npj climate and atmospheric science

www.nature.com/npjclimatsci

ARTICLE OPEN

Check for updates

Seasonal-to-decadal prediction of El Niño–Southern Oscillation and Pacific Decadal Oscillation

Jung Choi[1] and Seok-Woo Son[1]✉

The growing demand for skillful near-term climate prediction encourages an improved prediction of low-frequency sea surface temperature (SST) variabilities such as the El Niño-Southern Oscillation (ENSO) and Pacific Decadal Oscillation (PDO). This study assesses their seasonal-to-decadal prediction skills using large ensembles of the Coupled Model Intercomparison Project phases 5 and 6 retrospective decadal predictions. A multi-model ensemble reforecast successfully predicts ENSO over a year in advance. While its seasonal prediction skill in the following spring and summer is achieved by multi-model ensemble averaging of relatively smaller ensemble members, the multi-year prediction of winter ENSO needs a larger ensemble size. The PDO is significantly predicted at a lead time of five-to-nine years but such a long-lead prediction is sourced from external radiative forcing instead of initialization, as evidenced from uninitialized historical simulations. The effect of model initialization lasts only two years. These results confirm that both the model initialization and the proper estimate of near-term radiative forcing are required to improve the seasonal-to-decadal prediction in the Pacific Basin.

npj Climate and Atmospheric Science (2022)5:29 ; https://doi.org/10.1038/s41612-022-00251-9

INTRODUCTION

Near-term climate predictions that incorporate seasonal-to-decadal (S2D) time scales have recently received much attention from policymakers, stakeholders, and the climate science community in the context of climate risk management. International effort to address the demand for skillful climate prediction in the forthcoming decade has elevated[1]. It is now well documented that near-term climate predictions are influenced not only by boundary conditions (mainly greenhouse gases and aerosol concentrations) but also by initial conditions (mainly the ocean state)[2]. The initialized prediction systems have shown a promise in improving the S2D prediction skill[3,4,5]. The multi-model ensembles initialized with observations, subject to bias adjustment, have reproduced more reliable climate predictions compared with uninitialized climate projections[6,7,8].

The primary source of S2D prediction skill is the low-frequency variability of sea surface temperature (SST), such as the El Niño-Southern Oscillation (ENSO), Pacific Decadal Oscillation (PDO), and Atlantic Multidecadal Oscillation (AMO)[8]. In this context, S2D predictions have often been examined to ascertain their ability to predict SST variability in each ocean basin[9,10,11,12]. For instance, a set of retrospective decadal predictions (or hindcasts) from the Coupled Model Intercomparison Project phase 5 (CMIP5)[13] and phase 6 (CMIP6)[14] were recently utilized to quantify the multi-decadal predictability in the North Atlantic[15,16]. However, skill in predicting low-frequency SST variability in the Pacific that is characterized by ENSO and PDO has not been reported for the latest decadal hindcasts thus far.

Given the importance of ENSO on global climate variability, various approaches have been adopted to improve its prediction skill[17,18]. It has been reported that the state-of-the-art dynamical models have high skill in predicting ENSO up to 12-month lead times[19]. Ham et al.[17] demonstrated that ENSO could be predicted 18 months in advance when incorporating a machine learning technique. This prediction skill is higher than that of the dynamical forecasting systems. Most studies have focused on ENSO prediction with a maximum lead time of one or one-and-a-half years. The multi-year ENSO prediction, albeit with a great societal need, has been considered only in a few studies[18,20,21,22,23]. These studies have focused only on extreme ENSO events or employed a single forecasting system to evaluate the multi-year ENSO prediction skill.

As the leading mode of decadal SST variability in the North Pacific, the PDO has been treated as a prominent source of decadal climate prediction in the pan-Pacific region[24]. The PDO is driven not simply by internal ocean variability but by various nonlinear processes that span tropical and extratropical ocean-atmosphere interactions[25]. Owing to this complexity, its prediction skill has typically been evaluated using ensemble forecasts. It has been reported that the PDO can be predicted several years in advance[26,27,28]. However, this result is based on a limited number of models because of the enormous computing resource requirement. Kim et al.[9] assessed the PDO prediction skill using seven CMIP5 decadal hindcasts; however, PDO prediction skill has not been assessed in the latest CMIP6 decadal hindcasts.

The present study revisits S2D predictions of ENSO and PDO. Unlike previous studies, a very large ensemble of CMIP5 and CMIP6 retrospective decadal predictions is utilized. In particular, decadal hindcast extending over half a century are used to reduce uncertainty. The prediction skill arising from model initialization is identified by comparing the initialized ensemble predictions to the uninitialized historical simulations. By taking advantage of large ensemble sizes, the relative importance of the ensemble size versus the multi-model ensemble average in predicting multi-year ENSO and PDO indices is also evaluated.

RESULTS

Global sea surface temperature

The overall prediction skills of the annual-mean multi-model ensemble (MME) SST anomaly (SSTA) at the first two years, three-to-four-year, and five-to-nine-year average (YR1, YR2, YR3–4, and

[1]School of Earth and Environmental Sciences, Seoul National University, Seoul, South Korea. ✉email: seokwooson@snu.ac.kr

