



Quantify the Coupled GEFS Forecast Uncertainty for the Weather and Subseasonal Prediction

Yuejian Zhu¹ , Bing Fu², Bo Yang³, Hong Guan³, Eric Sinsky², Wei Li², Jiayi Peng², Xianwu Xue³, Dingchen Hou¹, Xin-Zhong Liang^{4,5} , and Sanghoon Shin⁴

¹NOAA/NWS/NCEP/EMC, College Park, MD, USA, ²IMSG at NOAA/NWS/NCEP/EMC, College Park, MD, USA, ³SRG at NOAA/NWS/NCEP/EMC, College Park, MD, USA, ⁴ESSIC, University of Maryland, College Park, MD, USA, ⁵Department of Atmospheric and Oceanic Science, University of Maryland, College Park, MD, USA

Key Points:

- The coupled Global Ensemble Forecast System has advantages over the uncoupled system for weather and subseasonal predictions
- The forecast uncertainty is quantified by different measures and diagnostic tools
- Model resolution is less important for extended range than weather forecasts

Correspondence to:

Y. Zhu,
Yuejian.Zhu@noaa.gov

Citation:

Zhu, Y., Fu, B., Yang, B., Guan, H., Sinsky, E., Li, W., et al. (2023). Quantify the coupled GEFS forecast uncertainty for the weather and subseasonal prediction. *Journal of Geophysical Research: Atmospheres*, 128, e2022JD037757. <https://doi.org/10.1029/2022JD037757>

Received 29 AUG 2022
Accepted 23 DEC 2022

Abstract The Global Ensemble Forecast System version 12 (GEFSv12) has been implemented into National Centers For Environmental Prediction operations since September 2020, which was uncoupled, but increased the horizontal resolution from 34 to 25 km, increased ensemble members from 21 to 31, and extended forecasts from 16 to 35 days. It significantly improved probabilistic forecast skills in many categories, such as precipitation, tropical storms, Madden-Julian Oscillation (MJO), etc. The improvements resulted from many aspects, including model resolution increase, dynamical core upgrade, advances in hybrid data assimilation and physical parameterizations, and more importantly, from using new stochastic schemes to improve forecast uncertainty. To further improve GEFS's sub-seasonal forecast skill, a coupled GEFS was built up on the Unified Forecast System prototype version 5 that fully couples an atmospheric model with a land surface model, ocean model, ice model, and wave model. A set of coupled GEFS experiments were conducted to test different horizontal resolutions at approximately 50 and 25 km while adjusting the stochastic parameterization schemes for the atmosphere to better represent forecast uncertainties. The experiments were run for a 2-year period from October 2017 to September 2019, with one initialization per week at Wednesday 00 UTC, 11 ensemble members, and were forecasted out to 35 days. The coupled GEFS significantly improves 500 hPa height anomaly correlation in week-1, week-2, and MJO skills compared to the current operational forecast. The forecast spread of tropical wind is greatly reduced by improved stochastic schemes and matches well with the forecast root mean square errors. The correlation of forecast error variance and ensemble variance is improved for the coupled GEFSs. Meanwhile, the spread of MJO has been greatly reduced for the coupled GEFSs to improve the MJO forecast uncertainty.

Plain Language Summary The Unified Forecast System is a new community-based, atmosphere-land-ocean-sea ice-wave-aerosol coupled, comprehensive Earth modeling system. A Global Ensemble Forecast System (GEFS; version 12; uncoupled) has been implemented into National Centers For Environmental Prediction operations since September 2020 with 31 members, out to 35 days to cover subseasonal prediction. A new fully coupled GEFS in this investigation has demonstrated an overall better performance around the weather, medium-range, and weeks 3 and 4 time scales. Forecast uncertainties are quantified by adjusting the coefficients and parameters of the stochastic schemes.

1. Introduction

Currently, the National Centers For Environmental Prediction (NCEP) operates two global ensemble forecast systems: the Global Ensemble Forecast System version 12 (GEFSv12) and the Climate Forecast System version 2 (CFSv2), which are optimized to address separately weather and seasonal prediction. During the past two decades, several important achievements have been made on both fronts owing to the advances in data assimilation and model/ensemble configurations (Saha et al., 2010, 2014; Wei et al., 2008; Zhu et al., 2017, 2018; Zhu, Li, Sinsky, et al., 2019; Zhou et al., 2016, 2017, 2019, 2022), however, the forecast gap between the weather and climate still needs to be filled. In recent years, to meet the evolving public demand, the National Oceanic and Atmospheric Administration (NOAA) has been developing a Unified Forecast System (UFS) with time scales spanning from sub-hourly analyses to seasonal predictions.

The NCEP's effort on subseasonal forecasts was first made through a "climate-down" approach, that is, using the climate model to output higher frequency forecasts for short lead times. Specifically, CFSv2, since its operation

in 2011 for seasonal climate prediction, has also provided 45-day forecasts to cover the subseasonal scale (Saha et al., 2014). In 2017, to support the NOAA Subseasonal Experiment (SubX) project (Guan et al., 2019; W. Li et al., 2019; Pegion et al., 2019; Zhu et al., 2018), the operational GEFS forecast was extended experimentally from 16 to 35 days using an optimal configuration suitable for subseasonal forecasts. This “weather-up” approach was based on the dramatic improvement in the skill of medium-range weather forecasts by GEFS after implementing an ensemble-based data assimilation system and adopting the ensemble analyses as the initial conditions for ensemble forecasts.

The ensemble-based forecast system deals with uncertainties in both initial conditions and model configurations. For sub-seasonal forecasts, the memory of initial conditions is gradually diminished with increasing forecast lead-time while slowly varying external forcings eventually take over (Vitart et al., 2017). As such, incorporating interactions between the atmosphere and slowly varying ocean and other components in the earth system is likely a key focus to potentially reduce the forecast gap. The GEFS, originally designed for weather forecasts, is an atmosphere-land coupled forecast system. Until version 11, GEFS prescribed the SST distribution by damping an initial analysis toward the observed climatology with a 90-day e-folding for a 16-day forecast (Zhu et al., 2017). To support the SubX project, as an intermediate approach toward the atmosphere-ocean coupling, GEFS updated the SST with a “two-tiered” method. The SST forecast from the coupled atmosphere-ocean model CFSv2 was used as an input to the uncoupled atmospheric model GEFS to simulate a one-way forcing from the ocean. The use of such two-tiered SST has proven to bring important skill gain for subseasonal forecasts (Zhu et al., 2017, 2018). In particular, the Madden-Julian Oscillation (MJO) forecast skill was improved by 1.7 days compared to the damping-to-climatology method (W. Li et al., 2019; Zhu et al., 2017). This skill gain indicates the need to incorporate the atmosphere-ocean interaction in subseasonal forecasts.

On 23 September 2020, the GEFSv12 was implemented into NCEP operations with the dynamical core transition from the spectral to finite volume (FV3) representation. The GEFSv12 has also increased the horizontal resolution from T574 (~34 km) in the GEFSv11 to C384 (~25 km), the forecast leads from 16-day to 35-day, the same vertical resolution of 64 hybrid levels with a model top at 0.2 hPa (approximately 54 km) and the ensemble size from 21 to 31 members (Zhou et al., 2022). Compared to the CFSv2, the GEFSv12 includes better initial conditions, higher spatial resolution, the latest upgraded GFS model dynamical and physical configurations, and optimal perturbed initial conditions and model stochastic parameterizations (Zhou et al., 2022). The GEFSv12 is the first UFS in NCEP operations. Several upgrades have been made from the GEFSv11 to the GEFSv12 that result in significant skill improvements, especially for the MJO forecast. In particular, the stochastic physics perturbation scheme has been updated from the stochastic total tendency perturbation (Hou et al., 2008) to a combination of stochastically perturbed physics tendencies (SPPTs, Buizza et al., 1999; Christensen et al., 2017; Palmer et al., 2009; Zhu, Li, Zhou, et al., 2019), and stochastic kinetic energy backscatter (SKEB, Shutts & Palmer, 2004; Shutts, 2005; Berner et al., 2009; Shutts, 2015; Zhu, Li, Zhou, et al., 2019). All these upgrades occurred in parallel with the improvements of the Global Forecast System model (GFS; Han et al., 2017) and the Global Data Assimilation System (GDAS; Kleist & Ide, 2015).

This study aims to quantify the GEFS forecast uncertainty of the atmosphere by introducing a fully coupled system, and further improve GEFS weather and subseasonal forecasts based on UFS coupled prototype version 5 (P5; Stefanova et al., 2022). Specifically, we configure, test, and evaluate the UFS coupled GEFS forecast system. The new model is a fully coupled atmosphere-land-ocean-ice-wave-aerosol ensemble forecast system. The atmosphere and land models are based on NCEP operational GFS version 15 (GFSv15) implemented on 12 June 2019 and the GEFSv12 implemented on 23 September 2020 (Zhou et al., 2022). The ocean model is MOM6 (Adcroft et al., 2019; Griffies et al., 2020; Held et al., 2019; White et al., 2009). The sea ice model is CICE6 (<https://cice-consortium-cice.readthedocs.io/en/cice6.0.0.alpha/>; CICE6 documentation: <https://readthedocs.org/projects/cice-consortium-cice/downloads/pdf/master/>). The wave model is WAVEWATCH III (Tolman, 2008). We test two horizontal model resolutions (about 0.5 and 0.25°) together with the adjusted stochastics schemes for a 2-year period (October 2017 - September 2019).

Section 2 gives the details of the model configuration and experiment setup including the initializations from the atmosphere, ocean, and sea ice. Section 3 describes the evaluation and diagnostic methods, and verification measures; Section 4 presents the evaluation and verification results against the GEFSv12 reforecast which is

Table 1
The Configurations of CGEFS-H and CGEFS-L Experiments

Models	Names	Version	Initial conditions	Initial perturbations	Stochastic schemes	Low resolution experiments	High resolution experiments
Atmos.	GFS	V15	GFSv15 retrospective analysis	EnKF analysis	SKEB SPPT	C192/L64	C384/L64
Ocean	MOM	6	CFS analysis	No	No	0.5° L75	0.25° L75
Sea ice	CICE	6	CPC analysis	No	No	0.5°	0.25°
Wave	WW3	7+	GFSv15 wind/ice forcing	No	No	0.5°	0.5°

our benchmark toward future operational implementation. The final section concludes with a discussion of the coupled GEFS experiments, as well as future direction.

2. The Configurations and Experiment Set Up

2.1. Reference—The GEFSv12 Reforecast

The GEFSv12 reforecasts (Phase 2; 2000–2019; Guan et al., 2022) were generated by NCEP Environmental Modeling Center (EMC) before the GEFSv12 was implemented into operations (23 September 2020). The selected fields of 20-year reforecast data could be accessed through Amazon Web Service (<https://noaa-gefs-retrospective.s3.amazonaws.com/index.html>). The GEFSv12 reanalyses (Hamill et al., 2022) were generated by Physical Sciences Laboratory from the FV3 GFS/Ensemble Kalman Filter (EnKF) hybrid analyses and EnKF 6-hr forecasts with the Incremental Analysis Update (Bloom et al., 1996) replay process. The GFS model system was consistent with the GEFSv12 reanalysis, reforecast and real-time operational forecast except for the Near-Surface Sea Temperature (NSST) which was replaced by Optimum Interpolation Sea Surface Temperature (Reynold & Smith, 1994) for the reanalysis (Hamill et al., 2022). The reforecasts have been configured to have the same horizontal resolution as operations, but it only ran once per day at 00UTC with five members (one unperturbed and four perturbed forecasts), out to 16 lead days, except for every Wednesday with 11 members (one unperturbed and 10 perturbed forecasts), out to 35 lead days.

The GEFSv12 is the first UFS implementation with a reforecast to support weather-to-subseasonal probabilistic prediction and hydrology applications. The GEFSv12 reforecast was using the same model configurations including dynamics and physical parameterizations as the operational GEFSv12 (Zhou et al., 2022). The stochastic schemes are using SKEB with an amplitude of 0.6 and SPPT with 5-scales (total sigma = 0.95) of each coefficient as (0.8; 0.4; 0.2; 0.08; 0.04) (Zhou et al., 2022).

