

19 **Abstract**

20 Seasonal forecasts are commonly issued in the form of anomalies as departures from average in a
21 specified multiyear reference period (climatology). The model climatology is estimated as
22 average of retrospective forecasts over the hindcast period. However, different operational
23 centers providing seasonal ensemble predictions use different hindcast periods based on their
24 own model climatology. In addition, the hindcast period of recently developed/upgraded newer
25 models tends to shift to the recent years. In this paper, we discuss recent challenges faced by the
26 APCC multi-model ensemble (MME) operations, especially changes in the hindcast period for
27 individual models. Based on the results of various sensitivity experiments for the MME
28 prediction, we proposed to change the hindcast period that is the most appropriate solution for
29 the APCC operations. It makes the newly developed models join the MME and increase the total
30 number of participating models, which facilitates the skill improvement of the MME prediction.

31

32 **Plain Language Summary**

33 In seasonal forecasting, it is well known that the MME, which combines different single-model
34 prediction from various operational and research centers, is more effective way to improve the
35 forecast skill. Since 2005, the APCC has been providing the MME seasonal forecasts and the
36 models participating in the APCC MME operations have been continuously changing.
37 Particularly, as the hindcast period of newly developed models is shifted to the latest, they
38 cannot participate in the operational MME forecasts because of discrepancy in climatologies.
39 However, over time, as the number of new models expected to have skillful forecasts gradually
40 increases, APCC faced the challenge of continuously reducing the number of participating
41 models or changing the hindcast period to the more recent years. Considering various aspects
42 such as the number of participating models, skills, and climatology period, we have selected the
43 most appropriate way for the APCC operations. Through this, the MME prediction skill has been
44 improved over most of the globe and seasons due to increasing of the number of participating
45 models, particularly inclusion of newer models.

46

47 **1 Introduction**

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

58 World Meteorological Organization (WMO) recommends climatology (normals) to be
59 estimated as 30-year averages computed for the most-recent 30-year period finishing in a year
60 ending with 0 (WMO, 2007), i.e., 1991-2020 at present. National meteorological and
61 hydrological services (NMHS) estimate forecasts as departures from these 30-year normals in
62 their locations. However, different operational and research centers use different hindcast periods
63 for estimation of the model climatology. Furthermore, hindcast periods of recently developed
64 and improved climate models, particularly beginning of the hindcast period, tends to be shifted to
65 recent years. The Asia-Pacific Economic Cooperation (APEC) Climate Center (APCC) is one of
66 major operational centers providing well-validated multi-model ensemble (MME) seasonal
67 forecasts. Since its establishment in 2005, APCC has been collecting dynamical ensemble
68 forecasts through multi-institutional cooperation and coordinating the MME prediction. At
69 present, 15 leading operational and research institutes from 11 countries are involved in the
70 APCC operational MME prediction. The MME operational centers, e.g., APCC (Min et al., 2014,
71 2017), the WMO Lead Center (WMOLC; Kim et al., 2021), the North American MME (NMME;
72 Becker et al., 2014; Kirtman et al., 2014), and the Copernicus Climate Change Service (C3S;
73 Manazanas et al., 2019) use a common hindcast period of all participating models, which results
74 in a relatively short period compared to that of single-model prediction systems.

75 As the hindcast period for recently developed newer models has gradually shifted to the
76 later years, the full range of hindcast periods for the dynamical models routinely running in
77 operational centers is widening, from early-1980s to late-2010s nowadays. However, the
78 common hindcast period is rather shortening due to shift of the newer models' hindcast period
79 beginning to the early 1990s. This raised new issue at APCC that combines all information from
80 different climate prediction systems, particularly in 2019. This is because some of contributing
81 models in the APCC MME prediction were changed to their upgraded versions in 2020, and their
82 hindcast periods were shifted to more recent years. That is, with implementation of new models,
83 if the common hindcast period, 1983-2010, were kept, the number of participating models in
84 MME would have been reduced and would be gradually reducing in future because recently
85 developed models that are expected to have better skill do not match with this common period. It
86 might lead to deterioration of the MME prediction skill. Therefore, the time has come for APCC
87 to consider the issue on the hindcast period that could affect the number of participating models
88 in MME and eventually the MME skill. In this study, we discuss challenges faced by MME
89 operations caused by upgrading of the participating model set. In particular, we focus on the
90 issue of a decrease in the number of participating models in the MME prediction with a shift to

91 the later years of the hindcast periods of recently developed models. We suggest most
92 appropriate solution for APCC operations based on several sensitivity experiments on the
93 different hindcast periods and different number of participating models in MME.

94

95 **2 Method and Data**

96 With the most recent joining of SYS8 from METFR, APCC currently collects the
97 ensemble predictions from 15 state-of-the-art climate models, and the models are being
98 continuously improved with the great efforts of their own operational and research centers. The
99 APCC MME model suits of 2019 and 2020 are listed in Table 1. In 2019, the real-time
100 operational MME prediction suit comprised eight models from APCC, BOM, CWB, JMA,
101 MSC/ECCC, NASA, NCEP, and PNU that matched with the common hindcast period of 1983-
102 2010. However, APCC was not able to involve recently upgraded models from CMCC, KMA,
103 and UKMO in the real-time operations due to the mismatch of the hindcast period. Furthermore,
104 several models were scheduled to be changed to their upgraded versions in 2020. To test the
105 sensitivity in terms of predictability as the participating models in MME change due to their
106 improvements, we performed several experiments with varying reference periods and
107 contributing models in MME. Here, the MME forecast is a simple average of individual models
108 with equal weights.

