Diagnostics of Tropical Variability for Numerical Weather Forecasts

MARIA GEHNE,[a,b] BRANDON WOLDING,[a,b] JULIANA DIAS,[b] AND GEORGE N. KILADIS[b]

[a] *CIRES, University of Colorado Boulder, Boulder, Colorado*
[b] *Physical Sciences Laboratory, NOAA, Boulder, Colorado*

(Manuscript received 15 December 2021, in final form 25 May 2022)

ABSTRACT: Tropical precipitation and circulation are often coupled and span a vast spectrum of scales from a few to several thousands of kilometers and from hours to weeks. Current operational numerical weather prediction (NWP) models struggle with representing the full range of scales of tropical phenomena. Synoptic to planetary scales are of particular importance because improved skill in the representation of tropical larger-scale features such as convectively coupled equatorial waves (CCEWs) has the potential to reduce forecast error propagation from the tropics to the midlatitudes. Here we introduce diagnostics from a recently developed tropical variability diagnostics toolbox, where we focus on two recent versions of NOAA's Unified Forecast System (UFS): operational GFSv15 forecasts and experimental GFSv16 forecasts from April to October 2020. The diagnostics include space–time coherence spectra to identify preferred scales of coupling between circulation and precipitation, pattern correlations of Hovmöller diagrams to assess model skill in zonal propagation of precipitating features, CCEW skill assessment, plus a diagnostic aimed at evaluating moisture–convection coupling in the tropics. Results show that the GFSv16 forecasts are slightly more realistic than GFSv15 in their coherence between precipitation and model dynamics at synoptic to planetary scales, with modest improvements in moisture convection coupling. However, this slightly improved performance does not necessarily translate to improvements in traditional precipitation skill scores. The results highlight the utility of these diagnostics in the pursuit of better understanding of NWP model performance in the tropics, while also demonstrating the challenges in translating model advancements into improved skill.

KEYWORDS: Forecast verification/skill; Diagnostics; Model evaluation/performance;
Numerical weather prediction/forecasting; Tropical variability

1. Introduction

While significant improvements have been made in numerical weather prediction (NWP) model forecasts in recent decades (Bauer et al. 2015), synoptic-scale forecasts still tend to be less skillful in the tropics than in midlatitude regions (Zhu et al. 2014). This is in part related to the difference in the dominant dynamics between the two regions. A weaker Coriolis force and stronger insolation at low latitudes means that waves strongly coupled to convection are the main driver of synoptic-scale weather in the tropics, whereas large-scale rotational dynamics are dominant in midlatitudes. Further, tropical precipitation is not driven by baroclinic waves associated with strong high and low pressure regions as in midlatitudes, but rather happens more spontaneously in the form of localized convective precipitation. NWP therefore relies more strongly on convective parameterizations in the tropics, which further contributes to forecast uncertainty (Selz and Craig 2015).

Deficient tropical skill has far-reaching impacts that are not confined to local model errors. When localized convection becomes organized on large enough scales it can influence weather patterns in the tropics over several days and, through teleconnections, impact weather in midlatitudes (Branstator 2014; Stan et al. 2017; Dias and Kiladis 2019). Within the tropics the large-scale dynamics that vary on time scales of days to weeks are dominated by easterly waves and tropical depressions, convectively coupled equatorial waves (CCEWs) and the Madden–Julian oscillation (MJO). On longer time scales El Niño–Southern Oscillation changes the sea surface temperature and atmospheric circulation over the tropical Pacific impacting weather systems across the globe. There is substantial observational evidence that variability in midlatitudes is correlated at lag with large-scale convective activity in the tropics. The MJO in particular has been shown to interact strongly with the subtropical jet stream, the Pacific/North American Oscillation, and the North Atlantic Oscillation through teleconnections (e.g., Ferranti et al. 1990; Lin et al. 2009; Stan et al. 2017). Through its influence on the subtropical jet, the MJO can also impact the frequency of blocking events (Hamill and Kiladis 2014). Precipitation has been shown to be modulated by the MJO, with extreme extratropical rainfall events becoming more common during active phases (Jones et al. 2004). Past studies have shown 1) that convection in the tropics can influence midlatitude variability (Branstator 2014; Stan et al. 2017) and 2) that improved forecast skill in the tropics translates to improvements in midlatitude forecast skill several days later (Jung et al. 2010; Dias and Kiladis 2019). By nudging the tropical atmosphere toward reanalysis, and thus limiting errors in the tropics, studies such as Ferranti et al. (1990), Jung et al. (2010), Dias et al. (2021) show improvements in mean absolute error in 500-hPa geopotential height in the Northern Hemisphere midlatitudes at weeks 3 and 4. While the findings from tropical nudging experiments should be viewed as an upper bound of what can be achieved, they also highlight the potential for a reduction in tropical errors to improve midlatitude forecasts. Therefore, improving the representation of tropical sources of subseasonal to seasonal (S2S) predictability is an important step

Corresponding author: Maria Gehne, maria.gehne@noaa.gov

DOI: 10.1175/WAF-D-21-0204.1

toward realizing some of the remote potential skill at longer lead times.

The potential impact of improvements of forecast skill in the tropics can in practice be estimated by assessing the predictability associated with the MJO and CCEWs, since these are organized disturbances that could presumably be improved in models. Current NWP models struggle to predict these large-scale phenomena well at longer lead times (Dias et al. 2018; Bengtsson et al. 2019; Yang et al. 2021). While the reasons for the deficient skill at low latitudes are not fully understood, a number of studies suggest that improvements in CCEWs and MJO simulations could potentially lead to improved tropical skill (Li and Stechmann 2020; Judt 2020). As these phenomena involve coupling of the large-scale circulation to convection it is helpful to keep in mind two more general questions when evaluating model forecasts in the tropics: 1) Given the correct dynamic and thermodynamic evolution, is model convection responding realistically? and 2) Given that convection is occurring within a CCEW, does the model propagate the disturbance correctly? Of course, dynamics, thermodynamics and convection are highly coupled in the tropics and errors in one part of the system will feed back onto the others. Evaluating this coupling allows better insight in tropical model errors and improvements.

Developing process oriented diagnostics for global circulation models (GCMs) has been a community focus in recent years (Maloney et al. 2019). In the tropics, one class of diagnostics has to do with convective transition, or the statistics that characterize the probability density functions of column water vapor (CWV) for precipitating points, the pickup of precipitation as a function of column water vapor, and the dependence of the moisture–precipitation relation on tropospheric temperature. Those diagnostics indicate that for a model to do well with convective transition, the convective parameterization must capture multiple aspects of the triggering of deep convection. This requires that the dependence of the deep convective plume on lower free tropospheric humidity by entrainment is well represented (Kuo et al. 2020). There is evidence that improved thermodynamic convection coupling in GCMs is related to improvements in CCEW variability (Weber et al. 2021). For NWP it has been shown that better representation of the interaction between cumulus convection and large-scale tropical circulation improves organized convection and leads to more coherently propagating waves in the tropics (Bengtsson et al. 2019, 2021). Recent studies of CCEWs in NWP by Dias et al. (2018) and Yang et al. (2021) showed that the models considered have a tendency to decay CCEW amplitude and propagate CCEWs too quickly. This has implications for high impact weather forecasts in the tropics, as episodes of above normal precipitation tend to be associated with CCEWs.

The goal of this work is to introduce diagnostics for the tropics aimed mainly at phenomena with high potential predictability. These metrics are needed to assess NWP forecast skill of large spatial and lower-frequency phenomena such as CCEWs as well as to assess the representation of physical processes that are important for their initiation and maintenance. These diagnostics will also aid in identifying sources of model error in the tropics and in decision making as to whether changes in the models have the potential to improve model skill with respect to CCEWs and the MJO. The motivation for these diagnostics is to go beyond statistical measures of model skill to be able to assess model performance of convectively coupled phenomena known to have potential predictability of several days to weeks. The diagnostics introduced below fall into two categories: process-level diagnostics and forecast performance evaluation diagnostics that specifically target convective variability are larger spatiotemporal scales. This effort is related to the Model Diagnostics Task Force (MDTF) effort (Maloney et al. 2019) in its goal for process-based diagnostics, but the focus here is on NWP models and diagnostics as a function of lead time, as discussed for example in the Model Evaluation Tools (MET) software package hosted at NCAR (Brown et al. 2021). The nature of NWP means that for very short term forecasts (out to about 1–2 days) initial conditions are most important for model performance. Diagnostics with lead time dependence allow us to distinguish between model and initialization errors. Other considerations are that, in general, operational NWP forecasts are initialized once or several times a day and run out to a few weeks and that model versions change frequently. Therefore, long (multiyear) time series of a particular version of NWP forecast are not commonly available. If the diagnostics are to be used to inform operational model development, they need to take that into account. To complement the analysis presented in this paper, we also include a brief description of a python module containing all diagnostics (section 4e). This software is available as a stand-alone package and several of the diagnostics are also included in the May 2021 release of METcalcpy (Win-Gildenmeister et al. 2021).

