New framework for benchmarking decadal predictions leveraging the PCMDI Metric Package with interactive visualization

Jung Choi[1], Jiwoo Lee[2], Kristin Chang[2], Paul A. Ullrich[2, 3], Peter J. Gleckler[2], and Sang-Yoon Jun[4]

[1]Seoul National University, Seoul, 08826, South Korea
[2]Lawrence Livermore National Laboratory, Livermore, 94550, United States
[3]UC Davis, Davis, 95616, United States
[4]Korea Polar Research Institute, Incheon, 21990, Korea

Correspondence to: Jiwoo Lee (lee1043@llnl.gov)

Abstract. Reliable climate predictions across multiple timescales are increasingly critical as climate-related risks continue to rise. With the growing number and diversity of climate prediction systems, systematic intercomparison has become essential. Here, we present a comprehensive evaluation framework based on the PCMDI Metric Package to assess the performance of multiple decadal climate prediction systems. Unlike uninitialized simulations, initialized predictions exhibit bias and predictive skill that evolve with forecast lead time. To address this, we introduce (1) model-by-lead-time portrait plots, which efficiently summarize metrics of global temperature, precipitation, and Arctic/Antarctic sea-ice extent, and (2) an HTML-based interactive visualization platform that provides detailed regional and seasonal diagnostics of model bias, skill scores, and ensemble spread for each model and lead time. Comparisons with uninitialized simulations further quantify the relative impacts of initialization and external forcing on prediction skill. The proposed framework provides a scalable and transparent approach for multi-model climate prediction assessments and can be readily extended to a wide range of operational and research forecasting systems.

1 Introduction

As society faces growing challenges related to extreme climate conditions, the demand for accurate climate predictions is increasing. To support climate-informed decision-making, operational forecast centers and research institutions around the world now routinely provide subseasonal-to-seasonal climate forecasts (e.g., Buontempo et al., 2022; Xue et al., 2025; Yhang et al., 2025) and long-term climate projections (e.g., IPCC, 2013) to inform climate adaptation and mitigation strategies. Decadal climate prediction, also known as near-term climate prediction, typically covers the period from one to ten years ahead. It bridges the gap between these seasonal forecasts and long-term projections (Meehl et al., 2021). As such, decadal climate prediction plays a critical role in enabling seamless climate services and guiding climate risk management (Dunstone et al., 2022; Solaraju-Murali et al., 2022; O'Kane et al., 2023).

In recognition of this need, major climate modeling centers began offering decadal climate prediction experiments in the early 2010s (Smith et al., 2013). Starting in 2021, the World Meteorological Organization (WMO) has provided annual climate outlooks extending up to five years ahead (Hermanson et al., 2022). Furthermore, the Decadal Climate Prediction Project (DCPP; Boer et al., 2016), which participated in the Coupled Model Intercomparison Project Phase 6 (CMIP6), is one of the coordinated international efforts that support the systematic production and comparison of initialized decadal prediction experiments. As decadal climate prediction is expected to remain a key component of the upcoming CMIP Phase 7 (Dunne et al., 2025), the number of participating prediction systems continues to increase.

Reliable decadal climate predictions require accurate information from both initial conditions and external forcing (Keenlyside et al., 2008; Corti et al., 2015; Klavans et al., 2021; Meehl et al., 2021). Initialization is particularly important for representing slow components of the climate system, such as the ocean, sea ice, and land surface, which can store anomalies and influence climate variability years in advance. These initialized climate predictions behave fundamentally differently from uninitialized climate projections. Since the observation-constrained initial state gradually drifts toward the model's climatology, initialized predictions exhibit time-evolving biases, systematic drift, and variations in prediction skill across lead times (Kharin et al., 2016; Sanchez-Gomez et al., 2016; Nadiga et al., 2019; Meehl et al., 2022). These lead-time-dependent characteristics add complexity to model evaluation and require specialized metrics capable of separating drift effects, externally forced signals, and actual predictive skill.

Over the past decades, the climate research community has developed a wide range of diagnostic frameworks and standardized metrics to evaluate long-term climate simulations, particularly for the CMIP historical experiments. These efforts include benchmarking-oriented diagnostics that quantify systematic model performance using standardized metrics and intercomparison frameworks (e.g., Gleckler et al., 2008; Eyring et al., 2020) and process-based diagnostics that investigate the physical mechanisms underlying model biases and variability (e.g., Maloney et al., 2019; Planton et al., 2021). Several community tools have been developed as reusable software packages, including the Program for Climate Model Diagnosis and Intercomparison (PCMDI) Metrics Package (PMP; Lee et al., 2014), the Earth System Model Evaluation Tool (ESMValTool; Eyring et al., 2020), and related CMIP diagnostics frameworks (Maher et al., 2025; Hassler et al., 2026).

Benchmarking diagnostics provide a standardized and reproducible framework for comparing model performance. However, most existing evaluation frameworks primarily target uninitialized historical simulations and long-term climate projections. These frameworks' current capabilities do not fully address the unique characteristics of initialized decadal predictions, where the forecast drifts, the time-evolving biases, and the lead-time-dependent prediction skill require their own customized and specialized evaluation strategies. To address this gap, we propose a new evaluation framework that could assess multi-model initialized decadal prediction systems in a systematic and comprehensive way, leveraging the established community tool for the CMIP evaluation, using some of the PMP's approach and capabilities.

2 Data and Methodology

In this section, we describe retrospective decadal predictions and observation datasets in Section 2.1. The measures of model bias, prediction skill, and spread are also presented in Sections 2.2 and 2.3.

2.1 Model and observation datasets

This study uses 11 models that participate in the DCPP. Each model includes at least three ensemble members and up to ten ensemble members, depending on the variables and simulations (see Table 1). We analyze retrospective predictions (hindcasts) initialized each November from 1960 to 2016. Two models (CanESM5 and IPSL-CM6A-LR) were initialized in January of the following year. Following the CMIP6 convention, each initialization is denoted as sYEAR; for instance, s1960 refers to the initialization in November 1960 or January 1961. These hindcasts use external boundary forcings from the CMIP6 historical experiment. Beyond the historical period (after 2015), the SSP2-4.5 scenario forcing is applied. Furthermore, we use historical simulations to evaluate the impact of external radiative forcing on predictive skills by comparing them with the initialized predictions.

The fifth-generation European Centre for Medium-Range Weather Forecasts atmospheric reanalysis (ERA5; Hersbach et al., 2020) is used as reference data for surface air temperature (TAS) from 1961 to 2021. Precipitation (PR) data are obtained from the Global Precipitation Climatology Project (GPCP; Huffman et al., 2023) for the period from 1979 to 2021. Sea ice concentration (SIC) is taken from the Hadley Centre Sea Ice and Sea Surface Temperature dataset (HadISST_ICE; Rayner et al., 2003) for the period of 1961–2021.

Table 1: Brief description of the models and three variables (TAS: surface air temperature, PR: precipitation, siconc: sea ice concentration) used in this study. Ten ensemble members are used, and variables with less than 10 members are indicated in parentheses.

No.	Model	Institution	Variables (No. of ensemble members)	
			dcppA-hindcasts	historical
1	CanESM5	CCCma, Canada	TAS, PR, siconc	TAS, PR, siconc
2	CMCC-CM2-SR5	CMCC, Italy	TAS, PR, siconc	TAS, PR, siconc
3	CNRM-ESM2-1	CNRM-CERFACS, France	TAS, PR, siconc	TAS, PR, siconc(5)
4	EC-Earth3	EC-Earth-Consortium	TAS, PR, siconc	TAS, PR, siconc(7)
5	FGOALS-f3-L	CAS, China	TAS(9), PR(9)	TAS(3), PR(3)
6	HadGEM3-GC31-MM	MOHC, UK	TAS, PR, siconc	TAS(4), PR(4), siconc(4)
7	IPSL-CM6A-LR	IPSL, France	TAS, PR, siconc	TAS, PR, siconc
8	MIROC6	MIROC, Japan	TAS, PR, siconc	TAS, PR, siconc
9	MPI-ESM1-2-HR	MPI, German	TAS, PR, siconc(5)	TAS, PR, siconc
10	MRI-ESM2-0	MRI, Japan	TAS, PR, siconc	TAS, PR, siconc(6)

11	NorCPM1	NCC, Norway	TAS, PR, siconc	TAS, PR , siconc

2.2 Measure of model bias

Model biases are evaluated for long-term climatology and trends. Bias in long-term climatology (hereafter, mean bias) is calculated for TAS and PR at each grid cell. It is defined as the difference between the model and observed climatology for each lead year during the common period of 1981–2010. For example, the mean bias for lead year 1 (LY1) is calculated using the simulations from s1980 to s2009, while that for lead year 10 (LY10) is calculated using the simulations from s1971 to s2000. Lead times up to ten years are analyzed, except for two models (CNRM-ESM2-1 and MRI-ESM2-0), which are evaluated up to five years.

To examine regional and seasonal variations, gridded monthly data are used to calculate the mean bias and then averaged over space and time. After annual averaging, the mean biases are averaged within five latitudinal zones and visualized as a function of lead years. These zones are defined as the Arctic (60–90° N), the Northern Hemisphere (NH) mid-latitudes (30–60° N), the tropics (30° S–30° N), the Southern Hemisphere (SH) mid-latitudes (30–60° S), and the Antarctic (60–90° S) (see also Table 2). Interactive plots show the global maps of mean biases, which are produced by averaging data over two-month periods to illustrate their regional and seasonal variations.

Table 2: Summary of the temporal and spatial definitions for the variables evaluated in this study. Sea-ice extent (SIE) followed the definition by Ivanova et al. (2016). Mean and trend biases are analyzed for Lead Year (LY) 1 to 10 and historical simulations (HIST). Skill scores are analyzed for LY1 to LY5 and the 5-year averages (LY1–5). The numbers and terms in parentheses in the first column refer to the periods and experimental names used for evaluating bias and skill scores, respectively.

Category	Variables	Region	Region boundaries	Lead times
Mean bias (1981–2020) and Trend bias (1979–2014)	TAS, PR	Arctic NH midlatitudes Tropics SH midlatitues Antarctic	60°–90°N, 0°–360°E 30°–60°N, 0°–360°E 30°S–30°N, 0°–360°E 30°–60°S, 0°–360°E 60°–90°S, 0°–360°E	From LY1 to LY10 and HIST
	SIE	Central Artic North Pacific North Atlantic Indian Ocean South Pacific South Atlantic	80°–90°N, 120°W–90°E; 65°–90°N, 90°E–120°W 45°–65°N, 90°E–120°W 45°–80°N, 120°W–90°E 90°–55°S, 20°–90°E 90°–55°S, 90°E–60°W 90°–55°S, 60°W–20°E	
Skill scores (s1960–s2016)	TAS, PR	Same as the above	Same as the above	From LY1 to LY5 and LY1–5
	SIE	Same as the above	Same as the above	

Bias in long-term trend (hereafter, trend bias) is defined as the difference in the least squares fitting coefficients between the model and observation. Trend biases in monthly TAS and PR are calculated at each grid cell for the common period of 1979–2014, and then averaged over space and time. Similar to the mean bias, the trend biases are averaged within the five latitudinal zones and visualized as a function of lead years. Interactive plots are also used to examine regional and seasonal variations in the trend bias.