109 We focus on 1-month lead 3-month mean (seasonal) MME forecasts of 2m temperature
110 and precipitation over globe and sub-regions: Northern Extratropics (NE; 20°N-90°N), Southern
111 Extratropics (SE; 20°S-90°S), Tropics (TR; 20°N-20°S), East Asia (EAs; 75°E-150°E, 15°N-
112 60°N), South Asia (SAs; 60°E-140°E, 10°S-35°N), North America (NA_m; 190°E-310°E, 10°N-
113 75°N), South America (SA_m; 270E-330E, 60°S-10°N), Australia (Aus; 110E-180E, 50°S-0°N),
114 and Northern Eurasia (NE; 25°E-190°E, 40°N-80°N). For skill assessment, we use the National
115 Center for Environmental Prediction (NCEP)-Department of Energy (DOE) Reanalysis 2 data
116 (Kanamitsu et al., 2002) for temperature and the Climate Anomaly System and Outgoing
117 longwave radiation Prediction Index data (CAMS-OPI, Janowiak and Xie, 1999) for
118 precipitation. All model forecasts and observations are interpolated onto a 2.5 x 2.5 common grid.
119 We use the anomaly pattern correlation coefficient (ACC) and temporal correlation coefficient
120 (TCC) to assess the prediction skill. We use the ACC-based skill score (Murphy, 1988) for
121 assessment of the prediction skill improvement and deterioration of the MME forecast with
122 another model set compared to the reference model set. The Student's t-test and the Mann-
123 Kendall test is used to assess the statistical significance of the difference between means and
124 trends.

125

126 **3 Results**

127 A history of operational exploiting of the seasonal prediction models exceeds two
128 decades. It has been marked by numerous upgrades of models, resolution, ensemble size, lead
129 time, etc. Operational long-range forecasting centers perform essential efforts to improve their
130 climate prediction system. Particularly, they tend to extend the period of hindcasts over which
131 climatology is estimated and to move it to more recent years. As shown in Fig. 1(a), the number
132 of models providing their ensemble forecasts to APCC and contributing to the real-time APCC
133 MME predictions vary from year to year, depending on the operational environmental at the time

134 of the given forecast issuing. Here, the proportion of models that do not participate in the real-
 135 time MME prediction has been gradually increasing and it was expected to increase up to nearly
 136 50% in 2020 (red line in Fig. 1(a)). It is noteworthy that about 80% of these models do not
 137 participate in real-time MME prediction because of inconsistency with the common hindcast
 138 period, 1983-2010 (black line in Fig. 1(a)). That is, model developers continue to improve their
 139 models and gradually move their hindcast periods to more recent years, however, with the
 140 current 1983-2010 MME hindcast remaining unchanged, the number of the models participating
 141 in the APCC real-time operations matching the current MME hindcast period will gradually
 142 decrease.

