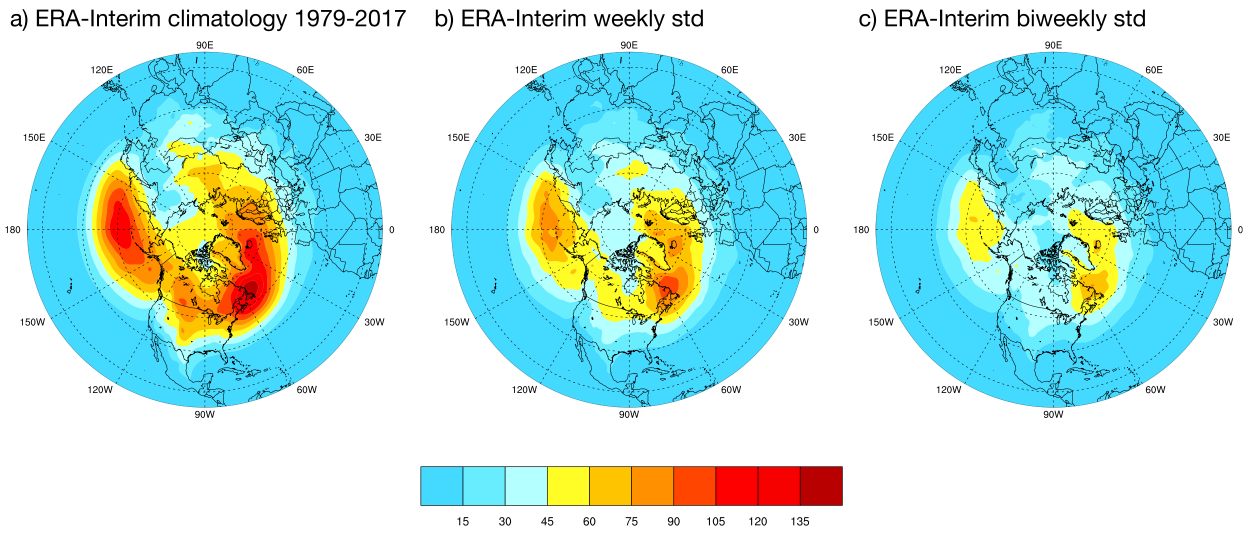

a) ERA-Interim climatology 1979-2017
b) ERA-Interim weekly std
c) ERA-Interim biweekly std
90E
120E
60E
150E
30E
180
0
150W
30W
120W
60W
90W
15
30
45
60
75
90
105
120
135

FIG. 1. (a) Climatology of the Northern Hemisphere extratropical cyclone activity (ECA; hPa^2) for 1979–2017 winters (December–February) based on all ERA-Interim reanalysis daily mean sea level pressure (MSLP) data. (b) Standard deviation of 7-day running mean ECA (hPa^2). (c) Standard deviation of 14-day running mean ECA (hPa^2).

location generates pressure perturbations as well as wind anomalies, temporal variance of pressure or wind at a grid point can capture the aggregate influence of extratropical cyclones over time. In this study, we apply a 24-h difference filter (Wallace et al. 1988) onto MSLP data to define ECA:

$$\mathrm{ECApp} = \overline{[\mathrm{MSLP}\,(t + 24\,\mathrm{h}) - \mathrm{MSLP}\,(t)]^2}, \qquad (1)$$

where t is any time step of Reanalysis or forecast dataset. Hence ECA is quantified at each grid point by the mean square of the 24-h difference of MSLP. The overbar represents averaging over time, which can be 1 week, 2 weeks, or 1 month. As shown by previous studies (e.g., Chang and Fu 2002; Wallace et al. 1988), the maxima from this 24-h difference filter, lie over locations where extratropical cyclones preferentially cross (see also Fig. 1a). Thus, the variance statistics of MSLP can be a good measure of ECA.

Both cyclone tracks and variance statistics have been used to evaluate ECA prediction by climate prediction models for subseasonal to seasonal forecasts. Befort et al. (2019) used cyclone tracks based on MSLP to assess seasonal prediction of ECA by the European Centre for Medium-Range Weather Forecasts (ECMWF) and British Meteorological Office climate prediction models in terms of ensemble mean cyclone track density, and found some skill on seasonal time scales over the North Atlantic, which is related to the North Atlantic Oscillation. Lukens and Berbery (2019) used cyclone tracks based on 850-hPa potential vorticity to assess subseasonal prediction of ECA by the National Centers for Environmental Prediction (NCEP) Climate Forecasting System, version 2 (CFSv2). They found that the root-mean-square errors in bias-corrected cyclone frequency and amplitude are close to or exceed one standard deviation, suggesting little prediction skill. However, Lukens and Berbery (2019) only used one single forecast member to estimate the prediction skill on subseasonal time scales. Usually the skill of an ensemble of multiple members has higher skill, as averaging over multiple members reduces the noise in the forecast data. Yang et al. (2015) assessed seasonal ECA predictions by the Geophysical Fluid Dynamics Laboratory climate prediction model based on MSLP variance. They found skill associated with the ENSO out to lead times of 9 months. Zheng et al. (2019; hereafter Z19) also used variance statistics based on MSLP to assess predictions of ECA in subseasonal to seasonal time scale (see below). The method to define ECA in this study [Eq. (1)] is the same as that in Z19. Z19 found that models that participated in the Seasonal to Subseasonal Prediction (S2S) project (Vitart et al. 2017) show significant prediction skill over East Asia, the central and eastern North Pacific, the central part of North America, Gulf of Mexico and the western Caribbean Sea, the central North Atlantic, as well as Scandinavia and the Norwegian Sea in week 3–4 predictions. The sources of predictability are mainly related to ENSO and the polar vortex. While the MJO can potentially be an important source of subseasonal predictability, the S2S models do not accurately capture the MJO's impact on ECA. In addition, Z19 did not find significant contributions directly from the QBO to ECA subseasonal predictions.

In this study, methods similar to Z19 will be applied to the models participating in the Subseasonal Experiment (SubX; Pegion et al. 2019) to evaluate model prediction skill. In addition, we will also compare model performance among S2S and SubX models. ECA is derived from MSLP data in this study. S2S provides instantaneous MSLP data at 0000 UTC each day, while SubX provides daily mean MSLP data. This results in different variability of ECA (see more details in section 2), which makes it inappropriate to combine the S2S and SubX models into a larger ensemble. Also, whether this difference in ECA variability between the SubX and S2S ensembles will lead to differences in prediction skill will be explored in this study. In addition, different models in the SubX and S2S ensembles perform reforecasts with different

TABLE 1. The description of SubX models that are used in this study. Note that for NCEP-CFSv2, we directly use the NCEP-CFSv2 reforecast and operational forecast, as MSLP is not archived in the SubX website. See main text for ensemble size of NCEP-CFSv2.

Model	Time range	Atmosphere model resolution	Reforecast frequency	Reforecast period	Reforecast Sizes	Ocean coupling	Sea ice coupling
Environmental and Climate Change Canada Global Ensemble Prediction System (ECCC-GEPS5)	Day 0–32	0.45° × 0.45°; 40 levels;	Weekly	1998–2017	4	No	No
National Centers for Environmental Prediction (NCEP) Environmental Modeling Center, Global Ensemble Forecast System (EMC-GEFS)	Day 0–35	T574L64 for 0–8 day and T382 for 8–35 day	Weekly	1999–2016	11	No	No
National Oceanic and Atmospheric Administration, Earth System Research Laboratory, Flow-Following Icosahedral Model (ESRL-FIM)	Day 0–32	~60 km; 64 vertical layers;	Weekly	1999–2016	4	Yes	Yes
National Aeronautics and Space Administration, Global Modeling and Assimilation Office, Goddard Earth Observing System (GMAO-GEOS)	Day 0–45	GEOS5–0.5° horizontal resolution; 72 vertical layers	Every 5 days	1999–2016	4	Yes	Yes
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)	Day 0–45	0.9° × 1.25°; L26	Weekly	1999–2016	3	Yes	Yes
National Centers for Environmental Prediction, Climate Forecast System, version 2 (NCEP-CFSv2)	Day 0–45	T126 L64	Every 6 h	—	—	Yes	Yes

initialization times and reforecast periods, which can lead to different prediction skill (sections 3 and 4). In this study, we will introduce a method to compare any SubX or S2S model to CFSv2 in a fair way. As CFSv2 is frequently initialized (every 6 h) with 6-hourly output available, one can construct a subsample of CFSv2 reforecasts, which has the same reforecast initialization time as that of any model. Then this subsample of CFSv2 can be fairly compared with that model. In this way, CFSv2 provides a bridge to compare the skill among different models. In section 2, the datasets and metrics to evaluate ECA predictions will be introduced. Prediction skill of ECA in SubX models will be evaluated in section 3. The comparison between SubX and S2S ensembles will be provided in section 4. The conclusions and some implications of this study will be discussed in section 5.

2. Data and methods

a. Data

1) SUBX AND S2S MODELS

At the time the analyses were performed, seven models from six participating modeling groups from the SubX dataset have available MSLP data with complete wintertime [December–January–February (DJF)] reforecasts. We make use of five of the models: ECCC-GEPS5, EMC-GEFS ESRL-FIM, GMAO-GEOS, and RSMAS-CCSM4 (see Table 1; we have problems processing the MSLP data from the other two models in SubX, see Text S1 in the online supplemental material). These five models have different ensemble sizes, reforecast initialization frequency, forecast time ranges and resolutions (see Table 1). Some of the models have coupled ocean and sea ice components, while others do not. We make use of daily MSLP data on a 1° × 1° horizontal resolution grid from the SubX dataset. The daily MSLP is the mean of 0000, 0600, 1200, and 1800 UTC of each day. ECA prediction will be evaluated during the overlapping period of SubX models (from DJF 1999/2000 to DJF 2015/16; 17 seasons in total).

