Recent Challenges in the APCC Multi-Model Ensemble Seasonal Prediction: Hindcast Period Issue

Young-Mi Min[1], Chang-Mook Lim[2], Vladimir Kryjov[3], and Daeun Jeong[1]

[1]APEC Climate Center (APCC)
[2]APEC Climate Center
[3]Pusan National University

March 17, 2024

Recent Challenges in the APCC Multi-Model Ensemble Seasonal Prediction: Hindcast Period Issue

Young-Mi Min[1], Chang-Mook Im[1], Vladimir N. Kryjov[2], and Daeun Jeong[1]

[1]APEC Climate Center, Busan, Republic of Korea

[2]Busan National University, Busan, Republic of Korea

Corresponding author: Young-Mi Min (ymmin@apcc21.org)

Key Points:

- APCC, which combines all the information from different ensemble prediction systems, recently faced challenges in hindcast period issues
- The proposed solution leads to an increase in the number of models contributing to MME prediction, particularly recently developed models
- It shows improved skills for both temperature and precipitation predictions over most of the globe and seasons

Abstract

Seasonal forecasts are commonly issued in the form of anomalies, which are departures from the average over a specified multiyear reference period (climatology). The model climatology is estimated as the average of the retrospective forecasts over the hindcast period. However, different operational centers that provide seasonal ensemble predictions use different hindcast periods based on their model climatology. Additionally, the hindcast periods of recently developed and upgraded newer models have shifted in the recent years. In this paper, we discuss the recent challenges faced by APCC multi-model ensemble (MME) operations, especially changes in the hindcast period for individual models. Based on the results of various experiments for MME prediction, we propose changing the hindcast period, which is the most appropriate solution for APCC operation. This makes the newly developed models join the MME and increases the total number of participating models, which facilitates the skill improvement of the MME prediction.

Plain Language Summary

In seasonal forecasting, it is well known that the MME, which combines different single-model predictions from various operational and research centers, is a more effective way to improve forecast skill. Since 2005, the APCC has provided the MME seasonal forecasts, and the models participating in the APCC MME operations have been continuously changing. In particular, as the hindcast periods of newly developed models shift to the latest, they cannot participate in operational MME forecasts because of climatological discrepancies. However, over time, as the number of new models expected to provide skillful forecasts gradually increases, the APCC faces the challenge of continuously reducing the number of participating models or changing the hindcast period to more recent years. Considering various aspects such as the number of participating models, skills, and climatology period, we selected the most appropriate method for APCC operation. Thus, the MME prediction skill has improved over most of the globe and seasons because of the increase in the number of participating models, particularly the inclusion of newer models.

1 Introduction

Seasonal forecasts are commonly expressed in terms of anomalies, as departures from the climatological mean and/or probabilities of an event occurring with respect to a climatological distribution (usually, tercile-based categorical forecasts). This allows users to see whether the predicted seasonal mean variables are anomalously positive or negative with respect to climatological means, and/or what probability of the events (e.g., above, near, or below-normal category) is expected. Therefore, climatology is used as a benchmark or reference against which the expected conditions are likely to be experienced. It also provides a way to remove systematic biases in forecasts from dynamical prediction systems by subtracting model climatology, because they are not perfect representations of the real world (Stockdale, 1997; Kumar et al., 2012). The model climatology is estimated using retrospective forecasts (hindcasts) over a specified long-term reference period.

World Meteorological Organization (WMO) recommends climatology (normals) to be estimated as 30-year averages computed for the most-recent 30-year period finishing in a year ending with 0 (WMO, 2007), i.e., 1991-2020 at present. National Meteorological and Hydrological Services (NMHSs) estimate forecasts as departures from these 30-year normals in their locations. However, different operational and research centers have different hindcast periods resulting in the use of different climatology periods for model climatology. Furthermore, the hindcast periods of recently developed and improved climate models, particularly beginning of the hindcast period, tend to shift to recent years. The Asia-Pacific Economic Cooperation (APEC) Climate Center (APCC) is one of the major operational centers providing well-validated multi-model ensemble (MME) seasonal forecasts. Since its establishment in 2005, APCC has collected dynamical ensemble forecasts through multi-institutional cooperation and coordinated MME predictions. At present, 15 leading operational and research institutes from 11 countries are involved in APCC operational MME prediction. MME operational centers, such as APCC (Min et al., 2014, 2017), WMO Lead Center for Long-Range Forecast (WMO LC-LRF; Kim et al., 2021), North American MME (NMME; Becker et al., 2014; Kirtman et al., 2014), and Copernicus Climate Change Service (C3S; Manazanas et al., 2019) use a common hindcast period for all participating models, which results in a relatively short period compared to that of single-model prediction systems. For example, APCC used the hindcast period in the early 20 years covering from early-1980s to the mid-2000s and extended it to 28 years in 2019, 1983-2010.

As the hindcast periods for recently developed newer models have gradually shifted to later years, the full range of hindcast periods for the dynamical models routinely running in operational centers has widened, from early-1980s to late-2010s nowadays. However, the common hindcast period is rather short because of shift in the newer models' hindcast periods beginning in the early 1990s. This raised a new issue at APCC, which combines all the information from different climate prediction systems, particularly in 2019. This is because some of the models included in the operational APCC MME prediction were expected to change to their upgraded versions in 2020, and their hindcast periods shifted to more recent years. That is, with the implementation of new models, if the common hindcast period, 1983-2010, were maintained, the number of participating models in the MME would have been reduced and would be gradually reduced in the future because recently developed models that are expected to have better skills do not match this common hindcast period. This may lead to deterioration of the MME prediction skill. Therefore, APCC has come to consider the issue of the hindcast period, which could affect the number of participating models in the MME and eventually the MME skill. This study discusses the challenges faced by MME operations caused by upgrading participating models. In particular, we focus on the decrease in the number of participating models in MME prediction with a shift to the later years of the hindcast periods of recently developed models. We suggest the most appropriate solution for the APCC operation based on several experiments with the different hindcast periods and different numbers of participating models in the MME.

2 Data and Method

2.1 Forecast data

With the most recent joining of System 8 from Met France (METFR; http://www.umr-cnrm.fr/IMG/pdf/system8-technical.pdf), APCC currently collects ensemble predictions from 15 state-of-the-art climate models, and the models are being continuously improved with great efforts from their own operational and research centers. The MME prediction system largely depends on operational changes for the modeling centers, and the participating models in the MME operation for each year and season differ slightly depending on the operational situation at that time. The collected models through the APCC multi-institutional cooperation for research and operation purposes in 2019 and 2020 are listed in Table 1. In 2019, the operational MME prediction comprised eight models from APCC (SCoPS; Ham et al., 2019), BOM (POAMA; Cottrill et al., 2013), CWB (GFST119; Paek et al., 2015), JMA (MRI-CPS2; Takaya et al., 2018), MSC/ECCC (CanSIP; Merryfield et al., 2013), NASA (GEOS-S2S-2; Molod et al., 2015), NCEP (CFSv2; Saha et al., 2014), and PNU (CGCMv1.0; Ahn & Kim, 2013) that matched with the common hindcast period of 1983-2010. The remaining six models could not be included in the MME because of different hindcast periods, although some were recently upgraded, for example, KMA (GloSea5GC2; Ham et al., 2019) and UKMO (GloSea5; MachLachlan et al., 2015). Furthermore, several models were scheduled to be changed to their upgraded versions in 2020 (e.g., POAMA to ACCESS-S (Hudson et al., 2017) in BOM, SPSv2 to SPSv3 (Sanna et al., 2017) in CMCC, and CanSIP to CanSIPv2 (Lin et al., 2020) in MSC/ECCC). To test sensitivity in terms of predictability as the participating models in MME change due to their improvements, we performed several experiments with varying reference periods and participating models in the MME, where the MME forecast is a simple average of individual models with equal weights.

2.2 Verification data and Metrics

We focus on 1-month lead 3-month mean (seasonal) MME forecasts of 2m temperature and precipitation over the globe (GL; 90°S-90°N) and sub-regions: Northern Extratropics (NE; 20°N-90°N), Southern Extratropics (SE; 20°S-90°S), Tropics (TR; 20°N-20°S), East Asia (EAs; 75°E-150°E, 15°N-60°N), South Asia (SAs; 60°E-140°E, 10°S-35°N), North America (NAm; 190°E-310°E, 10°N-75°N), South America (SAm; 270E-330E, 60°S-10°N), Australia (Aus; 110E-180E, 50°S-0°N), and Northern Eurasia (NEu; 25°E-190°E, 40°N-80°N). For skill assessment, we use the National Center for Environmental Prediction (NCEP)-Department of Energy (DOE) Reanalysis 2 data (Kanamitsu et al., 2002) for temperature and the Climate Anomaly System and Outgoing Longwave Radiation Prediction Index data (CAMS-OPI, Janowiak & Xie, 1999) for precipitation. For Nino 3.4 index, we use the optimum interpolation

(OI) version 2 monthly mean SST (Reynolds et al. 2002), obtained from the Climate Diagnostics Center of National Oceanic and Atmospheric Administration.