The paper is organized as follows. Section 2 gives details on model versions and verification data used. Section 3 shows how the model versions perform using commonly used verification methods before applying the tropical diagnostics to the same model versions in section 4. The results are discussed in section 5.

2. Model data and observations

We use output from two versions of NOAA's Unified Forecast System (UFS) for comparison: The operational model version GFSv15.1 (Yang and Tallapragada 2018; Maxson 2019, GFSv15) during April–October 2020 and real-time parallel forecasts of the GFSv16 (Yang et al. 2020; Farrar 2021, GFSv16) for the same time period. Both model versions are initialized every 6 h and run out to lead time 240 h. The model output is regridded to a regular grid at 1° resolution.

One of the main differences between the model versions used here and the model version assessed in Dias et al. (2018) involves the dynamical core. The GFS version in Dias et al. (2018) study used the global spectral model which has since (starting with GFSv15) been replaced with the finite volume cubed sphere dynamical core (FV3; Harris et al. 2021). Both GFSv15 and GFSv16 are using the FV3 dynamical core in nonhydrostatic mode, with 768 grid cells on a cube sphere tile (C768), which corresponds to a global horizontal resolution of about 13 km. GFSv15 has 64 vertical layers and a model top at 54 km, and uses a physics suite consisting of the GFDL cloud microphysics scheme (Chen and Lin 2011, 2013; Zhou et al. 2019), a first-order turbulent transport scheme which is using a "hybrid" eddy diffusion mass flux (EDMF) approach (Han and Pan 2011; Han et al. 2017), and shallow and deep cumulus parameterizations which are originally based on Arakawa and Schubert (1974), but have over the years seen substantial updates following e.g., Pan and Wu (1995), Han and Pan (2011), and Han et al. (2017). Land surface processes in the GFSv15 physics suite are described by the Noah land surface model (Ek et al. 2003), and shortwave and longwave radiative fluxes and heating rates are parameterized using the RRTMG radiation scheme (Mlawer et al. 1997; Clough et al. 2005; Iacono et al. 2008). Finally, gravity wave drag is simulated as described by Alpert et al. (1988). The GFSv15 analysis is obtained through the Global Data Assimilation System (GDAS), which uses a hybrid four-dimensional ensemble variational formulation "hybrid 4DEnVar" (Kleist and Ide 2015).

The GFSv16 uses the same horizontal grid as GFSv15, but has increased vertical resolution consisting of 127 vertical layers and a model top at 80 km. Major changes in GFSv16 physics parameterizations compared with GFSv15 are the introduction of the scale-aware turbulent kinetic energy–eddy diffusion mass flux (TKE-EDMF) atmospheric boundary layer turbulence scheme (Han and Bretherton 2019) and the addition of parameterizations of subgrid-scale nonstationary gravity wave drag (Yudin et al. 2016, 2018). Several updates were also made in the data assimilation system, including the use of a four-dimensional incremental analysis update (4D-IAU) technique (Lei and Whitaker 2016), assimilation of new satellite observations and improved quality control. More details on GFS model changes can be found at https://www.emc.ncep.noaa.gov/gmb/STATS/html/model_changes.html (accessed 9 April 2021).

The diagnostics below are applied as a function of lead time to assess how model initialization on the one hand and model physics on the other are able to represent important aspects of tropical dynamics. Model output used for these diagnostics are precipitation, divergence (computed from winds), temperature, and specific humidity at all lead times and pressure levels between 1000 and 100 hPa.

Observational precipitation data used for verification are Integrated Multi-satellitE Retrievals for GPM V6 (IMERG; Huffman et al. 2019), the PERSIANN climate data record (CDR) (PERSIANNCDR; Ashouri et al. 2015; Sorooshian et al. 2014) and the Tropical Rainfall Measuring Mission (TRMM) 3B42 (Huffman et al. 2007; Tropical Rainfall Measuring Mission 2011). ERA5 reanalysis data (ERA5; Hersbach et al. 2020, 2018) are used for verification of dynamical fields, ensuring that the verification is based on an entirely independent system from the forecasts. All of these datasets were regridded to 6-hourly, 1° resolution except for PERSIANNCDR, which is daily. Single-level variables used are precipitation, surface pressure and land sea mask. Zonal wind, meridional wind, temperature, and specific humidity are used at vertical levels between 1000 and 100 hPa.

3. Commonly used forecast verification methods

NWP verification routinely uses statistical metrics to assess model skill (Brown et al. 2021). Below we apply some of the most commonly used metrics to the GFS forecasts and compare model skill in midlatitudes and in the tropics, using IMERG precipitation and ERA5 for verification. While this is not the same setting as in operational verification, it allows a consistent comparison of the performance of the two model versions. All metrics are computed for two latitude bands 20°S–20°N (tropics) and 35°–50°N (Northern Hemisphere midlatitudes). The metrics are included here to illustrate how model performance with respect to these more standard metrics is not necessarily an indication of performance with respect to the tropical diagnostics metrics introduced in section 4.

a. *Precipitation*

1) Equitable threat score

The equitable threat score (ETS) measures the fraction of observed events that were correctly predicted, adjusted for hits associated with random chance (e.g., it is easier to correctly forecast rain occurrence in a wet climate than in a dry climate). The ETS is often used in the verification of rainfall in NWP models because its "equitability" allows scores to be compared more fairly across different climate regimes. Because it penalizes both misses and false alarms in the same way, it is less useful for distinguishing the source of forecast errors (Wilks 2011; Jolliffe and Stephenson 2012).

Figures 1a–c show that both models perform worse in the tropics at short lead times than in midlatitudes for different precipitation percentiles (50, 75, 95), where the 95th percentile refers to the top 5% of precipitation totals during the period over a region (tropics and Northern Hemisphere midlatitudes here) and correspondingly for the other percentiles. Tropical performance initially decays much more rapidly than in the midlatitudes, but at leads longer than around 5 days this reverses, consistent with previous results by Zhu et al. (2014). GFSv16 ETS is comparable or slightly improved compared to GFSv15 for both midlatitudes and tropics.

2) Frequency bias

Precipitation frequency bias (FBias) measures the ratio of the frequency of forecast events to the frequency of observed events, indicating whether the forecast system has a tendency to underforecast (FBias < 1) or overforecast (FBias > 1) events. FBias does not measure how well the forecast amplitude corresponds to the observations, but only measures relative frequencies.

Figures 1d–f show that the FBias is improved in GFSv16 in the 95th percentile and deteriorated for the lower percentiles, with GFSv16 overforecasting precipitation frequency in the tropics and underforecasting it in midlatitudes.