Unlike TAS and PR, the monthly SIC is averaged first over predefined regions. This regional averaging approach further enables consistent comparison across heterogeneous native ocean grids without requiring an additional remapping procedure. According to Ivanova et al. (2016), both the Arctic and the Antarctic are divided into three regions to examine the mean and trend biases (see also Table 2). In each domain, the sea ice extent (SIE) is defined as the region in which SIC exceeds 15%. Interactive plots display the SIC climatology and trend distribution. These maps are represented on the raw grid of model data as shading, together with the observed values shown as contours. The results are presented in bimonthly periods to highlight seasonal variations.

2.3 Measure of prediction skill and spread

Initialized decadal predictions are known to exhibit systematic forecast drift as the simulations diverge from the observation-constrained initial state and approach the model's preferred climatology. Following the standard procedure used in decadal prediction studies (Goddard et al., 2013; Choi and Son, 2022), the mean climatological bias for the period 1981–2010 is removed for each lead month before calculating the metrics.

In contrast, long-term trend biases are intentionally retained in the current framework. Unlike the mean climatological drift, the trend biases may evolve nonlinearly in lead time, reflecting the complex interactions between the externally forced signals, internally generated variability, and model adjustment processes. Rather than correcting these effects explicitly, this study compares evaluation metrics between initialized decadal hindcasts and uninitialized historical simulations.

To allow for as many initializations as possible for skill evaluation, we use the initialized hindcasts of s1960–s2016 (s1978–s2016 for precipitation). For LY1, the corresponding observations cover the period from 1961 to 2017. For lead year 5 (LY5), the corresponding observations span from 1965 to 2021. To verify the predictive performance of long-term changes, the skill scores are further calculated for an average of lead years one through five (LY1–5).

Deterministic prediction skill is evaluated quantitatively using two metrics: the anomaly correlation coefficient (ACC) and the mean squared skill score (MSSS). These metrics are defined as follows:

$$ACC_{\tau}(M,O) = \frac{\frac{1}{n}\sum_{j=1}^{n}(M_{j\tau}-\bar{M}_{\tau})(O_{j\tau}-\bar{O}_{\tau})}{\sqrt{\frac{1}{n}\sum_{j=1}^{n}(M_{j\tau}-\bar{M}_{\tau})^2}\sqrt{\frac{1}{n}\sum_{j=1}^{n}(O_{j\tau}-\bar{O}_{\tau})^2}}, \quad (1)$$

$$MSSS_{\tau}(M,O) = 1-\frac{MSE_{\tau}(M)}{MSE_{\tau}(\bar{O})} = 1-\frac{\frac{1}{n}\sum_{j=1}^{n}[(M_{j\tau}-\bar{M}_{\tau})-(O_{j\tau}-\bar{O}_{\tau})]^2}{\frac{1}{n}\sum_{j=1}^{n}(O_{j\tau}-\bar{O}_{\tau})^2} \quad (2)$$

where *M* and *O* are the ensemble-mean forecasts and observations, respectively. Since the annual average is performed for both *M* and *O*, the skill scores are not dependent on the seasons. The subscript *j* represents the initialization year, and *n* represents the total number of initializations. Therefore, *n* is set to 57 for TAS and SIE (39 for PR). The overbar indicates the long-term average over the entire period. The MSSS is a function of the mean squared error (MSE); mathematically, it combines the ACC and the conditional bias (Goddard et al., 2013). Therefore, while the ACC quantifies the model's predictive ability in terms of the phase of variability, the MSSS further estimates both the phase and amplitude of variability. Both the ACC and MSSS are 1 for perfect forecasts. The prediction skill of SIE is further evaluated using root mean square error (RMSE), defined as the square root of the numerator in the second term of Eq. (2).

In addition, the ratio of predictable components (RPC) between the real and model worlds is computed to measure the signal-to-noise paradox of climate forecasts (Weisheimer et al., 2024). It is defined by comparing the predictable component of the observations ($\mathrm{PC_{Obs}}$) with the predictable component of the model ($\mathrm{PC_{Model}}$) as follows:

$$\mathrm{RPC} = \frac{PC_{Obs}}{PC_{Model}} \geq \frac{ACC(M,O)}{\sqrt{\sigma^2_{signal}/\sigma^2_{total}}} \approx \frac{ACC(M,O)}{ACC(M,M')}. \tag{3}$$

The lower bound of $\mathrm{PC_{Obs}}$ can be estimated by the ACC between ensemble-mean forecasts (*M*) and observations (*O*), as the $\mathrm{ACC}^2(M, O)$ reflects the proportion of the observed variance explained by ensemble-mean forecasts (Eade et al., 2014). Similarly, $\mathrm{PC_{Model}}$ can be identical to the expected ACC between ensemble-mean forecasts (*M*) and individual forecasts (*M'*). If the RPC is significantly greater than one, then the observations are more predictable than the model ensemble forecasts, constituting the signal-to-noise paradox (Weisheimer et al., 2024).

2.4 Interactive visualization

Unlike uninitialized simulations (e.g., historical experiments), initialized decadal predictions exhibit evolution in model biases and prediction skill that depends on lead time. Consequently, evaluation outputs are substantially larger and more multidimensional, considering lead times, models, and metrics simultaneously. Therefore, static figures alone are not very efficient to explore and interpret the full evaluation results.

To address this challenge, in this study, we have developed an HTML-based interactive visualization framework following the PMP approach. We used the Python library Bokeh (https://bokeh.org/) to enable flexible exploration of the benchmarking diagnostics. The interface allows users to dynamically compare models, lead times, and diagnostic metrics through interactive selection and navigation tools. With this design, researchers can rapidly detect systematic model behaviors, lead-time drift, and differences between models—insights that are tough to catch using conventional static plots.

The framework supports reproducible and scalable evaluation workflows for large multi-model prediction archives such as CMIP6 DCPP. By leveraging browser-based interactive graphics, the visualization outputs can be easily shared, archived, and extended to future prediction systems and additional diagnostic metrics. All interactive figures are available at https://pcmdi.llnl.gov/metrics/dcpp (last access: 18 May 2026).

3 Results

3.1 Mean bias

Figure 1a presents a model-by-lead-time portrait plot of the mean bias in TAS. The rightmost column corresponds to the mean bias derived from the historical experiments (HIST), enabling a comparison between the time-dependent mean bias of initialized decadal hindcasts and the systematic climatological bias of HIST. The five inner boxes within each cell, from top to bottom, correspond to the Arctic, NH mid-latitudes, tropics, SH mid-latitudes, and Antarctic regions. This allows for an inspection of the latitude-dependent bias structure. In general, biases tend to be larger in the polar regions than in the low-to-mid latitudes.

Although some models exhibit clear lead-time-dependent bias growth (e.g., polar regions in EC-Earth3 and HadGEM3-GC31-MM), the evolution of forecast bias is not always visually pronounced in Fig. 1. This partly reflects the substantial inter-model spread represented using a common color scale, but also the rapid adjustment of initialized forecasts toward the model climatology, particularly in the ocean component, where mean biases can develop within the first several months of integration (Ma et al., 2021).

As lead time increases, initialized forecasts generally tend to drift from the observed initial state toward the model's preferred coupled equilibrium state. However, the evolution of forecast bias and skill is not necessarily monotonic, as it can be affected by initialization shocks and adjustment processes during the forecast period (e.g., Mulholland et al., 2015). Additionally, coupled ocean–atmosphere feedbacks and internally generated ocean variability may partially compensate for inherited model biases. Consequently, some models exhibit even decreasing biases with lead time, especially for tropical TAS (e.g., CanESM5, CMCC-CM2-SR5, FGOALS-f3-L in Fig. 1a), rather than the monotonic growth expected from climatological drift alone.

Figure 1b shows the mean bias in PR. Given the fact that simulated precipitation is affected by model physics and dynamics, as well as complicated nonlinear surface processes (Li et al., 2020; Tian and Dong, 2020), the magnitude of the precipitation bias does not necessarily correspond to the magnitude of the temperature bias. Instead, Fig. 1b showcases more of the known pattern of systematic precipitation biases in climate models than the substantial lead-time-dependent variations. Most models exhibit wet biases in the tropics and dry biases in the mid-latitudes. Tropical wet biases are commonly associated with the excessive double Intertropical Convergence Zone (ITCZ), which has been linked to multiple factors, including deficiencies in dynamical circulation and thermodynamic processes (Zhang et al., 2015; Samanta et al., 2019). In contrast, dry biases in the mid-latitudes are often related to land–atmosphere feedbacks associated with summer warm biases (Mueller and Seneviratne, 2014; Lin et al., 2017), as well as atmospheric circulation biases associated with storm-track position and intensity (Priestley et al., 2020; Schemm, 2023).

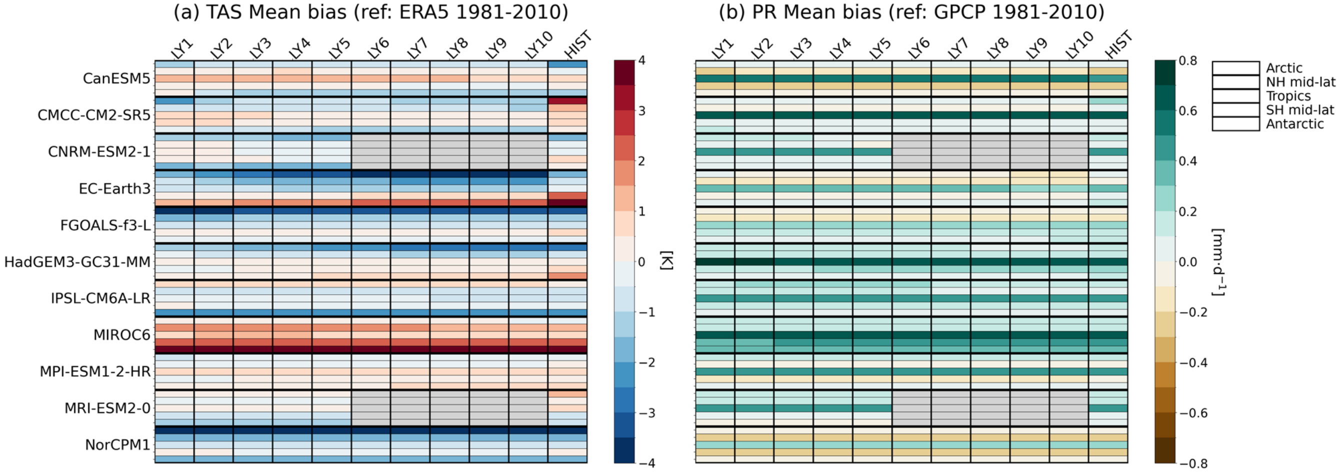

Figure 1: (a) Model-by-lead-time portrait plot showing the mean bias of TAS over the period 1981–2010. The mean bias is calculated as a function of LY over the globe and five subregions: Arctic (60°–90° N), the Northern Hemisphere mid-latitudes (30°–60° N), the Tropics (30° S–30° N), the Southern Hemisphere mid-latitudes (30°–60° S), and Antarctic (60°–90° S). The rightmost column indicates the mean bias from HIST for the same period. ERA5 is used as the reference dataset. Units are [K]. (b) Same as (a), but for the precipitation (PR). GPCP is used as the reference dataset. Units are [mm d^{-1}]. Mean bias is calculated for each month and each grid cell, and then annual and area averages are taken. Note that labels are top-aligned to optimize visibility within the scrolling web interface. All individual subplots relevant to portrait plots can be interactively visualized on the https://pcmdi.llnl.gov/metrics/dcpp (last access: 18 May 2026).