2.2. Experiments

There are two coupled experiments with similar configurations, but the horizontal resolutions for the atmosphere, ocean, and sea ice models are different (see Table 1). The low resolution coupled GEFS (CGEFS-L hereafter) is about 0.5° (approximately 50 km) and the high resolution coupled GEFS (CGEFS-H hereafter) is about 0.25° (approximately 25 km) for all model components.

The SPPT scheme and configurations in the coupled models are similar to the GEFSv12 reforecast, but the amplitudes (or coefficients) are reduced by 30% compared to the operational GEFSv12 (Zhou et al., 2022; Zhu, Li, Zhou, et al., 2019) in order to reduce tropical overdispersion. The adjusted coefficients are 0.56; 0.28; 0.14; 0.056; 0.028 for five scales, respectively. The SKEB scheme and set up are the same as the GEFSv12 reforecast, except that the amplitude is increased from 0.6 to 0.7 in order to make up for the slight perturbation loss from the 30% reduction in the 5-scale SPPT scheme (Zhou et al., 2022; Zhu, Li, Zhou, et al., 2019).

The experiment period spanned from 4 October 2017 to 25 September 2019 (2 years) initialized every Wednesday at 00 UTC (once per week; 104 total runs), and 11 (1 unperturbed and 10 perturbed) members with a uniform resolution out to 35 lead days.

3. The Methods of Evaluation and Diagnosis

3.1. Verification Metrics for Weather (Day-To-Day Verification)

The selected atmospheric fields (1,000 and 500 hPa heights; 10m, 850 and 250 hPa winds; 2m and 850 hPa temperature) are verified against the GEFSv12 reanalysis (Hamill et al., 2022) from lead day 1 to day 16. The probabilistic forecast verification is based on 10 climatologically equally-likely bins (Zhu et al., 1996). The measures include pattern anomaly correlation (AC) and root-mean-square (RMS) error of ensemble mean, ensemble spread, mean error and absolute error of ensemble mean, continuously ranked probability score (CRPS), ranked probability score (RPS), Brier score and its decompositions (reliability and resolution), Relative operating characteristics (ROC) score and Economic values (Toth et al., 2006; WMO et al., 2021; Zhu et al., 1996, 2002).

The scores are assessed over three main domains and four subdomains. The main domains are Northern hemisphere (NH; 20°-80°N), Southern hemisphere (SH; 20°-80°S), and TR (tropical area; 20°N-20°S). The subdomains are Pacific - North American (PNA), North America (NA), Europe (EU), and East Asia (EA). In this article, the selected most interesting statistics will be presented.

3.2. Weekly, Bi-Weekly, and Monthly Mean Scores

The weekly (week-1 and week-2), bi-weekly (weeks 3 & 4), and monthly mean statistics of 500 hPa height, are generated from the average of the forecast period (valid at each 00 UTC), and the corresponding period of analysis and climatology. Then the RMS error, absolute error, mean error (or bias), and anomaly correlation (AC) coefficient of the ensemble mean are calculated.

The proxy truth (or analysis) used for this verification is the GEFSv12 reanalysis. The climatology is from the 40-year NCEP/NCAR reanalysis (Kalnay et al., 1996).

3.3. MJO Evaluation

The MJO events are evaluated using the traditional Wheeler–Hendon index (Gottschalck et al., 2010; Lin et al., 2008; Wheeler & Hendon, 2004). The MJO forecast skill is defined as the bi-variate anomaly correlation between the analysis and forecasts real-time multivariate MJO index 1 and 2 (RMM1 and RMM2) over the period of the evaluation, which is calculated at each lead time. RMM1 and RMM2 are generated by first calculating the anomalies of the analysis and the forecast ensemble mean for the outgoing longwave radiation (OLR), 850 hPa zonal-wind component (U850), and the 200 hPa zonal-wind component (U200). Then a meridional mean is performed from 15°S to 15°N. These averaged anomalies are then normalized using 15.1 W/m² for OLR, 1.81 m/s for U850, and 4.81 m/s for U200 (Gottschalck et al., 2010). Then the anomalies for each field are projected onto the Wheeler-Hendon Empirical Orthogonal Functions (EOFs) to produce two projection coefficients (Wheeler & Hendon, 2004). Finally, these projection coefficients are normalized by using their respective observed standard deviations provided by Wheeler & Hendon (2004), which generates the RMM1 and RMM2 (Gottschalck et al., 2010). The ensemble mean OLR, U850, and U200 are verified against the same variables from the GDAS analysis. The long-term climatology used to generate the forecast and analysis anomalies is calculated from the NCEP/NCAR reanalysis for U200 and U850 and from the NCAR Interpolated Outgoing Longwave Radiation dataset (Liebmann & Smith, 1996) for the OLR, both for the period 1981–2010. The long-term mean and average of the previous 120 days are removed from the climatology to eliminate long-term trends and seasonal variability.

3.4. Diagnosis of the Ratio of Error and Spread

Diagnosis of the error-spread relation is performed with the statistics package developed by Kolczynski et al. (2011). Ensemble spread and RMSE of the ensemble mean against the corresponding verification data (the GEFSv12 reanalysis) are utilized to generate horizontal and vertical cross section maps of spread-error ratios.

To further explore the relationship between *distributions* of error variance and *distributions* of ensemble variance in the 2-year experiments, error-ensemble variance pairs for all 1° × 1° grid points at each forecast time are ordered by the ensemble variance. The ordered pairs are then grouped into equally populated bins. The average of the ensemble variances within each bin is used as the representative ensemble variance. Meanwhile, this kind of grouping also provides a practical approach to estimating the representative error variance within each bin by

assuming that samples with similar ensemble variance correspond to similar error variance distributions. Scatter plots generated by this procedure will be used to depict the error-spread variance relationship.

3.5. Tropical Cyclones

The tropical cyclone (TC) track error (mean absolute error) has been calculated to compare the ensemble mean position which is the weighted center from each ensemble member and the best (or observed) tracks from the National Hurricane Center (NHC)/Joint Typhoon Warning Center. A similar method is used to calculate the TC intensity error but through a maximum 10m wind speed. The ensemble spread of TC position is a “standard deviation” of all ensemble member positions around their mean position.

3.6. SST and Surface Fluxes

The RMS error of ensemble mean and spread for SST and surface fluxes for a particular forecast lead-time (day) is calculated using the same method as 3.1 (day-to-day verification). The proxy truths are The Operational Sea Surface Temperature and Ice Analysis (documentation: <https://ghrsst-pp.metoffice.gov.uk/ostia-website/index.html>, data accessed through: <http://podaac-ftp.jpl.nasa.gov/allData/ghrsst/data/L4/GLOB/UKMO/OSTIA/2019/001/20190101-UKMO-L4HRfnd-GLOB-v01-fv02-OSTIA.nc.bz2>) for SST, the European Center for Medium-Range Weather Forecasts Reanalysis Version 5 (ERA5; Hersbach et al., 2020) for the latent and sensible heat fluxes, and the Clouds and the Earth's Radiant Energy System product for radiation fluxes from NASA Langley Research Center (Rutan et al., 2015).

4. The Evaluation and Verification Results

4.1. The Scores of 500 hPa Geopotential Height Ensemble Mean and Probabilistic Forecast

The pattern anomaly correlation coefficient (hereafter AC score) of 500 hPa geopotential height ensemble mean for Northern Hemisphere (NH) (Figure 1a) and Southern Hemisphere (SH) (Figure 1b) demonstrates the large scale ensemble performance. Overall, both CGEFS-H and CGEFS-L are better than GEFSv12 reforecast for all lead-times. The AC score of NH for the coupled GEFS ensemble mean provides a useful skillful forecast of nearly 10 days (60% AC score threshold), which extends 6–12 hr beyond the uncoupled forecast (the GEFSv12 reforecast). For SH, the CGEFS-H has a most useful skillful forecast (60% AC score threshold). A slight degradation of CGEFS-H from CGEFS-L of NH may indicate that the high resolution model does not take advantage of ensemble mean extended-range forecasts due to imperfect numerical schemes and physical parameterizations. Small scale waves could introduce more noise than signal, which increases large-scale error from the interactions of the different scales.

The probabilistic forecast is generated from a set of perturbed and unperturbed initialized forecasts. There are many methods to evaluate probabilistic forecasts. The Continuous Ranked Probability Score (CRPS) is one of them to measure both the forecast reliability and resolution. In this investigation, we have generated only 10 perturbed and one unperturbed forecast due to limited computational resources. The comparison is based on the same ensemble size and the same reference (and climatology) of all experiments (GEFSv12 reforecast, CGEFS-H and CGEFS-L) which could represent the differences in the probabilistic forecast performances. CRPS is similar to the RMS error but it counts the error (or distance) between the observation and probabilistic distribution through continuous ranking (self ranking). The score could be converted to the skill, namely CRPSS when climatological CRPS is calculated and referred.

Figure 1 also shows the 500 hPa geopotential height CRPSS for both hemisphere extratropical domains (Figure 1c is for NH; Figure 1d is for SH). Similar to Figures 1a and 1b (500 hPa geopotential height AC score), the coupled GEFS (CGEFS-L and CGEFS-H) performs better than the GEFSv12 reforecast in both hemispheres and at all lead times. Second, the high resolution coupled GEFS (CGEFS-H) consistently performs better than the low resolution coupled GEFS (CGEFS-L) for the first 2 weeks which is slightly different from NH AC scores. The improvement of coupled GEFS could be due to multiple factors, but importantly, the ensemble spread has been improved by adjusting the stochastic parameterization schemes (reduced SPPT and increased SKEB). It will be discussed in Sections 4.3 and 4.4.

The ROC scores are based on the ROC curves relative to the threats from predefined 10 climatologically equally-likely bins (Zhu et al., 1996). The ROC curve is the probability of detection (or hit rate) against the prob-

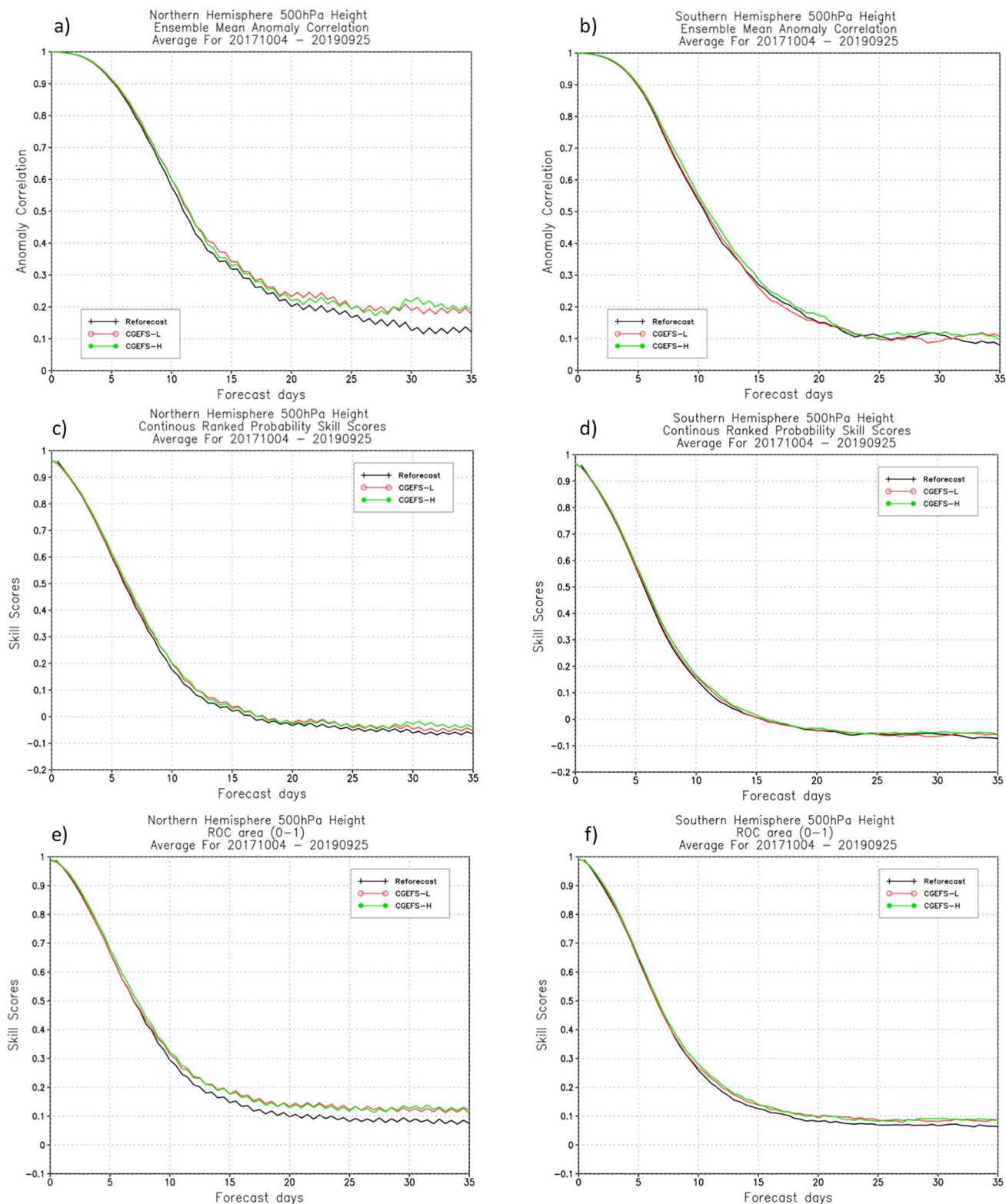


Figure 1. Two-year average scores of 500 hPa geopotential height as a function of forecast lead time (days) for the Global Ensemble Forecast System v12 reforecast (black), CGEFS-L (red), and CGEFS-H (green). (a) Anomaly correlation (AC) coefficient of ensemble mean for the Northern hemisphere (NH) (20°N–80°N) domain; (b) AC score for the Southern hemisphere (SH) (20°S–80°S) domain; (c) Continuous ranked probability skill score (CRPSS) for NH; (d) CRPSS for SH; (e) Relative operating characteristics (ROC) score for NH; (f) ROC score for SH.

