Comparison of Subseasonal Predictive Skill of the CFS and GEFS Operational Systems at NCEP and the Impact of Bias Correction

YANRU YANG,[a] JIE FENG,[a,b] YUEJIAN ZHU,[c] YUNTAO WEI,[a] AND JIANPING LI[d,e]

[a] *Department of Atmospheric and Oceanic Sciences and Institute of Atmospheric Sciences, Fudan University, Shanghai, China*
[b] *Shanghai Academy of Artificial Intelligence for Science, Shanghai, China*
[c] *CMA Earth System Modeling and Prediction Centre, China Meteorological Administration, Beijing, China*
[d] *Frontiers Science Center for Deep Ocean Multi-spheres and Earth System, Key Laboratory of Physical Oceanography, Academy of Future Ocean, College of Oceanic and Atmospheric Sciences, Center for Ocean Carbon Neutrality, Ocean University of China, Qingdao, China*
[e] *Laboratory for Ocean Dynamics and Climate, Qingdao Marine Science and Technology Center, Qingdao, China*

(Manuscript received 29 June 2025, in final form 7 March 2026, accepted 6 May 2026)

ABSTRACT: For subseasonal predictive skill, the relative merit of atmospheric model configuration and surface–atmosphere coupling has not been fully explored. The Climate Forecast System, version 2 (CFSv2), and Global Ensemble Forecast System, version 12 (GEFSv12), are two operational forecast systems used at the National Centers for Environmental Prediction for subseasonal prediction. CFSv2 incorporates a fully coupled system, whereas GEFSv12 utilizes simplified one-way surface boundary forcings to drive its atmospheric component. Nevertheless, GEFSv12 features a more advanced atmospheric model than CFSv2. This raises the question of how the atmospheric-only model and the surface–atmosphere coupling influence subseasonal forecast skill. Therefore, a comprehensive comparison of deterministic forecast skill between these two systems is conducted to better understand the relative merit of these two factors in subseasonal forecasting. Our evaluation reveals that, prior to bias correction (BC), GEFSv12 generally exhibits higher predictive skill than CFSv2 across most variables and regions. However, this performance contrast is largely reversed following BC, with CFSv2 outperforming GEFSv12 in the majority of cases. Diagnostic analyses suggest that the superior pre-BC skill of GEFSv12 is primarily attributable to its smaller systematic biases, probably stemming from its more advanced atmospheric model configuration. In contrast, the post-BC results indicate that CFSv2's relative strength lies in its more accurate representation of surface–atmosphere coupling, which enhances its subseasonal forecast skill despite its less advanced atmospheric component. These findings imply that further improvements in subseasonal prediction may require not only enhanced simulation of coupling processes but also continued reduction in biases associated with atmospheric model deficiencies.

SIGNIFICANCE STATEMENT: The Climate Forecast System, version 2 (CFSv2), and Global Ensemble Forecast System, version 12 (GEFSv12), are two subseasonal systems differing in their atmospheric model configuration and handling of surface–atmosphere interactions. However, the relative merits of atmospheric model configuration versus coupling processes of these two systems remain unexplored. This study provides a thorough and quantitative comparison of the subseasonal predictive skill of CFSv2 and GEFSv12. The main results reveal that GEFSv12, with a more advanced atmospheric model, outperforms CFSv2 before bias correction, while CFSv2 reversely displays overall higher subseasonal predictive skill than GEFSv12 after removing larger systematic bias. This underscores that advancing subseasonal prediction skill will likely require both improved representation of surface–atmosphere coupling processes and continued mitigation of biases arising from deficiencies in atmospheric model components.

KEYWORDS: Forecast verification/skill; Model comparison; Model evaluation/performance

1. Introduction

Subseasonal predictions encompass atmospheric forecasts that bridge the gap between medium-range weather predictions (up to 2 weeks) and short-term climate forecasts (up to one season). The subseasonal predictions have garnered significant attention due to the growing need for enhanced disaster prevention and risk mitigation, particularly in response to the increasing frequency and severity of extreme weather and climate events (Vitart et al. 2012; White et al. 2017, 2022). The primary challenge in subseasonal forecasting lies in the complexity arising from multiple sources of uncertainties, including those related to initial conditions, model configurations, and boundary forcings from factors such as oceanic, terrestrial, and stratospheric processes (Vitart et al. 2017). Consequently, quantifying the effects of these uncertainties and improving our understanding of subseasonal predictability are crucial for enhancing forecast skill on subseasonal time scales.

As the forecast horizon extends beyond the medium range (3–10 days), the slowly evolving components of Earth system

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/WAF-D-25-0134.s1.

Corresponding author: Jie Feng, fengjiefj@fudan.edu.cn

DOI: 10.1175/WAF-D-25-0134.1

TABLE 1. Model configuration of operational GEFSv12 and CFSv2.

		CFSv2	GEFSv12
Atmospheric model	Resolution	T126L64 (∼100 km horizontal, 64 vertical)	C384L64 (∼25 km horizontal, 64 vertical)
	Integration time	45 days	35 days
	Dynamical core	Spectral dynamical core	FV3
	Microphysics	Zhao–Carr microphysics	GFDL microphysics
	Ensemble member	4 perturbed members at each of 0000, 0600, 1200, and 1800 UTC	30 perturbed members initialized at 0000 UTC
Surface–atmosphere processes		Two-way interactions	One-way forcings
	Land	Noah land model	Noah LSM model (Zhu et al. 2023)
	Ocean	Modular Ocean Model, version 4 (MOM4)	NSST + two-tiered SST
	Sea ice	A three-layer interactive global sea ice model	Daily sea ice analysis from the Interactive Multisensor Snow and Ice Mapping System (Ramsay 1998)

adjacent to the atmosphere, such as land, ocean, and sea ice, play an amplifying role in modulating atmospheric evolution through coupling processes (Buizza and Leutbecher 2015; Richter et al. 2024). Consequently, leading operational centers, including the National Centers for Environmental Prediction (NCEP) and the European Centre for Medium-Range Weather Forecasts (ECMWF), have prioritized developing coupled models to enhance forecast accuracy at subseasonal scales (Vitart et al. 2017).

NCEP as one of the leading operational centers worldwide made a major advancement in 2004 by launching the Climate Forecast System (CFS), the first quasi-global, fully coupled atmosphere–ocean–land operational prediction system for S2S forecasts. In 2011, the CFS was upgraded to version 2 (CFSv2), employing a more advanced coupling framework that explicitly integrates atmosphere, ocean, sea ice, and land components (Saha et al. 2014). Systematic evaluations have highlighted the outstanding performance of CFSv2 across multiple variables and forecast metrics, as well as its skill in predicting key atmospheric phenomena such as the Madden–Julian oscillation (MJO; Saha et al. 2014), heat waves (Tang et al. 2023), and blocking events (Jia et al. 2014). Nevertheless, the atmospheric model component of the CFSv2 still relies on an outdated version of the Global Forecast System (GFS) model with a spectral dynamical core and has remained largely unchanged since 2007 (see more details in Table 1; Wang et al. 2014; Zhou et al. 2022). The lack of continuous upgrades to key model configurations including the physics package may constrain the prediction skill of CFSv2, particularly on subseasonal time scales.

In response to the limitations of the coupled CFSv2 system and the growing social demand for improved atmospheric forecasts on subseasonal scales, the NCEP has upgraded the Global Ensemble Forecast System, version 12 (GEFSv12), since 2020, which is still an atmosphere-only system. Although GEFS is originally designed for medium-range forecasts, GEFSv12 has been extended to support subseasonal forecasting, incorporating improvements in the dynamical core, horizontal resolution, and microphysics (Zhou et al. 2022; see section 2a for details). In addition, the system now has upgraded the module that simulates air–sea interactions, using bias-corrected sea surface temperature (SST) from CFSv2 forecast output as boundary conditions to simulate one-way ocean-to-atmosphere forcing (Zhu et al. 2018; Zhou et al. 2022). The GEFSv12 now provides routine subseasonal ensemble forecasts with lead times of up to 1 month.

Numerous studies have evaluated the subseasonal forecasting skill of the GEFS. Guan et al. (2022) demonstrated that the upgraded GEFSv12 outperforms its predecessor GEFSv10 across multiple performance metrics. Nevertheless, studies have shown that some variables such as 2-m temperature of GEFS present remarkable forecast bias, indicating the necessity of bias correction to further enhance forecast skill, particularly on subseasonal forecasts (Guan et al. 2019, 2022). Moreover, Zhu et al. (2023) quantified forecast uncertainty in both the coupled and uncoupled versions of GEFS, highlighting the crucial role of coupling processes in improving subseasonal prediction performance.

Although the subseasonal forecast skill of CFSv2 and GEFSv12 has been individually evaluated in numerous studies, direct comparisons of their forecast performance remain limited. For instance, Chang et al. (2022) assessed subseasonal forecasts of extratropical storminess and reported that GEFSv12 outperforms CFSv2, while a combined ensemble of the two systems yields the highest overall skill. Similarly, Saminathan et al. (2024) analyzed forecasts of temperature and precipitation over India within 15 days, revealing that CFSv2 exhibits superior skill for precipitation, whereas GEFSv12 performs better for temperature, and emphasizing the importance of bias correction in enhancing forecast performance. Furthermore, Tierra et al. (2023) demonstrated that GEFSv12 provides more accurate tropical cyclone forecasts over the Philippines during the first 2 weeks, whereas CFSv2 surpasses GEFSv12 in forecast skill at weeks 3–4.

Despite these contributions, two critical aspects remain underexplored. First, there is a lack of comprehensive and quantitative evaluations by comparing the subseasonal predictive skill of CFSv2 and GEFSv12 across multiple variables, spatial domains, and vertical levels. Existing assessments, often based on limited variable selections or specific events, may give biased conclusions regarding the relative performance of the two systems. Second, the forecast bias becomes increasingly significant in long-range (e.g., weeks 3–4) subseasonal forecasts and can notably impact the predictive skill. The influence of forecast biases on the relative performance of CFSv2 and GEFSv12 has not been thoroughly investigated in these assessments.

Although both CFSv2 and GEFSv12 employ temporally evolving SST fields for subseasonal prediction, they differ markedly in two major aspects: 1) their atmospheric model configurations and 2) interactions between Earth's surface (ocean, land, or ice) and the atmosphere at the lowest levels (referred to as surface–atmosphere coupling hereafter; Saha et al. 2014; Wang et al. 2014; Zhou et al. 2022). Given the two major differences between the operational systems CFSv2 and GEFSv12 and their use of shared SST fields in forecasts (Zhu et al. 2018; Zhou et al. 2022), a comprehensive comparison of their predictive skill at subseasonal time scales may offer valuable insights into the relative merit of atmospheric model configuration and surface–atmospheric coupling.

Building upon the aforementioned motivation, this study provides a detailed evaluation and comparison of the subseasonal predictive skill of CFSv2 and GEFSv12. We compare the predictive skill of their deterministic (or single) forecasts aiming to elucidate the influence of the processing of surface–atmosphere interactions and the atmospheric model configuration. Although disentangling the individual contributions of these factors to subseasonal forecast skill is inherently challenging, given their simultaneous differences between the two systems, we attempt a preliminary attribution through a process-oriented evaluation of characteristic atmospheric phenomena such as the MJO. We focus on three key atmospheric variables, including temperature, zonal wind, and geopotential height, across different spatial domains and vertical levels, assessing their forecast errors and predictive temporal limits globally. In particular, we examine the predictive skill of each system before and after the removal of model forecast biases, to better illustrate model bias impacts on their relative performance of forecast accuracy.

Section 2 gives a comparison of the system configurations between CFSv2 and GEFSv12 and an introduction of the datasets used in this study. Section 3 outlines the methodologies used in this research. Then, the main results regarding the spatial-mean and spatial distribution of the subseasonal predictive skill, both before and after bias correction, are presented in section 4. Finally, the results are summarized and discussed in section 5.

2. Data

a. Comparison of system configurations

This study employs two operational forecast datasets from NCEP: CFSv2 and GEFSv12. The key differences in the configurations of these two systems are summarized in Table 1. The operational CFSv2 (Saha et al. 2014) is a fully coupled atmosphere–ocean–sea ice–land model. The system is initialized four times daily at 0000, 0600, 1200, and 1800 UTC, generating four ensemble members at each cycle (https://cfs.ncep.noaa.gov/cfsv2.info/Operational.CFSv2.info.ppt). It provides forecasts extending up to 45 days with an output frequency of 6 h across all integration times. The model features a model resolution of T126L64 (approximately 100 km in the horizontal, 64 vertical levels on sigma pressure hybrid layers). Its atmospheric model component utilizes a spectral dynamical core and Zhao–Carr microphysics.

In contrast, the operational GEFSv12, introduced in 2020, is an atmosphere-only system that incorporates one-way forcing from ocean and sea ice (Zhou et al. 2022). This configuration differs substantially from the fully coupled framework of Earth system components in CFSv2. Nevertheless, the temporally varying SST used in GEFSv12 throughout the forecast lead times is sourced from the corresponding output of CFSv2, narrowing the differences in the effect of air–sea interactions between the two systems. The SST forcing data in GEFSv12's operation are additionally processed by bias correction using the CFS Reanalysis (CFSR) climatology and the CFSv2 model climatology (see more details in Zhu et al. 2017; Qi et al. 2022) as a reference. However, a systematic evaluation by Zhu et al. (2017) found that this bias correction has only a limited impact on subseasonal forecast skill. Our diagnostic shows that the spatial-mean absolute SST difference between the two systems from 50°S to 50°N remains only approximately 0.2 K averaged over the first 35 days.

Due to the computational expense of running 30 perturbed ensemble forecasts alongside one control forecast, GEFSv12 forecasts are initialized daily at 0000 UTC, with integration time extending to a slightly shorter period of 35 days. Noticeably, GEFSv12 employs the Finite-Volume Cubed-Sphere Dynamical Core (FV3; Harris et al. 2020), which was selected for the new generation of operational medium-range prediction system GFSv15.1 at NCEP since 2019 following a competitive evaluation of several candidates. In addition, GEFSv12 features more advanced GFDL microphysics and a higher spatial resolution of C384L64 (approximately 25 km in the horizontal, 64 vertical levels on sigma pressure hybrid layers).