To compare model prediction skill between SubX and S2S models, we make use of the six models from the S2S dataset evaluated by Z19 (CMA, CNR-ISAC, CNRM, ECCC-GEM, ECMWF, and HMCR; see Table 2). Similar to SubX models, there are also significant differences in the setup among the S2S models. MSLP data are available on a 1.5° × 1.5° horizontal resolution grid at 0000 UTC at each forecast day for S2S

TABLE 2. The description of S2S models that are used in this study. Note that for NCEP-CFSv2, we directly use the NCEP-CFSv2 reforecast and operational forecast. See main text for ensemble size of NCEP-CFSv2. Also note that NCEP-CFSv2 and ECCC-GEPS5 are not included in the S2S ensemble in Z19. As NCEP-CFSv2 ensemble is "initialized" every day (see main text), we also combine NCEP-CFSv2 into the S2S MME in this study. As combining ECCC-GEPS5 into the S2S MME will significantly reduce the number of available MME cases, ECCC-GEPS5 is not combined into the S2S MME. But we will compare ECCC-GEPS5 prediction skill with individual models in the S2S ensemble.

Model	Time range	Atmosphere model resolution	Reforecast frequency	Reforecast period	Reforecast Sizes	Ocean coupling	Sea ice coupling
China Meteorological Administration (CMA)	Day 0–60	T106 L40	Daily	1994–2014	4	Yes	Yes
Institute of Atmospheric Sciences and Climate of the National Research Council (CNR-ISAC) (model version date 6 Jun 2017)	Day 0–32	0.75° × 0.56° L54	Every 5 days	1981–2010	5	No	No
Météo-France/Centre National de Recherche Meteorologiques (CNRM)	Day 0–61	T255 L91	4 times a month	1993–2014	15	Yes	Yes
Environment and Climate Change Canada Model (ECCC-GEM) version: GEM Jan-2016	Day 0–32	0.45° × 0.45° L40	Weekly	1995–2014	4	No	No
European Centre for Medium-Range Weather Forecasts (ECMWF) Model version: CY43R3	Day 0–46	Tco639/319 L91	Twice a week	1997–2016	11	Yes	Yes
Hydrometeorological Centre of Russia (HMCR)	Day 0–61	1.1° × 1.4° L28	Weekly	1985–2010	10	No	No
National Centers for Environmental Prediction, Climate Forecast System, version 2 (NCEP-CFSv2)	Day 0–45	T126L64	Every 6 h	—	—	Yes	Yes
Environment and Climate Change Canada (ECCC-GEPS5) Model version: GEPS5 Sep 2018	Day 0–32	0.35° × 0.35° L45	Weekly	1997–2016	4	No	No

models. Z19 evaluated S2S model predictions over their overlap period from DJF 1997/98 to 2009/10. Here, when comparing models across SubX and S2S, only reforecasts in DJF from 1999/2000 to 2009/10 will be evaluated for both datasets.

CFSv2 (Saha et al. 2014; also see Table 1) participates in both the SubX and S2S projects. As MSLP data from CFSv2 is not archived in the SubX dataset, here we directly use the 6-hourly MSLP forecast data from CFSv2 reforecasts (1999–2011) and operational forecasts (2011–16). These reforecasts (or operational forecasts) are initialized every 6 h, and MSLP data are available on a 1° × 1° horizontal resolution grid every 6 h. In this study, CFSv2 is combined into both the SubX multimodel ensemble (MME) and S2S MME. To be consistent with other SubX models, when comparing with SubX models or constructing the SubX MME, CFSv2 ECA is calculated by using daily mean MSLP (average of MSLP at 0000, 0600, 1200, and 1800 UTC). Similarly, when comparing with S2S models or constructing the S2S MME, CFSv2 ECA is calculated by using MSLP at 0000 UTC, which is also regrided from 1° resolution to 1.5° resolution. As discussed in the introduction, frequent reforecasts with 6-hourly output available makes CFSv2 the bridge to compare different SubX and S2S models.

CFSv2 only provides one hindcast at each 6-hourly initialization time. One common way to combine the CFSv2 members into an ensemble is to use the lagged ensemble method (e.g., Chen et al. 2010, 2013; Riddle et al. 2013; Zhu et al. 2013). Chen et al. (2013) showed that the optimal number of lagged ensemble members that should be included in a lagged-ensemble is determined by a balance between two competing factors: increase in prediction skill due to a larger ensemble, and degradation of skill due to inclusion of members with longer lead times. They also showed that the optimal number depends critically on the variable predicted. Here, we use a 16-member lagged ensemble of CFSv2 (a lagged ensemble by using all reforecasts initialized within 4 days) for two reasons. First, when we construct the SubX or S2S MME, only the reforecasts initialized within 4 days are included (see more details in section 2b and Z19), the construction of the 16-member CFSv2 lagged ensemble is consistent with the way we construct the SubX or S2S MME. More importantly, as we will show later, the weeks 3–4 prediction skill of the lagged CFSv2 ensemble is still marginally increasing as we add members with longer lead time up to 16 members. Therefore, we believe that it is reasonable to construct the 16-member CFSv2 ensemble for weeks 3–4 prediction of ECA. Given that all other SubX models used in this study have initialization times that are separated by more than 4 days (see Table 1), CFSv2 is the only model that lagged ensemble is used.

2) Reanalysis and other datasets

ECA calculated from European Centre for Medium-Range Weather Forecasts (ECMWF) interim reanalysis (ERA-Interim; Dee et al. 2011). MSLP is used as the verification for ECA hindcasts in this study. When verifying SubX data, daily averages of MSLP are calculated from 6-hourly ERA-Interim MSLP on a 0.75° × 0.75° grid, and then regrided onto a 1° × 1° horizontal resolution grid. MSLP data at 0000 UTC is regrided to a 1.5° × 1.5° grid when verifying S2S models.

The phase of ENSO is defined by the Niño-3.4 index, which is obtained from the National Oceanic and Atmospheric Administration (NOAA) Earth System Research Laboratory (ESRL) website. This index is calculated from the Hadley Centre's Sea Ice and Sea Surface Temperature (SST) dataset (HadISST1; Rayner et al. 2003).

b. Methods

1) Definition of ECA

As discussed in the introduction, ECA is defined by applying a 24-h difference filter on MSLP data [Eq. (1)]. Daily MSLP is used for SubX data, while 0000 UTC instantaneous MSLP is used for S2S data. The winter climatology of ECA, as well as variability on weekly, and biweekly (week 3–4 prediction skill is evaluated in this study) time scales, are shown in Fig. 1. The variability is defined as the standard deviation of weekly or biweekly ECA during DJF in reanalysis. Figure 1a shows that ECA and its variability are maximized over the midlatitude oceanic basins, along a band extending from the western Pacific, across North America, the Atlantic, into northern Europe. Most of the contribution of ECApp is from extratropical cyclones and anticyclones, with almost no contribution from tropical cyclones, and the climatological ECApp is very small in the tropics (Fig. 1a). This is also the case for other seasons, including summer (not shown). Previous studies have shown that monthly and seasonal variations in ECA as defined by (1) are well correlated with variations in precipitation and weather extremes (e.g., Chang et al. 2015; Yang et al. 2015; Ma and Chang, 2017) in many midlatitude regions.

Although the climatologies of ECA computed using daily mean MSLP (Fig. 1a) and 0000 UTC MSLP (Fig. 1a in Z19) data look very similar, the amplitudes of reanalysis winter biweekly variability, (Fig. 1c and Fig. S1f in the online supplemental material; Fig. S1f is the same as Fig. 1c in Z19) are very different. The differences in Fig. 1c and Fig. S1f are not due to the differences in spatial resolution. As shown in Figs. S1c and S1d, compared with Figs. 1b and 1c (or Figs. S1a,b, which are the same), ECA variability computed on a 1.0° grid and on a 1.5° grid are almost identical, which is not surprising since the two different data resolutions merely represent regridding from the same original data. Thus, the differences in the amplitudes of weekly or biweekly variability are due to the use of daily mean MSLP versus 0000 UTC instantaneous MSLP to calculate ECA. The reason is that 0000 UTC MSLP is noisier than daily mean MSLP. Hence the variance of ECA calculated from 0000 UTC MSLP (Fig. S1f; or Fig. 1c in Z19) is larger than that calculated from daily mean MSLP (Fig. 1c). As using 0000 UTC MSLP or daily mean MSLP leads to different variability of ECA in reanalysis data, certainly it will also lead to different ECA variability calculated from S2S and SubX models. With different amplitude of variability in ECA simply because of the way MSLP is archived but not because of the models themselves, it is inappropriate to combine ECA from SubX and S2S models into a larger ensemble and then calculate ensemble mean. Also, with the differences in ECA variability, whether using daily mean MSLP or 0000 UTC MSLP will lead to differences in the prediction skill of ECA remains unclear. This will be examined in section 4.

