Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE
10.1029/2018JD028506

Key Points:
- There is a capability to predict subseasonal timescale from the extended Global Ensemble Forecast System (GEFS)
- Improving model uncertainty representation for the tropics is important to MJO prediction
- The impacts of SST and tropical convection are also important for tropical prediction on the subseasonal timescale

Correspondence to:
Y. Zhu,
yuejian.zhu@noaa.gov

Citation:
Zhu, Y., Zhou, X., Li, W., Hou, D., Melhauser, C., Sinsky, E., et al. (2018). Toward the improvement of subseasonal prediction in the National Centers for Environmental Prediction Global Ensemble Forecast System. *Journal of Geophysical Research: Atmospheres*, *123*, 6732–6745. https://doi.org/10.1029/2018JD028506

Received 9 FEB 2018
Accepted 29 MAY 2018
Accepted article online 9 JUN 2018
Published online 5 JUL 2018

Toward the Improvement of Subseasonal Prediction in the National Centers for Environmental Prediction Global Ensemble Forecast System

Yuejian Zhu[1], Xiaqiong Zhou[2], Wei Li[2], Dingchen Hou[1], Christopher Melhauser[2], Eric Sinsky[2], Malaquias Peña[4], Bing Fu[2], Hong Guan[3], Walter Kolczynski[2], Richard Wobus[2], and Vijay Tallapragada[1]

[1]EMC, NCEP, NWS, NOAA, College Park, MD, USA, [2]IMSG at EMC, NCEP, NWS, NOAA, College Park, MD, USA, [3]SRG at EMC, NCEP, NWS, NOAA, College Park, MD, USA, [4]Department of Civil and Environmental Engineering, University of Connecticut, Storrs, CT, USA

Abstract In order to provide ensemble-based subseasonal (weeks 3 and 4) forecasts to support the operational mission of the Climate Prediction Center, National Centers for Environmental Prediction, experiments have been designed through the Subseasonal Experiment (SubX) project to investigate the predictability in both tropical and extratropical regions. The control experiment simply extends the current operational Global Ensemble Forecast System (GEFS; version 11) from 16 to 35 days. In addition to the control, the parallel experiments will be mainly designed to focus on three areas: (1) improving model uncertainty representation for the tropics through stochastic physical perturbations; (2) considering the impact of the ocean by using a two-tiered sea surface temperature approach; and (3) testing a new scale-aware convection scheme to improve the model physics for tropical convection and Madden-Julian Oscillation (MJO) forecasts. All experiments are initialized every 5 days at 0000 UTC during the period of May 2014–May 2016 (25 months). In the tropics, MJO forecast skill has been improved from an average of 12.5 days (control) to nearly 22 days by combining all three modifications to GEFS. In the extratropics, the ensemble mean anomaly correlation of 500-hPa geopotential height improved over weeks 3 and 4. In addition, the Continuous Ranked Probability Score (of the Northern Hemisphere raw surface temperature (land only) is improved as well. A similar result is found in the Contiguous United States precipitation, although forecast skill is extremely low. Our results imply that calibration may be important and necessary for surface temperature and precipitation forecast for the subseasonal timescale due to the large systematic model errors.

1. Introduction

The National Oceanic and Atmospheric Administration (NOAA) is working on a unified modeling system through the National Weather Service Next Generation Global Prediction System project (NOAA National Weather Service, 2014) to provide the best possible guidance to a wide customer base, including emergency managers, forecasters, and the aviation community, to protect life and property and enhance the national economy. In 2016, the National Academies of Sciences, Engineering, and Medicine published the strategies for subseasonal to seasonal forecasts through the next generation Earth system prediction (National Academies of Sciences, Engineering, and Medicine, 2016). The dominant factors that impact model prediction for medium-term (7 to 10 days) forecasts are slightly different than for extended range (multiweeks) forecasts. A numerical weather prediction system (including an ensemble forecast system) heavily relies on atmospheric initial conditions (including initial uncertainties) and model parameterizations for a short-range forecast. As forecast lead time increases, the impact of the ocean and other external forcing in the Earth system become important; thus, they cannot be neglected (Li et al., 2009; Liu et al., 2016). Various studies have shown that the oceanic variation at the subseasonal timescale is primarily related to thermodynamic forcing (Li et al., 2001; Ling et al., 2015). The impacts of large-scale ocean currents are of secondary importance (Takaya et al., 2010).

Subseasonal forecasts span the time period between weather and seasonal (and/or climate) forecasts. Two of the leading systems (the European Centre for Medium-Range Weather Forecasts global ensemble system and the National Centers for Environmental Prediction (NCEP) Climate Forecast System version 2 (CFS v2)) are extending weather forecast to cover subseasonal timescale but having limited value, such as

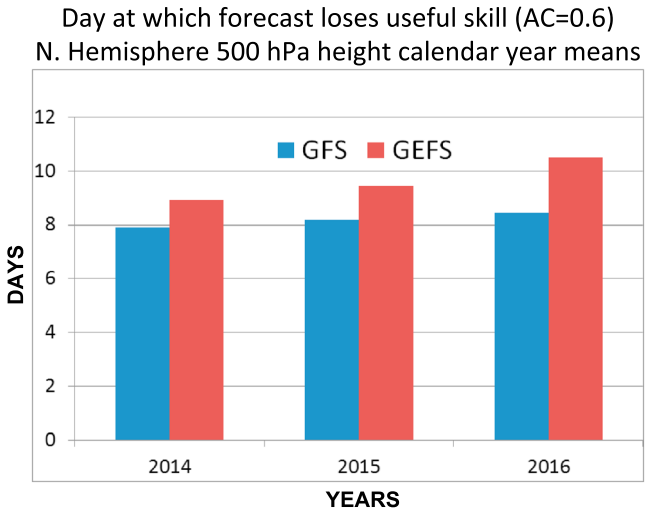

Figure 1. Lead time, in days, at which forecast loses useful skill (anomaly correlation [AC] = 0.6) for Northern Hemisphere (20°N–80°N) 500-hPa geopotential height of Global Forecast System (GFS) forecast (blue) and Global Ensemble Forecast System (GEFS) ensemble mean forecast (red) averaged over each year in 2014–2016.

surface temperature and precipitation for Northern Hemisphere (NH) extropical area (Saha et al., 2014; Vitart, 2014). Currently, there are no optimal configurations of numerical weather or climate models that can provide skillful forecast covering the subseasonal timescale. With the ultimate goal to improve forecast skill and deliver useful numerical guidance at subseasonal timescales, we explore the forecast skill of an extended-range version of the NCEP Global Ensemble Forecast System (GEFS). An early effort has already shown positive results when extending the forecast lead of the GEFS and updating the underlying sea surface temperature (SST) boundary conditions (Zhu et al., 2017).

The NCEP GEFS has been widely used as probabilistic forecast guidance for weather (0–7 days) and week-2 anomaly forecast since it was implemented into operation in 1992 (Zhou et al., 2016, 2017). The products include upper-air atmospheric variables, surface weather elements, precipitation, and hurricane tracks. The operational GEFS has provided more skillful numerical guidance compared to the operational high-resolution deterministic Global Forecast System (GFS; Figure 1) based on length of forecast exceeding 0.6 mean pattern anomaly correlation (AC, and thereafter; Wilks, 2011) in the NH (20°N–80°N) 500-hPa geopotential height for the past 3 years. The GEFS ensemble mean improves from 8.92 days of skillful forecast (AC $>$ 0.6) in 2014 (GEFS v10) to 10.52 days in 2016 (GEFS v11) in contrast to the GFS that improve from 7.9 days in 2014 to 8.45 days in 2016. Improvement of the numerical guidance is achieved through model upgrades that include enhanced data assimilation, model physics, initial perturbations, stochastic physics perturbations, and model resolution.

In contrast to current operational CFS v2 (Saha et al., 2014), there are several advantages in extending GEFS to cover the subseasonal timescale, including (1) improved initial perturbations using an ensemble Kalman filter (EnKF) data assimilation system (Zhou et al., 2017), to better represent observation and analysis uncertainties; (2) increased horizontal resolution for weather, allowing small-scale process to be resolved and more realistic interactions around multiple scales; (3) advanced model physics with various stochastic physics perturbation schemes to represent model uncertainties (Han et al., 2017); (4) increased ensemble size (i.e., GEFS currently runs 80 + 4 members each day) to provide more reliable probabilistic guidance; (5) suitable configuration (ensemble size and frequency) for real-time reforecasts/hindcasts for calibration; and (6) seamless forecasts across weather and seasonal timescales.

Based on the performance of GEFS for the extended range forecast in our early investigation on SST forcing (Zhu et al., 2017), this study builds on those results and provides an additional comprehensive understanding from upgraded GEFS model with higher resolution, taking into account the key factors that may contribute to improve forecast skill in the extended range for GEFS. In this investigation, the NCEP operational GEFS v11 (Zhou et al., 2017) is used as the base configuration. The goal is to focus and improve our common understanding in three different physical processes (or scientific areas). The first area is to improve tropical forecast uncertainty. Unlike extratropical baroclinic systems, the tropics exhibit less potential error growth modes from optimal initial perturbations (Toth & Kalnay, 1993, 1997). The lack of error growth in the tropics is especially pronounced in the current GEFS due to its underdispersion (Zhou et al., 2017). To improve this deficiency, we use a combination of stochastic physical perturbation schemes to represent tropical forecast uncertainty. The second area is to more realistically consider the day-to-day variability of the underlying SST using a two-tiered SST approach. This allows for a one-way interaction between the ocean and atmosphere that may be an improvement in representing the impact of ocean forcing (Melhauser et al., 2016; Zhu et al., 2017). The third area is to test a new convection scheme with a scale-aware parameterization that can more realistically represent the vertical motion, radiative impact, and cloud and precipitation processes associated with tropical convection (Han et al., 2017).