This comparison highlights the fundamental differences in model configurations between these two systems, particularly in terms of 1) atmospheric model and 2) dealing with surface–atmosphere processes (see the summary in Table 1). It should be noted that the two systems differ in multiple configuration aspects, which complicates the attribution of their forecast skill. For clarity, the discrepancies in the dynamical core, model resolution, and physical parameterizations are collectively categorized as differences in their atmospheric model components. Meanwhile, variations related to surface–atmosphere interaction processes—such as coupling or one-way forcing—are defined as differences in surface–atmosphere processes. Notice that this study focuses on comparing the subseasonal forecast skill between GEFSv12 and CFSv2, which may arise from these two major differences. However, attributing the results to specific subsets within each category would require extensive additional numerical experiments, which is beyond the scope of this work.

b. Forecast and verification data

In this study, we assess the subseasonal predictive skill for key atmospheric variables at various vertical levels, including zonal wind and temperature at 200 hPa (U200 and T200) and 850 hPa (U850 and T850), geopotential height at 500 hPa (H500), and 10-m zonal wind (U10m), which are representative of large-scale circulation features across the upper and lower troposphere and are useful for diagnosing model performance. We evaluate the operational forecast datasets from GEFSv12 (accessible at https://registry.opendata.aws/noaa-gefs) and CFSv2 (available at https://registry.opendata.aws/noaa-cfs). These datasets are initialized at 3-day intervals from 1 October 2020 to 1 October 2023, yielding a total of 366 samples. All samples are initialized daily at 0000 UTC, with lead times up to 34 days at an output interval of 12 h. It should be noted that the GEFSv12 system provides 30 ensemble forecast members, whereas the CFSv2 system includes only four members. Given this substantial discrepancy in ensemble size between the two systems, we choose not to evaluate the performance of the ensemble-mean forecasts. Nevertheless, to assess the sensitivity of our conclusions to this choice, we conducted an additional analysis in which the ensemble means of both systems were calculated using four members, matching the ensemble size of CFSv2. This test yielded the same overall conclusion. We therefore focus on comparing the single-member forecasts from CFSv2 and GEFSv12 to ensure a relatively fair assessment. For simplification and computational efficiency, 10 perturbed members are randomly selected from the 30 available in GEFSv12, and one ensemble member is selected from the four available in CFSv2 for evaluation. Sensitivity tests further indicate that the predictive skill averaged over the selected perturbed members is largely insensitive to the number of ensemble members chosen (not shown), consistent with the findings from previous studies (Ye et al. 2023; Zhou et al. 2022; Kim et al. 2014).

The data for verifying both the GEFSv12 and CFSv2 forecasts use the NCEP–NCAR reanalysis data (Kanamitsu et al. 2002) valid at the same time. To standardize the comparison and ensure reliability, all forecast and verification datasets were interpolated onto a 2.5° × 2.5° grid, consistent with the resolution of reanalysis data. Prior to calculating forecast errors, a 5-day moving average is applied to both the forecast and verification data. This preprocessing step is a common practice in previous research, aimed at filtering out unpredictable daily variability beyond a 2-week horizon while preserving the predictable atmospheric information on subseasonal scales (Buizza and Leutbecher 2015; Xiao et al. 2024; J. Xie et al. 2020, 2024).

c. Data for bias correction

To implement an effective bias correction for the forecast data during the verification period from 2020 to 2023, GEFSv12 utilizes a 7-yr reforecast dataset preceding the verification period, i.e., from September 2012 to October 2019. This design is to ensure independence between the datasets for verification and bias correction. This reforecast dataset shares the same output interval and forecast lead times as the operational GEFSv12 (see section 2a for details). However, it includes only 10 perturbed members and 1 control forecast, available once per week on Wednesdays. For simplicity, we randomly selected 5 out of the 10 ensemble members for bias correction, maintaining a 12-h output interval consistent with the operational forecasts. For CFSv2, bias correction is based on operational forecast data from September 2012 to October 2019, aligned with the period and valid times used for GEFSv12. This approach was adopted due to the limited temporal coverage of the CFSv2 reforecast dataset, which only extends through 2011 (Saha et al. 2014). Notably, the bias correction for CFSv2 relies on a single ensemble member initialized at 0000 UTC, ensuring consistency between the correction and assessment datasets.

3. Methodology

In the following methodology, all procedures are applied to the operational forecasts from 2020 to 2023. When bias-corrected results are presented, the correction is performed using independent datasets from 2012 to 2019, as described in section 2c.

a. Calculation of RMSE and ACC

Root-mean-square error (RMSE) and anomaly correlation coefficient (ACC) are widely used metrics for evaluating forecast accuracy, one focusing on the difference between forecast and observation fields and the other measuring the spatial similarity of their anomaly fields (Li et al. 2018; Zhang et al. 2019; Pegion et al. 2019; Zhu et al. 2023; Liu et al. 2024). In this study, sample-mean RMSE and ACC are employed to comprehensively assess the predictive skill of all selected variables for a specified domain.

We categorize the globe into three main regions: the Northern Hemisphere (NH; 30°–70°N), the Southern Hemisphere (SH; 30°–70°S), and the tropics (TRO; 30°S–30°N) following Feng et al. (2023). The forecast and observation data are first preprocessed by applying a 5-day moving average centered on the third day. For the calculation of RMSE, the errors between the forecasts and corresponding observations are computed for each grid point, followed by calculating the regional-mean error variance for the NH, SH, and TRO. Finally, the RMSE is obtained by averaging the squared errors across all 366 cases for both GEFSv12 and CFSv2. The calculation of RMSE is written as

$$\mathrm{RMSE} = \frac{1}{366}\sum_{n=1}^{366}\sqrt{\overline{(f_n - a_n)^2}}. \qquad (1)$$

The metric ACC is calculated as follows:

$$\mathrm{ACC} = \frac{1}{366}\sum_{n=1}^{366}\frac{\overline{(f_n - c_n)(a_n - c_n)}}{\sqrt{\overline{(f_n - c_n)^2}\,\overline{(a_n - c_n)^2}}}, \qquad (2)$$

where n represents each individual case; f is the forecast value; a is the observed value; and c denotes the climatological value, which is using the NCEP–NCAR reanalysis dataset spanning the period from 2000 to 2019. The overbar indicates a regional average. Notice that for GEFSv12, the average RMSE and ACC for each individual ensemble member are evaluated for a fair comparison with the deterministic forecast of CFSv2, rather than the ensemble mean. The assessment of the ensemble-mean forecast skill for the subseasonal atmosphere is beyond our scope in this study because this may additionally introduce the influence of ensemble initialization for the comparison of the two systems.

b. Quantification of predictive limit

The predictive limit refers to the time beyond which the forecast, on average, loses all skill, reflecting the intrinsic unpredictability of the system due to its chaotic nature. This metric is a critical indicator of a forecast system's predictive capability. In many studies, the predictive limit is defined as the point at which the RMSE exceeds 90% or 95% of the saturated error variance (Dalcher and Kalnay 1987; Li and Ding 2011). Saturated error variance represents the maximum potential error in a prediction, constrained by the inherent limitations of the model. However, since the saturated error variance depends on the specific forecast model, it can lead to unfair comparisons between the predictive skill of GEFSv12 and CFSv2. Therefore, we follow Li et al. (2018) and adopt the 90% global attractor radius (GAR) of atmospheric variables as a more objective threshold for determining the predictive limit for both systems. The GAR is defined as the statistically averaged root-mean-square distance between two arbitrarily selected states from the historical data of observed atmosphere, which is calculated as follows:

$$\mathrm{GAR} = \sqrt{E(\|\mathbf{x} - \mathbf{y}\|^2)}, \tag{3}$$

where $E(\cdot)$ represents the mathematical expectation, $\|\cdot\|$ is the L_2 norm of a vector, and $\mathbf{x}$ and $\mathbf{y}$ are two randomly selected variable states from its long-time historical series. Li et al. (2018) demonstrated that the GAR is quantitatively equivalent to the square root of twice the climatological variance of the observed atmosphere, which is an invariant quantity for a given dynamic system (see Li et al. 2018 for further details). In a perfect model, the saturated error equals the GAR, as demonstrated by Li et al. (2018) and Feng et al. (2019). The GAR in this study is calculated using the NCEP–NCAR reanalysis dataset spanning the period from 2000 to 2019.

As for the metric ACC, the effective predictive limit is defined as the maximum forecast lead time at which the ACC remains above a skill threshold (typically 0.5), indicating that the forecast retains useful predictive skill relative to climatology. This definition is consistent with numerous previous studies and operational practice (e.g., Murphy and Epstein 1989; Chowdary et al. 2011; He et al. 2019).

4. Results

In the results section, we evaluate and compare the subseasonal predictive skill of CFSv2 and GEFSv12, focusing on the forecast RMSE and ACC and their corresponding predictive temporal limits, before and after the bias correction. The analysis includes both the spatial mean and gridpoint distribution of predictive skill for selected variables and domains, providing a comprehensive assessment. Most results will be presented in terms of the metric RMSE as ACC offers qualitatively similar conclusions. Finally, we interpret the mechanisms responsible for the observed skill differences between the two systems.

a. Predictive skill without bias correction

1) SPATIAL-MEAN PREDICTIVE SKILL

In this section, we first evaluate the spatial-mean predictive skill of both CFSv2 and GEFSv12 without applying bias correction, which provides a baseline for comparison with results after bias correction. Figure 1 shows the forecast RMSE of CFSv2 (blue solid lines) and GEFSv12 (red solid lines) as a function of lead time for key upper- and midlevel atmospheric variables across different regions. The results for lower levels and for ACC are shown in Figs. S1–S3 in the online supplemental material. Their corresponding predictive temporal limits, determined by 90% of the GAR for RMSE and 0.5 for ACC (termed as "threshold" hereafter), are summarized in Table 2.

The results in Fig. 1 and Fig. S1 show that the differences in forecast RMSE between CFSv2 and GEFSv12 are smaller within the first week but increase significantly over a 3–4-week period, indicating that model configuration differences have a greater impact over extended time ranges. For dynamical variables, such as wind and geopotential height, forecast error differences between the two systems are minimal in the extratropics and statistically insignificant. Both models' saturated forecast errors converge toward the climatological GAR (black dashed lines). In contrast, the saturated forecast errors for T200 and T850 in the NH and SH exceed the GAR by 22% on average, implying greater model deficiencies for thermodynamic variables. In the tropics, forecast errors are 21% above the GAR for all variables on average, significantly larger than those in the extratropics (approximately 9% on average), highlighting substantial model limitations in tropical regions.

Despite the overall similarities in forecast performance between the CFSv2 and GEFSv12 systems, they exhibit statistically significant differences (at the 0.05 level) in forecast error beyond 1 week, particularly for variables with notable model deficiencies. These variables include T200 in both the NH and SH, T850 in the SH, and all the variables in the tropics. Among these nine variables with significant differences, GEFSv12 demonstrates significantly lower RMSE in six of them compared to CFSv2, except for three zonal wind variables, U200, U850, and U10m in the tropics. Moreover, forecast error differences between CFSv2 and GEFSv12 are generally more pronounced at upper levels than at lower levels, possibly due to the greater variability of upper-level atmosphere. The relative performance of these variables' RMSE between CFSv2 and GEFSv12 is generally consistent with their ACC results, which demonstrates that GEFSv12 exhibits overall superior predictive skill.

We then compare the predictive limits between CFSv2 and GEFSv12 (Table 2) estimated based on their forecast RMSE and ACC (see section 3a). Overall, the predictive limits derived from ACC are generally shorter than those based on RMSE across all variables due to their different criteria for determining the predictability limit. Nevertheless, both methods yield broadly consistent results, with predictive limits mostly ranging from 1 to 2 weeks, despite the application of 5-day smoothing. This suggests significant potential for improving subseasonal forecasts at NCEP. The tropics generally exhibit the lowest predictive limits for nearly all variables compared to the NH and

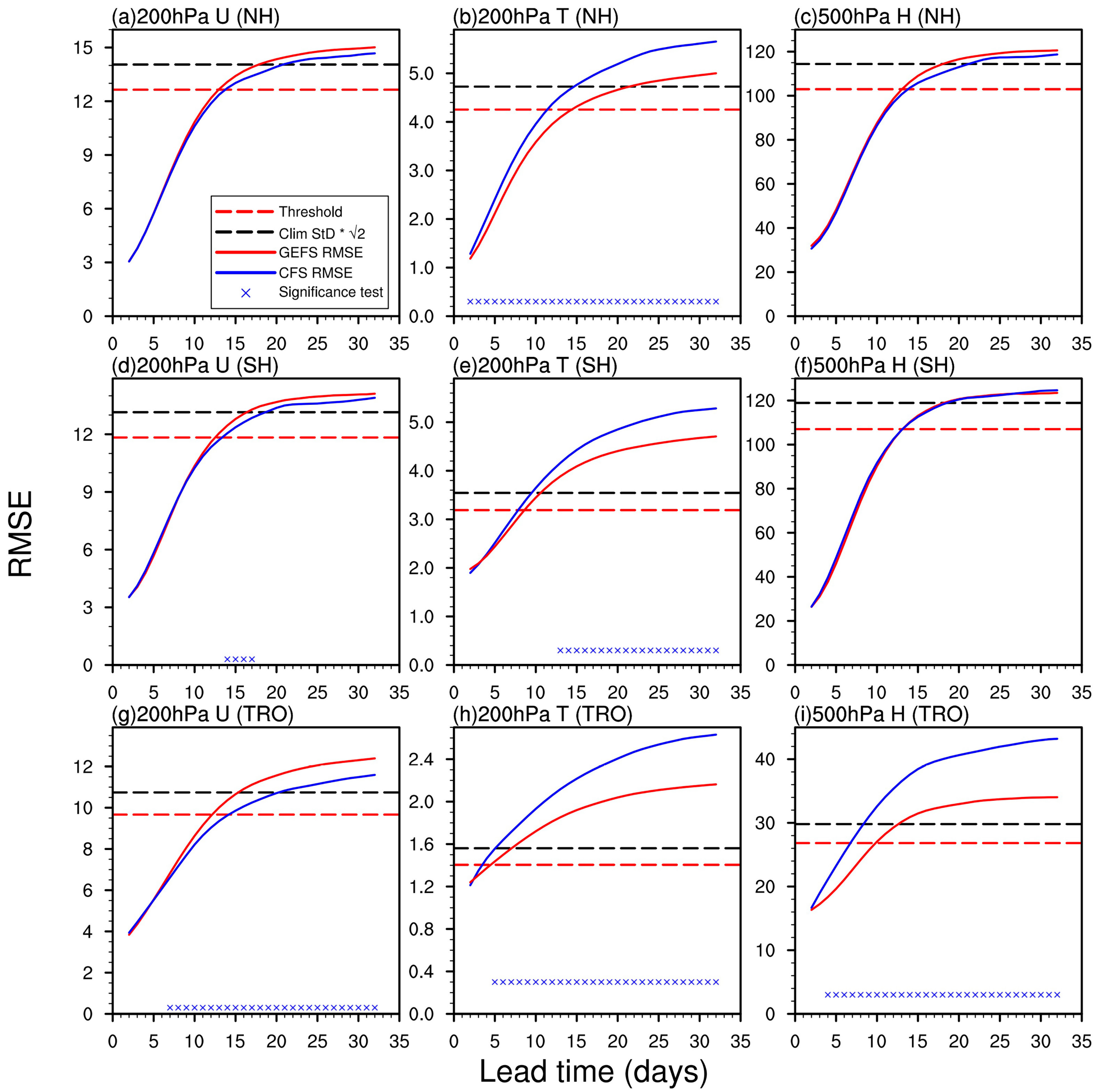

FIG. 1. The 3-yr average RMSE of variables in the NH, SH, and TRO as a function of forecast lead time for GEFSv12 (red solid line) and CFSv2 (blue solid line). (left) 200-hPa zonal wind (m s^{-1}), (center) 200-hPa temperature (K), and (right) 500-hPa geopotential height (gpm). (a)–(c) The NH; (d)–(f) the SH; and (g)–(i) the TRO. The black dashed line represents the saturated error variance, the red dashed line indicates 90% of the saturated error variance, and the dots indicate significant differences between the two systems at the 95% confidence level, based on a *t* test.

SH, particularly for temperature and geopotential height fields. Notably, the forecast horizon for T200 in the tropics is limited to only 3–4 days.