2) Climatology and anomalies of ECA

Bias corrections for subseasonal forecasts are important as the model bias can become dominant on subseasonal time scales (e.g., Monhart et al. 2018). For SubX models, a model climatology that depends on the model initialization time and forecast day (e.g., forecast day 1, forecast day 2…) is defined to correct the model bias of ECApp but not MSLP. Similar to Z19, all the reforecast of ECA of one single model can be written as $\mathrm{ECApp}_{\mathrm{model}}(y, d, n, f)$, where y is year, d is initialization day during each year, f is the forecast lead day. $n = 1, \ldots, N$, where N is the number of ensemble members for each model. The model climatology at each grid point, which depends on the reforecast initialization time and forecast day, is obtained by averaging all the years and removing the first four harmonics of the annual cycle:

$$\mathrm{ECApp}_{\mathrm{cli}}(d,f) = \frac{\sum_{y}\sum_{n} \mathrm{ECApp}_{\mathrm{model}}(y,d,n,f)}{N \times Y}. \quad (2)$$

Here, Y is the total number of years. Model anomalies are then defined as the deviation from model climatology:

$$\mathrm{ECApp}_{\mathrm{ano}}(y,d,n,f) = \mathrm{ECApp}_{\mathrm{model}}(y,d,n,f) - \mathrm{ECApp}_{\mathrm{cli}}(d,f). \quad (3)$$

Note that each model has its own climatology $\mathrm{ECApp}_{\mathrm{cli}}(d, f)$, and the anomaly of each model $\mathrm{ECApp}_{\mathrm{ano}}$ is the deviation from the model's own climatology $\mathrm{ECApp}_{\mathrm{cli}}(d, f)$. Following Z19, ECA model climatology and anomaly for the S2S models are defined similarly. Also, as discussed in Z19, the similar method can be applied to define reanalysis climatology and anomalies, except that there is only one ensemble member for the reanalysis and the reanalysis climatology does not depend on the forecast day.

3) Combining different models into an MME

It has been shown by previous studies that in terms of prediction skill, a multimodel ensemble (MME) usually outperforms a single model (e.g., Hagedorn et al. 2005; Smith et al. 2013; Becker et al. 2014). In addition, Z19 shows that combining S2S models into an MME is beneficial to ECA subseasonal forecast. Here, we also combine the SubX models into an MME. Since the reforecasts are initialized on different dates for different SubX models, we follow a procedure similar to Z19 to construct the MME. During the overlapping winter seasons of the SubX models (DJF from 1999/2000 to 2015/16), for every day and every model, we define the lead time of the reforecast. The lead time at any day of a model forecast is the gap between this day and the initialization time of the nearest reforecast earlier than this day. We select the days as day 0 of the MME if that day satisfies all the following requirements: 1) Every model has a lead time less than or equal to 4 days. 2) If continuous days satisfy the requirement (1), only the earliest day is selected (to make the lead time smallest). 3) The averaged lead time of the 6 S2S models is smaller than 1.5 days. There are 182 cases that we can combine the SubX models into an MME in 17 winter seasons. The procedure here is the same as Z19, except for the additional requirement 3. This requirement is added not only to reduce the lead time of the models (which is important for the prediction skill, see sections 3 and 4), but also to reduce the MME case frequency from about twice a week to about once a week. As the reforecasts of four SubX models are initialized once a week, we want to make the frequency of MME cases to be about once a week in order to avoid using one run of any model in multiple MME cases.

4) Prediction skill of ECA

We use the anomaly correlation coefficient (ACC) to assess the model prediction skill. The association between the anomalies in forecast and analysis can be represented by the ACC. When calculating the ACC, we use the ensemble mean (EM) of a single model, or the EM of an MME (every model member is weighted equally). The ACC at any grid point can be written as

$$\mathrm{ACC}=\frac{\sum\limits_{y}\sum\limits_{d}\mathrm{ECApp}_{\mathrm{ano}}^{\mathrm{EM}}(y,d)\mathrm{ECApp}_{\mathrm{ano}}^{\mathrm{obs}}(y,d)}{\sqrt{\sum\limits_{y}\sum\limits_{d}[\mathrm{ECApp}_{\mathrm{ano}}^{\mathrm{EM}}(y,d)]^2\sum\limits_{y}\sum\limits_{d}[\mathrm{ECApp}_{\mathrm{ano}}^{\mathrm{obs}}(y,d)]^2}}, \quad (4)$$

where $\mathrm{ECApp}_{\mathrm{ano}}^{\mathrm{EM}}(y, d)$ represents the ensemble mean of model forecast anomalies, and $\mathrm{ECApp}_{\mathrm{ano}}^{\mathrm{obs}}(y, d)$ is the anomaly in the reanalysis data. Following Z19, only weekly or biweekly ACC is calculated. We have also examined the Heidke skill score, but the results are consistent with those using ACC as well as the results presented in Z19 and thus are not shown (see Text S2 in the supplemental material).

5) Comparison of prediction skill between CFSv2 and other models

As the reforecasts are initialized differently among the SubX models, usually the number of reforecasts and the reforecast initialization times are different between any of two SubX models. After combining the SubX models into an MME, the number of reforecasts is the same. But different models have different lead time in the MME. As lead time can degrade the forecast skill (see section 3), it may not be fair to directly compare the forecast skill of two models using the MME cases. As CFSv2 ensemble is available every day, here we develop a way to directly compare CFSv2 versus any other model. For any model in the SubX ensemble, say EMC-GEFS, we just use a subsample of the CFSv2 ensemble. We make this subsample of CFSv2 reforecasts have the same reforecast initialization dates as EMC-GEFS. Then the forecast skill calculated within this subsample of CFSv2 reforecast can be fairly compared with EMC-GEFS, since they are both initialized during the same dates. Note that here we still use the same method mentioned above in this section to calculate forecast skill (ACC), and the bias corrections of each model are still based on the model's own climatology. This method can show the forecast skill of any model relative to CFSv2.

3. ECA predictions by SubX models

Weekly ACC for the MME of weeks 1–4 is shown in Fig. 2. The prediction skill decreases from week 1 to 4, as the ACC is above 0.6 almost everywhere in the midlatitudes in week 1 (Fig. 2a), and the highest ACC during week 4 (Fig. 2d) is only 0.3–0.4. Consistent with the S2S models (see Z19), starting from week 2 (Figs. 2b–d), high ACC is found over east Asia, the central and eastern North Pacific, the Bering Sea and Alaska, central North America, the Gulf of Mexico and western Caribbean Sea, and the North Atlantic along 30°–45°N and 60°–75°N. Following Z19, weeks 3–4 (week 3 and week 4 combined) prediction will be the main focus here.

Weeks 3–4 ECA ACC for several models and the MME are shown in Fig. 3. Note that as discussed in Z19, models with small ensemble sizes (e.g., fewer than five ensemble members) generally have relatively low prediction skill. Therefore, the models with smaller ensemble size in the SubX ensemble, which are ECCC-GEPS5, ESRL-FIM, GMAO-GEOS, and RSAMS-CCSM4, are combined into one larger ensemble. ACC of these four models is shown in Fig. S2. This four-model ensemble (Fig. 3e; equivalent to Fig. 3b without EMC-GEFS) with 15 members, has better prediction skill than the 11-member EMC-GEFS (Fig. 3c), but has lower ACC than the 16-member CFSv2. CFSv2 has the best prediction skill among the SubX models, in terms of single model performance. The MME (Fig. 3a) has better prediction skill than any single model. Combining the models other than CFSv2 (Fig. 3b), provides similar prediction skill compared to CFSv2.

Figure 4a shows the ACC averaged over the NH (north of 10°N) of individual models and the MME for the 182 MME cases. The *x* axis represents the size of the ensemble, which is different for CFSv2 and for the other models. For the other models in Fig. 4a, if the value of the *x* axis is equal to *x*, we select *x* members from the total *N* members randomly for 200 times and use the average ACC of these 200 samples as the ordinate. For CFSv2, as different ensemble members have different lead times, we use the ACC of the latest member for $x = 1$, the ACC of the ensemble combining the latest two members for $x = 2$, and so on. Thus, the CFSv2 line (cyan line) looks noisy as there is less averaging for each point for the line. Nevertheless, we can see that the prediction skill for CFSv2 generally increases as the number of lagged-ensemble members is increased, but the rate of increase slows considerably as