The experiment design and configurations will be described in section 2. The forecast skill will be evaluated in section 3, and an evaluation of ensemble size will be discussed in section 4. Summary and conclusions along with future work will be presented in section 5.

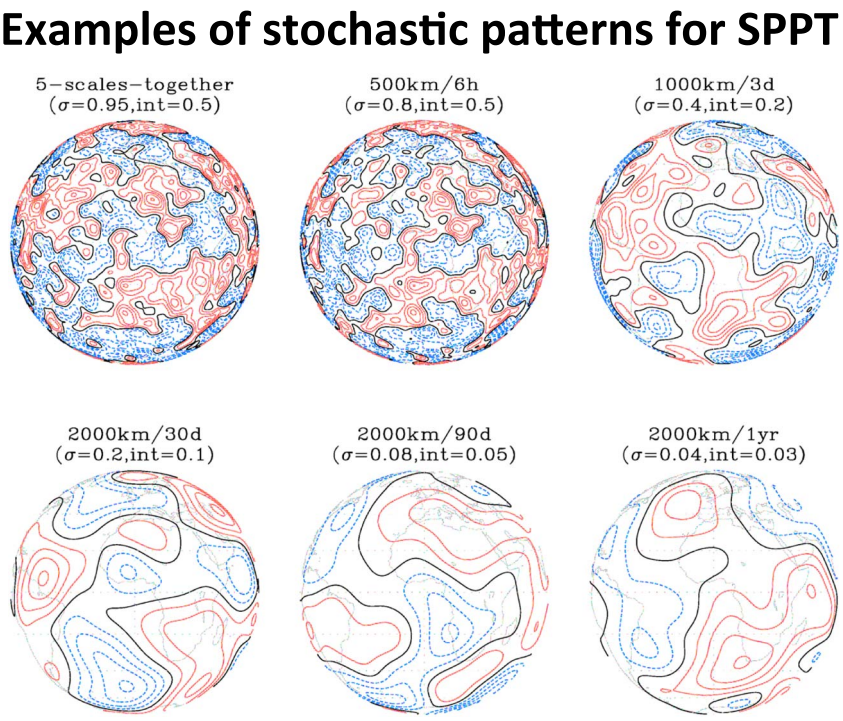

Figure 2. Five-scale random patterns used in stochastic perturbed physics tendencies (SPPT). On the top of each plot, the numbers (except for upper left; nondimension) represent the scales of spatial and temporal perturbations with the maximum amplitude and contour intervals in the bracket.

2. Experiment Design and Configurations

In this paper, we test three configurations to explore the forecast skill of GEFS on subseasonal prediction. In the design of each experiment, we combined the changes based on earlier investigations on the effect of some of the configurations (Han et al., 2017; Melhauser et al., 2016; Zhu et al., 2017). Although it is useful to independently examine the impact of each configuration change for a full experiment period, running these permutations for the full experiment period would be too computationally expensive with a high-resolution GEFS and 21 ensemble members.

2.1. Operational GEFS v11 and Extended Forecast—"CTL"

As a control, we apply a simple extension of forecast lead based on the current operational GEFS (v11). The operational GEFS v11 was implemented on 2 December 2015. It uses a reduced horizontal resolution version of the NCEP GFS Global Spectral Model (with semi-Lagrangian dynamics) v12.0 (GSM). GEFS is initialized four times per day (0000, 0600, 1200, and 1800 UTC) and integrated to 16 days (Han et al., 2017; Sela, 1980). The horizontal resolution is approximately 34 km for days 0–8 and 52 km for days 8–16 with 64 hybrid vertical levels. More details of GEFS v11 can be found in Zhou et al. (2017), Melhauser et al. (2016), and Zhu et al. (2017). For the extended GEFS forecast in this study, the horizontal resolution of forecast lead days 16–35 is the same as days 8–16 (about 52 km). In addition, the GEFS uses the same SST forcing as the GFS, which is initialized with the Real Time Global (RTG) SST analysis (Gemmill et al., 2007) and damped to analysis climatology (90-day e-folding; Melhauser et al., 2016; Zhu et al., 2017) during model integration. The sea ice concentration is initialized from the daily 0000 UTC sea ice analysis from the Interactive Multisensor Snow and Ice Mapping System (Ramsay, 1998).

2.2. New Stochastic Perturbed Physics Schemes—"SPs"

The first experiment replaces the stochastic total tendency perturbation (STTP; Hou et al., 2008) used in the control (and the operational GEFS) with a suite of three "stochastic physics" (SP) methods. This SP suite introduces a better representation of model uncertainty into the model integration. This is a natural first

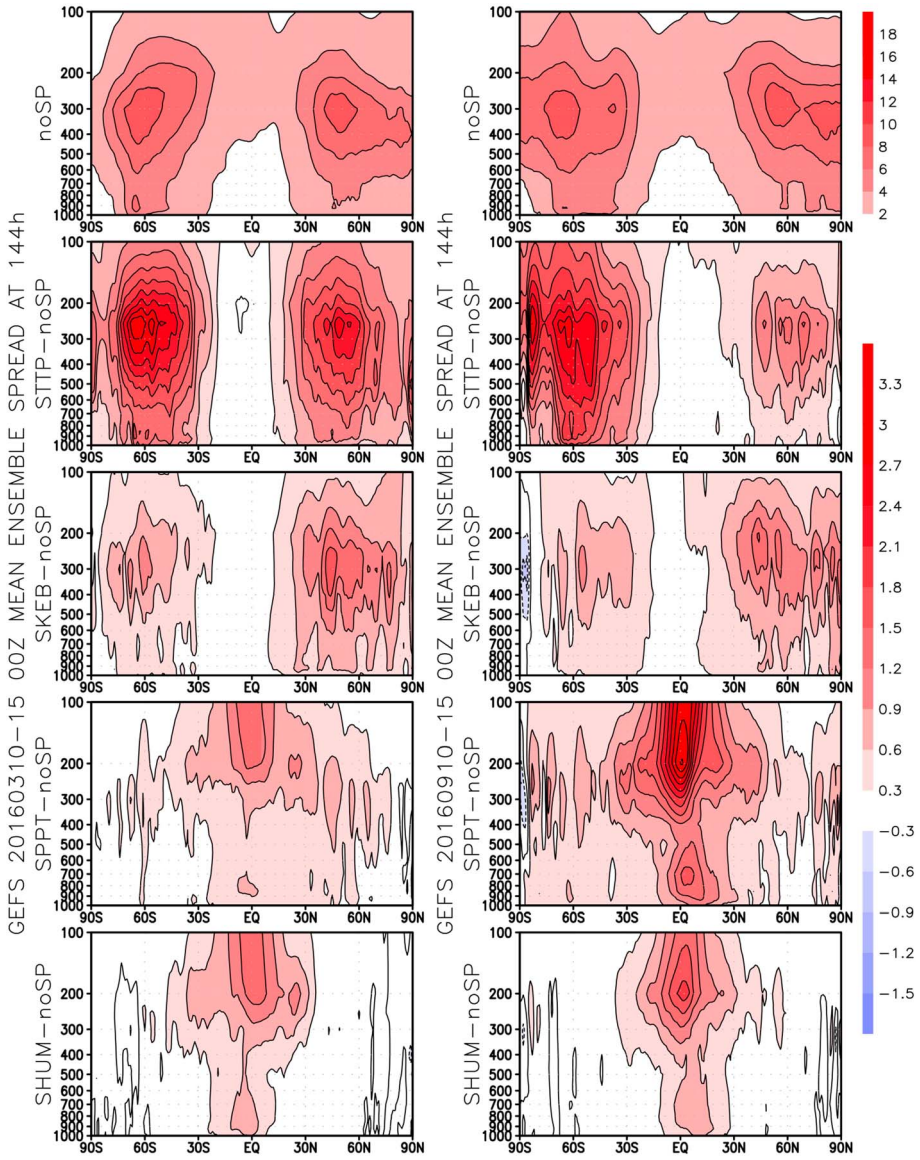

Figure 3. Global meridional cross section showing the impact of stochastic perturbations for 144-hr forecasts from the average of six spring initializations (left) and six fall initializations (right). Paneled are zonal wind spread (m/s) from no stochastic perturbations (noSP; top), and the difference of Stochastic Total Tendency Perturbation (STTP; upper middle), stochastic kinetic energy backscatter (SKEB; middle), stochastically perturbed parameterization tendencies (SPPT; lower middle), and stochastic perturbed humidity (SHUM; bottom) from noSP. GEFS = Global Ensemble Forecast System.

experiment, as this SP suite is planned for inclusion in the next configuration of the operational GEFS, and the stochastic physics schemes are already a part of the operational GFS data assimilation system.