3) Fractions skill score

The fractions skill score (FSS) compares the forecast and observed precipitation fraction over increasing areas. It ranges from 0 (complete mismatch) to 1 (perfect match). The value of FSS above which the forecasts are considered to have

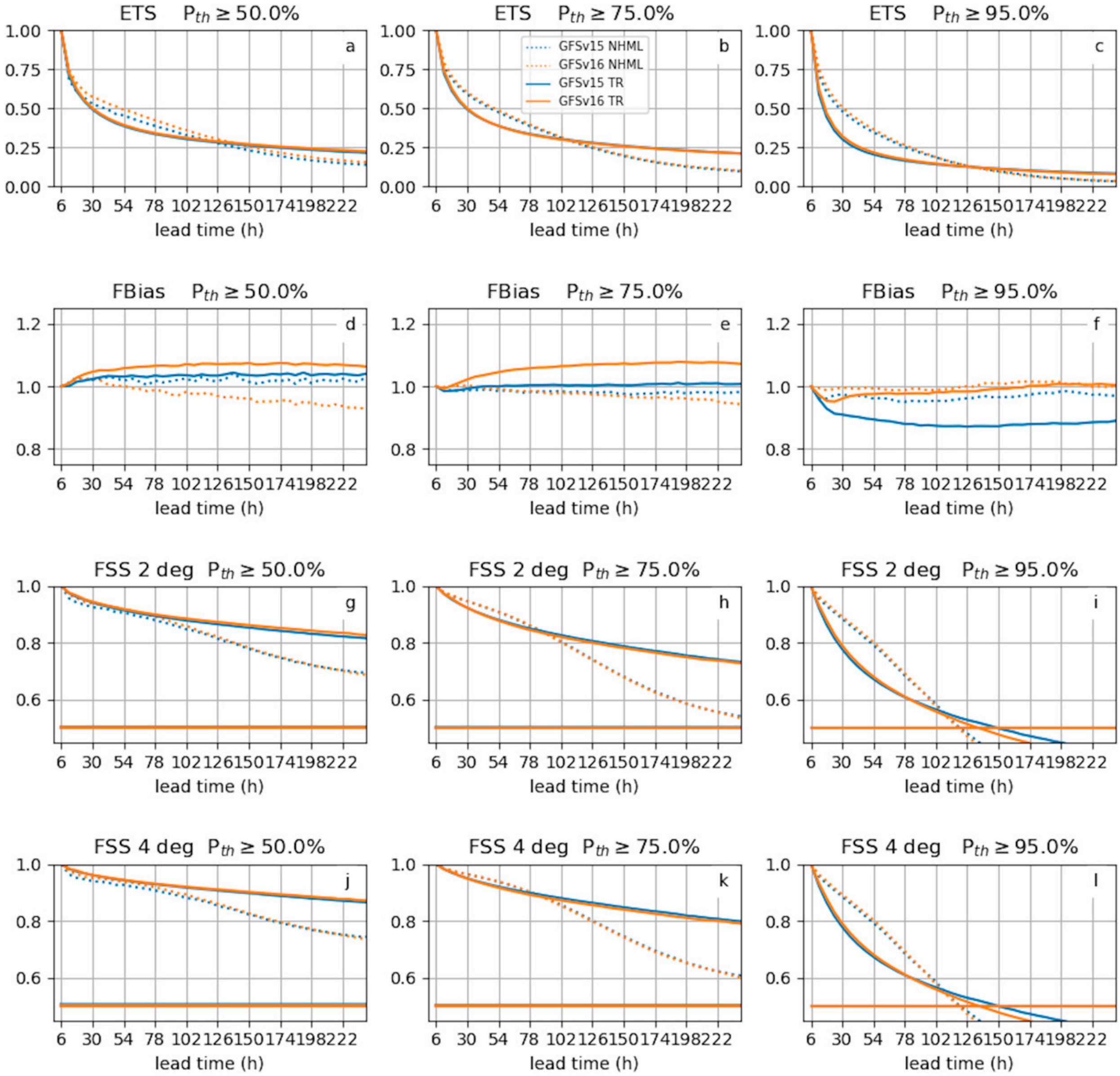

FIG. 1. Precipitation skill scores for tropics (20°S–20°N) and Northern Hemisphere midlatitudes (35°–50°N). Shown are (a)–(c) equitable threat score, (d)–(f) frequency bias, and fraction skill score for area sizes of (g)–(i) 2° squared and (j)–(l) 4° squared for GFSv15 (blue lines) and GFSv16 (orange lines) for tropics (20°S–20°N, solid lines) and Northern Hemisphere midlatitudes (35°–50°N, dotted lines). Horizontal lines indicate the value of $\mathrm{FSS}_{\mathrm{useful}}$ for both model versions.

better than random skill is given by $\mathrm{FSS}_{\mathrm{useful}} = 0.5 + f_o/2$, where f_o is the domain average observed fraction. The smallest area for which $\mathrm{FSS} \geq \mathrm{FSS}_{\mathrm{useful}}$ can be considered the "skillful scale," this will depend on lead time as well. As the area used to compute the fractions gets larger, the score will asymptote to a value that depends on the ratio between the forecast and observed frequencies of the event. The closer the asymptotic value is to 1, the smaller the forecast bias. The score is most sensitive to rare events (e.g., small rain areas).

The tropical FSS is lower than midlatitude FSS for a given area and across percentiles until about 4–5-day lead time for both GFSv15 and GFSv16 (Figs. 1g–l). For longer lead times the tropical FSS eventually outperforms FSS in midlatitudes, with the cross-over happening at earlier lead times for lower percentiles. The decrease in FSS is steeper initially for precipitation forecasts in the tropics than in midlatitudes. In the tropics, FSS is slightly improved in GFSv16 over GFSv15 for higher percentile thresholds and shorter lead times.

b. Tropospheric and near-surface verification statistics

Next we consider verification metrics that assess skill of dynamic and thermodynamic variables in the troposphere.

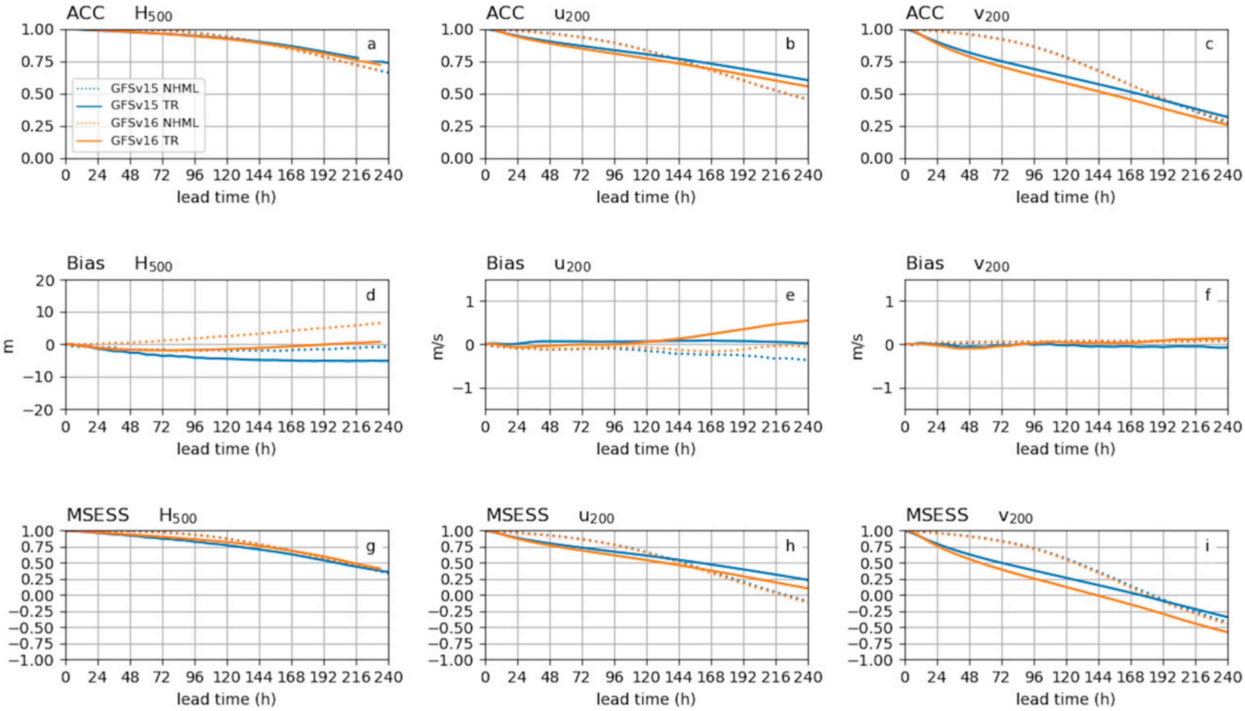

FIG. 2. Tropospheric skill scores for GFSv15 (blue lines) and GFSv16 (orange lines) tropics (20°S–20°N, solid lines) and Northern Hemisphere midlatitudes (35°–50°N, dashed lines). Shown are (a)–(c) anomaly correlation, (d)–(f) bias, and (g)–(i) mean square error skill score for (left) 500-hPa geopotential height, (center) 200-hPa zonal wind, and (right) 200-hPa meridional wind.

The metrics considered here are anomaly correlation (ACC), bias and mean square error skill score (MSESS). These metrics are described in more detail below.