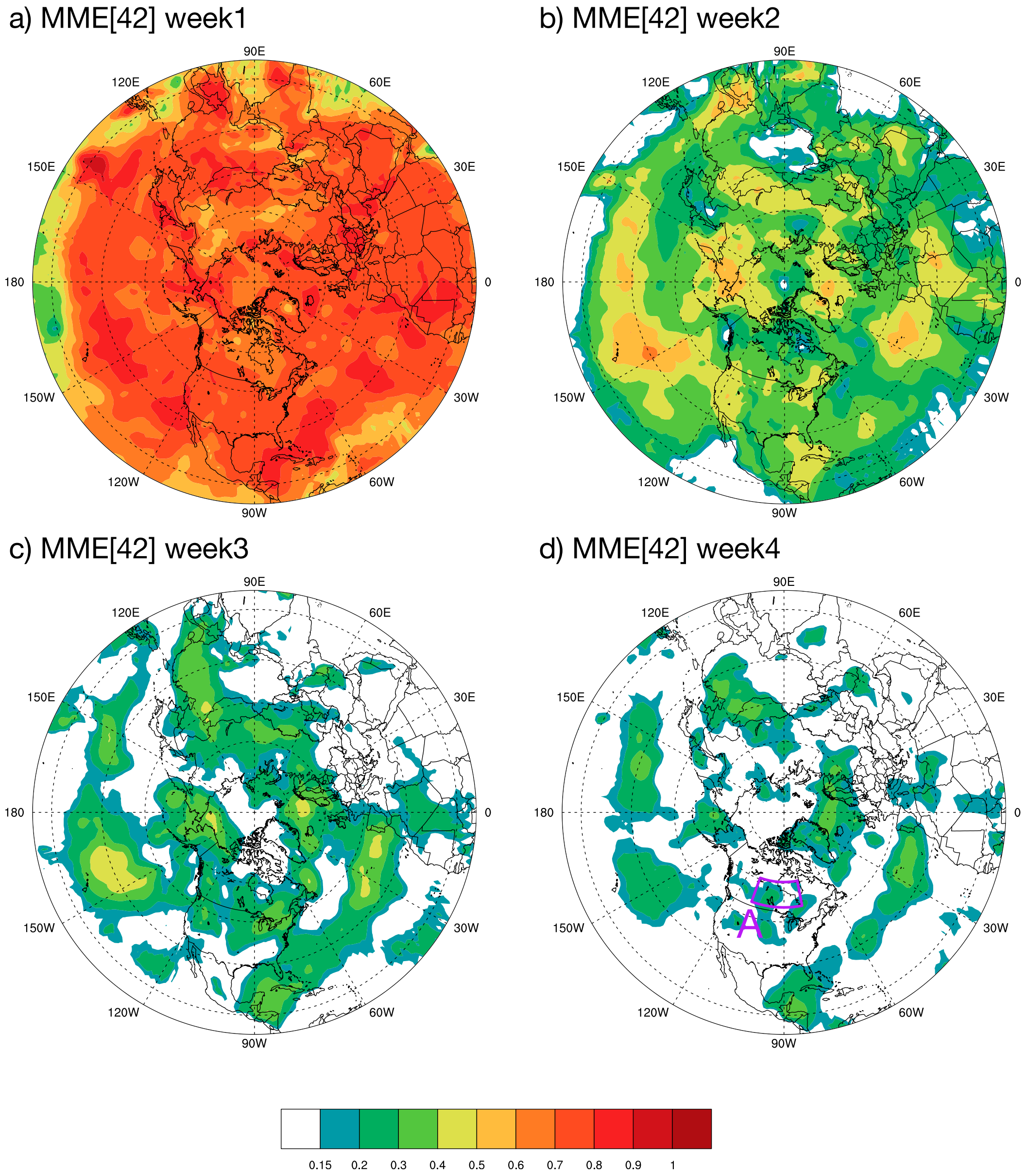
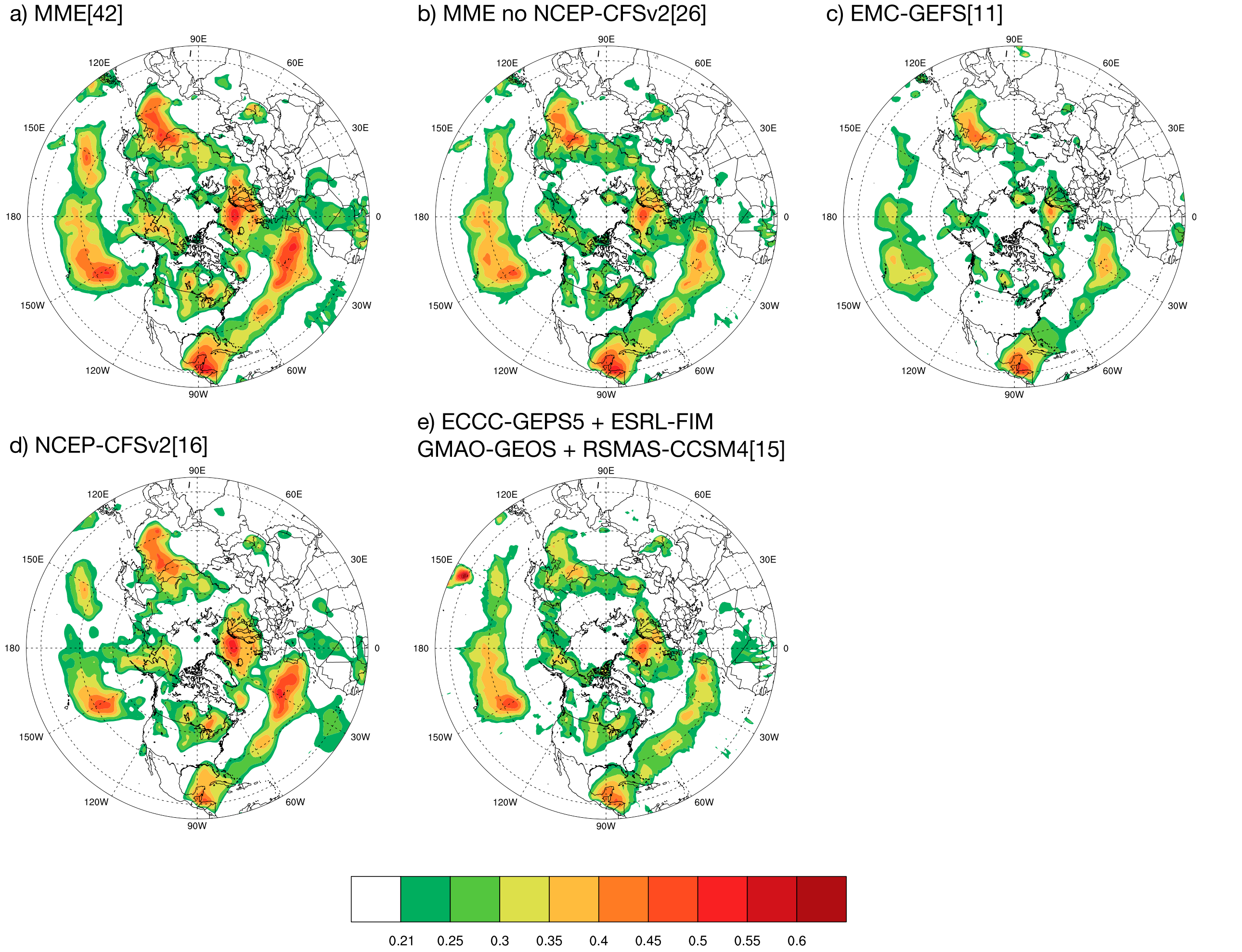


FIG. 2. (a)–(d) Prediction skill [anomaly correlation coefficient (ACC)] of multimodel ensemble (MME, 42 ensemble members) of extratropical cyclone activity for week 1–4, respectively. The region A (50.5°–60.5°N, 110.5°–78.5°W) is plotted in (d). See Text S3 for definition of region A. For the 182 cases that are investigated here, a correlation of 0.15 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 the 95% level.

the number of members reaches 16. Thus a 16-member lagged ensemble for CFSv2 is appropriate.

The MME without CFSv2 (dashed blue line), which has 26 members, has similar ACC compared to the 16-member CFSv2. It is quite clear that the MME (dashed red line) has the best ACC. The five SubX models other than CFSv2, have relatively similar performance, as the red, green, dark green, purple, and blue solid lines are clustered together. EMC-GEFS ensemble has high ACC due to larger ensemble size. And these five models have lower ACC than CFSv2. One potential reason is that with the MME cases here, the first member of CFSv2 has no lead time (as it is initialized every 6 h), while all members of the other models generally have nonzero lead times. In Fig. 4c, we use all the available cases in CFSv2 (every day) in DJF and show the NH averaged prediction skill of CFSv2 ensemble for lead time from 0 to 3 days. We use the same 16 members at lead time 0. For lead time 1 only 12 members (the latest 4 members excluded) are used, and so on. Therefore, the ensemble size is


a) MME[42]

b) MME no NCEP-CFSv2[26]

c) EMC-GEFS[11]

d) NCEP-CFSv2[16]

e) ECCC-GEPS5 + ESRL-FIM GMAO-GEOS + RSMAS-CCSM4[15]

FIG. 3. (a) Prediction skill [anomaly correlation coefficient (ACC)] of week 3–4 extratropical cyclone activity for MME (42 members). (b)–(d) As in (a), but for MME without NCEP-CFSv2 (26 members), EMC-GEFS (11 members), and NCEP-CFSv2 (16 members), respectively. (e) As in (a), but for the MME of ECCC-GEPS5, ESRL-FIM, GMAO-GEOS, and RSMAS-CCSM4. For the 182 cases that are investigated here, a correlation of 0.21 is significant at the 95% level. Note the over most of the regions, autocorrelation with 2-week lag of biweekly ECA is not significant at the 95% level. As the average interval between each case is about a week, the estimated degree of freedom is 91 (half of 182).

smaller for longer lead time. It is very clear that the prediction skill systematically decreases with longer lead time. Thus, because of different lead times, it is not fair to compare CFSv2 with other models using the MME cases (Fig. 4a).

To better compare the other SubX models to CFSv2, we plot the ACC of each individual model using all the available cases of this model with 0 lead time (solid lines in Fig. 4b). With the method discussed in section 2b, we use subsamples of CFSv2 to compare the prediction skill of one model and CFSv2. Each subsample of CFSv2 has the same initialization times as one of the models. The ACC of the subsamples of CFSv2 are plotted in dashed lines. The CFSv2 subsample corresponding to each SubX model is plotted in the same color as that model. For example, GMAO-GEOS is plotted in the blue solid line. The subsample of CFSv2 corresponding to GMAO-GEOS, is shown in the blue dashed line. Thus, the comparison between one SubX model and CFSv2 can be achieved by comparing the solid lines and the dashed lines having the same color. By comparing the solid lines and the dashed lines, it is clear that CFSv2 still outperforms the other models. However, the margin between CFSv2 in Fig. 4b is smaller than that in Fig. 4a. This is because for the five models other than CFSv2, lead time is zero in Fig. 4b while lead time is generally nonzero in Fig. 4a. Lead time degrades the forecast skill, so the skill of these models is lower in Fig. 4a. Note that different subsamples of CFSv2 (different dashed lines in Fig. 4b) also show some variations in ACC, especially for the red dashed line with 8–11 ensemble member. This shows that even for the same model, different selection of reforecast initialization dates can also result in differences in prediction skill even when the same reforecast period is considered. Note that this difference is of the same order of magnitude (~0.01) as the model-to-model difference in forecast skill.