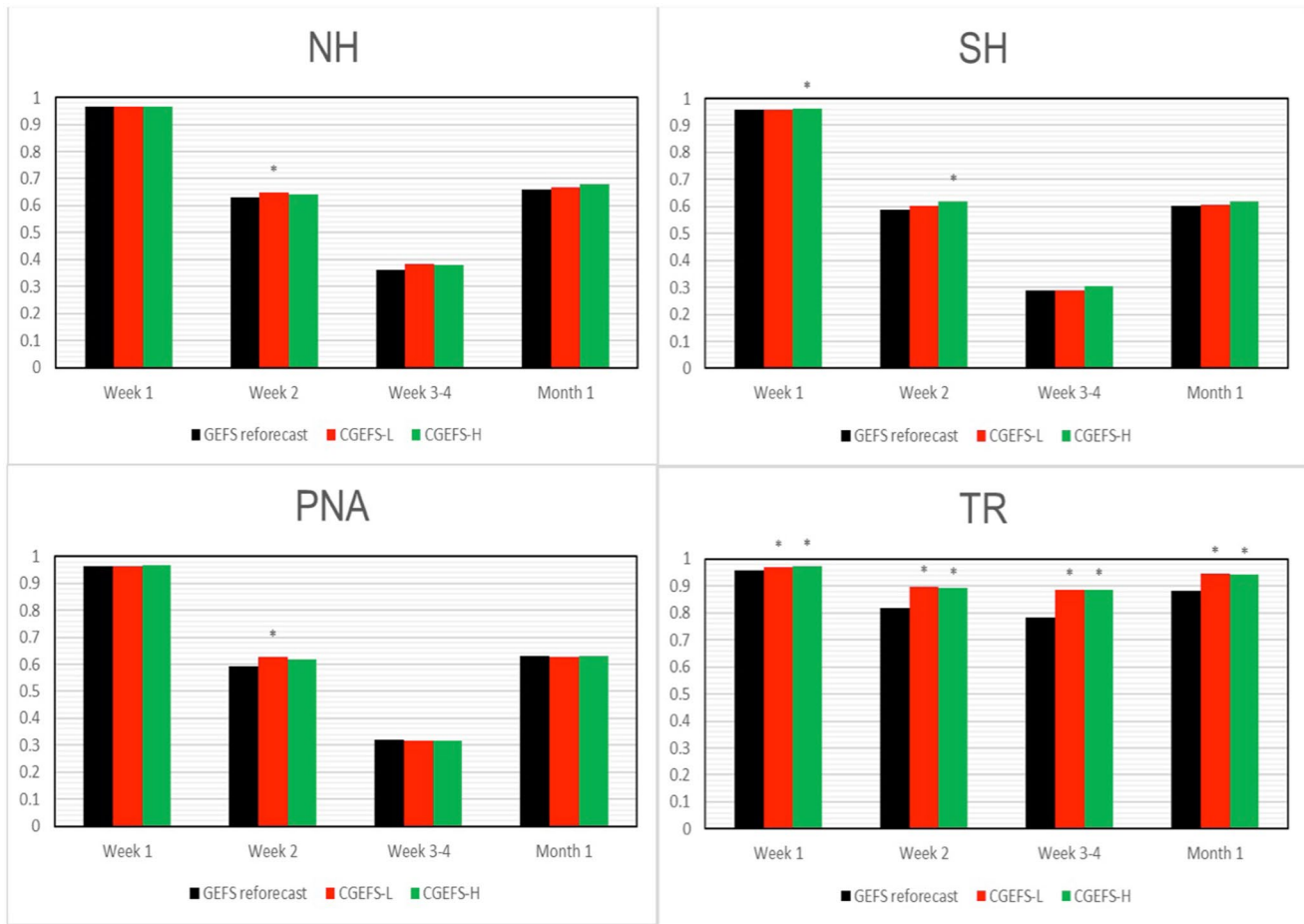


Figure 2. Two-year averaged 500 hPa geopotential height anomaly correlation (AC) scores for the GEFSv12 reforecast (black), CGEFS-L (red) and CGEFS-H (green). The asterisks (*) signify that the difference between CGEFS and the reforecast average AC score is statistically significant at 95%. The AC scores are for week-1, week-2, weeks 3&4, monthly, NH (top left), SH (top right), TR (bottom right), and Pacific-North American (bottom left) domains.

ability of false detection (or false alarm rate) for a given event (10 equally). The area under the curve ("AUC") is defined as the area below the ROC curve and above the diagonal (corresponding to climatological forecasts), with a perfect score of 1. Therefore, the ROC scores are measuring the forecast resolution which is one of the two forecast attributes (reliability and resolution; Toth et al., 2006; WMO et al., 2021).

For the domain averages of NH and SH, the ROC scores (Figures 1e and 1f) of CGEFS-H and CGEFS-L are better than GEFSv12 reforecast for all lead-time. For short lead-time (day 1–7), CGEFS-H takes advantage of the skills from higher model resolution (25 km) than lower model resolution (CGEFS-L; 50 km). It is similar to the 500 hPa geopotential height of the ensemble mean, CGEFS-H does not take much advantage over CGEFS-L for extended time-range which indicates the model resolution is less important for subseasonal forecast.

To summarize the performance of the global forecast, Figure 2 shows the average AC scores of week-1 (days 1–7); week-2 (days 8–14), weeks 3&4 (days 15–28), and month (days 1–30) for NH, SH, TR, and PNA domains. The asterisk (*) signs on the top of bar plots indicate a statistically significant difference at 95% exists between the GEFSv12 reforecast and CGEFS. Overall, coupled ensembles (CGEFS-L and CGEFS-H) are better than the un-coupled ensemble (the GEFSv12 reforecast) for most weekly-averaged lead-times and all the domains. In the tropical domain, the coupled ensembles are significantly better than the un-coupled ensemble for all the weekly-averaged lead-times. These improvements mainly occur for two reasons. First, the AC score improvements are mainly due to the coupling of the atmosphere to the ocean and sea ice, which improves the lower boundary conditions essential to the prediction of the atmospheric circulation, for example, Rossby waves and tropical convection. Second, the enhanced model performance benefits from the adjusted coefficients of the

stochastic schemes (SPPT and SKEB), which improve the representation of forecast uncertainty and ensemble mean solution.

4.2. Tracks of the Tropical Cyclones

The number of tropical cyclones cases is insufficient to draw any solid conclusions (and/or significance tests) since we have run the ensembles once per week only, but we still could review the performance of the three experiments in terms of TC tracks of the ensemble mean error, ensemble spread and intensity error for three major TC basins.

Based on the combinations of three basins (North-West Pacific, East Pacific, and Atlantic ocean), TC track error (Figure 3a) of CGEFS-H is better (smaller error) than the GEFSv12 reforecast out to 96 hr. After 96 hr, the track error is slightly worse, but the sample size is limited as well during these lead times. CGEFS-H has a very similar performance to CGEFS-L for short lead times. For longer lead times, CGEFS-H track error is slightly lower than CGEFS-L, but the sample size is limited as well. However, the GEFSv12 reforecast and two coupled experiments show the ensemble spread close to the track error out to 120 hr.

The TC intensity is highly dependent on the model resolution (Figure 3b). There is a substantial degradation in CGEFS-L compared to the GEFSv12 reforecast and CGEFS-H due to a coarse model resolution (approximately 50 km compared to 25 km resolution). CGEFS-H is much better than the GEFSv12 reforecast out to 48 hr which should be beneficial for a fully coupled system. After 48 hr, there is minor degradation in CGEFS-H compared to the GEFSv12 reforecast, but the sample size is very limited for the longer lead times.

4.3. RMS Error and Ensemble Spread of Zonal Winds

A common method to measure an ensemble system is to evaluate the RMS error of ensemble mean and ensemble spread statistically on a specified domain and averaged over individual cases for a specified lead-time. A good ensemble system should have a forecast spread that is close to its RMS error of ensemble mean statistically, or a full distribution of ensemble forecast should cover all possible outcomes, with no or fewer outliers. The performance of the high and low-level troposphere zonal wind is an important measure not only to account for the accuracy of atmospheric momentum, but also to indicate whether the model can predict large scale circulation (interaction of tropical and subtropical area), the position and intensity of the high-level jet and low-level jet streams, and other large scale atmospheric phenomena.

Generally, both coupled ensemble systems (CGEFS-L and CGEFS-H) show similar or better RMS errors for both hemispheres of 850 hPa and 250 hPa zonal winds compared to the uncoupled ensemble system (Figures 4a–4d and 4e). In particular, after the stochastic schemes are tuned from the GEFSv12 reforecast (reduced SPPT's contribution and increased SKEB's contribution), the tropical spreads are substantially reduced to match the RMS error of ensemble mean in terms of domain and time average (Figures 4c and 4f). Meanwhile, a tropical spread difference of CGEFS-L and CGEFS-H could indicate that the model perturbations are highly dependent on the model resolutions through the SPPT (tendency perturbation) scheme (Figures 4c and 4f). An improvement of the tropical zonal winds indicates a better prediction of tropical circulations, tropical convection, and a beneficial effect on MJO's prediction (see discussion of MJO).

