Sensitivities of Subseasonal Unified Forecast System Simulations to Changes in Parameterizations of Convection, Cloud Microphysics, and Planetary Boundary Layer

BENJAMIN W. GREEN,[a,b] ERIC SINSKY,[c,d] SHAN SUN,[b] VIJAY TALLAPRAGADA,[d] AND GEORG A. GRELL[b]

[a] *Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado*
[b] *NOAA/OAR/Global Systems Laboratory, Boulder, Colorado*
[c] *I. M. System Group, Inc., College Park, Maryland*
[d] *NOAA/NWS/NCEP/Environmental Modeling Center, College Park, Maryland*

(Manuscript received 11 December 2022, in final form 23 June 2023, accepted 23 June 2023)

ABSTRACT: NOAA has been developing a fully coupled Earth system model under the Unified Forecast System framework that will be responsible for global (ensemble) predictions at lead times of 0–35 days. The development has involved several prototype runs consisting of bimonthly initializations over a 7-yr period for a total of 168 cases. This study leverages these existing (baseline) prototypes to isolate the impact of substituting (one-at-a-time) parameterizations for convection, microphysics, and planetary boundary layer on 35-day forecasts. Through these physics sensitivity experiments, it is found that no particular configuration of the subseasonal-length coupled model is uniformly better or worse, based on several metrics including mean-state biases and skill scores for the Madden–Julian oscillation, precipitation, and 2-m temperature. Importantly, the spatial patterns of many "first-order" biases (e.g., impact of convection on precipitation) are remarkably similar between the end of the first week and weeks 3–4, indicating that some subseasonal biases may be mitigated through tuning at shorter time scales. This result, while shown for the first time in the context of subseasonal prediction with different physics schemes, is consistent with findings in climate models that some mean-state biases evident in multiyear averages can manifest in only a few days. An additional convective parameterization test using a different baseline shows that attempting to generalize results between or within modeling systems may be misguided. The limitations of generalizing results when testing physics schemes are most acute in modeling systems that undergo rapid, intense development from myriad contributors—as is the case in (quasi) operational environments.

KEYWORDS: Hindcasts; Coupled models; Model errors; Model evaluation/performance; Parameterization

1. Introduction

Numerical prediction of the future state of the atmosphere (as part of the larger Earth system) theoretically can range in temporal scale from minutes (e.g., Stensrud et al. 2009; Benjamin et al. 2016) to millennia (e.g., Sepulchre et al. 2020). In current practice, atmospheric prediction falls into one of two distinct regimes. The first regime is often termed "numerical weather prediction" (NWP), for time scales less than ~2 weeks, which corresponds to the theoretical deterministic predictability limit in the midlatitudes (Lorenz 1969), beyond which the memory of the atmospheric initial conditions is lost. The second regime is often termed "climate modeling" and covers time scales of no less than ~3 months. In this regime, the mean state of the atmosphere is regulated by fluctuations in slowly evolving external forcings including sea surface temperature (SST) and net radiation at the top of the atmosphere.

Atmospheric prediction at the subseasonal time scale falls in the temporal gap between NWP and climate modeling. Efforts to close this gap by researching and providing subseasonal forecasts have existed for over a decade (e.g., Brunet et al. 2010; Kirtman et al. 2014; Vitart et al. 2017; Sun et al. 2018a,b; Pegion et al. 2019). Advancements in subseasonal prediction have focused on phenomena such as the Madden–Julian oscillation (e.g., Rodney et al. 2013) and various global teleconnections that link regions of high and low predictability (e.g., Stan et al. 2017; Yamagami and Matsueda 2020). The concept of "forecasts of opportunity" (Jones et al. 2011), which aims to identify situations in which subseasonal predictability may be enhanced relative to a long-term baseline, has also proved to be invaluable.

In the United States, the National Oceanic and Atmospheric Administration (NOAA) is the government agency tasked with providing official forecasts across a range of time scales to a variety of stakeholders. At the subseasonal time scale (weeks 3–4), NOAA provides operational 2-m temperature forecasts and experimental precipitation forecasts to users in a wide range of sectors including but not limited to agriculture, wildfire, environmental resource management, public health, and energy. Since 2011, NOAA has been running the Climate Forecast System, version 2 (CFSv2; Saha et al. 2014), coupled Earth system model in operations to provide subseasonal to seasonal forecasts. As can be inferred from its name, CFSv2 was designed as a "climate model" rather than an NWP model (cf. two paragraphs previous); moreover, it has remained essentially unchanged during its operational lifetime—in contrast with the NWP-based Global Forecast System (GFS), which provides forecasts out to 16 days and has been updated at least six times since 2011.

Sinsky's current affiliation: Lynker at NOAA/NWS/NCEP/Environmental Modeling Center, College Park, Maryland.

Corresponding author: Benjamin W. Green, ben.green@noaa.gov

DOI: 10.1175/MWR-D-22-0338.1

In recent years, NOAA has used the Unified Forecast System (UFS) framework to consolidate its operational modeling products. Eventually, the Global Ensemble Forecast System (GEFS) will be a fully coupled (atmosphere, ocean, sea ice, wave, chemistry, and land surface) Earth system model providing global ensemble forecasts for lead days 0–35. A major step toward this goal was the September 2020 release of GEFSv12 (Zhou et al. 2022). This release provides 35-day ensemble forecasts initialized once per day at 0000 UTC but instead of a full ocean model, GEFSv12 uses an SST that is taken from CFSv2: specifically, from the raw CFSv2 SST forecast the historical model bias (as functions of initialization date, lead time, and location) is removed in order to obtain a predictive, bias-corrected CFSv2 SST field for use in GEFSv12. Ongoing work at NOAA, led by the Environmental Modeling Center (EMC), has produced eight different iterations (sets) of prototype coupled subseasonal hindcasts; each set of hindcasts contains the same 168 initial condition dates. As explained below, this coupled prototype framework provides the starting point for the physics sensitivity tests that are the focus of the present study.

One of the most consequential components of atmospheric models based on physical laws (as opposed to various empirical models) is the representation of processes not explicitly resolved on the model grid, that is, subgrid-scale physical parameterization schemes. A nonexhaustive list of processes represented by these physics schemes includes convection (except for the highest resolution models, e.g., Miyamoto et al. 2013), cloud microphysics, planetary boundary layer, surface fluxes, and radiative transfer. Subgrid-scale parameterizations require some set of assumptions to be made and thus are inherently imperfect; different assumptions and approaches yield different physics schemes. Parameterization schemes are often compatible with multiple atmospheric dynamical cores, and one dynamical core is often compatible with multiple parameterizations of the same process,[1] giving users and developers the flexibility to choose which physics schemes to use. The choice of physics scheme is known to have a tremendous impact on the state of the simulated atmosphere both in NWP [e.g., for planetary boundary layer (Hu et al. 2010); convective parameterization (Gómez et al. 2011); surface fluxes in a tropical cyclone (Green and Zhang 2013); cloud microphysics (McMillen and Steenburgh 2015)] and in long-range climate modeling [e.g., convective parameterization (Song and Zhang 2018); cloud microphysics (Lohmann et al. 2007; Hazra et al. 2017); a combination of boundary layer, shallow convection, and cloud macrophysics (Bogenschutz et al. 2013)]. The papers cited in the previous sentence represent a miniscule fraction of the peer-reviewed literature documenting physics sensitivity.

Despite the abundance of studies documenting the sensitivity of various atmospheric simulations to subgrid-scale physics at both the NWP and climate time scales, there is relatively little *controlled* physics sensitivity research[2] at the subseasonal time scale (Green et al. 2017; Zhu et al. 2018; Guimarães et al. 2020; Kaur et al. 2022). This is due to a combination of two factors. First, subseasonal modeling in general is a much newer research area as compared with the more than one-half of a century spent developing NWP and climate models. Second, subseasonal-length simulations require tremendous computational resources, which is a problem that becomes exacerbated when trying to build up sufficiently large sample sizes for robust results.[3] Due to a lack of controlled physics experiments at the subseasonal time scale, the temporal evolution of the impact of physics changes (i.e., differences at day 1 vs those at week 1 vs those at weeks 3–4) was poorly understood.

The aforementioned lack of understanding on the temporal evolution in the subseasonal (to seasonal) time scale is not limited just to physics sensitivity. At the NWP time scale, it is well-known that errors become saturated on the order of ~2 weeks (e.g., Dalcher and Kalnay 1987). Moreover, on the climate time scale, Ma et al. (2021) showed that model biases related to parameterizations of moist processes seen in multiyear climate runs develop as quickly as a few days. At the subseasonal–seasonal time scales, however, the annual cycle becomes a factor: the mean state of the Earth system is very different in the boreal winter (austral summer) than in the boreal summer (austral winter). Thus, there is a period of time centered on 6 months, but with somewhat unknown bounds at both ends, for which model errors may differ substantially from what is found in the first few days (or after a year). The present work provides an opportunity to examine whether the annual cycle begins to impact the model biases at the subseasonal time scale.

The purpose of this study is to document in a controlled setting the impacts of changing various physical parameterizations on subseasonal forecasts within the coupled UFS framework. Specifically, using the "Prototype 5" experiment as a baseline (Stefanova et al. 2022, hereinafter S22), three additional sets of experiments were conducted: one in which the convection scheme was changed; one in which the cloud microphysics scheme was changed; and one in which the planetary boundary layer scheme was changed. An additional test swapping out the convection scheme was performed but with "Prototype 7" as a baseline (S22), providing an opportunity

[1] As part of the transition to the UFS framework, NOAA has adopted the Common Community Physics Package to facilitate community-based atmospheric physics scheme development across a variety of dynamical models. More information can be found online (https://dtcenter.org/community-code/common-community-physics-package-ccpp).

[2] Here, controlled physics sensitivity research means that physics schemes are modified (or replaced) *within the same modeling system* such that differences between two sets of runs can be attributed to the changes in the physics. As with any Earth system model, nonlinear interactions must be considered when attempting to isolate the impacts of individual physical processes, and looking at aggregates of many cases is preferred to a single case study.

[3] In climate model physics tests, one way to collect a "large" sample is to conduct an extremely long simulation and treat each annual cycle as a quasi-independent realization (e.g., Jia et al. 2010).

to determine which conclusions may be baseline-specific versus those that may be more generalized.

The remainder of this paper is organized as follows. Section 2 details the data and experimental methodology, including descriptions of the overall model, the various physics schemes tested, and verification strategy. Section 3 presents results, first focusing on Prototype 5 and then comparing Prototype 5 with Prototype 7. A discussion and the conclusions are provided in section 4.

2. Data and experimental methods

a. *UFS description*

The UFS is an evolving system with numerous configurations; results presented in this research utilized two prototype versions of the "fully coupled" UFS: "Prototype 5" and "Prototype 7" (S22). For brevity, only the most essential information about these prototypes is described here; interested readers are referred to S22 for a complete description.

In Prototype 5 (hereinafter "P5"), UFS was configured as an Earth system model that coupled models of the atmosphere [Finite-Volume Cubed-Sphere Dynamical Core (FV3; Harris et al. 2021)], ocean [version 6 of the Modular Ocean Model (MOM6; Adcroft et al. 2019)], sea ice [version 6 of the Los Alamos Sea Ice Model (CICE6)], and a surface wave model (WAVEWATCH III; WAVEWATCH III Development Group 2016). MOM6 has 75 hybrid depth-isopycnal vertical coordinates and shares with CICE6 a 0.25° tripolar grid mesh; WAVEWATCH III uses a regular 0.5° grid.