All model forecasts and observations were interpolated onto a 2.5 x 2.5 common grid. We used the anomaly pattern correlation coefficient (ACC) and temporal correlation coefficient (TCC) to assess the prediction skill. We used the ACC-based relative skill difference to assess the prediction skill improvement and deterioration of the MME forecasts with another model set compared to the reference model set. The statistical robustness of the skill difference was verified using a bootstrap resampling method with 500 Montel-Carlo simulations. This method involves estimating the distribution of a statistic by randomly resampling and using it to evaluate statistical significane (Wilks, 1995, 1997; Stephenson and Doblas-Reyes, 2000; Min et al. 2017). Student's t-test and the Mann-Kendall test (Mann, 1945; Kendall, 1975) were used to assess the statistical significance of the difference between means and trends of observations and predictions. All forecast data from individual models are expressed in the form of anomalies as departures from the model climatology. As verification data, we used observed anomalies to represent deviations from the observed climatology. Consequently, model bias does not affect forecast skill. However, the use of anomalies, which implies bias correction, enhances the role of the correct estimation of model and observed climatologies.

Table 1. Collected models through APCC multi-institutional cooperation in 2019 and 2020

	2019		2020	
Institute	Model	Hindcast Period	Model	Hindcast Period
APCC	**SCoPS**	1982-2013	SCoPS	1982-2013
BCC	CSM_1.1m	1991-2015	CSM_1.1m	1991-2015
BOM	**POAMA**	1983-2011	ACCESS-S	1990-2012
CMCC	SPSv2	1993-2016	SPSv3	1993-2016
CWB	**GFST119**	1982-2011	GFST119	1982-2011
HMC	SL-AV	1985-2010	SL-AV	1985-2010
JMA	**MRI-CPS2**	1979-2014	MRI-CPS2	1979-2014
KMA	GloSea5GC2	1991-2010	GloSea5GC2	1991-2016
MGO	MGOAM-2	1979-2004	MGOAM-2	1979-2004
MSC/ECCC	**CanSIP**	1981-2010	CanSIPv2	1981-2010
NASA	**GEOS-S2S-2**	1981-2016	GEOS-S2S-2	1981-2016
NCEP	**CFSv2**	1982-2010	CFSv2	1982-2010

PNU	**CGCMv1.0**	1980-2018	CGCMv1.0	1980-2019
UKMO	GloSea5	1993-2016	GloSea5	1993-2016

The bold text in 2019 indicates the models that participated in the operational APCC MME prediction based on 1983-2010 climatology.

3 Results

More than two decades have passed since dynamical prediction systems have been operationally exploited for seasonal forecasting. Operational long-range forecasting centers make essential efforts to improve climate prediction systems. In particular, they tend to extend the period of hindcasts over which climatology is estimated and move it to more recent years. As shown in Fig. 1, the number of models providing ensemble forecasts to APCC and the number of models participating in the operational MME prediction vary from year to year, depending on the operational situations at the time. The proportion of models not included as part of the operational MME prediction has been gradually increasing and was expected to increase to nearly 50% by 2020 (red line in Fig. 1). Recently, the reason why some of the models could not participate in the MME has been mainly due to inconsistencies with the common hindcast period, and the proportion of these models has gradually increased over time (black line in Fig. 1). In other words, model developers continue to improve their model by gradually shifting their hindcast periods to more recent years. However, if the current common hindcast period for the APCC MME does not change, the number of models participating in APCC MME operation will gradually decrease. A more important issue is the MME skill, which is affected by the mean skill of individual models and models' diversity (Yoo & Kang, 2005; Alessandri et al., 2018). If the number of participating models in the MME prediction continues to decrease, particularly by excluding recently developed and improved newer models, it may lead to a decrease in MME skill.

When faced with this issue in 2019, APCC examined changes in MME skills if the common hindcast period was maintained, considering expected model changes scheduled for 2020. As shown in Table 1, under the condition of the current 28-year hindcast period, the BOM's new model with a recent hindcast period (1990-2012), ACCESS-S, was expected to be unable to participate in the MME operation in 2020, and in the case of MSC/ECCC, CanSIP was scheduled to be upgraded to CanSIPv2 with the 1981-2010 hindcast period. Therefore, it was expected that CanSIPv2 would continue to participate in MME operations. To examine differences in MME skill due to model changes, we compared the expected MME hindcast skill with seven models in the 2020 version, considering BOM's and MSC/ECCC's model changes (experiment), to the MME hindcast skill with eight models in the 2019 version (reference: APCC, BOM, CWB, JMA, MSC/ECCC, NASA, NCEP, and PNU) for the common 28-year hindcast period (1983-2010). We were able to perform the hindcasts of the new models scheduled to be changed in 2020 because APCC collects a new version of the hindcast before the newer model is applied to the MME operation and prepares various aspects from an operational perspective. Fig. 2 shows the relative skill difference of the experimental MME hindcast compared with that of the reference MME hindcast. The ACC-based relative skill difference (%) was estimated as the difference between the ACCs of the experimental and reference forecasts, divided by the ACC of the reference forecasts. The relative skill difference is mainly negative, which indicates a deterioration in the MME skill caused by the expected models' changes for 2020. The skill of experimental forecasts for both global temperature and precipitation decreased across almost all seasons. This is also true for the sub-regions in terms of 12-season averages (annual means), with the exception of temperature in South America. That is, it was clearly expected that if the 28-year hindcast period was maintained in 2020, the MME prediction skill would ultimately decrease owing to a decrease in the number of participating models (from eight to seven), despite the MSC ECCC's model being replaced by CanSIPv2, which has a higher prediction skill than its previous version, CanSIP (Fig. 3). These results served as the motivation for the various considerations and experiments in this study to increase the number of participating models and consequently improve the MME prediction skill.

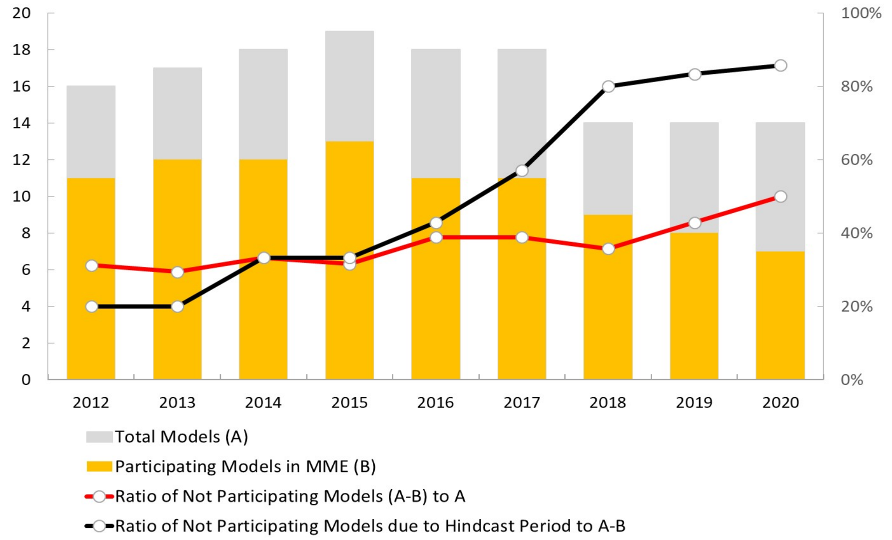

Figure 1. Changes in the number of models providing their seasonal forecasts to APCC (grey bar; A) and the number of models participating in the operational APCC MME prediction (yellow bar; B) in 2012-2020. Red lines indicate the proportion of models not participating in the operational MME prediction to the total models ((A-B)/A). Black lines represent the proportion of models not participating in MME due to inconsistency of common hindcast period to not participating models in MME. The values for 2020 refer to the expected changes if the 28-year (1983-2010) hindcast period for MME prediction continues in 2020.