Regarding the differences in predictive limit between CFSv2 and GEFSv12, the two systems show comparable predictive limits for variables with similar forecast errors, such as U200, H500, U850, and U10m in both the NH and SH. However, for most other variables, CFSv2 exhibits significantly larger forecast RMSE values, resulting in its predictive limits being generally shorter by 1–3 days compared to those of GEFSv12. This is particularly evident for T200 in all domains, T850 in the SH, and H500 and T850 in the tropics. This is further supported by the ACC scores, which are higher for GEFSv12 than for CFSv2 across nearly all variables.

2) SPATIAL DISTRIBUTION OF PREDICTIVE LIMIT

To have an overview of the predictive skill across the global regions, we assess the spatial distribution of predictive limits based on RMSE for both CFSv2 and GEFSv12 from 80°S to

TABLE 2. Spatial-mean predictive limits for CFSv2 and GEFSv12 before BC (days).

			U200	T200	H500	U850	T850	U10m
NH	CFSv2	RMSE	13.7	11.4	13.8	12.1	10.3	13.0
		ACC	9.5	7.7	9.4	8	8	7.3
	GEFSv12	RMSE	12.9	14.4	13.0	11.6	10.8	13.3
		ACC	9.5	8.9	9.4	8.1	8.4	7.4
SH	CFSv2	RMSE	13.4	7.8	13.1	12.3	7.2	12.2
		ACC	9.4	9.3	9.3	7.8	7.1	7.1
	GEFSv12	RMSE	12.5	8.6	13.1	12.0	8.3	12.7
		ACC	9.6	9.7	9.8	8.2	7.8	7.5
TRO	CFSv2	RMSE	14.2	3.5	6.8	10.8	8.5	11.8
		ACC	10	3.2	5.7	7.8	6.9	7.6
	GEFSv12	RMSE	12.1	4.5	9.8	10.6	9.0	10.7
		ACC	9.9	3.4	7.8	8.1	7.7	7.7

80°N with variables consistent with the spatial mean given in the previous part. The gridpoint predictive limit is determined by identifying the lead time at which the gridpoint RMSE exceeds 90% of the local saturated error variance. Notably, we applied a horizontal Gaussian filtering of gridpoint predictive limit with a radius of 300 km (i.e., the standard deviation of the Gaussian distribution is 300 km) centered at each grid point (Wahr et al. 1998; Swenson and Wahr 2006). This approach avoids abrupt changes in predictive limits between adjacent grids due to the noises in forecast errors, thereby yielding a more coherent and physically interpretable map of predictive limits.

Figures 2 and 3 depict the spatial distribution and zonal mean of predictive limits for upper- and lower-level variables of CFSv2 and GEFSv12, respectively. These two systems display generally similar spatial distributions, with predictive limits generally shorter in tropical regions compared to extratropical areas, especially for T200, H500, and U10m. This can be largely attributed to significant systematic bias in forecast models for tropical regions (see discussion in section 4b). Both models exhibit relatively higher predictive limits, generally exceeding 20 days, for T850 over the tropical eastern Pacific and for U10m over the tropical western Pacific. This is likely influenced by large-scale ocean–atmosphere interactions (see further discussion in section 4c). Additionally, the spatial distribution of predictive limits for U10m is notably affected by complex topography. For instance, the Arctic regions demonstrate relatively higher predictive limits beyond 1 month, potentially due to the predictability source from boundary forcing of sea ice (Domeisen et al. 2022; Sun et al. 2015). In contrast, areas near high-altitude terrain—such as the Tibetan Plateau, eastern Africa, and the Rocky Mountains—show significantly lower predictive limits of less than 4 days.

The relative differences in zonal-mean predictive limit between CFSv2 and GEFSv12 are generally consistent with the spatial-mean differences across different variables. In particular,

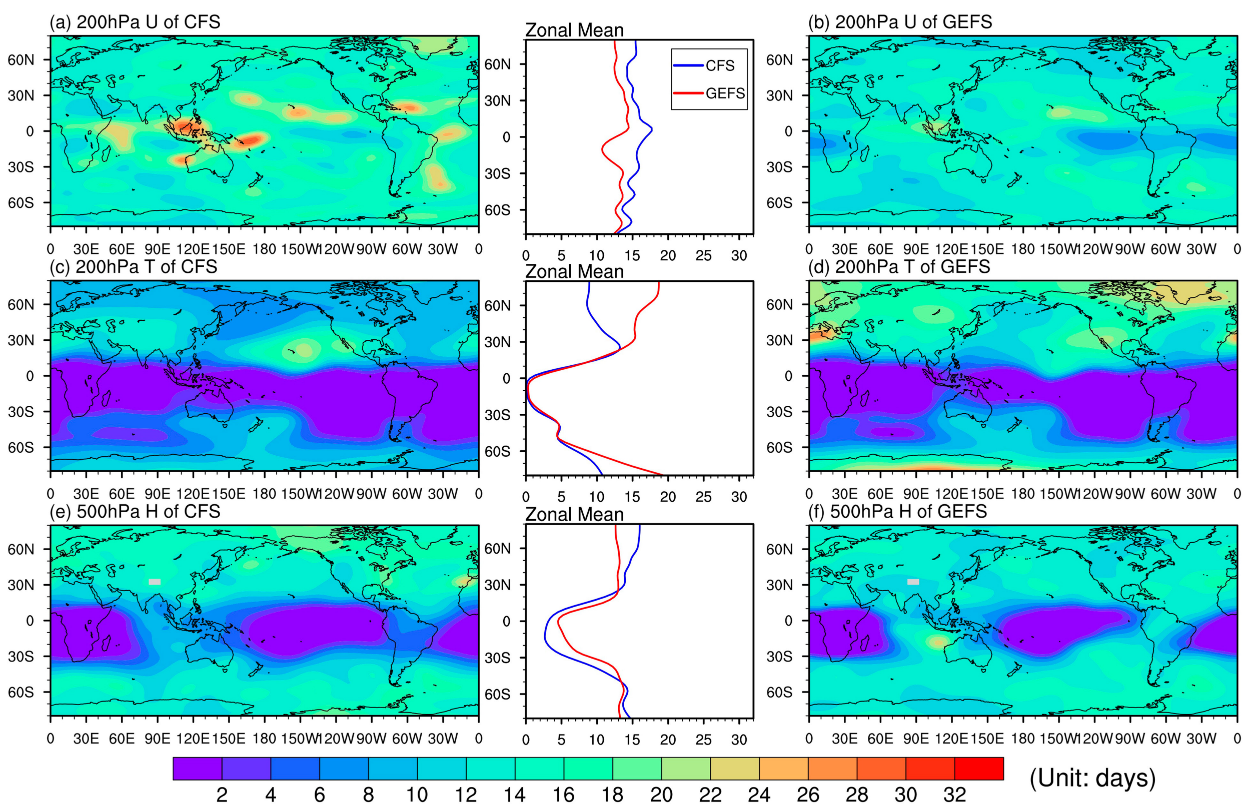

FIG. 2. Spatial distribution of the predictive limit (day) for (left) CFSv2 and (right) GEFSv12. (center) The zonal mean of the predictive limit for CFSv2 (blue solid line) and GEFSv12 (red solid line). The variables represented are as follows: (a),(b) 200-hPa zonal wind, (c),(d) 200-hPa temperature, and (e),(f) 500-hPa geopotential height.

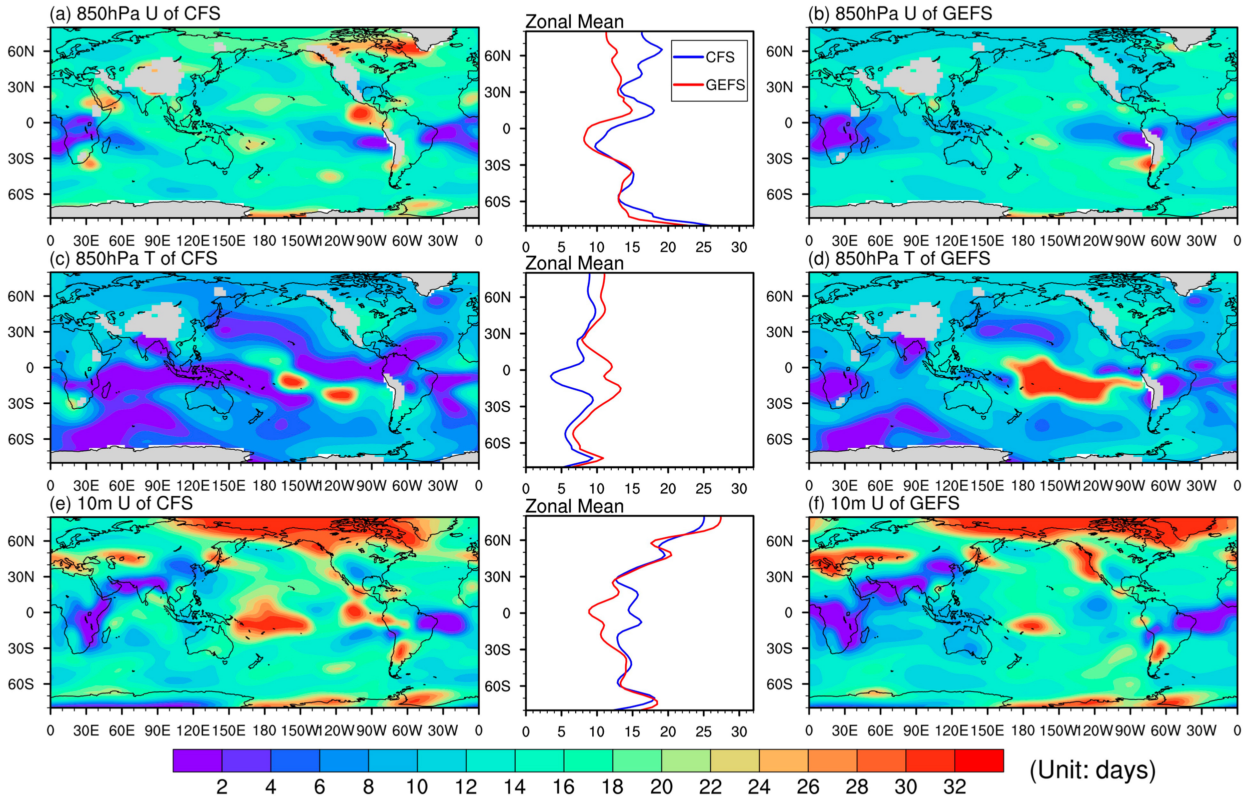

FIG. 3. As in Fig. 2, but corresponding to (a),(b) 850-hPa zonal wind, (c),(d) 850-hPa temperature, and (e),(f) 10-m zonal wind.

for temperature variables such as T200 and T850, GEFSv12 presents higher predictive limits than CFSv2 across the NH, SH, and TRO domains. Overall, we have two hypotheses to explain the superior forecast performance of GEFSv12 relative to CFSv2. First, GEFSv12 incorporates a more advanced atmospheric model configuration, including an improved dynamical core, higher spatial resolution, and more sophisticated microphysics schemes, which may enhance its ability to simulate atmospheric variability. Second, GEFSv12 may be characterized by smaller systematic forecast biases due to atmospheric model advances mentioned above compared to CFSv2. The specific role of model bias can be further clarified through bias-correction experiments applied to both systems. Accordingly, the following sections will examine the impact of model bias on forecast skill by applying bias correction to both systems.

b. *Forecast bias*

In the following sections, we evaluate and compare the forecast performance of CFSv2 and GEFSv12 after applying bias correction at each lead time. Biases in the temporal-mean forecast states relative to observational references constitute a major source of model forecast error. Prior to the performance comparison, we present the temporal evolution and spatial distribution of forecast biases for both CFSv2 and GEFSv12 (see more details about the bias-correction method in appendix A). Although the forecast biases of CFSv2 and GEFSv12 have been individually assessed for specific variables, a systematic comparison of their forecast biases has not yet been conducted before.

Figure 4 presents the spatial-mean bias, calculated as the absolute value of bias, as a function of lead time for both systems. Given that all six variables exhibit similar patterns, the analysis focuses on zonal wind (Figs. 4a–c) and temperature (Figs. 4d–f) at upper levels as representative variables. The bias in the 200 hPa is generally larger than that in a lower level (not shown). The results reveal that the bias of both systems generally increases with forecast time, except for the T200 in the tropics for GEFSv12. Additionally, CFSv2 consistently demonstrates larger domain-averaged bias compared to GEFSv12 across nearly all lead times, regions, and variables. Notably, by day 16, the growth rate of model bias tends to decelerate, marking a point of relative stability.

To gain deeper insights, we also analyzed the spatial distribution of forecast bias across grid points for all lead times (figures not shown). The results reveal that the spatial patterns of bias are generally consistent across different forecast days. Considering the observed deceleration in the growth rate of bias around day 16, we present the gridpoint distribution of forecast bias specifically for this lead time as a representative example, shown for CFSv2 (Fig. 5) and GEFSv12 (Fig. 6) across all six variables. To facilitate a consistent comparison across regions, the forecast bias at each grid point is standardized using the climatological standard deviation, enabling an evaluation of relative bias.

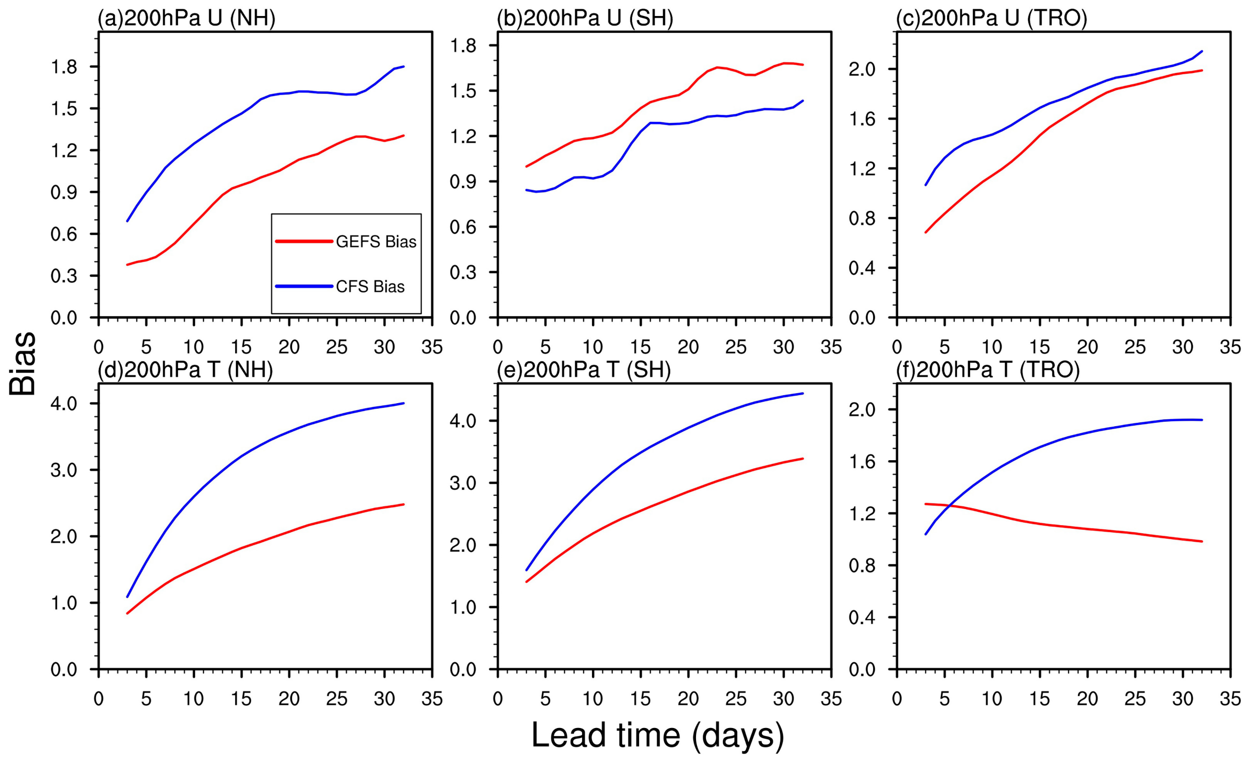

FIG. 4. The 3-yr average bias of variables in the NH, SH, and TRO as a function of forecast lead time for GEFSv12 (red solid line) and CFSv2 (blue solid line). (top) The 200-hPa zonal wind bias (m s^{-1}) and (bottom) the temperature bias (K).