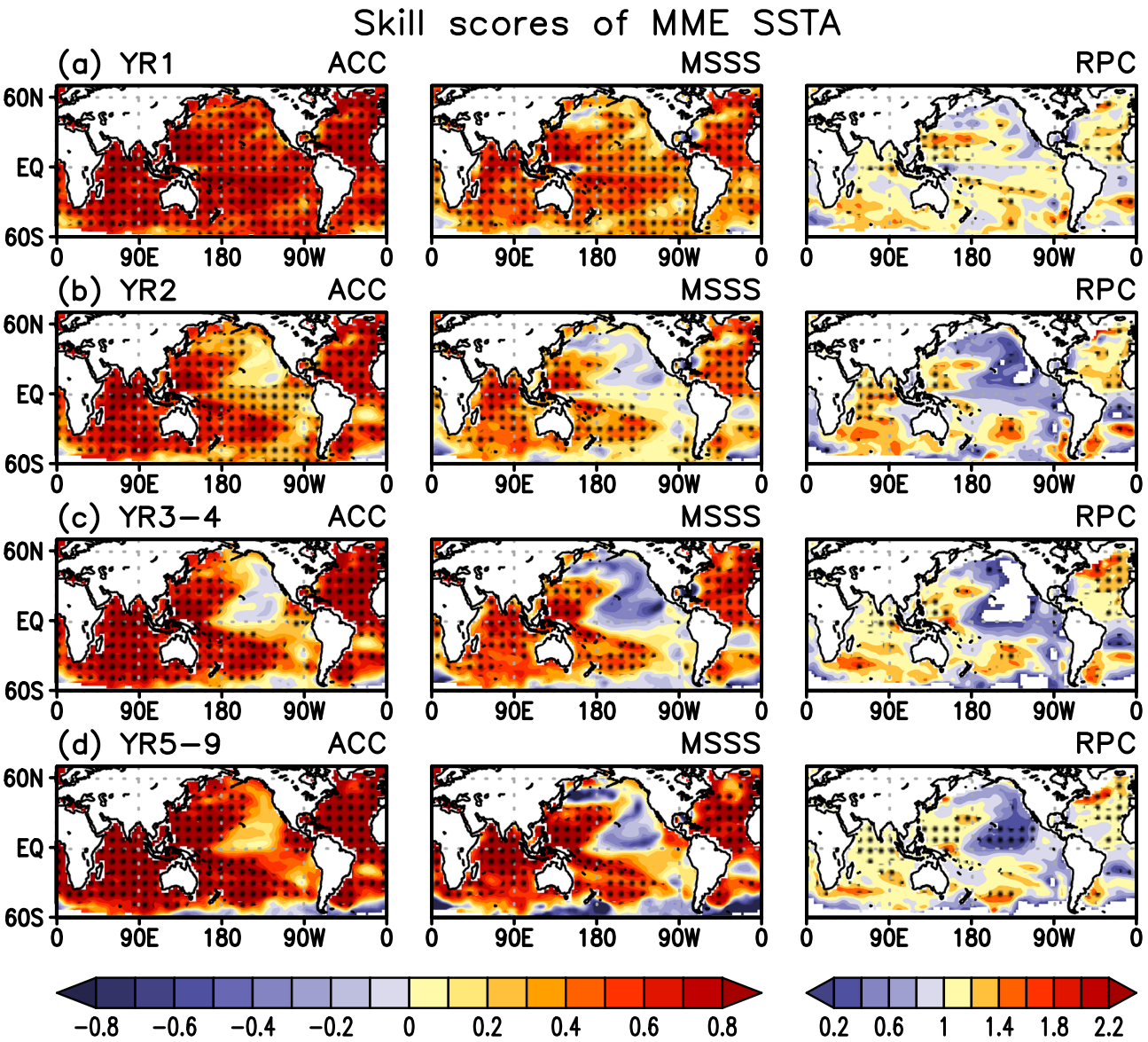

Fig. 1 Prediction skill of sea surface temperature. a ACC, MSSS, and RPC of MME SSTA predictions at a one-year lead time (YR1). Dotted regions denote where ACC and MSSS are significantly greater than 0, or RPC is significantly different to one at the 95% confidence level. Note that RPC is calculated only when ACC is positive. The regions with negative ACC are masked out as they imply no predictive skill. **b, c**, and **d** Same as (**a**) but for MME SSTA predictions at two-year (YR2), three-to-four-year (YR3–4), and five-to-nine-year (YR5–9) lead times, respectively. A non-parametric bootstrap method is applied to test statistical significance.

YR5–9, respectively) are shown in Fig. 1. At YR1, statistically significant prediction skills are found in almost all regions in both the anomaly correlation coefficient (ACC) and mean-squared skill score (MSSS) metrics. As lead time increases, high skill scores gradually decrease but not uniformly in space. For instance, the prediction error in the North Pacific arises from the western coast of North America and subsequently extends to the central Pacific Basin (Fig. 1b–c). The error also increases in the Southern Ocean. The MSSS decreases more rapidly compared to the ACC, and becomes statistically insignificant in the eastern North Pacific and the Southern Ocean at lead times longer than two years.

It is noticeable from Fig. 1c–d that the skill score maps are very similar between YR3–4 and YR5–9. The regions with high skill scores, such as the western Pacific, North Atlantic, and Indian Ocean, are persistently found. This is partly due to the long-term trend driven by external radiative forcing. As shown in Supplementary Fig. 1c–d, similar skill score maps are found at YR3–4 and YR5–9 in the uninitialized MME. More importantly, these regions appear even at YR1 in the uninitialized MME, confirming that it is driven by external radiative forcing.

To separate the prediction skill by external radiative forcing from the one by model initialization, the skill differences between the initialized and uninitialized MMEs are examined in Fig. 2. The model initialization results in overall skill improvement at YR1 (see the hatched regions). However, at YR2, its impact remains only in the limited regions such as the Indian Ocean and the tropical central Pacific. Except for these regions, the skill difference is rather small from YR2 to YR5–9. This result indicates that external radiative forcing, mostly due to greenhouse gas and aerosol concentration changes, is one of the major sources of interannual-to-decadal climate prediction[29,30,31]. It is noteworthy that both ACC and MSSS show significant differences at YR5–9 over the western coast of South America. Despite their differences, the skill scores of the initialized MME are still very low in this region (Fig. 1d).