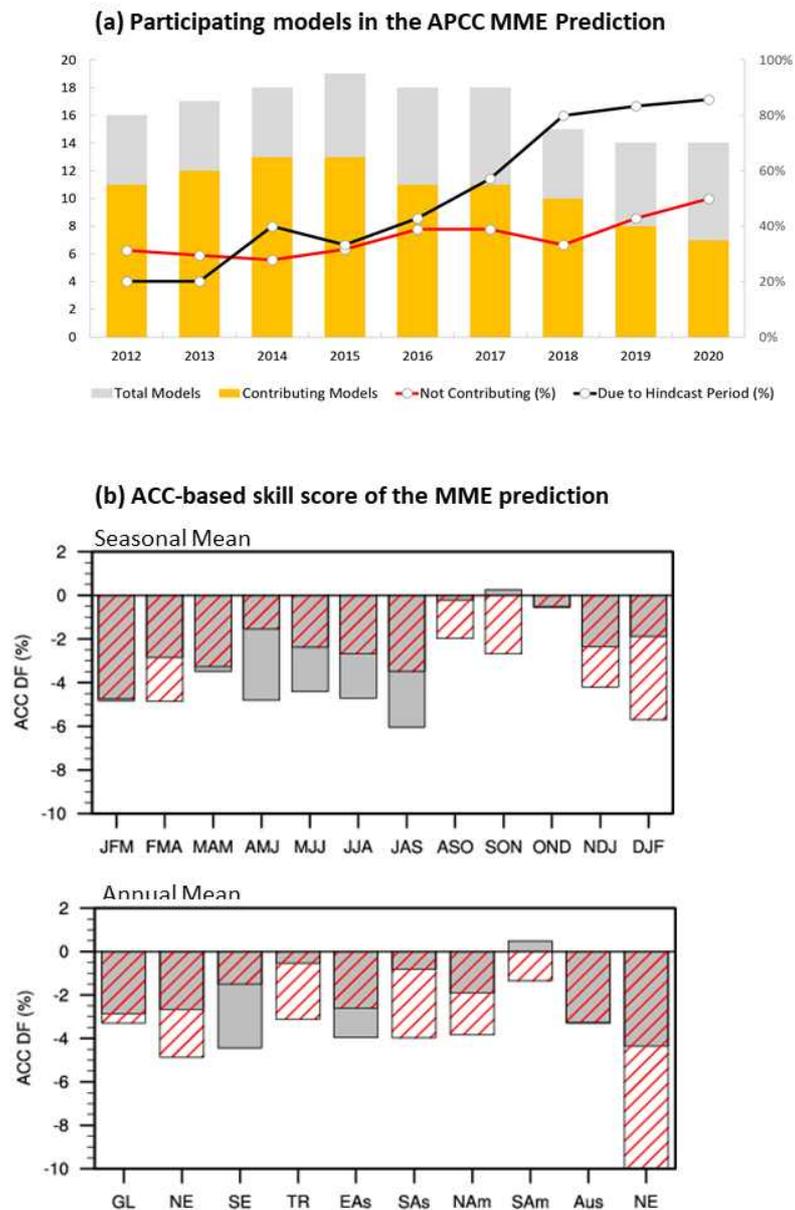
143

144 **Table 1.** Evolved models in the real-time APCC MME prediction in 2019 (v2019) and in 2020
 145 (v2020)

Institute	v2019		v2020	
	Model Name	Hindcast Period	Model Name	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	1982-2013	MGOAM-2	1982-2013
MSC/ECCC	CanSIP	1981-2010	CanSIPsv2	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	1980-2018
UKMO	GloSea5	1993-2016	GloSea5	1991-2016

146 Bold in 2019 indicates the models contributing to the real-time APCC MME operation.

147 * Not participating model due to the inconsistency of the hindcast experiment (i.e., the AMIP-
 148 type)



149

150 **Figure 1.** (a) Changes in the number of models providing their seasonal forecasts to APCC (grey
 151 bar) and participating in the real-time operational APCC MME prediction (yellow bar) for the
 152 period 2012-2019 and expected changes in 2020. Lines indicate the proportion of models that do
 153 not participate in the MME operation (red, %) and the proportion of the not-participating models
 154 due to the common hindcast period mismatch (black, %). (b) ACC-based skill score of the MME
 155 predictions comprising forecasts from seven models of v2020 in respect to that comprising eight
 156 models of v2019 for the period of 1983-2010 of seasonal mean temperature and precipitation
 157 forecasts over globe and annual mean for several sub-regions.

158

159 The more important issue is the MME skill that is affected by the mean skill of individual
160 models and models' diversity (Yoo and Kang 2005; Alessandri et al., 2018). If the number of
161 contributing models in the MME predictions continues to reduce, particularly by excluding of
162 recently developed and improved newer models, it may result in the MME skill decreases. Fig.
163 1(b) shows the ACC-based skill score of the MME predictions comprising forecasts from seven
164 models of v2020 in respect to that comprising eight models of v2019 (Table 1) selected under
165 condition that the MME system matches the common 28-year hindcast period (1983-2010). The
166 skill score is mainly negative that indicates deterioration in the MME skill caused by the
167 decrease in the number of participating models even by one. The hindcast skill of both global
168 temperature and precipitation using v2020 decreases across almost all seasons. It is also true for
169 sub-regions in terms of annual means, with exception of temperature in South America. Thus, if
170 we keep the 28-year long climatology period of 1983-2010, we will not be able to incorporate
171 recently developed/upgraded models to the MME prediction system and as a result the MME
172 prediction skill may continue to decrease along with decrease in the number of the models
173 participating in real-time MME operations.