A comparison of forecast biases in CFSv2 (Fig. 5) and GEFSv12 (Fig. 6) reveals a notably similar spatial distribution of biases across all variables, despite the two systems utilizing distinctly different atmospheric model configurations and surface–atmosphere interaction frameworks. This similarity in forecast biases may be attributed to their consistent use of SST over time, as GEFSv12 employs SST predictions from CFSv2 as a one-way boundary forcing for atmospheric evolution.

Overall, the magnitude of forecast biases is consistently larger in CFSv2 than in GEFSv12 across all variables. These larger biases in CFSv2 may stem from two primary sources. First, they may reflect limitations inherent in the older GFS atmospheric model configuration used within CFSv2. Second, deficiencies in the surface–atmosphere coupling framework could contribute systematic errors, particularly through misrepresentation of surface energy and mass fluxes between the atmosphere and other Earth system components (Rai and Saha 2018; Goswami et al. 2014). By contrast, the reduced biases in GEFSv12 likely arise from advancements in model configuration, including higher horizontal resolution, updated physical parameterizations, and the implementation of the more advanced FV3 dynamical core (Guan et al. 2022; Zhou et al. 2022). Nonetheless, a more definitive attribution of forecast bias sources will require targeted numerical experiments and comprehensive diagnostic analyses.

For both upper- and lower-level zonal winds (U200 and U850), forecast biases predominantly originate from tropical regions, particularly over the eastern Pacific extending into the Atlantic Ocean and North Africa, as well as the Maritime Continent. In these tropical areas, the forecast bias for U200 and U850 averages around 0.33 times the climatological standard deviation, compared to approximately 0.18 times in extratropical regions. The biases are relatively more pronounced in CFSv2 than in GEFSv12 within the tropics. To further diagnose these discrepancies, we examine lead–longitude composites of outgoing longwave radiation (OLR) and 850-hPa zonal wind anomalies associated with strong MJO events (phase-2 initializations; amplitude > 1; 15°S–15°N) in original forecasts. The analysis indicates that CFSv2 exhibits weaker and less coherent eastward propagation compared to GEFSv12 (figure not shown), which may contribute to CFSv2's larger bias. Such deficiencies of CFSv2 are possibly associated with limitations of the convective parameterization schemes (Weber and Mass 2017) and misrepresentation in air–sea coupling processes (Wang et al. 2014).

For T200, T850, and H500, both CFSv2 and GEFSv12 exhibit globally pervasive negative biases, with the most pronounced errors concentrated in tropical regions. In particular, forecast biases in CFSv2 across much of the tropics exceed 1.5 times the climatological standard deviation, whereas those in GEFSv12 largely remain within one standard deviation. The substantial temperature biases in the tropics observed in CFSv2 may stem from deficiencies in its convective parameterization (Weber and Mass 2017; Goswami et al. 2014;

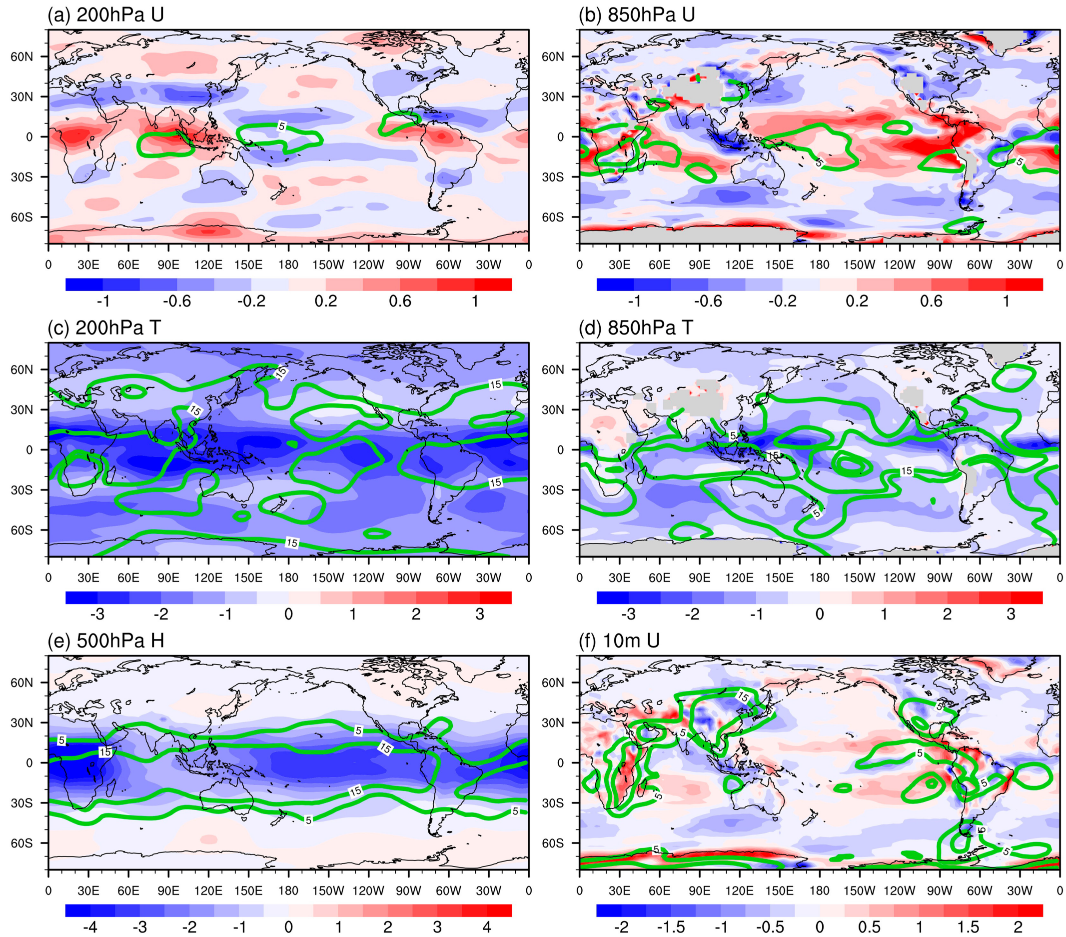

FIG. 5. Bias of CFSv2 at the 16th forecast day, standardized by the climatological standard deviation (shading). The contour lines represent the differences in predictive limits (days) with and without BC; only the 5- and 15-day contours are shown.

Hazra et al. 2016). Given that convective activity is most vigorous in the tropics, inadequacies in representing these processes can lead to distorted vertical heating profiles and cloud fractions. Such errors further propagate into significant biases in radiative effect, ultimately affecting the representation of upper-tropospheric circulation and thermodynamic structures. Additionally, substantial forecast biases in the 10-m zonal wind of both systems appear closely associated with complex topographic features, possibly arising from unresolved orographic flows and limitations in boundary layer parameterization (Huang et al. 2015), particularly over regions characterized by steep elevation gradients, such as the Tibetan Plateau.

c. Predictive skill with bias correction

Given the observed differences in forecast bias between CFSv2 and GEFSv12, it is of interest to evaluate and compare their forecast performance after bias correction. By removing these model biases to some extent, we enhance the comparability of the forecast performance for both systems, thereby providing a clearer assessment of their predictive skill regarding atmospheric internal variability.

1) SPATIAL-MEAN PREDICTIVE SKILL

Figures 7 and 8 display the domain-averaged RMSE and ACC, respectively, for the same mid- and upper-level variables as in Fig. 1, but with bias correction applied. The results for the low-level variables are shown in Figs. S4 and S5. The forecast RMSE values with model bias removed are overall reduced for all variables. This reduction is particularly pronounced for the saturated values of forecast RMSE. Specifically, the forecast saturated values for both CFSv2 and GEFSv12 are now closely aligned with the GAR, exhibiting only a 5% difference on average. This small difference in saturated forecast errors is in stark contrast with the substantial differences observed in temperature and tropical variables in Fig. 1 and Fig. S1, highlighting significant systematic biases in their forecasts. Notably, the RMSE for H500 shows minimal change before and after bias correction in mid- and high latitudes of both the NH and SH (Figs. 7c,f). This is consistent with the negligible forecast bias of H500 in both the NH and SH (see Figs. 5e and 6e) and previous studies that assume that the simulation of upper-level atmospheric circulation in mid- and high latitudes is nearly bias free in current general circulation models (Buizza and Leutbecher 2015). All

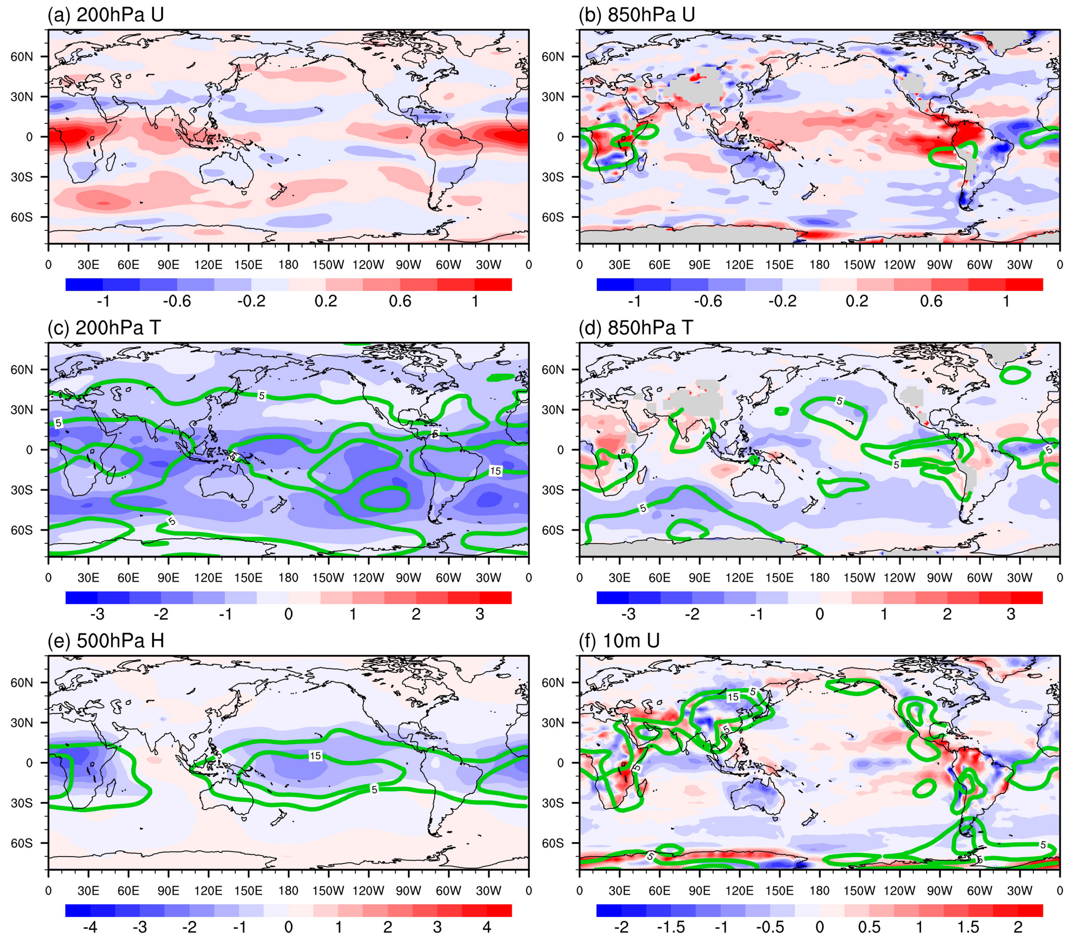

FIG. 6. As in Fig. 5, but for GEFSv12.

results in Figs. 7 and 8 underscore the effectiveness of the applied bias-correction method.

Although the gap in forecast error (Fig. 7 and Fig. S4) between CFSv2 and GEFSv12 narrows after bias correction, the variables exhibiting statistically significant differences remain largely consistent with those observed before correction, especially beyond the 2-week lead time. Notably, in stark contrast to the precorrection results, CFSv2 now demonstrates lower RMSE values for all variables showing significant differences in forecast error compared to GEFSv12. For ACC evaluation (Fig. 8 and Fig. S5), the overall inferior forecast skill of CFSv2 compared to GEFSv12 prior to bias correction now transfers to a negligible difference between the bias-corrected forecast skill of the two systems. The similarity is not an artifact of temporal averaging, as the daily mean ACC also displays highly consistent behavior between the two systems (figure not shown). This shift indicates that, once biases are removed, the fully coupled CFSv2 system exhibits overall superior predictive skill in predicting atmospheric internal variability compared to the atmosphere-only GEFSv12, which operates under a one-way forcing from Earth's surface. Previous studies (e.g., Ji and Toepfer 2016; Magnusson et al. 2022) have shown that the GFS with the upgraded FV3 dynamic core presents higher forecast skill compared to the version with the old spectral dynamical core given the same initial condition. Consequently, it is plausible that the relative improvement in CFSv2's performance, despite retaining the older dynamical core, possibly stems from its fully coupling framework that simulates the interactive processes between the atmosphere and other Earth system components, such as the ocean and sea ice. This coupling is crucial for capturing the critical interactions that influence atmospheric variability on subseasonal time scales.