Figure 2 displays the mean bias of SIE in the Arctic and Antarctic regions. While the observed SIE exhibits a similar magnitude between the two regions (Arctic: ~11.98 × 10^6 km^2; Antarctic: ~12.59 × 10^6 km^2; the Arctic SIE is approximately 95% of the Antarctic SIE), the mean bias and inter-model spread tend to be larger in the Antarctic. In some models, the SIE mean bias is closely related to the TAS mean bias in the polar regions. For example, EC-Earth3 shows an expansion of Arctic sea ice, especially in the North Atlantic. This is likely associated with the growth of the Arctic cold bias at longer lead times (see also Fig. 1a). Conversely, the Antarctic warm bias in MIROC6 is pronounced and corresponds to widespread sea ice retreat. HadGEM3-GC31-MM exhibits a greater reduction in Antarctic sea ice in HIST than in the initialized decadal hindcasts. This is consistent with the stronger warm bias in this region (see also Fig. 1a).

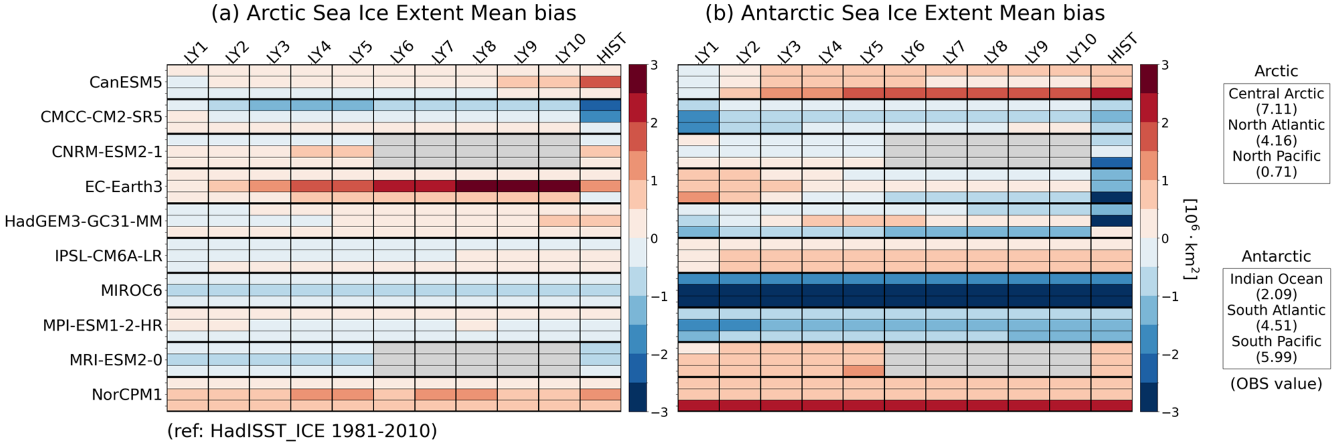

Figure 2: Model-by-lead-time portrait plot showing the mean bias in sea-ice extent over the period 1981–2010. Area averages are firstly taken for three predetermined subregions: (a) the Arctic and (b) the Antarctic, for each month. Then, the annual averages are taken at each LY. The rightmost column indicates the mean bias from HIST for the same period. HadISST_ICE is used as the reference dataset. Units are [10^6 km^2]. Numbers in parentheses on the right side of the figure represent the observed values.

As illustrated in Figs. 1 and 2, the model-by-lead-time portrait plots are useful for summarizing of overall mean bias and spread across models. However, they are limited to capturing regional and seasonal characteristics. To provide a more comprehensive view of mean biases, we offer an HTML-based interactive visualization map of mean bias. Figures 3 and 4 present examples of interactive visualizations for TAS and Arctic sea-ice mean biases of EC-Earth3, respectively. These present global maps of mean bias averaged over two-month periods at each lead time and model. For example, EC-Earth3 shows a pronounced cold bias over the North Atlantic during the cold season at LY10 (Fig. 3), accompanied by a significant overestimation of sea-ice coverage relative to observations (Fig. 4). Although not shown here, all models exhibit a wet bias in the tropics, primarily linked to the summer ITCZ. The dry bias in the mid-latitudes varies by season and region (https://pcmdi.llnl.gov/pmp-preliminary-results/graphics/dcpp/mean_bias/fig1b_interactive_PR_mean_bias_portrait_plot.html, last access: 18 May 2026).

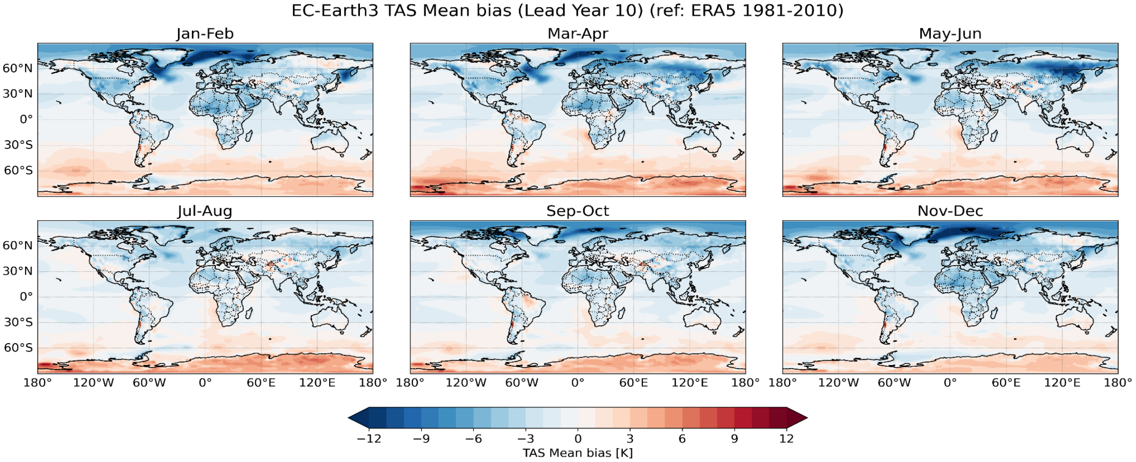

Figure 3: This example figure shows the spatial distribution of the mean bias in TAS, averaged over two-month periods, for the EC-Earth3 model at LY10. All cases are available at https://pcmdi.llnl.gov/pmp-preliminary-results/graphics/dcpp/mean_bias/fig1a_interactive_TAS_mean_bias_portrait_plot.html, last access: 18 May 2026).

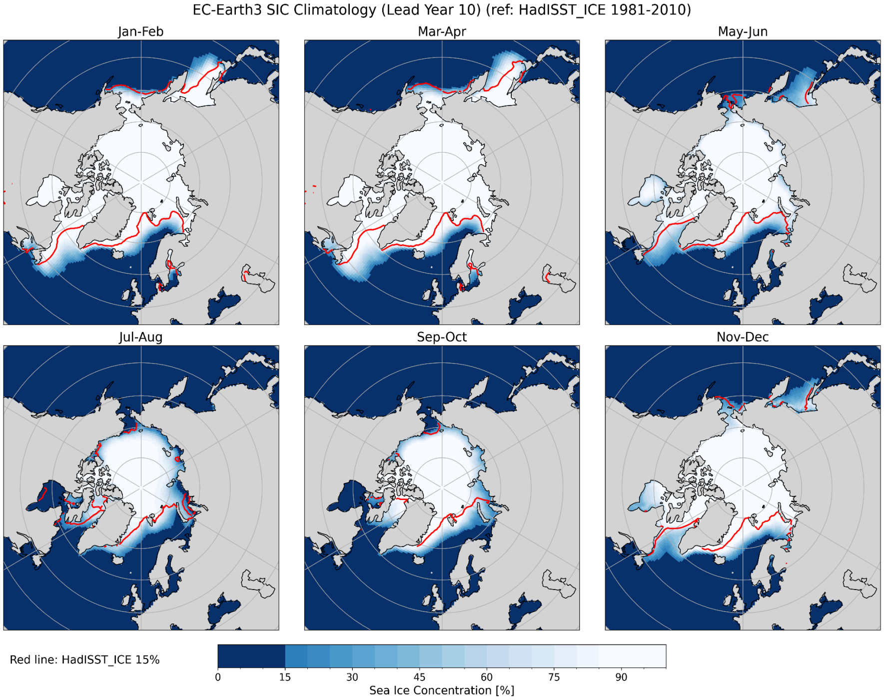

Figure 4: This example figure illustrates the spatial distribution of the climatology of Arctic sea-ice concentration, averaged over two-month periods of 1981–2010. It is shown for EC-Earth3 model at LY10. Red line represents the observed value of

15% from HadISST_ICE for the same period. All cases are available at https://pcmdi.llnl.gov/pmp-preliminary-results/graphics/dcpp/mean_bias/fig2a_interactive_Arctic_mean_bias_portrait_plot.html, last access: 18 May 2026).

3.2 Trend bias

Figures 5 and 6 show the long-term trend bias for the period from 1979 to 2014. During this period, a warming trend is observed in all regions except the Antarctic. The observed warming trends are as follows: Arctic 0.57, NH mid-latitudes 0.28, Tropics 0.13, and SH mid-latitudes 0.04 K per 10 yr (see also Fig. S1). This latitudinal structure reasonably reflects the Arctic amplification (e.g., Smith et al., 2019; Screen et al., 2025). Most models exhibit a positive bias, indicating that they simulate a stronger warming trend. This feature is particularly pronounced in the Arctic, where some models, such as CanESM5, HadGEM3-GC31-MM, and IPSL-CM6A-LR, show an increasing positive bias with lead time (Fig. 5a). The warmer bias in the Arctic is consistently larger during the cold season than during the warm season across all models (https://pcmdi.llnl.gov/pmp-preliminary-results/graphics/dcpp/trend_bias/fig5a_interactive_TAS_trend_bias_portrait_plot.html, last access: 18 May 2026). This discrepancy between observations and models of Arctic warming trends may be due to imperfect sea ice processes within the models, differences in internal variability, and changes in observed trends over the analysis period (Huang et al., 2019; Ye and Messori, 2021; Chylek et al., 2023).

In contrast, observations show a very weak cooling trend over the Antarctic (–0.03 K per 10 yr). This is likely because temperature trends in Antarctica are not stationary over time; rather, they are greatly influenced by inter-decadal variability in the surrounding atmosphere and ocean (Jun et al., 2020; Dalaiden et al., 2021; Sato and Simmonds, 2021). The models, however, tend to simulate a warming trend and thus show a positive bias. The Antarctic warm bias is more prominent in sea-ice regions, especially in the South Atlantic during the cold season, rather than over the continent itself (see subplots at https://pcmdi.llnl.gov/pmp-preliminary-results/graphics/dcpp/trend_bias/fig5a_interactive_TAS_trend_bias_portrait_plot.html, last access: 18 May 2026). This indicates that current climate models have difficulty representing sea ice-related climate feedback and inter-decadal variability in the Antarctic region.

Trend bias in PR is most prevalent in the NH mid-latitudes for all models (Fig. 5b). Although the observed trend during this period is negative (–0.25 mm d^{-1} per 100 yr), all models fail to simulate the observed drying trend, even at short lead times. For instance, a significant drying trend is present in the western North Atlantic and western North America from January to February (Fig. S2). However, all models show a positive (wet) bias, that exceeds the magnitude of the observed trend, even within the one-year forecast range (Fig. 7, see subplots at https://pcmdi.llnl.gov/pmp-preliminary-results/graphics/dcpp/trend_bias/fig5b_interactive_PR_trend_bias_portrait_plot.html, last access: 18 May 2026). Given the fact that the models are initialized in November or January, this suggests that relatively coarse-resolution climate models fail to capture key wintertime NH mid-latitudes precipitation processes, even within the seasonal prediction timescales. In contrast to the NH mid-latitudes, the tropics exhibit a positive observed trend (0.13 mm d^{-1} per 100 yr), yet the models demonstrate a dry bias. This bias largely results from the models' inability to reproduce the observed wetting trend over the Indian Ocean during September–December (see also the same subplots).