Results are shown below for the same latitude bands as above and the tropospheric variables: 500-hPa height (h500) and 200-hPa winds (u200, v200). We also considered winds at 850 hPa with qualitatively similar results. Near-surface variables considered here are 2-m air temperature (T2m), 2-m relative humidity (q2m), and 10-m zonal wind (u10m).

1) ANOMALY CORRELATION

The ACC is computed by first removing the time mean of the verification (the analysis in this case) and the time mean of the forecast at each lead time. This removes the impact of model bias from model drift. The anomalies are then used to compute a pattern correlation between the forecast and verification at each time and the result is averaged over all times.

Figure 2a shows that for both model versions ACC for h500 is slightly lower in the tropics than midlatitudes until lead time 120 h, after which the ACC in the tropics is higher. However, since tropical temperature gradients are weak, h500 does not have large variability within the tropics, making the ACC metric not very informative in that region. For u200 and v200 (Figs. 2b,c), ACC is deteriorated in GFSv16 in the tropics and comparable to GFSv15 in midlatitudes.

Figures 3a,c show comparable skill in both model versions in T2m and U10m ACC in the tropics with the tropics having lower ACC out to 96-h lead time and higher ACC after that. GFSv16 skill in midlatitudes is improved after 5-day lead time. Figure 3b for q2m shows comparable ACC in GFSv16 compared to GFSv15 in both regions. We also considered land points only and observed very similar behavior (not shown). In general, the near-surface variables are not well constrained by the model but are important forecast variables for end users, which is the reason we show them here.

2) BIAS

To compute the mean bias the verification mean is removed from the forecast and then the time average is taken at each lead time. Bias is improved for h500, u200, v200 in the tropics (Figs. 2d–f) in model version GFSv16 until at least 5-day lead time. This is similar for bias in T2m, q2m, and u10m (Figs. 3d–f). Considering land only, bias is improved in GFSv16 for q2m in both tropics and mid latitudes, bias is improved for q2m in the tropics until 72-h lead time and increases for T2m and u10m.

3) MEAN SQUARE ERROR SKILL SCORE

The MSESS is computed based on the MSE at each lead time and the verification variance is used as an estimate of the climatological MSE (Murphy 1988; Jolliffe and Stephenson 2012).

MSESS is lower in the tropics than midlatitudes for both model versions for h500, u200 and v200 (Figs. 2g–i) and GFSv16 performs worse in the tropics and comparable in midlatitudes. MSESS performance for the near-surface variables

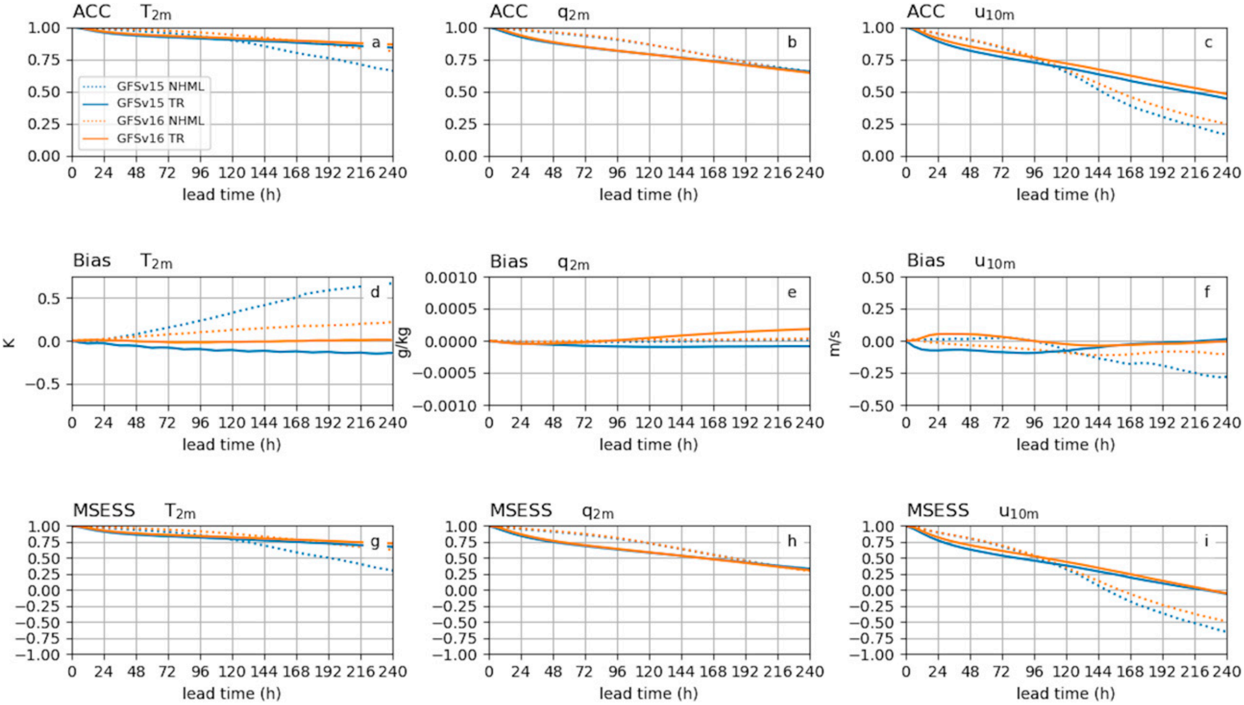

FIG. 3. Surface skill scores for GFSv15 (blue lines) and GFSv16 (orange lines) tropics (20°S–20°N, solid lines) and Northern Hemisphere midlatitudes (35°–50°N, dashed lines). Shown are (a)–(c) anomaly correlation, (d)–(f) bias, and (g)–(i) mean square error skill score for (left) 2-m temperature, (center) 2-m specific humidity, and (right) 10-m zonal wind.

(Figs. 3g–i) is similar to that seen for the near-surface ACC. Land only MSESS is very similar to the results shown in Figs. 3g–i as well.

Summarizing the results of this section, tropical forecast skill tends to lag midlatitude skill early in the forecast when skill tends to be high based on metrics commonly used for verification. At longer lead times skill in the tropics tends to be higher than midlatitude skill, but overall skill is low.

4. Tropical diagnostics

Biases in precipitation means and variances, when verified against IMERG, tend to be large where precipitation is large and where model precipitation is dominated by parameterized and not large-scale precipitation (Fig. 4). Precipitation errors in the tropics are largest in the intertropical convergence zone (ITCZ) and South Pacific convergence zone regions with model mean precipitation exceeding IMERG. Both model versions and ERA5 underestimate mean precipitation near the equator in the central Pacific. This error pattern leads to more of a double ITCZ pattern in this region than is seen in IMERG. Model and ERA5 variance are smaller than IMERG precipitation variance (not shown). This, together with the larger mean precipitation indicates that too much of the model precipitation falls at rates that are too small compared to IMERG, a common issue with model and reanalysis precipitation (Stephens et al. 2010; Pendergrass and Hartmann 2014; Gehne et al. 2016). This can be seen in Figs. 4d,f,h,j,l where the frequency of occurrence of light precipitation exceeds IMERG over most ocean regions and by about 30% over the subtropical oceans. While the mean errors are smaller for GFSv16 at lead time 6h (Figs. 4e,g), there is an increase in the occurrence of light rain compared to GFSv15 (Figs. 4f,h).

Based on the considerations above and in the previous section, we argue that having diagnostics available that can help in identifying forecast error sources in the tropics related to moisture–convection–circulation coupling will be highly beneficial. In the following sections we introduce several diagnostics that assess the coupling between dynamics and convection to better understand model dynamics and errors in the tropics and how to improve them. This is by no means an exhaustive suite of possible diagnostics for tropical convection, but rather an adaptation to NWP evaluation of diagnostics that have been successfully used in both observational studies and climate model evaluation.

a. *Hovmöller diagrams and pattern correlation*

Hovmöller diagrams are time–longitude plots of latitude band averages and were first introduced to analyze troughs and ridges in 500-hPa height observations in midlatitudes (Hovmöller 1949). Since then they have been widely used in NWP and for identifying zonal propagation characteristics of large-scale tropical phenomena (e.g., Kiladis et al. 2009; Persson 2017; Dias et al. 2018).