In P5, FV3 had a horizontal resolution of ~25 km and 64 hybrid vertical layers. Atmospheric physics within FV3 utilized the Common Community Physics Package framework, with default parameterizations including the scale-aware Simplified Arakawa–Schubert (SAS) scheme for convection (Han and Pan 2011), hybrid Eddy-Diffusivity Mass-Flux scheme (hereinafter GFS-PBL; Han et al. 2016) for planetary boundary layer (PBL), the Geophysical Fluid Dynamics Laboratory (GFDL) cloud microphysics scheme (Zhou et al. 2019; Harris et al. 2020), the Rapid Radiative Transfer Model for General Circulation Models (RRTMG; Iacono et al. 2008) for both shortwave and longwave, and the Noah land surface model (Mitchell et al. 2005). The physics time step was 450 s. To account for the mismatch in grid meshes, variables in P5 were passed between component models by conservative grid interpolation followed by a nearest neighbor fill.

Atmospheric initial conditions in P5 were from the Climate Forecast System Reanalysis (CFSR; Saha et al. 2010). Ocean initial conditions were obtained from the Climate Prediction Center (CPC; S22). Sea ice initial conditions were derived from the CPC Sea Ice Initialization System (Collow et al. 2019; Liu et al. 2019). Wave initial conditions were created through a spin-up process forced by CFSR 10-m wind speeds and ice concentrations (S22).

Prototype 7 (hereinafter "P7") was run in the fall of 2021 (a full year after P5) and featured significant changes from P5. These changes include but are not limited to: an increase in FV3 vertical layers to 127 (model top raised from ~54 to ~80 km) and reduction in atmospheric physics time step to 300 s (note: P7 had the same horizontal resolution as P5); the implementation of a "fractional land mask" (S22) allowing for rigorous air–sea flux conservation; and updates to essentially all of the atmospheric physics (e.g., Han et al. 2021) including a change in land surface model from Noah to Noah-multiparameterization ("Noah-MP"; Niu et al. 2011; Yang et al. 2011). There were also updates to the ocean, sea ice, and wave models. Another major change that is particularly relevant for the results presented in this study is that atmospheric (and wave) initial conditions previously provided by CFSR for P5 were changed to the GEFSv12 Reanalysis (Hamill et al. 2022). It is important to note that for both P5 and P7, all simulations can be thought of as "cold starts" in that there was no native cycling data assimilation.

TABLE 1. Overview of the four experimental configurations used to test the impact of various physics schemes on subseasonal hindcasts within the P5 framework. More details are provided in the main text.

Expt	Convective scheme	PBL scheme	Microphysics scheme
CNTL	saSAS	GFS-PBL	GFDL
GF	GF	GFS-PBL	GFDL
MYNN	saSAS	MYNN	GFDL
THMP	saSAS	GFS-PBL	Thompson

As described in the previous section, evaluation of a coupled subseasonal modeling system requires a sufficiently large sample of cases initialized throughout the year and across multiple years. The sample size of cases is constrained by limited computational resources. Therefore, all coupled UFS subseasonal prototypes conducted by NOAA's EMC—as well as the additional physics sensitivity test experiments conducted for this research, described in section 2b—follow the same protocol: bimonthly initializations (on the 1st and 15th of every month) for a 7-yr period from 1 April 2011 to 15 March 2018, giving a total of 168 initial condition dates. Hindcasts (retrospective forecasts) from each initialization date are integrated out to 5 weeks (35 days).

b. *Sensitivity test setup*

The overall theme of this study is to investigate the sensitivity of subseasonal simulations (hindcasts) to changes in atmospheric physical parameterizations of (i) convection; (ii) cloud microphysics; and (iii) planetary boundary layer. This is primarily accomplished through three additional sets of experiments that use EMC's P5 as a common baseline but use an alternative parameterization for one of the three aforementioned subgrid-scale processes. Table 1 summarizes the four sets of P5-based experiments; "CNTL" is the control (i.e., baseline) P5 against which the other experiments are compared. More detailed descriptions of the physics schemes evaluated in the experiments are given below. It should be noted that all of the parameterizations tested are in a constant state of development and tuning, such that the results gleaned from these experiments may appear different as schemes are updated. To get a sense of the impact of changing baseline [i.e., accounting for the fact that model (physics) development is

always occurring], section 3b repeats the "CNTL" and "GF" experiments except using the framework of P7 instead of P5.

1) Convective parameterizations

The scale-aware SAS scheme was used as EMC's default convective parameterization scheme for both P5 and P7. More complete details of the philosophy behind this scheme can be found in Han and Pan (2011), Han et al. (2017, 2020, 2021). The alternative convective parameterization tested in this study is the Grell–Freitas (GF; Grell et al. 1991; Grell and Devenyi 2002; Grell and Freitas 2014; Freitas et al. 2021). Both SAS and GF have their origins in the same mass-flux convective parameterization originally implemented in the Rapid Update Cycle model and described by Grell (1993). A detailed description of the relationships of the two schemes and how they developed over the last few decades can be found in Lin et al. (2022).

The GF scheme is a more recent development with significant new additions and fundamental changes relative to previous versions [such as the Grell and Devenyi (2002) scheme]. It introduced scale-awareness, aerosol-awareness, and a very different way to derive normalized mass flux. GF envisioned that probability density functions should characterize the mean statistical properties of deep, congestus, and shallow convection and used a beta function to estimate each profile [see Freitas et al. (2021) for details]. Entrainment and detrainment rates are then derived from the normalized profiles, assuming statistically representative initial entrainment rates for each type. While many other modifiable parameters in GF and SAS are still used with the same names (e.g., autoconversion and cloud water detrainment constants), they have different values in the schemes. In addition, mixed-phase microphysics impacts were included in GF (Freitas et al. 2021). Variants of GF and SAS were tested in the frameworks of both P5 and P7.

2) Cloud microphysics parameterizations

The Geophysical Fluid Dynamics Laboratory (GFDL) microphysics scheme was used by EMC for P5 and P7. This scheme has its roots from one first described by Lin et al. (1983) and is described more fully in Zhou et al. (2019) and Harris et al. (2020). Over the past decade, much of the development of this single-moment six-category microphysics scheme has been focused on GFDL's global cloud-resolving models. The Thompson microphysics parameterization (e.g., Thompson et al. 2008) is single moment for snow and graupel and double moment for cloud ice, rain, and cloud water (Thompson and Eidhammer 2014). The Thompson scheme has been mainly developed for regional weather applications but has recently garnered interest in the global modeling community. In fact, independent of the results of this study, EMC elected to use Thompson microphysics for the baseline Prototype 8 of the coupled UFS. Number concentrations for microphysics species are diagnosed in single-moment schemes but predicted in double-moment schemes. An idealized study of a squall line (Morrison et al. 2009) found the double-moment scheme to produce higher precipitation rates than the single-moment scheme in the stratiform region but lower rates in the convective region. Conversely, a head-to-head comparison of Thompson and GFDL microphysics in FV3 (run in a regional configuration with 3-km horizontal grid spacing) found that the Thompson scheme had much *heavier* precipitation in convective storms than the GFDL scheme (Potvin et al. 2019). These conflicting results make it impossible to predict a priori the behavior of different microphysics schemes in global subseasonal-length simulations. It should be noted that when the Thompson scheme was implemented for testing within P5 (Table 1), two of the initial condition dates failed; further investigation revealed that P5 was somewhat prone to crashes in other parts of the model and that common efforts to circumvent crashes (such as reducing the model physics time step) proved unsuccessful. Thus, 166 cases of P5 with Thompson microphysics are compared with 168 cases of the baseline P5; this small difference in case size (~1.2% fewer cases for Thompson) should not substantially impact any of the results presented herein.

3) Planetary boundary layer parameterizations

The GFS PBL scheme ("GFS-PBL"; Han et al. 2016) was used as the default planetary boundary layer parameterization by EMC in both P5 and P7. The Mellor–Yamada–Nakanishi–Niino (MYNN) boundary layer scheme originated half a century ago (Mellor and Yamada 1974) and evolved with a primary focus on limited-area (as opposed to global) modeling applications (e.g., Nakanishi 2001; Nakanishi and Niino 2004, 2006, 2009; Olson et al. 2019). The main differences between GFS-PBL and MYNN are as follows. First, MYNN estimates its own subgrid-scale clouds, which are used to help regulate the turbulent mixing within cloudy environments, and these clouds are also used for coupling to the radiation scheme; in contrast, GFS-PBL does not produce its own clouds, instead relying on a stand-alone shallow-cumulus scheme that will also perform the nonlocal mixing in shallow-cumulus environments. Second, in dry (noncloudy) boundary layers, MYNN uses the same mass-flux scheme as used for shallow-cumulus environments: a 10-plume (assuming none of the plumes condense in a dry environment) dynamical model to perform the nonlocal mixing across the PBL top, allowing for the same physics to be used seamlessly across dry-cloudy boundary layers. The GFS-PBL uses a single plume mass-flux scheme in dry PBLs where the single plume is terminated at the PBL top, not allowing entrainment processes to be represented by its mass-flux scheme. For MYNN testing in P5 (Table 1), one of the initial condition dates failed for similar reasons as the two failures in Thompson described above. Thus, 167 cases of P5 with MYNN PBL are compared with 168 cases of the baseline P5; this represents a difference in sample size of less than 1% and should not impact the findings.

c. *Postprocessing and verification*

All model output variables were postprocessed to a regular $1° \times 1°$ grid, consistent with EMC's own internal evaluation practices for subseasonal forecasts. Output in the vertical