4.4. Diagnosis of the Ensemble Forecast Uncertainties

To continue the discussion of the relationship between ensemble spread and RMS error to quantify the forecast uncertainty, a set of ensemble diagnostic methods have been utilized to investigate the quality of ensemble uncertainties (or spreads) and to assess the impacts of physical parameterization or other model changes. First, Figures 5 and 6 present maps of the ratio of ensemble spread to RMS error to demonstrate the spatial distribution of these relationships statistically. We use 144 hr (6-day) forecasts as an example because this lead-time is around the peak of error growth (highest error growth rate) and within the most challenging period of model physical parametrizations (Zhu, Li, Zhou, et al., 2019). For the boreal summer half-year (April - September; 52 total cases; Figure 5), the ratios range from -20% to $+20\%$ (0 means perfect) for most areas, except for the tropical region 850 hPa zonal wind of the GEFSv12 reforecast which shows overdispersion (ratio greater than $+20\%$) in part of the domain. As a consequence of the reduction of SPPT's contribution, the tropical area has

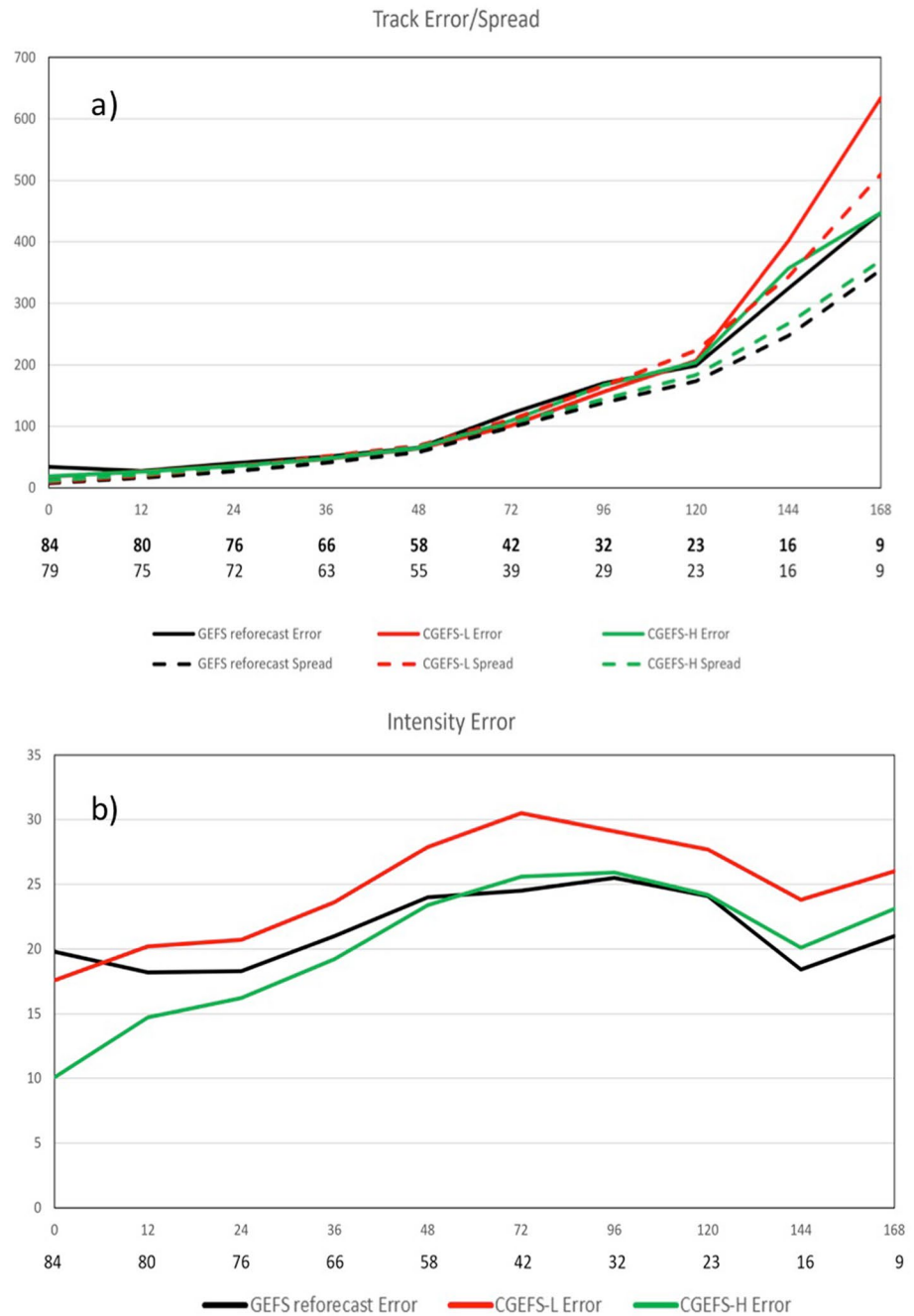


Figure 3. (a). Tropical cyclone tracks (solid line) and spread (dash line) for all three basins of three experiments (the GEFSv12 reforecast (Black), CGEFS-L (red), and CGEFS-H (green)). The first, second, and third rows of *x*-axis labels are respectively forecast lead hours, numbers of cases used to calculate track error (bolded; 2017–2019), and number of cases used to calculate spread (2018–2019). The *y*-axis is track error (unit: Nautical Mile (NM)). (b), Tropical cyclone intensity error comparison of GEFS reforecast (black), CGEFS-L (red) and CGEFS-H (green) for the average of all domains. The first and second rows of *x*-axis are respectively forecast lead hours and numbers of cases used to calculate the intensity error. The *y*-axis is intensity error (unit: knot = Nautical Mile per Hour).

an expected spread-error ratio for both coupled ensemble systems (CGEFS-L and CGEFS-H). For the boreal winter half-year (October - March; 52 total cases; Figure 6), the spatial distribution of the ratios is very similar to summer, and overdispersion of tropical 850 hPa zonal wind in the GEFSv12 reforecast is also prominent. The 850 hPa tropical spread is reduced after adjusting the SPPT scheme and shows a nearly perfect spread-error ratio

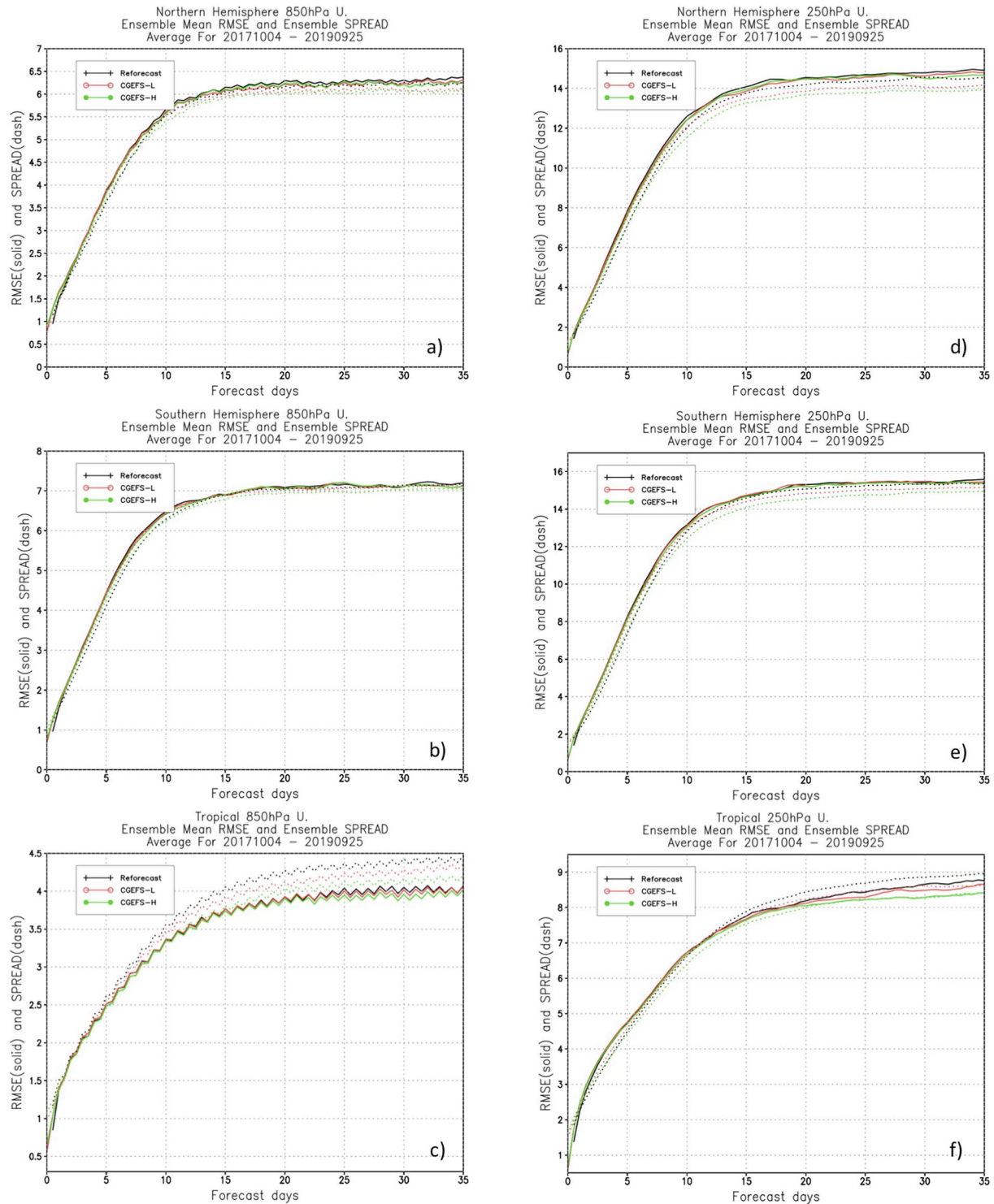


Figure 4. The root mean square errors of the ensemble mean (solid lines) and ensemble spread (dashed lines) of 850 hPa zonal wind for NH (a), SH (b), and TR (c) and of 250 hPa zonal wind for NH (d), SH (e) and TR (f). Comparison of reforecast - black; CGEFS-L - red; CGEFS-H - green.

for the coupled GEFS. However, winter 850 hPa tropical overdispersion in the reforecast mainly appears around the southern hemisphere tropical land areas (e.g., Amazon basin, North Africa, and South-East Asia continent). The combined impacts of the annual march of the seasons, Walker circulation, and land-sea contrast shift the precipitation centers to these land areas during the winter half-year. The overdispersion centers coincide with the

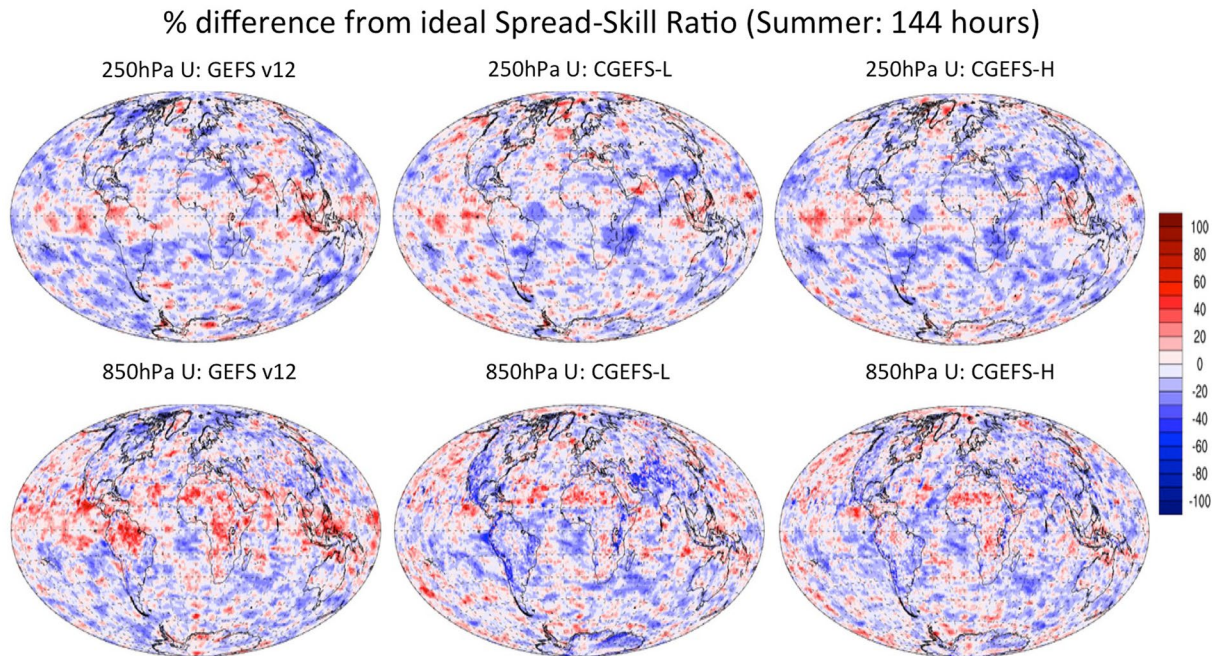


Figure 5. The boreal summer 6 months (April - September) spread-error ratio map of 850 hPa zonal winds (upper row) and 250 hPa zonal winds (lower row) at 144 hr (6 days) forecast for the GEFSv12 reforecast (left column), CGEFS-L (middle column) and CGEFS-H (right column).

large precipitation centers, suggesting that the large tendency from the convection parameterization scheme may contribute to the overdispersion centers at these locations.