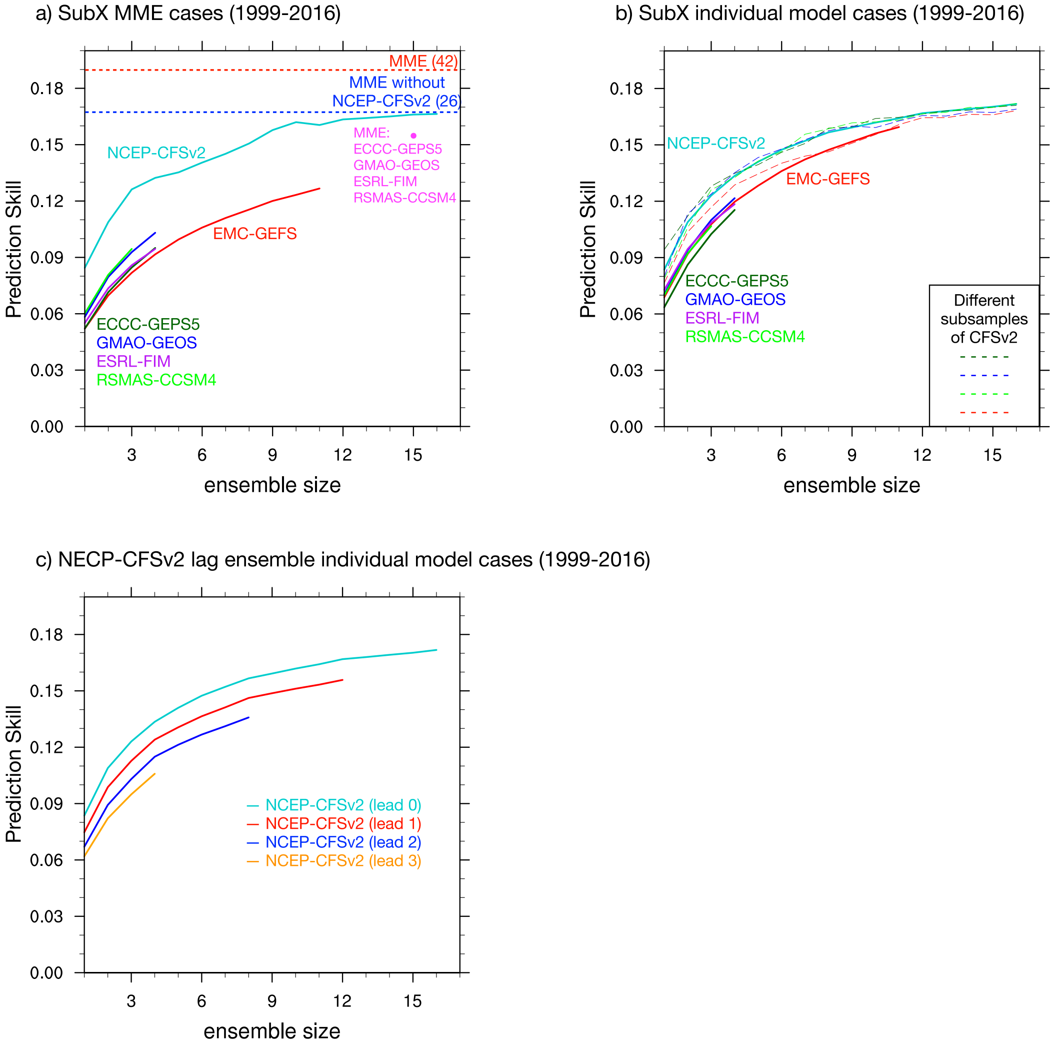


FIG. 4. (a) Area-averaged ACC over all the grid points north of 10°N. The solid lines show the averaged ACC as a function of ensemble size. Note that the mean of x axis is different for CFSv2 (see main text). The magenta dot shows the prediction skill ECCC-GEPS5, ESRL-FIM, GMAO-GEOS, and RSMAS-CCSM4 combined ensemble. The dashed red line and dashed blue line shows the averaged ACC of MME and MME without CFSv2, respectively. (b) The solid lines are as in (a), but using the all available cases of individual models with no lead time, instead of using MME cases. The dashed lines are the ACC of different subsamples of CFSv2. The dashed lines and solid lines in the same color have the same reforecast cases. Note the EMC-GEFS and ESRL-FIM have the same cases, and the corresponding CFSv2 subsample is only shown in the dashed red line. (c) As in (b), but for 16-member NCEP-CFSv2, with all members [cyan line; the same as the cyan line in (b)], with members has at least 1-day lead (red line), 2-days lead (blue line), and 3-days lead (orange line).

We also make use of the Heidke skill score (HSS) as an alternative measure to evaluate the prediction skill of the models (see Text S2). The conclusions we reach are very similar to those in Z19. In addition, as discussed in Z19, for the S2S models, the source of predictability of ECA is mainly from ENSO and the stratospheric polar vortex. We use similar methods and find that the source of predictability for the SubX models also mainly comes from these two phenomena (see Text S3–S5). These have been extensively discussed in Z19 and will not be the focus of this study. In short, the SubX models and S2S models have similar spatial pattern of prediction skill, while the source of predictability is also similar (mainly from ENSO and the stratospheric polar vortex. In the following section, we will focus on comparing the skill of the SubX and S2S models.

4. Comparing S2S and SubX models

a. *Differences in SubX and S2S reforecast data*

There are several differences in SubX and S2S data, which can potentially lead to differences in prediction skill. As discussed in section 2, the MSLP is daily mean in the SubX

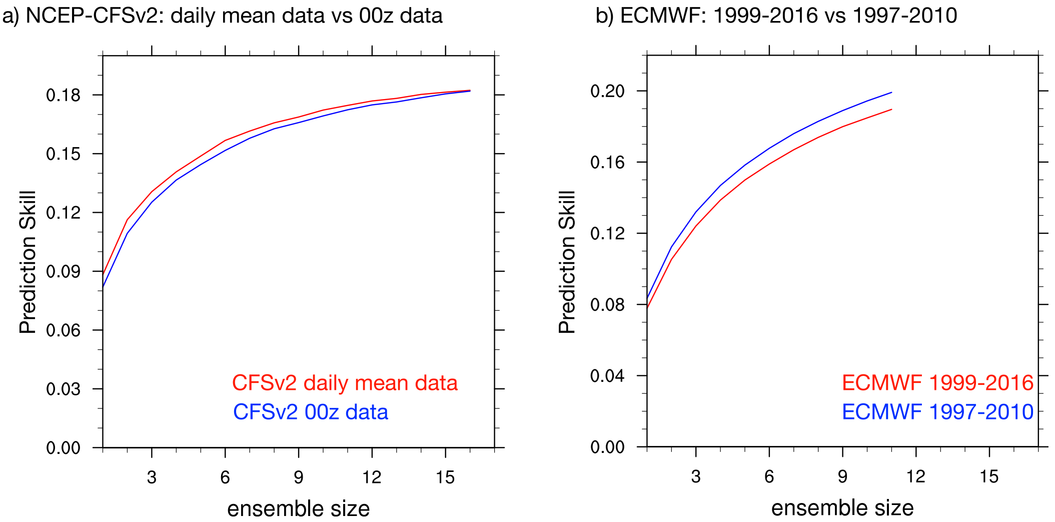


FIG. 5. (a) NH averaged (north of 10°N) CFSv2 week 3–4 ACC during 1999–2016 using all available CFSv2 cases (no lead time) with daily mean data (red line) and 0000 UTC data (blue line). The meaning of horizontal axis is the same as that in Fig. 4c. (b) NH averaged (north of 10°N) ECMWF week 3–4 ACC using all available ECMWF cases (no lead time) during 1999–2016 (red; SubX overlapping period) and 1997–2010 (blue; S2S overlapping period).

ensemble archive while 0000 UTC instantaneous MSLP is archived in the S2S ensemble. In Fig. 5a, we compare the ACC averaged over the NH (north of 10°N) by using daily mean MSLP (red line) and 0000 UTC MSLP (blue line) for all available CFSv2 cases (daily) during 1999–2016. The verification reanalysis data are ECA calculated from ERA-Interim daily mean MSLP and 0000 UTC MSLP, respectively. The prediction skill from ECA computed based on the different MSLP data are different, especially when only using the first few members of CFSv2. The skill attained by using daily mean data (red) is a bit higher, possibly because daily mean data are less noisy, as discussed in section 2. While this difference is not large (<0.01), it is still about the same order of magnitude as model to model difference in skill (Fig. 4b). The difference in data availability (0000 UTC versus daily mean MSLP) is one factor to be considered when comparing the SubX and S2S models.