After applying bias correction, the predictive limits derived by both RMSE and ACC measurements, as shown in Table 3, indicate a significant improvement in forecast performance for all variables in both CFSv2 and GEFSv12. Overall, the ACC-based predictive limits are shorter than those derived by RMSE as the results prior to correction. For variables exhibiting relatively smaller bias, such as U200, U850, and U10m, both systems show relatively modest enhancements in their predictions across all regions, as well as for H500 in the extratropics, with predictive limits extending by approximately 2 days for both metrics. However, for variables characterized by significant forecast biases, such as temperature and H500 in the tropics, CFSv2 exhibits more significant enhancements of predictive limits compared to GEFSv12. This improvement

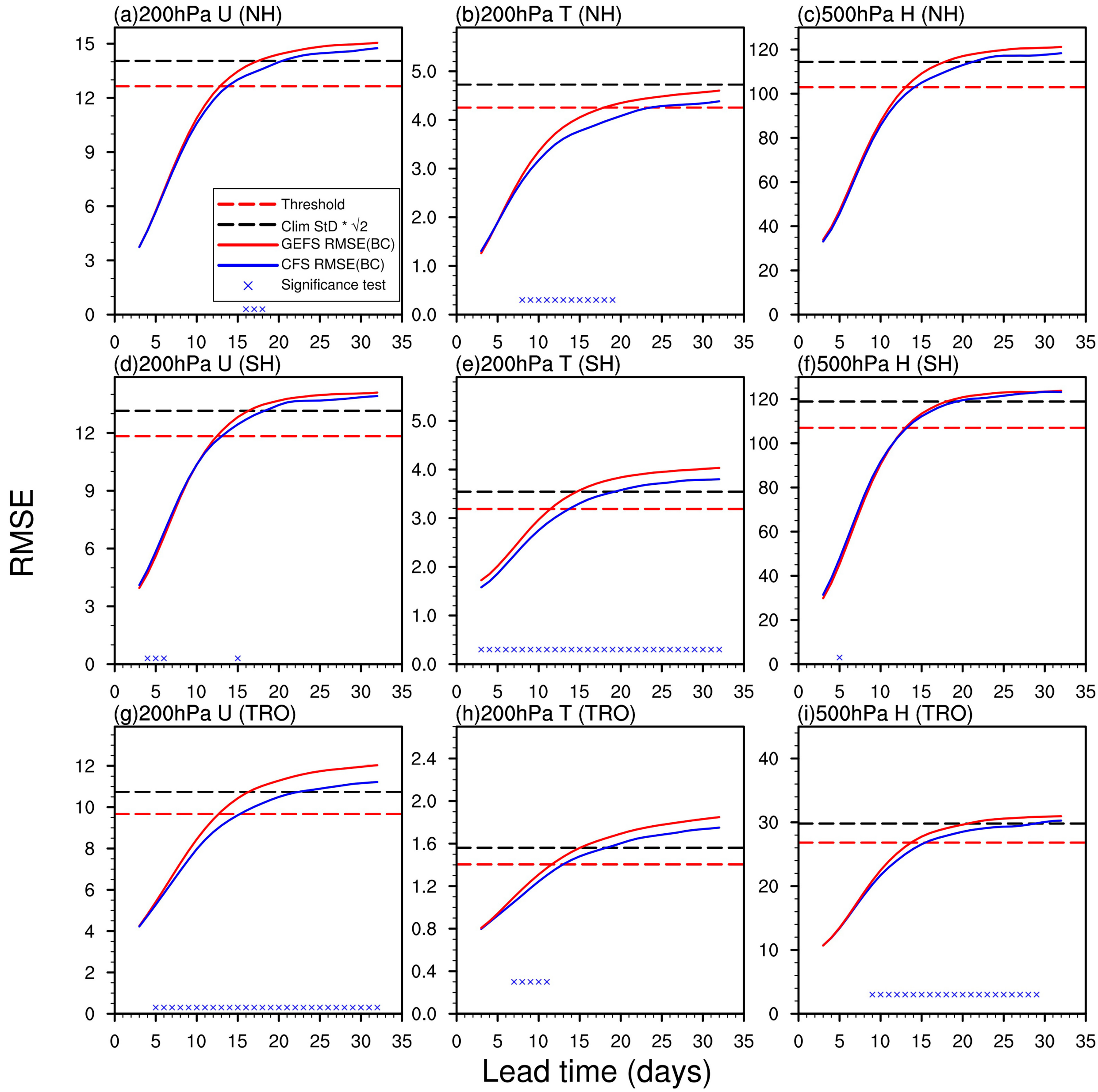

FIG. 7. As in Fig. 1, but for results after applying BC.

is particularly notable in the upper atmospheric layers. For example, in CFSv2, the RMSE-based forecast horizon for T200 in the NH extends to 23 days—double the forecast limit prior to bias correction. Additionally, the predictive limits for T200 in the tropics extend by approximately 10 days (RMSE) and 8 days (ACC).

A notable outcome of this part is that CFSv2 generally outperforms GEFSv12 across the majority of variables after correcting for forecast bias, marking an overall reversal from previous results. Although numerous studies have demonstrated the critical role of bias correction in enhancing subseasonal prediction skill based on evaluations of either CFSv2 or GEFSv12 individually, our findings reveal that bias correction may actually reverse the relative predictive skill between the two systems for some variables. In summary, the more advanced atmospheric model configuration in GEFSv12 (i.e., the FV3) likely contributes to its reduced forecast biases. However, after bias correction, GEFSv12 does not exhibit a clear advantage over CFSv2 in subseasonal forecast skill (3–4 weeks) based on our sample-mean evaluations, despite their comparable performance at medium-range lead times (1–2 weeks). This finding suggests that advancements in fully coupled model frameworks may play a more critical role in subseasonal prediction than improvements in the atmospheric model alone. Nonetheless, it is

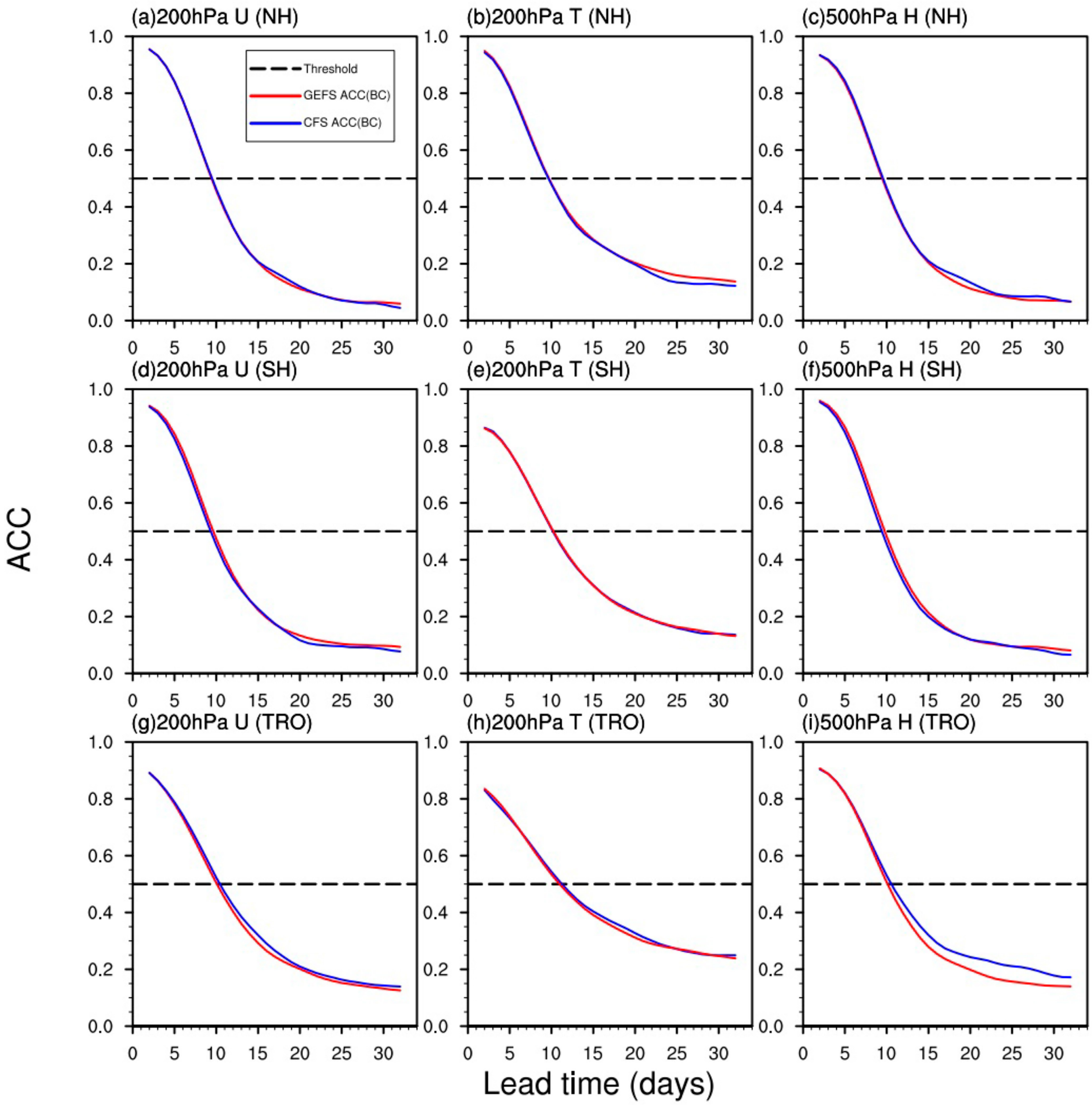

FIG. 8. The 3-yr average ACC of variables in the NH, SH, and TRO as a function of forecast lead time for GEFSv12 (red solid line) and CFSv2 (blue solid line). (left) 200-hPa zonal wind, (center) 200-hPa temperature, and (right) 500-hPa geopotential height. (a)–(c) The NH; (d)–(f) the SH; and (g)–(i) the TRO. The black dashed line represents the threshold of 0.5.

important to recognize that the enhanced atmospheric configuration in GEFSv12 may still yield improvements for specific weather phenomena or in particular regions. Addressing these possibilities will require targeted case studies and process-oriented analyses in future work.

2) SPATIAL DISTRIBUTION OF PREDICTIVE LIMIT

Figures 9 and 10 illustrate the spatial distribution of predictive limits for the CFSv2 and GEFSv12 models after bias correction according to forecast RMSE, separately for the upper and lower levels. The findings indicate that, for both systems and across all variables, the predictive limits at individual grid points and the zonal mean have generally expanded following bias correction. Notably, CFSv2 exhibits longer zonal-mean predictive limits than GEFSv12 across most latitudes and variables (cf. black and red solid lines), as highlighted by the spatial-mean results.

For the U200 variable (Figs. 9a,b), the predictive limits in the extratropics for both CFSv2 and GEFSv12 show minimal

TABLE 3. Spatial-mean predictive limits for CFSv2 and GEFSv12 after BC (days).

			U200	T200	H500	U850	T850	U10m
NH	CFSv2	RMSE	13.8	23.6	14.1	12.3	11.6	14.3
		ACC	9.5	9.6	9.6	8	8.8	7.4
	GEFSv12	RMSE	12.9	17.9	13.0	11.7	11.4	14.3
		ACC	9.4	9.7	9.5	8.1	8.8	7.6
SH	CFSv2	RMSE	13.4	13.7	13.2	12.5	9.5	13.3
		ACC	9.3	10.2	9.4	7.9	8.1	7.4
	GEFSv12	RMSE	12.5	11.4	13.1	12.1	10.0	13.8
		ACC	9.6	10.2	9.8	8.3	8.5	7.8
TRO	CFSv2	RMSE	15.3	13.0	15.5	12.4	12.0	13.2
		ACC	10.4	11.3	10.6	8.3	8.8	7.9
	GEFSv12	RMSE	12.7	11.6	13.8	11.5	10.9	11.9
		ACC	10.1	11	10.1	8.4	8.7	8

change before and after bias correction. However, in the equatorial regions, the predictive limit of CFSv2 is notably extended by approximately 3 days, while GEFSv12 shows negligible change as a result of the larger tropical forecast bias in CFSv2 (cf. Figs. 5 and 6). Consequently, CFSv2 achieves a predictive limit about 5 days longer than GEFSv12 in the tropics, reaching approximately 20 days. This extension is primarily observed across the region spanning from the northern Indian Ocean, eastward through the Maritime Continent, to the western Pacific Ocean. This area is likely influenced by the MJO, which serves as a key predictability source (Kim et al. 2018; Zhang et al. 2013). Similarly, for the H500 (Figs. 9e,f) and U850 (Figs. 10a,b) variables, the bias correction results in an approximately 7- and 5-day longer predictive limit for CFSv2 compared to GEFSv12 near equator, respectively, a reversal of the pattern observed prior to bias correction. These findings underscore the positive impact of incorporating air–sea coupling in the tropics on subseasonal predictability.

For T200 (Figs. 9c,d), the extension of predictive limits is particularly pronounced not only in the tropics but also across mid- and high-latitude regions, especially in the Arctic, Antarctic, and Tibetan Plateau, for both CFSv2 and GEFSv12. This extension ranges from 5 to 20 days, suggesting a notable forecast bias for T200 in high-latitude regions. This bias may be linked to the limited capability of forecast models to accurately represent atmosphere–topography interactions, especially over complex terrains like mountains and sea ice (Zadra et al. 2018). Nonetheless, CFSv2 extends predictive limits more effectively than GEFSv12 in both the extratropics and the tropics following bias correction. Both CFSv2 and GEFSv12 exhibit a local minimum in predictive limits over the Maritime Continent and western Pacific, likely due to the predictability barrier posed by large-scale convective cloud systems (Wang et al. 2019). Interestingly, a wavelike pattern of maximum predictive limit is evident from the central Pacific Ocean extending toward North America in both systems, possibly associated with the Pacific–North

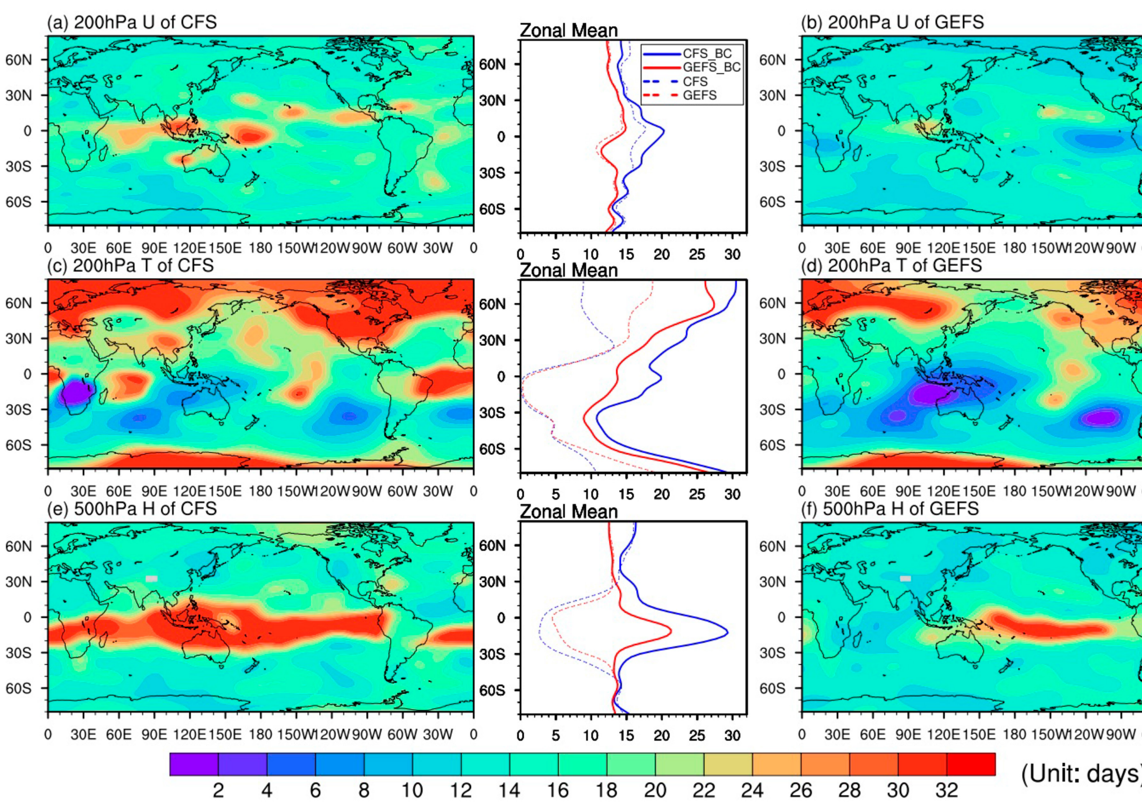

FIG. 9. As in Fig. 2, but for results after applying BC. The dashed lines represent the nonbias-corrected results for zonal-mean predictive skill.