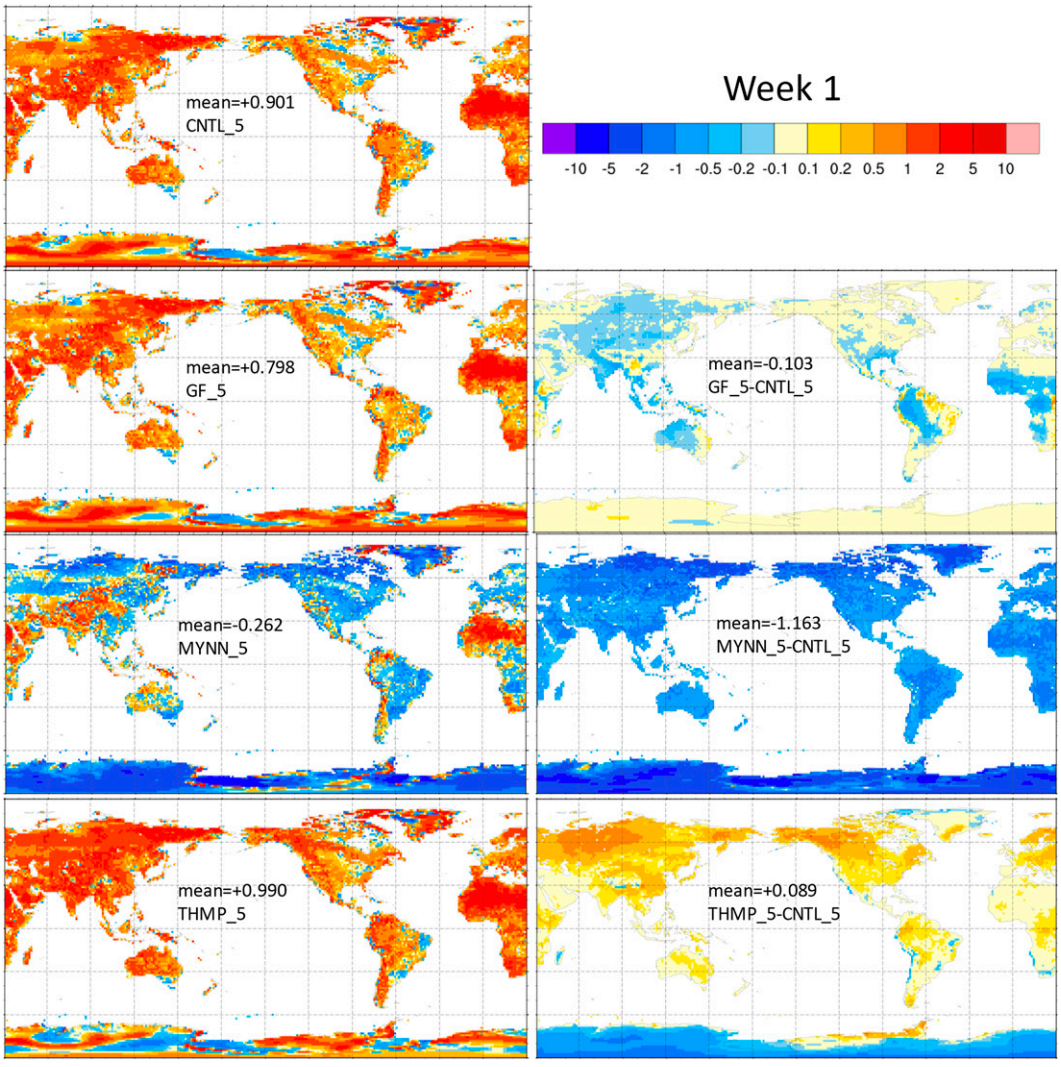

FIG. 1. (left) Annual T2m biases in lead week 1 for each of the four P5 experiments (cf. Table 1). (right) Difference fields between each of the three sensitivity experiments (GF_5, MYNN_5, and THOM_5) and CNTL_5 for lead week 1. Units are kelvins.

dimension was mapped to standard isobaric levels. In the temporal dimension, model output was available every 6 h; however, this paper presents results that are averaged daily, weekly (e.g., "week 1" is the first 7 days), or biweekly (specifically, weeks 3 and 4 combined). The use of longer analysis/verification windows with increasing lead time is common in subseasonal prediction (Zhu et al. 2014). To maintain a large sample size, results presented incorporate all initialization dates rather than subsetting by initialization season [e.g., June–July–August (JJA) for boreal summer].

For this particular work, a variety of verification databases were used. Precipitation was obtained from the Tropical Rainfall Measuring Mission 3B42 product (TRMM; Liu et al. 2012) for the global evaluation and the Climatology-Calibrated Precipitation Analysis (CCPA; Hou et al. 2014) for the coterminous United States (CONUS) evaluation. All other data, including 2-m temperature ("T2m") over land, as well as kinematic and thermodynamic variables on isobaric surfaces, were verified against reanalyses used to generate initial conditions: specifically, CFSR for P5 and GEFSv12 Reanalysis for P7. The change in verification source between prototypes is arguably necessary for T2m, because this variable is a highly diagnosed product that is sensitive to factors such as surface land use properties and near-surface vertical resolution. By verifying T2m against the analysis used to create the model's initial conditions, therefore, prognostic "biases" are more reflective of upscale error growth rather than simply highlighting areas of rugged terrain and/or sharp gradients in bulk land surface properties.

The next section shows model verification of a multitude of fields using various metrics. Mean-state biases are shown in map form not only for the variables of interest to end-users of subseasonal forecasts (precipitation and T2m), but also for 200-hPa zonal wind ("U200") and for geopotential height at

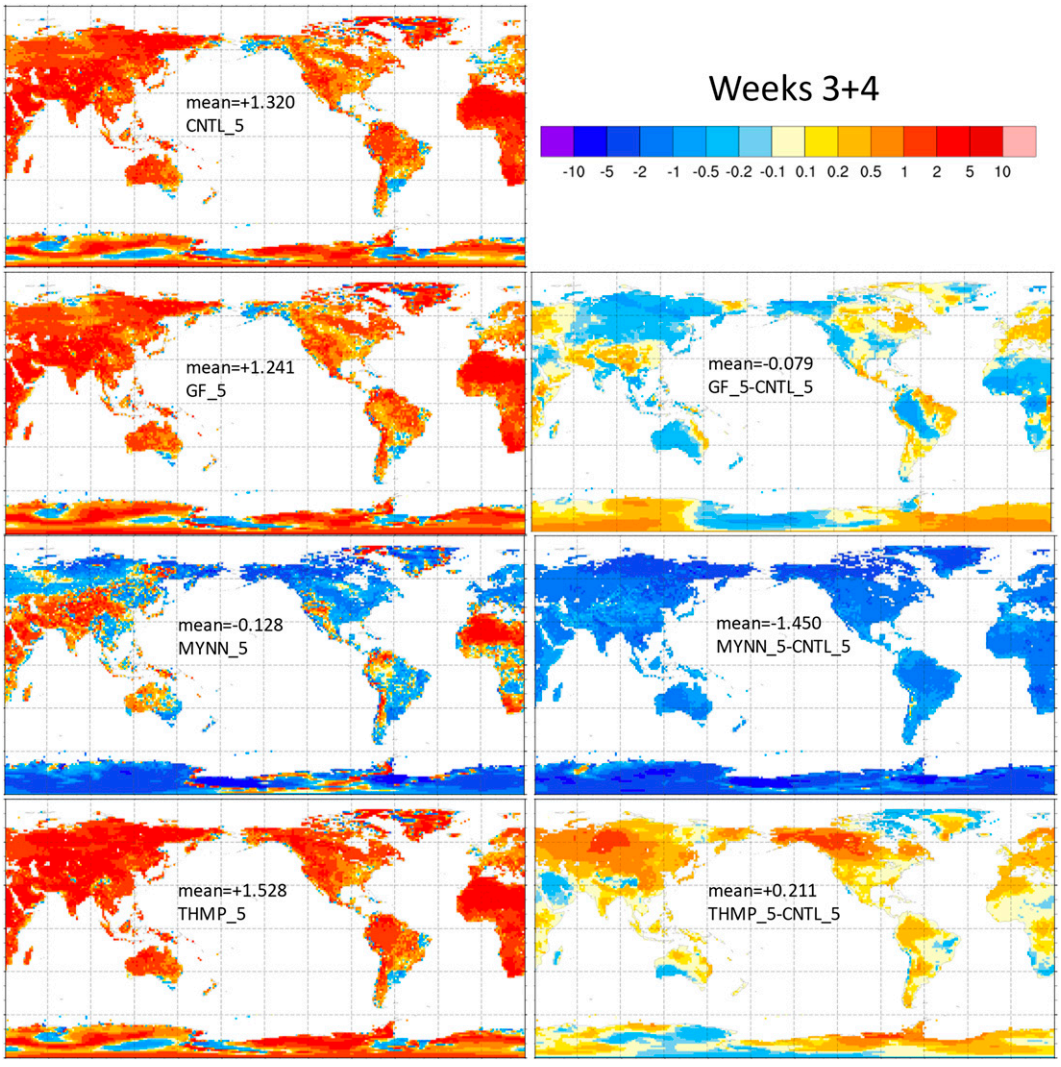

FIG. 2. As in Fig. 1, but for lead weeks 3–4.

500 hPa ("Z500"). Biases as a function of height and lead time are shown for temperature over the entire globe as well as the tropics (20°S–20°N) and extratropics (20°–80° latitude). Anomaly correlation coefficients – both for Z500 in the Northern Hemisphere and for the bivariate real-time multivariate Madden–Julian oscillation (RMM; Wheeler and Hendon 2004) – will also be presented for the experiments. Heidke Skill Scores of precipitation forecasts further demonstrate the extremely limited predictive capabilities of this field at subseasonal time scales.

3. Results

a. *Impact of physics changes relative to a fixed baseline (prototype 5)*

1) MEAN-STATE BIASES

Interrogation of mean-state biases is a critical component of model evaluation and development. Although postprocessed bias correction is critical in an operational setting, this procedure assumes that all biases are independent and thus does not actually address root cause, or account for the inherent interdependencies between variables. Instead, developers of (atmospheric) models aim to minimize bias through changes to the code (typically, subgrid-scale physical parameterization schemes) based on physical reasoning. As an example, Kim et al. (2019) established, in several subseasonal-to-seasonal prediction models, a partial association between Madden–Julian oscillation (MJO; Madden and Julian 1971, 1972) propagation biases through the Maritime Continent and Indo-Pacific biases in lower-tropospheric moisture and surface precipitation; it stands to reason, therefore, that reduction of these biases (through, e.g., modification of the PBL scheme) might improve the simulation of MJO propagation. The primary aim of the present study, however, is more to serve as a starting point for identifying the impact of changing various parameterization schemes (convection, cloud microphysics, and PBL) on mean-state biases, rather than in-depth process analyses.