In addition, the overlapping period of the S2S models (both in this study and in Z19) is 1997–2010, which is different from the SubX overlapping period (1999–2016). Model prediction skill during different time periods can be different, as shown in Fig. 5b. We use all cases with no lead time from ECMWF model during 1997–2010 (S2S overlapping period) and 1999–2016 (SubX overlapping period) to calculate the NH (north of 10°N) averaged prediction skill. The skill during 1997–2010 (blue) is higher than that during 1999–2016 (red). The spatial patterns of the week 3–4 prediction skill during 1999–2016 and 1997–2010 are shown in Figs. 6a and 6c, respectively. One of the reasons that may lead to the differences in model prediction skill during different time periods is that predictability of ECA during different time periods can be different. As discussed in Z19, ENSO is one of the major sources of predictability of week 3–4 ECA. We plot the absolute value of ACC between ENSO and reanalysis ECA during the two time periods (Figs. 6b,d) during the same reforecast cases in Figs. 6a and 6c. The value of ACC between ENSO and ECA indicates the potential predictability from ENSO. The spatial patterns in Figs. 6b and 6d are similar to the prediction skill in Figs. 6a and 6c. Note that models are able to capture the correlation between ENSO and ECA as shown by Z19. The correlation between ENSO and ECA is higher during 1997–2010 (Fig. 6d) than that during 1999–2016 (Fig. 6b). This can potentially be one of the reasons why prediction skill during 1997–2010 (Fig. 6c) is higher than that during 1999–2016 (Fig. 6a). We also plot the correlations between ENSO and ECA in individual EMC-GEFS members (Fig. S9), which have similar spatial patterns compared to the Reanalysis. Different models have different reforecast initialization time, which also leads to different prediction skill. These have been discussed in section 3. When using the dates in the MME forecast, different models have different lead time, and lead time will degrade the skill (Fig. 4c). When using no lead time to compare different models, then different initialization time of each model leads to different reforecast cases, resulting in differences in skill (Fig. 4b cyan line and dashed lines).

The three factors mentioned above, related to differences in data availability (0000 UTC versus daily mean MSLP), different reforecast time period, and different initialization time, can all lead to differences in ACC of the order of 0.01 averaged over NH, which is comparable to intermodel skill differences (e.g., Figs. 4a,b). Therefore, to fairly compare the skill of different models, one need to use either 0000 UTC or daily mean MSLP data for all the different models, with the same reforecast cases during the same reforecast time period. As CFSv2 is available daily with 6-hourly output, we can fairly compare the skill between CFSv2 and any other model.

b. *CFSv2 versus other models*

As discussed above, a fair way to compare the skill of different models is to use the same MSLP data (0000 UTC or daily

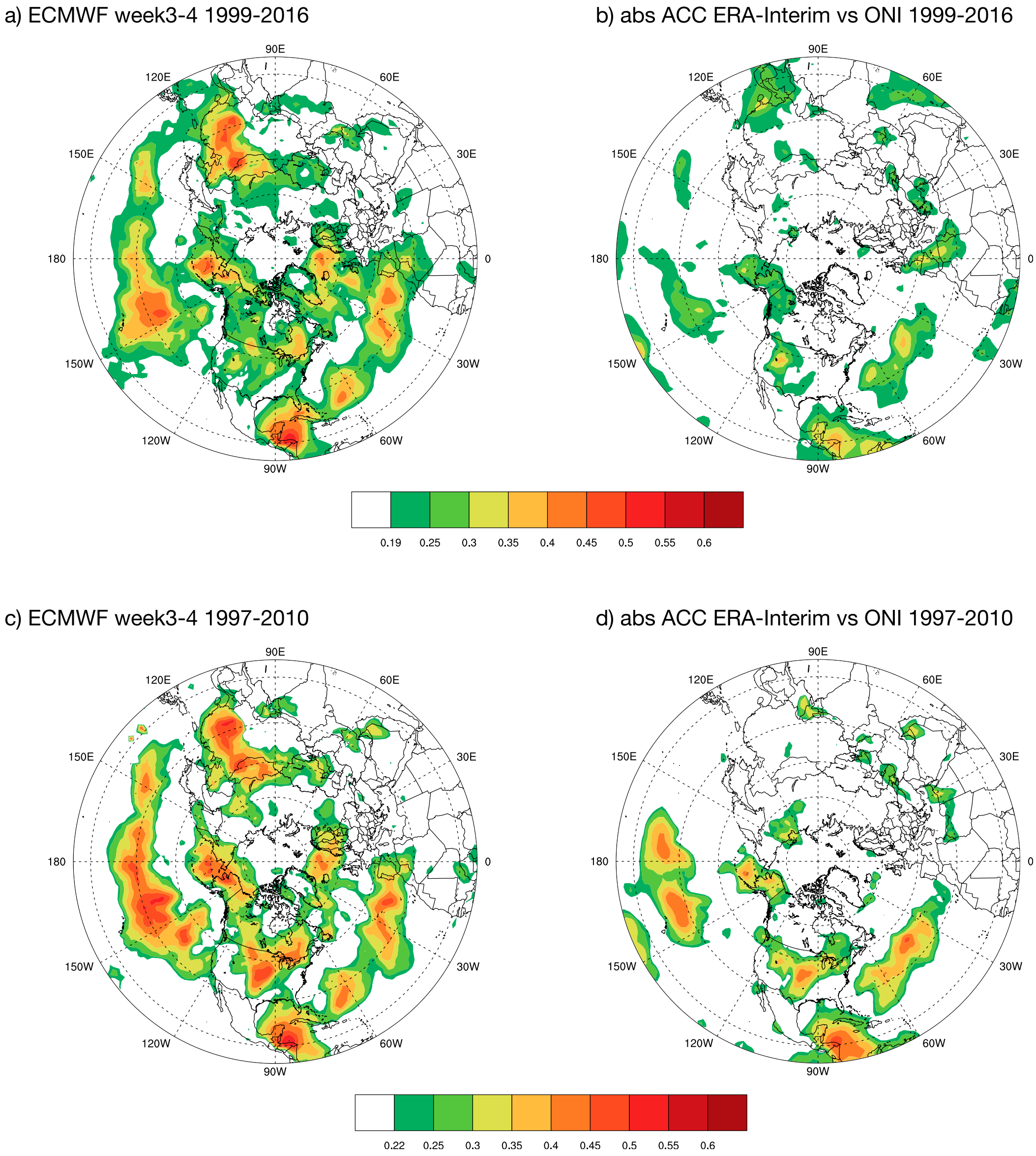

a) ECMWF week3-4 1999-2016
b) abs ACC ERA-Interim vs ONI 1999-2016
90E
120E
60E
150E
30E
180
0
150W
30W
120W
60W
90W
0.19
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
c) ECMWF week3-4 1997-2010
d) abs ACC ERA-Interim vs ONI 1997-2010
0.22
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6

FIG. 6. (a) As in Fig. 3, but for ECMWF prediction skill (no lead time) during 1999–2016. (b) Absolute value of ACC between ERA-Interim week 3–4 extratropical cyclone activity and winter season mean (December–February) Niño-3.4 index (also known as ONI) in same ECMWF cases in (a) during 1999–2016. (c) As in (a), but for 1997–2010. (d) As in (b), but for the ECMWF cases in (c) during 1997–2010.

mean) with the same initialization time during the same reforecast period. Similar to section 3 (Fig. 4b), for any model other than CFSv2, we can directly compare this model with CFSv2 by using a subsample of CFSv2 with the same reforecast cases as this model. For all the SubX and S2S models, this comparison is performed during the overlapping period of the SubX and S2S ensembles, which is DJF from 1999/2000 to 2009/10 (Fig. 7a for the S2S ensemble and Fig. 7b for the SubX ensemble). So, the comparison between any model and CFSv2, is during the same time period (1999–2010), with the same reforecast cases (reforecast initialization times) between this model and CFSv2 (no lead time). Similar to section 3a, the comparison between any model and CFSv2 can be done by comparing the dashed lines (CFSv2) and the solid lines (the other model) in the same color. For example, one can compare the ACC over the NH for ECMWF (solid green line in Fig. 7a) and CFSv2 (dashed green line in Fig. 7a). The spread of the dashed lines denotes the variability of CFSv2 ACC in different subsamples of CFSv2. This variability of ACC is smaller than the intermodel differences of ACC in the S2S ensembles (Fig. 7a). The skill of any model minus CFSv2 skill is shown in the subpanel in Fig. 7a, providing an alternative illustration of the comparison between any model and CFSv2.