The current method of modeling model uncertainty, STTP, introduces additional ensemble spread by adding a random combination of the 6-hr tendency from other ensemble members and has been used in operations since GEFS v9 (2010). The impact of STTP is largely constrained to the extratropical region and the boreal winter; there are only minor impacts in the tropical region. To replace STTP, a suite of three widely accepted SP methods are used: (1) stochastic kinetic energy backscatter (SKEB; Shutts & Palmer, 2004; Shutts, 2005; Berner et al., 2009); (2) stochastically perturbed parameterization tendencies (SPPT; Buizza et al., 1999; Palmer et al., 2009); and (3) stochastic perturbed humidity (SHUM; Tompkins & Berner, 2008). These schemes have already been implemented in the hybrid-EnKF data assimilation system of GFS, making them readily available for use in the GEFS. SPPT perturbs the combined tendencies of wind, temperature, and water vapor produced by the

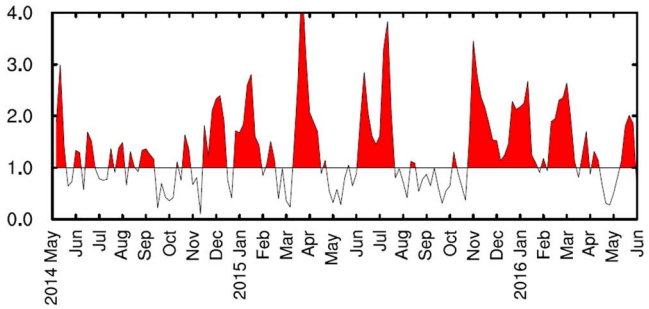

Figure 4. Amplitude of Madden-Julian Oscillation during May 2014 to May 2016 from Global Data Assimilation System analysis data. The resolution of the time series is 5 days.

GFS physical parameterizations (excluding clear-air radiation). Our implementation of SPPT combines five different random patterns with different correlation length scales and/or timescales to determine the perturbations. The patterns are uniform in the vertical, except they are reduced in magnitude near the surface and tapered to zero near and above the tropopause. An example of the individual independent scale random patterns and combined five-scale random pattern is shown in Figure 2 (values of sigma represent maximum amplitude).

SHUM introduces stochastic perturbation to represent model uncertainty of humidity. It perturbs the near-surface humidity state to present subgrid distribution and variability through convection. SHUM uses the same random pattern generator as SPPT, but only a single spatial-temporal scale is used. The perturbation is maximized in the lowest model level and decreases exponentially with height.

SKEB is used to represent uncertainty in representing the upscale impacts from subgrid-scale processes. This is done via a stream function forcing based on the dissipation rate. Unlike other implementations of SKEB, the GFS implementation of SKEB only considers numerical dissipation (diffusion). Perturbations are generated independently on each vertical level and then vertically smoothed to provide some vertical coherence. The inclusion of SKEB improves the power spectrum of kinetic energy of the global model, which otherwise exhibits damped power near the truncation frequency.

Figure 3 shows the impact of the individual stochastic schemes on ensemble spread at 144 hr (average of six spring cases and six fall cases) when compared to ensemble spread without introducing stochastic perturbations (noSP, top row). The stochastic system used in the control (and operational GEFS), STTP, produces additional spread in the extratropics but has little impact in the tropics (row 2). Of the components in the stochastic physics suite, SKEB produces additional spread in similar areas as STTP (though slightly muted; row 3). The other two components, SPPT (row 4) and SHUM (row 5), both increase spread in the tropics where the control system is deficient. SPPT also has an impact in the spring/summer hemisphere. In combination, these stochastic schemes generally improve forecast uncertainty, particularly in the tropics, which may lead to improvements in tropical forecast skill.

2.3. Two Tiered SST Approach—"SPs + SST_bc"

In addition to the new stochastic perturbed physics schemes discussed in section 2.2, this experiment uses a "two-tiered" SST approach for lower boundary conditions over ocean by considering an evolving ocean SST state gradually increasing with lead time. The two-tiered SST approach relaxes the RTG SST analysis to a bias-corrected SST prediction from the operational CFS v2 (Saha et al., 2014) rather than the climatological SST that is used in the control configuration. The relaxation SST is updated to the forecast CFS SST every 24 hr of model integration (Melhauser et al., 2016; appendix I of Zhu et al., 2017). While we consider this an interim configuration in place of true coupling, this method should still be an improvement over relaxation to climatology. Compared to the operational version of the SST scheme that relaxes the initial SST analysis to climatological SST, the lower boundary of atmospheric model receives evolving bias-corrected SSTs from a coupled model; thus, the boundary SST is represented more realistically. Also, compared to a fully coupled forecast system, the computational resources are significantly reduced since an ocean and sea ice model are not required. While this is not a true coupled approach, the impact of using predicted ocean variation from the fully coupled system (CFS v2) might be determined through various evaluations, such as the Madden-Julian Oscillation (MJO) skill and other tropical low-level forecast elements of the extended GEFS forecast. Meanwhile, work is ongoing at NCEP to produce a fully coupled GEFS (atmosphere, ocean, sea ice, wave, and aerosol).

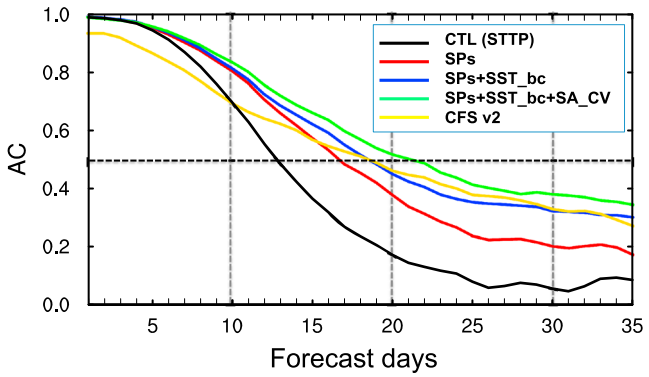

Figure 5. Madden-Julian Oscillation RMM1 + RMM2 forecast skill for CTL (black), SPs (red), SPs + SST_bc (blue), SPs + SST_bc + SA_CV (green), and CFSv2 (yellow) during the period of May 2014 to May 2016. AC = anomaly correlation.

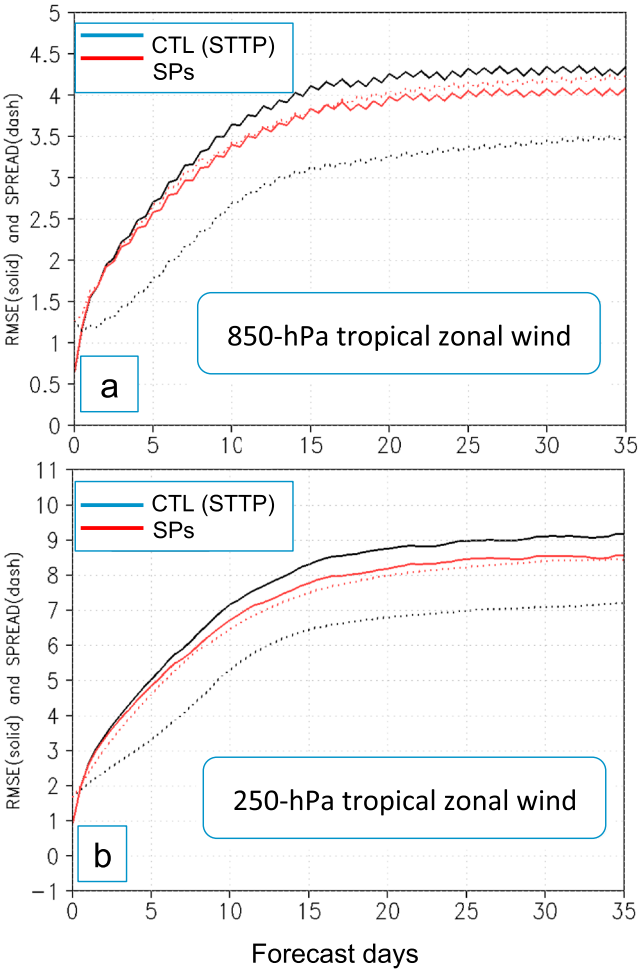

Figure 6. Root mean square error (RMSE) of the ensemble mean (solid) and the ensemble spread (dash) plotted every 12 hr out to 35 days for 850- (top) and 250-hPa (bottom) tropical (20°N–20°S) zonal wind (m/s) during the May 2014 to May 2016 period comparing CTL (black) and SPs (red).

2.4. New Scale Aware Convective Scheme—"SPs + SST_bc + SA_CV"

The final component tested for this extended forecast system is an upgraded Simplified Arakawa-Schubert cumulus parameterization scheme that is both scale- and aerosol-aware (Han et al., 2017). This upgraded scheme was recently implemented into the operational GFS v14 on 19 July 2017. Since this change would presumably be included in the future GEFS upgrade as well, it is logical to consider this change here.

The main changes in the upgraded scheme include the following: (1) the rain conversion rate decreases with decreasing air temperature above the freezing level; (2) convective adjustment time in deep convection is proportional to the convective turnover time, with convective available potential energy approaching zero after the adjustment time; (3) cloud base mass flux in the shallow convection scheme is modified to use convective turnover time as the convective adjustment timescale; (4) convective inhibition in the subcloud layer is an additional trigger condition to suppress unrealistic spotty rainfall, especially over high terrain during the summer; and (5) convective cloudiness is enhanced by suspended cloud condensate in an updraft (Han et al., 2017).