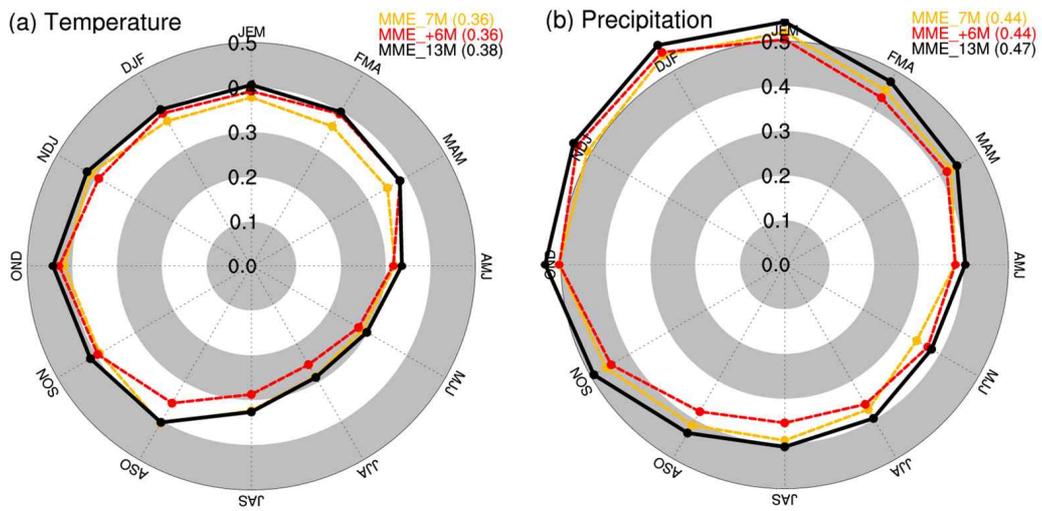
174 APCC considered several solutions to solve this hindcast issue and to take advantage of a
175 large set of models participating in the MME prediction. The first solution would be the use of
176 forecast anomalies with respect to climatologies estimated over the models' own hindcast
177 periods, which varies among the groups producing the model forecast, like IRI ENSO forecast.
178 That is, all the models can participate in the MME prediction by using forecast anomalies with
179 respect to different base periods and, consequently, to the different climatologies. However,
180 discrepancies may arise if the climatologies are significantly different. We have assessed
181 significance of the difference between climatologies estimated over two periods, 1983-2010 and
182 1993-2016, which cover the common hindcast period and the most recent hindcast period of the
183 14 models in v2019, based on the Student's *t*-test (Fig. S1). The results show that the differences
184 between the two climatologies of seasonal mean temperature in observations are statistically
185 significant in many regions. The most significant differences are evident in the high latitudes of
186 Northern and Southern Hemispheres throughout all seasons. These are the regions of the mostly
187 pronounced global warming accelerating in the recent years. It is also evident in the South Indian
188 Ocean in MAM and JJA and in the Western Pacific in SON and DJF. Furthermore, for the model
189 with the longest hindcast period spanning from the early 1980s to the most recent years the
190 differences between climatologies from the periods 1983-2010 and 1993-2016 are statistically
191 significant (not shown). Thus, the first solution may cause another issue in forecast anomalies
192 due to the significant differences in the climatologies because of the different reference
193 (hindcast) periods of individual models, and eventually in the MME prediction that combine the
194 forecast anomalies of individual models (Wallace and Arribas, 2012). Furthermore, this solution
195 does not suit the users of our seasonal forecasts. The users formulate their local forecasts in
196 terms of anomalies in respect to their local normals estimated over the 30-year periods
197 appointed/defined by WMO. As a rule, they make their local corrections to the MME forecasts
198 accounting for the difference between the local normals estimated over, e.g., 1991-2020 and
199 local MME climatology estimated over, e.g., 1983-2010. However, this solution does not provide
200 distinctness of the period of MME climatology that makes impossible any corrections.
201 Particularly, another issue arises as to which a reference period should be applied to observation
202 to assess the MME forecasts combined with models using different climatologies.

203 In this situation, we suggested an alternative solution that is to change the current
204 hindcast period to unified 1991-2010, for which almost all of the models are included. Although,

205 the 20-year climatology is shorter, it is comparable with the climatologies of other MME groups
206 (e.g., WMOLC (1993-2009; 17 years), C3S (1993-2016; 24 years). In case of models of CMCC
207 and HMC whose data start from 1993, it was treated as missing values of 1991-1992 to allow
208 more models to participate in MME and to extend the common hindcast period. To estimate the
209 forecast skill according to the changes in the number of participating models, we examine the
210 skill of MME hindcast (1991-2010) in three different model combinations within model suites of
211 2020 (v2020). Those are the seven models matched with the 1983-2010 hindcast period (+7M;
212 APCC, CWB, JMA, NASA, NCEP, PNU, MSC/ECCC); additional six models matched with the
213 1991-2010 (+6M; BCCM, HMC, KMA, UKMO, BOM, CMCC); the whole model set
214 comprising all 13 models(13M). The diagrams shown in Fig. 2 demonstrate that the skills of the
215 MMEs based on 7M (MME_7M) and +6M (MME_+6M) are comparable with each other,
216 showing ACC=0.36 (0.44) for annual mean temperature (precipitation) for both MMEs.
217 Meanwhile, the MME comprising all 13 models (MME_13M) definitely outperforms both
218 MME_7M and MME_+6M for both temperature and precipitation and for all 12 seasons
219 throughout the year. The skill improvement of MME_13M forecasts as compared with
220 MME_7M for both variables appears in most oceans and lands, only except for precipitation in
221 the Arctic region (Fig. 3). This improvement is mainly due to increase in the number of models
222 and corresponding increase of diversity of the contributing models and the relatively high skills
223 of newly contributing models (Yoo and Kang 2005; Alessandri et al., 2018; Fig. S2).