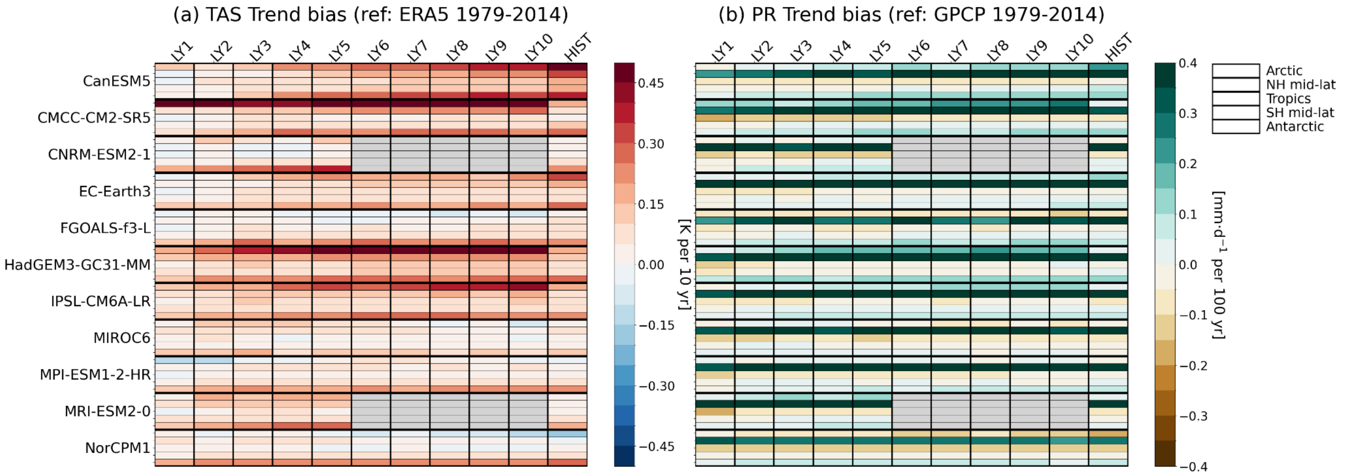

Figure 5. Same as Fig. 1, but for the trend bias over the period 1979–2014. The trend bias is defined as the difference in trend between the model and the observations. Units are [K per 10 yr] for (a) TAS and [mm d^{-1} per 100 yr] for (b) PR. Trend bias is calculated for each month and each grid cell, and then annual and area averages are taken.

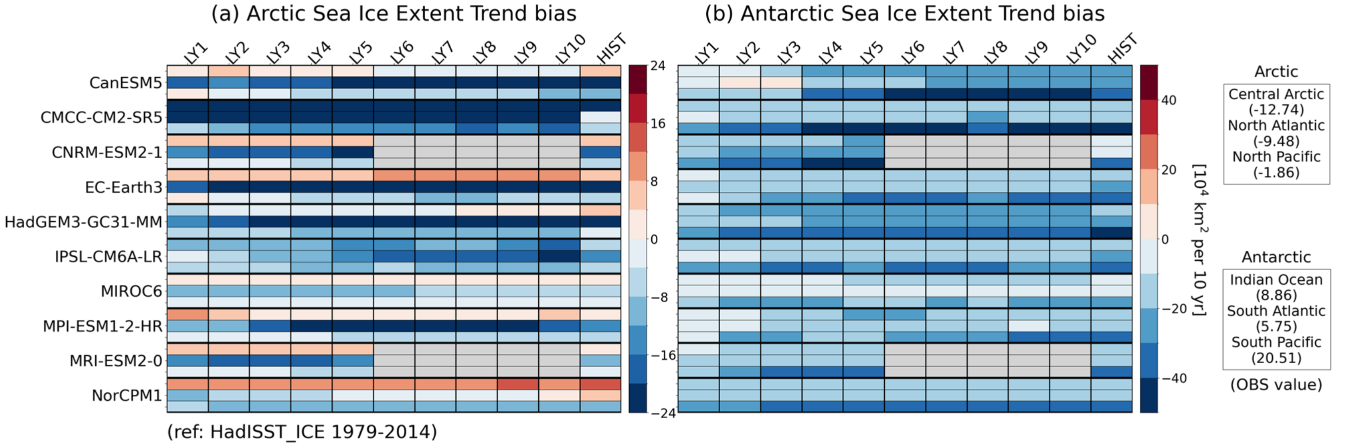

Figure 6. Same as Fig. 2, but for the trend bias of (a) the Arctic and (b) the Antarctic sea-ice extent. Units are [10^4 km^2 per 10 yr].

Observations show the strongest negative SIE trend in the Central Arctic (–12.74 × 10^4 km^2 per 10 yr), particularly during September–October (see subplots at https://pcmdi.llnl.gov/pmp-preliminary-results/graphics/dcpp/trend_bias/fig6a_interactive_Arctic_trend_bias_portrait_plot.html, last access: 18 May 2026). A negative annual-mean trend is also observed in the North Atlantic, particularly from November to June. The trend bias is generally negative in the North Atlantic and North Pacific, but eight out of 10 models exhibit a positive bias in the Central Arctic (Fig. 6a). Interestingly, models with strong positive biases in Arctic temperature trends, such as CMCC-CM2-SR5, HadGEM3-GC31-MM, and IPSL-CM6A-LR, also exhibit strong biases in sea-ice shrinkage trends, particularly in the North Atlantic.

An expansion trend of SIE is observed in the Antarctic during the period of 1979–2014, particularly in the South Pacific (20.51 × 10^4 km^2 per 10 yr) (see also contours in Fig. 8). Most models fail to reproduce this increasing trend, instead showing a negative trend bias. At shorter lead times, some models (e.g., CanESM, EC-Earth3, and MRI-ESM2-0) simulate the positive trend, resulting in a relatively small trend bias (Fig. 6b, see also subplots at https://pcmdi.llnl.gov/pmp-preliminary-results/graphics/dcpp/trend_bias/fig6b_interactive_Antarctic_trend_bias_portrait_plot.html, last access: 18 May 2026). However, as the lead time increases, the models show a transition toward a negative trend and an increase in the magnitude of the bias. These results suggest that improving sea ice initialization may enhance prediction skills, at least for shorter lead times.

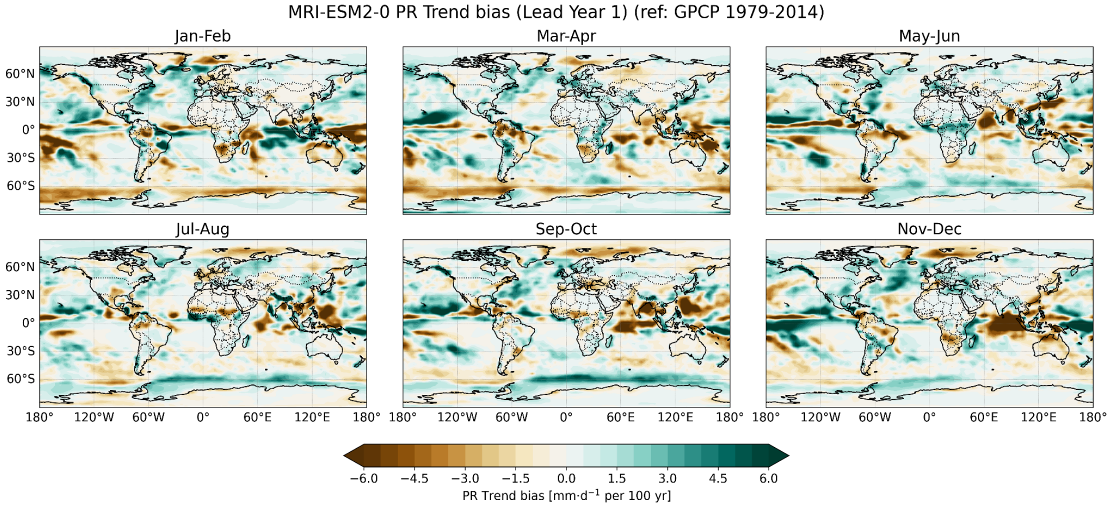

Figure 7: This example figure shows the spatial distribution of the trend bias in PR, averaged over two-month periods, for MRI-ESM2-0 model at LY1. All cases are available at https://pcmdi.llnl.gov/pmp-preliminary-results/graphics/dcpp/trend_bias/fig5b_interactive_PR_trend_bias_portrait_plot.html, last access: 18 May 2026).

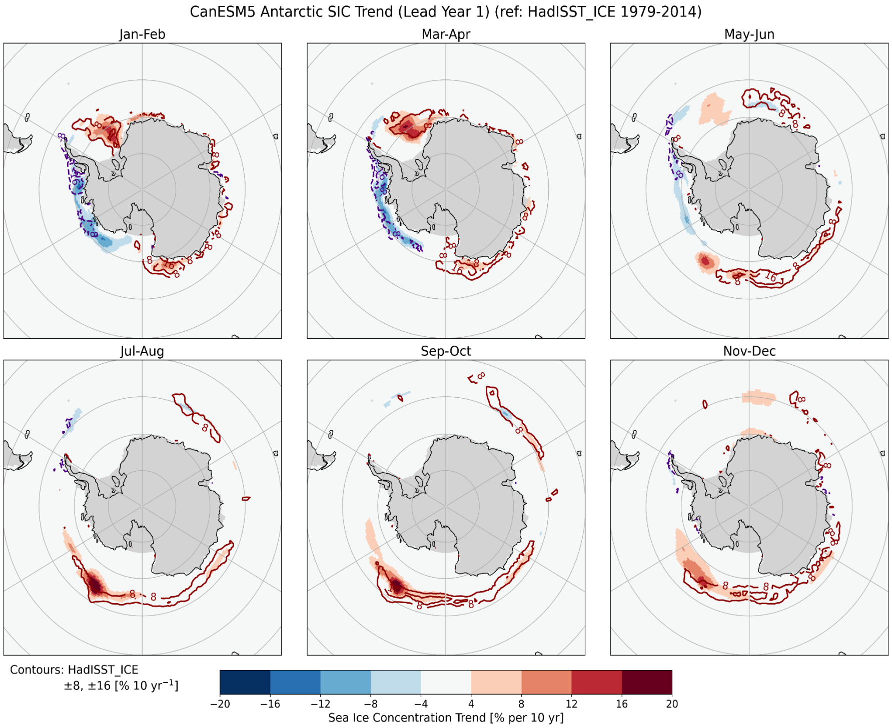

Figure 8. This example figure illustrates the spatial distribution of the trend in Antarctic sea-ice concentration, averaged over two-month periods of 1979–2014. It is shown for CMCC-CM2-SR5 model at LY1. Red (purple) lines represent the observed trend of 8 and 16% (–8 and –16%) per 10 years from HadISST_ICE for the same period. All cases are available at https://pcmdi.llnl.gov/pmp-preliminary-results/graphics/dcpp/trend_bias/fig6b_interactive_Antarctic_trend_bias_portrait_plot.html, last access: 18 May 2026).