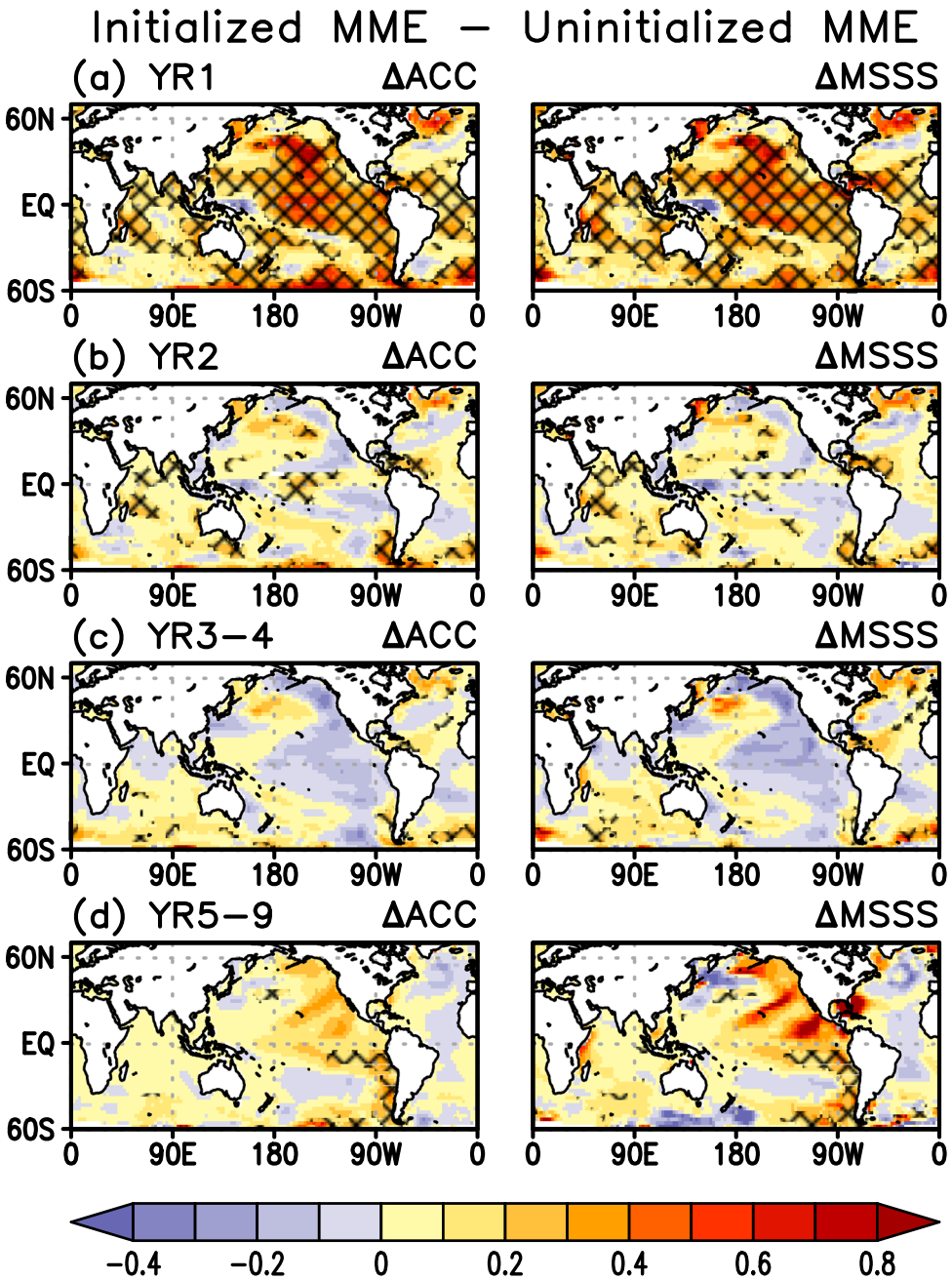

Fig. 2 Skill differences between the initialized and uninitialized MMEs. (left) ACC and (right) MSSS differences between the initialized and uninitialized MMEs of (**a**) YR1 SSTA, (**b**) YR2 SSTA, (**c**) YR3–4 SSTA, and (**d**) YR5–9 SSTA. The cross indicates that the positive skill difference (i.e., skill improvements due to model initialization) is statistically significant at the 95% confidence level. Note that the skill differences at YR5–9 in (**d**) are computed using only 46 initializations (see Supplementary Table 2 and Supplementary Fig. 5 for further details).

The errors in the signal-to-noise ratio are estimated with the ratio of predictable components (RPC; right panels in Fig. 1 and Supplementary Fig. 1). A statistically significant underconfident feature (RPC > 1, a low signal-to-noise ratio) is more widely observed than an overconfident feature (RPC < 1, a high signal-to-noise ratio) at YR1. For YR1 and YR2 SSTA prediction, only the initialized experiments show the significant RPC > 1 regions in the western North Pacific (compare Fig. 1a–b and Supplementary Fig. 1a–b). It indicates that this region's low signal-to-noise ratio likely results from erroneously capturing internal variability even in the initialized experiments. It is noticeable that the RPC maps are very similar between the initialized and uninitialized experiments for lead times longer than three years as in skill score maps. This result again suggests a critical role of external radiative forcing in predicting SSTA at interannual-to-decadal time scales.

ENSO

The prediction skill of three-month running mean NINO3.4 index is summarized in Fig. 3. The MME skill scores show a sharp decrease in the spring then a gradual increase in the winter, consistent with the well-known 'spring predictability barrier' for ENSO prediction[32]. Nevertheless, MME ACCs are statistically significant at the 95% confidence level at 25–27 lead months (January to March of the third year; JFM3). A statistically significant MME MSSS is also

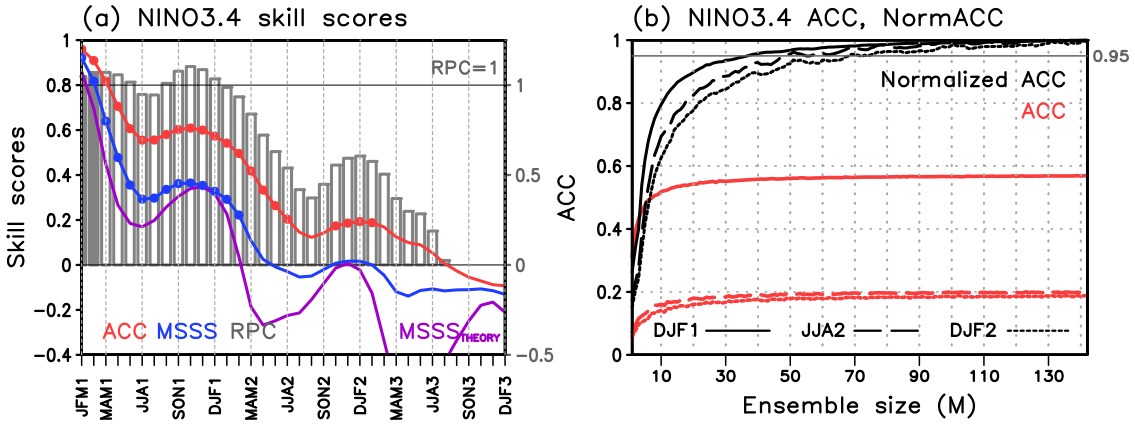

Fig. 3 Prediction skill of ENSO and ensemble size effect. a Skill scores of three-month running averaged NINO3.4 SSTA from MME prediction. ACC, MSSS, and RPC are shown by red, blue lines, and gray bars, respectively. The closed circles and filled bars denote the statistically significant skill scores (ACC and MSSS) and RPC at the 95% confidence level. The purple line represents the theoretical estimation of MSSS ($MSSS_{THEORY}$). **b** ACCs of DJF1, JJA2, and DJF2 (12–14, 18–20, and 24–26 lead months, respectively) NINO3.4 SSTA as a function of ensemble sizes (red lines). The black lines represent the normalized ACCs by the Min-Max normalization method. A five-member running average is applied to smooth the normalized ACCs.

found at 14–16 lead months in February to April of the second year (FMA2). Until these lead times, the RPC is close to one (gray bars in Fig. 3). These results indicate that ENSO is reliably predicted by the MME with a maximum lead time of 14–16 months.