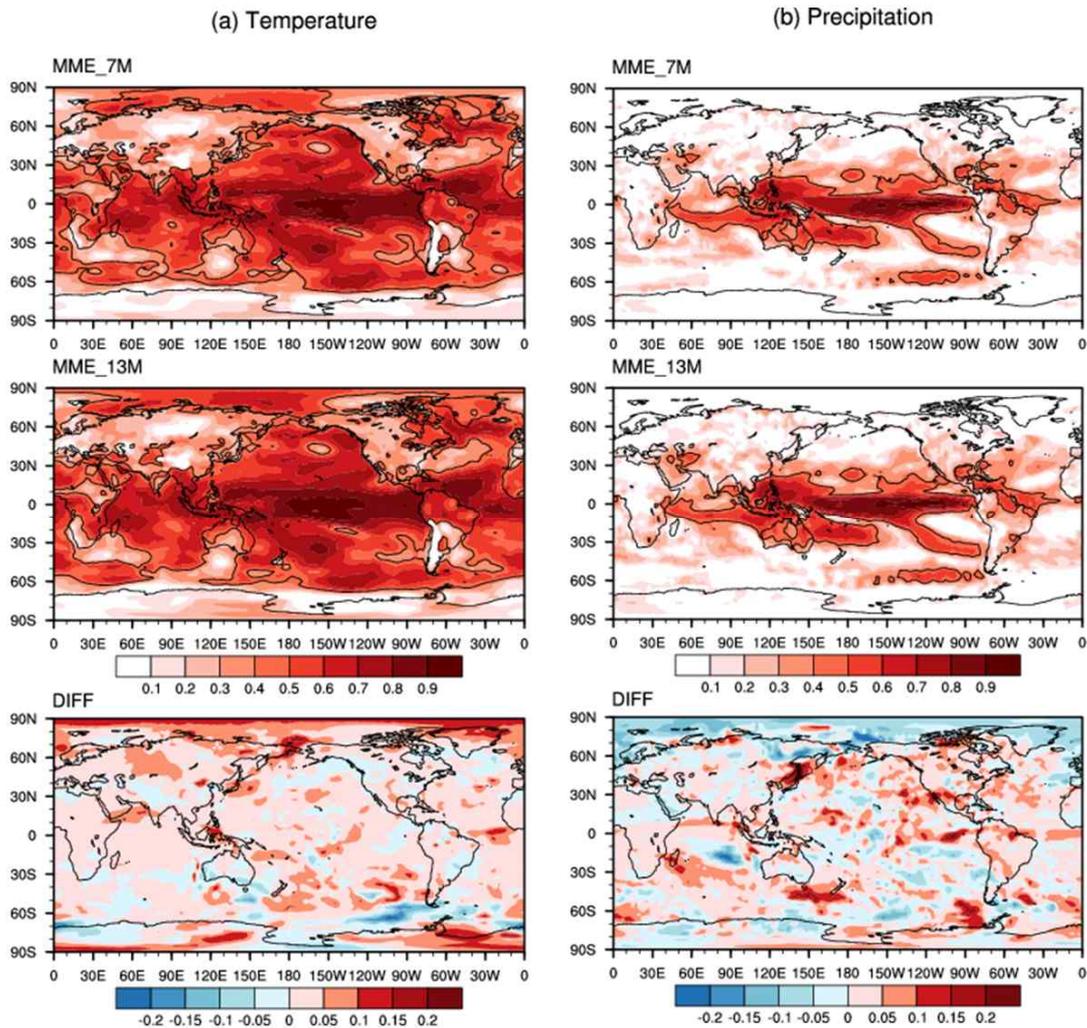
224 Based on the results from the hindcast experiments, we have changed the common base
225 period to 1991-2010 for APCC MME operations since 2020, which is covered by almost all the
226 models (Oper). Finally, we assessed the MME skill of the real-time forecasts for 2020JFM-
227 2022DJF because it is important for operational centers to assess whether the skill improvement
228 exists in real-time forecast as well as hindcast, although the 3-year periods is too short for the
229 collection of a sufficient number of real-time forecasts to obtain some well-grounded conclusions.
230 For comparison, we produced the real-time MME forecasts for the period of 2020-22 with the
231 models (Exp) that could participate if the hindcast period has not been changed. By changing the
232 hindcast period to 1991-2010, the number of participating models in the real-time MME
233 operations in 2020-22 increased by 100%, and recently the difference between Oper and Exp has
234 gradually widened (Fig. 4a). As a result, substantial improvement in temperature over the globe
235 is observed in recent years and the skill increases by approximately 2.8% for temperature (Fig.
236 4b). However, precipitation shows little change in a global scale. Based on the results from
237 hindcast and real-time forecasts, it turned out that the change of common hindcast period for
238 MME prediction in 2020 was an appropriate action for APCC operations.