A further investigation of the ratio through a spatial distribution is generated from latitude-vertical (200–1,000 hPa) cross section maps of the zonal mean for the same forecast lead-time (144-hr) (Figure 7). There is overdispersion over the tropics for all vertical levels in the GEFSv12 reforecast. After adjusting the SPPT and SKEB schemes, the spread-error ratio for the zonal mean has remarkably improved for the two coupled GEFS experiments.

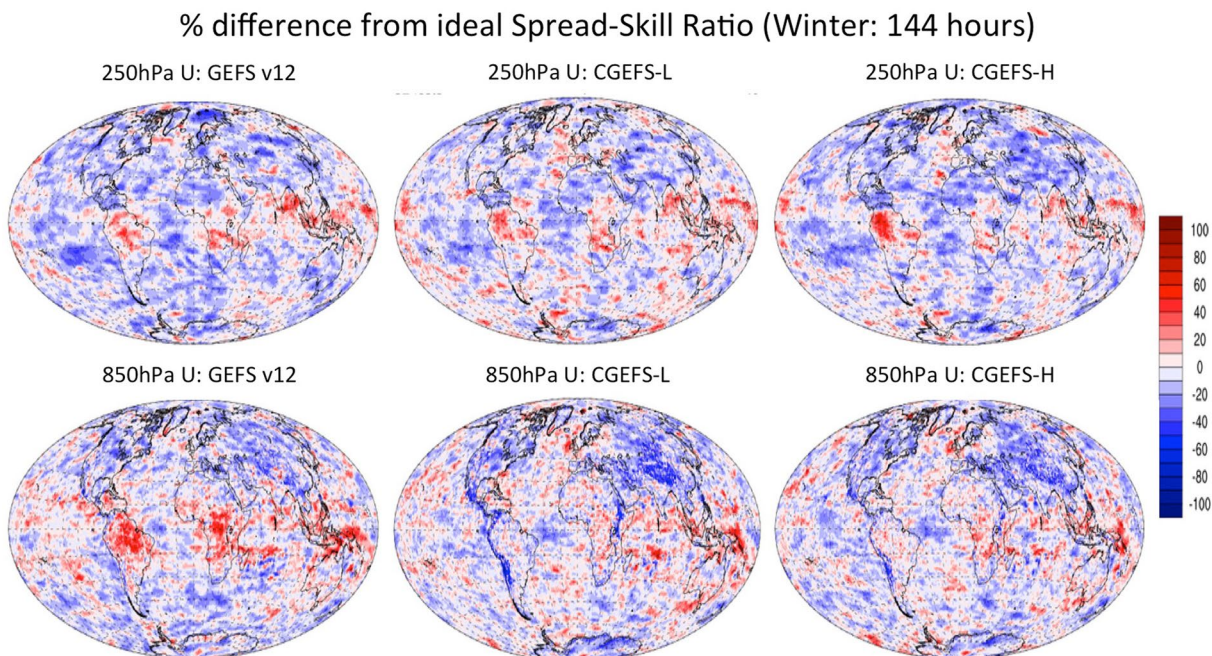


Figure 6. The same as Figure 7, but for boreal winter 6 months (October - March).

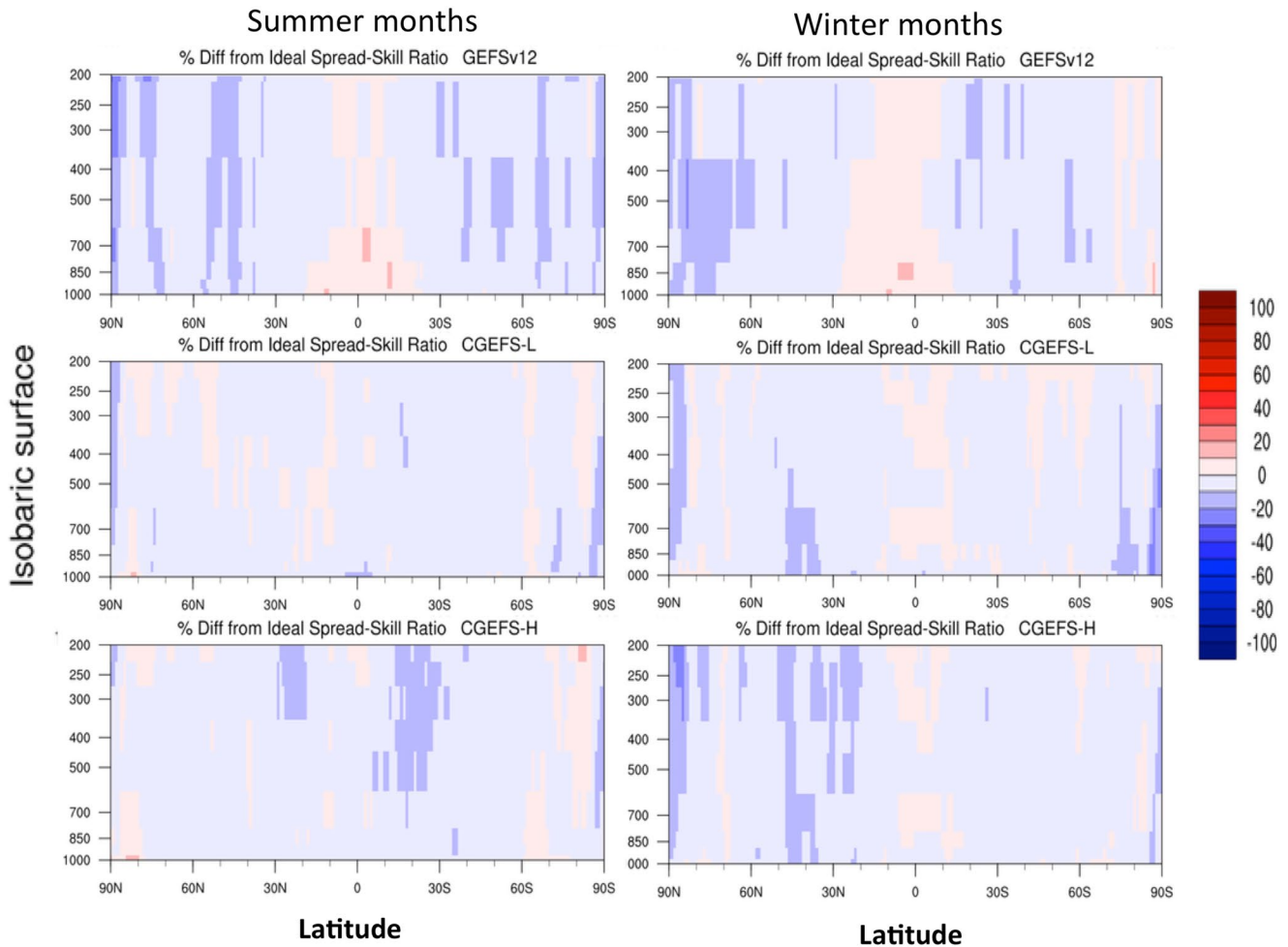


Figure 7. The vertical cross section of the ratio for boreal summer 6 months (left column) and boreal winter 6 months (right column) of zonal wind from surface (1,000 hPa) to 200 hPa vertically, for 144 hr (6 days) forecasts, and for the GEFSv12 reforecast (top), CGEFS-L (middle) and CGEFS-H (bottom).

Figures 5–7 show the statistical averages of the error-spread ratio. It can measure the performance of “reliability” (statistical average), but not “resolution” or “sharpness” (Toth et al., 2006; WMO et al., 2021). In order to analyze the performance of the “resolution”, a Linear Variance Method model (LVM; Kolczynski et al., 2011) has been introduced. Scatter plots are used to pair the ensemble variance (sorted and divided by bin) and the corresponding ensemble mean error variance, which can count the case-dependent error-spread relationships, specifically for sudden weather regime changes. A line that is close to the diagonal indicates that the stochastic scheme is capable of responding to the system changes covering a wide ensemble variance regime properly (Figure 8).

Slopes (1 is perfect) and intercepts (0 is perfect) from linear regression are used to quantitatively evaluate the performance of the ensemble systems (blue solid line in Figure 8). There are only limited ensemble members (11 members) in our experiments, which is not sufficient to represent full (true) ensemble uncertainty. In the real-time operational forecasts, 31 ensemble members are generated for each initial time. Measures of uncertainty from limited member ensemble forecasts can be calibrated to get a more accurate representation of the actual uncertainty. Slope and intercept are adjusted for idealized infinite ensemble size based on the LVM regression dilution theory (dashed line in Figure 8, also see appendix from Kolczynski et al. (2011) for a detailed description of the adjustment algorithm). A nearly perfect adjusted slope and intercept indicate that the original correlations are good for such an ensemble size. Based on Figure 8, 250 hPa tropical zonal wind at 144-hr (6-day) as an example, the original slope (and adjusted slope) and original intercept (and adjusted intercept) indicate the CGEFS-L is better than the GEFSv12 reforecast, and CGEFS-H is better than CGEFS-L for both boreal winter and summer. Overall, CGEFS-H is the best, and our new parameters setting works very well.

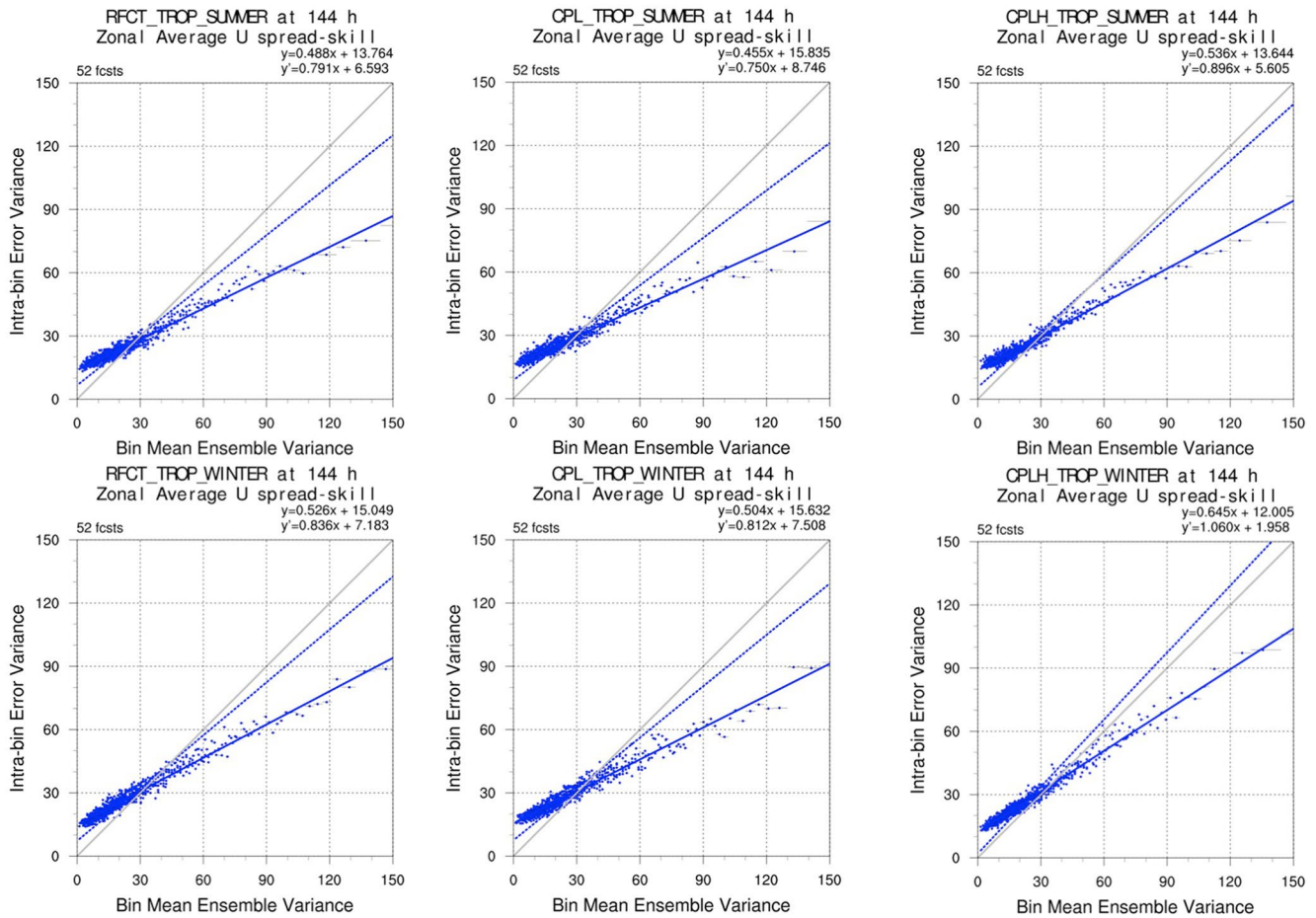


Figure 8. The scatter plot of ensemble bin variance and mean error variance for boreal summer (6 months; upper row) and boreal winter (6 months; lower row) of tropical zonal winds on the 250 hPa and forecast lead-time at 144 hr (7 days). The GEFSv12 reforecast is on the left column, CGEFS-L is on the middle column and CGEFS-H is on the right column.