Figures 5a–d show precipitation averaged between 10°S and 10°N for IMERG, ERA5, and model precipitation at 6h lead time. Upon initial inspection, the correspondence

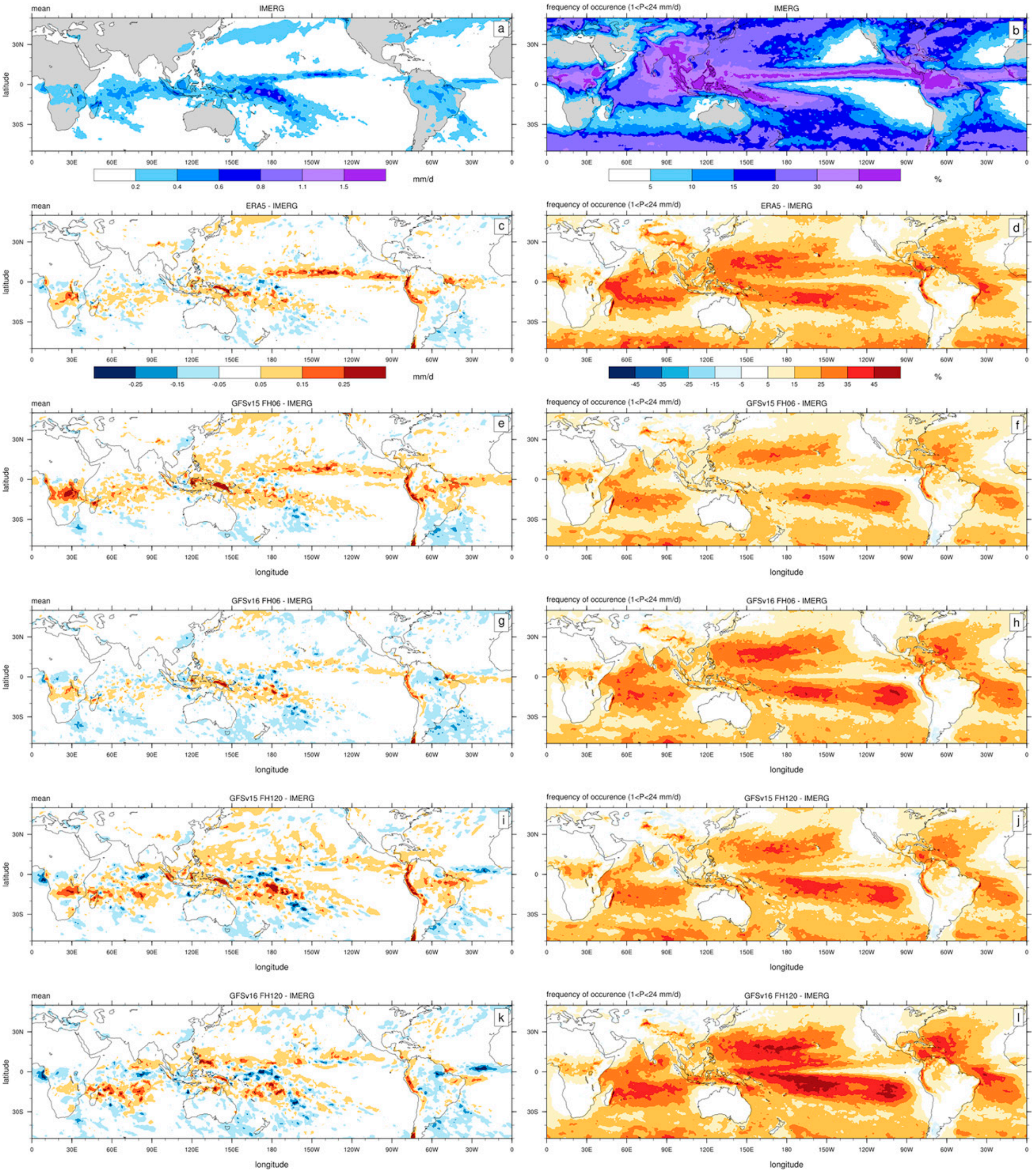

FIG. 4. Precipitation mean and frequency of occurrence of light rain (1–24 mm day^{-1}) from April through October 2020 for (a) IMERG mean and (b) IMERG frequency of occurrence of precipitation rates $1 < P < 24$ mm day^{-1}. Differences with IMERG means are shown for (c) ERA5 precipitation, (e) GFSv15 FH06, (g) GFSv16 FH06, (i) GFSv15 FH120, and (k) GFSv16 FH120. Difference with IMERG frequency of light precipitation occurrence are shown for (d) ERA5 precipitation, (f) GFSv15 FH06, (h) GFSv16 FH06, (j) GFSv15 FH120, and (l) GFSv16 FH120.

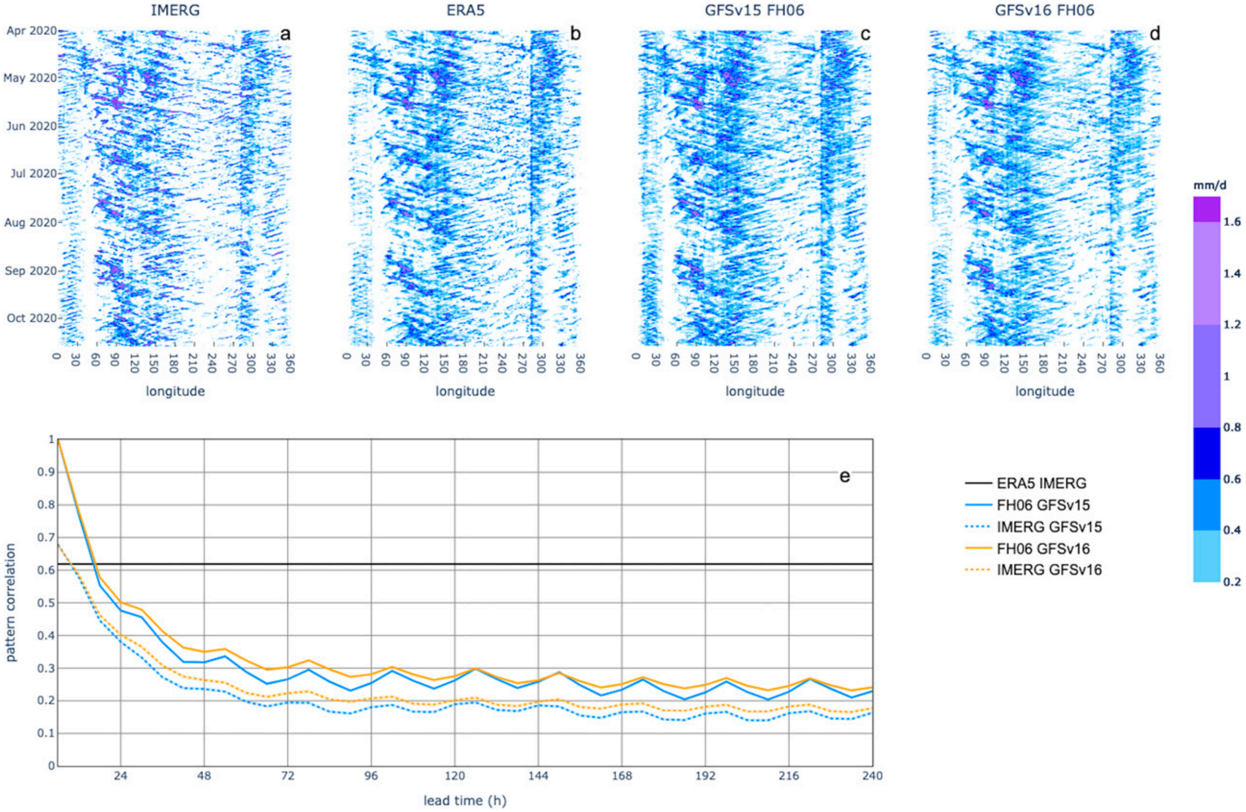

FIG. 5. Hovmöller diagrams of 6-hourly precipitation averaged from 10°S to 10°N for (a) IMERG, (b) ERA5, (c) GFSv15, and (d) GFSv16. (e) Pattern correlation of latitude averages (10°S–10°N) between ERA5 and IMERG (black horizontal line), GFSv15 and GFSv15 FH06 (solid blue), GFSv15 and IMERG (dotted blue), GFSv16 and GFSv16 FH06 (solid orange), and GFSv16 and IMERG (dotted orange) are shown. The 95% confidence intervals are shown in shading.

between model precipitation and IMERG/ERA5 is high and large-scale eastward-propagating disturbances can easily be identified by eye in all four panels. The larger areas with light blue shading indicate that both models and ERA5 precipitation (Figs. 5b–d) tend to have more light precipitation than IMERG (Fig. 5a).