239

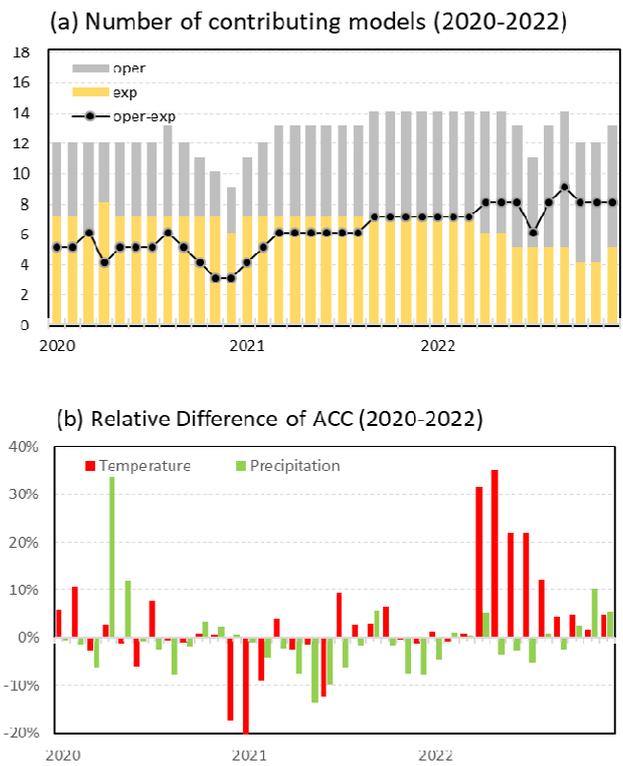


240

241 **Figure 2.** ACC for MME hindcast (1991-2010) with 7 models, additionally contributing 6
 242 models, and with 13 models of (a) seasonal mean temperature and (b) precipitation over globe.
 243 Annual mean ACC for each MME is shown in parentheses.
 244



246 **Figure 3.** Spatial distribution of 12-season averaged (annual mean) temporal correlation
 247 coefficients (TCCs) for MME hindcast (1991-2010) with 7 models and with 13 models of (a)
 248 seasonal mean temperature and (b) precipitation. The contour lines enclose the areas where the
 249 TCCs are statistically significant at the 5% level using the two-tailed Student’s t-test. The skill
 250 differences (DIFF) shows the difference two MMEs (MME_13M minus MME_7M).
 251
 252



254

255 **Figure 4.** (a) Number of participating models in real-time MME operations for 2020-22 and ratio
 256 of its change from Exp to Oper. (b) Relative difference of ACC from Exp to Oper for global
 257 temperature and precipitation.

258

259 5 Conclusions

260 Construction of the MME is a compromise between the number of participating models
 261 and length of a common hindcast period. Increase in the number of participating models with
 262 sufficient model diversity provided decreases random and model formulation errors. On the other
 263 hand, increase in the length of the common hindcast period decreases errors in climatology but
 264 increase random and model formulation errors in MME forecasts because of decrease in the
 265 number of participating models. In this situation, permanent development and improvement of
 266 the newly developed/improved models, being a sign of progress, is gradually destroying existing
 267 compromise and requires achievement of a new compromise with participation of the newer
 268 models. In the early of 2020s, APCC faced new challenges while maintaining a common
 269 hindcast period exceeding 20 years that maintained for many years, resulting that the proportion
 270 of the models that could not participate in the operational MME prediction due to their hindcast
 271 periods started in the late 1980s-early 1990s achieved about 50% by 2020. Based on the results
 272 of several experiments, we proposed a solution to change the common hindcast period to 1991-
 273 2010, which is the most appropriate way for APCC operation, reflecting the recently developed
 274 models. That is, by changing the reference period for MME prediction, APCC provides
 275 opportunities for participation in operational MME prediction system for the newly

276 developed/upgraded models, which results in a twofold increase in the number of participating
277 models and improvement of the forecast skill. Certainly, the payment for the increase in the
278 number of the participating models is shortening of the common hindcast period down to 20
279 years, which is less than 30 years recommended by WMO for climatology estimation but which
280 is comparable with other MME prediction producing centers, e.g., WMOLC and C3S.

281 Although not the scope of this study, the most important issue in recent years is that since
282 late 2021, the NMHSs worldwide have used the WMO recommended 1991-2020 normals
283 ([https://www.wmo.int/edistrib_exped/grp_prs/_en/08791-2019-CLW-CLPA-DMA-
284 CLIN8110_en.pdf](https://www.wmo.int/edistrib_exped/grp_prs/_en/08791-2019-CLW-CLPA-DMA-CLIN8110_en.pdf)). Meanwhile, there are still some limitations to match with the WMO
285 recommended period of normal and nowadays, no climate center providing MME seasonal
286 forecasts to the NHMSs uses climatology matching with the WMO references. The NMHSs
287 needs to adjust the model predicted departures from the model climatologies to departures from
288 the WMO normals. So, adjustment of the period of the MME climatology to the period of the
289 WMO normals would be one of the forthcoming research tasks.

290

291 **Acknowledgments**

292 This research was supported by APEC Climate Center. The authors acknowledge the APCC
293 MME Producing Centres (PCs) for making their hindcast/forecast data available for analysis and
294 the APCC for collecting and archiving them and for organizing APCC MME prediction. We also
295 thanks to the APCC MME PCs to discuss this issue together at the 3rd APCC MME Providers'
296 meeting in 2019.

297

298 **Data Availability Statement**

299 The single-model and MME predictions used in the study are available at the platform-based
300 climate data service, CLimate Information toolkit (CLIKs; <https://cliks.apcc21.org>). The NCEP-
301 DOE reanalysis 2 was obtained from the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA,
302 from their website (<https://psl.noaa.gov/data/gridded/index.html>). The CAMS OPI monthly
303 estimates were available at [https://www.cpc.ncep.noaa.gov/products/global_precip/html/wpage.
304 cams_opi.html](https://www.cpc.ncep.noaa.gov/products/global_precip/html/wpage.cams_opi.html).