4.5. The MJO Indexes

The MJO performance of the three different configurations is demonstrated in Figures 9 and 10. Since MJO is a dominant mode of sub-seasonal predictability, MJO index (RMM1+RMM2; RMM1; RMM2) and its associated components (U850; U200 and OLR) are of major importance when evaluating the capability of the forecast system on a sub-seasonal timescale. Compared to the GEFSv12 reforecast, which has already built a high standard (Guan et al., 2022; Zhou et al., 2022) for MJO skills, the coupled GEFS experiments (both CGEFS-L and CGEFS-H) show better skills for all lead-times (Figure 9) except for a fluctuation around lead days 17–23 of RMM1. Meanwhile, the RMM index, the bi-variate anomaly correlation between the analysis and forecast RMM1 and RMM2 over the period of the evaluation, increases from ~23 to ~24 days which is based on the 50% anomaly correlation threshold. All these improvements are likely due to the atmosphere-ocean coupling and the optimum stochastic physical perturbations. When comparing low resolution coupling (CGEFS-L) to high-resolution coupling (CGEFS-H), the result indicates the low resolution is better than the high resolution for RMM2 (Figure 9) for all lead times. The high correlation of RMM2 means good predictability for the high amplitude of the maritime continent (Phase 4 & 5). Based on the studies (Li et al., 2020; Ling et al., 2019; Madden & Julian, 1972), the maritime continent region is key to the success of predicting MJO. Many numerical models are facing challenges regarding tropical convection over the maritime continent region including CGEFS-H. To improve the forecast performance for high resolution coupled models on subseasonal timescales over the maritime continent, future work should be done on high resolution convection schemes.

For ensemble probabilistic evaluation, the RMS error and ensemble spread of MJO (multi-variance) have been calculated to compare the uncertainties of tropical multivariable prediction (Figure 10). There are two important

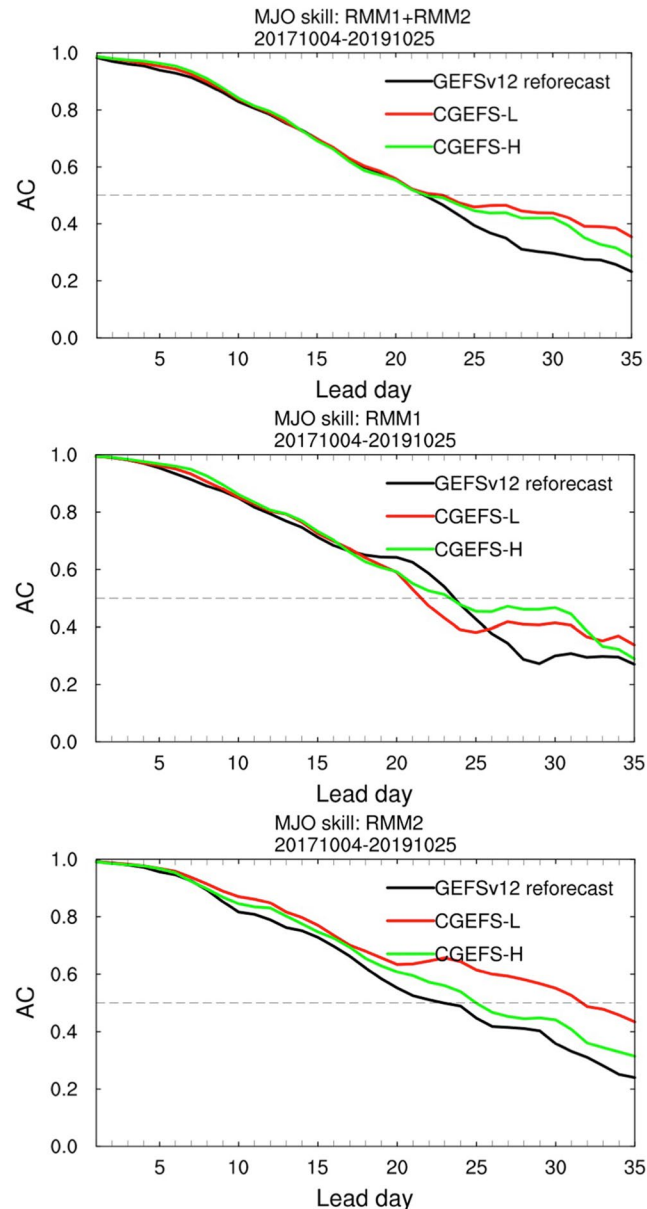


Figure 9. Madden-Julian Oscillation (MJO) skills for the GEFSv12 reforecast (black), CGEFS-L (red) and CGEFS-H (blue). The combined MJO index (RMM1+RMM2) is on the top, the RMM1 index is in the middle, and the RMM2 index is on the bottom. The MJO forecast skill is defined as the bi-variate anomaly correlation between the analysis and forecast real-time multivariate MJO index 1 and 2 (RMM1 and RMM2) over the period of the evaluation, which is calculated at each lead time.

characteristics of this evaluation. First, CGEFS-H has a smaller RMS error than the CGEFS-L and GEFSv12 reforecast for all lead times. Second, both coupling systems (CGEFS-L and CGEFS-H) reduce the ensemble spread substantially which matches much more closely to the RMS error. However, there is still overdispersion in the ensemble spread which is similar to what has been seen in the tropical 850 hPa zonal wind (Figure 4c).

4.6. SST and Surface Fluxes

Key differences have been found between the uncoupled GEFS (GEFSv12 reforecast) and the two coupled GEFS experiments (CGEFS-H and CGEFS-L). In order to diagnose the differences and pros/cons of the atmosphere-ocean coupling, the performance of SST and surface fluxes are evaluated for the global ocean (50°S-50°N) and Nino 3.4 domains (Figures 11 and 12). For the global ocean, both coupling systems (CGEFS-L and CGEFS-H) have

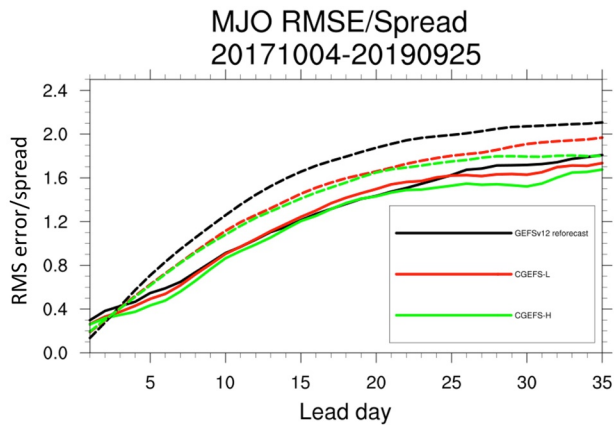


Figure 10. The root mean square errors of ensemble mean (solid lines) and spread (dashed lines) of the combined Madden-Julian Oscillation (MJO) index for the GEFSv12 reforecast (black), CGEFS-L (red) and CGEFS-H (blue). The unit of RMS error is based on the RMM index of MJO events (Unit: non-dimension).

near-zero or positive biases, whereas the GEFSv12 reforecast has a negative bias. The SST bias for the Nino 3.4 domain is similar to that of the global ocean domain for the GEFSv12 reforecast and two experiments. Overall, the RMS errors from the global and Nino 3.4 ocean domains for the coupled systems are smaller than the uncoupled systems for short lead times and slightly larger than reforecast for extended lead times. The ensemble spread of the coupled system grows gradually, but it is still under-dispersive (Figure 11). However, it is important to note that the coupled experiments have no initial perturbations and stochastics for the ocean.

The surface latent heat flux (LHTFL) and downward shortwave radiation flux have been selected as examples to demonstrate the differences between coupled GEFS and uncoupled GEFS because these two fluxes are dominant in ocean surface energy budget (Figure 12). For latent heat flux, all of the experiments have positive biases which indicate either surface wind stress or relative humidity vertical gradients are larger than reality. The bias differences between the GEFSv12 reforecast and the two experiments are relatively small compared to the overall magnitude of the bias. Furthermore, CGEFS-H has the smallest bias for global latent heat flux. The RMS error and spread are very similar as well between the three experiments. The ensemble spreads are in reasonable agreement with the RMS errors. For downward shortwave

radiation, CGEFS-H bias is larger than the GEFSv12 reforecast, but CGEFS-L has a smaller bias than the reforecast. A larger positive bias of downward shortwave radiation may indicate the presence of fewer clouds. The global RMS errors of downward shortwave flux in the two coupled experiments are very close to the GEFSv12 reforecast, while their spread is slightly reduced.

5. Discussion and Summary

For the 2-year experiments, high and low resolution coupled global ensembles (CGEFS-L and CGEFS-H) have been compared to the un-coupled GEFS (GEFSv12 reforecast) for day-to-day forecasts, weekly, bi-weekly and monthly averaged forecasts from ensemble mean (similar to deterministic) and probabilistic forecasts in terms of the forecast distribution. The TC track error/spread/intensity and MJO skills (RMM1+RMM2, RMS error and spread) are verified to demonstrate the forecast skills and predictability. The predictions of the SST and surface fluxes are evaluated with respect to overall changes stemming from the coupling of atmospheric, ocean, and sea ice models. Moreover, the diagnostics to measure the relationship of forecast uncertainty and RMS error have been performed, which helps to complete the forecasts and find an optimal configuration for both ensemble initial perturbations and the model stochastic schemes for 25 km (CGEFS-H) and 50 km (CGEFS-L) model resolutions. The investigation is more focused on the model stochastic schemes, rather than initial perturbations.

Overall, the coupled GEFS (CGEFS-L and CGEFS_H) has extended prediction skills of 500 hPa geopotential height and MJO mostly from the GEFSv12 reforecast (or current operational GEFS). Coupled GEFS is significantly better than uncoupled GEFS for week-2 of both hemispheres (extra-tropical area) and significantly better than uncoupled GEFS for all lead-time for tropical areas. CGEFS-H is better than CGEFS-L for weather prediction or short range forecasts, the skills are very similar for the extended range (week-2 and week 3 & 4) forecasts. These results indicate that (a) the fully coupled GEFSs improve the forecast skills and (b) The model resolution is very important for weather forecasting but has little impact on extended range forecast.

The TC track and intensity error have been evaluated for a limited number of cases. The TC track error of CGEFS-H is slightly lower than GEFSv12 reforecast and CGEFS-L for short lead times, while the intensity is much better than the GEFSv12 reforecast and CGEFS-L. Apparently, higher resolution models should provide more accurate TC intensity.