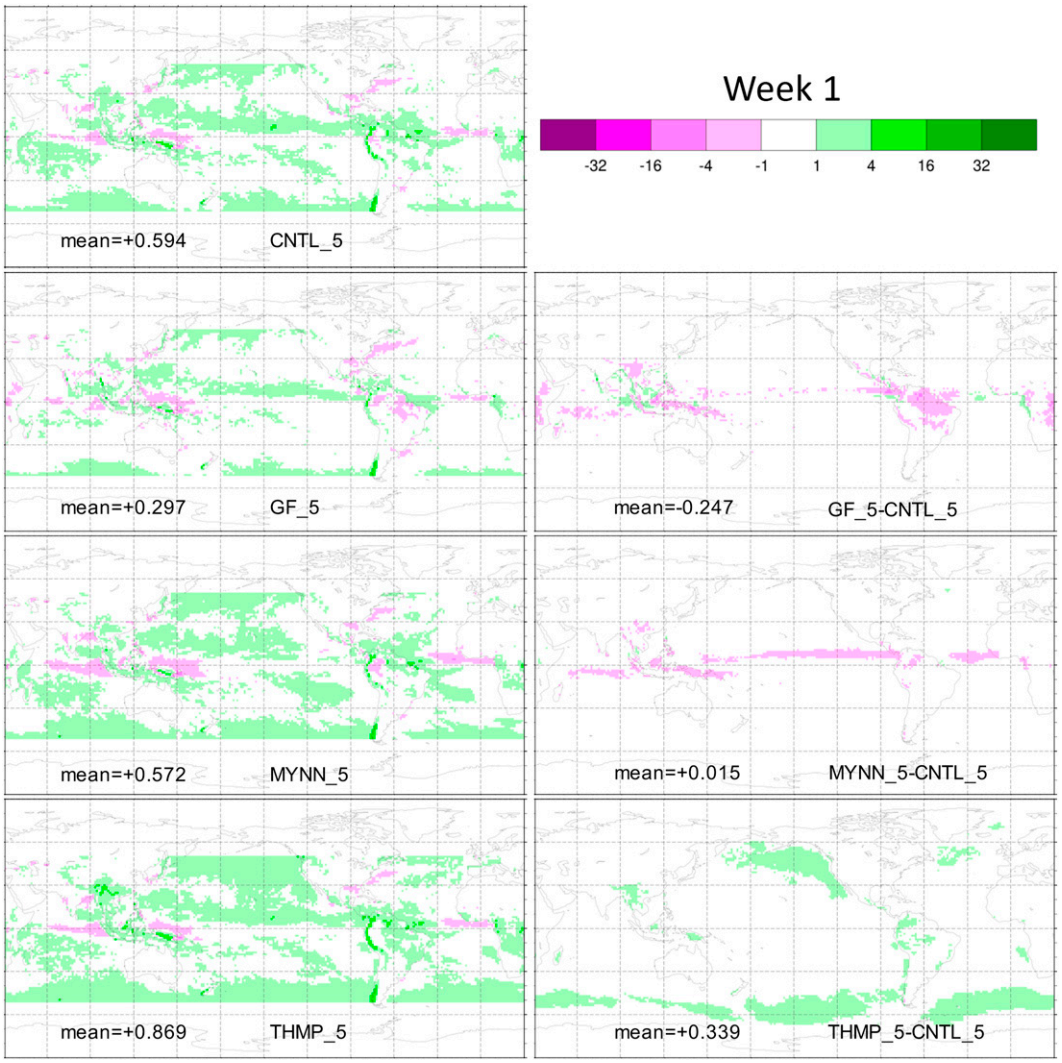

FIG. 3. As in Fig. 1, but for precipitation (mm day^{-1}). Note that biases in the left column do not cover the high latitudes, whereas differences in the right column are global.

Figures 1 and 2 show biases of T2m over land (verified against CFSR) for lead week 1 and lead weeks 3–4, respectively, for each of the P5 experiments. Difference maps between each of the three alternative physics runs (GF, MYNN, THMP) and CNTL are also plotted. In week 1, MYNN is the most different from CNTL, not surprising because the PBL scheme controls near-surface turbulent transport. MYNN is globally colder than CNTL. In contrast, GF has a very similar T2m field to CNTL in week 1, which is not unexpected because convective schemes do not have a direct, immediate impact on dry regions. However, because the convective scheme contributes to the prognostic equations of the model (rather than being purely diagnostic), its effects will eventually be seen globally at longer lead times due to upscale error growth. This is what is seen in Fig. 2: all three alternative physics runs show differences from CNTL globally at weeks 3–4, with generally larger magnitude than week 1. The most surprising result is in comparison of Figs. 1 and 2: the bias patterns established in week 1 are strikingly similar to those in weeks 3–4. For example, in MYNN for both week 1 and weeks 3–4 the African continent has generally warm biases at its northern and southern ends, with cool biases in between; another example can be seen in the GF minus CNTL plots, with a streak of blue (GF colder than CNTL) oriented northwest to southeast across South America. This result has significant implications for future development of subseasonal prediction models, as discussed later. It should be noted that these figures are not meant to convey superiority or weakness of any particular scheme. Model development, particularly within UFS, has proceeded at a rapid pace; section 3b's comparison of CNTL and GF within the P7 framework highlights how some results are changed by the underlying model baseline.

Figures 3 and 4 display precipitation biases for week 1 and weeks 3–4, respectively. Note that "mean" biases for each experiment only cover the TRMM domain, whereas differences between experiments are computed globally. Compared to

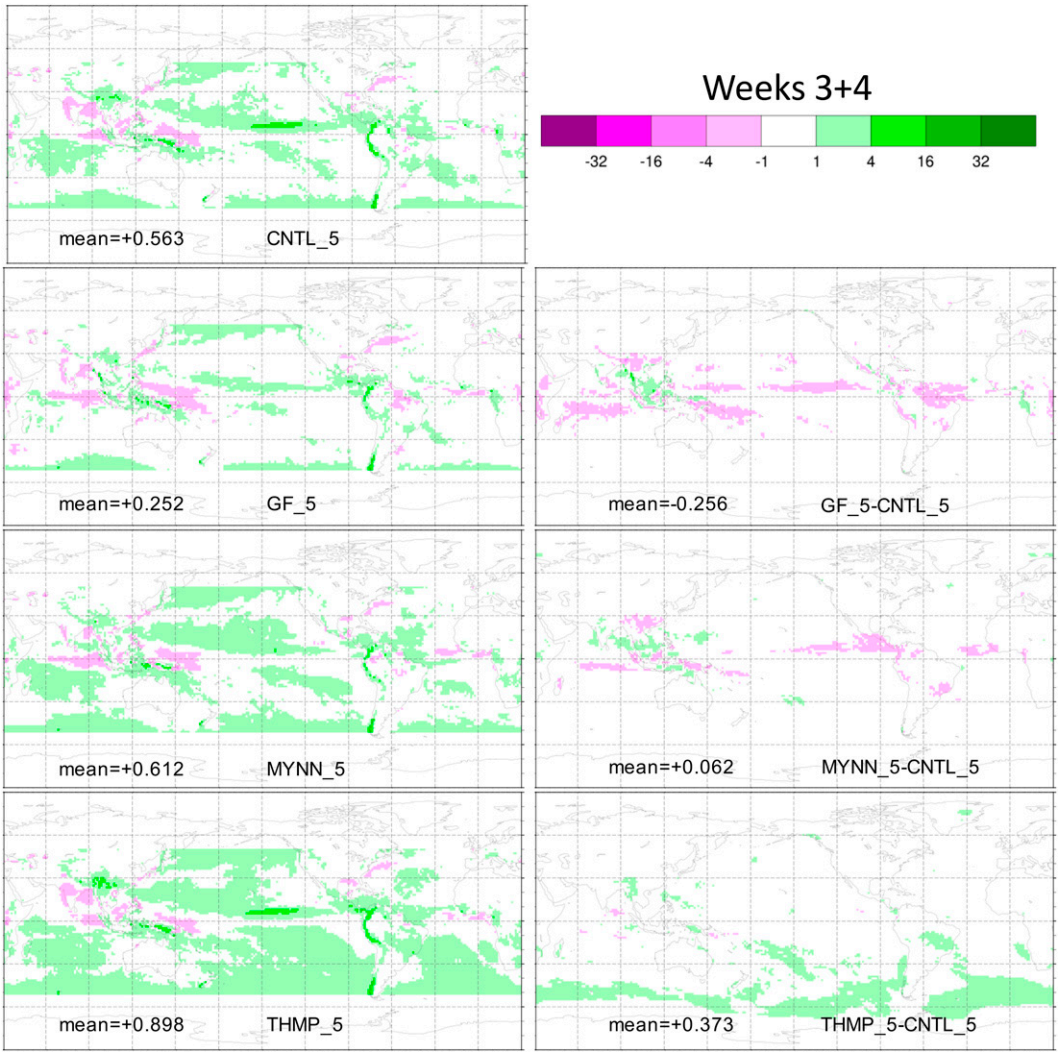

FIG. 4. As in Fig. 3, but for lead weeks 3–4.

CNTL, GF has more precipitation near the Maritime Continent but less in other tropical areas (including South America)—similar patterns exist at both week 1 and weeks 3–4. THMP has substantially more precipitation than CNTL over the extratropics, although the spatial patterns are somewhat different between lead times. Because the dominant mechanism of precipitation in the tropics (extratropics) is convective (stratiform) processes, it makes sense that GF (THMP) differs most from CNTL in these respective regions. In contrast, MYNN has globally slightly more precipitation than CNTL despite sizeable areas of less precipitation in the tropics. Planetary boundary layer processes do not have a "first order" direct effect on precipitation in the same manner as representations of convection or cloud microphysics but do have a "second order" indirect effect (particularly over the ocean) in terms of moisture transport to the free atmosphere. Overall, there is some similarity between the bias patterns in week 1 versus weeks 3–4, but not nearly of the same level as for T2m. Thus, while there may be some utility in deducing subseasonal precipitation biases through shorter runs, this should be limited to modifications in first-order processes (e.g., convection and cloud microphysics) and applied with more caution.

Subseasonal forecasting relies heavily on various global teleconnections (e.g., Black et al. 2017; Mariotti et al. 2020). It is worthwhile to look at the biases of mass and momentum fields [here, zonal winds at 200 hPa (U200) and 500-hPa geopotential height (Z500)] because systematic errors in representing teleconnections will hamper predictions. Because these free-atmosphere, quasi-balanced fields evolve more slowly than T2m or precipitation, their biases at week 1 are very small (not shown) and thus are not considered further. Figure 5 shows Z500 and U200 biases for each P5 experiment at weeks 3–4. For Z500, all experiments have a net positive bias, and high latitudes often show a high bias. MYNN has a negative bias in the tropics, which combined with its extratropical bias pattern suggests a weaker midlatitude height gradient. GF has a stronger positive bias in the tropics than the midlatitudes, suggesting a stronger height gradient – this is consistent with