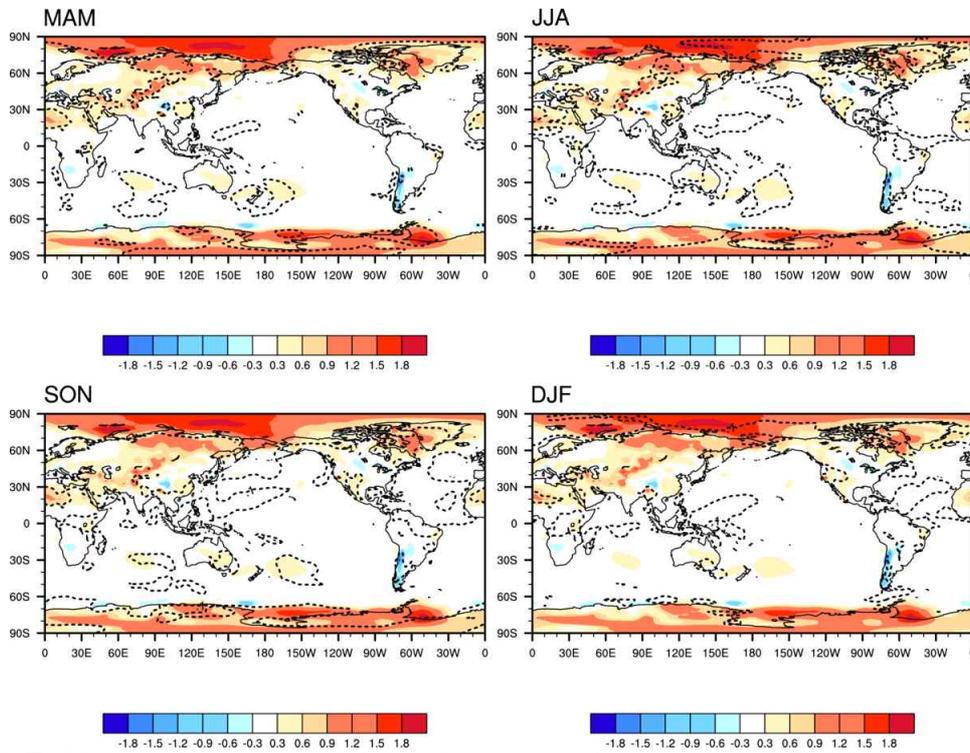
305

306 **References**

- 307 Alessandri, A., Catalano, F., Lee, J. Y., Wang, B., Lee, D. Y., Yoo, J. H., & Weisheimer, A.
308 (2018). Grand European and Asian-Pacific multi-model seasonal forecasts: maximization
309 of skill and of potential economical value to end-users. *Climate Dynamics*, 50, 2719-2738.
310 <https://doi.org/10.1007/s00382-017-3766-y>
- 311 Becker, E., den Dool, H. V., & Zhang, Q. (2014). Predictability and forecast skill in NMME.
312 *Journal of Climate*, 27(15), 5891-5906. <https://doi.org/10.1175/JCLI-D-13-00597.1>
- 313 Janowiak, J. E., & Xie, P. (1999). CAMS_OPI: a global satellite-raingauge merged product for
314 real-time precipitation monitoring applications. *Journal of Climatology*, 12:3335-3342.
315 [https://doi.org/10.1175/1520-0442\(1999\)012<3335:COAGSR>2.0.CO;2](https://doi.org/10.1175/1520-0442(1999)012<3335:COAGSR>2.0.CO;2)
- 316 Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.-K., Hnilo, J. J., Fiorino, M., & Potter, G. L.
317 (2002). NCEP-DOE AMIP-II Reanalysis (R-2). *Bulletin of the American Meteorological*
318 *Society*, 83:1631-1643. <https://doi.org/10.1175/Bams-83-11-1631>
- 319 Kim, G., Ahn, J. B., Kryjov, V. N., Lee, W. S., Kim, D. J., and Kumar, A (2021). Assessment of
320 multi-model ensemble methods for WMO LC-LRFMME. *International Journal of*
321 *Climatology*, 41, E2462-E248. <https://doi: 10.1002/joc.6858>.
- 322 Kirtman, B. P., Min, D., Infanti, J. M., et al. (2014). The North American Multimodel Ensemble:
323 Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal
324 Prediction. *Bulletin of the American Meteorological Society*, 95(4), 585-601.
325 <https://doi.org/10.1175/BAMS-D-12-00050.1>
- 326 Kryjov, V. N., & Min, Y.-M. (2016). Predictability of the wintertime Arctic Oscillation based on
327 autumn circulation. *International Journal of Climatology*, 36, 4181–4186,
328 <https://doi.org/10.1002/joc.4616>
- 329 Kumar, A., Chen, M., Zhang, L., Yang, W., Xue, Y., Wen, C., Marx, L., & Huang, B. (2012). An
330 analysis of the nonstationarity in the bias of sea surface temperature forecasts for the
331 NCEP Climate Forecast System (CFS) version 2. *Monthly Weather Review*, 140, 3003-
332 3016. <https://doi.org/10/1175/MWR-D-11-00335.1>
- 333 Lim, J., Dunstone, N. J., Scaife, A. A., and Smith D. M., (2019). Skillful seasonal prediction of
334 Korean winter temperature, *Atmospheric Science Letter*, 20, <https://doi.org/10.1002/asl.881>
- 335 Manzanas, R., Gutierrez, J. M., Bhend, K., Hemri, S., Doblas-Reyes, F. J., Torralba, V., Penabad,
336 E., & Brookshaw, A. (2019). Bias adjustment and ensemble recalibration methods for
337 seasonal forecasting: a comprehensive intercomparison using the C3S dataset. *Climate*
338 *Dynamics*, 53:1287-1305. <https://doi.org/10.1007/s00382-019-04640-4>
- 339 Min, Y. M., Kryjov, V. N., & Oh, S. M. (2014). Assessment of APCC multimodel ensemble
340 prediction in seasonal climate forecasting: Retrospective (1983-2003) and real-time