Consistent with previous studies[10,19], the MME skill score is typically higher than individual models (Supplementary Fig. 2). The only exception is the CanCM4 which outperforms the MME ACC in one-year ENSO prediction. The CanCM4 shows a comparable ACC to the MME at relatively short lead times, but surpasses the MME at 8–10 lead months (ASO1) to 21–23 lead months (SON2). This superior performance of the CanCM4 is well documented in Barnston et al.[19]. In terms of MSSS, the highest skill score is found in the MME until the spring of the second year. The MSSS difference among the models is substantially large in the spring and summer, but becomes small in the winter.

As described in the Methods section, a higher prediction skill of the MME than individual models might result from a large ensemble size (i.e., the ensemble size effect) and/or a cancelation of model errors through multi-model ensemble average (i.e., the multi-model ensemble average effect). To test the ensemble size effect, the ACCs are presented in Fig. 3b as a function of ensemble sizes at lead times of DJF1, JJA2, and DJF2 (12–14, 18–20, and 24–26 lead months, respectively). The normalized ACC shows a rapid saturation ($M \sim 40$) with increasing ensemble sizes at DJF1. No significant skill difference is found between $M = 40$ and $M = 142$. This result indicates that only 40 ensemble members are needed for skillful one-year ENSO prediction. This number is comparable to the ensemble size for the climate model to robustly capture ENSO characteristics ($M > 50$) in Lee et al.[33]. However, almost twice as many ensemble members ($M > 70$) are required to extract the maximum prediction skill of the winter ENSO at a two-year lead time (see the normalized ACC of DJF2). It is noteworthy that although MME ACC in JJA2 is comparable to that in DJF2 (0.20 and 0.19, respectively), their normalized ACCs exhibit different curves with increasing ensemble sizes. For instance, the normalized ACC at JJA2 reaches 0.95 with 50–60 ensemble members, while that at DJF2 does so with 70–80 ensemble members. A faster skill saturation at JJA2 than at DJF2 suggests that a higher MME skill in this season is not simply caused by a large ensemble size.

To verify the multi-model ensemble average effect, the theoretically-estimated MSSS ($MSSS_{THEORY}$) in Eq. (8) is compared to the practically-calculated MSSS (Fig. 3a). It is found that MME MSSS is larger than $MSSS_{THEORY}$, particularly in the spring and summer. This indicates that the perfect model assumption is rejected, implying that the ensemble variance is not identical to the error variance of ensemble-mean prediction. It can be translated that the model-dependent errors are effectively reduced by the multi-model ensemble average. Dunstone et al.[34] recently showed that the combination of two prediction systems has a higher prediction skill at predicting the NINO3.4 index in the extended boreal summer from May to September (MJJAS) (7–11 lead months) than a single prediction system for any given ensemble size. Our result also suggests that multi-model ENSO prediction is better than the individual systems in the spring and summer when the ENSO prediction barrier appears.

The ENSO prediction skills are also computed by taking the two different MME methods—1) MME by considering all ensemble members with an equal weighting (same to Fig. 3) and 2) MME computed from each model's ensemble mean. The results are not sensitive to the MME methods (Supplementary Fig. 3a). The skill scores are also separately computed for the CMIP5 and CMIP6 MMEs (compare red and blue lines in Supplementary Fig. 3a). Although they do not show a substantial difference at the one-year lead time, the springtime skill scores in the second year are slightly improved in the CMIP6, especially for predicting ENSO amplitude measured by MSSS. It may result from a larger ensemble size used for CMIP6 than CMIP5 MMEs.

PDO

The PDO prediction skill is illustrated in Fig. 4 as a function of forecast lead times. The MME ACC of PDO index is statistically significant at the 95% confidence level at YR1, YR2, and YR5–9. As shown in Fig. 2, the PDO prediction skill at lead times longer than three years is mostly due to external radiative forcing. This result indicates that the PDO prediction skill is determined by both model initialization at a short lead time and external radiative forcing at a long lead time. However, the MSSS score is statistically significant only at YR1. This indicates that only the PDO phase is qualitatively predicted in the current S2D prediction systems. All RPCs are lower than one, indicating an overconfident prediction with a high signal-to-noise ratio, but they are not statistically different from one at the 95% confidence level.

The sensitivity of the PDO ACCs with ensemble sizes is demonstrated in Fig. 4b. At YR1, the skill saturation of PDO is somewhat faster than that of ENSO. The normalized ACC reaches 0.95 with 30 ensemble members (solid black line). At YR2, more ensemble members are needed for skill saturation. The normalized

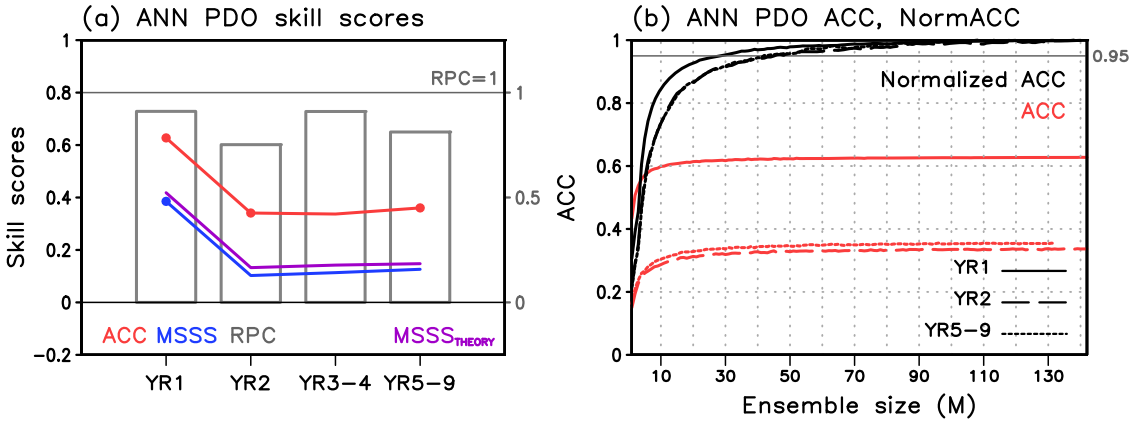

Fig. 4 Prediction skill of PDO and ensemble size effect. Same as Fig. 3 but for the annual-mean PDO index. Note that significantly different RPC from one is not found in (**a**). (**b**) ACCs (red) and normalized ACCs (black) of annual-mean PDO index at one-year (YR1), two-year (YR2), and five-to-nine-year (YR5–9) lead times as a function of ensemble sizes. Only the statistically significant ACCs are tested. The normalized ACCs at YR2 and YR5–9 overlap with each other.