Figure 5e shows the pattern correlation of model precipitation averaged between 10°S and 10°N with model precipitation at 6-h lead time and IMERG precipitation for the entire evaluation period. Confidence intervals are computed by random subsampling of the data 1000 times, computing the pattern correlation, and picking the top and bottom 2.5% percentiles. The correlation between IMERG and ERA5 precipitation is 0.63 and does not change with lead time. The moderate correlations between IMERG and ERA5 likely stem from the lack of direct assimilation of satellite observations of rain rates in ERA5 in combination with the substantial impact of model physics in reanalysis rain rates. Initially model correlations are highest with the precipitation at 6-h lead time, are comparable with the IMERG-ERA5 correlation between 6 and 18 h and drop below 0.5 by 30-h lead time for all cases. The pattern correlation is higher for GFSv16 precipitation for all cases and correlations with precipitation at 6-h lead time drop at a slightly slower rate for GFSv16.

Overall GFSv16 shows higher correlation with its own precipitation at 6-h lead time, ERA5 (not shown), and IMERG precipitation. The precise reasons for this would be challenging to pinpoint because, in addition to changes in model physics related to clouds and precipitation, the transition from GFSv15 to GFSv16 also included changes in vertical resolution and in the data assimilation system. What is clear is that much potential skill in precipitation forecasts is already lost during the first few hours after initialization. Even provided with our best estimate of the dynamical and thermodynamic state of the system, the model loses a lot of forecast skill because model convection is not producing the correct quantity or spatial distribution of precipitation.

While the fast drop in initial skill might be partially related to drift toward the models' own climatology and/or chaotic processes, the next few sections show that much of the tropical skill loss is likely related to biases in how the models couple convection to dynamics and thermodynamics along with struggling to evolve those aspects of the system correctly.

b. CCEW activity and skill

Identifying convectively coupled equatorial waves (CCEWs) in forecasts has been done following several different approaches. These include padded filtering of model forecasts

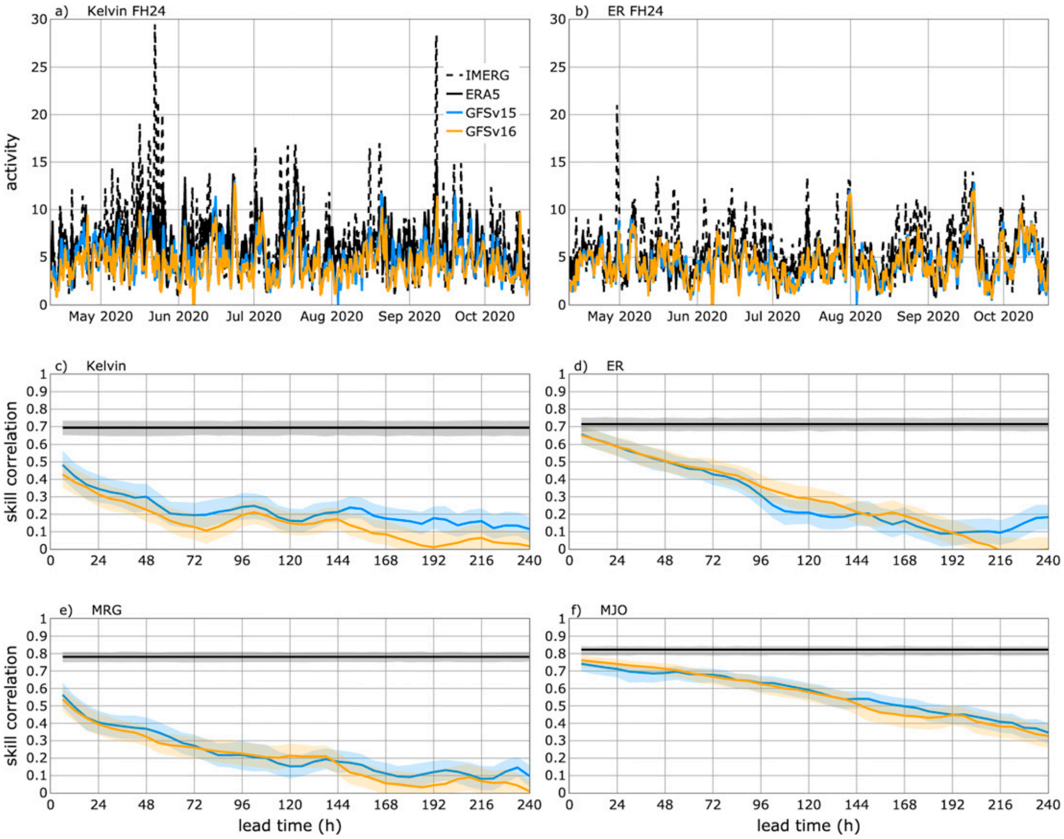

FIG. 6. CCEW activity and skill compared to IMERG precipitation (black dashed) for ERA5 (black solid), GFSv15 (blue), and GFSv16 (orange). (a) Kelvin activity and (b) ER activity for the period December 2019–March 2020. Skill correlation between model activity and IMERG activity for (c) Kelvin activity, (d) ER activity, (e) MRG activity, and (f) MJO activity. In (c)–(f) the skill correlation between IMERG and ERA5 is shown at all lead times (black solid) for comparison. The 95% confidence intervals are shown in shading.

(Janiga et al. 2018; Schreck et al. 2020), projection onto spatial patterns derived from observations (Gottschalck et al. 2010) and projection of model forecasts onto theoretical spatial structures (Yang et al. 2021). Here we use an approach similar to the one applied by Gottschalck et al. (2010) for the MJO where we derive empirical orthogonal functions (EOFs) based on IMERG precipitation data as the basis for assessing CCEW activity. This approach has the advantage of avoiding the application of relatively low-frequency temporal filters to short model forecast time series. We define EOF based indices for Kelvin, equatorial Rossby (ER), mixed Rossby–gravity (MRG) waves, and the MJO. While there are already MJO indices and EOFs available that are based on OLR and winds such as the RMM and OMI (Wheeler and Hendon 2004; Kiladis et al. 2014), for consistency with the CCEW skill assessment we use precipitation EOFs for the MJO here. See appendix A for more details.

CCEW activity (Figs. 6a,b) are shown for Kelvin and ER waves and skill (Figs. 6c–f) for Kelvin, ER, MRG and the MJO. ERA5 skill for the MJO and MRG waves is about 0.8, but only about 0.7 for Kelvin and ER waves. This relatively low skill for Kelvin and ER waves even for ERA5 could be due to only a few strong events that occurred during this time period (Fig. 6a). There are a few strong ER events (Fig. 6b), which appear to be equally well captured by ERA5, GFSv15, and GFSv16. This is also evident in the model skill correlation where Kelvin skill is below 0.5 by 12-h lead time (Fig. 6c), while MJO skill stays above 0.5 past 5-day lead time (Fig. 6f). GFSv16 has slightly higher skill correlation values for the first 24 h into the forecast for ER and MJO.