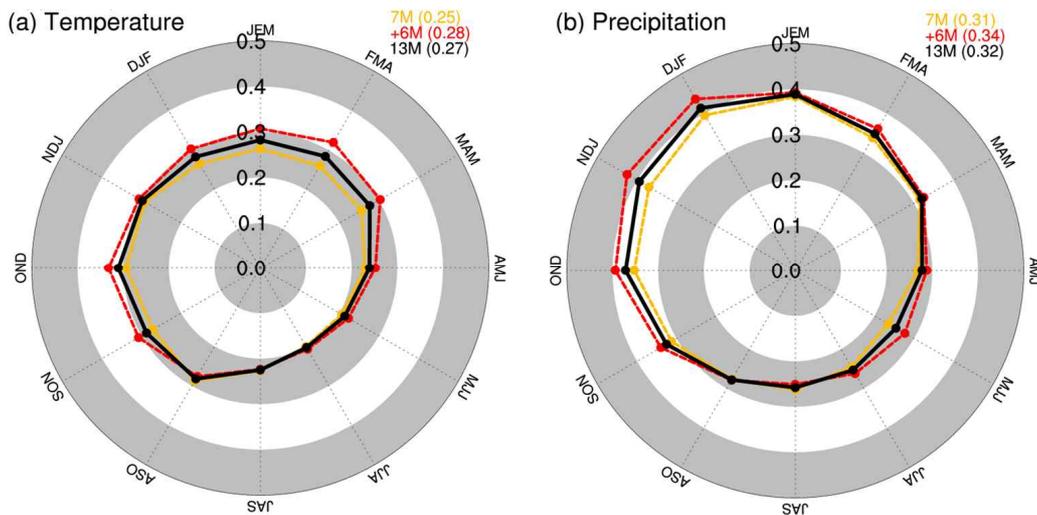
- 341 forecasts (2008-2013). *Journal of Geophysical Research: Atmospheres*, 199, 12,132-
342 12,150. <https://doi.org/10.1002/2014JD022230>
- 343 Min, Y. M., Kryjov, V. N., Oh, S. M., & Lee, H. J. (2017) Skill of real-time operational forecasts
344 with the APCC multi-model ensemble prediction system during the period 2008-2015.
345 *Climate Dynamics*, 49(11–12), 4141–4156. <https://doi.org/10.1007/s00382-017-3576-2>
- 346 Murphy, A. H. (1988). Skill scores based on the mean squared error and their relationships to the
347 correlation coefficient. *Mon. Weather Rev.*, 116, 2417–24. [https://doi.org/10.1175/1520-
348 0493\(1988\)116<2417:SSBOTM>2.0.CO;2](https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2)
- 349 Stockdale, T. N. (1997). Coupled ocean-atmosphere forecasts in the presence of climate drift.
350 *Monthly Weather Review*, 235, 809-818. [https://doi.org/10.1175/1520-
351 0493\(1997\)125<0809:COAFIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1997)125<0809:COAFIT>2.0.CO;2)
- 352 Wallace, E., Arribas, A. (2012). Forecasting with reference to a specific climatology. *Monthly*
353 *Weather Review*, 140, 3795-3802. <https://doi.org/10.1175/MWR-D-12-00159.1>
- 354 World Meteorological Organization (2007). *The role of climatological normals in a changing*
355 *climate*. WCDMP-No. 61, WMO/TD-No. 1377, Geneva, World Meteorological
356 Organization
- 357 Yoo, J. H., & Kang, I. S. (2005). Theoretical examination of a multi-model composite for
358 seasonal prediction. *Geophysical Research Letter*, 32, L18707.
359 <https://doi.org/10.1029/2005GL023513>
360

361



362
363
364
365
366
367
368

Figure S1. Differences between two climatologies over the period of 1979-2014 and 1993-2016 (dashed line) and trends of seasonal mean temperature for the whole 37-year period (shading). Differences and trends are only displayed at a 10% significance level.



369
370
371
372
373

Figure S2. The averaged ACC for hindcast (1991-2010) with 7 models, additionally contributing 6 models, and all 13 models of (a) seasonal mean temperature and (b) precipitation over globe. Annual mean ACC for the average of individual models' skill is shown in parentheses.