Retrospective runs of the NCEP GFS v14 implementation (an approximately 13-km horizontal resolution deterministic forecast; Han et al., 2017) using this updated Simplified Arakawa-Schubert scheme have improved precipitation forecasts for Contiguous United States (CONUS) at short lead times (equitable threat score is higher, and frequency bias is reduced; not shown).

2.5. Experiment Configurations

Each of the four experiments is initialized every 5 days at 0000 UTC for a 25-month period from May 2014 to May 2016. This cadence allows us to sample each season over multiple years while limiting computational load and also satisfies the requirements of one of the primary customers for subseasonal forecasts, Climate Prediction Center (~1/week). Each experiment runs 20 perturbed forecasts and a control (unperturbed) forecast out to 35 days. The ensemble resolution for all experiments is identical to the operational GEFS out to 16 days, with the low-resolution portion extended to 35 days. The initial analyses (or initial condition) and perturbations are also identical for all experiments and taken from the NCEP GFS v13 data assimilation system (with the hybrid analysis—4DEnVar hybrid analysis), which was implemented in May 2016.

3. Evaluation

The verification in this study will evaluate the performance (forecast skill) difference for the three designed experiments compared to CTL. The evaluation covers tropical and extratropical areas, including upper atmosphere and surface variables applying deterministic (ensemble mean) and probabilistic (ensemble distribution) verifications. The evaluation varies from single variables to phenomena that combine the effect of multiple variables such as MJO. For extended range forecasts, such as week 2 and weeks 3 and 4, an average over the period (or an accumulation within the period for precipitation) has been used to evaluate forecast skill. All verifications are on 1° × 1° or 2.5° × 2.5° lat-lon grid points weighted by latitude. The proxy truths for verification use the NCEP Global Data Assimilation System analysis and the Climatological Calibrated Precipitation Analysis (Hou et al., 2014). The skills are defined with respect to NCEP/NCAR 40-year reanalysis (Kalnay et al., 1996) climatology.

Table 1
Days of Useful Madden-Julian Oscillation Forecast Skill (0.5 of Amplitude Correlation Coefficient) for the Four Experiments Over the Full Period (1 May 2014 to 26 May 2016) and for a Weak (May 2014 to March 2015) and Strong Period (April 2015 to March 2016)

Configurations	Weak	Strong	25 months
CTL (STTP)	12.2	12.8	12.5
SPs	15.8	18	16.8
SPs + SST_bc	17	19.5	18.5
SPs + SST_bc + SA_CV	18+	23+	22.0

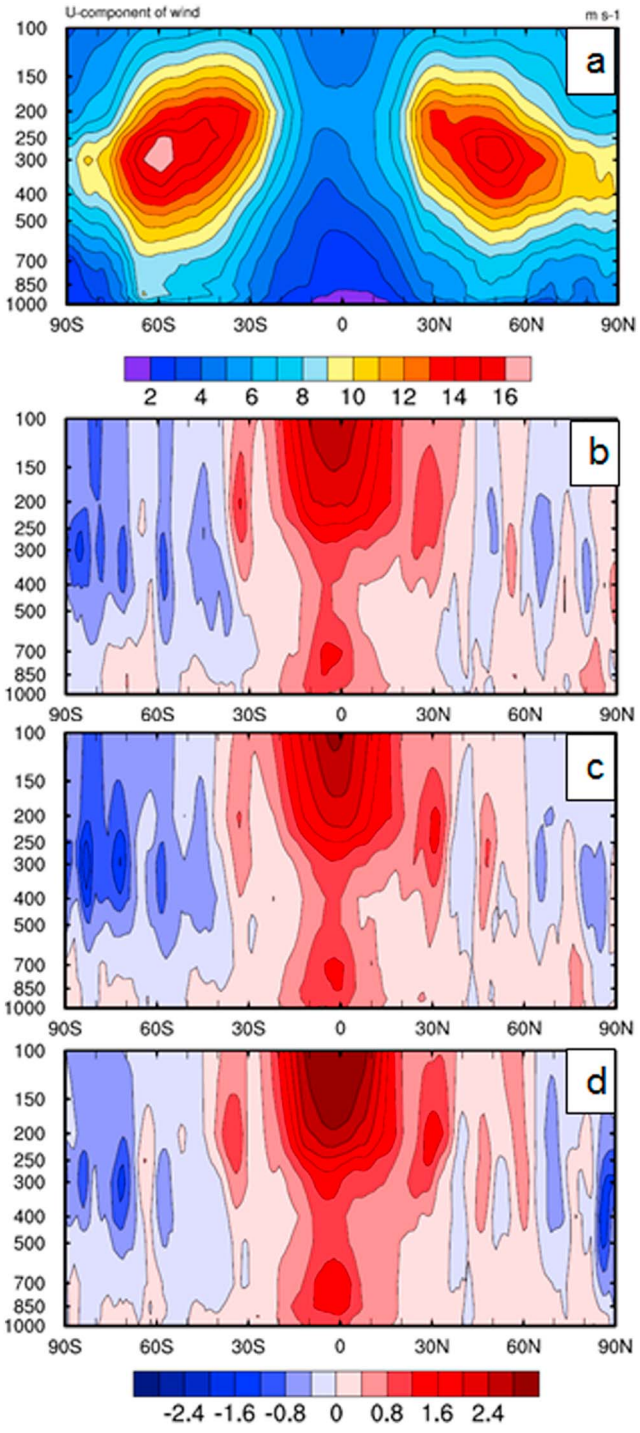

Figure 7. Global meridional cross section of the zonal wind spread (m/s) at 360 forecast hours (15 days) for (a) CTL, (b) SPs minus CTL, (c) SPs + SST_bc minus CTL; and (d) SPs + SST_bc + SA_CV minus CTL. The result is calculated using six cases starting 1 March 2016 and initialized every 5 days.

3.1. Tropical Evaluation—Forecast Skill of MJO

As one of the sources of predictability for the subseasonal timescale (or extended range forecast), MJO prediction skill is always considered a key metric when evaluating the forecast capability of the operational models (Kim et al., 2014; Ling et al., 2015; Neena et al., 2014; Wang et al., 2015). In this study, we evaluated the Wheeler Hendon MJO skill (Lin et al., 2008; Wheeler & Hendon, 2004), which is defined as the bivariate anomaly correlation between the analysis and forecast of two principal component time series (Real-time Multivariate MJO—RMM1 and RMM2) from combined empirical orthogonal functions of the MJO components using outgoing long wave radiation (OLR) and zonal wind at 200- and 850-hPa, respectively, that is,

$$\mathrm{AC}(\tau)=\frac{\Sigma_{i=1}^{N}[a_1(t)f_1(t,\tau)+a_2(t)f_2(t,\tau)]}{\sqrt{\Sigma_{i=1}^{N}[a_1^2(t)+a_2^2(t)]}\sqrt{\Sigma_{i=1}^{N}[f_1^2(t,\tau)+f_2^2(t,\tau)]}}, \tag{1}$$

where $f_1(t,\tau)$ and $f_2(t,\tau)$ are the RMM1 and RMM2 of the forecast at lead day τ initialized at day t and $a_1(t)$ and $a_2(t)$ are the RMM1 and RMM2 of the analysis data corresponding to the forecast at day t.

To calculate the AC score, we first calculated the daily mean of the zonal wind and OLR using 7-point average (-3, -2, -1, 0, 1, 2, and 3 day(s)) centered at 00*Z* of the validation time. Since at the time of the evaluation we did not have the long term GEFS hindcast available, we used the climate data averaged between 1999 and 2011 from NCEP/NCAR reanalysis as a reference to obtain the zonal wind and OLR anomaly. To obtain the RMMs, we projected the Wheeler Hendon MJO structure (i.e., the first two principal components) of 850-hPa, 250-hPa zonal wind, and OLR anomaly to both forecast and analysis data (Gottschalck et al., 2010). The AC score is then calculated by correlating the RMM1 and RMM2 between forecast data at each lead day and the analysis at the corresponding validation time.

The MJO strength during the experimental period is shown in Figure 4. Over the 2-year period, there are several significant MJO events with the strongest occurring in April and July of 2015. The MJO skill of the four experiments and CFS v2 (Saha et al., 2014) during the period of 1 May 2014 to 26 May 2016 (Figure 5) indicates that using SPPT combined with SHUM and SKEB (SPs hereafter) outperformed the STTP (CTL) on MJO forecast skill by about 4 days (AC $\geq$ 50%), improving the MJO skill from 12.5 to 16.8 days. Since a difference occurs without changing the model physics or its external forcing, an improvement is mainly contributed by SPs. The decrease in forecast error and increase in spread of the upper level (250 hPa) and lower level (850 hPa) tropical zonal winds are providing a better representation of tropical forecast uncertainties. This improved representation of uncertainty enhances the overall performance of tropical circulation as illustrated in the day-to-day verification out to 35 days (Figure 6). Figure 6 also indicates that the errors are saturated around 25 days, which means that there is no value to use daily forecasts for longer lead times.