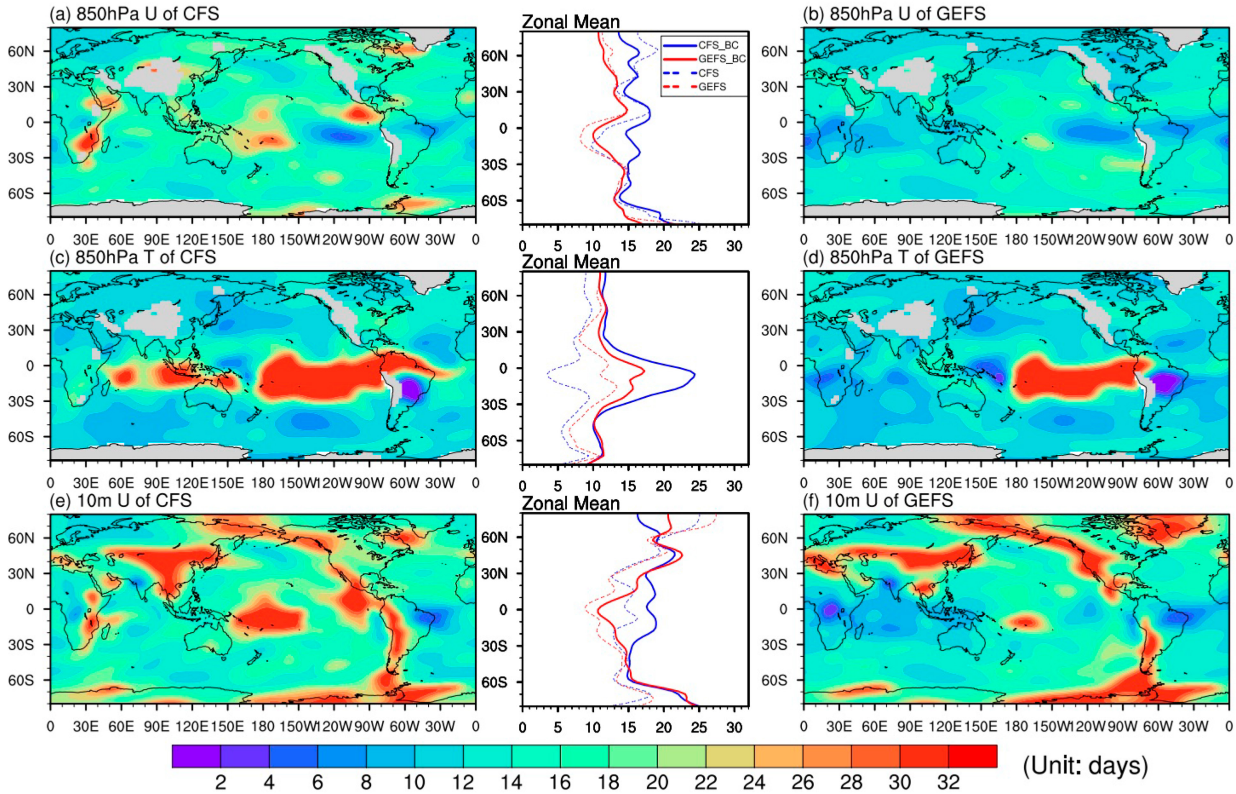

FIG. 10. As in Fig. 3, but for results after applying BC. The dashed lines represent the nonbias-corrected results for zonal-mean predictive skill.

America (PNA) teleconnection pattern (Wang and Robertson 2019). Furthermore, CFSv2 displays generally longer predictive limits along this wave pattern, suggesting that improvements in surface–atmosphere two-way coupling in the tropics may enhance predictability in extratropical regions.

For T850 (Figs. 10c,d), unlike upper-level temperatures, the predictive limit varies minimally in the mid- and high latitudes before and after bias correction for both CFSv2 and GEFSv12. In the tropics, however, the predictive limit of CFSv2 extends by nearly 20 days following bias correction, in contrast to a roughly 5-day extension for GEFSv12. This considerable improvement in CFSv2 may be linked to its enhanced ability to simulate air–sea coupling processes associated with the MJO (see more details in section 4d), particularly across the Indian Ocean, Maritime Continent, and western Pacific region. Furthermore, predictive limits in the eastern tropical Pacific exceed 30 days for both systems, likely reflecting the influence of predictability signals originating from El Niño–Southern Oscillation (ENSO) forcing (Vitart et al. 2017; R. Xie et al. 2020). For the U10m variable (Figs. 10e,f), bias correction primarily enhances predictive limits in mid- and high-latitude land regions with complex topography for both CFSv2 and GEFSv12. Additionally, the predictive limits in the extratropics are approximately 5–10 days longer than those in the tropics.

The comparison of predictive limits before and after bias correction reveals a significant impact of model bias on predictive skill at subseasonal time scales, especially beyond 2 weeks. Notably, removing the larger forecast bias in CFSv2 does not necessarily lead to its higher subseasonal forecast skill compared to GEFSv12. The forecast skill comparison after bias correction further highlights the advantage of CFSv2's inclusion of surface–atmosphere coupling processes. This conclusion is also supported by Zhu et al. (2023), who compared coupled and uncoupled versions of GEFSv12 and found that the coupled system generally outperforms the uncoupled one across many variables at subseasonal time scales.

d. *MJO forecast skill in bias-corrected forecasting systems*

The MJO is an atmospheric phenomenon arising from strong air–sea interactions (Wei et al. 2018) and represents a dominant mode of intraseasonal variability in the tropics. It serves as a major source of subseasonal predictability and exerts substantial influence on weather and climate across both tropical and extratropical regions (Wei et al. 2023). Our analysis has shown significantly higher predictive skill of CFSv2 for all variables compared to GEFSv12 in the tropics. These distinctions are particularly evident along the MJO's main path, which stretches from the Indian Ocean to the western Pacific. This suggests that the improved predictive skill of these variables in the fully coupled CFSv2 might be related to its enhanced simulation of the MJO following removal of bias.

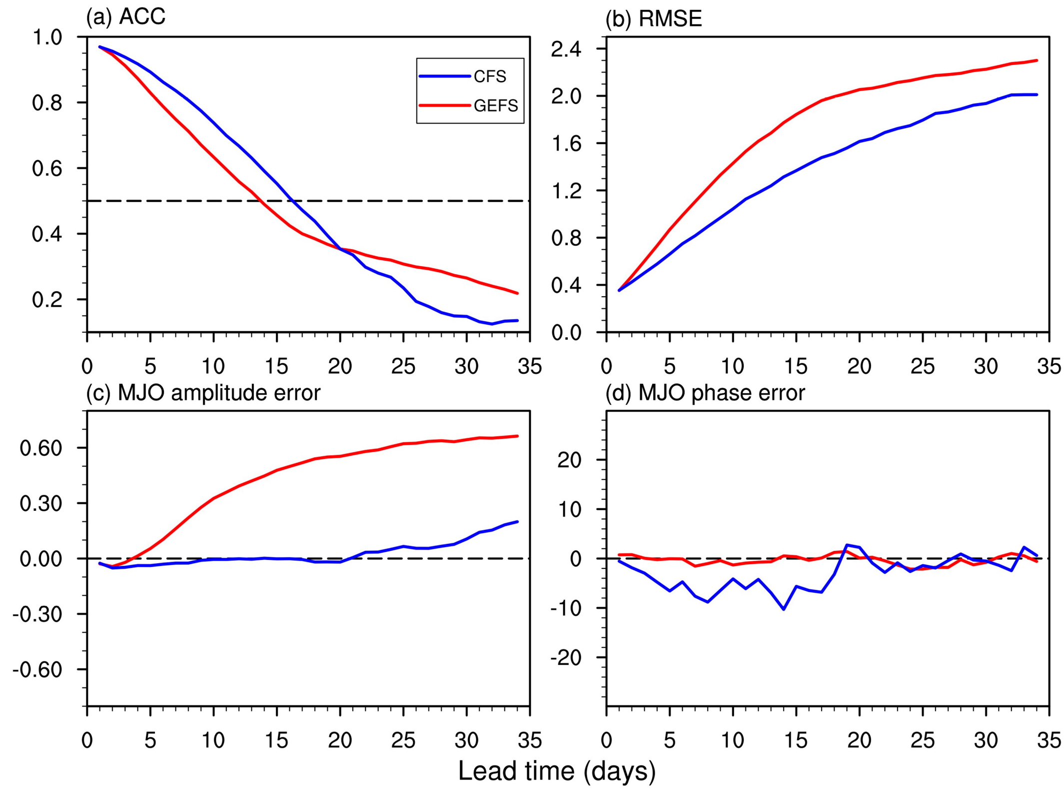

FIG. 11. Evaluation of MJO forecast skill based on the RMM index, derived from the first two principal components (RMM1 and RMM2) of combined OLR, 850-hPa zonal wind, and 200-hPa zonal wind anomalies. (a) ACC, (b) RMSE, (c) amplitude error, and (d) phase error as a function of lead time for CFSv2 (blue) and GEFSv12 (red).

To substantiate this conjecture, we first validate the superior simulation of the MJO in bias-corrected CFSv2, as compared to bias-corrected GEFSv12. Following previous studies (e.g., Huang et al. 2026), we quantify the MJO using the forecasted real-time multivariate MJO (RMM) index (Wheeler and Hendon 2004; see appendix B for details). Figure 11a shows the bivariate ACC, where a threshold of 0.5 is used to define the effective predictive limit. For CFSv2, the predictive skill extends up to 17 days, which is 3 days longer than GEFSv12. Note that the RMM skill represented here is lower than in previous studies (e.g., Zhu et al. 2023; Saha et al. 2014), primarily because we assess the average forecast skill of individual ensemble members rather than the ensemble mean. Across all lead times, CFSv2 demonstrates smaller errors compared to GEFSv12 (Fig. 11b). The reduced RMSE in CFSv2 is mainly associated with the reduced error of RMM amplitude (Fig. 11c vs Fig. 11d). Therefore, the RMM index is better forecasted in CFSv2, as compared to GEFSv12.

The contrasts in the amplitude and phase errors between the two models motivate us to further examine the anomalous convection and circulation of the MJO. Figure 12 shows meridionally averaged (15°S–15°N) intraseasonal anomalies of OLR and U850 from observations (OBS) and model forecasts initialized in RMM phase 2 with strong MJO activity (initial amplitude > 1.0). The OBS shows an eastward-propagating MJO mode characterized by deep convection coupled with planetary-scale zonal winds (Fig. 12a). In CFSv2, the coupled system of the MJO is well forecasted, but with much slower phase speed, which is consistent with its much smaller error in RMM amplitude but larger error in RMM phase (see Fig. 12d). The GEFSv12 also forecasts an eastward-propagating mode with similar phase speed as the OBS, but the deep convection decays quickly and encounters a strong barrier effect of the Maritime Continent, due mainly to a westward intrusion of dry signals (Fig. 12c). These contrasts indicate that the MJO forecast skill is enhanced in CFSv2 after BC, as compared to GEFSv12.

Then, we examine to what extent the enhanced simulation of the MJO in CFSv2 can contribute to its higher subseasonal forecast skill following BC. We applied a regression-based method to remove the components associated with the RMM index from U200, T200, and H500 and subsequently recalculated their gridpoint ACC. The results (not shown) reveal that the predictive skill of both CFSv2 and GEFSv12 decreases notably by over 50% averaged over all lead times once the MJO-related signals are excluded. This indicates that the MJO can be efficiently translated into subseasonal forecast skill of U200, T200, and H500. Consequently, that CFSv2 has a higher subseasonal forecast skill than GEFSv12 should be partly explained by the better simulation skill of the MJO in CFSv2 than in GEFSv12.

The superior postcorrection performance of CFSv2 underscores the importance of atmosphere–ocean coupling and/or

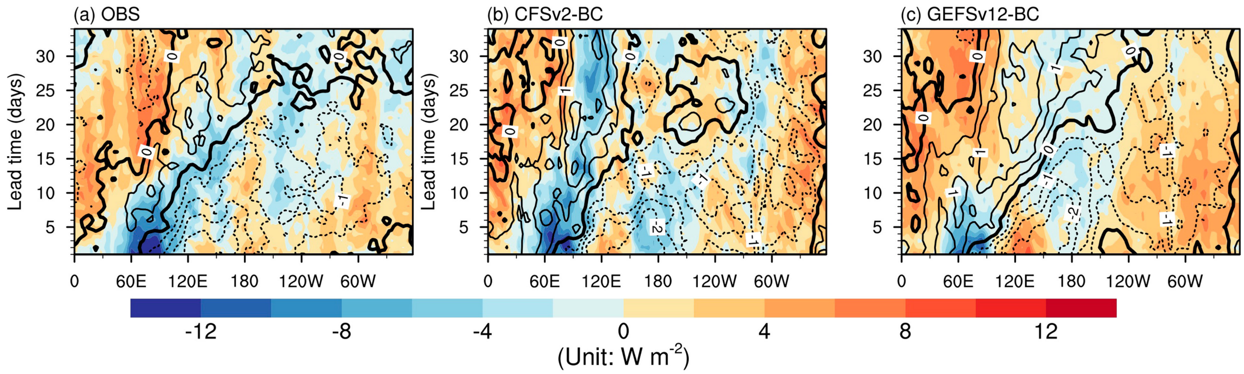

FIG. 12. Lead time–longitude composite diagrams of OLR (shading; W m^{-2}) and U850 anomalies (contours; solid for westerlies, dashed for easterlies; m s^{-1}) averaged over 15°S–15°N, initialized at MJO phase 2 with amplitude > 1 (57 cases selected from 366 events). (a) OBS, (b) CFSv2 after BC (CFSv2-BC), and (c) GEFSv12 after BC (GEFSv12-BC).

other coupled processes for sustaining MJO variability. Based on the theoretical studies of Wang and Xie (1998) and Wei et al. (2018), the air–sea coupling influences the MJO simulation in two aspects: slows down the eastward propagation speed and strengthens the convection. The comparison of Figs. 12b and 12c has indicated slower propagation and stronger deep convection in CFSv2, which thus is likely associated with the embedded air–sea coupling. Taken together, these findings highlight the critical role of surface–atmosphere coupling in enhancing the MJO simulation and also the forecast skill of subseasonal time scales.

5. Summary and discussion

Subseasonal forecasts are influenced by both initial conditions and boundary conditions, making it imperative to accurately represent atmospheric processes and their interactions with surface boundaries such as land, ocean, and ice. While quantifying the impact of the atmospheric model configurations and surface–atmosphere interactions remains a central focus of subseasonal forecasting research, there is limited comparative analysis on the relative importance of these factors in subseasonal forecast skill.