Performance of GFSv16 is slightly improved over GFSv15 for ER and MJO in this diagnostic during the first 48 h of the forecast. Skill for these wave types for GFSv16 is comparable to IMERG-ERA5 correlation until 12-h lead time. It is interesting to note that the skill correlations are much higher initially than for the Hovmöller pattern correlations (Fig. 5). This is conceivably due to the EOFs picking up larger zonal scales of variability which the models can forecast more robustly (at least at short lead times) than the smaller scales. Considering EOFs for OLR and zonal wind at 200 and 850 hPa, computed by regressing ERA5 winds and observed OLR onto the precipitation principal components, we see similar behavior (not shown). Skill for OLR is higher than precipitation for

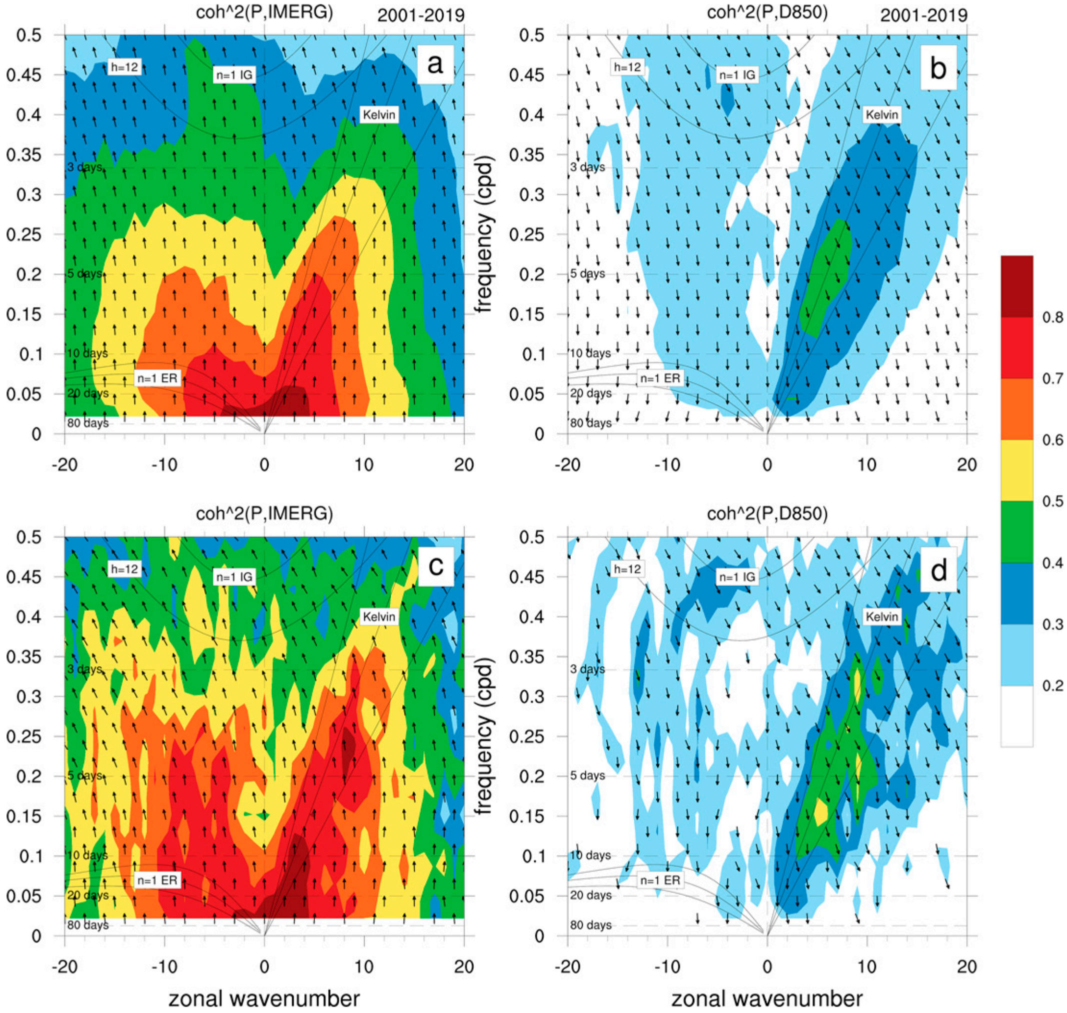

FIG. 7. Symmetric space–time coherence spectra 15°S–15°N for (a) ERA5 and IMERG precipitation from 2001 to 2019, (b) ERA5 precipitation and ERA5 divergence at 850 hPa from 2001 to 2019, (c) ERA5 and IMERG precipitation from April to October 2020, and (d) ERA5 precipitation and ERA5 divergence at 850 hPa from April to October 2020. Shading shows coherence squared, and arrows show the phase between the two variables. Lines show theoretical linear shallow-water dispersion curves for equivalents heights of 12, 25, and 50 m. The window length for the spectra was 46 days with a 20-day overlap. There are no data at periods longer than 46 days. Phase arrows are oriented as follows: up arrow—variables are in phase, right arrow—first variable leads, down arrow—variables are out of phase, and left arrow—second variable leads.

Kelvin (by 0.2) and ER (by 0.15) and comparable for MRG and MJO, while skill for zonal winds is much higher, initially above 0.9. There is no significant difference between the model versions except that GFSv16 ER u200 outperforms GFSv15 after lead time 144 h and GFSv15 Kelvin u200 outperforms GFSv16 after lead time 168 h.

c. *Space–time coherence spectra*

Coherence-squared wavenumber-frequency spectra between two variables highlight temporal and spatial scales where the two variables have significant correlation. Coherence-squared spectra are computed at each latitude by performing a two dimensional fast Fourier transform on both variables of interest to generate the two-dimensional frequency–zonal wavenumber power, quadrature, and cospectra. Latitudinal averages (either symmetric or antisymmetric across the equator) are then taken (Wheeler and Kiladis 1999; Hendon and Wheeler 2008). We only show results for the symmetric spectra, but the same general conclusions apply to the antisymmetric spectra as well. The coherence squared is the sum of the squared co and quadrature spectra divided by the product of the two power spectra. As an example, Fig. 7a shows the long-term coherence squared between ERA5 and IMERG precipitation. If there was perfect agreement between the reanalysis and satellite precipitation the coherence squared values would be equal to 1 and the phase arrows would point straight up. Instead, Fig. 7a shows that ERA5 and IMERG precipitation agree more along the CCEW dispersion curves and low wavenumbers and less at higher

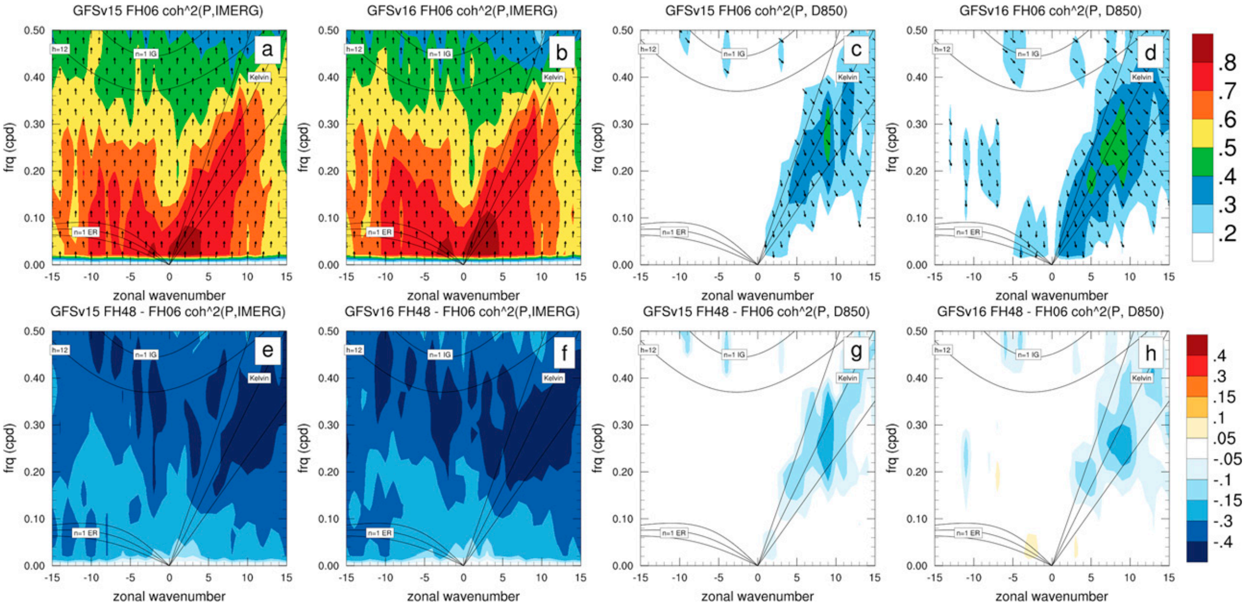

FIG. 8. Symmetric space–time coherence spectra 15°S–15°N between (a) GFSv15 precipitation at FH06 (FH = forecast hour) and IMERG precipitation, (b) GFSv16 precipitation at FH06 and IMERG precipitation, (c) GFSv15 precipitation and divergence at 850 hPa (D850) at FH06, and (d) GFSv16 precipitation and D850 at FH06. Shading shows coherence squared, and arrows show the phase between the two variables. Difference of coherence squared at FH48 and coherence squared at FH06 is shown for (e) GFSv15 precipitation and IMERG precipitation, (f) GFSv16 precipitation and IMERG precipitation, (g) GFSv15 precipitation and D850, and (h) GFSv16 precipitation and D850.

frequencies. In addition, at higher frequencies IMERG tends to lead ERA5 precipitation as indicated by the phase arrows pointing slightly to the left of straight up. Figure 7b shows the long-term symmetric coherence-squared spectrum between ERA5 divergence at 850 hPa and precipitation. Large values of coherence-squared can be seen along theoretical dispersion curves for Kelvin, ER, MRG, westward inertio-gravity (WIG) waves, and the MJO. Because of the 46-day low-frequency cutoff (chosen for consistency with the model spectra below), the MJO is only marginally resolved. Phase arrows show that low-level convergence leads precipitation maxima by an 8th of a cycle for the higher-frequency waves and becomes more in phase for lower frequencies. For the shorter period waves such as WIGs and MRGs this translates to 850-hPa convergence leading precipitation by 6–9 h. Close inspection shows that this holds for the longer period waves as well, but because the lifetime of those disturbances is much longer, the lag represents a much smaller fraction of the cycle.