3.3 Skill scores

Figure 9 shows the ACC, MSSS, and RPC of the annual-mean TAS using all available initialized decadal hindcasts. These metrics are first calculated at each grid cell and then averaged over the five regions. The top row presents the MME results. For the skill scores (ACC and MSSS), the MME clearly outperforms most individual models (Figs. 9a and 9b). Overall, prediction skill is highest in the tropics and lowest in the Antarctic. As the ACC only considers the sign of the predicted variability, its value remains high across all lead times due to the strong warming trend. In contrast, the MSSS, which estimates the magnitude of the predicted variability in addition to the sign, drops sharply after LY1. Meanwhile, the prediction skill is enhanced for the LY1–5 prediction compared with the LY1 prediction due to the temporal smoothing, consistent with previous studies (e.g., Hermanson et al., 2022; Choi et al., 2026).

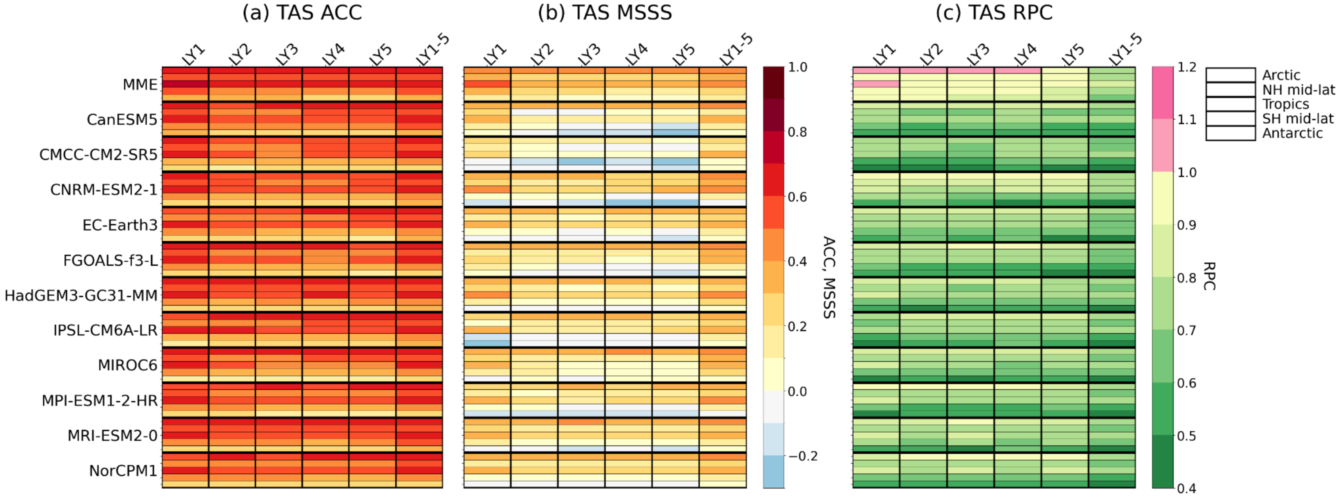

Figure 9: Model-by-lead-time portrait plot showing the (a) ACC, (b) MSSS, and (c) RPC of TAS for the multi-model ensemble mean (MME) and each model at LY1, LY2, LY3, LY4, LY5, and LY1–5. Skill scores are calculated using the annual-mean TAS at each grid cell, and then averaged across the regions. RPC is only calculated in regions where ACC > 0.

Figure 10 illustrates an example of the interactive plot for the grid-cell-based ACC of the MME at LY1. The hatched regions in Figs. 10a–c indicate areas where the ACC is below 0.6, a commonly used empirical threshold for limited practical prediction skill. ACC usually exceeds 0.6 over major ocean basins, but tends to fall below 0.6 over land and in the Southern Ocean (Fig. 10a). Since these values are influenced by initialization and external forcing, Figs. 10b–d are included to separate these effects. Figures 10b and 10c show the ACC from the initialized decadal hindcasts and HIST, respectively, for the same evaluation period (1965–2014) to ensure a fair comparison. The difference between Figs. 10b and 10c highlights the impact of initialization on the ACC (see Fig. 10d). As in previous studies, initialization has a significant impact on the skill scores in the tropical Pacific and North Atlantic regions (Meehl et al., 2016; Bilbao et al., 2021; Choi and Son, 2022). Skill enhancement in the tropical Pacific is evident in all models at LY1. Additionally, a strong effect is found near 60° S in the South Pacific. This region is dominated by long-term variability in the Antarctic deep-sea bottom waters, and improved prediction performance through initialization has been reported (Zhang et al., 2017). This can be interpreted as an improvement associated with sea ice initialization, as discussed in Fig. 6b. These enhancements are consistent across all models in the South Pacific at both LY1 and LY1–5. However, they diminish rapidly as lead time increases (see subplots at https://pcmdi.llnl.gov/pmp-preliminary-results/graphics/dcpp/skill_score/fig9a_interactive_TAS_skill_ACC_portrait_plot.html, last access: 18 May 2026).

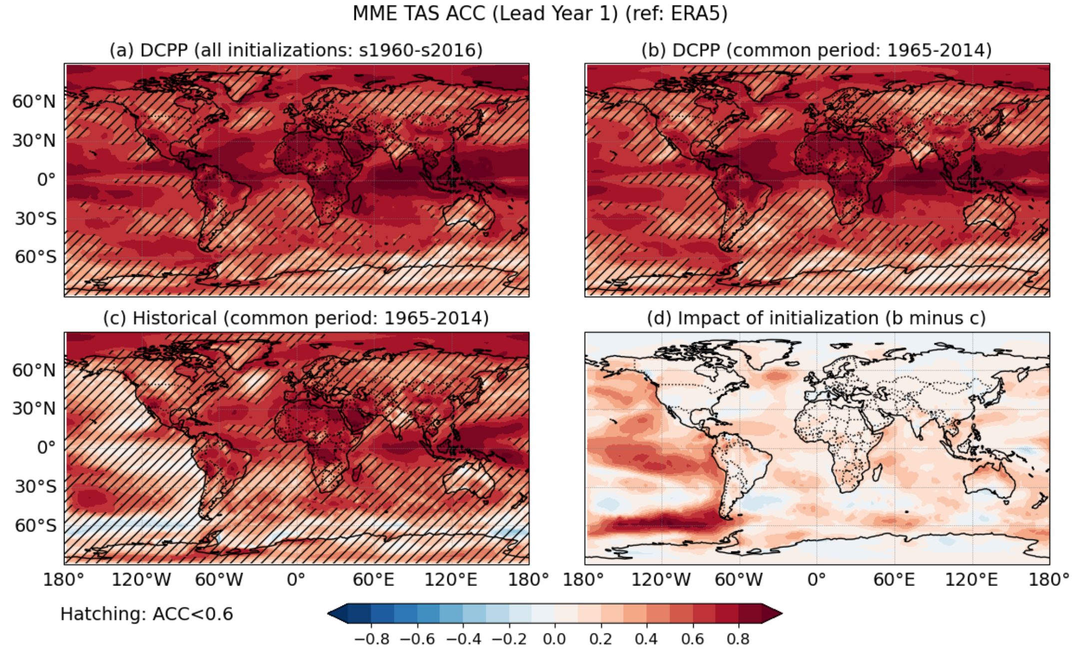

Figure 10: This example figure illustrates the spatial distribution of the MME at LY1 of TAS ACC. (a) ACC calculated from all available DCPP experiments initialized from 1960 to 2016. (b and c) Same as (a) but for the common evaluation period of 1965–2014 from DCPP and HIST simulations, respectively. Regions with ACC below 0.6 are hatched to indicate limited practical prediction skill. (d) The difference between (b) and (c) represents the added skill due to initialization. Red regions indicate areas where initialization improves prediction skill.

Figure 9c shows the RPC values for the MME and the individual models. Most models have RPC values below one because they have fewer than ten ensemble members. In contrast, the MME, which has a total of 109 ensemble members and yields an RPC greater than one in the tropics and the Arctic. The RPCs greater than one of the MME LY1 in the tropics mainly occur in the Northern Hemisphere subtropics (Fig. 11a). In the Arctic, high RPC values appear in the Central Arctic region. However, the HIST also shows high values in the same region, suggesting that the higher RPCs in the Arctic arise from the external forcing rather than initialization. RPCs greater than one are also found in the western Antarctic; however, this is offset by small values in the eastern Antarctic through zonal-mean averaging (Fig. 9c). These high RPCs are associated with improvements in ACC due to initialization (compare Figs. 11b–d).

It is interesting to note that the RPC of MME at LY1–5 drops below one in most regions, even though the ACC is high at this lead time (Figs. S3a and S3b). Since ACC is used as the numerator in the RPC definition, the decrease in RPC is likely due to a substantial increase in the denominator, which is the predictable component of the model (see Eq. (3)). In other words, this indicates that the long-term variability of individual ensemble members more closely agrees with the ensemble-mean forecasts than the ensemble-mean forecasts reproduce the observed long-term variability. On the other hand, the HIST exhibits RPCs greater than one where the high ACC exists (Fig. S3c). These results suggest that the initialization enhances inter-member consistency, particularly for the LY1–5 predictions, compared to the historical experiment.

For precipitation, meaningful prediction skill exists only over the tropical Pacific at MME LY1 (Fig. S4), and no significant results emerge at longer lead times.

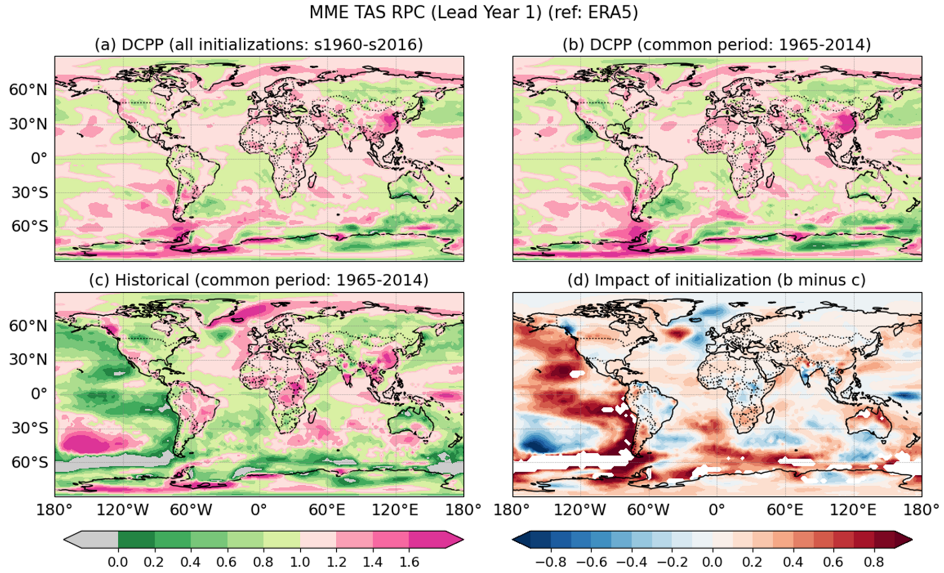

Figure 11: Same as Fig. 10 but for the RPC.

Figures 12 and 13 show the prediction skill of the Arctic and Antarctic SIE, as measured by ACC, MSSS, and RMSE. Overall, the Arctic SIE has a higher ACC and MSSS than the Antarctic SIE. In contrast, the RMSE is lower. Notably, the Central Arctic exhibits ACC values greater than 0.8 for MME in all lead times and for all individual models in LY1–5 (Fig. 12a). These high ACCs are attributable to the pronounced declining trend in this region, particularly in September (see also Fig. S5a). The lowest prediction skill is found in the North Pacific, even MSSS shows a negative value in the region (Fig. 12b). This is likely because the models fail to reproduce the observed fluctuations in the 1960s and 2010s (see also Fig. S5c). Although the climatological area is smaller in the North Atlantic than in the Central Arctic, RMSE is larger in the North Atlantic (Fig. 12c). This is due to errors in the spring, when the SIE is at its widest. The Central Arctic maintains its maximum area throughout the analysis period with negligible spring variability, while the North Atlantic SIE exhibits significant spring variability (Figs. S5a and S5b).