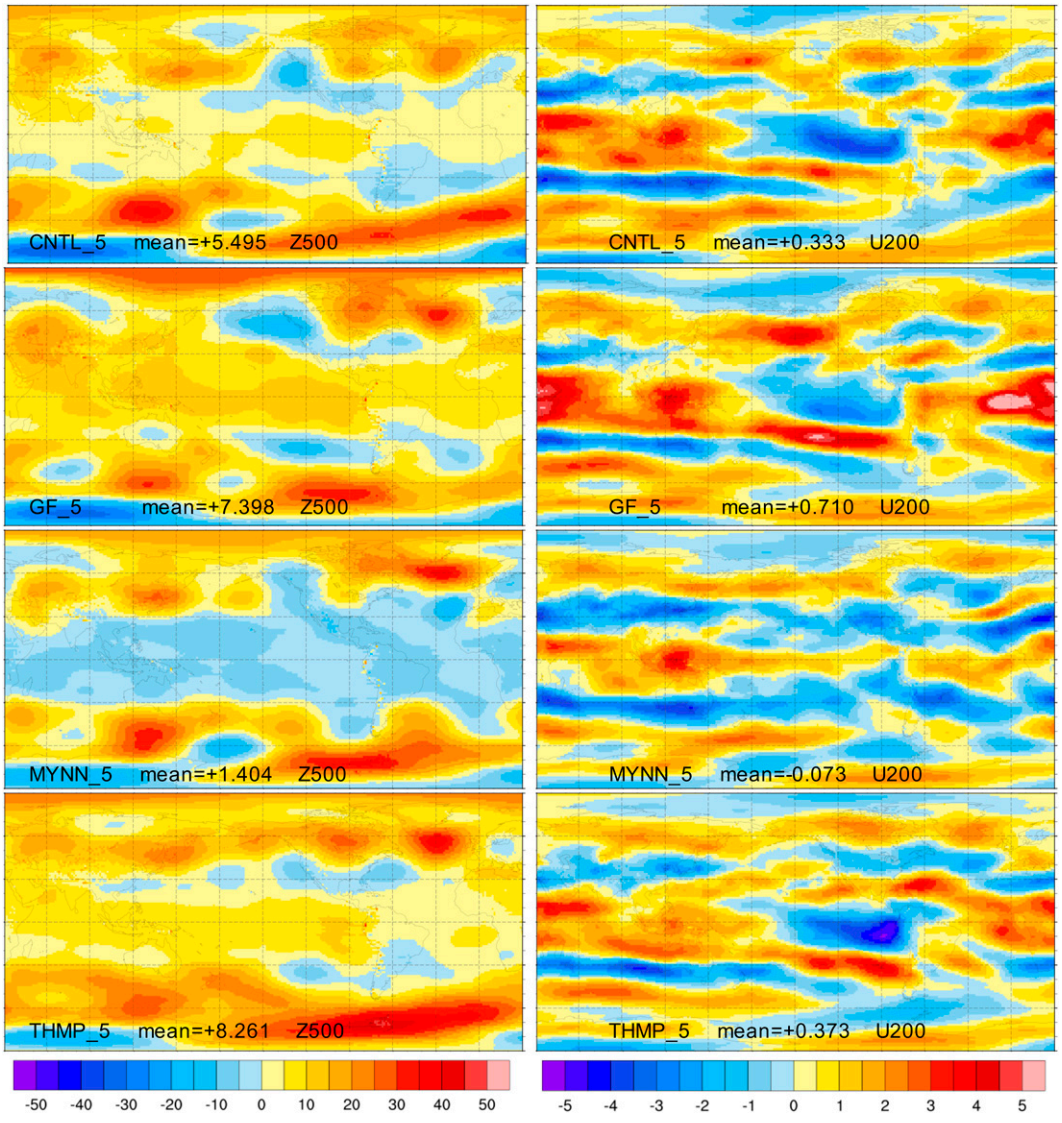

FIG. 5. Annual biases for lead weeks 3–4 for each of the four P5 experiments in (left) Z500 (m) and (right) U200 (m s^{-1}). Note the different color-bar scales.

stronger U200 westerlies in the northern Pacific. Otherwise, U200 biases overall are characterized by considerable spatial variability between positive and negative sign, and the U200 difference fields (experiment minus CNTL, not shown) are even noisier. The lack of any coherent pattern in the U200 bias maps suggests that parameterizations of convection, cloud microphysics, and planetary boundary layer may not systematically impact the upper-tropospheric circulation.

Another way to interrogate mean-state biases is to look at vertical height versus lead time plots aggregated over various areas such as the global, hemispheric (20°–80° latitude), and tropical (20°S–20°N) domains. Figure 6 does this for temperature; looking at the figure, biases composited globally are very similar to those in the Northern Hemisphere (first two rows) and the Southern Hemisphere (not shown). This figure also shows how the sign of the temperature bias for each pressure level becomes established within the first ~7 days and then grows in magnitude throughout the remainder of the 35-day forecast. Thus, for temperature, model tuning for subseasonal-length biases could potentially be accomplished via shorter (7–10 day) integrations, which would be a major computational savings. The nontropical biases for CNTL, GF, and THOM are all very similar; MYNN biases in the troposphere—particularly the lowest 300 hPa (i.e., closest to the boundary layer)—are substantially different from the other three experiments, as would be expected when changing the PBL scheme. In the tropics, GF is most different than CNTL between 500 and 100 hPa, showing the impact of convective parameterization on free-tropospheric temperature. Similarly, THOM differs most from CNTL between 300 and 100 hPa, demonstrating the impact of the cloud microphysics scheme on upper-tropospheric temperature. Somewhat surprising is that MYNN is cooler than

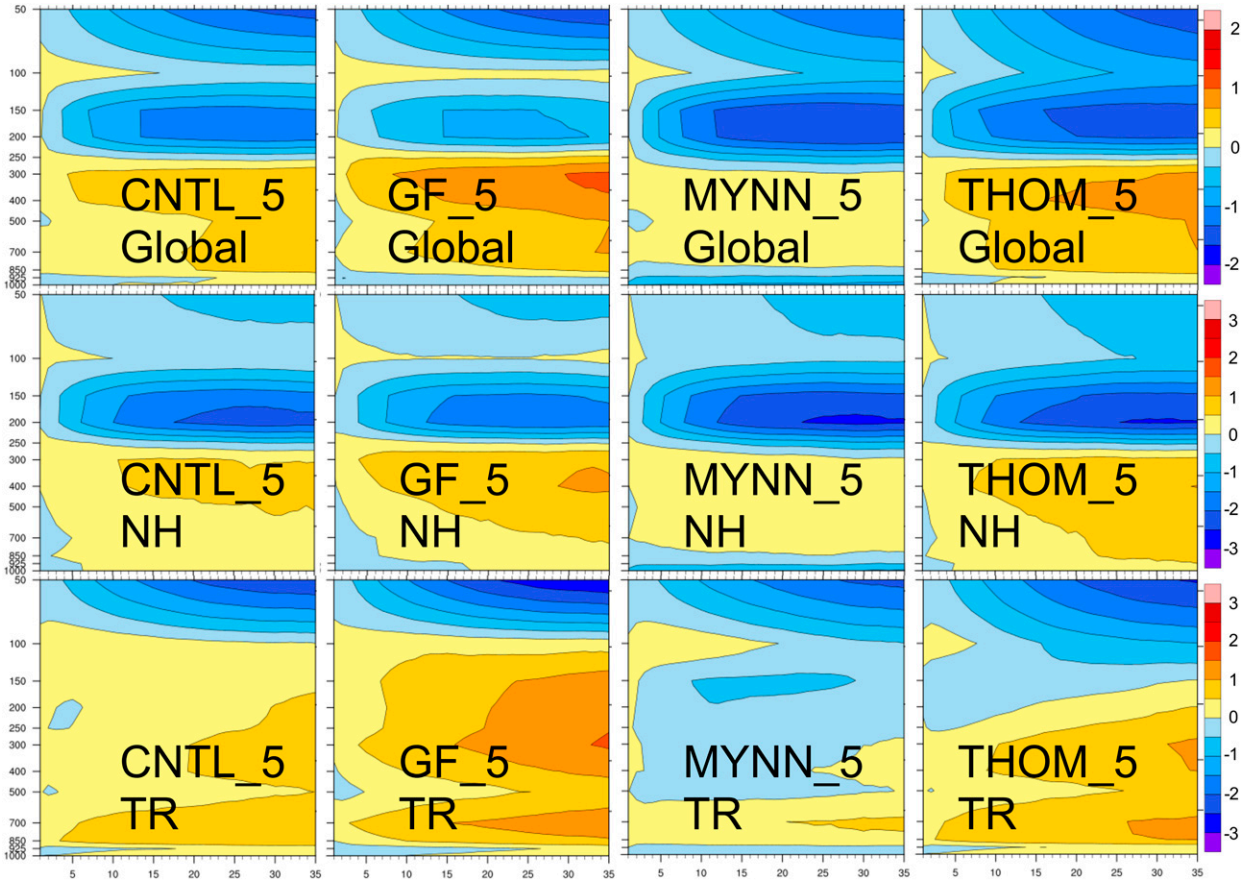

FIG. 6. Pressure vs lead-time temperature biases (K) for each of the four P5 experiments (left) CNTL_5, (left center) GF_5, (right center) MYNN_5, and (right) THOM_5 for the (top) global mean, (middle) Northern Hemisphere (20°–80°N), and (bottom) tropics (20°S–20°N). Note the different color-bar scales.

CNTL not just in the tropical boundary layer but also in much of the free atmosphere. It is speculated that the free atmospheric differences are a result of secondary effects, namely, different representations of PBL flux impact the organization/distribution of tropical convection and in turn affect the temperature, but a more thorough investigation is beyond the scope of this paper. Specific humidity biases were also examined (not shown); because moisture decreases rapidly going poleward, global composites most resembled those of the tropics and indicated—as expected—that changing PBL scheme had the greatest impact in the lowest levels of the atmosphere, whereas convection and cloud microphysics had larger impacts in the free atmosphere.

2) SKILL SCORES

Beyond examining mean-state biases, it is useful to compare the performance of the various physics schemes within the P5 framework in terms of skill score. Figure 7 shows the anomaly correlation coefficients for Z500 in the Northern Hemisphere (specifically, 20°–80°N) for the first 16 days of all forecasts. Beyond this time, Z500 skill is very low as expected due to intrinsic predictability limits. In general, changing the physics scheme has no statistically significant impact on Z500 anomaly correlations. However, this is not the case for THOM, which within the P5 framework has statistically significantly worse skill than CNTL in the first 192 h in the Northern Hemisphere. In the Southern Hemisphere (not shown), no experiment was statistically significantly different from CNTL for Z500 skill. Despite the lower (P5) skill in THOM over the Northern Hemisphere, EMC chose to use Thompson over GFDL microphysics for P8 because parallel testing with later versions of UFS (at shorter forecast lead times) revealed Thompson to produce better forecasts than GFDL (not shown). As detailed further in section 3b, extreme caution must be used in trying to generalize the findings of this paper, namely: results herein should not be used to draw conclusions about which physics schemes are better or worse than others. Physics schemes are constantly being updated, and the other components of models that interact with these schemes are also constantly being updated.

Although a systematic investigation into the effects of varying physics schemes on MJO is beyond the scope of this paper, it would be remiss not to provide a cursory look at forecast performance. Figure 8 shows the anomaly correlation of the bivariate RMM index for each of the P5 experiments.

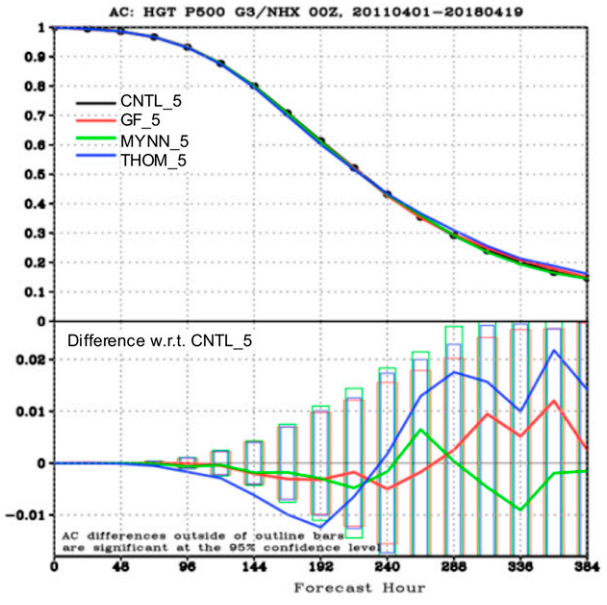

FIG. 7. (top) Anomaly correlation (AC) coefficients for Northern Hemisphere (20°–80°N) Z500 in the first 384 forecast hours for each of the four P5 experiments. (bottom) Differences in GF_5, MYNN_5, and THOM_5 with respect to CNTL_5.

The P5 experiments have 17–20 days of skillful MJO forecasts, using the typical correlation threshold of 0.5. Cursory investigation of RMM phase and amplitude errors (not shown) reveal an amplitude too weak in the first ~7 days, and then too strong for the remainder of the 35-day forecast; all experiments are also too slow in weeks 3 and 4. The THOM experiment performs the best in terms of RMM, although it was shown that this experiment had the lowest Z500 skill (Fig. 7). There is no reason to expect that a particular model configuration with the best (or worst) Z500 skill will perform similarly for RMM, but the THOM result underscores the difficulties faced by operational centers in implementing changes to a model: some metrics may improve, but others may degrade.