At NCEP, two operational systems—CFSv2 and GEFSv12—are used for subseasonal forecasting. These two systems primarily differ in their atmospheric model configurations and surface–atmosphere coupling processes. The two major differences between CFSv2 and GEFSv12 offer an opportunity to examine their relative importance for subseasonal forecasts. To address this, we conduct a comprehensive quantitative evaluation of the 3-yr operational spatial-mean (NH, SH, and TRO) and gridpoint predictive skill of deterministic forecasts in terms of RMSE and ACC across multiple variables for both CFSv2 and GEFSv12. Considering the substantial biases often present in subseasonal forecasts, we also examine the impact of model bias by applying bias correction to both systems. The results of this analysis offer important insights into the strengths and weaknesses of current S2S systems and provide valuable guidance for improving the accuracy and reliability of future subseasonal prediction systems.

The results indicate that, prior to bias correction, the forecast errors of CFSv2 and GEFSv12 differ more significantly for temperature than for wind and geopotential height. Particularly, in regions and for variables where the two systems exhibit statistically significant differences in forecast errors, GEFSv12 generally demonstrates smaller forecast errors, higher ACC values, and longer practical predictive limits across most variables, including T850, T200, and H500. This suggests that GEFSv12 exhibits overall superior predictive skill on subseasonal scales compared to CFSv2 (see the solid error curves in the schematic of Fig. 13). This is partially contributed by GEFSv12's overall smaller forecast bias, which may be associated with its more advanced atmospheric model configuration.

The key finding in this study is that once bias correction is applied, CFSv2 generally outperforms GEFSv12 for almost all variables across global regions in contrast to the overall superior performance of GEFSv12 before bias correction (see the dashed error curves in Fig. 13). This result highlights the critical role of surface–atmosphere coupling in subseasonal forecasts. While more advanced atmospheric models, such as GEFSv12, may provide advantages in short- to medium-range forecasts, achieving performance comparable to or even exceeding that of CFSv2 in certain regions and variables, the configuration of the atmospheric model alone may be less influential for extended-range predictions. Instead, the ability of the forecasting system to represent surface–air interaction processes appears to play a more important role in determining skill at subseasonal time scales.

The improvement in CFSv2 following bias correction is particularly evident in the tropics and regions with complex topography. In tropical areas, the predictive horizon generally extends from a few days to over 12 days for CFSv2 after bias correction, which is more significant than the extension for GEFSv12. Detailed analysis reveals that the longer predictive limits of CFSv2 in the tropics after bias correction are likely attributed to its more accurate representation of key air–sea coupling processes, such as the MJO. The enhanced accuracy in simulating these dominant variabilities in CFSv2 provides

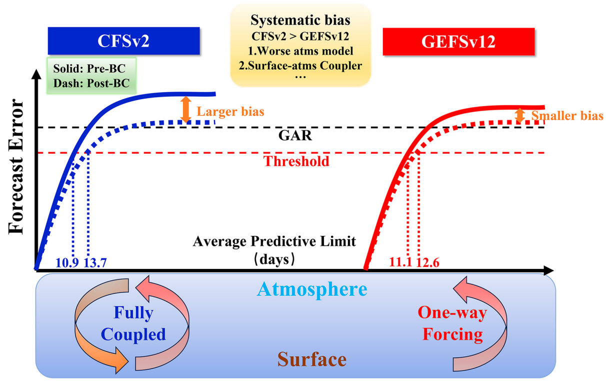

FIG. 13. The schematic diagram summarizing the main findings of this study. The dashed black and red lines indicate the GAR and the predictive limit threshold for both systems, defined as 90% of the GAR. The blue solid and dashed curves represent the RMSE pre-BC and post-BC in CFSv2, respectively, while the red solid and dashed curves correspond to the same but for GEFSv12. The post-BC forecast errors (dashed curves) closely approach the GAR for both systems compared to the pre-BC errors. The x axis denotes the predictive limits of the two systems, averaged across six variables (U200, U850, U10m, T200, T850, and H500) over all domains (SH, NH, and TRO). Orange arrows depict the model biases for GEFSv12 and CFSv2, with CFSv2 exhibiting a larger systematic bias. The yellow box outlines possible reasons for the larger systematic bias in CFSv2, which may stem from deficiencies in its atmospheric model and the surface–atmosphere coupler.

critical predictability sources for subseasonal forecasts. These findings further underscore the critical role of air–sea coupling processes, particularly in the tropics, in improving subseasonal prediction skill.

Our study primarily evaluates the predictive skill of key atmospheric variables in GEFSv12 and CFSv2, with a particular focus on the critical roles of surface–atmosphere coupling in enhancing the subseasonal forecast skill. However, several limitations should be acknowledged. First, the bias correction employed here only addresses mean-state errors and does not fully account for variance-related biases, such as those linked to seasonality or regional dependence. Additionally, this study focuses on deterministic forecast skill rather than ensemble-mean forecasts, which limits the robustness of the results. Furthermore, it does not explore the mechanisms behind extended predictability in specific regions, nor does it account for seasonal variations in forecast performance. Finally, because GEFSv12 and CFSv2 differ in multiple aspects of their system configurations, our comparison, without definite controlled sensitivity experiments, can only provide a coarse attribution, broadly distinguishing between atmospheric model configuration and coupling differences. Future research should prioritize further isolating the contributions of atmospheric model configuration and coupling processes through carefully designed sensitivity experiments, to enhance our understanding of surface–atmosphere interactions and coupled model simulations.

Acknowledgments. The authors sincerely thank Dr. Qianqian Qi from the CMA Earth System Modeling and Prediction Centre and Can Cao from Fudan University for their invaluable insights and constructive feedback during the evaluation of the MJO prediction skill. This research was supported by the National Natural Science Foundation of China (42375058 and 42288101).

Data availability statement. This study utilized ensemble forecast data from the operational Global Ensemble Forecast System (GEFS), provided by NCEP. These datasets can be accessed at https://registry.opendata.aws/noaa-gefs. The operational data from the Climate Forecast System (CFS) at NCEP are publicly available for the most recent 5 years at https://registry.opendata.aws/noaa-cfs, while previous data can be retrieved from https://www.ncei.noaa.gov/data/climate-forecast-system/access/operational-9-month-forecast/6-hourly-by-pressure. For historical reforecast data, GEFS reforecast datasets are available at https://registry.opendata.aws/noaa-gefs-reforecast. The study also relied on the NCEP–NCAR Reanalysis 2 dataset, which is accessible at https://psl.noaa.gov/data/gridded/data.ncep.reanalysis2.html, and the NOAA outgoing longwave radiation (OLR) dataset, available at https://www.ncei.noaa.gov/data/outgoing-longwave-radiation-daily/access.

APPENDIX A

Bias Correction

The algorithm of bias correction in this study follows the approach outlined by Buizza and Leutbecher (2015). For each operational forecast, the forecast bias as a function of lead time is calculated using their own historical forecast data and corresponding reference data within the 7 years (2012–19) preceding the verification period (2020–23). CFSv2 applies bias correction using its own operational data ($\mathrm{fcst}^{\mathrm{BC}}$), while GEFSv12 uses its reforecast data ($\mathrm{refcst}^{\mathrm{BC}}$) due to significant upgrades to its real-time operational system during this period, including the introduction of a new atmospheric dynamical core in 2019. Specifically, for a given forecast valid time (e.g., 1 October 2020), the bias is calculated as the average over 7 years (y = September 2012–October 2019) of the 5-week mean centered on that valid time ($\Delta w = -2, -1, 0, 1, 2$), minus the corresponding reanalysis data processed in the same way. For GEFSv12, which has five ensemble members ($j = 1, 2, \ldots, 5$), the bias for each member is averaged across the five members at the end. The bias functions on this given valid time for the two systems are expressed as follows:

$$\langle \mathrm{CFS_{bias}}(\tau) \rangle = \frac{1}{35} \sum_{t \in y} \sum_{\Delta w=-2}^{2} \{\mathrm{fcst}^{\mathrm{BC}}(t;\, \Delta w,\, \tau) - a(t;\, \Delta w + \tau)\},$$

$$\langle \mathrm{GEFS_{bias}}(\tau) \rangle = \frac{1}{175} \sum_{j=1}^{5} \sum_{t \in y} \sum_{\Delta w=-2}^{2} \{\mathrm{refcst}_j^{\mathrm{BC}}(t;\, \Delta w,\, \tau) - a(t;\, \Delta w + \tau)\},$$

where τ represents the lead time of τ days.

After calculating the bias, bias-corrected forecast data can be obtained by subtracting the computed bias from the original forecasts. The bias-corrected GEFSv12 operational data include 10 ensemble members ($i = 1, 10$), and the final formulas for bias correction of the two systems are

$$\mathrm{CFS}^{\mathrm{BC}}(\tau) = \mathrm{fcst}^{\mathrm{CFS}}(\tau) - \mathrm{CFS}_{\mathrm{bias}}(\tau),$$

$$\mathrm{GEFS}_i^{\mathrm{BC}}(\tau) = \mathrm{fcst}_i^{\mathrm{GEFS}}(\tau) - \mathrm{GEFS}_{\mathrm{bias}}(\tau).$$

APPENDIX B

MJO Evaluation

The MJO is the dominant mode of intraseasonal variability in the tropics and is considered a major source of global predictability on the subseasonal time scale. Assessing MJO forecast skill can provide critical insights into the differences in subseasonal predictive performance between CFSv2 and GEFSv12. The most widely used metric for evaluating the MJO forecast skill is the RMM index (Wheeler and Hendon 2004; Seo et al. 2009; Wang et al. 2015), which takes into account OLR, U200, and U850.

The calculation of the RMM index in this study follows the standard algorithm in numerous prior studies (Kim et al. 2018; Ren et al. 2023; Vitart and Balmaseda 2024). First, anomalies of these three variables are calculated by removing the climatology for all lead times and subtracting the mean of the most recent 120 days for each calendar date. The meridional mean of these anomalies is then computed from 15°N to 15°S. The resulting averaged anomalies are normalized using 15.1 W m^{-2} for OLR, 1.81 m s^{-1} for U850, and 4.81 m s^{-1} for U200 (Gottschalck et al. 2010). These normalized values are projected onto observed empirical orthogonal functions (EOFs) to obtain two projection coefficients (Wheeler and Hendon 2004). Finally, these two projection coefficients are divided by their respective observed standard deviations (7.38 for the first and 7.26 for the second) to produce the normalized RMM1 and RMM2 indices for every single prediction.

In this study, the long-term climatology of OLR, U200, and U850 is calculated from the NCAR interpolated outgoing longwave radiation dataset and NCEP–NCAR reanalysis for zonal wind, spanning from 1990 to 2019. The verified RMM indices are generated using the same data as that used for climatology calculations, covering the same time period as the operational forecasts. MJO skill is evaluated at a daily temporal resolution.

To assess MJO forecast skill, the RMSE and bivariate ACC between predicted and observed RMMs are used. These metrics are calculated as follows:

$$\mathrm{ACC}(\tau) = \frac{\sum_{t=1}^{N}[a_1(t+\tau)b_1(t,\tau) + a_2(t+\tau)b_2(t,\tau)]}{\sqrt{\sum_{t=1}^{N}[a_1^2(t+\tau) + a_2^2(t+\tau)]}\sqrt{\sum_{t=1}^{N}[b_1^2(t,\tau) + b_2^2(t,\tau)]}},$$

$$\mathrm{RMSE}(\tau) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}[|a_1(t+\tau) - b_1(t,\tau)|^2 + |a_2(t+\tau) - b_2(t,\tau)|^2]},$$

where $a_1(t+\tau)$ and $a_2(t+\tau)$ represent the observed RMM1 and RMM2, respectively, at time $t+\tau$ and $b_1(t,\tau)$ and $b_2(t,\tau)$ are the corresponding forecasted values at time t with a lead time of τ days. The term N, the number of predictions, equals 366. An ACC value of 0.5 is typically considered the threshold for MJO skill (Kim et al. 2018), and the predictive limit of the MJO is defined as the point when the ACC falls below 0.5.

Additionally, MJO forecast skill can be influenced by both the amplitude and phase of the oscillation. Errors in these aspects are identified using metrics from Rashid et al. (2011). The amplitude of the MJO is calculated as

$$\mathrm{AMP}(t,\tau) = \sqrt{\mathrm{RMM1}(t,\tau)^2 + \mathrm{RMM2}(t,\tau)^2}.$$

The amplitude error indicates how quickly a forecast loses the MJO signal's amplitude. Negative (positive) amplitude errors imply weaker (stronger) amplitude relative to observations. The calculation is

$$\mathrm{ERR}_{\mathrm{amp}}(\tau) = \frac{1}{N}\sum[\mathrm{AMP}_{\mathrm{prd}}(t,\tau) - \mathrm{AMP}_{\mathrm{obs}}(t+\tau)].$$

The phase error measures the angular difference between the predicted and observed phases in the RMM phase-space diagram. Negative (positive) phase errors suggest slower (faster) propagation in the forecast relative to observations. Phase error is computed as

$$\mathrm{ERR}_{\mathrm{phase}}(\tau) = \frac{1}{N}\sum\tan^{-1}\left[\frac{a_1(t+\tau)b_2(t,\tau) - a_2(t+\tau)b_1(t,\tau)}{a_1(t+\tau)b_1(t,\tau) + a_2(t+\tau)b_2(t,\tau)}\right].$$

REFERENCES

Buizza, R., and M. Leutbecher, 2015: The forecast skill horizon. *Quart. J. Roy. Meteor. Soc.*, **141**, 3366–3382, https://doi.org/10.1002/qj.2619.

Chang, E. K. M., Y. Pan, W. Wang, and C. Zheng, 2022: Developing and evaluating week 2 and weeks 3-4 outlook tools for extratropical storminess. *Front. Earth Sci.*, **10**, 963779, https://doi.org/10.3389/feart.2022.963779.

Chowdary, J. S., S.-P. Xie, J.-J. Luo, J. Hafner, S. Behera, Y. Masumoto, and T. Yamagata, 2011: Predictability of northwest Pacific climate during summer and the role of the tropical Indian Ocean. *Climate Dyn.*, **36**, 607–621, https://doi.org/10.1007/s00382-009-0686-5.

Dalcher, A., and E. Kalnay, 1987: Error growth and predictability in operational ECMWF forecasts. *Tellus*, **39A**, 474–491, https://doi.org/10.1111/j.1600-0870.1987.tb00322.x.

Domeisen, D. I. V., and Coauthors, 2022: Advances in the subseasonal prediction of extreme events: Relevant case studies across the globe. *Bull. Amer. Meteor. Soc.*, **103**, E1473–E1501, https://doi.org/10.1175/BAMS-D-20-0221.1.

Feng, J., J. Li, J. Zhang, D. Liu, and R. Ding, 2019: The relationship between deterministic and ensemble mean forecast errors revealed by global and local attractor radii. *Adv. Atmos. Sci.*, **36**, 271–278, https://doi.org/10.1007/s00376-018-8123-5.