In the Antarctic, where long-term trends are much weaker than in the Arctic (see also Fig. S6), both ACC and MSSS are generally low (Figs. 13a and 13b). RMSE is highest in the South Pacific because its climatological extent is the largest of the thress regions (Fig. 13c). Some models (e.g., CanESM5 and EC-Earth3) that realistically reproduce the positive SIE trend in the South Pacific as discussed in Figs. 6b and 8, show relatively high ACCs at LY1. Since this high ACC does not appear in the historical simulations, especially in September, it can be interpreted as an improvement due to initialization (see also Fig. S6c). A similar enhancement related to initialization is also found in the Indian Ocean in September (Fig. S6a).

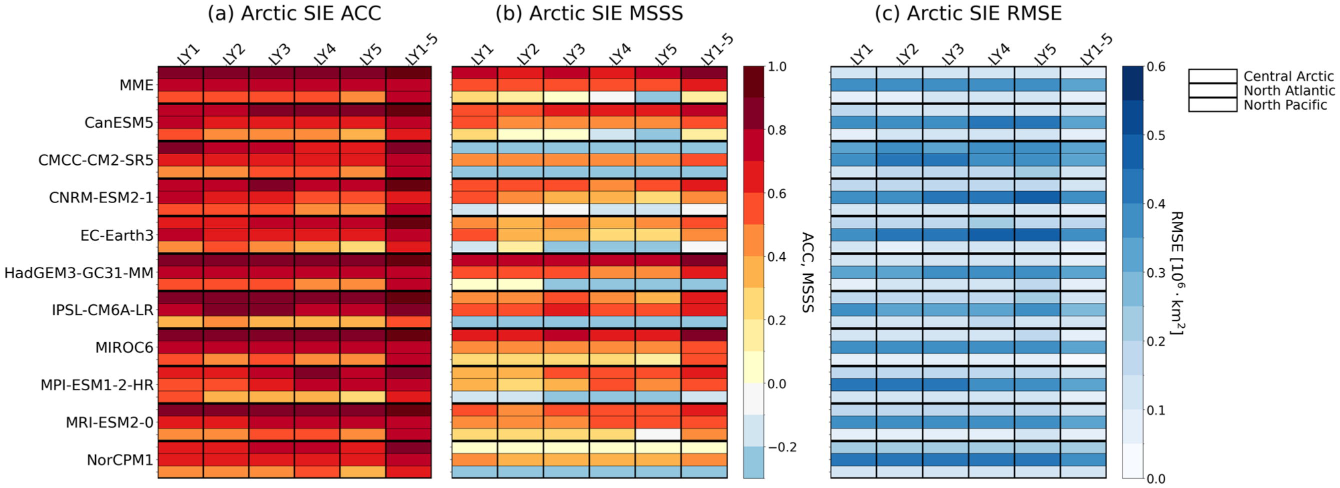

Figure 12: Model-by-lead-time portrait plot showing the Arctic SIE (a) ACC, (b) MSSS, and (c) RMSE for the MME and each model at LY1, LY2, LY3, LY4, LY5, and LY1–5. SIE is firstly defined, and then the skill scores are calculated.

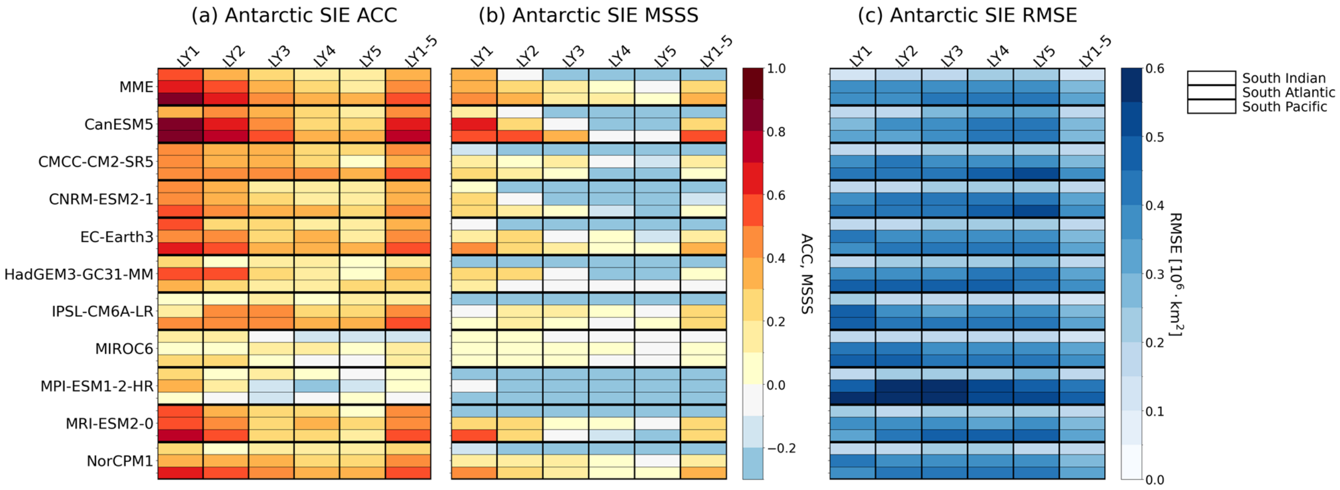

Figure 13: Same as Fig. 12 but for the Antarctic SIE.

4 Summary and discussion

In this study, we introduce a comprehensive framework for evaluating and comparing the performance of multiple initialized decadal climate prediction systems using the PCMDI Metric Package. A key feature of initialized predictions, unlike the uninitialized simulations, is that both model biases and prediction skill evolve with forecast lead time. This intrinsic behaviour substantially increases the volume and complexity of evaluation results. To efficiently diagnose these characteristics, we developed two complementary visualization tools:

(1) a model-by-lead-time portrait plot, which summarizes model biases and skill scores across five latitude-based regions for global temperature and precipitation, and across three sectors for Arctic and Antarctic sea-ice extent; and

(2) an HTML-based interactive visualization platform, which provides regional and seasonal diagnostics of model bias, skill scores, and ensemble spread for each model and lead time.

The portrait plots enable rapid identification of model-to-model differences and the evolution of biases and skill scores with forecast lead time. The interactive diagnostics further allow direct comparisons between initialized and uninitialized simulations, thereby helping to distinguish the relative contributions of initialization and external forcing to prediction performance.

Applying this portrait-based framework reveals several robust characteristics of decadal climate predictions. Surface air temperature biases exhibit substantial inter-model spread and a pronounced evolution with lead time, indicating a clear drift toward each model's biased climatology despite initialization. This gradual evolution likely reflects the influence of slowly varying processes associated with climate feedbacks, ocean memory, and model drift toward their preferred coupled equilibrium states. In contrast, precipitation biases show a more consistent and systematic spatial pattern across models. The relatively stable structure of these biases suggests that precipitation errors are primarily driven by higher-frequency atmospheric variability and localized convective processes that establish systematic biases early in the forecast period. This uniform structure implies common deficiencies in model physics, such as convection, cloud processes, and land–atmosphere coupling.

Trend-bias diagnostics also reveal systematic deficiencies in representing long-term climate changes regionally. Notably, the persistent underestimation of observed precipitation trends in the Northern Hemisphere mid-latitudes, even at short lead times, suggests that important mechanisms governing regional hydroclimate change may be inadequately represented in current prediction systems. These results underscore the importance of evaluating both mean-state biases and long-term trend accuracy when assessing decadal climate predictions. Finally, the results highlight the strong coupling between temperature and sea-ice predictions. Inter-model differences in sea ice extent bias generally scale with temperature bias, which emphasizes the role of polar feedback processes.

The lead-time-dependent skill diagnostics further demonstrate the distinct roles of initialization and externally forced variability in decadal prediction systems. In many regions, initialized predictions show enhanced ACC and MSSS at short lead times compared to uninitialized historical simulations, indicating the contribution of initialization to near-term prediction skills.

However, this improvement often decreases with increasing lead time as internally generated variability progressively reduce the influence of the initial state. The comparison between initialized and uninitialized simulations provides a useful diagnostic framework for separating transient initialization benefits from the underlying forced climate response.

Overall, the proposed framework provides a scalable and transparent approach for multi-model evaluation and can be easily extended to future prediction systems, including CMIP Phase 7, as well as operational and research forecasting systems. In addition, Future extensions of the framework should incorporate statistical significance diagnostics and uncertainty quantification for lead-time-dependent skill metrics and trend analyses. Although the present framework is primarily designed for benchmarking diagnostics rather than process-based attribution, future work should extend the framework toward process-based diagnostics that more directly connect model errors with underlying physical mechanisms.

Code and data availability

The ERA5 is obtained from the C3S web server (https://cds.climate.copernicus.eu/, last access: 18 May 2026). Global Precipitation Climatology Project (GPCP) data provided by the NOAA PSL, Boulder, Colorado, USA, from their website (https://psl.noaa.gov/data/gridded/data.gpcp.html, last access: 18 May 2026). The Hadley Centre Sea Ice dataset is obtained from the Met Office Hadley Centre web server (https://www.metoffice.gov.uk/hadobs/hadisst/, last access: 18 May 2026). Model outputs can be obtained from CMIP6 ESGF MetaGrid (https://aims2.llnl.gov/search, last access: 18 May 2026). Analysis codes are publicly available on GitHub (https://github.com/PCMDI/DCPP_PMP, last access: 18 May 2026) and the version of the codes used for this paper is archived at Zenodo (https://doi.org/10.5281/zenodo.18692307, last access: 18 May 2026).

Author contributions

JC led the conceptualization of the study, performed the data curation, formal analysis, investigation, and methodology development, developed the software, conducted validation, produced the visualizations, and wrote the original draft of the manuscript. JL contributed to the conceptualization, investigation, and methodology development, contributed to the software and validation, produced visualizations, supervised the research, and reviewed and edited the manuscript. KC contributed to the conceptualization and methodology development, contributed to the software and validation, produced visualizations, and reviewed and edited the manuscript. PAU contributed to the conceptualization of the study, acquired funding, supervised the research, and reviewed and edited the manuscript. PJG contributed to the conceptualization, supervised the research, and reviewed and edited the manuscript. SYJ contributed to the methodology development, software, and validation, and reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Acknowledgements

We thank the modeling centers 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 CMIP6 and ESGF. JC was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00342219). Work of JL, KC, PU, and PG was performed under the auspices of the U.S. DOE by the Lawrence Livermore National Laboratory (LLNL) (contract no. DE-AC52-07NA27344) and their efforts were supported by the Regional and Global Model Analysis (RGMA) program of the U.S. Department of Energy (DOE) Office of Science (OS), Biological and Environmental Research (BER) program. SYJ was supported by the Korea Meteorological Administration Research and Development Program under Grant (RS-2025-02222417).

References

Adler, R.F., Gu, G., Sapiano, M., Wang, J.-J. and Huffman, G. J.: Global Precipitation: Means, Variations and Trends During the Satellite Era (1979–2014), Surv Geophys, 38, 679–699, https://doi.org/10.1007/s10712-017-9416-4, 2017.