ACC of the YR2 PDO index reaches 0.95 when 40–50 ensemble members are used. A similar ensemble size is needed for skill saturation at YR5–9. Note a relatively small ensemble size for PDO prediction skill saturation compared to ENSO. It may result from the reduced noise in the annual or multi-year mean PDO index.

The $MSSS_{THEORY}$ of the PDO index is quite similar to the practical MSSS (compare purple and blue solid lines in Fig. 4a). It indicates that the multi-model PDO prediction does not always guarantee a high skill than a single model prediction when their ensemble sizes are comparable. It may hint that there are common errors in predicting PDO in the current S2D prediction models. In accordance with Kim et al.[35], the multimodel ensemble average effect is not significant in addressing systematic errors over the North Pacific.

The PDO skill scores are also separately evaluated for the CMIP5 and CMIP6 MMEs (Supplementary Fig. 3b). At YR1 and YR2, the CMIP5 MME shows higher skill scores than the CMIP6 MME. It is partly explained by the fact that two CMIP5 models (CanCM4 and MIROC5) surpass the MME at these lead times (Supplementary Fig. 4). At longer lead times, they show comparable skill scores, indicating no major improvement in PDO prediction in the latest S2D prediction systems.

SUMMARY AND DISCUSSION

The present study evaluates the S2D prediction skills of the Pacific SSTA, focusing on ENSO and PDO. Two deterministic metrics (ACC and MSSS) and RPC are utilized to assess the prediction skill and its spread. The results are compared to those of the uninitialized experiments to quantify the skill improvement due to the model initialization versus external radiative forcing.

The overall skill scores of MME SSTA decrease inhomogeneously in space as the lead time increases. The skill scores rapidly decline in the regions where no skills are found in the uninitialized MME (e.g., the eastern Pacific coast and the Southern Ocean). The model initialization results in successful SSTA predictions at lead times of up to two years, particularly in the tropical Pacific. However, the skillful prediction at lead times longer than three years is mostly explained by external radiative forcing. In the regions where external radiative forcing is important, a large inter-member diversity in predicting SSTA with a low signal-to-noise ratio (RPC > 1) is also found at decadal time scales. This result indicates that both the initialization and the proper estimate of near-term radiative forcing are required to improve the S2D prediction.

For ENSO prediction, the skill score is significantly high at a lead time of one year in predicting the ENSO amplitude and at two years in predicting the ENSO phase, respectively. The MME skill scores are higher than individual models. The improved MME ENSO prediction skill in the boreal spring and summer with 40 ensemble members is partly related to the cancelation of model errors through the multimodel ensemble average effect. The theoretically estimated prediction skills below the practical ones confirm that the multimodel prediction system has the potential to mitigate the spring predictability barrier of ENSO. It is also found that multi-year ENSO prediction requires a larger ensemble size, more than 70 ensemble members. Unlike ENSO, the PDO prediction skill follows the theoretical estimation, indicating that individual model errors are not effectively eliminated by the multimodel ensemble average. More than 50 ensemble members are needed to extract the maximum prediction skill of YR5–9 PDO index despite the skill score is not very high.

The PDO prediction skill is considerably more limited than the AMO prediction[9] because of its shorter timescale. The physically indistinguishable resemblance between the ENSO and PDO[25,36] and systematic model bias in the extratropical Pacific Basin[37] may also contribute to the lower prediction skill of the PDO. Regardless of its cause(s), the limited PDO prediction skill implies that the PDO-related regional and global climate variabilities may not be well represented in the decadal prediction systems. In this regard, a sophisticated and advanced post-processing technique, applied to improve the AMO prediction[15] and/or a statistical error correction[35], may help to improve the PDO prediction.

The CMIP6 DCPP provides various experiments that are not explored in this study. Further studies using additional experiments, especially those designed to investigate the idealized impacts of the Pacific and Atlantic decadal variability, would be desirable. This would provide an opportunity to improve our understanding of S2D climate variability and its prediction.

METHODS

Data

Six CMIP5 and ten CMIP6 models are used in this study (Supplementary Table 1). Only the retrospective predictions (or hindcasts) initialized every year from 1960/1961 winter to 2009/2010 winter are used. Each model consists of 50 initializations with three to ten ensemble members. In total, 142 ensemble members are available. Among 16 models, seven models are initialized in January. The other nine models are initialized in November. Despite this difference in initialized months, the forecast lead time of multimodel ensemble is counted from January. For example, one-month and one-year lead times denote January and January-December average, respectively (see Supplementary Table 2).

The individual ensemble members are interpolated to a uniform 2.5° × 2.5° grid and are considered independently. To remove the model drift,

model bias is defined as the difference between the modeled and observed climatology at each lead time. Then, the lead time-dependent model bias is corrected to individual ensemble members before the analysis[38]. The multi-model ensemble (MME) is then calculated by averaging bias-corrected ensemble members with an equal weighting. The prediction skill is quantified by comparing the MME with the observed SST anomaly (SSTA) of the National Oceanic and Atmospheric Administration (NOAA) Extended Reconstructed Sea Surface Temperature version 5 (ERSSTv5)[39], interpolated to a uniform 2.5° × 2.5° grid. The anomaly at each grid point is defined by the deviation from the monthly climatology over the period of 1961–2020.