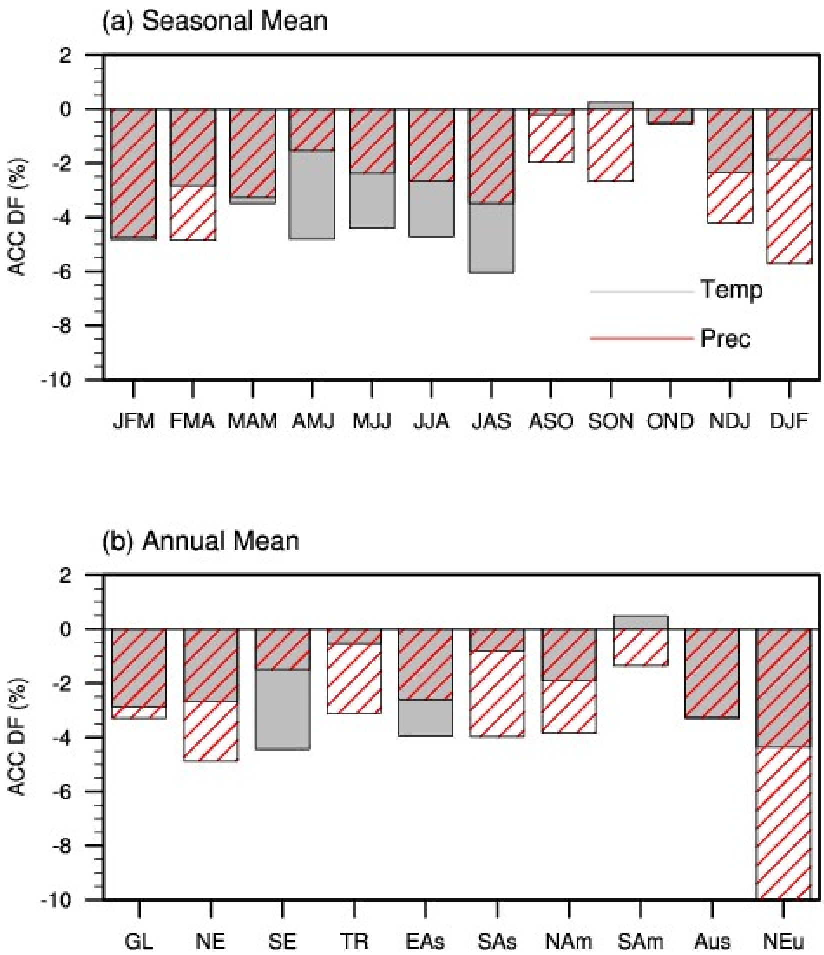

Figure 2. (a) Relative skill difference of the experimental MME hindcasts in 2020 to the reference MME hindcasts in 2019 of 3-month (seasonal) mean temperature and precipitation forecasts over the globe and (b) 12-season averaged (annual mean) forecasts for several sub-regions for the common period of 1983-2010.

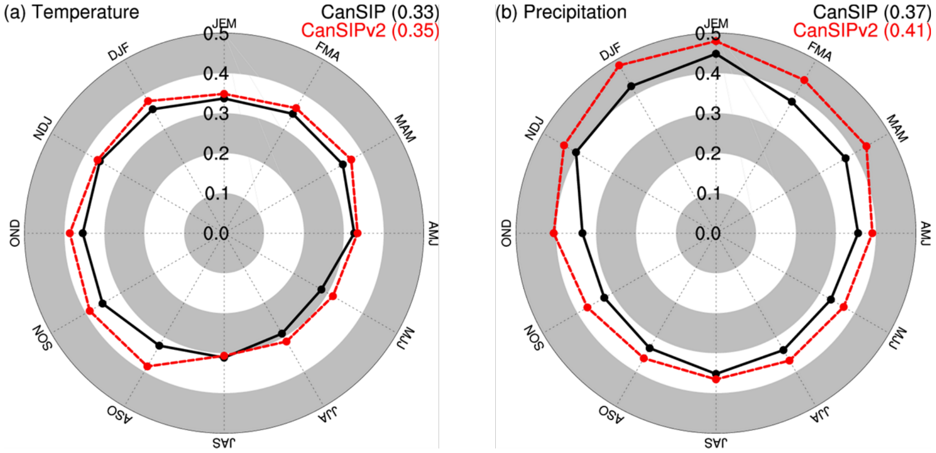

Figure 3. Anomaly pattern correlation coefficients (ACCs) for seasonal mean temperature and precipitation forecasts over the globe of CanSIP and CanSIPv2 for the common period of 1983-2010. The annual mean ACCs for each model are shown in parentheses.

APCC considered several solutions to solve this hindcast issue and took advantage of a large set of models participating in the MME prediction. The first solution would be the use of forecast anomalies with respect to climatologies estimated over the models' own hindcast periods, which vary among the groups producing the model forecast, such as the IRI ENSO forecast (http://iri.columbia.edu/our-expertise/climate/forecasts/enso/current/?enso-sst_table). That is, all models can participate in MME prediction by using forecast anomalies with respect to different base periods and, consequently, to the different climatologies. However, discrepancies may arise if the climatologies differs significantly. We assessed the significance of the difference between climatologies estimated over two periods, 1983-2010 and 1993-2016, which covered the common hindcast period and the most recent hindcast period of the 14 models in the 2019 version, at a 10% significance level based on the Student's *t*-test. The results showed that the differences between the two climatologies of seasonal mean temperature in the observation were statistically significant in many regions (Fig. 4). The most significant differences were evident in the high latitudes of the Northern and Southern Hemispheres throughout all seasons. In these regions, global warming has significantly accelerated in recent years. This is also evident in the South Indian Ocean in MAM and JJA and in the Western Pacific in SON and DJF. Furthermore, for the model with the longest hindcast period spanning from the early 1980s to the most recent years, the differences between climatologies from the periods 1983-2010 and 1993-2016 were also statistically significant (not shown). Thus, the first solution may cause another issue in forecast anomalies because of the significant differences in climatologies due to the different reference (hindcast) periods of individual models, and eventually in the MME prediction that combines the forecast anomalies of individual models (Wallace & Arribas, 2012). Furthermore, this solution is not suitable for users who utilize our seasonal forecasts, such as, NMHSs. Users formulate their local forecasts in terms of anomalies with respect to their local normals estimated over the 30-year period appointed/defined by WMO. As a rule, for their local area of interest, they perform corrections to MME forecasts to account for the difference between the normals estimated over, e.g., 1991-2020 and MME climatology estimated over, e.g., 1983-2010. However, this solution does not provide a reference for the MME climatology period, which may confuse users performing regional/local corrections.

The second solution would be to separate the models into two groups, with hindcast periods specific to each group, and the difference in climatology between the two groups should not be significant. Climatology-I is specified for the current common hindcast period (1983-2010) covered by most models so far. The common hindcast period covered by the newer models (1993-2010) is specified as Climatology-II. As shown in Fig. 4, the difference between Climatology-I and II is not statistically significant most of the globe and seasons. This indicates that the newly developed and recently upgraded models may participate in MME prediction using Climatology-II. This is slightly different from the first solution, as the difference between the two climatologies is not statistically significant, which can reduce some of the confusion in the user's post-processing and interpretation of our forecasts. However, another issue arises as to which a reference period should be applied to observations to assess the MME forecasts combined with two groups of models using different climatologies.