Finally, the skill of precipitation in CONUS is examined through the Heidke skill score (HSS), with three categories (below, near, and above normal) and based on the CCPA climatology. Note that model climatology was not considered at all. The HSS for precipitation falls off dramatically by Week 2, and the only location with HSS above 50% is California (not shown). Figure 9 shows HSS aggregated over CONUS for each lead week and each experiment. Clearly, direct prediction of precipitation at subseasonal time scales over the United States is a daunting challenge. Aggregating weeks 3 and 4 together does not change skill compared to either of the two weeks separately. No single experiment is consistently better (or worse) than the others: thus, swapping physics schemes within the same baseline model framework does not have a clear impact on CONUS precipitation prediction. In the next subsection, the impact of changing the baseline is briefly examined.

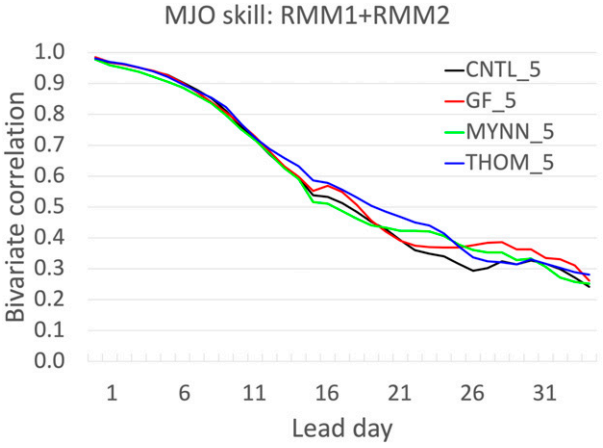

FIG. 8. Bivariate correlation (between forecast and observed) of RMM index for each of the four P5 experiments as a function of forecast lead day.

b. *Impact of changing baseline*

So far, results have been presented relative to a common baseline (coupled UFS prototype 5, P5). Within P5, it was shown that mean-state bias patterns at weeks 3–4 are often established within the first week, and that no physics scheme tested is uniformly superior or inferior. However, Earth system models are constantly undergoing simultaneous development on multiple fronts (initialization strategies, numerics, subgrid-scale physics, etc.), and it is computationally infeasible to conduct one-at-a-time tests for each and every change made to a modeling system. Moreover, the highly nonlinear and interconnected nature of the Earth system guarantees that any one change will have global impacts. Thus, when multiple (seemingly beneficial) changes are made at once, there is no guarantee that the net result will be uniformly positive. Similarly, any conclusions drawn from the P5-based experiments shown above are strictly valid within the P5 framework: subsequent developments in the coupled UFS prototype framework may impact, for example, how changing the convective scheme changes results. To demonstrate

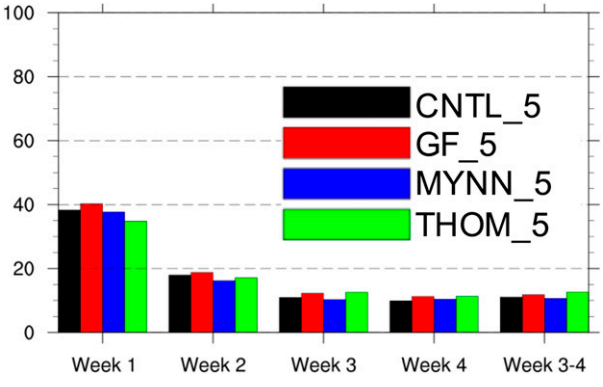

FIG. 9. Precipitation HSS over CONUS for each of the four P5 experiments by lead week; note that the combined weeks 3–4 forecast HSS is also shown.

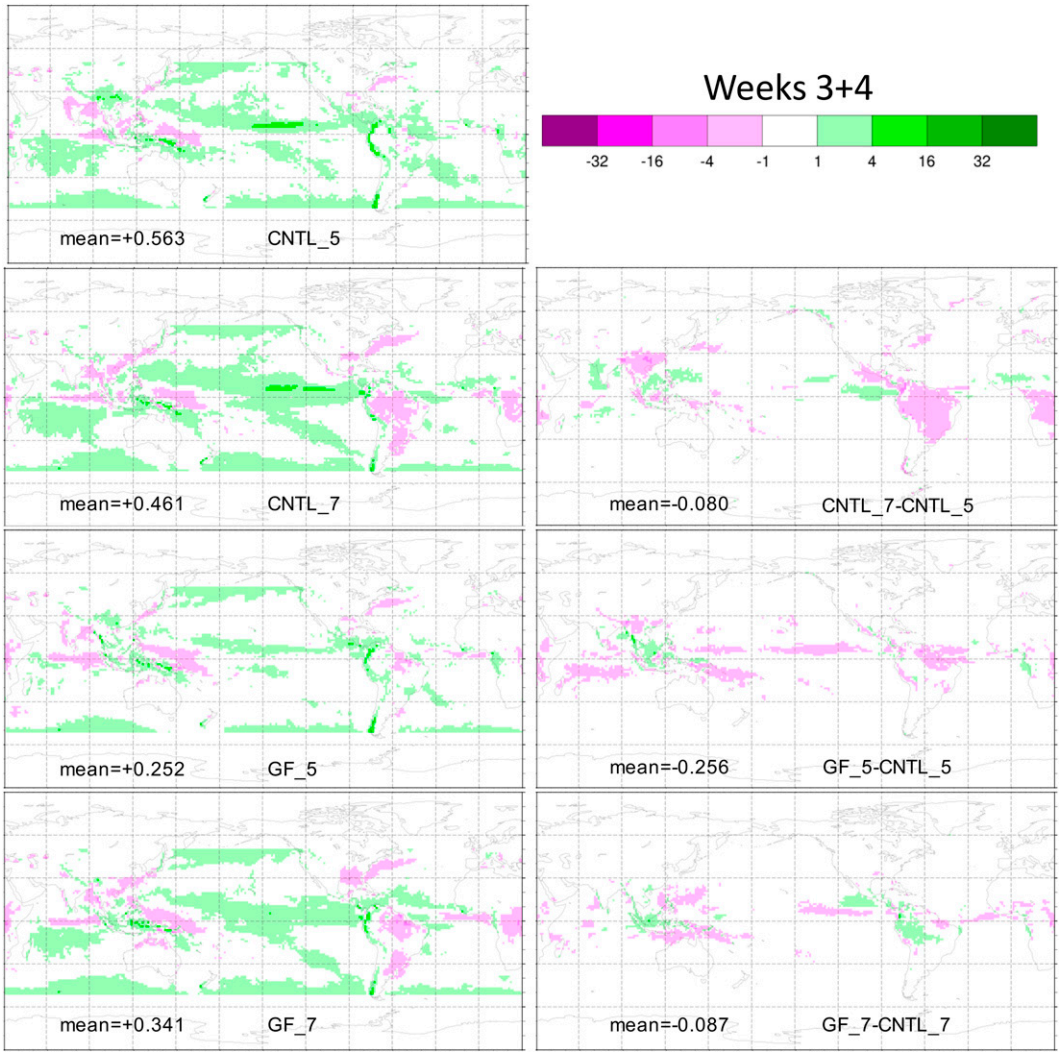

FIG. 10. (left) Annual precipitation biases at lead weeks 3–4 for CNTL_5, CNTL_7, GF_5, and GF_7. (right) Differences between various experiments (specifically, CNTL_7 − CNTL_5, GF_5 − CNTL_5, and GF_7 − CNTL_7). Units are millimeters per day.

this, some results from the baseline Prototype 7 (P7), as well as a P7-based experiment with GF convection, are presented here.

Figure 10 is similar to Fig. 4 in showing weeks 3–4 precipitation biases; in Fig. 10, however, CNTL and GF experiments for both P5 and P7 are plotted. The second row of the figure shows how the precipitation biases change from P5 to P7. It should be noted that the different baselines have changed due to differences that include, but are not limited to, initial condition fields, physics time steps, vertical levels, and updates to the physics schemes. The baseline P7 has less precipitation globally than the baseline P5, but within the GF framework the opposite is true: GF_7 has more global precipitation than GF_5. In comparing Figs. 4 and 10, it is seen that SAS has more precipitation than GF *for a given baseline* (i.e., CNTL_7 is wetter than GF_7, and CNTL_5 is wetter than GF_5). Looking at the bottom two rows of Fig. 10, it is obvious that comparisons between GF and the default SAS are baseline dependent (due, in part, to changes in both GF and SAS over a year of model development)—particularly over South America and the landmasses of the Maritime Continent. Another substantial change is that whereas GF_5 has cooler T2m than CNTL_5 over much of Africa (cf. Fig. 2), GF_7 has warmer T2m than CNTL_7 over that same region (not shown). Thus, readers are warned against extrapolating specific results from P5 to other coupled UFS prototypes, much less to changing physics within other, non-UFS, modeling systems. The main finding that did hold true between P5 and P7, while not explicitly shown for P7, is that mean-state biases generally develop their spatial (horizontal and vertical) patterns within the first week or so and then simply grow in magnitude through the subseasonal period.

Figure 11 shows how skill scores can change across baselines, using the RMM as an example (cf. Fig. 8). Whereas in

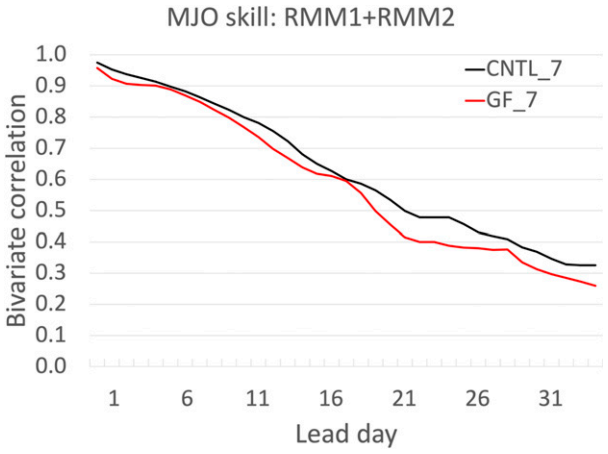

FIG. 11. As in Fig. 8, but for the two P7 experiments.

P5 both GF and CNTL reached a bivariate correlation of 0.5 at ~19 days (with GF slightly better than CNTL), Fig. 11 shows that in P7 the 0.5 correlation threshold is reached at 20 days for GF and at 22 days for CNTL. Although a more thorough examination of MJO representation in these simulations is beyond the scope of the present paper, it is likely that changes to the convective parameterizations (GF and SAS) as well as other changes between P5 and P7 all contributed to the differences in RMM results.

4. Discussion and conclusions

This study presented results from experiments that tested the sensitivity of coupled (atmosphere, sea ice, ocean, wave) Earth system simulations at subseasonal time scales (up to 35 days) to changes in parameterizations of atmospheric convection, cloud microphysics, and planetary boundary layer. This provides a comprehensive set of controlled subseasonal-length physics sensitivity experiments to be examined. One of the key motivations for these experiments is the ongoing work by NOAA to upgrade their operational Global Ensemble Forecast System (GEFS)—a modeling system that sees fairly regular updates and is heavily relied on for forecasts in the first 2 weeks—to a fully coupled Earth system model with predictions out to 35 days. Currently, operational coupled subseasonal forecasts at NOAA are handled by the Climate Forecast System, version 2 (CFSv2), which has not been updated in over a decade.