The ENSO prediction skill is evaluated by computing three-month running mean NINO3.4 index (SSTA averaged over 170°–120°W and 5° S–5° N) for lead times up to three years. The PDO prediction skill is computed with the annual-mean PDO index. The PDO index is calculated as follows. The observed PDO pattern is first defined as the leading empirical orthogonal function (EOF) of monthly ERSSTv5 SSTA in the North Pacific (poleward of 20°N) for 1961–2020. Global-mean SSTA is removed at each grid point before conducting the EOF analysis. Since the leading EOF pattern in models often does not agree with the observed EOF pattern[40], we employed the common basis function approach by projecting model anomalies onto the observed pattern. In other words, the annual-mean PDO index in each hindcast is derived by projecting the bias-corrected annual-mean SSTA of each hindcast onto the observed PDO pattern. The resultant PDO index is evaluated for the first two years, three-to-four-year, and five-to-nine-year average (YR1, YR2, YR3–4, and YR5–9, respectively). While the first three predictions (YR1, YR2, and YR3–4) are tested with all 142 ensemble members, the YR5–9 prediction is tested with 132 ensemble members as the MRI-ESM2-0 provides only five-year-long hindcasts.

In order to estimate prediction skills originating from external radiative forcing, the MME of uninitialized historical simulations is also examined. A total of 121 ensemble members from ten CMIP6 models are used (Supplementary Table 1). Since the historical simulations cover the period up to 2014, the comparison to the ensemble hindcasts is limited to 1961–2014 for YR5–9 (see Supplementary Table 2 for further details).

Measure of prediction skill and spread

The prediction skill is quantitatively evaluated using the two deterministic metrics. Following the US CLIVAR decadal predictability working group[41], the anomaly correlation coefficient (ACC) and mean-squared skill score (MSSS) are particularly used. These metrics are defined as follows:

$$\mathrm{ACC}(\tau) = \frac{\frac{1}{n}\sum_{j=1}^{n}(H_{j\tau} - \overline{H}_\tau)(O_{j\tau} - \overline{O}_\tau)}{\sqrt{\frac{1}{n}\sum_{j=1}^{n}(H_{j\tau} - \overline{H}_\tau)^2}\sqrt{\frac{1}{n}\sum_{j=1}^{n}(O_{j\tau} - \overline{O}_\tau)^2}} \quad (1)$$

where O and H are the observations and a set of hindcasts, respectively. The subscript j is the initialization year, n is the total number of initializations ($n = 50$), and τ is the forecast lead time. The climatological averages of the observation and hindcasts are computed by

$$\overline{O}_\tau = \frac{1}{n}\sum_{j=1}^{n} O_{j\tau} \text{ and } \overline{H}_\tau = \frac{1}{n}\sum_{j=1}^{n} H_{j\tau} \quad (2)$$

The MSSS is based on the mean squared error (MSE). It is obtained as

$$\mathrm{MSSS}(\tau) = 1 - \frac{\mathrm{MSE}_H(\tau)}{\mathrm{MSE}_{\overline{O}}(\tau)} \quad (3)$$

where

$$\mathrm{MSE}_H(\tau) = \frac{1}{n}\sum_{j=1}^{n}[(H_{j\tau} - \overline{H}_\tau) - (O_{j\tau} - \overline{O}_\tau)]^2 \text{ and } \mathrm{MSE}_{\overline{O}}(\tau) = \frac{1}{n}\sum_{j=1}^{n}(O_{j\tau} - \overline{O}_\tau)^2 \quad (4)$$

The ACC estimates the linear association between the observations and hindcasts (i.e., the phase of variability), while the MSSS further quantifies the variance ratio of the hindcasts relative to the observations (i.e., the amplitude of variability). Therefore, the MSSS is a more constrained metric than the ACC.

The ratio of predictable components (RPC) is also computed to measure the ensemble spread[42]. The RPC is defined by comparing the predictable component in the observations (PC_O) and that in the hindcasts (PC_H).

$$\mathrm{RPC}(\tau) = \frac{\mathrm{PC}_O}{\mathrm{PC}_H} \geq \frac{\mathrm{ACC}(\tau)}{\sigma_{signal}/\sigma_{total}} \quad (5)$$

where σ_{signal} and σ_{total} are the expected standard deviations of the predictable signal and the total variability in the hindcasts, respectively. They are computed from the ensemble mean hindcasts and the individual members at lead time τ. Since PC_O cannot be directly estimated, its lower bound is used from the ACC as the ACC^2 reflects the proportion of the observed variance accounted for by the model hindcasts. For a perfect forecast, the RPC is expected to be one. A small RPC of <1 indicates an overconfident forecast, in which individual ensemble members agree well with each other (high signal-to-noise ratio) but the ensemble mean forecasts do not capture the observed variations (low correlation). A large RPC of >1 denotes an underconfident forecast, in which the ensemble mean forecast agrees well with the observed variations (high correlation) but individual ensemble members do not agree well with each other (low signal-to-noise ratio)[42]. In this study, RPC is computed only when ACC is positive.

Significance

A non-parametric bootstrap resampling method, without any assumptions on data distribution, is utilized to test the statistical significance of the prediction skill[41]. This method allows for a more objective criterion to determine a practically useful skill score than a certain threshold value. An additional 1,000 bootstrapped hindcasts are created from a finite ensemble size M (=142) and a finite number of validation years N (=50), as follows. 1) The N cases are randomly selected with replacement. To take autocorrelation into account, this is done in blocks of five consecutive years. 2) For each case, M ensemble members are randomly selected with replacement. The ensemble mean is then computed from these M samples. 3) The evaluation metrics (i.e., ACC, MSSS, RPC, and skill difference) are computed with the resultant ensemble mean prediction. 4) The steps 1)-3) are repeated 1,000 times to create a sample distribution of the evaluation metrics.

For the ACC and MSSS, the p-value is defined as the ratio of negative value from the bootstrapped sample distribution on the basis of a one-tailed test of the hypothesis that the prediction skill is greater than 0. Similarly, only the positive skill difference (initialized MME minus uninitialized MME) is assessed as it represents an improvement by the model initialization. For example, if the p-value is smaller than or equal to 0.05, the skill score and skill difference are set to be statistically significant at the 95% confidence level. The RPC is significantly different from one at the 95% confidence level when the 2.5–97.5% percentile, obtained from the bootstrapped sample distribution, does not cross one[15,42].

Impact of the ensemble size and multi-model ensemble average

The sensitivity of the prediction skill to the ensemble size is also evaluated. The ACCs are first calculated as a function of ensemble sizes using a bootstrap resampling methodology. For each validation year, M members are randomly selected with replacement. The ACC is then computed with the resultant ensemble mean. This resampling is repeated 1,000 times for each M (here, M varies from 1 to 142). An average of 1,000 bootstrapped ACCs with an ensemble size M is denoted as ACC_M.