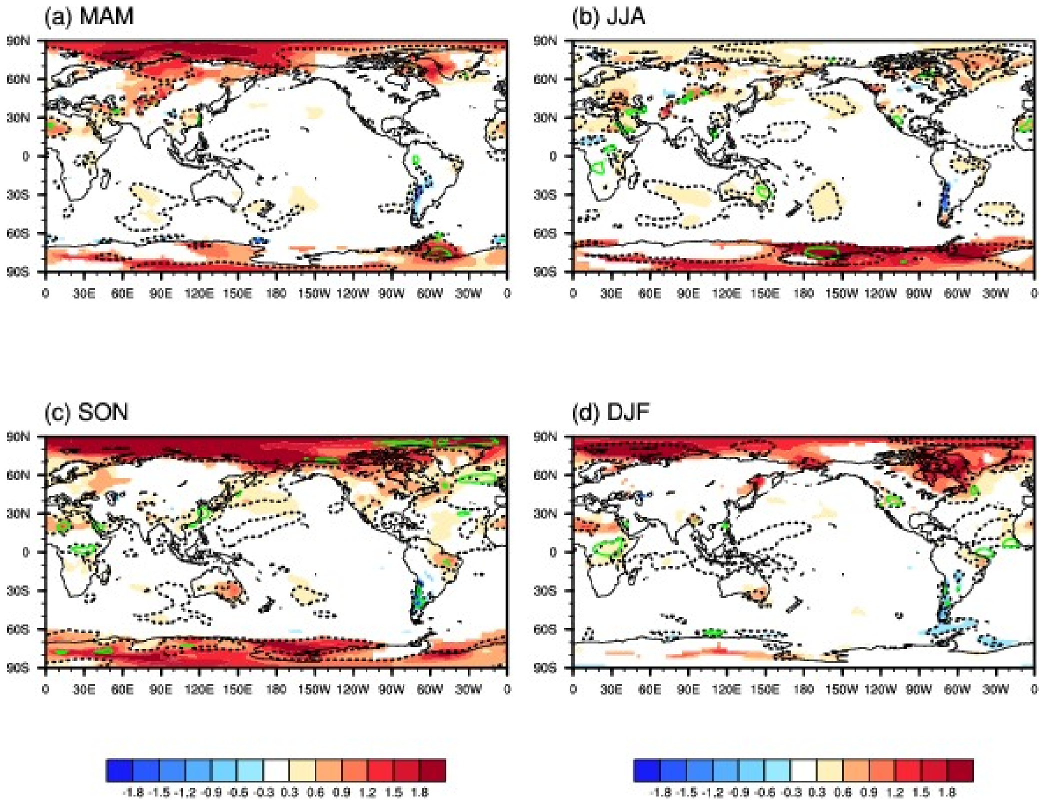

Figure 4. Differences between two climatologies over the period 1983-2010 and 1993-2016 (black dashed line) and trends of observed seasonal mean temperature for the entire 34-year period 1983-2016 (shading). Differences and trends were only displayed at a 10% significance level using Student's t-test and Mann-Kendall test. The green lines represent statistically significant differences in the climatology between the 1983-2010 and 1993-2010 periods.

In this situation, we suggest an alternative solution that is to change the current hindcast period to a unified 1991-2010, for which almost all models could be included. Models of CMCC and UKMO, starting with data from 1993, were treated as missing values for 1991-1992 to allow more models to participate in the MME and extend the hindcast period by at least 20 years. According to the guidelines for objective seasonal forecasting by WMO (2020), hindcast periods shorter than about 20 years may suffer from inadequate sample sizes to allow a robust estimation of skill. In addition, it was mentioned that a shorter hindcast period impacts the merging of information coming from different models using different hindcast periods, especially for MME approaches, because the anomalies and forecast quality are calculated with respect to the hindcast period. Additionally, in terms of prediction skill, increasing the number of participating models, by treating the 2-year period as missing for both models, had a positive effect on improving the MME hindcast skill (not shown). To estimate the forecast skill according to the changes in the number of participating models as the hindcast period for MME climatology changes to unified 1991-2010, we further examined the skill of the MME hindcast in three different model combinations within the model suites of the 2020 version in Table 1. Table 2 shows detailed descriptions of the three different model sets of the MME experiments. Here, 7M was composed of the same models as the experimental MME hindcast results based on the 28-year climatology shown in Fig. 2. However, in this experiment, the 20-year climatology was used to compare the MME prediction skill with all models, including the newly joined models owing to the change in the hindcast period to 1991-2010.

Table 2. Description of three different model suites of MME hindcasts in the 2020 version.

Experiment	Description
7M	7 models expected to continuously participate in MME for 2020 if the current 1983-2010 hindcast period is maintained (APCC, CWB, JMA, MSC/ECCC, NASA, NCEP, PNU)
+6M	Additional 6 models expected to newly participate in MME for 2020 by changing the hindcast period to unified 1991-2010 (BCC, BOM, CMCC, HMC, KMA, UKMO)
13M	All 13 models expected to participate in MME for 2020 by changing the hindcast period to unified 1991-2010

Under the condition of the 1991-2010 hindcast period, the diagrams shown in Fig. 5a and b demonstrate that the skills of the MMEs based on 7M (MME_7M) and +6M (MME_+6M) were comparable, showing ACC=0.36 (0.44) for annual mean temperature (precipitation) for both MMEs. By changing the hindcast period to 1991-2010, the MME consisting of all 13 models (MME_13M) clearly outperformed MME_7M and MME_+6M for both temperature and precipitation over all 12 seasons. The skill improvement of MME_13M forecasts compared with that of MME_7M for both annual mean temperature and precipitation appears not only in the oceans but also on land, with the exception of precipitation in the Arctic region (Fig. 6), where the precipitation is relatively low, and there is significant uncertainty in observations. Consequenctly, the decrease in forecasting skill for precipitation in this region was not considered a significant concern in the paper. Most of these skill improvements in terms of temporal correlation coefficients demonstrated statistical robustness at the 10% significance level in a bootstrap test with 500 Monte-Carlo simulations, particularly evident in regions where the prediction skills are relatively low.

To conduct a detailed examination across seasons and regions, we calculated the ACC-based relative skill difference between the MME_13M and MME_7M for each season and region (Fig. 7). Our analysis revealed a notable enhancement in the forecast skill of MME_13M for temperature during boreal winter seasons, demonstrating its statistical robustness. Notable from a regional perspective, improvements beyond the tropical Pacific are significant, for example, North America for temperature and East Asia, South America and Australia for precipitation. There is variation in skill improvement across seasons and variables. Although the details of this finding are beyond the scope of this study, a potential explanation lines in the inclusion of three models within +6M: UKMO's GloSea5, KMA's GloSea5GC2, and BOM's ACCESS-S, the latter two being developed based on UKMO's GloSea5. It is widely recognized that Glosea5-based models exhibit similar overall model biases and prediction skills. Notably, these models demonstrate high performance in predicting Northern extratropical atmospheric circulation (e.g., Kang et al., 2014; MacLachlan et al., 2015; Scaife et al., 2014; Ham and Jeong, 2021) and the associated temperatures (e.g., Kryjov & Min, 2016; Lim et al., 2019). These findings significantly enhance the forecast skill of MME_13M for boreal winter temperature. However, the improvement in MME_13M prediction skill for summer temperatures was minimal compared to winter, as +6M showed limited improvement in predicting summer temperatures. Conversely, improvements in precipitation were robust across most seasons, with particularly significant enhancements observed during boreal summer seasons. For precipitation, the greatest variability is observed in tropical regions, where it is closely linked to convective activity influenced by ENSO conditions (e.g., Ropelewski and Halpert, 1987; Collins et al., 2010). Consequently, the largest model errors typically occur during spring and summer, particularly when SST forcing is weak or during the ENSO transition phase (e.g., Jin et al. 2008; Wang et al., 2009; Min et al., 2017). In contrast, the strong manifestation of ENSO conditions tends to occur during winter, leading to already commendable accuracy in winter precipitation forecasts, even with older models. In such situation, when the precipitation forecasting skill of +6M is moderate across all seasons, the improvement in precipitation of MME_13M appears to be more significant during the boreal summer seasons, when prediction skill is relatively lower, compared to winter.

Consequently, these skill improvements of MME_13M were mainly due to the higher mean skill of the newly participating models (+6M; mostly recently developed/upgraded models) to MME by changing the hindcast period, compared to the mean skill of the originally participating models (7M) for both temperature and precipitation across all seasons (Fig. 5c, d). In addition, MME_13M, which represents a moderate level when averaging the skills of all 13 models, showed the highest skill because of the increase in the number of models and the corresponding increase in the diversity of the contributing models (Yoo & Kang, 2005; Alessandri et al., 2018). In other words, by changing the hindcast period to the unified 1991-2010, models with relatively high skill can contribute to the MME, which can increase the total number of participating models in the MME and ultimately improve the MME efficiency, thereby improving the prediction skill of MME_13M compared to MME_7M.