In addition to the uncertainty associated with the model dynamics and physics, as an uncoupled forecast system, error in the underlying SST forcing is a significant factor limiting the forecast skill (Wang et al., 2015). As previously discussed, the SST forcing in the operational GEFS uses the RTG analysis damped to the climatology; therefore, the day-to-day variability of the SST cannot be well represented. To more realistically

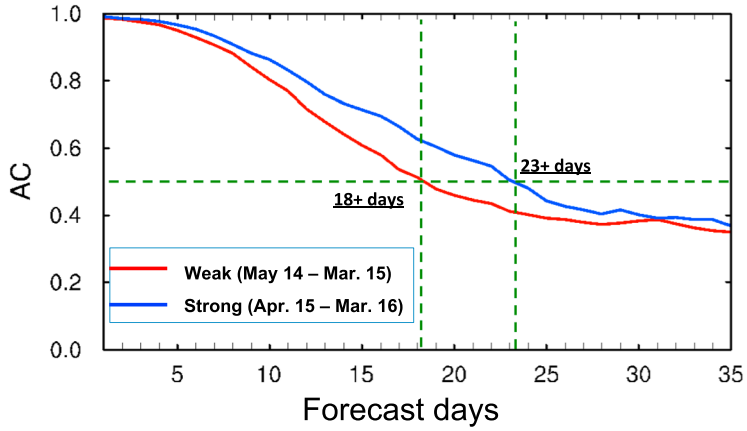

Figure 8. Madden-Julian Oscillation (MJO) RMM1 + RMM2 forecast skill calculated for a weak MJO period (May 2014 to March 2015) and a strong MJO period (April 2015–March 2016) for experiment SPs + SST_bc + SA_CV. AC = anomaly correlation.

represent the underlying SST forcing, an updated SST from the bias-corrected CFSv2 forecast is used to replace the climatological SST. When combined with SPs, the updated SST forcing (i.e., SPs + SST_bc) further improves the MJO skill of the extended GEFS by 1.7 days (Figure 5; Table 1).

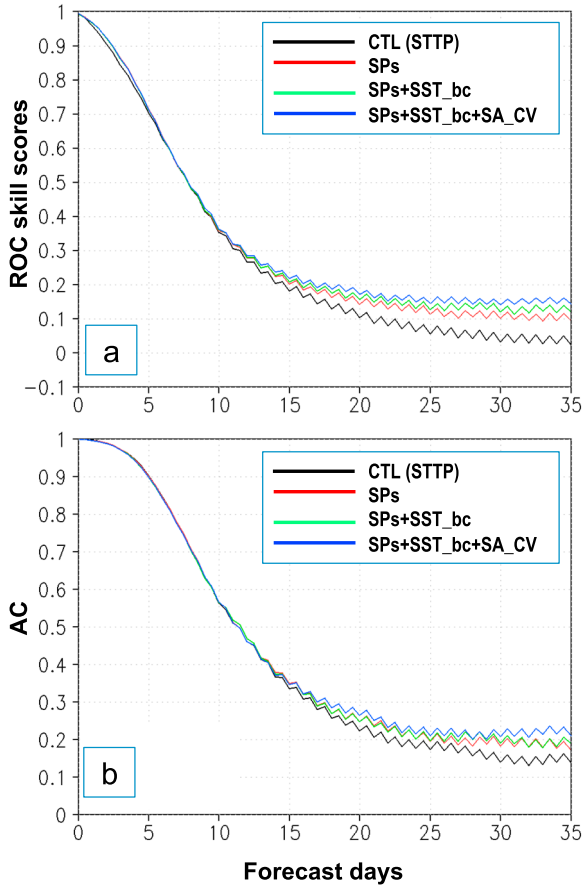

Figure 9. Daily relative operating characteristics (ROC; top) scores and anomaly correlation (AC; bottom) scores out to 35 days for Northern Hemisphere (20°N–80°N) 500-hPa geopotential heights for CTL (black), SPs (red), SPs + SST_bc (green), and SPs + SST_bc + SA_CV (blue).

The model physics have been found to impact the MJO skill (DeMott et al., 2014; Jiang et al., 2015, 2016; Kim & Maloney, 2017). To examine the effect of the model physics, especially the convective parameterization, we updated the convection package to the updated scale-aware Simplified Arakawa-Schubert convection scheme before applying the SPs in the GEFS 35-day experiment. This further improves the MJO skill for additional 3 days compared to the SPs + SST_bc (Figure 5; Table 1).

To further explore the reason for the improvement in each experiment, in addition to examining the root mean square error and spread in tropical zonal wind at higher and lower levels for SPs scheme, we also investigate the variation of the profile of zonal wind spread as a function of latitude for all experiments (Figure 7). Compared to the CTL (Figure 7a), the SP experiment shows an enhanced zonal wind spread over the tropics and most of NH midlatitudes. This enhancement is even stronger in SPs + SST_bc and strongest in SPs + SST_bc + SA_CV. Given the fact that the ensemble spread is comparable to the forecast error in SPs scheme (Figure 6) and other SP schemes (figure not shown), we hypothesize that the increase of the zonal wind spread over the tropics contributes to the improvement of the MJO forecast skill through an improvement of the zonal wind forecast skill. The two-tiered SST and new convection scheme provide additional improvement in MJO forecast skill. The addition of these two physical processes could be providing improvement of the large-scale circulation, which would lead to these improved scores. Since it is hard to separate the effect solely due to the two-tiered SST or the new convection scheme from the effect of the stochastic schemes due to our experimental design, we compare the combined effect as each component is added to help explain the reason of the improvement for the MJO forecast skill.

In contrast to operational CFS v2 that is fully coupled to ocean and sea ice, the MJO forecast skill has been calculated from the same period for 16 time-lagged ensemble members (Figure 5). GEFS has an overall better performance for short lead-time (out to 10 days) due to high resolution, latest model version, and others. CFS v2 shows a most useful MJO skill (≥0.5) out

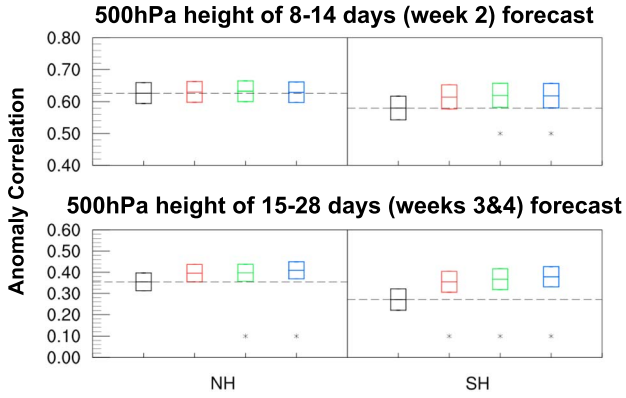

Figure 10. Ensemble-mean anomaly correlation for Northern Hemisphere (NH; 20°N–80°N) and Southern Hemisphere (SH; 20°S–80°S) 500-hPa geopotential height from May 2014 to May 2016 for CTL (black), SPs (red), SPs + SST_bc (green) and SPs + SST_bc + SA_CV (blue) for (a) days 8–14 (upper panel) and (b) days 15–28 (weeks 3 and 4, lower panel). Panels (c) and (d) are the same as (a) and (b) except for the Southern Hemisphere (20°S–80°S). Average scores of period are shown by straight lines in the vertical boxes matching the color of the experiment. The high/low box boundaries are representing the 95% confidence interval of the score differences between a given experiment and CTL, and "*" indicates a result is significantly better than CTL at 95% confidence level.

to 19 days, which is much better than GEFS CTL but not as good as GEFS with all science improvements (SPs + SST_bc + SA_CV).

In order to show the difference of the MJO forecast skill between strong and weak periods, we separated the 25-month experiment period into an averaged stronger MJO period that covers April 2015 to March 2016 and an averaged weaker period that covers May 2014 to March 2015 (Figure 8). As expected, the MJO forecast is more skillful during a stronger MJO amplitude period than a weaker period, with a difference of ~5 days. This is consistent for all experiments (not shown; aggregate numbers provided in Table 1). But the different skill should be attributed to the difference in the MJO propagation (i.e., phase speed prediction).

Table 1 summarizes the MJO forecast skill for the 2-year average and the comparison between the strong and weak MJO period. Overall, the SPs + SST_bc + SA_CV is the best configuration, with 22 days of MJO skill. The second best is the SPs + SST_bc. The largest contribution to the improvement is from the new stochastic scheme, which largely reduced the error over tropical circulation. The convection scheme and external boundary forcing contribute to additional improvement, but since we also use stochastic physics in those experiments for the evaluation, the effect of the convection scheme and external forcing can only be considered in combination with stochastic physics.

3.2. Extratropical Evaluation—Forecast Skill of 500-hPa Geopotential Height

Although the MJO forecast skill is considered to be one of the key metrics to evaluate the performance of a forecast system on the subseasonal timescale, the ultimate goal of these experiments is to examine the impact on extratropical forecast skill. As such, it is important to determine if the improvements contributed by SPs over the tropics extend into the extratropics. SPs shows substantial improvement according to the relative operating characteristics area every 12-hr lead time for 500-hPa geopotential height over the Northern Hemisphere (Figure 9a). The forecast skill is enhanced further for SPs + SST_bc and the SPs + SST_bc-SA_CV, with SPs + SST_bc-SA_CV having the highest skill, especially at longer lead times.