Independent tests for convection (GF replacing the default SAS), cloud microphysics (Thompson replacing the default GFDL), and planetary boundary layer (MYNN replacing the default GFS-PBL) were compared with the benchmark Prototype 5 (P5) run by EMC in late 2020. An additional GF versus SAS test was conducted against the benchmark Prototype 7 (P7) run by EMC in late 2021. Within the P5 framework, it was found that no experiment was unilaterally better or worse than any other in terms of mean-state biases (of 2-m temperature, precipitation, mass, and momentum fields) and skill scores (500-hPa anomaly correlation, RMM for the MJO, and precipitation).

The most important finding was that mean-state biases tend to manifest their spatial pattern (in both the horizontal and vertical) within the first week, and then simply grow in amplitude through the remainder of the 35-day simulation period. As expected, the differences in mean states between any two experiments were most pronounced for fields in which the changed physics parameterization would have a first-order direct effect (e.g., changing the PBL scheme had the most profound impact on near-surface temperature). These results have tremendous implications for subseasonal model development, namely, shorter runs on the order of 1–2 weeks may be conducted instead to provide insight into how proposed changes might impact forecast biases at longer lead times (however, bias improvement may not translate into improved skill scores). With shorter runs, the computational savings could be reallocated to more testing and/or larger sample sizes. A somewhat surprising result was that the bias fields of all the experimental runs (and their differences from each other) for zonal winds at 200 hPa were very noisy in the horizontal, signaling that it may be difficult to isolate biases related to global teleconnections that are so vital for subseasonal prediction.

The finding that subseasonal mean-state biases are similar to those within the first week is consistent with the results of Ma et al. (2021), who compared short (3-day) runs with a 16-yr-long climate run and found that many error fields at 3 days were very similar to the average errors from the long-term simulation. However, Ma et al. (2021) did not consider the impact of the annual cycle; for example, their results are often restricted to certain months of the year (e.g., November–April in their Figs. 4–6). The present study indicates that the impacts of the annual cycle on bias evolution are not evident through the first 5 weeks.

Although not unexpected, the results of this work serve as yet another warning against generalization of any apparent benefits or deficiencies in a particular physics parameterization. For example, it was found in P5 that GF yielded cooler 2-m temperatures than SAS over much of Africa, but the opposite was true in P7. Contradictory results were also seen in RMM skill. With the rapid pace of development in many Earth modeling systems like UFS, isolation of cause is extremely difficult.

Future work should focus on the quantification of mean-state error growth from initialization through the subseasonal time scale. This would allow model developers to determine which biases they can attempt to tune through shorter, more computationally feasible simulations. Another avenue for research is to examine the physics sensitivity at seasonal time scales and, just as important, to examine the bias evolution—particularly through the first year, covering one annual cycle with particular focus on the ~6-month period (especially comparing a winter season with a summer season). Also, process-oriented diagnostics can be employed on the hindcast sets described here to gain a deeper understanding of the role of various physics schemes on

subseasonal-length simulations (including their state- and location-dependent predictability).

Acknowledgments. This material is based upon work supported by the Climate Testbed within the NOAA/OAR Weather Program Office under Award NA20OAR4590314 (specifically for coauthors Green and Sinsky). The first author is also supported by funding from NOAA OAR Awards NA17OAR4320101 and NA22OAR4320151. The authors acknowledge the NOAA Research and Development High Performance Computing Program for providing computing and storage resources that have contributed to the research results reported within this article (http://rdhpcs.noaa.gov). CFSR data are obtained from the National Centers for Environmental Information. Josh Hacker and three anonymous reviewers are thanked for their beneficial comments during formal peer review. Dave Turner is thanked for providing internal review; Lydia Stefanova and Wei Li contributed plotting scripts; and Fanglin Yang, Hannah Barnes, Joseph Olson, Anders Jensen, and Stanley Benjamin provided helpful suggestions.

Data availability statement. The data associated with the figures produced in this paper are stored on NOAA mass-store tape and are available upon request from the lead author.

REFERENCES

Adcroft, A., and Coauthors, 2019: The GFDL global ocean and sea ice model OM4.0: Model description and simulation features. *J. Adv. Model. Earth Syst.*, **11**, 3167–3211, https://doi.org/10.1029/2019MS001726.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Black, J., N. C. Johnson, S. Baxter, S. B. Feldstein, D. S. Harnos, and M. L. L'Heureux, 2017: The predictors and forecast skill of Northern Hemisphere teleconnection patterns for lead times of 3–4 weeks. *Mon. Wea. Rev.*, **145**, 2855–2877, https://doi.org/10.1175/MWR-D-16-0394.1.

Bogenschutz, P. A., A. Gettelman, H. Morrison, V. E. Larson, C. Craig, and D. P. Schanen, 2013: Higher-order turbulence closure and its impact on climate simulations in the Community Atmosphere Model. *J. Climate*, **26**, 9655–9676, https://doi.org/10.1175/JCLI-D-13-00075.1.

Brunet, G., and Coauthors, 2010: Collaboration of the weather and climate communities to advance subseasonal-to-seasonal prediction. *Bull. Amer. Meteor. Soc.*, **91**, 1397–1406, https://doi.org/10.1175/2010BAMS3013.1.

Collow, T., Y. Liu, W. Wang, and A. Kumar, 2019: Overview of the CPC Sea Ice Initialization System (CSIS) and its use in experimental sea ice forecasting at the NOAA Climate Prediction Center. *44th NOAA Annual Climate Diagnostics and Prediction Workshop*, Durham, NC, NOAA, 85–88, https://www.nws.noaa.gov/ost/climate/STIP/44CDPW/44cdpw-TCollow.pdf.

Dalcher, A., and E. Kalnay, 1987: Error growth and predictability in operational ECMWF forecasts. *Tellus*, **39A**, 474–491, https://doi.org/10.3402/tellusa.v39i5.11774.

Freitas, S. R., G. A. Grell, and H. Li, 2021: The Grell-Freitas (GF) convection parameterization: Recent developments, extensions, and applications. *Geosci. Model Dev.*, **14**, 5393–5411, https://doi.org/10.5194/gmd-14-5393-2021.

Gómez, I., F. Pastor, and M. J. Estrela, 2011: Sensitivity of a mesoscale model to different convective parameterization schemes in a heavy rain event. *Nat. Hazards Earth Syst. Sci.*, **11**, 343–357, https://doi.org/10.5194/nhess-11-343-2011.

Green, B. W., and F. Zhang, 2013: Impacts of air–sea flux parameterizations on the intensity and structure of tropical cyclones. *Mon. Wea. Rev.*, **141**, 2308–2324, https://doi.org/10.1175/MWR-D-12-00274.1.

——, S. Sun, R. Bleck, S. G. Benjamin, and G. A. Grell, 2017: Evaluation of MJO predictive skill in multiphysics and multimodel global ensembles. *Mon. Wea. Rev.*, **145**, 2555–2574, https://doi.org/10.1175/MWR-D-16-0419.1.

Grell, G. A., 1993: Prognostic evaluation of assumptions used by cumulus parameterizations. *Mon. Wea. Rev.*, **121**, 764–787, https://doi.org/10.1175/1520-0493(1993)121<0764:PEOAUB>2.0.CO;2.

——, and D. Devenyi, 2002: A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. *Geophys. Res. Lett.*, **29**, 1693, https://doi.org/10.1029/2002GL015311.

——, and S. R. Freitas, 2014: A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling. *Atmos. Chem. Phys.*, **14**, 5233–5250, https://doi.org/10.5194/acp-14-5233-2014.

——, Y.-H. Kuo, and R. J. Pasch, 1991: Semiprognostic tests of cumulus parameterization schemes in the middle latitudes. *Mon. Wea. Rev.*, **119**, 5–31, https://doi.org/10.1175/1520-0493(1991)119<0005:STOCPS>2.0.CO;2.

Guimarães, B. S., C. A. S. Coelho, S. J. Woolnough, P. Y. Kubota, C. F. Bastarz, S. N. Figueroa, J. P. Bonatti, and D. C. de Souza, 2020: Configuration and hindcast quality assessment of a Brazilian global sub-seasonal prediction system. *Quart. J. Roy. Meteor. Soc.*, **146**, 1067–1084, https://doi.org/10.1002/qj.3725.

Hamill, T. M., and Coauthors, 2022: The reanalysis for the Global Ensemble Forecast System, version 12. *Mon. Wea. Rev.*, **150**, 59–79, https://doi.org/10.1175/MWR-D-21-0023.1.

Han, J., and H.-L. Pan, 2011: Revision of convection and vertical diffusion schemes in the NCEP Global Forecast System. *Wea. Forecasting*, **26**, 520–533, https://doi.org/10.1175/WAF-D-10-05038.1.

——, M. L. Witek, J. Teixeira, R. Sun, H.-L. Pan, J. K. Fletcher, and C. S. Bretherton, 2016: Implementation in the NCEP GFS of a hybrid eddy-diffusivity mass-flux (EDMF) boundary layer parameterization with dissipative heating and modified stable boundary layer mixing. *Wea. Forecasting*, **31**, 341–352, https://doi.org/10.1175/WAF-D-15-0053.1.

——, W. Wang, Y. C. Kwon, S.-Y. Hong, V. Tallapragada, and F. Yang, 2017: Updates in the NCEP GFS cumulus convection schemes with scale and aerosol awareness. *Wea. Forecasting*, **32**, 2005–2017, https://doi.org/10.1175/WAF-D-17-0046.1.

——, W. Li, F. Yang, E. Strobach, W. Zheng, and R. Sun, 2021: Updates in the NCEP GFS cumulus convection, vertical turbulent mixing, and surface layer physics. NCEP Office Note 505, 18 pp., https://doi.org/10.25923/cybh-w893.

Han, J.-Y., S.-Y. Hong, and Y. C. Kwon, 2020: The performance of a revised simplified Arakawa-Schubert (SAS) convection scheme in the medium-range forecasts of the Korean

Integrated Model (KIM). *Wea. Forecasting*, **35**, 1113–1128, https://doi.org/10.1175/WAF-D-19-0219.1.

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

——, X. Chen, W. Putman, L. Zhou, and J.-H. Chen, 2021: A scientific description of the GFDL finite-volume cubed-sphere dynamical core. NOAA Tech. Memo. OAR GFDL-2021(001), 109 pp., https://doi.org/10.25923/6nhs-5897.

Hazra, A., H. S. Chaudhari, S. K. Saha, and S. Pokhrel, 2017: Effect of cloud microphysics on Indian summer monsoon precipitating clouds: A coupled climate modeling study. *J. Geophys. Res. Atmos.*, **122**, 3786–3805, https://doi.org/10.1002/2016JD026106.

Hou, D., and Coauthors, 2014: Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of Stage IV toward CPC gauge-based analysis. *J. Hydrometeor.*, **15**, 2542–2557, https://doi.org/10.1175/JHM-D-11-0140.1.

Hu, X.-M., J. W. Nielsen-Gammon, and F. Zhang, 2010: Evaluation of three planetary boundary layer schemes in the WRF Model. *J. Appl. Meteor. Climatol.*, **49**, 1831–1844, https://doi.org/10.1175/2010JAMC2432.1.

Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough, and W. D. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *J. Geophys. Res.*, **113**, D13103, https://doi.org/10.1029/2008JD009944.

Jia, X., C. Li, N. Zhou, and J. Ling, 2010: The MJO in an AGCM with three different cumulus parameterization schemes. *Dyn. Atmos. Oceans*, **49**, 141–163, https://doi.org/10.1016/j.dynatmoce.2009.02.003.

Jones, C., J. Gottschalck, L. M. V. Carvalho, and W. Higgins, 2011: Influence of the Madden–Julian oscillation on forecasts of extreme precipitation in the contiguous United States. *Mon. Wea. Rev.*, **139**, 332–350, https://doi.org/10.1175/2010MWR3512.1.

Kaur, M., A. K. Sahai, R. Phani, S. Joseph, R. Mandal, A. Dey, and R. Chattopadhyay, 2022: Multi-physics schema for subseasonal prediction of the Indian summer monsoon. *Climate Dyn.*, **58**, 669–690, https://doi.org/10.1007/s00382-021-05926-2.

Kim, H., M. A. Janiga, and K. Pegion, 2019: MJO propagation processes and mean biases in the SubX and S2S reforecasts. *J. Geophys. Res. Atmos.*, **124**, 9314–9331, https://doi.org/10.1029/2019JD031139.

Kirtman, B. P., and Coauthors, 2014: The North American Multimodel Ensemble: Phase-1 seasonal-to-interannual prediction; Phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, https://doi.org/10.1175/BAMS-D-12-00050.1.

Lin, J., and Coauthors, 2022: Atmospheric convection. *Atmos.–Ocean*, **60**, 422–476, https://doi.org/10.1080/07055900.2022.2082915.

Lin, Y.-L., R. D. Farley, and H. D. Orville, 1983: Bulk parameterization of the snow field in a cloud model. *J. Climate Appl. Meteor.*, **22**, 1065–1092, https://doi.org/10.1175/1520-0450(1983)022<1065:BPOTSF>2.0.CO;2.

Liu, Y., W. Wang, A. Kumar, and T. Collow, 2019: Assessment of CPC Sea Ice Initialization System (CSIS) and CPC weekly experimental sea ice forecasts. *44th NOAA Annual Climate Diagnostics and Prediction Workshop*, Durham, NC, NOAA, 74–77, https://www.nws.noaa.gov/ost/climate/STIP/44CDPW/44cdpw-YLiu.pdf.

Liu, Z., D. Ostrenga, W. Teng, and S. Kempler, 2012: Tropical Rainfall Measuring Mission (TRMM) precipitation data and services for research and applications. *Bull. Amer. Meteor. Soc.*, **93**, 1317–1325, https://doi.org/10.1175/BAMS-D-11-00152.1.

Lohmann, U., P. Stier, C. Hoose, S. Ferrachat, S. Kloster, E. Roeckner, and J. Zhang, 2007: Cloud microphysics and aerosol indirect effects in the global climate model ECHAM5-HAM. *Atmos. Chem. Phys.*, **7**, 3425–3446, https://doi.org/10.5194/acp-7-3425-2007.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21A**, 289–307, https://doi.org/10.3402/tellusa.v21i3.10086.

Ma, H.-Y., and Coauthors, 2021: A multi-year short-range hindcast experiment with CESM1 for evaluating climate model moist processes from diurnal to interannual timescales. *Geosci. Model Dev.*, **14**, 73–90, https://doi.org/10.5194/gmd-14-73-2021.

Madden, R. A., and P. R. Julian, 1971: Detection of a 40–50 day oscillation in the zonal wind in the tropical Pacific. *J. Atmos. Sci.*, **28**, 702–708, https://doi.org/10.1175/1520-0469(1971)028<0702:DOADOI>2.0.CO;2.

——, and ——, 1972: Description of global-scale circulation cells in the tropics with a 40–50 day period. *J. Atmos. Sci.*, **29**, 1109–1123, https://doi.org/10.1175/1520-0469(1972)029<1109:DOGSCC>2.0.CO;2.

Mariotti, A., and Coauthors, 2020: Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bull. Amer. Meteor. Soc.*, **101**, E608–E625, https://doi.org/10.1175/BAMS-D-18-0326.1.

McMillen, J. D., and W. J. Steenburgh, 2015: Impact of microphysics parameterizations on simulations of the 27 October 2010 Great Salt Lake–effect snowstorm. *Wea. Forecasting*, **30**, 136–152, https://doi.org/10.1175/WAF-D-14-00060.1.

Mellor, G. L., and T. Yamada, 1974: A hierarchy of turbulence closure models for planetary boundary layers. *J. Atmos. Sci.*, **31**, 1791–1806, https://doi.org/10.1175/1520-0469(1974)031<1791:AHOTCM>2.0.CO;2.

Mitchell, K., and Coauthors, 2005: The community Noah Land-Surface Model (LSM) user's guide public release version 2.7.1. Tech. Doc., NOAA, 26 pp., https://ral.ucar.edu/document-or-file/noah-lsm-users-guide.

Miyamoto, Y., Y. Kajikawa, R. Yoshida, T. Yamaura, H. Yashiro, and H. Tomita, 2013: Deep moist atmospheric convection in a subkilometer global simulation. *Geophys. Res. Lett.*, **40**, 4922–4926, https://doi.org/10.1002/grl.50944.

Morrison, H., G. Thompson, and V. Tatarskii, 2009: Impact of cloud microphysics on the development of trailing stratiform precipitation in a simulated squall line: Comparison of one- and two-moment schemes. *Mon. Wea. Rev.*, **137**, 991–1007, https://doi.org/10.1175/2008MWR2556.1.

Nakanishi, M., 2001: Improvement of the Mellor-Yamada turbulence closure model based on large-eddy simulation data. *Bound.-Layer Meteor.*, **99**, 349–378, https://doi.org/10.1023/A:1018915827400.

——, and H. Niino, 2004: An improved Mellor-Yamada level-3 model with condensation physics: Its design and verification. *Bound.-Layer Meteor.*, **112**, 1–31, https://doi.org/10.1023/B:BOUN.0000020164.04146.98.

——, and ——, 2006: An improved Mellor-Yamada level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Bound.-Layer Meteor.*, **119**, 397–407, https://doi.org/10.1007/s10546-005-9030-8.

——, and ——, 2009: Development of an improved turbulence closure model for the atmospheric boundary layer. *J. Meteor. Soc. Japan*, **87**, 895–912, https://doi.org/10.2151/jmsj.87.895.

Niu, G.-Y., and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J. Geophys. Res.*, **116**, D12109, https://doi.org/10.1029/2010JD015139.

Olson, J. B., J. S. Kenyon, W. A. Angevine, J. M. Brown, M. Pagowski, and K. Suselj, 2019: A description of the MYNN-EDMF scheme and the coupling to other components in WRF-ARW. NOAA Tech. Memo. OAR GSD-61, 42 pp., https://doi.org/10.25923/n9wm-be49.

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

Potvin, C. K., and Coauthors, 2019: Systematic comparison of convection-allowing models during the 2017 NOAA HWT Spring Forecasting Experiment. *Wea. Forecasting*, **34**, 1395–1416, https://doi.org/10.1175/WAF-D-19-0056.1.

Rodney, M., H. Lin, and J. Derome, 2013: Subseasonal prediction of wintertime North American surface air temperature during strong MJO events. *Mon. Wea. Rev.*, **141**, 2897–2909, https://doi.org/10.1175/MWR-D-12-00221.1.

Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91**, 1015–1058, https://doi.org/10.1175/2010BAMS3001.1.

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

Sepulchre, P., and Coauthors, 2020: IPSL-CM5A2—An Earth system model designed for multi-millenial climate simulations. *Geosci. Model Dev.*, **13**, 3011–3053, https://doi.org/10.5194/gmd-13-3011-2020.

Song, X., and G. J. Zhang, 2018: The roles of convection parameterization in the formation of double ITCZ syndrome in the NCAR CESM: I. Atmospheric processes. *J. Adv. Model. Earth Syst.*, **10**, 842–866, https://doi.org/10.1002/2017MS001191.

Stan, C., D. M. Straus, J. S. Frederiksen, H. Lin, E. D. Maloney, and C. Schumacher, 2017: Review of tropical-extratropical teleconnections on intraseasonal timescales. *Rev. Geophys.*, **55**, 902–937, https://doi.org/10.1002/2016RG000538.

Stefanova, L., and Coauthors, 2022: Description and results from UFS coupled prototypes for future global, ensemble and seasonal forecasts at NCEP. NCEP Office Note 510, 201 pp., https://doi.org/10.25923/knxm-kz26.

Stensrud, D. J., and Coauthors, 2009: Convective-scale Warn-on-Forecast System: A vision for 2020. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1500, https://doi.org/10.1175/2009BAMS2795.1.

Sun, S., R. Bleck, S. G. Benjamin, B. W. Green, and G. A. Grell, 2018a: Subseasonal forecasting with an icosahedral, vertically quasi-Lagrangian coupled model. Part I: Model overview and evaluation of systematic errors. *Mon. Wea. Rev.*, **146**, 1601–1617, https://doi.org/10.1175/MWR-D-18-0006.1.

——, B. W. Green, R. Bleck, and S. G. Benjamin, 2018b: Subseasonal forecasting with an icosahedral, vertically quasi-Lagrangian coupled model. Part II: Probabilistic and deterministic forecast skill. *Mon. Wea. Rev.*, **146**, 1619–1639, https://doi.org/10.1175/MWR-D-18-0007.1.

Thompson, G., and T. Eidhammer, 2014: A study of aerosol impacts on clouds and precipitation development in a large winter cyclone. *J. Atmos. Sci.*, **71**, 3636–3658, https://doi.org/10.1175/JAS-D-13-0305.1.

——, P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, https://doi.org/10.1175/2008MWR2387.1.

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

WAVEWATCH III Development Group, 2016: User manual and system documentation of WAVEWATCH III version 5.16. NOAA/NWS/NCEP/MMAB Tech. Note 329, 326 pp. + appendixes, https://polar.ncep.noaa.gov/waves/wavewatch/manual.v5.16.pdf.

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

Yamagami, A., and M. Matsueda, 2020: Subseasonal forecast skill for weekly mean atmospheric variability over the Northern Hemisphere in winter and its relationship to midlatitude teleconnections. *Geophys. Res. Lett.*, **47**, e2020GL088508, https://doi.org/10.1029/2020GL088508.

Yang, Z.-L., and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 2. Evaluation over global river basins. *J. Geophys. Res.*, **116**, D12110, https://doi.org/10.1029/2010JD015140.

Zhou, L., S.-J. Lin, J.-H. Chen, L. M. Harris, X. Chen, and S. L. Rees, 2019: Toward convective-scale prediction within the next generation global prediction system. *Bull. Amer. Meteor. Soc.*, **100**, 1225–1243, https://doi.org/10.1175/BAMS-D-17-0246.1.

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

Zhu, H., M. C. Wheeler, A. H. Sobel, and D. Hudson, 2014: Seamless precipitation prediction skill in the tropics and extratropics from a global model. *Mon. Wea. Rev.*, **142**, 1556–1569, https://doi.org/10.1175/MWR-D-13-00222.1.

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