Based on the results of the hindcast experiments, we changed the common base period to 1991-2010 for APCC MME operation from 2020, which is covered by almost all the models (Oper). Finally, we assessed the MME skills of real-time forecasts from 2020JFM to 2023JAS using the most recently updated observations. Real-time forecast verification is important for operational centers to assess whether skill improvement exists in real-time forecasts as well as in hindcasts, although this period is too short for the collection of a sufficient number of real-time forecasts to obtain well-grounded conclusions. We first assessed Oper's forecast skill for both variables, indicating a strong dependence on ENSO strength, which reaches its peak in boreal winter and serves as one of the key sources of predictability for seasonal forecasts (Fig. 8a; Wang et al. 2009; Barnston et al. 2010; Min et al. 2017). For example, relatively high levels of the forecast skills were observed during the boreal autumn and winter seasons of 2020/21 and 2021/22, coinciding with moderate La Nina events. Towards mid-2023, a strong El Nino was developing, accompanied by an improvement in forecasting skill. Meanwhile, in 2022 the strong negative Nino 3.4 SST anomaly persisted into spring, summer, and autumn, providing strong persistent forcing that governed skillful seasonal forecast. Conversely, the relatively low skills were observed during the transition and/or ENSO-neutral phases of 2020, 2021 and 2023. For comparison, we produced MME forecasts for the same periods as the models that would have participated in the MME if the 1983-2010 hindcast period had not changed (Exp). By changing the hindcast period to 1991-2010, the number of participating models in the real-time MME operations in 2020JFM-2023JAS increased by 100%, and the difference between Oper and Exp gradually widened (Fig. 8b). The improvement or degradation in forecast skill by Oper fluctuates across seasons and years under limited data set conditions. However, an encouraging finding for real-time forecasts is the significant enhancement in Oper manifested from mid-2022, coinciding with a widening disparity in the number of participating models between Oper and Exp. That is, as the models continued to improve, along with the hindcast period shifted, it was clear that if the 1983-2010 hindcast period had been maintained, the number of participating models in the MME operations would have gradually decreased, leading to a subsequent decline in forecast skill. As a result, from the preliminary results of the real-time forecasts, substantial improvements in temperature over the globe have been observed in recent years; however, the prediction of precipitation still remains a difficult problem, with little change on a global scale (Fig. 8c). Given that the assessment for real-time forecast has been based on limited data, more detailed analysis is needed to determine the causes of the improvement and decrease in forecast skill for further study as more data become available. Based on the results from hindcasts and real-time forecasts, the change in the common hindcast period to 1991-2010 for MME prediction in 2020 was an appropriate action for APCC operation from a long-term perspective.

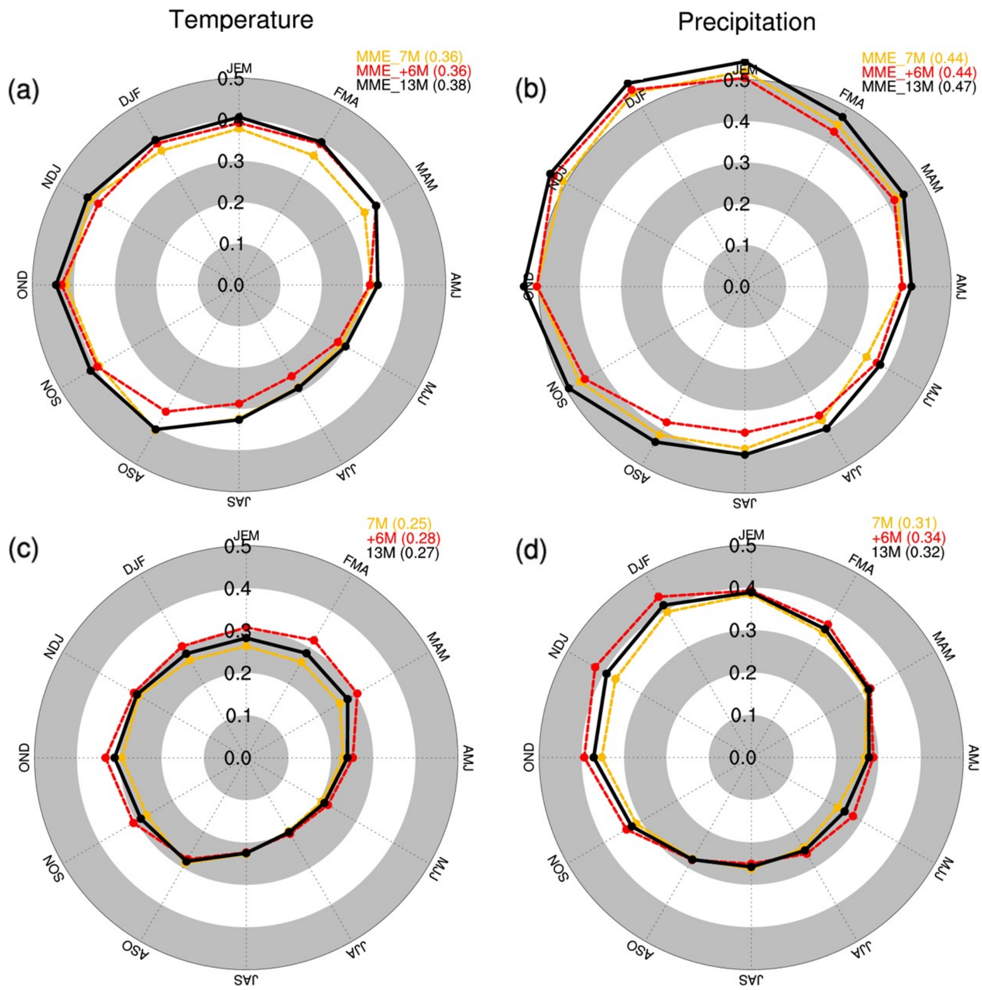

Figure 5. (a, b) ACCs of MME hindcasts (1991-2010) with different model combinations (MME_7M, MME_+6M, and MME_13M) and (c, d) average ACCs of the participating models for each combination (7M, +6M, and 13M), for seasonal mean temperature and precipitation forecasts over the globe. The annual mean ACCs for each MME and the average of models' skills are shown in parentheses.

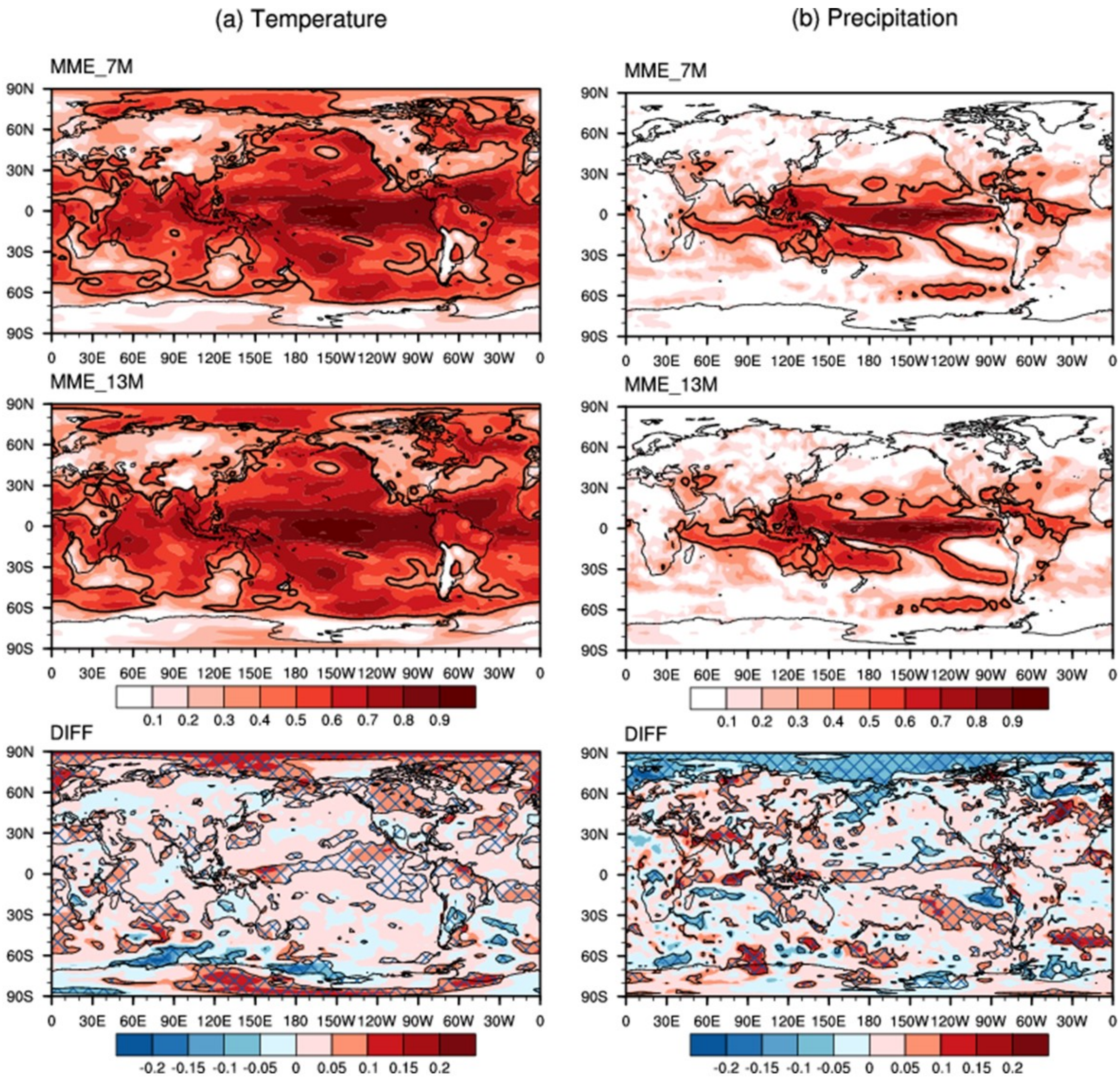

Figure 6. Spatial distributions of annual mean temporal correlation coefficients (TCCs) for the MME hindcast (1991-2010) with 7 models (MME_7M) and 13 models (MME_13) of seasonal mean temperature and precipitation. The contour lines enclose the areas in which the TCCs are statistically significant at the 5% level using a two-tailed Student's t-test. The skill differences (DIFF) indicate the differences between the two MMEs (MME_13M minus MME_7M), with the skill difference being statistically robust at the 10% significance level in a bootstrap test with 500 Monte-Carlo simulations.