The forecast uncertainties (or ensemble spread) for the coupled GEFS (CGEFS-L and CGEFS-H) are better than the GEFSv12 reforecast, which is comparable to RMS errors after adjusting the parameters of the SPPT and SKED stochastic schemes. An improvement of ensemble spread results directly in an enhancement in the probabilistic forecast skill. The tropical forecast spread of zonal winds may still be slightly larger than it should be

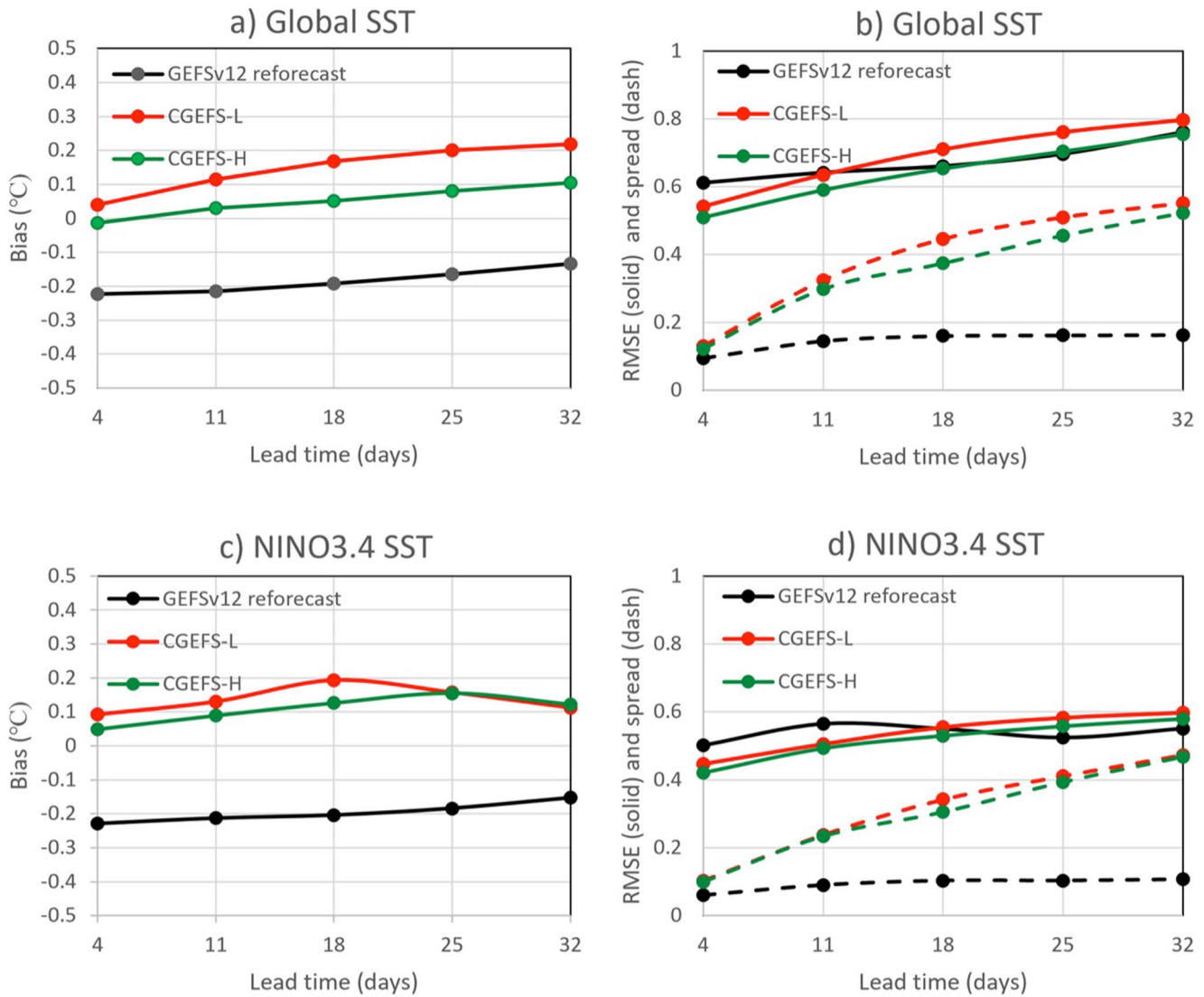


Figure 11. The SST bias, root mean square error and spread of the global (50°N-50°S and ocean only) and Nino 3.4 domains for the GEFSv12 reforecast (black), CGEFS-L (red) and CGEFS-H (green or blue).

even though it has been reduced a lot. An MJO skill could be improved furtherly if the spread of tropical winds (including MJO spread) is reduced slightly. For the extended range forecast, the variability of ensemble spread is highly reliant on the model dynamics, physics and stochastic perturbation schemes. Therefore, the results could be changed if the model is upgraded significantly.

By using LVM, the correlation of model error variance and ensemble forecast variance is clearly favored by the high resolution coupled model (CGEFS) for tropical zonal winds. The LVM is also able to assimilate the optimum relationship (or correlation) if there is an infinite ensemble size. It is different from the common measurement of error-spread ratio, which is based on the statistics (“reliability”) of the domain and time average, the LVM could detect the true correlation (“resolution”) of the variability on the forecast error and forecast spread.

The next experiment has been planned to adopt a new UFS GFS package and others which include the NOAA MP model to replace the NOAA LSM model, recall NSST back to assimilate diurnal variation (night cooling) of SST and CA (cellular automata) to advance physical perturbation from the convection scheme. Meanwhile, the vertical resolution will be increased from 64 to 97 hybrid levels with a model top of 0.01 hPa (~54 km).

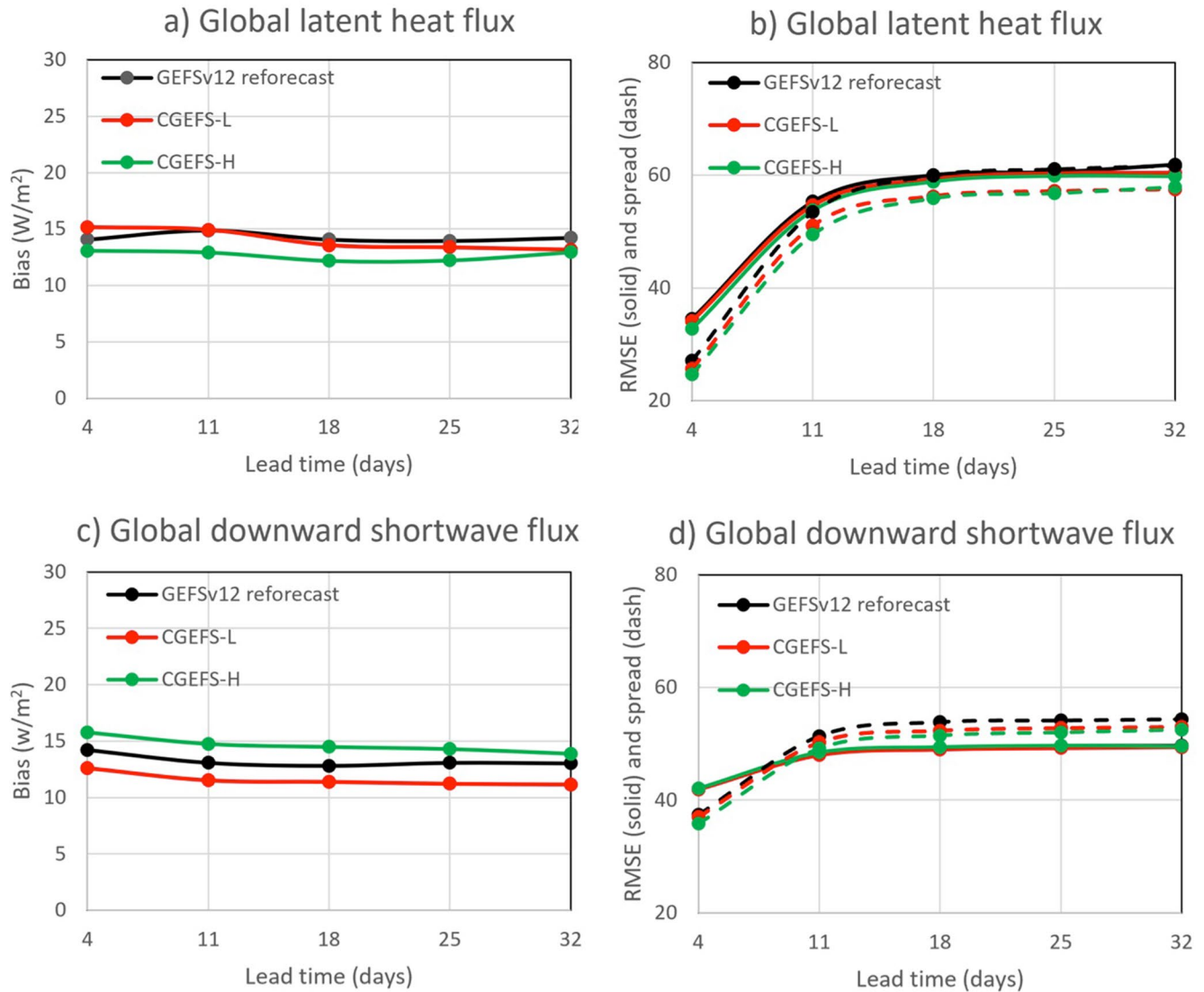


Figure 12. The downward shortwave radiation flux and surface latent heat flux 's bias, root mean square error and spread of the global (50°N-50°S and ocean only; top) and Nino 3.4 domains (bottom) for the GEFSv12 reforecast (black), CGEFS-L (red), and CGEFS-H (green).

Data Availability Statement

The climatology and analysis are from NCEP-NCAR Reanalysis (Kalnay et al., 1996). The reforecast and real-time operation data could be accessed through either International Research Institute for Climate and Society (IRI) of Columbia University at <http://iridl.ldeo.columbia.edu/SOURCES/.Models/.SubX/.EMC/.GEFS/> or Amazon Web-Service (free) at <https://noaa-gefs-retrospective.s3.amazonaws.com/index.html>. The Operational Sea Surface Temperature and Ice Analysis: documentation (OSTIA) and data can be accessed through UKMet office website at <https://ghrsst-pp.metoffice.gov.uk/ostia-website/index.html> and JPL website at <http://podaac-ftp.jpl.nasa.gov/allData/ghrsst/data/L4/GLOB/UKMO/OSTIA/2019/001/20190101-UKMO-L4HRfnd-GLOB-v01-fv02-OSTIA.nc.bz2>. The latent and sensible heat fluxes, and clouds are provided by the ECMWF Reanalysis Version 5 (ERA5; Hersbach et al., 2020). The Earth's Radiant Energy System (CERES) product for radiation fluxes are from NASA Langley Research Center (Rutan et al., 2015). The Outgoing Longwave Radiation (OLR) dataset is from NCAR (Liebmann & Smith, 1996).

Acknowledgments

The authors would like to thank Drs. Lydia Stefanova, Jiande Wang at EMC for evaluation and configuration support; to thank two anonymous reviewers for providing valuable recommendations; to thank the EMC coupled modeling team for developing the coupled UFS prototype; to thank Drs. Vijay Tallapragada, Fanglin Yang, Avichal Mehra, Matthew Rosencrans, and Phillip Pegion at EMC (and CPC, PSL) for valuable discussion pertaining to the design of our experiments. This study is partially supported by NOAA's Weather Program Office (WPO)'s Climate Testbed (CTB) project (U8R1CRS-P00).