Bilbao, R., Wild, S., Ortega, P., Acosta-Navarro, J., Arsouze, T., Bretonnière, P.-A., Caron, L.-P., Castrillo, M., Cruz-García, R., Cvijanovic, I., Doblas-Reyes, F. J., Donat, M., Dutra, E., Echevarría, P., Ho, A.-C., Loosveldt-Tomas, S., Moreno-Chamarro, E., Pérez-Zanon, N., Ramos, A., Ruprich-Robert, Y., Sicardi, V., Tourigny, E., and Vegas-Regidor, J.: Assessment of a full-field initialized decadal climate prediction system with the CMIP6 version of EC-Earth, Earth Syst. Dynam., 12, 173–196, https://doi.org/10.5194/esd-12-173-2021, 2021.

Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., Kushnir, Y., Kimoto, M., Meehl, G. A., Msadek, R., Mueller, W. A., Taylor, K. E., Zwiers, F., Rixen, M., Ruprich-Robert, Y., and Eade, R.: The Decadal Climate Prediction Project (DCPP) contribution to CMIP6, Geosci. Model Dev., 9, 3751–3777, https://doi.org/10.5194/gmd-9-3751-2016, 2016.

Buontempo, C., Burgess, S. N, Dee, D., Pinty, B., Thépaut, J.-N., Rixen, M., Almond, S., Armstrong, D., Brookshaw, A., Alos, A. L., Bell, B., Bergeron, C., Cagnazzo, C., Comyn-Platt, E., Damasio-Da-Costa, E., Guillory, A., Hersbach, H., Horányi, A., Nicolas, J., Obregon, A., Ramos, E. P., Raoult, B., Muñoz-Sabater, J., Simmons, A., Soci, C., Suttie, M., Vamborg, F., Varndell, J., Vermoote, S., Yang, X., and de Marcillaand, J. G.: The Copernicus Climate Change Service: Climate Science in Action, Bull. Amer. Meteor. Soc., 103, E2669–E2687, https://doi.org/10.1175/BAMS-D-21-0315.1, 2022.

Choi, J., Son, S.-W., Ham, Y.-G., Lee, J., and Kim, H.: Seasonal-to-Interannual Prediction Skills of Near-Surface Air Temperature in the CMIP5 Decadal Hindcast Experiments, J. Climate, 29, 1511–1527, https://doi.org/10.1175/JCLI-D-15-0182.1, 2016.

Choi, J., and Son, S.-W.: Seasonal-to-decadal prediction of El Niño–Southern Oscillation and Pacific Decadal Oscillation, npj Clim Atmos Sci, 5, 29, https://doi.org/10.1038/s41612-022-00251-9, 2022.

Choi, J., Jun, S.-Y., Son, S.-W., Hyun, Y.-K., Lee, J.-R., Lee, J., Boo, K.-O., and Park, B.-J.: Near-term Climate Prediction of Agricultural Thermal Conditions in East Asia, Adv. Atmos. Sci., 43, 631–644, https://doi.org/10.1007/s00376-025-4471-0, 2026.

Chylek, P., Folland, C. K., Klett, J. D., Wang, M., Lesins, G., and Dubey, M. K.: High values of the Arctic Amplification in the early decades of the 21st century: Causes of discrepancy by CMIP6 models between observation and simulation, J. Geophysical Research: Atmospheres, 128, e2023JD039269, https://doi.org/10.1029/2023JD039269, 2023.

Corti, S., Palmer, T., Balmaseda, M., Weisheimer, A., Drijfhout, S., Dunstone, N., Hazeleger, W., Kröger, J., Pohlmann, H., Smith, D.. von Storch, J.-S., and Wouters, B.: Impact of Initial Conditions versus External Forcing in Decadal Climate Predictions: A Sensitivity Experiment, J. Climate, 28, 4454–4470, https://doi.org/10.1175/JCLI-D-14-00671.1, 2015.

Dalaiden, Q., Goosse, H., Rezsöhazy, J., and Thomas, E. R.: Reconstructing atmospheric circulation and sea-ice extent in the West Antarctic over the past 200 years using data assimilation, Clim. Dyn., 57, 3479–3503, https://doi.org/10.1007/s00382-021-05879-6, 2021.

Dunne, J. P., Hewitt, H. T., Arblaster, J. M., Bonou, F., Boucher, O., Cavazos, T., Dingley, B., Durack, P. J., Hassler, B., Juckes, M., Miyakawa, T., Mizielinski, M., Naik, V., Nicholls, Z., O'Rourke, E., Pincus, R., Sanderson, B. M., Simpson, I. R., and Taylor, K. E.: An evolving Coupled Model Intercomparison Project phase 7 (CMIP7) and Fast Track in support of future climate assessment, Geosci. Model Dev., 18, 6671–6700, https://doi.org/10.5194/gmd-18-6671-2025, 2025.

Dunstone, N., Lockwood, J., Solaraju-Murali, B., Reinhardt, K., Tsartsali, E. E., Athanasiadis, P. J., Bellucci, A., Brookshaw, A., Caron, L.-P., Doblas-Reyes, F. J., Früh, B., González-Reviriego, N., Gualdi, S., Hermanson, L., Materia, S., Nicodemou, A., Nicolì, D., Pankatz, K., Paxian, A., Scaife, A., Smith, D., and Thornton, H. E.: Towards Useful Decadal Climate Services, Bull. Amer. Meteor. Soc., 103, E1705–E1719, https://doi.org/10.1175/BAMS-D-21-0190.1, 2022.

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

Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., Arnone, E., Bellprat, O., Brötz, B., Caron, L.-P., Carvalhais, N., Cionni, I., Cortesi, N., Crezee, B., Davin, E. L., Davini, P., Debeire, K., de Mora, L., Deser, C., Docquier, D., Earnshaw, P., Ehbrecht, C., Gier, B. K., Gonzalez-Reviriego, N., Goodman, P., Hagemann, S., Hardiman, S., Hassler, B., Hunter, A., Kadow, C., Kindermann, S., Koirala, S., Koldunov, N., Lejeune, Q., Lembo, V., Lovato, T., Lucarini, V., Massonnet, F., Müller, B., Pandde, A., Pérez-Zanón, N., Phillips, A., Predoi, V., Russell, J., Sellar, A., Serva, F., Stacke, T., Swaminathan, R., Torralba, V., Vegas-Regidor, J., von Hardenberg, J., Weigel, K., and Zimmermann, K.: Earth System

Model Evaluation Tool (ESMValTool) v2.0 – an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP, Geosci. Model Dev., 13, 3383–3438, https://doi.org/10.5194/gmd-13-3383-2020, 2020.

Hassler, B., Hoffman, F. M., Beadling, R., Blockley, E., Bo, H., Lee, J., Lembo, V., Lewis, J., Lu, J., Madaus, L., Malinina, E., Medeiros, B., Pokam, W., Scoccimarro, E., Swaminathan, R.: Systematic Benchmarking of Climate Models: Methodologies, Applications, and New Directions, Reviews of Geophysics, 64, e2025RG000891, https://doi.org/10.1029/2025RG000891, 2026.

Hermanson, L., Smith, D., Seabrook, M., Bilbao, R., Doblas-Reyes, F., Tourigny, E., Lapin, V., Kharin, V. V., Merryfield, W. J., Sospedra-Alfonso, R., Athanasiadis, P., Nicoli, D., Gualdi, S., Dunstone, N., Eade, R., Scaife, A., Collier, M., O'Kane, T., Kitsios, V., Sandery, P., Pankatz, K., Früh, B., Pohlmann, H., Müller, W., Kataoka, T., Tatebe, H., Ishii, M., Imada, Y., Kruschke, T., Koenigk, T., Karami, M. P., Yang, S., Tian, T., Zhang, L., Delworth, T., Yang, X., Zeng, F., Wang, Y., Counillon, F., Keenlyside, N., Bethke, I., Lean, J., Luterbacher, J., Kolli, R. K., and Kumar, A.: WMO Global Annual to Decadal Climate Update: A Prediction for 2021–25, Bull. Amer. Meteor. Soc., 103, E1117–E1129, https://doi.org/10.1175/BAMS-D-20-0311.1, 2022.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, S.-N.: The ERA5 global reanalysis, Quart. J. Roy. Meteor. Soc., 146, 1999-2049, https://doi.org/10.1002/qj.3803, 2020.

Huang, J., Ou, T., Chen, D., Luo, Y., and Zhao, Z.: The amplified Arctic warming in the recent decades may have been overestimated by CMIP5 models, Geophy. Res. Lett., 46, 13338–13345, https://doi.org/10.1029/2019GL084385, 2019.

Huffman, G. J., Adler, R. F., Behrangi, A., Bolvin, D. T., Nelkin, E. J., Gu, G., and Ehsani, M. R.: The New Version 3.2 Global Precipitation Climatology Project (GPCP) Monthly and Daily Precipitation Products, J. Climate, 36, 7635–7655, https://doi.org/10.1175/JCLI-D-23-0123.1, 2023.

IPCC: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp, 2013.

Ivanova, D. P., Gleckler, P. J., Taylor, K. E., Durack, P. J., and Marvel, K. D.: Moving beyond the Total Sea Ice Extent in Gauging Model Biases, J. Climate, 29, 8965–8987, https://doi.org/10.1175/JCLI-D-16-0026.1, 2016.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res., 113, D06104, doi:10.1029/2007JD008972, 2008:

Goddard, L., Kumar, A., Solomon, A., Smith, D., Boer, G., Gonzalez, P., Kharin, V., Merryfield, W., Deser, C., Mason, S. J., Kirtman, B. P., Msadek, R., Sutton, R., Hawkins, E., Fricker, T., Hegerl, G., Ferro, C. A. T., Stephenson, D. B., Meehl, G. A., Stockdale, T., Burgman, R., Greene, A. M., Kushnir, Y., Newman, M., Carton, J., Fukumori, I., and Delworth, T.: A verification framework for interannual-to-decadal predictions experiments, Clim. Dyn., 40, 245–272, https://doi.org/10.1007/s00382-012-1481-2, 2013.

Jun, S.-Y., Kim, J.-H., Choi, J., Kim, S.-J., Kim, B.-M., and An, S.-I.: The internal origin of the west-east asymmetry of Antarctic climate change, Sci. Adv., 6, eaaz1490, https://doi.org/10.1126/sciadv.aaz1490, 2020.

Keenlyside, N. S., Latif, M., Jungclaus, J., Kornblueh, L., and Roeckner, E.: Advancing decadal-scale climate prediction in the North Atlantic sector, Nature, 453, 84–88, https://doi.org/10.1038/nature06921, 2008.

Kharin, V. V., Boer, G. J., Merryfield, W. J., Scinocca, J. F., and Lee, W.-S.: Statistical adjustment of decadal predictions in a changing climate, Geophys. Res. Lett., 39, L19705, https://doi.org/10.1029/2012GL052647, 2012.

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, 4, 22, https://doi.org/10.1038/s41612-021-00177-8, 2021.

Lee, J., Gleckler, P. J., Ahn, M.-S., Ordonez, A., Ullrich, P. A., Sperber, K. R., Taylor, K. E., Planton, Y. Y., Guilyardi, E., Durack, P., Bonfils, C., Zelinka, M. D., Chao, L.-W., Dong, B., Doutriaux, C., Zhang, C., Vo, T., Boutte, J., Wehner, M. F., Pendergrass, A. G., Kim, D., Xue, Z., Wittenberg, A. T., and Krasting, J.: Systematic and objective evaluation of Earth system models: PCMDI Metrics Package (PMP) version 3, Geosci. Model Dev., 17, 3919–3948, https://doi.org/10.5194/gmd-17-3919-2024, 2024.