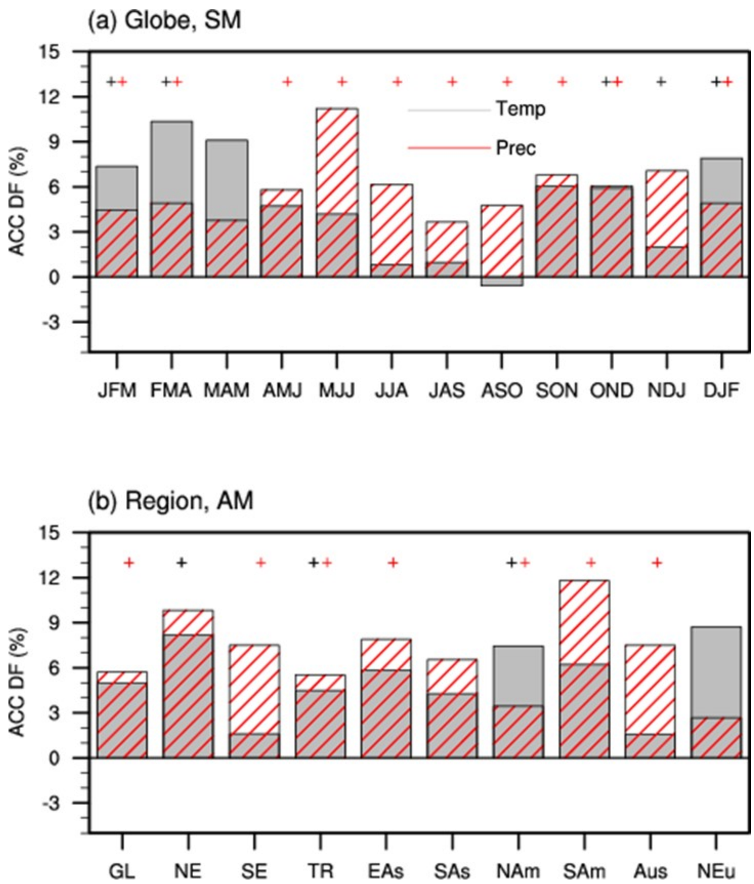

Figure 7. (a) ACC-based relative skill difference of MME_13M hindcasts to MME_7M hindcasts of seasonal mean temperature and precipitation forecasts over the globe and (b) annual mean forecasts for several sub-regions for the period of 1991-2010. The black and red crosses mark the seasons and regions for which the relative skill difference is statistically robust at the 10% significance level in a bootstrap test with 500 Monte-Carlo simulations.

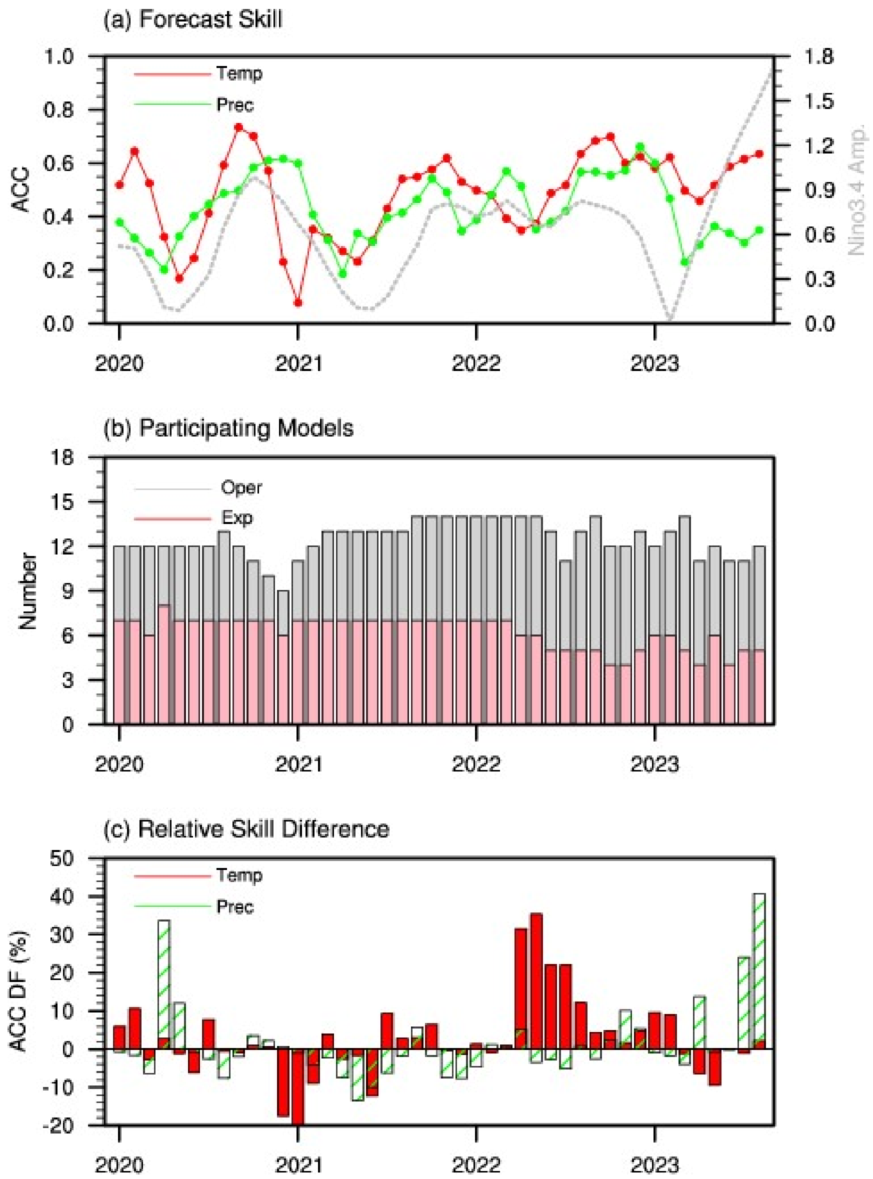

Figure 8. (a) ACCs of real-time operational MME forecasts (Oper) for global temperature and precipitation for 2020JFM-2023JAS. The grey line indicates the amplitude (absolute value) of 3-month mean Nino 3.4 Index. (b) Number of participating models in Oper and experimental forecasts (Exp) and (c) Relative skill difference of ACCs from Exp to Oper for global temperature and precipitation.

4 Conclusions

The construction of the MME is a compromise between the number of participating models and the length of the common hindcast period. An increase in the number of participating models with sufficient model diversity decreases random and model formulation errors in MME forecasts (e.g., DelSole et al., 2014; Yang et al., 2016). On the other hand, an increase in the length of the common hindcast period decreases errors in climatology but increase random and model formulation errors because of a decrease in the number of participating models in the MME prediction (e.g., Shi et al., 2015). In this situation, as the hindcast periods of recently developed and improved models have shifted to the latest, APCC faced new challenges in 2019 while continuing to maintain a common hindcast period for many years. As a result, the proportion of models that could not participate in operational MME prediction was expected to be approximately 50% by 2020 because their hindcast periods started in the mid-1980s to early 1990s. Based on the results of several experiments, we proposed a solution to change the common hindcast period to a unified 1991-2010, which is the most appropriate method for APCC operation, reflecting recently developed models. That is, by changing the reference period for MME prediction, APCC provides opportunities for participation in operational MME prediction for newly developed/upgraded models, resulting in a double increase in the number of participating models and improvement in the MME prediction skill.