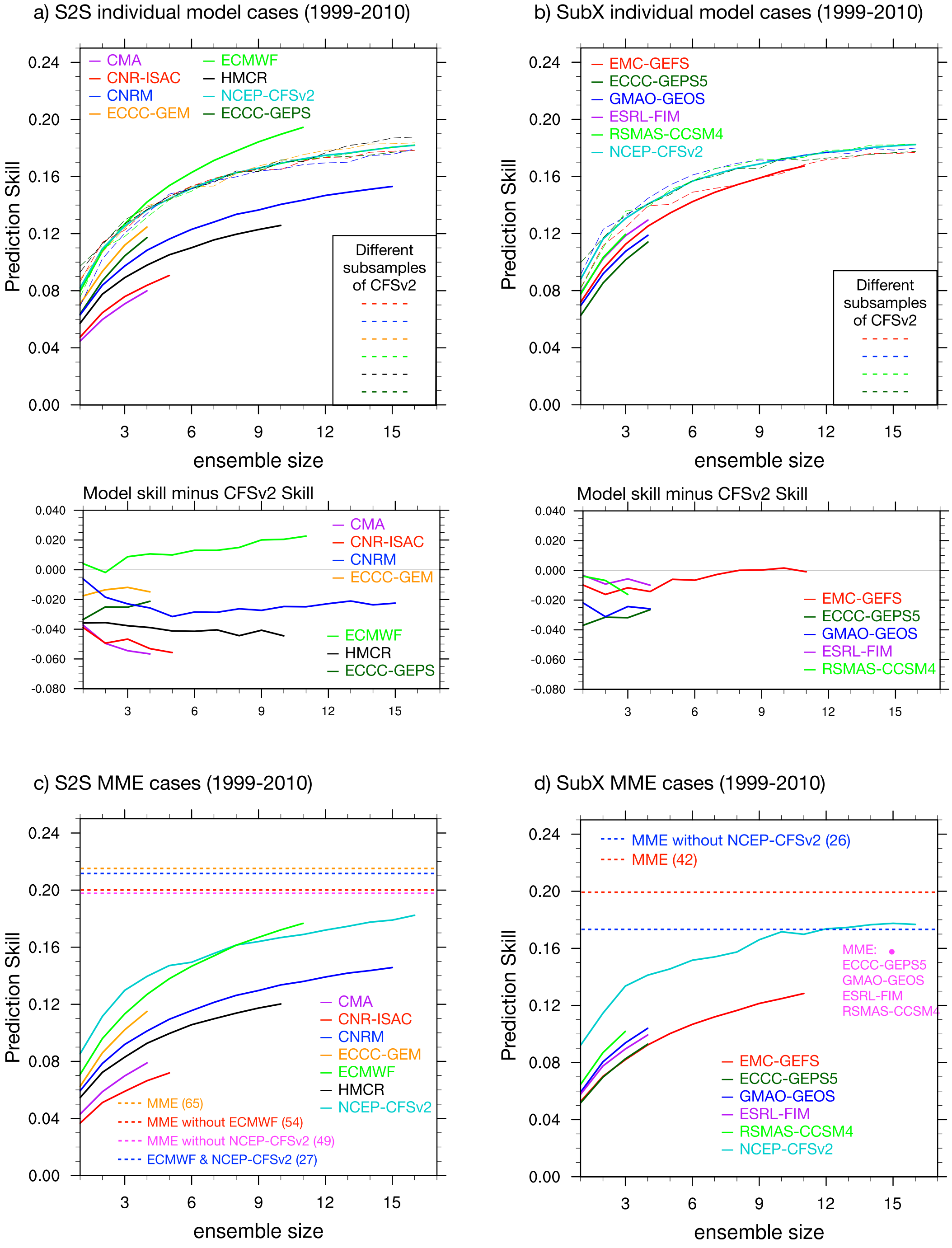


FIG. 7. (a) As in Fig. 4b, but for S2S models of individual S2S models cases in 1999–2010. The solid lines correspond to the prediction skill using the all available cases of individual models with no lead time. The dashed lines are the ACC of different subsamples of CFSv2. The dashed lines and solid lines in the same color have the same reforecast cases. The subpanel below shows the skill differences between any model and CFSv2. (b) As in Fig. 4b, but during 1999–2010 instead of 1999–2016. (c) As in Fig. 4a, but for the S2S ensemble during 1999–2010. The dashed magenta line, dashed red line, dashed blue line, and dashed orange line show the averaged ACC over NH of MME without NCEP-CFSv2, MME without ECMWF, ECMWF and NCEP-CFSv2 combined ensemble, and MME, respectively. (d) As in Fig. 4a, but for 1999–2010 instead of 1999–2016.

In the S2S ensemble (Fig. 7a), it is clear that ECMWF and CFSv2 outperform the other models. The ACC of the 11-member ECMWF with no lead time is higher than the 16-member CFSv2. However, this does not necessarily mean that ECMWF is a better model than CFSv2 in predicting the ECA. As different CFSv2 members have different lead times, when the ensemble size becomes larger, the averaged lead time of CFSv2 also becomes larger. Thus, it is not exactly a fair comparison again between CFSv2 and ECMWF. In addition, the ACC of the first three members of CFSv2 (solid cyan line in Fig. 7a) is almost the same as that for ECMWF with the three ensemble members. So, it is not clear that, if ECMWF and CFSv2 reforecasts are initialized with the same frequency with the same number of ensemble members, which model has better prediction of ECA. Nevertheless, in practice, given that we can only form lagged ensembles using CFSv2, ECMWF ensemble with all members having zero lag is expected to outperform lagged ensembles formed using CFSv2 forecasts. This suggests an advantage of initializing multiple forecast members at the same time. For the SubX ensemble (Fig. 7b), similar to Fig. 4b, CFSv2 outperforms all the other SubX models. The skill between EMC-GEFS and CFSv2 is close with about 8–11 members. Similarly, this may be due to that CFSv2 has longer lead time when the ensemble size is larger, and EMC-GEFS takes the advantage of initializing multiple forecast members at the same time. Note that intermodel spread in skill is smaller among SubX models compared to S2S models.

c. The S2S MME and the SubX MME

As discussed in section 2b, the ECA of the SubX and S2S models cannot be combined into a ‘‘grand’’ ensemble. Here we try to compare the MME skill of the S2S and SubX during their overlapping time period in Figs. 7c and 7d. Note that, any comparisons among the models within Fig. 7c or Fig. 7d, or comparisons of the MME skill across Figs. 7c and 7d, are not completely ‘‘fair.’’ Models have different lead time when they are combined into the MME. Also, SubX MME and S2S MME have different reforecast cases during 1999–2010. Here, the focus is more on the relative contribution of different models to the MME skill. In addition to the S2S MME used in Z19, CFSv2 is also included in the S2S MME in our analysis here. Also note that ECCC-GEPS5, which is in both SubX and S2S datasets, is not combined into the S2S MME (see Text S6). The spatial patterns of ACC of the S2S models are shown in Fig. S11, which are very similar to those shown in Z19. For the S2S ensemble (Fig. 7c), ECMWF and CFSv2 outperform the other S2S models. The ensemble mean ACC of CFSv2 is now slightly better than the ensemble mean of ECMWF, which is different from Fig. 7a. This is likely due to the fact that there is a larger average lead time for ECMWF when combining the models into the MME (also see discussions above). The averaged ACC over NH for the MME without CFSv2 (magenta dashed line in Fig. 7c) is very close to that for the MME without ECMWF (red dashed line in Fig. 7c). It is not surprising that the full S2S ensemble (orange dashed line in Fig. 7c) has the best prediction skill. Note that if we only combine the two best models, which are ECMWF and CFSv2, the prediction skill over NH of this 27-member ensemble (blue dashed line) is comparable to that of the full S2S ensemble. This shows that ECMWF and CFSv2 are probably the key contributors to the ACC in the S2S MME.

In the SubX MME (Fig. 7d), as discussed in section 3a, CFSv2 outperforms the other SubX models. The prediction skill of the full S2S MME (65 members; orange dashed line in Fig. 7c) is better than the full SubX MME (42 members; red dashed line in Fig. 7d). The comparison across Figs. 7c and 7d is not entirely fair, as the reforecast initialization times in the MME cases are different and the data used is different (daily mean for SubX versus 0000 UTC instantaneous for S2S). However, the ACC variability due to different case selections (e.g., spread of dashed lines in Figs. 7a,b) and due to different MSLP data (Fig. 5a) probably will not change the conclusion that S2S MME outperforms SubX MME. If we exclude the ECMWF model from the S2S MME (red dashed line in Fig. 7c), the prediction skill is then almost the same as the SubX MME (red dashed line in Fig. 7d). This suggests that S2S MME outperforms SubX MME because it benefits from the relatively skillful ECMWF model.

5. Conclusions and discussion

In this study, we evaluate the prediction skill of extratropical cyclone activity (ECA) on subseasonal time scales by models that participated in the Subseasonal Experiment (SubX) and those used in the Seasonal to Subseasonal Prediction (S2S) project. Consistent with Z19, the regions where the anomaly correlation coefficient (ACC) is high are over east Asia, the central and eastern North Pacific, central North America, the Gulf of Mexico and western Caribbean Sea, the central North Atlantic, as well as Scandinavia and the Norwegian Sea. CFSv2 has the best prediction skill among the SubX models. The other models have relatively similar prediction skill when the ensemble size is the same, and EMC-GEFS has relatively higher prediction skill as it has a relatively larger ensemble size. Consistent with previous studies, we find that combining different CFSv2 members into a lagged ensemble improves the forecast skill, with forecasts up to 4 days old still marginally improving the skill of the lagged ensemble for predicting weeks 3–4 ECA. Thus a 16-member CFSv2 lagged ensemble is used in this study.

The SubX and S2S models have different configurations. The forecast initialization time is different, and the SubX archives daily mean MSLP while S2S archives instantaneous MSLP. The large differences in variability of ECA because of the different archived data make it inappropriate to combine the SubX models and S2S models into a ‘‘grand’’ ensemble. As different forecast initialization times, different MSLP data availability (0000 UTC versus daily mean), and different reforecast time periods can all lead to different model prediction skill, it is not straightforward to compare the ECA prediction skill between the S2S and SubX models. As CFSv2 has frequent forecast and reforecast initialized every 6 h, with 6-hourly output available, we are able to compare the ECA prediction skill between any model and CFSv2 in a fair way by using a subsample of CFSv2. This allows the CFSv2 to be used as a baseline to evaluate the skill of the various models.

The setup of the CFSv2 reforecast, which is different from many other models, makes CFSv2 the best choice to be constructed into an ensemble that is ''initialized'' daily. Note that here we are emphasizing CFSv2 because we want to develop a fair way to compare the skill of different models. Our methods and findings should not be treated as implications about whether it is better to initialize reforecast every 6 h with one model member (as CFSv2 does), or to initialize reforecast with an ensemble of members with longer intervals (a few days) between reforecast initialization times (the way of most other models).