Li, J.-L. F., Xu, K.-M., Richardson, M., Lee, W.-L., Jiang, J. H., Yu, J.-Y., Wang, Y.-H., Fetzer, E., Wang, L.-C., Stephens, G., and Liang, H.-C.: Annual and seasonal mean tropical and subtropical precipitation bias in CMIP5 and CMIP6 models, Environ. Res. Lett., 15, 124068, https://doi.org/10.1088/1748-9326/abc7dd, 2020.

Lin, Y., Dong, W., Zhang, M., Xie, Y., Xue, W., Huang, J., and Luo, Y.: Causes of model dry and warm bias over central U.S. and impact on climate projections. Nat. Commun., 8, 881, https://doi.org/10.1038/s41467-017-01040-2, 2017.

Ma, H.-Y., Siongco, A. C., Klein, S. A., Xie, S., Karspeck, A. R., Raeder, K., Anderson, J. L., Lee, J., Kirtman, B. P., Merryfield, W. J., Murakami, H., and Tribbia, J. J.: On the Correspondence between Seasonal Forecast Biases and Long-Term Climate Biases in Sea Surface Temperature, J. Climate, 34, 427–446, https://doi.org/10.1175/JCLI-D-20-0338.1, 2021.

Maher, N., Phillips, A. S., Deser, C., Wills, R. C. J., Lehner, F., Fasullo, J., Caron, J. M., Brunner, L., Beyerle, U., and Jeffree, J.: The updated Multi-Model Large Ensemble Archive and the Climate Variability Diagnostics Package: new tools for the study of climate variability and change, Geosci. Model Dev., 18, 6341–6365, https://doi.org/10.5194/gmd-18-6341-2025, 2025.

Maloney, E. D., Gettelman, A., Ming, Y., Neelin, J. D., Barrie, D., Mariotti, A., Chen, C.-C., Coleman, D. R. B., Kuo, Y.-H., Singh, B., Annamalai, H., Berg, A., Booth, J. F., Camargo, S. J., Dai, A., Gonzalez, A., Hafner, J., Jiang, X., Jing, X., Kim, D., Kumar, A., Moon, Y., Naud, C. M., Sobel, A. H., Suzuki, K., Wang, F., Wang, J., Wing, A. A., Xu, X., and

Zhao, M.: Process-Oriented Evaluation of Climate and Weather Forecasting Models, Bull. Amer. Meteor. Soc., 100, 1665–1686, https://doi.org/10.1175/BAMS-D-18-0042.1, 2019.

Meehl, G. A., Hu, A., and Teng, H.: Initialized decadal prediction for transition to positive phase of the Interdecadal Pacific Oscillation, Nat. Commun., 7, 11718, https://doi.org/10.1038/ncomms11718, 2016.

Meehl, G. A., Richter, J. H., Teng, H., Capotondi, A., Cobb, K., Doblas-Reyes, F., Donat, M. G., England, M. H., Fyfe, J. C., Han, W., Kim, H., Kirtman, B. P., Kushnir, Y., Lovenduski, N. S., Mann, M. E., Merryfield, W. J., Nieves, V., Pegion, K., Rosenbloom, N., Sanchez, S. C., Scaife, A. A., Smith, D., Subramanian, A. C., Sun, L., Thompson, D., Ummenhofer, C. C., and Xie, S.-P.: Initialized Earth System prediction from subseasonal to decadal timescales, Nat. Rev. Earth & Environ., 2, 340–357, https://doi.org/10.1038/s43017-021-00155-x, 2021.

Meehl, G.A., Teng, H., Smith, D., Yeager, S., Merryfield, W., Doblas-Reyes, F., and Glanville, A. A.: The effects of bias, drift, and trends in calculating anomalies for evaluating skill of seasonal-to-decadal initialized climate predictions, Clim. Dyn., 59, 3373–3389, https://doi.org/10.1007/s00382-022-06272-7, 2022.

Mueller, B., and Seneviratne, S. I.: Systematic land climate and evapotranspiration biases in CMIP5 simulations, Geophys. Res. Lett., 41, 128–134, https://doi.org/10.1002/2013GL058055, 2014.

Mulholland, D. P., Laloyaux, P., Haines, K., and Balmaseda, M. A.: Origin and Impact of Initialization Shocks in Coupled Atmosphere–Ocean Forecasts, Mon. Wea. Rev., 143, 4631–4644, https://doi.org/10.1175/MWR-D-15-0076.1, 2015.

Nadiga, B. T., Verma, T., Weijer, W., and Urban, N. M.: Enhancing skill of initialized decadal predictions using a dynamic model of drift, Geophy. Res. Lett., 46, 9991–9999, https://doi.org/10.1029/2019GL084223, 2019.

O'Kane, T. J., Scaife, A. A., Kushnir, Y., Brookshaw, A., Buontempo, C., Carlin, D., Connell, R. K., Doblas-Reyes, F., Dunstone, N., Förster, K., Graça, A., Hobday, A. J., Kitsios, V., van der Laan, L., Lockwood, J., Merryfield, W. J., Paxian, A., Payne, M. R., Reader, M. C., Saville, G. R., Smith, D., Solaraju-Murali, B., Caltabiano, N., Carman, J., Hawkins, E., Keenlyside, N., Kumar, A., Matei, D., Pohlmann, H., Power, S., Raphael, M., Sparrow, M. and Wu, B.: Recent applications and potential of near-term (interannual to decadal) climate predictions, Front. Clim., 5, 1121626, https://doi.org/10.3389/fclim.2023.1121626, 2023.

Planton, Y. Y., Guilyardi, E., Wittenberg, A. T., Lee, J., Gleckler, P. J., Bayr, T., McGregor, S., McPhaden, M. J., Power, S., Roehrig, R., Vialard, J., and Voldoire, A.: Evaluating Climate Models with the CLIVAR 2020 ENSO Metrics Package, Bull. Amer. Meteor. Soc., 102, E193–E217, https://doi.org/10.1175/BAMS-D-19-0337.1, 2021.

Priestley, M. D. K., Ackerley, D., Catto, J. L., Hodges, K. I., McDonald, R. E., and Lee, R. W.: An Overview of the Extratropical Storm Tracks in CMIP6 Historical Simulations. J. Climate, 33, 6315–6343, https://doi.org/10.1175/JCLI-D-19-0928.1, 2020.

Purich, A., Cai, W., England, M. H., and Cowan, T.: Evidence for link between modelled trends in Antarctic sea ice and underestimated westerly wind changes, Nat. Commun., 7, 10409, https://doi.org/10.1038/ncomms10409, 2016.

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

Samanta, D., Karnauskas, K. B., and Goodkin, N. F.: Tropical Pacific SST and ITCZ biases in climate models: Double trouble for future rainfall projections?, Geophy. Res. Lett., 46, 2242–2252, https://doi.org/10.1029/2018GL081363, 2019.

Sanchez-Gomez, E., Cassou, C., Ruprich-Robert, Y., Fernandez, E., and Terray, L.: Drift dynamics in a coupled model initialized for decadal forecasts, Clim. Dyn., 46, 1819–1840, https://doi.org/10.1007/s00382-015-2678-y, 2016.

Sato, K., and Simmonds, I.: Antarctic skin temperature warming related to enhanced downward longwave radiation associated with increased atmospheric advection of moisture and temperature, Environ. Res. Lett., 16, 064059, https://doi.org/10.1088/1748-9326/ac0211, 2021.

Schemm, S.: Toward eliminating the decades-old “too zonal and too equatorward” storm-track bias in climate models. Journal of Advances in Modeling Earth Systems, 15, e2022MS003482. https://doi.org/10.1029/2022MS003482, 2023.

Screen, J. A., Audette, A., Blackport, R., Deser, C., England, M., Feldl, N., Gervais, M., Hay, S., Kushner, P. J., Liang, Y.-C., Msadek, R., Mudhar, R., Sigmond, M., Smith, D., Sun, L., and Yu, H.: Causes and consequences of Arctic amplification elucidated by coordinated multimodel experiments, Commun. Earth Environ., https://doi.org/10.1038/s43247-025-03052-z, 2025.

Smith, D. M., Eade, R., and Pohlmann, H.: A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction, Climate Dyn., 41, 3325–3338, https://doi.org/10.1007/s00382-013-1683-2, 2013.

Smith, D. M., Screen, J. A., Deser, C., Cohen, J., Fyfe, J. C., García-Serrano, J., Jung, T., Kattsov, V., Matei, D., Msadek, R., Peings, Y., Sigmond, M., Ukita, J., Yoon, J.-H., and Zhang, X.: The Polar Amplification Model Intercomparison Project (PAMIP) contribution to CMIP6: investigating the causes and consequences of polar amplification, Geosci. Model Dev., 12, 1139–1164, https://doi.org/10.5194/gmd-12-1139-2019, 2019.

Solaraju-Murali, B., Bojovic, D., Gonzalez-Reviriego, N., Nicodemou, A., Terrado, M., Caron, L.-P., Doblas-Reyes, F.-J.: How decadal predictions entered the climate services arena: an example from the agriculture sector, Climate Services, 27, 100303, https://doi.org/10.1016/j.cliser.2022.100303, 2022.

Tian, B., and Dong, X.: The double-ITCZ bias in CMIP3, CMIP5, and CMIP6 models based on annual mean precipitation, Geophy. Res. Lett., 47, e2020GL087232, https://doi.org/10.1029/2020gl087232, 2020.

Xue, Y., Simon, S. M., Anderson, J. R., Pegion, P., Barton, N. P., Baggett, C. F., Stan, C., Johnson, N. C., Akella, S., Becker, E., Mehra, A., Olsen, M., Shevliakova, E., Whitaker, J. S., and Zhu, J.: Advancing NOAA’s Subseasonal and Seasonal Applications and Enhancing Collaboration among Stakeholders, Modelers, and Researchers, Bull. Amer. Meteor. Soc., 106, E1295–E1302, https://doi.org/10.1175/BAMS-D-25-0060.1, 2025.

Ye, K., and Messori, G.: Inter-model spread in the wintertime Arctic amplification in the CMIP6 models and the important role of internal climate variability, Global and Planetary Change, 204, 103543, https://doi.org/10.1016/j.gloplacha.2021.103543, 2021.

Yhang, Y.-B., Lim, C.-M., and Jeong, D.: APEC climate center multi-model ensemble dataset for seasonal climate prediction, Sci. Data, 12, 303, https://doi.org/10.1038/s41597-025-04643-3, 2025.

Weisheimer, A., Baker, L. H., Bröcker, J., Garfinkel, C. I., Hardiman, S. C., Hodson, D. L. R., Palmer, T. N., Robson, J. I., Scaife, A. A., Screen, J. A., Shepherd, T. G., Smith, D. M., and Sutton, R. T.: The Signal-to-Noise Paradox in Climate Forecasts: Revisiting Our Understanding and Identifying Future Priorities, Bull. Amer. Meteor. Soc., 105, E651–E659. https://doi.org/10.1175/BAMS-D-24-0019.1, 2024.

Zhang, L., Delworth, T. L., Yang, X., Gudgel, R. G., Jia, L., Vecchi, G. A., and Zeng, F.: Estimating Decadal Predictability for the Southern Ocean Using the GFDL CM2.1 Model, J. Climate, 30, 5187–5203. https://doi.org/10.1175/JCLI-D-16-0840.1, 2017.

Zhang, X., Liu, H., and Zhang, M.: Double ITCZ in coupled ocean-atmosphere models: From CMIP3 to CMIP5. Geophy. Res. Lett., 42, 8651–8659. https://doi.org/10.1002/2015GL065973, 2015.