However, some questions remain regarding whether the 20-year hindcast period is sufficient to represent the climatological means. Because the operational MME center incorporates predictions from various models, it is inevitable that the hindcast period for the MME is shorter than that for individual models. The suggested 20-year climatology is comparable to the climatologies of other MME groups for seasonal forecasting (e.g., WMO LC-LRF (1993-2009; 17 years) and C3S (1993-2016; 24 years)). Although WMO recommends that the hindcast period should be as long as possible (WMO, 2019) and that a short period may affect the estimation of anomalies and forecast skill of MME, especially those that integrate predictions from various models, even the WMO LC-LRF currently uses a common 17-year hindcast period in performing MME by integrating outputs from 16 Global Producing Centres' (GPC) models. That is, there are still realistic limitations or gaps in the hindcast period of producing centers that match the WMO recommendation. The differences in hindcast periods for each model mainly stem from when the models were developed and the production schedule for its operation. For example, the hindcast period of recently developed models has shifted to more recent years, whereas the hindcast period of models that were developed relatively early and have continued to be maintained mostly covers the hindcast period of 1980s to mid-to-late 2010s.

Moreover, in terms of the production schedules, some systems follow a so-called "on the fly" approach, generating a new set of hindcasts every time a new forecast is produced (e.g., PNU). In some models, fixed hindcasts are produced before the system becomes operational and remain unchanged throughout its operational lifetime (e.g., NCEP). Each method has its own advantages, and each modeling center produces hindcasts in a manner that is appropriate for their operational situation. This issue can be fundamentally solved by making further efforts to extend or shift the hindcast period at each modeling center, along with improvements in other modeling components. As part of these efforts, APCC, as one of the MME model providers, is currently working to expand the period for the APCC's in-house model, SCoPS, to mid to late 2010s. Another aspect of the APCC's efforts as an MME center is to encourage MME model providers to expand the hindcast period to the latest through regularly held APCC MME Model Providers' Meetings. However, these problems cannot be solved in a short time and may not be feasible on the operational situation of each modeling center. In this situation, this study is significant in that we addressed the critical and practical challenges recently faced by operational MME centers due to the hindcast issue and provided various approaches that MME groups can consider to solve these problems.

Finally, although not within the scope of this study, the most important issue in recent years is that since late 2021, NMHSs worldwide have used the WMO recommended 1991-2020 normals (https://www.wmo.int/edistrib_exped/grp_prs/_en/08791-2019-CLW-CLPA-DMA-CLIN8110_en.pdf). However, there are still some limitations to matching with the WMO-recommended normal period; currently no climate center providing MME seasonal forecasts to the NHMSs uses a climatology matching with the WMO references. In particular, the recent period in which the difference between the model climatology (e.g., 1991-2010) and the WMO normal (e.g., 1991-2020) appears is the period when global warming is accelerating. Therefore, forecast anomalies based on a more recent reference climate may be more relevant in the context of climate change (WMO, 2020). It is more difficult to make seasonal forecasts during periods of strong climate trends, and the warming trends are important effects that should not be discarded. Therefore, further studies needed on the methodologies for adjusting and correcting (or calibrating) the climatology in models to the WMO normal, including recent periods.

Acknowledgments

This study was supported by APEC Climate Center. The authors acknowledge the APCC MME Producing Centres (PCs) for making their hindcast/forecast data available for analysis and the APCC for collecting and archiving them and for organizing the APCC MME prediction. We also thank the APCC MME PCs for discussing this issue at the 3rd APCC MME Providers' Meeting in 2019.

Data Availability Statement

The single-model and MME predictions used in this study are available from the Climate Service Integration Platform, Climate Information toolkit (CLIK; https://cliks.apcc21.org). The National Center for Environmental Prediction – Department of Energy (NCEP-DOE) reanalysis 2 was obtained from the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA (available online at https://psl.noaa.gov/data/gridded/index.html). The monthly precipitation was acquired from the NOAA/NCEP climate anomaly monitoring system – outgoing longwave radiation precipitation index (CAMS OPI; available online at https://www.cpc.ncep.noaa.gov/products/global_precip/html/wpage. cams_opi.html). For Nino 3.4 index, we use the optimum interpolation (OI) version 2 monthly mean SST (OI SSTv2; available online at https://www.psl.noaa.gov/data/gridded/data.noaa.oisst.v2.html).

References

Ahn, J. B., & Kim, H. J. (2013). Improvement of one-month lead predictability of the wintertime AO using a realistically varying solar constant for a CGCM. *Meteorological Applcations*, *21*(2), 415-418. https://doi.org/10.1002/met.1372

Alessandri, A., Catalano, F., Lee, J. Y., Wang, B., Lee, D. Y., Yoo, J. H., & Weisheimer, A. (2018). Grand European and Asian-Pacific multi-model seasonal forecasts: maximization of skill and of potential economical value to end-users. *Climate Dynamics*, *50*, 2719-2738. https://doi.org/10.1007/s00382-017-3766-y

Barnston, A. G., Shuhua, L., Mason, S. J., DeWitt, D. G., Goddard, L., & Gond, X. (2010) Verification of the first 11 years of IRI's seasonal climate forecast. *Journal of Applied Meteorology and Climatology*, *49*, 493-520. https://doi.org/10.1175/1520-0493(2003)131<1509:SAIVOT>2.0.CO;2

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

Collins, M., An, S. I., Cai, W., Ganachaud, A., Guilyardi, E., Jin, F. F., Jochum, M., Lengaigne, M., Power, S., Timmermann, A., Vecchi, G., & Wittenberg, A. (2010). The impact of Global Warming on the Tropical Pacific Ocean and El Nino. *Nature Geoscience*, *3*(6), 391-397.

Cottrill, A., Hendon, H. H., Lim, E.-P., Langford, S., Shelton, K., et al. (2013). Seasonal forecasting in the Pacific using the coupled model POAMA-2. *Weather and Forecasting*, *28*(3), 668-680. https://doi.org/10.1175/WAF-D-12-00072.1

DelSole, T., Nattala, J., & Tippett, M. K. (2014). Skill improvement from increased ensemble size and model diversity. *Geophysical Researh Letters*, *41*, 7331–7342. https://doi.org/10.1002/2014GL060133

Ham, H., Lee, S.-M., Hyun, Y.-K., & Kim, Y. (2019). Performance assessment of monthly ensemble prediction data based on improvement of climate prediction system at KMA. *Atmosphere*, *29*(2), 149-164. https://doi.org/10.14191/Atmos.2019.29.2.149

Ham, S. Y., Im, A.-Y., Kang, S., Jeong, H., & Jeong, Y. (2019). Correction to : A newly developed APCC ScoPS and its prediction of East Asia seasonal climate variability. *Climate Dynamics*, *53*, 3703-3704. https://doi.org/10.1007/s00382-019-04894-y

Ham, S. Y. And Jeong, Y (2021) Characteristics of subseasonal winter prediction skill assessment of GloSea5 for East Asia. Atmopshere, 12(10), 1311. https://doi.org/10.3390/atmos12101311

Hudson, D., Alves, O., Hendon, H. H., Lim, E.-P., et al. (2017). ACCESS-S1 The new Bureau of Meteorology multi-week to seasonal prediction system. *Journal of Southern Hemisphere Earth System Science*, *67*(3). https://doi.org/10.22499/3.6703.001

Janowiak, J. E,, & Xie, P. (1999). CAMS_OPI: a global satellite-raingauge merged product for real-time precipitation monitoring applications [Dataset]. *Journal of Climatology*, 12(11), 3335-3342. https://doi.org/10.1175/1520-0442(1999)012<3335:COAGSR>2.0.CO;2

Jin, E. K., Kinter III, J. L., Wang, B., Pack, C. K., Kang, I. S., Kirtman, B. P., Kug, J.-S., Kumar, A., Juo, J.-J., Schemm, J., Sukla, J., & Yamagata T. (2008). Current status of ENSO prediction skill in coupled ocean-atmosphere models. *Climate Dynamics*, *31*(6), 647-664. https://doi.org/10.1007/s00382-008-0397-3

Kang, D., Lee, M. I., Im, J., Kim, D., Kim, H. M., Kang, H. S., Schubert, S. D., Arribas, A., & MacLachlan, C. (2014). Prediction of the Arctic Oscillation in boreal winter by dynamical seasonal forecasting systems. *Geophysical Research Letters*. *41*, 3577–3585. https://doi.org/10.1002/2014GL060011

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.-K., Hnilo, J. J., Fiorino, M., & Potter, G. L. (2002) [Dataset]. NCEP-DOE AMIP-II Reanalysis (R-2). *Bulletin of the American Meteorological Society*, 83(11), 1631-1643. https://doi.org/10.1175/Bams-83-11-1631

Kendall, M. G. (1975). Rank Correlation Methods. New York, NY, Oxford University Press

Kim, G., Ahn, J. B., Kryjov, V. N., Lee, W. S., Kim, D. J., & Kumar, A. (2021). Assessment of multi-model ensemble methods for WMO LC-LRFMME. *International Journal of Climatology, 41*, E2462-E248. https://doi.org/10.1002/joc.6858.