The AC of 500-hPa geopotential height is a commonly used measure to estimate pattern similarities of the large-scale circulation and is used in this study to quantify potential skill over the extratropics (Wilks, 2011). The AC of 500-hPa geopotential height for SPs, SPs + SST_bc, and SPs + SST_bc + SA_CV shows improvement compared to the CTL, with SPs + SST_bc + SA_CV having the highest AC (Figure 9b). In addition to examining the Northern Hemisphere 500-hPa geopotential height as a function of lead time, it is also helpful to examine AC as a function of verification date so that the variations can be detected with respect to the date for a given lead time (Figure 10). The 8–14-day lead time mean is used to evaluate week 2, and the 15–28-day lead time mean is used to evaluate weeks 3 and 4.

Over the Northern Hemisphere statistics (Figure 10), for the 500-hPa height forecast, there is 15% improvement compared to the CTL when using the combined SPs, CFS v2 SST, and scale-aware convection schemes (SPs + SST_bc + SA_CV) for the weeks 3 and 4 time period. The improvement over the Southern Hemisphere (SH) is about 40% from this configuration. Although the improvement for week 2 forecasts over the two hemispheres is not as large as weeks 3 and 4, we do see a positive improvement using the three different configurations (SPs; SPs + SST_bc; SPs + SST + SA_CV) compared to the CTL version.

There is a small (0.3%) improvement in skill in the SPs + SST_bc + SA_CV experiment compared to the CTL experiment during week 2 over the 2-year period (Figure 10). However, during weeks 3 and 4, when SPs replace the operational STTP, there is a 3.1% improvement. There is an additional 0.2% improvement when CFS v2 SSTs are added and an additional 0.9% improvement when the scale-aware convection scheme is added (Figure 10). SPs + SST_bc and SPs + SST_bc + SA_CV show statistically significant improvement relative to CTL during weeks 3 and 4. In the SH, the overall skill is slightly less than the NH, especially for weeks 3 and 4. However, the improvement in skill over the SH is much larger than for the NH (Figure 10). There is a 3–4% improvement for week 2 and a 40% (or 0.1 in AC score) for weeks 3 and 4 over the SH.

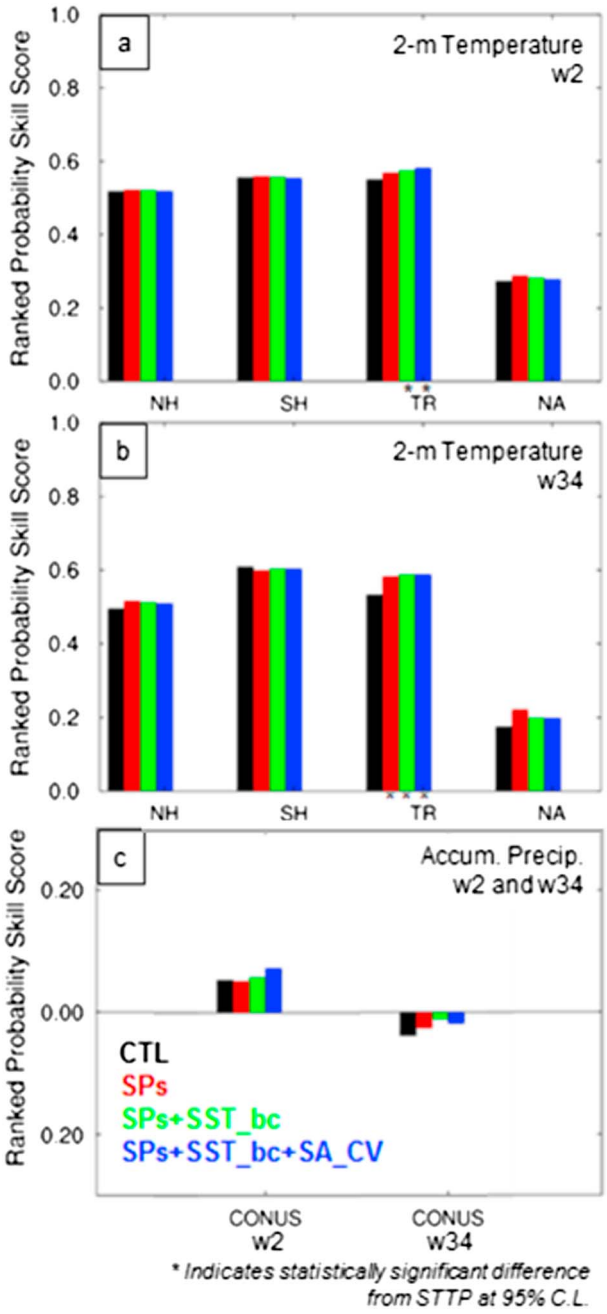

Figure 11. Ranked probability skill scores for CTL (black), SPs (red), SPs + SST_bc (green), and SPs + SST_bc + SA_CV (blue) during the period of May 2014 to May 2016 for (a) land-only 2-m temperature week 2 (days 8–14) average; (b) land-only 2-m temperature weeks 3 and 4 (days 15–28) average; and (c) Contiguous United States (CONUS)-only accumulated precipitation week 2 (days 8–14; left) and weeks 3 and 4 (days 15–28; right) average.

There is a similar pattern of improvement over the SH, but for both week 2 and weeks 3 and 4. The addition of SPs contributes the most to the forecast improvement (4.2%) during week 2 as well as the most (5.3%) during weeks 3 and 4 compared to the additional contributions of CFS v2 SSTs and scale-aware convection scheme (Figure 10). These results imply that the use of SPs improves the forecast skill the most over the extratropical region, especially during weeks 3 and 4, but since the SST_bc and SA_CV were only tested in combination with SPs and the improvements may not combine linearly, it is possible either would have as large an improvement as SPs if added alone.

3.3. Evaluation of 2-M Temperature and Precipitation

The global land-only 2-m temperature Ranked Probabilistic Skill Scores (RPSS; Toth et al., 2003; Melhauser et al., 2016) were calculated for the NH, SH, tropics, and Northern America (NA) over the experiment period for two lead-time averages (week 2 average [Figure 11a] and weeks 3 and 4 average [Figure 11b]). The tropics and SH have the highest RPSS for both week 2 and weeks 3 and 4 average lead times with NA having the lowest. Comparing week 2 with weeks 3 and 4, the RPSS remains similar for the tropics and SH for both shorter and longer lead times, but the NH and NA show a ~0.05–0.1 reduction in skill. Except for the SH, stochastic physics improves the results compared with CTL, although the only statistically significant (95% confidence level) improvement in week 2 is found in the tropics for SPs + SST_bc and SPs + SST_bc + SA_CV. In weeks 3 and 4, all experiments show statistically significant improvement in the tropics. That improvement is strongest in the tropics is not a surprise since all three configuration changes, in some way, improve tropical forecasts either by increasing ensemble spread, introducing additional SST information, or improving convection parameterization, a dominant process in the tropics. Although NA shows improvement for all experiments, none of these are statistically significant at the 95% confidence level for either lead-time averages, but SPs skill is improved for weeks 3 and 4 at the 90% confidence level.

The CONUS precipitation forecasts for week 2 and weeks 3 and 4 averages have been verified against the Climatological Calibrated Precipitation Analysis (Hou et al., 2014). There are positive RPSS for the week 2 average (Figure 11c; left) and negative RPSS for the weeks 3 and 4 average (Figure 11c; right). The lower (or negative) skill for accumulated precipitation compared with 2-m temperatures is no surprise, due to lower predictability for extended range precipitation forecasts and imperfections in the model physical parameterizations causing the model forecast to be biased. Adding SPs has little impact on week 2 RPSS, but SST_bc adds a small improvement, and adding SA_CV generates further improvement. No experiment provides skill for CONUS precipitation in weeks 3 and 4.

4. Impact of Ensemble Size on Extended Range Forecasts

Evaluations have been made for the 21-member GEFS based on the three different configurations (experiments) and a control experiment. It has been found that the experiment SPs + SST_bc + SA_CV generally shows the highest skill for many forecast variables and measures. Therefore, this experiment has been selected as the configuration for the National Multimodel Ensemble and Subseasonal Experiment (SubX) projects in real time (once per week on Wednesday 0000 UTC). The configuration of this final experiment is also used for an 18-year GEFS reforecast (1999–2016), which will be used for forecast calibration.