References

Adcroft, A., Anderson, W., Balaji, V., Blanton, C., Bushuk, M., Dufour, C. O., et al. (2019). The GFDL global ocean and sea ice model OM4.0: Model description and simulation features. *Journal of Advances in Modeling Earth Systems*, *11*(10), 3167–3211. GFDL Ocean Circulation Model. <https://doi.org/10.1029/2019MS001726>

Berner, J., Shutts, G. J., Leutbecher, M., & Palmer, T. N. (2009). A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *Journal of the Atmospheric Sciences*, *66*(3), 603–626. <https://doi.org/10.1175/2008jas2677.1>

Bloom, S. C., Takacs, L. L., da Silva, A. M., & Ledvina, D. (1996). Data assimilation using incremental analysis updates. *Monthly Weather Review*, *124*(6), 1256–1271. [https://doi.org/10.1175/1520-0493\(1996\)124<1256:DAUIAU>2.0.CO;2](https://doi.org/10.1175/1520-0493(1996)124<1256:DAUIAU>2.0.CO;2)

Buizza, R., Miller, M., & Palmer, T. (1999). Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of Royal Meteorological Society*, *125*(560), 2887–2908. <https://doi.org/10.1002/qj.49712556006>

Christensen, H. M., Lock, S.-J., Moroz, I. M., & Palmer, T. M. (2017). Introducing independent dataset patterns into the stochastically perturbed parametrization tendencies (SPPT) scheme. *Quarterly Journal of Royal Meteorological Society, Part A*, *143*(706), 2168–2181. <https://doi.org/10.1002/qj.3075>

Gottschalck, J., Wheeler, M., Weickmann, K., Vitart, F., Savage, N., Lin, H., et al. (2010). A framework for assessing operational Madden-Julian oscillation: A CLIVAR MJO working group project. *Bulletin of American Meteorological Society*, *91*(9), 1247–1258. <https://doi.org/10.1175/2010bams2816.1>

Griffies, S., Adcroft, A., & Hallberg, R. (2020). A primer on the vertical Lagrangian-Remap method in ocean models based on finite volume generalized vertical coordinates. *Journal of Advances in Modeling Earth Systems*, *12*(10). <https://doi.org/10.1029/2019MS001954>

Guan, H., Zhu, Y., Sinsky, E., Fu, B., Li, W., Zhou, X., et al. (2022). GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. *Monthly Weather Review*, *150*(3), 647–665. <https://doi.org/10.1175/MWR-D-21-0245.1>

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

Hamill, T. M., Whitaker, J. S., Shlyaeva, A., Bates, G., Fredrick, S., Pegion, P., et al. (2022). The reanalysis for the global ensemble forecast system, version 12. *Monthly Weather Review*, *150*(1), 59–79. <https://doi.org/10.1175/MWR-D-21-0023.1>

Han, J., Wang, W., Kwon, Y. C., Hong, S.-Y., Tallapragada, V., & Yang, F. (2017). Updates in the NCEP GFS cumulus convection schemes with scale and aerosol awareness. *Weather Forecasting*, *32*(5), 2005–2017. <https://doi.org/10.1175/WAF-D-17-0046.1>

Held, I. M., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J., et al. (2019). Structure and performance of GFDL's CM4.0 climate model. *Journal of Advances in Modeling Earth Systems*, *11*(11), 3691–3727. <https://doi.org/10.1029/2019MS001829>

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horanyi, A., Munoz-Sabater, J., et al., (2020). The ERA5 global reanalysis [Dataset]. *Quarterly Journal of Royal Meteorological Society*, *146* (730), 1999–2049. <https://doi.org/10.1002/qj.3803>

Hou, D., Toth, Z., Zhu, Y., & Yang, W. (2008). Evaluation of the impact of the stochastic perturbation schemes on global ensemble forecast. In *Proceedings of the 19th Conference on Probability and Statistics*. American Meteorological Society. Retrieved from <https://ams.confex.com/ams/88Annual/webprogram/Paper134165.html>

Kalnay, E., and co-authors. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, *77*(3), 437–471. [https://doi.org/10.1175/1520-0477\(1996\)077<0437:TNYRP>2.0.CO;2](https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2)

Kleist, D. T., & Ide, K. (2015). An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Part II: 4D EnVar and hybrid variants. *Monthly Weather Review*, *143*(2), 452–470. <https://doi.org/10.1175/MWR-D-13-00350.1>

Kolczynski, W., Stauffer, D., Haupt, S. E., Altman, N. S., & Deng, A. (2011). Investigation of ensemble variance as a measure of true forecast variance. *Monthly Weather Review*, *139*(12), 3954–3963. <https://doi.org/10.1175/mwr-d-10-05081.1>

Li, K., Yu, W., Yang, Y., Feng, L., Liu, S., & Li, L. (2020). Spring barrier to the MJO eastward propagation. *Geophysical Research Letters*, *46*(13), e2020GL087788. <https://doi.org/10.1029/2020GL087788>

Li, W., Zhu, Y., Zhou, X., Hou, D., Sinsky, E., Melhauser, C., et al. (2019). Evaluating the MJO forecast skill from different configurations of NCEP GEFS extended forecast. *Climate Dynamics*, *52*(7–8), 4923–4936. <https://doi.org/10.1007/s00382-018-4423-9>

Liebmann, B., & Smith, C. A. (1996). Description of a Complete (Interpolated) Outgoing Longwave Radiation Dataset. [Dataset]. *Bulletin of American Meteorological Society*, *77*, 1275–1277. Retrieved from <https://www.jstor.org/stable/26233278>

Lin, H., Brunet, G., & Derome, J. (2008). Forecast skill of the Madden-Julian oscillation in two Canadian atmospheric models. *Monthly Weather Review*, *136*(11), 4130–4149. <https://doi.org/10.1175/2008MWR2459.1>

Ling, J., Zhao, Y., & Chen, G. (2019). Barrier effect on MJO propagation by the maritime continent in the MJO Task Force/GEWEX atmospheric system study models. *Journal of Climatology*, *32*(17), 5529–5547. <https://doi.org/10.1175/JCLI-D-18-0870.1>

Madden, R., & Julian, P. (1972). Description of global-scale circulation cells in the tropics with a 40–50 Day period. *Journal of the Atmospheric Sciences*, *29*, 1109–1123. [https://doi.org/10.1175/1520-0469\(1972\)029<1109:DOGSCC>2.0.CO;2](https://doi.org/10.1175/1520-0469(1972)029<1109:DOGSCC>2.0.CO;2)

Palmer, T. R. B., Reyes, F. D., Jung, T., Leutbecher, M., Shutts, G., Steinheimer, M., & Weisheimer, A. (2009). Stochastic parametrization and model uncertainty. *ECMWF Technical Memorandum*, *598*, 42. <http://www.ecmwf.int/publications/>

Pegion, K., Kirtman, B. P., Becker, E., Collins, D. C., LaJoie, E., Burgman, R., et al. (2019). The subseasonal experiment (SubX): A multi-model subseasonal prediction experiment. *Bulletin of American Meteorological Society*, *100*(10), 2043–2060. <https://doi.org/10.1175/BAMS-D-18-0270.1>

Reynold, D., & Smith, T. M. (1994). Improved global Sea Surface temperature analyses using optimum interpolation. *Journal of Climate*, *7*, 929–948. [https://doi.org/10.1175/1520-0442\(1994\)007<0929:IGSSTA>2.0.CO;2](https://doi.org/10.1175/1520-0442(1994)007<0929:IGSSTA>2.0.CO;2)

Rutan, D. A., Kato, S., Doelling, D. R., Rose, F. G., Nguyen, L. T., Caldwell, T. E., & Loeb, N. G. (2015). CERES synoptic product: Methodology and validation of surface radiant flux [Dataset]. *Journal of Atmospheric and Oceanic Technology*, *32*(6), 1121–1143. <https://doi.org/10.1175/JTECH-D-14-00165.1>

Saha, S., Moorthi, S., Pan, H. L., Wu, X., Wang, J., Nadiga, S., et al. (2010). The NCEP climate forecast system reanalysis. *Bulletin of American Meteorological Society*, *91*(8), 1015–1057. <https://doi.org/10.1175/2010BAMS3001.1>

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., et al. (2014). The NCEP climate forecast system version 2. *Journal of Climate*, *27*(6), 2185–2208. <https://doi.org/10.1175/JCLI-D-12-00823.1>

Shutts, G. (2005). A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quarterly Journal of Royal Meteorological Society*, *131*(612), 3079–3102. <https://doi.org/10.1256/qj.04.106>

Shutts, G. (2015). A stochastic convective backscatter scheme for use in ensemble prediction systems. *Quarterly Journal of the Royal Meteorological Society: Part A*, *141*(692), 2602–2616. <https://doi.org/10.1002/qj.2547>

- Shutts, G., & Palmer, T. N. (2004). *The use of high-resolution numerical simulations of tropical circulation to calibrate stochastic physics schemes*. In *Proc. Workshop on Simulation and Prediction of Intra-seasonal Variability with Emphasis on the MJO* (pp. 83–102). ECMWF. Retrieved from <https://www.ecmwf.int/en/learning/workshops-and-seminars/past-workshops/2003-simulation-prediction-intra-seasonal-variability>
- Stefanova, L., Meixner, J., Wang, J., Mehra, A., Worthen, D., Sun, S., et al. (2022). *Description and results from UFS-coupled prototypes for future global, ensemble and seasonal forecasts at NCEP*. NCEP Office Notes 510. <https://doi.org/10.15923/knxm-kz26>
- Tolman, H. L. (2008). *User manual and system documentation of WAVEWATCH III version 3.14* (Vol. 268, p. 192). NOAA/NWS/NCEP MMAB Tech.
- Toth, Z., Talagrand, O., & Zhu, Y. (2006). In T. N. Palmer, & R. Hagedorn (Eds.), (pp. 584–595). Cambridge University Press. The attributes of forecast systems. In *Book of: Predictability of Weather and Climate*.
- Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., et al. (2017). The Sub-seasonal to seasonal (S2S) prediction project database. *Bulletin of American Meteorological Society*, 98(1), 163–176. <https://doi.org/10.1175/BAMS-D-16-0017>
- Wei, M., Toth, Z., Wobus, R., & Zhu, Y. (2008). Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus*, 60A(1), 62–79. <https://doi.org/10.1111/j.1600-0870.2007.00273.x>
- Wheeler, M. C., & Hendon, H. (2004). An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Monthly Weather Review*, 132, 1917–1932. [https://doi.org/10.1175/1520-0493\(2004\)132<1917:AARMMI>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<1917:AARMMI>2.0.CO;2)
- White, L., Adcroft, A., & Hallberg, R. (2009). High-order regridding–remapping schemes for continuous isopycnal and generalized coordinates in ocean models. *Journal of Computational Physics*, 228(23), 8665–8692. <https://doi.org/10.1016/j.jcp.2009.08.016>
- WMO, Zhu, Y., & others. (2021). *Guidelines on ensemble prediction system postprocessing*. WMO publication. No. 1254 Retrieved from http://library.wmo.int/doc_num.php?explnum_id=10726
- Zhou, X., Zhu, Y., Fu, B., Hou, D., Peng, J., Luo, Y., & Li, W. (2019). The development of the next NCEP global ensemble forecast system. *STI Climate Bulletin*, 159–163.
- Zhou, X., Zhu, Y., Hou, D., Fu, B., Li, W., Guan, H., et al. (2022). The introduction of the NCEP global ensemble forecast system version 12. *Weather Forecasting*, 37(6), 1069–1084. <https://doi.org/10.1175/WAF-D-21-0112.1>
- Zhou, X., Zhu, Y., Hou, D., & Kleist, D. (2016). Comparison of the ensemble transform and the ensemble Kalman Filter in the NCEP global ensemble forecast system. *Weather Forecasting*, 31, 2058–2074.
- Zhou, X., Zhu, Y., Hou, D., Luo, Y., Peng, J., & Wobus, D. (2017). The NCEP global ensemble forecast system with the EnKF initialization. *Weather Forecasting*, 32(5), 1989–2004. <https://doi.org/10.1175/waf-d-17-0023.1>
- Zhu, Y., Iyengar, G., Toth, Z., Tracton, M. S., & Marchok, T. (1996). *Objective evaluation of the NCEP global ensemble forecasting system. Preprints, 15th Conf. on Weather Analysis and Forecasting*. Amer. Meteor. Soc. J79–J82.
- Zhu, Y., Li, W., Sinsky, E., Guan, H., Zhou, X., & Fu, B. (2019). An investigation of prediction and predictability of NCEP global ensemble forecast system (GEFS). *STI Climate Bulletin*, 154–158.
- Zhu, Y., Li, W., Zhou, X., & Hou, D. (2019). (pp. 317–328). Springer Atmospheric Science. Stochastic representation of NCEP GEFS to improve subseasonal forecast. In *the Book of Current Trends in the Representation of Physical Processes in Weather and Climate Models*.
- Zhu, Y., Toth, Z., Wobus, R., Richardson, D., & Mylne, K. (2002). On the economic value of ensemble based weather forecasts. *Bulletin of the American Meteorological Society*, 83(1), 73–83. [https://doi.org/10.1175/1520-0477\(2002\)083<0073:tevoeb>2.3.co;2](https://doi.org/10.1175/1520-0477(2002)083<0073:tevoeb>2.3.co;2)
- Zhu, Y., Zhou, X., Li, W., Hou, D., Melhauser, C., Sinsky, E., et al. (2018). Toward the improvement of sub-seasonal prediction in the NCEP global ensemble forecast system (GEFS). *Journal of Geophysical Research: Atmospheres*, 123(13), 6732–6745. <https://doi.org/10.1029/2018JD028506>
- Zhu, Y., Zhou, X., Pena, M., Li, W., Melhauser, C., & Hou, D. (2017). Impact of sea surface temperature forcing on weeks 3 and 4 forecast skill in the NCEP Global Ensemble Forecast System. *Weather Forecasting*, 32, 2159–2173. <https://doi.org/10.1175/WAF-D-17-0093.1>