Kirtman, B. P., Min, D., Infanti, J. M., 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

Kryjov, V. N., & Min, Y.-M. (2016). Predictability of the wintertime Arctic Oscillation based on autumn circulation. *International Journal of Climatology*, *36*, 4181–4186. https://doi.org/10.1002/joc.4616

Kumar, A., Chen, M., Zhang, L., Yang, W., Xue, Y., Wen, C., Marx, L., & Huang, B. (2012). An analysis of the nonstationarity in the bias of sea surface temperature forecasts for the NCEP Climate Forecast System (CFS) version 2. *Monthly Weather Review*, *140*, 3003-3016. https://doi.org/10/1175/MWR-D-11-00335.1

Lin, H., William, J., Muncaster, R., et al. (2020). The Canadian Seasonal to Interannual Prediction System Version 2 (CanSIPv2), *Weather & Forecasting*, *35*(4), 1317-1343. https://doi.org/10.1175/WAF-D-19-0259.1

Lim, J., Dunstone, N. J., Scaife, A. A., and Smith D. M. (2019). Skillful seasonal prediction of Korean winter temperature. *Atmospheric Science Letter*, *20*, https://doi.org/10.1002/asl.881

MacLachlan, C., Arribas, A., Peterson, K. A., Maidens, A., Fereday, D., Scaife, A. A., Gordon, M., Vellinga, M., Williams, A., Comer, R. E., Camp, J., Xavier, P., & Madec, G. (2015). Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system. *Q. J. R. Meteorol. Soc*. *141*, 1072–1084. https://doi.org/10.1002/qj.2396

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica,* 13(3), 245-259. https://doi.org/10.2307/1907187

Manzanas, R., Gutierrez, J. M., Bhend, K., Hemri, S., Doblas-Reyes, F. J., Torralba, V., Penabad, E., & Brookshaw, A. (2019). Bias adjustment and ensemble recalibration methods for seasonal forecasting: a comprehensive intercomparison using the C3S dataset. *Climate Dynamics*, 53, 1287-1305. https://doi.org/10.1007/s00382-019-04640-4

Merryfield, W. J., et al. (2013). The Canadian seasonal to interannual prediction system. Part I: models and initialization. *Monthly Weather Review*, 141, 2910–2945. doi:10.1175/MWR-D-12-00216.1

Min, Y. M., Kryjov, V. N., & Oh, S. M. (2014). Assessment of APCC multimodel ensemble prediction in seasonal climate forecasting: Retrospective (1983-2003) and real-time forecasts (2008-2013). *Journal of Geophysical Research: Atmospheres*, 199, 12,132-12,150. https://doi.org/10.1002/2014JD022230

Min, Y. M., Kryjov, V. N., Oh, S. M., & Lee, H. J. (2017). Skill of real-time operational forecasts with the APCC multi-model ensemble prediction system during the period 2008-2015 [Dataset]. *Climate Dynamics*, *49*(11–12), 4141–4156. https://doi.org/10.1007/s00382-017-3576-2

Molod, A. M., Takacs, L. L., Suarez, M., & Bacmeister, J. (2015). Development of the GEOS-5 atmospheric general circulation model: evolution from MERRA to MERRA2. *Geoscientific Model Development*, 8, 1339-1356. https://doi.org/10.5194/gmd-8-1339-2015

Paek, H., Yu, J.-Y., Hwu, J.-W., Lu, M.-M., & Gao, T. (2015). A source of AGCM bias in simulating the western Pacific subtropical high: Different sensitivities to the two types of ENSO. *Monthly Weather Review*, *143*(6), 2348-2362. https:/doi.org/10.1175/MWR-D-14-00401.1

Reynolds, R. W., Rayner, N. A., Smith, T. M., Stokes, D. C., & Wang, W. (2002) An improved in situ and satellite SST analysis for climate [Dataset]. *Journal of Climate*, *15*, 1609-1625. https://doi.org/10.1175/1520-0442(2002)015<1609:AIISAS>2.0.CO;2

Ropelewski, C. F., & Halpert, M. S. (1987) Global and Regional Scale Precipitation Patterns Associated with the El Nino/Southern Oscillation. *Monthly Weather Review*, *115*(8), 1606-1626. https://doi.org/10.1175/1520-0493(1987)115<1606:GARSPP>2.0.CO;2

Saha, S., et al. (2014). The NCEP climate forecast system version 2. *Journal of Climate*, *27*, 2185-2208. https://doi.org/10.1175/JCLI-D12-00823.1

Sanna, A., Borrelli, A., Pthanasiadis, P., Materia, S., Storto, A., Tibaldi, S., & Gualdi, S. (2017). CMCC-SPS3: the CMCC Seasonal Prediction System 3. Centro Euro-Mediterraneo sui Cambiamenti Climatici. CMCC Tech. Rep. RP0285, p61

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 Letter*, *41*, 2514-2519. https://doi.org/10.1002/2014GL059637

Shi, W., Schaller, N., MacLeod, D., Palmer, T. N., & Weisheimer, A. (2015). Impact of hindcast length on estimates of seasonal climate predictability. Geophyiscal Research Letters, *42*(5). doi:10.1002/2014GL062829

Stephenson, D. B., & Doblas-Reyes, F. J. (2000). Statistical methods for interpreting Monte Carlo ensemble forecasts, *Tellus (A)*, *52*, 300–322. https://doi.org/10.3402/tellusa.v52i3. 12267

Stockdale, T. N. (1997). Coupled ocean-atmosphere forecasts in the presence of climate drift. *Monthly Weather Review*, *235*, 809-818. https://doi.org/10.1175/1520-0493(1997)125<0809:COAFIT>2.0.CO;2

Takaya, Y., Hirahara, S., Yasuda, T., Matsueda, S., et al. (2018). Japan Meteorological Agency/Meteorological Research Institute-Coupled Prediction System version 2 (JMA/MRI-CPS2): atmosphere-land-ocean-sea ice coupled prediction system for operational seasonal forecasting. *Climate Dynamics*, *50*(1), 1-15. https://doi.org/10.1007/s00382-017-3638-5

Wang, B., et al. (2009) Advance and prospectus of seasonal prediction: assessment of the APC/CliPAS 14-model ensemble retrospective seasonal prediction (1980-2004). Climate Dynamics, 33, 93-117. https://doi.org/10.1007/s00382-008-0460-0

Wallace, E., Arribas, A. (2012). Forecasting with reference to a specific climatology. *Monthly Weather Review*, *140*, 3795-3802. https://doi.org/10.1175/MWR-D-12-00159.1

Wilks, D. S. (1995). *Statistical Methods in the Atmospheric Sciences*, 2nd ed. Elsevier Academic Press: San Diego, CA.

Wilks. D. S. (1997). Resampling hypothesis tests for autocorrelated fields. *Journal of Climatology*, *10*, 65–82. https://doi.org/10.1175/1520-0442(1997)010<0065:RHTFAF> 2.0.CO;2.

WMO (2007). *The role of climatological normals in a changing climate*. WCDMP-No. 61, WMO/TD-No. 1377, Geneva, World Meteorological Organization

WMO (2019) Manual on the Global Data-processing and Forecasting System. WMO-No. 485, Geneva, Switzerland

World Meteorological Organization (2020). Guidance on Operational Practices for Objective Seasonal Forecasting. WMO-No. 1246, Geneva, Switzerland

Yang, D., Yang, X. O., Xie, Q., Zhang, Y., Ren, X., & Tang, Y. (2016). Probabilistic versus deterministic skill in predicting the western North Pacific-East Asian summer monsoon variability with multimodel ensembles. *Journal of Geophysical Research-Atmosphere*, *121*(3), 1079-1103. https://doi.org/10.1002/2015JD023781

Yoo, J, H., & Kang, I. S. (2005). Theoretical examination of a multi-model composite for seasonal prediction. *Geophysical Research Letter*, *32*, L18707. https://doi.org/10.1029/2005GL023513