Due to limited computing resources, an optimum GEFS configuration (including the reforecasts) in terms of resolution and ensemble size (Ma et al., 2012), years of reforecast, and reforecast frequency need to be

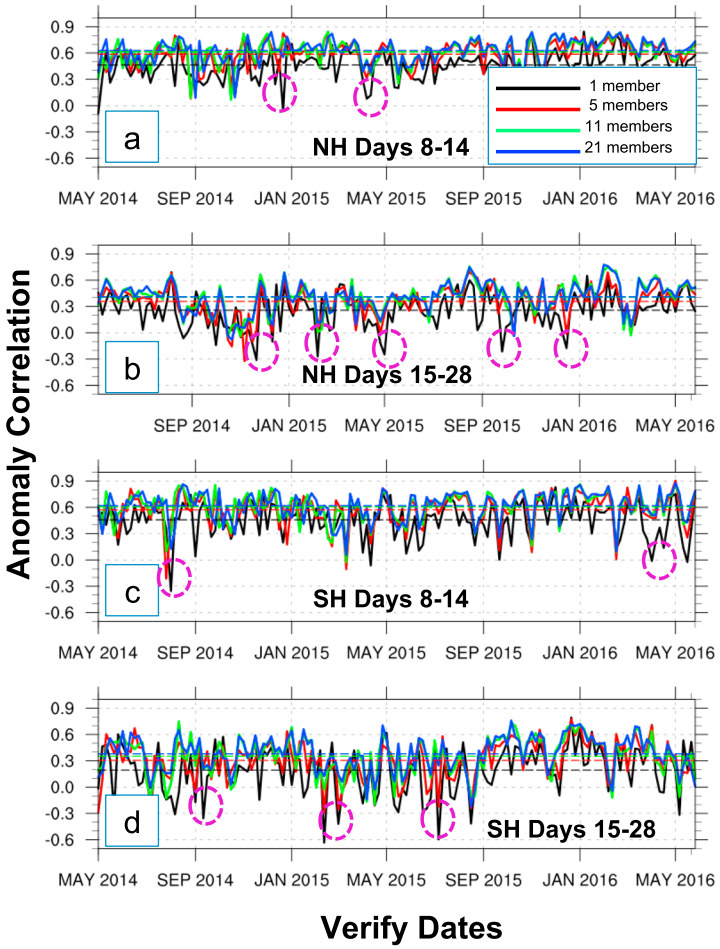

Figure 12. Ensemble-mean anomaly correlation time series for Northern Hemisphere (NH; 20°N–80°N) 500-hPa geopotential height from May 2014 to May 2016 for 1 member (black), 5 members (red), 11 members (green), and 21 members (blue) for (a) days 8–14 and (b) days 15–28 (weeks 3 and 4). Panels (c) and (d) are the same as (a) and (b) except for the Southern Hemisphere (SH; 20°S–80°S). Average scores are shown by straight dashed lines matching the color of different member sizes.

determined. In this section, an impact of ensemble size will be discussed for week 2 and weeks 3 and 4 average forecast skills in order to provide valuable information for reforecast configuration.

One-member (ensemble control; unperturbed member), 5-member (1 ensemble control and 4 perturbation members), 11-member (1 ensemble control and 10 perturbation members), and 21-member experiments are compared to test the impact of ensemble size. For the perturbation members of the 5-member and 11-member experiments, 4 perturbed members and 10 perturbed members are randomly selected from a pool of 20 perturbed members, respectively. Unlike deterministic forecasts, an ensemble mean (5 members, 11 members, and 21 members) will eliminate many "drop-out" cases (e.g., highlighted features in Figure 12). During both week 2 and weeks 3 and 4, the 21-member experiment has the greatest AC scores compared to experiments with reduced ensemble sizes (and deterministic as well) for the NH and SH 500-hPa geopotential height and the tropical zonal winds of 850- and 250-hPa (Figure 13). From Figure 13, we find that the 11-member and 21-member ensembles agree relatively well, thus indicating that an 11-member ensemble is satisfactory in reproducing expected features of a 21-member ensemble. Overall, an ensemble system remains more skillful than a deterministic model for the SPs + SST_bc + SA_CV configuration. The significance of the degradation due to decreased ensemble size was also determined (Figure 13). In the tropics during week 2, there is significant degradation in AC scores for the five-member and control-only experiment relative to the 21-member experiment for both the upper- and lower-level zonal wind (Figure 13). During weeks 3 and 4, however, the degradation is not significant for the five-member experiment for the upper level zonal

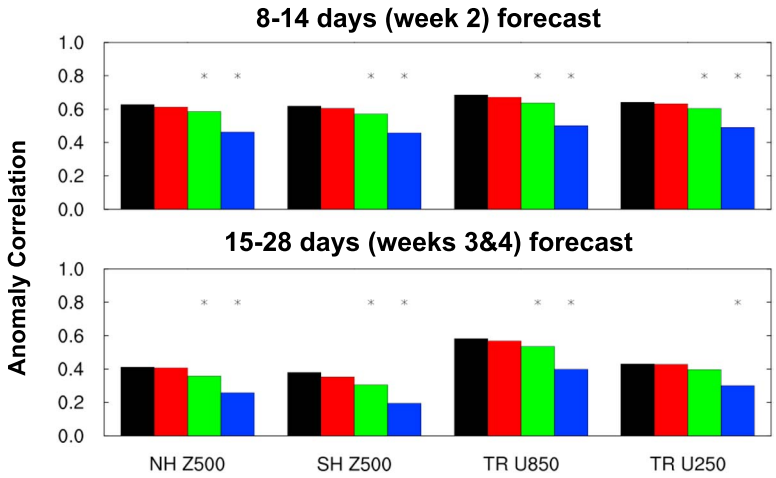

Figure 13. Ensemble-mean anomaly correlation for Northern Hemisphere (NH; 20°N–80°N) and Southern Hemisphere (SH; 20°S–80°S) height; tropical (TR; 20°N–20°S) 850- and 250-hPa zonal winds from May 2014 to May 2016 for 21 members (black), 11 members (red), 5 members (green), and 1 member (blue) for (a) days 8–14 (week 2; top panel) and (b) days 15–28 (weeks 3 and 4; bottom panel). The "*" represents the results that are significantly degraded from the 21-member ensemble experiment at the 95% confidence level.

wind. In the extratropics during week 2 and weeks 3 and 4, the AC score of the five-member experiment is significantly degraded for NH. All AC scores for 11 members are lower than for 21 members, but the difference is not significant.

5. Summary and Discussion

The three experiments, which emphasize three different physical processes (i.e., tropical forecast uncertainties, the ocean forcing on the atmosphere, and tropical deep convection) to improve forecast skill on the subseasonal timescale, have been evaluated and compared with the control version forecast. The purpose of this experiment design was to determine the benefit to the weeks 3 and 4 forecast skill through the combination of all three areas based on some preliminary investigation. Due to the limitation of computational resources that are required for a comprehensive investigation, we did not perform a comprehensive experiment to focus on each physical process independently.

Based on the analysis of MJO forecast skill, the experiment SPs + SST_bc + SA_CV shows ~22 days skillful forecast (defined by AC scores greater than 0.5), thus is considered to be the best configuration. A large gain in MJO forecast skill is found after introducing new stochastic physics perturbation schemes. Updating the underlying SST and a new convection scheme added further positive impact to the MJO forecast skill. As such, we speculate that the improvement of the representation of the forecast uncertainties over the tropics through introducing new stochastic physics perturbation schemes leads to the biggest gain in MJO skill. The improvement of the MJO skill in SPs + SST_bc (SPs + SST_bc + SA_CV) should be considered as a combined effect of the SST and SPs (SST, new convection scheme, and SPs). We cannot make any conclusions in terms of the relative importance of the SPs versus SST and convection scheme due to a lack of independent experiments.

In the extratropics, large-scale AC of 500-hPa geopotential heights for SPs + SST_bc + SA_CV indicates that AC scores for NH weeks 3 and 4 forecast improved significantly (by an additional 5.4%) from CTL (35.5%) with no significant difference for week 2 forecast.

In terms of 2-m temperature and accumulated precipitation, we have seen some improvement for NA and CONUS for week 2 and weeks 3 and 4 in all experiments versus CTL, except SP precipitation in week 2, but the improvement is marginal compared to the improvement for upper atmosphere variables. A future study will consider analyzing each component independently to assess the relative impacts. Over the subseasonal timescale, systematic model errors in both 2-m temperature and accumulated precipitation can dominate the forecast error. The results may indicate that forecast skill could be improved if we have sufficient model hindcast (or reforecast) to calibrate subseasonal forecasts.

In additional to the real-time run, another big task for the extended GEFS is the long-term reforecast for the calibration of the real-time forecast. For the reforecast, the ensemble size is an important part of configurations that should be carefully considered. We performed some tests of the dependence of the forecast skill score on the ensemble size and found that the skills degraded significantly for both week 2 and weeks 3 and 4 forecast when decreasing the ensemble size from 21 members to 5 members; the difference from 21 members to 11 members is relatively small. This indicates that 11 members are sufficient for the GEFS 18-year reforecast. Our intention is to minimize the use of computational resources while still providing a high-quality and reliable reforecast.

We investigated the capability of extended GEFS to potentially improve subseasonal prediction through advancing three physical processes for the operational system. Because we focus on the subseasonal time, the impact of the initial condition uncertainty and the associated initial perturbations are gradually reduced (Zhu, 2005). Having a better representation of the forecast uncertainty, mainly over the tropics, improves the forecast skill of tropical variables including the MJO. Updating the underlying SST and convection scheme leads to further improvement of the tropical forecast skill. While we focus on the subseasonal forecast skill using an uncoupled forecast system in this study, a fully coupled system may further improve forecast skills. The purpose of this study is to help shed light on some key components providing subseasonal forecast skill. As an intermediate solution before a fully coupled system is available, a suitable forecast system that can provide a significant improvement in forecast skill compared to the current operational configuration should be considered. As seen in this study, by updating the three physical processes in the forecast system, we obtain larger improvement in the upper atmosphere of the tropics and extratropics compared to extratropical surface temperature and precipitation. More investigation in the future is required to explore the impact of sea ice on high-latitude forecast skill, the tele-connection of tropical and extratropical pattern to improve NA's hazard (and/or extreme) forecasts, and in tropics, examine the potential predictability of tropical cyclones at this timescale, including genesis.