To determine the minimum ensemble size when the prediction skill is saturated, the ACCs are standardized by the Min-Max normalization as

$$\text{Normalized } \mathrm{ACC}_M = \frac{\mathrm{ACC}_M - \mathrm{ACC}_{\mathrm{Min}}}{\mathrm{ACC}_{\mathrm{Max}} - \mathrm{ACC}_{\mathrm{Min}}} \quad (6)$$

where $\mathrm{ACC}_{\mathrm{Min}}$ and $\mathrm{ACC}_{\mathrm{Max}}$ are defined by $\mathrm{ACC}_{M=1}$ and $\mathrm{ACC}_{M=142}$, respectively. The normalized ACC_M gets transformed to a value between 0 to 1. The minimum ensemble size is defined when the normalized ACC_M reaches 0.95.

In order to quantify the impacts of the ensemble average on the prediction skill, the theoretical estimation of the ensemble-mean MSE is calculated following Murphy[43] under perfect model conditions. Mathematically, the error variance of an ensemble-mean hindcast ($\overline{H}_M$) can be written with the averaged error variance of individual ensemble members (H_i) as follows:

$$\mathrm{MSE}_{\overline{H}_M}(\tau) = \left(\frac{M+1}{2M}\right)\langle \mathrm{MSE}_{H_i}(\tau)\rangle \quad (7)$$

where <> denotes an average of MSE over individual ensemble members, i.e., $\frac{1}{M}\sum_{i=1}^{M}\mathrm{MSE}_{H_i}$. The $\mathrm{MSE}_{\overline{H}_M}(\tau)$ is smaller than $\langle \mathrm{MSE}_{H_i}(\tau)\rangle$ when $M>1$, indicating that the ensemble-mean hindcast is theoretically superior to the individual hindcasts. For the infinite ensemble size ($M\to\infty$), the error variance of the ensemble-mean hindcast becomes proportional to half of the averaged error variance of individual ensemble members.

Equation (7) indicates that the prediction error can be reduced by increasing the ensemble size M. It can also be reduced by canceling the model's systematic errors through the ensemble average. In terms of the prediction skill, the skill score increases by increasing the ensemble size and decreasing the model errors by averaging many ensemble members. Note that the former, the so-called ensemble size effect, is not necessarily related with the latter, the so-called ensemble average effect, which cannot be expressed as a function of the ensemble size.

The formulation described above can also be written in terms of the MSSS:

$$\begin{aligned}\mathrm{MSSS}_{\overline{H}_M}(\tau) &= 1-\frac{\mathrm{MSE}_{\overline{H}_M}(\tau)}{\mathrm{MSE}_{\overline{O}}(\tau)} = 1-\left(\frac{M+1}{2M}\right)\frac{\langle \mathrm{MSE}_{H_i}(\tau)\rangle}{\mathrm{MSE}_{\overline{O}}(\tau)}\\ &= 1-\left(\frac{M+1}{2M}\right)(1-\langle \mathrm{MSSS}_{H_i}(\tau)\rangle) = MSSS_{THEORY}\end{aligned} \tag{8}$$

When M is set to 142 (132 for YR5–9), the theoretical estimation of MSSS ($\mathrm{MSSS}_{\mathrm{THEORY}}$) can be obtained from the average value of individual ensemble members' MSSSs. Here it should be stated that $\mathrm{MSSS}_{\mathrm{THEORY}}$ is not necessarily the upper limit of model prediction skill when individual ensemble members and their average have a relatively poor skill (i.e., small $<\mathrm{MSSS}_{H_i}>$). Since this theoretical estimation assumes that the individual ensemble members are composed of the same model, a practical ensemble-mean prediction skill could become higher than $\mathrm{MSSS}_{\mathrm{THEORY}}$ if the model errors are corrected by the *multi-model* ensemble average.

DATA AVAILABILITY

The original CMIP5/6 database can be downloaded from the Earth System Grid Federation (ESGF) server (https://esgf-node.llnl.gov). The ERSSTv5 is obtained from the NOAA web interface (https://www.ncei.noaa.gov/products/extended-reconstructed-sst).

CODE AVAILABILITY

Methods have been fully mentioned in the Methods section and codes are available upon reasonable request from jungchoi@snu.ac.kr.

Received: 6 July 2021; Accepted: 21 February 2022;
Published online: 14 April 2022