——, J. Wang, G. Dai, F. Zhou, and W. Duan, 2023: Spatiotemporal estimation of analysis errors in the operational global data assimilation system at the China Meteorological Administration using a modified SAFE method. *Quart. J. Roy. Meteor. Soc.*, **149**, 2301–2319, https://doi.org/10.1002/qj.4507.

Goswami, B. B., M. Deshpande, P. Mukhopadhyay, S. K. Saha, S. A. Rao, R. Murthugudde, and B. N. Goswami, 2014: Simulation of monsoon intraseasonal variability in NCEP CFSv2 and its role on systematic bias. *Climate Dyn.*, **43**, 2725–2745, https://doi.org/10.1007/s00382-014-2089-5.

Gottschalck, J., and Coauthors, 2010: A framework for assessing operational Madden–Julian oscillation forecasts: A CLIVAR MJO working group project. *Bull. Amer. Meteor. Soc.*, **91**, 1247–1258, https://doi.org/10.1175/2010BAMS2816.1.

Guan, H., Y. Zhu, E. Sinsky, W. Li, X. Zhou, D. Hou, C. Melhauser, and R. Wobus, 2019: Systematic error analysis and calibration of 2-m temperature for the NCEP GEFS reforecast of the subseasonal experiment (SubX) project. *Wea. Forecasting*, **34**, 361–376, https://doi.org/10.1175/WAF-D-18-0100.1.

——, and Coauthors, 2022: GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. *Mon. Wea. Rev.*, **150**, 647–665, https://doi.org/10.1175/MWR-D-21-0245.1.

Harris, L., and Coauthors, 2020: GFDL SHiELD: A unified system for weather-to-seasonal prediction. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002223, https://doi.org/10.1029/2020MS002223.

Hazra, A., H. S. Chaudhari, and A. Dhakate, 2016: Evaluation of cloud properties in the NCEP CFSv2 model and its linkage with Indian summer monsoon. *Theor. Appl. Climatol.*, **124**, 31–41, https://doi.org/10.1007/s00704-015-1404-3.

He, Z., P. Hsu, X. Liu, T. Wu, and Y. Gao, 2019: Factors limiting the forecast skill of the boreal summer intraseasonal oscillation in a subseasonal-to-seasonal model. *Adv. Atmos. Sci.*, **36**, 104–118, https://doi.org/10.1007/s00376-018-7242-3.

Huang, H.-Y., S. B. Capps, S.-C. Huang, and A. Hall, 2015: Downscaling near-surface wind over complex terrain using a physically-based statistical modeling approach. *Climate Dyn.*, **44**, 529–542, https://doi.org/10.1007/s00382-014-2137-1.

Huang, Y., Y. Wei, and H.-L. Ren, 2026: Modulation of ENSO on the MJO forecast. *Atmos. Res.*, **330**, 108510, https://doi.org/10.1016/j.atmosres.2025.108510.

Ji, M., and F. Toepfer, 2016: Dynamical core evaluation test report for NOAA's Next Generation Global Prediction System (NGGPS). NOAA National Weather Service Tech. Doc., 93 pp., https://doi.org/10.25923/ztzy-qn82.

Jia, X., S. Yang, W. Song, and B. He, 2014: Prediction of wintertime Northern Hemisphere blocking by the NCEP Climate Forecast System. *Acta Meteor. Sin.*, **28**, 76–90, https://doi.org/10.1007/s13351-014-3085-8.

Kanamitsu, M., W. Ebisuzaki, J. Woollen, S.-K. Yang, J. J. Hnilo, M. Fiorino, and G. L. Potter, 2002: NCEP–DOE AMIP-II reanalysis (R-2). *Bull. Amer. Meteor. Soc.*, **83**, 1631–1644, https://doi.org/10.1175/BAMS-83-11-1631.

Kim, H., P. J. Webster, V. E. Toma, and D. Kim, 2014: Predictability and prediction skill of the MJO in two operational forecasting systems. *J. Climate*, **27**, 5364–5378, https://doi.org/10.1175/JCLI-D-13-00480.1.

——, F. Vitart, and D. E. Waliser, 2018: Prediction of the Madden–Julian oscillation: A review. *J. Climate*, **31**, 9425–9443, https://doi.org/10.1175/JCLI-D-18-0210.1.

Li, J., and R. Ding, 2011: Temporal–spatial distribution of atmospheric predictability limit by local dynamical analogs. *Mon. Wea. Rev.*, **139**, 3265–3283, https://doi.org/10.1175/MWR-D-10-05020.1.

——, J. Feng, and R. Ding, 2018: Attractor radius and global attractor radius and their application to the quantification of predictability limits. *Climate Dyn.*, **51**, 2359–2374, https://doi.org/10.1007/s00382-017-4017-y.

Liu, F., and Coauthors, 2024: Opportunities and barriers for skillful subseasonal prediction of East Asian summer precipitation. *Bull. Amer. Meteor. Soc.*, **105**, E2216–E2230, https://doi.org/10.1175/BAMS-D-24-0055.1.

Magnusson, L., and Coauthors, 2022: Skill of medium-range forecast models using the same initial conditions. *Bull. Amer. Meteor. Soc.*, **103**, E2050–E2068, https://doi.org/10.1175/BAMS-D-21-0234.1.

Murphy, A. H., and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.*, **117**, 572–582, https://doi.org/10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2.

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

Qi, Q., Y. Zhu, J. Chen, H. Tian, and H. Tong, 2022: Error diagnosis and assessment of sub-seasonal forecast using the GRAPES-GFS model. *Chin. J. Atmos. Sci.*, **46**, 327–345, https://doi.org/10.3878/j.issn.1006-9895.2008.20157.

Rai, A., and S. K. Saha, 2018: Evaluation of energy fluxes in the NCEP Climate Forecast System version 2.0 (CFSv2). *Climate Dyn.*, **50**, 101–114, https://doi.org/10.1007/s00382-017-3587-z.

Ramsay, B. H., 1998: The interactive multisensor snow and ice mapping system. *Hydrol. Processes*, **12**, 1537–1546, https://doi.org/10.1002/(SICI)1099-1085(199808/09)12:10/11<1537::AID-HYP679>3.0.CO;2-A.

Rashid, H. A., H. H. Hendon, M. C. Wheeler, and O. Alves, 2011: Prediction of the Madden–Julian oscillation with the POAMA dynamical prediction system. *Climate Dyn.*, **36**, 649–661, https://doi.org/10.1007/s00382-010-0754-x.

Ren, H.-L., Y. Wei, and S. Zhao, 2023: Low-frequency variability in the real-time multivariate MJO index: Real or artificial? *J. Climate*, **36**, 2073–2089, https://doi.org/10.1175/JCLI-D-22-0368.1.

Richter, J. H., and Coauthors, 2024: Quantifying sources of subseasonal prediction skill in CESM2. *npj Climate Atmos. Sci.*, **7**, 59, https://doi.org/10.1038/s41612-024-00595-4.

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

Saminathan, S., S. Mitra, S. Singh, R. Mandal, and S. Joseph, 2024: Multimodel comparison and assessment of short to medium range precipitation and temperature forecasts over India: Implications towards forecasting of meteorological indices in India. *Int. J. Climatol.*, **44**, 1758–1777, https://doi.org/10.1002/joc.8409.

Seo, K.-H., W. Wang, J. Gottschalck, Q. Zhang, J.-K. E. Schemm, W. R. Higgins, and A. Kumar, 2009: Evaluation of MJO forecast skill from several statistical and dynamical forecast models. *J. Climate*, **22**, 2372–2388, https://doi.org/10.1175/2008JCLI2421.1.

Sun, L., C. Deser, and R. A. Tomas, 2015: Mechanisms of stratospheric and tropospheric circulation response to projected Arctic sea ice loss. *J. Climate*, **28**, 7824–7845, https://doi.org/10.1175/JCLI-D-15-0169.1.

Swenson, S., and J. Wahr, 2006: Post-processing removal of correlated errors in GRACE data. *Geophys. Res. Lett.*, **33**, L08402, https://doi.org/10.1029/2005GL025285.

Tang, S., S. Qiao, T. Feng, P. Fan, J. Liu, J. Zhao, and G. Feng, 2023: Predictability of the unprecedented 2022 late summer

Yangtze River Valley and Tibetan Plateau heatwaves by the NCEP CFSv2. *Atmos. Res.*, **296**, 107053, https://doi.org/10.1016/j.atmosres.2023.107053.

Tierra, M. C. M., T.-T. Lo, H.-C. Tsai, and M. Q. Villafuerte, 2023: Evaluation of multiweek tropical cyclone forecasts in the Philippines. *Wea. Forecasting*, **38**, 2095–2110, https://doi.org/10.1175/WAF-D-22-0173.1.

Vitart, F., and M. A. Balmaseda, 2024: Sources of MJO teleconnection errors in the ECMWF extended-range forecasts. *Quart. J. Roy. Meteor. Soc.*, **150**, 2028–2044, https://doi.org/10.1002/qj.4688.

——, A. W. Robertson, and D. L. T. Anderson, 2012: Subseasonal to Seasonal Prediction project: Bridging the gap between weather and climate. *WMO Bull.*, **61**, 23–28.

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

Wahr, J., M. Molenaar, and F. Bryan, 1998: Time variability of the Earth's gravity field: Hydrological and oceanic effects and their possible detection using GRACE. *J. Geophys. Res.*, **103**, 30 205–30 229, https://doi.org/10.1029/98JB02844.

Wang, B., and X. Xie, 1998: Coupled modes of the warm pool climate system. Part I: The role of air–sea interaction in maintaining Madden–Julian oscillation. *J. Climate*, **11**, 2116–2135, https://doi.org/10.1175/1520-0442-11.8.2116.

Wang, L., and A. W. Robertson, 2019: Week 3–4 predictability over the United States assessed from two operational ensemble prediction systems. *Climate Dyn.*, **52**, 5861–5875, https://doi.org/10.1007/s00382-018-4484-9.

Wang, S., A. H. Sobel, M. K. Tippett, and F. Vitart, 2019: Prediction and predictability of tropical intraseasonal convection: Seasonal dependence and the Maritime Continent prediction barrier. *Climate Dyn.*, **52**, 6015–6031, https://doi.org/10.1007/s00382-018-4492-9.

Wang, W., M.-P. Hung, S. J. Weaver, A. Kumar, and X. Fu, 2014: MJO prediction in the NCEP Climate Forecast System version 2. *Climate Dyn.*, **42**, 2509–2520, https://doi.org/10.1007/s00382-013-1806-9.

——, A. Kumar, J. X. Fu, and M.-P. Hung, 2015: What is the role of the sea surface temperature uncertainty in the prediction of tropical convection associated with the MJO? *Mon. Wea. Rev.*, **143**, 3156–3175, https://doi.org/10.1175/MWR-D-14-00385.1.

Weber, N. J., and C. F. Mass, 2017: Evaluating CFSv2 subseasonal forecast skill with an emphasis on tropical convection. *Mon. Wea. Rev.*, **145**, 3795–3815, https://doi.org/10.1175/MWR-D-17-0109.1.

Wei, Y., F. Liu, M. Mu, and H.-L. Ren, 2018: Planetary scale selection of the Madden–Julian oscillation in an air-sea coupled dynamic moisture model. *Climate Dyn.*, **50**, 3441–3456, https://doi.org/10.1007/s00382-017-3816-5.

——, H.-L. Ren, B. Xiang, Y. Wang, J. Wu, and S. Wang, 2023: Diverse MJO genesis and predictability. *Bull. Amer. Meteor. Soc.*, **104**, E792–E809, https://doi.org/10.1175/BAMS-D-22-0101.1.

Wheeler, M. C., and H. H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, **132**, 1917–1932, https://doi.org/10.1175/1520-0493(2004)132<1917:AARMMI>2.0.CO;2.

White, C. J., and Coauthors, 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteor. Appl.*, **24**, 315–325, https://doi.org/10.1002/met.1654.

——, and Coauthors, 2022: Advances in the application and utility of subseasonal-to-seasonal predictions. *Bull. Amer. Meteor. Soc.*, **103**, E1448–E1472, https://doi.org/10.1175/BAMS-D-20-0224.1.

Xiao, H., P. Xu, and L. Wang, 2024: The unprecedented 2023 North China heatwaves and their S2S predictability. *Geophys. Res. Lett.*, **51**, e2023GL107642, https://doi.org/10.1029/2023GL107642.

Xie, J., J. Yu, H. Chen, and P.-C. Hsu, 2020: Sources of subseasonal prediction skill for heatwaves over the Yangtze River basin revealed from three S2S models. *Adv. Atmos. Sci.*, **37**, 1435–1450, https://doi.org/10.1007/s00376-020-0144-1.

——, P.-C. Hsu, Y. Hu, H. Zhang, and M. Ye, 2024: Advancing subseasonal surface air temperature and heat wave prediction skill in China by incorporating scale interaction in a deep learning model. *Geophys. Res. Lett.*, **51**, e2024GL111076, https://doi.org/10.1029/2024GL111076.

Xie, R., M. Mu, and X. Fang, 2020: New indices for better understanding ENSO by incorporating convection sensitivity to sea surface temperature. *J. Climate*, **33**, 7045–7061, https://doi.org/10.1175/JCLI-D-19-0239.1.

Ye, J., and Coauthors, 2023: Evaluation and process-oriented diagnosis of the GEFSv12 reforecasts. *J. Climate*, **36**, 4255–4274, https://doi.org/10.1175/JCLI-D-22-0772.1.

Zadra, A., and Coauthors, 2018: Systematic errors in weather and climate models: Nature, origins, and ways forward. *Bull. Amer. Meteor. Soc.*, **99**, ES67–ES70, https://doi.org/10.1175/BAMS-D-17-0287.1.

Zhang, C., J. Gottschalck, E. D. Maloney, M. W. Moncrieff, F. Vitart, D. E. Waliser, B. Wang, and M. C. Wheeler, 2013: Cracking the MJO nut. *Geophys. Res. Lett.*, **40**, 1223–1230, https://doi.org/10.1002/grl.50244.

Zhang, F., Y. Q. Sun, L. Magnusson, R. Buizza, S.-J. Lin, J.-H. Chen, and K. Emanuel, 2019: What is the predictability limit of midlatitude weather? *J. Atmos. Sci.*, **76**, 1077–1091, https://doi.org/10.1175/JAS-D-18-0269.1.

Zhou, X., and Coauthors, 2022: The development of the NCEP Global Ensemble Forecast System version 12. *Wea. Forecasting*, **37**, 1069–1084, https://doi.org/10.1175/WAF-D-21-0112.1.

Zhu, Y., X. Zhou, M. Peña, W. Li, C. Melhauser, and D. Hou, 2017: Impact of sea surface temperature forcing on weeks 3 and 4 forecast skill in the NCEP Global Ensemble Forecasting System. *Wea. Forecasting*, **32**, 2159–2174, https://doi.org/10.1175/WAF-D-17-0093.1.

——, and Coauthors, 2018: Toward the improvement of subseasonal prediction in the National Centers for Environmental Prediction Global Ensemble Forecast System. *J. Geophys. Res. Atmos.*, **123**, 6732–6745, https://doi.org/10.1029/2018JD028506.

——, and Coauthors, 2023: Quantify the coupled GEFS forecast uncertainty for the weather and subseasonal prediction. *J. Geophys. Res. Atmos.*, **128**, e2022JD037757, https://doi.org/10.1029/2022JD037757.