Acknowledgments

The authors would like to thank Xingren Wu, Wanqiu Wang, Jongil Han, Xu Li, and Ruiyu Sun at EMC (and Climate Prediction Center) for valuable discussion pertaining to the design of our experiments and, additionally, all of the help from the EMC ensemble team members. The authors would also like to thank Shrinivas Moorthi and Edward Strobach for valuable EMC internal reviews and three anonymous reviewers. This study is partially supported through National Weather Service OSTI and NOAA's Climate Program Office (CPO)'s Modeling, Analysis, Predictions, and Projections (MMAP) program. The forecast data could be accessed through http://iridl.ldeo.columbia.edu/SOURCES/.Models/.SubX/.EMC/.GEFS/.

References

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

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

DeMott, C. A., Stan, C., Randall, D. A., & Branson, M. D. (2014). Intraseasonal variability in coupled GCMs: The roles of ocean feedbacks and model physics. *Journal of Climate*, *27*(13), 4970–4995. https://doi.org/10.1175/JCLI-D-13-00760.1

Gemmill, W., Katz, B., & Li, X. (2007). The daily real-time, global sea surface temperature-high-resolution analysis: RTG_SST_HR. NCEP Office Note, #260 1–39. Retrieved from http://polar.ncep.noaa.gov/mmab/papers/tn260/MMAB260.pdf

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 the American Meteorological Society*, *91*(9), 1247–1258. https://doi.org/10.1175/2010BAMS2816.1

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 and Forecasting*, *32*(5), 2005–2017. https://doi.org/10.1175/WAF-D-17-0046.1

Hou, D., charles, M., Luo, Y., Toth, Z., Zhu, Y., Krzysztofowicz, R., et al. (2014). Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of stage IV toward CPC gauge-based analysis. *Journal of Hydrometeorology*, *15*(6), 2542–2557. https://doi.org/10.1175/JHM-D-11-0140.1

Hou, D., Toth, Z., Zhu, Y., & Yang, W. (2008). Evaluation of the impact of the stochastic perturbation schemes on global ensemble forecast. Proc. 19th Conf. on Probability and Statistics, New Orleans, LA, Amer. Meteor. Soc. Retrieved from https://ams. confex.com/ams/88Annual/webprogram/Paper134165.html

Jiang, X., Waliser, D. E., Xavier, P. K., Petch, J., Klingaman, N. P., Woolnough, S. J., et al. (2015). Vertical structure and physical processes of the Madden-Julian Oscillation: Exploring key model physics in climate simulations. *Journal of Geophysical Research: Atmospheres*, *120*, 4718–4748. https://doi.org/10.1002/2014JD022375

Jiang, X., Zhao, M., Maloney, E. D., & Waliser, D. E. (2016). Convective moisture adjustment time scale as a key factor in regulating model amplitude of the Madden-Julian Oscillation. *Geophysical Research Letters*, *43*, 10,412–10,419. https://doi.org/10.1002/2016GL070898

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., et al. (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

Kim, D., & Maloney, E. (2017). Simulation of the Madden-Julian Oscillation using general circulation models. In *The global monsoon system* (3rd ed., pp. 119–130). Singapore: World Scientific.

Kim, H.-M., Webster, P. J., Toma, V. E., & Kim, D. (2014). Predictability and prediction skill of the MJO in two operational forecasting systems. *Journal of Climate*, *27*(14), 5364–5378. https://doi.org/10.1175/JCLI-D-13-00480.1

Li, H., Luo, L., Wood, E. F., & Schaake, J. (2009). The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting. *Journal of Geophysical Research*, *114*, D04114. https://doi.org/10.1029/2008JD010969

Li, W., Yu, R., Liu, H., & Yu, Y. (2001). Impacts of diurnal cycle of SST on the intraseasonal variation of surface heat flux over the western Pacific warm pool. *Advances in Atmospheric Sciences*, *18*(5), 793–806.

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, T., Xu, M., Liang, X.-Z., Wang, J. X. L., & Noh, Y. (2015). A multilevel ocean mixed layer model resolving the diurnal cycle: Development and validation. *Journal of Advances in Modeling Earth Systems*, *7*, 1680–1692. https://doi.org/10.1002/2015MS000476

Liu, X., Wu, T., Yang, S., Li, T., Jie, W., Zhang, L., et al. (2016). MJO prediction using the sub-seasonal to seasonal forecast model of Beijing Climate Center. *Climate Dynamics*, *48*(9-10), 3283–3307. https://doi.org/10.1007/s00382-016-3264-7

Ma, J., Zhu, Y., Wobus, D., & Wang, P. (2012). An effective configuration of ensemble size and horizontal resolution for the NCEP GEFS. *Advances in Atmospheric Sciences*, *29*(4), 782–794. https://doi.org/10.1007/s00376-012-1249-y

Melhauser, C., Li, W., Zhu, Y., Zhou, X., Pena, M., & Hou, D. (2016). Exploring the impact of SST on the extended range NCEP Global Ensemble Forecast System. *STI Climate Bulletin*, 30–34. Retrieved from http://www.nws.noaa.gov/ost/climate/STIP/41cdpw_digest.htm

National Academies of Sciences, Engineering, and Medicine (2016). *Next generation Earth system prediction: Strategies for subseasonal to seasonal forecasts*. Washington, DC: The National Academies Press. https://doi.org/10.17226/21873

Neena, J. M., Lee, J. Y., Waliser, D., Wang, B., & Jiang, X. (2014). Predictability of the Madden-Julian Oscillation in the intraseasonal variability hindcast experiment (ISVHE). *Journal of Climate*, *27*(12), 4531–4543. https://doi.org/10.1175/JCLI-D-13-00624.1

NOAA National Weather Service (2014). Next generation of global prediction system (NGGPS). The documentations are available through online: https://www.weather.gov/sti/stimodeling_nggps

Palmer, T. N., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G., et al. (2009). Stochastics parametrization and model uncertainty (Tech. Rep. ECMWF RD Tech. Memo. 598, 42 pp). Retrieved from http://www.ecmwf.int/publications/

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

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

Sela, J. (1980). Spectral modeling at the National Meteorological Center. *Monthly Weather Review*, *108*(9), 1279–1292. https://doi.org/10.1175/1520-0493(1980)108<1279:SMATNM>2.0.CO;2

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

Shutts, G., & Palmer, T. N. (2004). The use of high-resolution numerical simulations of tropical circulation to calibrate stochastic physics schemes. Proc. ECMWF/CLIVAR simulation and prediction of intra-seasonal variability with emphasis on the MJO, Reading, United Kingdom, European Centre for Medium-Range Weather Forecasts (pp. 83–102).

Takaya, Y., Vitart, F., Balsamo, G., Balmaseda, M., Leutbecher, M., & Molteni, F. (2010). Implementation of an ocean mixed layer model in IFS, Proc. ECMWF Technical Memorandum No. 622 (pp. 34).

Tompkins, A. M., & Berner, J. (2008). A stochastic convective approach to account for model uncertainty due to unresolved humidity variability. *Journal of Geophysical Research*, *113*, D18101. https://doi.org/10.1029/2007JD009284

Toth, Z., & Kalnay, E. (1993). Ensemble forecasting at NMC: The generation of perturbations. *Bulletin of the American Meteorological Society*, *74*(12), 2317–2330. https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2

Toth, Z., & Kalnay, E. (1997). Ensemble forecasting at NCEP and the breeding method. *Monthly Weather Review*, *127*, 3297–3318.

Toth, Z., Talagrand, O., Candille, G., & Zhu, Y. (2003). Probability and ensemble forecasts. In I. T. Joliffe & D. B. Stephenson (Eds.), *Forecast verification–A practitioner's guide in atmospheric science* (pp. 137–163). New York: John Wiley & Sons.

Vitart, F. (2014). Evolution of ECMWF sub-seasonal forecast skill scores. *Quarterly Journal of the Royal Meteorological Society*, *140*(683), 1889–1899. https://doi.org/10.1002/qj.2256

Wang, W., Kumar, A., Fu, J. X., & Hung, M.-P. (2015). What is the role of the sea surface temperature uncertainty in the prediction of tropical convection associated with the MJO? *Monthly Weather Review*, *143*(8), 3156–3175. https://doi.org/10.1175/MWR-D-14-00385.1

Wheeler, M., & Hendon, H. H. (2004). An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Monthly Weather Review*, *132*(8), 1917–1932. https://doi.org/10.1175/1520-0493(2004)132<1917:AARMMI>2.0.CO;2

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (p. 676). Cambridge, MA: Academic Press.

Zhou, X., Zhu, Y., Hou, D., Luo, Y., Peng, J., & Wobus, D. (2017). The NCEP Global Ensemble Forecast System with the EnKF initialization. *Weather and Forecasting*, *32*(5), 1989–2004. https://doi.org/10.1175/WAF-D-17-0023.1

Zhou, X. Y., Zhu, D. H., & Kleist, D. (2016). Comparison of the ensemble transform and the ensemble Kalman filter in the NCEP Global Ensemble Forecast System. *Weather and Forecasting*, *31*, 2058–2074.

Zhu, Y. (2005). Ensemble forecast: A new approach to uncertainty and predictability. *Advances in Atmospheric Sciences*, *22*(6), 781–788. https://doi.org/10.1007/BF02918678

Zhu, Y., Zhou, X., Pena, M., Li, W., Melhauser, C., & Hou, D. (2017). Impact of sea surface temperature forcing on weeks 3 & 4 forecast skill in the NCEP global ensemble forecasting system. *Weather and Forecasting*, *32*(6), 2159–2174. https://doi.org/10.1175/WAF-D-17-0093.1