REFERENCES

1. Kushnir, Y. et al. Towards operational predictions of the near-term climate. *Nat. Clim. Change* **9**, 94–101 (2019).
2. Kirtman, B. et al. *Near-term Climate Change: Projections and Predictability. In: Climate Change 2013*: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA (2013).
3. Smith, D. M. et al. Improved surface temperature prediction for the coming decade from a global climate model. *Science* **317**, 796–799 (2007).
4. Keenlyside, N. et al. Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature* **453**, 84–88 (2008).
5. Mochizuki, T. et al. Pacific decadal oscillation hindcasts relevant to near-term climate prediction. *Proc. Natl Acad. Sci.* **107**, 1833–1837 (2010).
6. Doblas-Reyes, F. et al. Initialized near-term regional climate change prediction. *Nat. Comm.* **4**, 1715 (2013).
7. Meehl, G. A. & Teng, H. CMIP5 multi-model hindcasts for the mid-1970s shift and early 2000s hiatus and predictions for 2016–2035. *Geophys. Res. Lett.* **41**, 1711–1716 (2014).
8. Meehl, G. A. et al. Initialized Earth System prediction from subseasonal to decadal timescales. *Nat. Rev. Earth. Environ.* **2**, 340–357 (2021).
9. Kim, H.-M., Webster, P. J. & Curry, J. A. Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts. *Geophys. Res. Lett.* **39**, L10701 (2012).
10. Choi, J. et al. Seasonal-to-Interannual Prediction Skills of Near-Surface Air Temperature in the CMIP5 Decadal Hindcast Experiments. *J. Clim.* **29**, 1511–1527 (2016a).
11. Choi, J. et al. Potential for long-lead prediction of the western North Pacific monsoon circulation beyond seasonal time scales. *Geophys. Res. Lett.* **43**, 1736–1743 (2016b).
12. Smith, D. M. et al. Robust skill of decadal climate predictions. *Npj Clim. Atmos. Sci.* **2**, 13 (2019).
13. Meehl, G. A. et al. Decadal Climate Prediction: An Update from the Trenches. *Bull. Am. Meteor. Soc.* **95**, 243–267 (2014).
14. Boer, G. J. et al. The Decadal Climate Prediction Project (DCPP) contribution to CMIP6. *Geosci. Model Dev.* **9**, 3751–3777 (2016).
15. Smith, D. M. et al. North Atlantic climate far more predictable than models imply. *Nature* **583**, 796–800 (2020).
16. Borchert, L. F. et al. Improved decadal predictions of North Atlantic subpolar gyre SST in CMIP6. *Geophy. Res. Lett.* **48**, e2020GL091307 (2021).
17. Ham, Y.-G., Kim, J.-H. & Luo, J.-J. Deep learning for multi-year ENSO forecasts. *Nature* **573**, 568–572 (2019).
18. Petrova, D., Ballester, J., Koopman, S. J. & Rodó Multiyear Statistical Prediction of ENSO Enhanced by the Tropical Pacific Observing System. *J. Clim.* **33**, 163–174 (2020).
19. Barnston, A. G. et al. Deterministic skill of ENSO predictions from the North American Multimodel Ensemble. *Clim. Dyn.* **53**, 7215–7234 (2019).
20. Chen, D. et al. Predictability of El Niño over the past 148 years. *Nature* **428**, 733–736 (2004).
21. Luo, J.-J., Masson, S., Behera, S. K. & Yamagata, T. Extended ENSO Predictions Using a Fully Coupled Ocean–Atmosphere Model. *J. Clim.* **21**, 84–93 (2008).
22. Chikamoto, Y. et al. Skilful multi-year predictions of tropical trans-basin climate variability. *Nat. Comm.* **6**, 6869 (2015).
23. Luo, J.-J., Liu, G., Hendon, H., Alves, O. & Yamagata, T. Inter-basin sources for two-year predictability of the multi-year La Niña event in 2010–2012. *Sci. Rep.* **7**, 2276 (2017).
24. Mantua, N. J. et al. A Pacific Interdecadal Climate Oscillation with Impacts on Salmon Production. *Bull. Am. Meteor. Soc.* **78**, 1069–1080 (1997).
25. Newman, M. et al. The Pacific Decadal Oscillation, Revisited. *J. Clim.* **29**, 4399–4427 (2016).
26. Mehta, V. M., Mendoza, K. & Wang, H. Predictability of phases and magnitudes of natural decadal climate variability phenomena in CMIP5 experiments with the UKMO HadCM3, GFDL-CM2.1, NCAR-CCSM4, and MIROC5 global earth system models. *Clim. Dyn.* **52**, 3255–3275 (2019).
27. Wiegand, K. N., Brune, S. & Baehr, J. Predictability of multiyear trends of the Pacific Decadal Oscillation in an MPI-ESM hindcast ensemble. *Geophy. Res. Lett.* **46**, 318–325 (2019).
28. Boer, G. J. & Sospedra-Alfonso, R. Assessing the skill of the Pacific Decadal Oscillation (PDO) in a decadal prediction experiment. *Clim. Dyn.* **53**, 5763–5775 (2019).
29. Doblas-Reyes, F. J., Hagedorn, R., Palmer, T. N. & Morcrette, J. J. Impact of increasing greenhouse gas concentrations in seasonal ensemble forecasts. *Geophys. Res. Lett.* **33**, L07708 (2006).
30. Yeager, S. G. et al. Predicting Near-Term Changes in the Earth System: A Large Ensemble of Initialized Decadal Prediction Simulations Using the Community Earth System Model. *Bull. Am. Meteor. Soc.* **99**, 1867–1886 (2018).
31. Klavans, J. M., Cane, M. A., Clement, A. C. & Murphy, L. N. NAO predictability from external forcing in the late 20th century. *Npj Clim. Atmos. Sci.* **3**, 22 (2021).
32. Webster, P. J. & Yang, S. Monsoon and ENSO: Selectively Interactive Systems. *Quart. J. Roy. Meteor. Soc.* **118**, 877–926 (1992).
33. Lee, J. et al. Robust evaluation of ENSO in climate models: How many ensemble members are needed? *Geophys. Res. Lett.* **48**, e2021GL095041 (2021).
34. Dunstone, N. et al. Skilful interannual climate prediction from two large initialised model ensembles. *Environ. Res. Lett.* **15**, 094083 (2020).
35. Kim, H.-M., Ham, Y.-G. & Scaife, A. A. Improvement of Initialized Decadal Predictions over the North Pacific Ocean by Systematic Anomaly Pattern Correction. *J. Clim.* **27**, 5148–5162 (2014).
36. Lienert, F. & Doblas-Reyes, F. J. Decadal prediction of interannual tropical and North Pacific sea surface temperature. *J. Geophy. Res.: Atmosphere* **118**, 5913–5922 (2013).
37. Henley, B. J. et al. Spatial and temporal agreement in climate model simulations of the Interdecadal Pacific Oscillation. *Environ. Res. Lett.* **12**, 044011 (2017).
38. ICPO, 2011: *Data and Bias Correction for Decadal Climate Predictions*. CMIP-WGCM-WGSIP Decadal Climate Prediction Panel. International CLIVAR Project Office Publication Series No. 150, 5pp. https://www.wcrp-climate.org/images/key_deliverables/decadal_prediction/documents/DCPP_Bias_Correction.pdf
39. Huang, B. et al. Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5): Upgrades, Validations, and Intercomparisons. *J. Clim.* **30**, 8179–8205 (2017).
40. Lee, J. et al. Quantifying the agreement between observed and simulated extratropical modes of interannual variability. *Clim. Dyn.* **52**, 4057–4089 (2019).

41. Goddard, L. et al. A verification framework for interannual-to-decadal predictions experiments. *Clim. Dyn.* **40**, 245–272 (2013).
42. Eade, R. et al. Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophys. Res. Lett.* **41**, 5620–5628 (2014).
43. Murphy, J. M. The impact of ensemble forecasts on predictablity. *Quart. J. Roy. Meteor. Soc.* **114**, 463–493 (1988).

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2017R1E1A1A01074889) and the Ministry of Education (NRF-2019R1I1A1A01057039). We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP5/6. We thank the climate modeling groups (listed in Supplementary Table 1 of this paper) for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP5/6 and ESGF. We acknowledge the anonymous reviewers for their constructive comments.

AUTHOR CONTRIBUTIONS

J.C. processed the CMIP5 and CMIP6 datasets and led analysis with comments from S.-W.S. J.C. and S.-W.S. discussed the results and wrote the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41612-022-00251-9.

Correspondence and requests for materials should be addressed to Seok-Woo Son.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.