In terms of single model performance, ECMWF and CFSv2 are the two best models that are examined in the study. The prediction skill of the S2S ensemble is better than the SubX ensemble when averaged over the Northern Hemisphere, which is probably because ECMWF is included in the S2S ensemble but not in the SubX ensemble. It should be noted that as ECMWF has 51 members in the operational forecast and only 11 members in the reforecast. Although some of the SubX models also have more members in real time forecast than in reforecast, the increase of the ensemble size is not as large as the increase for ECMWF. These suggest that the S2S ensemble could provide higher prediction skill in real time forecasts.

As discussed in section 4, combining the two best models (ECMWF and CFSv2) in S2S provides prediction skill comparable with the entire S2S ensemble. Practically, only combining two models is much easier than constructing the MME for the entire S2S ensemble. Note that if we only use ECMWF and CFSv2, as CFSv2 ensemble is available every day, one can just use the days when ECMWF initializes a forecast to construct this two-model ensemble. As there will be no lead time for ECMWF, the prediction skill can be even higher. However, both ECMWF and CFSv2 have different real-time forecast configurations compared to their reforecast configurations. CFSv2 has 16 members per day in the real time forecast (4 members in the reforecast). As ECMWF has 51 members in real time forecast (11 members in reforecast), the ECMWF ensemble probably will still outperform the CFSv2 ensemble in the real time forecast. An interesting question that can be explored in the future study is: How skillful would a real-time version of this two-model ensemble be for analyzing and predicting ECA?

MSLP, which is used to calculate ECA, is archived differently in S2S (instantaneously at 0000 UTC) and SubX (daily average) models. In this study, we develop a method to compare the model prediction skill of ECA by using CFSv2. Our results suggest that the prediction skill of ECA is somewhat sensitive to whether the MSLP data are archived daily or instantaneously, with daily mean data resulting in smoother ECA fields that may be predicted with slightly higher skill. Similar to MSLP, variables like geopotential height and wind field are also archived differently in S2S and SubX. The method developed in this study could be useful in comparing prediction skill of these variables in the future between models with different setups.

Acknowledgments. The authors thank three anonymous reviewers for comments that help to improve this paper. This research has been conducted as part of the NOAA MAPP S2S Prediction Task Force and supported by NOAA Grant NA16OAR4310070 while the first author was at Stony Brook University. The ERA-Interim reanalysis data, SubX data, and S2S data are cited in the reference list. The Niño-3.4 index is available from the NOAA ESRL website (https://www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries/Nino34/).

REFERENCES

Becker, E., H. van den Dool, and Q. Zhang, 2014: Predictability and forecast skill in NMME. *J. Climate*, **27**, 5891–5906, https://doi.org/10.1175/JCLI-D-13-00597.1.

Befort, D. J., and Coauthors, 2019: Seasonal forecast skill for extratropical cyclones and windstorms. *Quart. J. Roy. Meteor. Soc.*, **145**, 92–108, https://doi.org/10.1002/qj.3406.

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

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

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

——, Y. Guo, X. Xia, and M. Zheng, 2013: Storm-track activity in IPCC AR4/CMIP3 model simulations. *J. Climate*, **26**, 246–260, https://doi.org/10.1175/JCLI-D-11-00707.1.

——, C. Zheng, P. Lanigan, A. M. W. Yau, and J. D. Neelin, 2015: Significant modulation of variability and projected change in California winter precipitation by extratropical cyclone activity. *Geophys. Res. Lett.*, **42**, 5983–5991, https://doi.org/10.1002/2015GL064424.

Chen, M., W. Wang, and A. Kumar, 2010: Prediction of monthly mean temperature: The roles of atmospheric and land initial conditions and sea surface temperature. *J. Climate*, **23**, 717–725, https://doi.org/10.1175/2009JCLI3090.1.

——, ——, and ——, 2013: Lagged ensembles, forecast configuration, and seasonal predictions. *Mon. Wea. Rev.*, **141**, 3477–3497, https://doi.org/10.1175/MWR-D-12-00184.1.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

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

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

Froude, L. S. R., 2010: TIGGE: Comparison of the prediction of Northern Hemisphere extratropical cyclones by different ensemble prediction systems. *Wea. Forecasting*, **25**, 819–836, https://doi.org/10.1175/2010WAF2222326.1.

——, L. Bengtsson, and K. I. Hodges, 2007a: The predictability of extratropical storm tracks and the sensitivity of their prediction to the observing system. *Mon. Wea. Rev.*, **135**, 315–333, https://doi.org/10.1175/MWR3274.1.

——, ——, and ——, 2007b: The prediction of extratropical storm tracks by the ECMWF and NCEP ensemble prediction systems. *Mon. Wea. Rev.*, **135**, 2545–2567, https://doi.org/10.1175/MWR3422.1.

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

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting–I. Basic concept. *Tellus*, **57A**, 219–233, https://doi.org/10.3402/tellusa.v57i3.14657.

Haynes, P. H., M. E. McIntyre, T. G. Shepherd, C. J. Marks, and K. P. Shine, 1991: On the ''downward control'' of extratropical diabatic circulations by eddy-induced mean zonal forces. *J. Atmos. Sci.*, **48**, 651–678, https://doi.org/10.1175/1520-0469(1991)048<0651:OTCOED>2.0.CO;2.

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

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

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

Lee, Y. Y., and G. H. Lim, 2012: Dependency of the North Pacific winter storm tracks on the zonal distribution of MJO convection. *J. Geophys. Res.*, **117**, D14101, https://doi.org/10.1029/2011JD016417.

Lukens, K. E., and E. H. Berbery, 2019: Winter storm tracks and related weather in the NCEP climate forecast system weeks 3–4 reforecasts for North America. *Wea. Forecasting*, **34**, 751–772, https://doi.org/10.1175/WAF-D-18-0113.1.

Ma, C., and E. K. Chang, 2017: Impacts of storm-track variations on wintertime extreme weather events over the Continental United States. *J. Climate*, **30**, 4601–4624, https://doi.org/10.1175/JCLI-D-16-0560.1.

Monhart, S., C. Spirig, J. Bhend, K. Bogner, C. Schär, and M. A. Liniger, 2018: Skill of subseasonal forecasts in Europe: Effect of bias correction and downscaling using surface observations. *J. Geophys. Res. Atmos.*, **123**, 7999–8016, https://doi.org/10.1029/2017JD027923.

Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment. *Bull. Amer. Meteor. Soc.*, **100**, 2043–2060, https://doi.org/10.1175/BAMS-D-18-0270.1.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.*, **108**, 4407, https://doi.org/10.1029/2002JD002670.

Riddle, E. E., A. H. Butler, J. C. Furtado, J. L. Cohen, and A. Kumar, 2013: CFSv2 ensemble prediction of the wintertime Arctic Oscillation. *Climate Dyn.*, **41**, 1099–1116, https://doi.org/10.1007/s00382-013-1850-5.

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

Scaife, A. A., and Coauthors, 2012: Climate change projections and stratosphere–troposphere interaction. *Climate Dyn.*, **38**, 2089–2097, https://doi.org/10.1007/s00382-011-1080-7.

Shaw, T. A., and Coauthors, 2016: Storm track processes and the opposing influences of climate change. *Nat. Geosci.*, **9**, 656–664, https://doi.org/10.1038/ngeo2783.

Smith, D. M., and Coauthors, 2013: Real-time multi-model decadal climate predictions. *Climate Dyn.*, **41**, 2875–2888, https://doi.org/10.1007/s00382-012-1600-0.

Stockdale, T. N., and Coauthors, 2010: Understanding and predicting seasonal-to-interannual climate variability—The producer perspective. *Proc. Environ. Sci.*, **1**, 55–80, https://doi.org/10.1016/j.proenv.2010.09.006.

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

Vitart, F., and Coauthors, 2017: The Subseasonal to Seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

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

Walter, K., and H. F. Graf, 2005: The North Atlantic variability structure, storm tracks, and precipitation depending on the polar vortex strength. *Atmos. Chem. Phys.*, **5**, 239–248, https://doi.org/10.5194/acp-5-239-2005.

Wang, J., H. M. Kim, and E. K. Chang, 2018a: Interannual modulation of Northern Hemisphere winter storm tracks by the QBO. *Geophys. Res. Lett.*, **45**, 2786–2794, https://doi.org/10.1002/2017GL076929.

——, ——, ——, and S. W. Son, 2018b: Modulation of the MJO and North Pacific storm track relationship by the QBO. *J. Geophys. Res. Atmos.*, **123**, 3976–3992, https://doi.org/10.1029/2017JD027977.

Yang, X., and Coauthors, 2015: Seasonal predictability of extratropical storm tracks in GFDL's high-resolution climate prediction model. *J. Climate*, **28**, 3592–3611, https://doi.org/10.1175/JCLI-D-14-00517.1.

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

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

——, ——, ——, ——, and ——, 2019: Subseasonal to seasonal prediction of wintertime Northern Hemisphere extratropical cyclone activity by S2S and NMME models. *J. Geophys. Res. Atmos.*, **124**, 12 057–12 077, https://doi.org/10.1029/2019JD031252.

Zhu, J., B. Huang, M. A. Balmaseda, J. L. Kinter III, P. Peng, Z.-Z. Hu, and L. Marx, 2013: Improved reliability of ENSO hindcasts with multi-ocean analyses ensemble initialization. *Climate Dyn.*, **41**, 2785–2795, https://doi.org/10.1007/s00382-013-1965-8.