To demonstrate the effect of using a short time period Figs. 7c and 7d show the same observed coherence spectra, but for the model verification period from April to October 2020. The smaller sample size leads to much noisier spectra. However, the small sample is still able to resolve peaks along the Kelvin wave dispersion curves and regions of higher coherence values indicating WIG and ER activity.

Next, we compute coherence spectra of the forecast time series by lead time. For each lead time we have a continuous 6-hourly time series that spans the verification period. Changes in coherence spectra with increasing lead time indicates changes in the strength of the coupling between precipitation and dynamics. Initially larger coherence values tend to be located near CCEW dispersion curves and at lower frequencies and larger spatial scales (Figs. 8a,b). This indicates that model precipitation in both GFSv15 and GFSv16 in the first 12–24 h past initialization is largely able to initialize and maintain large-scale CCEW events. Comparing Figs. 8a and 8b to Fig. 7c, it appears that the model tends to have peaks at slightly higher frequencies than the reanalysis and observations. The coherent evolution of observed and modeled precipitation decreases rapidly with lead time, which is shown as differences in coherence between lead times of 48 and 6 h in Figs. 8e and 8f. This is likely related to the model propagating convectively coupled phenomena at the wrong speed along with the model not being able to maintain those phenomena for long lead times. Based on the phase relationship, modeled precipitation leads observed precipitation indicating that the model is propagating convectively coupled phenomena too quickly. The decrease in coherence squared from 6- to 48-h lead time is most pronounced in the regions of CCEW dispersion curves and higher frequencies and wavenumbers.

In addition, coherence spectra not only allow insight in how the model represents the coupling between dynamics and precipitation, but also whether the coupling happens at the correct scales (Figs. 8c,d). The phase relationship in regions with significant coherence can be used to infer whether large-scale dynamics drive precipitation in a manner consistent with observations.

Coherence between precipitation and divergence is stronger at 850 hPa than at 200 hPa for ERA5 (not shown). The same is true for the model at initial time (Figs. 8c,d) and while there are distinct peaks in coherence along CCEW dispersion curves, overall the model coherence tends to be lower. By 48-h lead time GFSv15 shows decreased coherence between precipitation and 850-hPa divergence and the two distinct peaks in the Kelvin wave band have decreased by 50%–75% (Fig. 8g). In contrast, model version GFSv16 initially has stronger coherence between precipitation and 850-hPa divergence and is still able to represent at least the lower-frequency portion of the Kelvin wave peak at 48-h lead time (Fig. 8h) and the decrease in coherence squared is less than for GFSv15. At 200 hPa GFSv15 tends to have too strong coherence-squared in the ER region compared to ERA5, which is alleviated in GFSv16 but again, the coherence-squared values decrease with lead time (not shown).

These diagnostics suggests that while both model versions are able to initialize CCEWs, the coupling between moisture and dynamics is too weak even at initial time. In addition, at longer lead time precipitation is not coupled strongly to the near-surface dynamics, although this is improved in GFSv16 which starts out with stronger coupling initially and has a slower decrease in coherence squared at longer leads. The diagnostic also shows that both model versions propagate CCEWs too quickly, as was suggested in the results of section 4a above and in previous versions of the GFS (Dias et al. 2018; Bengtsson et al. 2019). Figure 8 shows almost no coherence at very high frequencies. Variability at higher frequencies and wavenumbers does not contribute much to S2S predictability although this activity could still be a source of feedback to the larger scales (Garfinkel et al. 2021). This may be an indication of a source for these differences not accounted for by the traditional statistical metrics that are perhaps capturing more of the high-frequency/small-scale error versus larger spatiotemporal scale error. In addition, if the traditional metrics are capturing more of the high-frequency error, that makes them potentially less relevant for identifying skill on S2S time scales.

Next we examine the vertical structure of coherence-squared within CCEW bands. Based on the decay in coherence with lead time for the Kelvin wave band shown in section 4c, we focus on Kelvin waves here, but the approach for other waves is analogous. Similarly to the coherence squared at a single level shown above we compute coherence squared spectra of wave filtered precipitation with dynamical and thermodynamical variables at all vertical levels. A detailed description of the technique can be found in appendix B.

Figure 9 shows the vertical structure of coherence-squared of Kelvin filtered 6-hourly precipitation with dynamical fields during the verification period. The vertical structure of the Kelvin wave coherence squared is very similar when using either IMERG or ERA5 precipitation and ERA5 dynamical fields (Figs. 9a,b) with stronger overall coherence for the latter. Divergence has two peaks in coherence-squared with precipitation, one around 950–900 hPa, the main inflow level, and one just above 200 hPa near the outflow level of deep convection that spans the depth of the troposphere. The phase relation shows that precipitation lags convergence at the lower peak and leads divergence at the upper peak, both by less than 1/8 of a cycle. A deep layer of high coherence-squared between zonal wind and precipitation exists from the surface to about 750 hPa with precipitation leading (lagging) a maximum (minimum) in zonal wind. This indicates that zonal advection throughout this layer may play a role in the moisture build-up prior to deep convection in the CCEW Kelvin wave. Temperature shows strong coherence-squared with precipitation near the surface and a secondary peak between 650 and 600 hPa, both with precipitation leading a minimum in temperature. The peak at 650–600 hPa is likely in part related to cooling due to melting just below the freezing level. Specific humidity and precipitation coherence-squared has a peak between 500 and 200 hPa with precipitation lagging specific humidity from the surface to 650 hPa and leading above that. These phase relationships between precipitation and humidity are well-known features of convectively coupled waves (e.g., Kiladis et al. 2009). Meridional wind does not have significant coherence-squared with Kelvin filtered precipitation. Comparing Figs. 9a and 9b demonstrates that while the general vertical profiles are consistent, establishing clear benchmarks/targets for these coherence profiles may not be feasible. Uncertainty between Figs. 9a and 9b is larger than between Figs. 9b and 9c or 9d.

Both model versions at 6-h lead time (Figs. 9c,e) have similar overall vertical structures when compared to the verification. Main differences are that the height of the secondary peak with temperature is slightly lower (around 750–700 hPa). The low-level divergence peak in coherence-squared for GFSv15 and GFSv16 is stronger than the upper-level peak, similar to ERA5. The coherence-squared peak for specific humidity is not as well defined as in ERA5 and extends from about 750–300 hPa. The main difference between coherence-squared of model dynamical variables with IMERG or model precipitation at 6-h lead time is that the low-level peak in divergence tends to be slightly stronger when using model precipitation, suggesting stronger coupling between model precipitation and model divergence (not shown).

By 48-h lead time (Figs. 9d,f) the maximum coherence-squared values for the divergence peaks have decreased. This decrease is more pronounced when using IMERG instead of model precipitation (not shown). The zonal wind coherence-squared peak is shallower, from the surface to about 900 hPa. Temperature and specific humidity show a general decrease in coherence-squared, but the vertical structure is still maintained.

Taken together these observations indicate that the model is able to initialize the vertical structure of the CCEW Kelvin wave as identified from IMERG precipitation and ERA5 and propagate that information into the forecast for a few days. By 120 lead time (not shown) the near surface divergence is in-phase with precipitation instead of leading by 1/8 of a cycle. Overall the vertical structure of coherence-squared observed near model initialization is increasingly washed out with lead time. This lack of coherent structure between precipitation and model dynamics and thermodynamics is without a doubt one root cause of the decrease in model skill in predicting CCEWs.

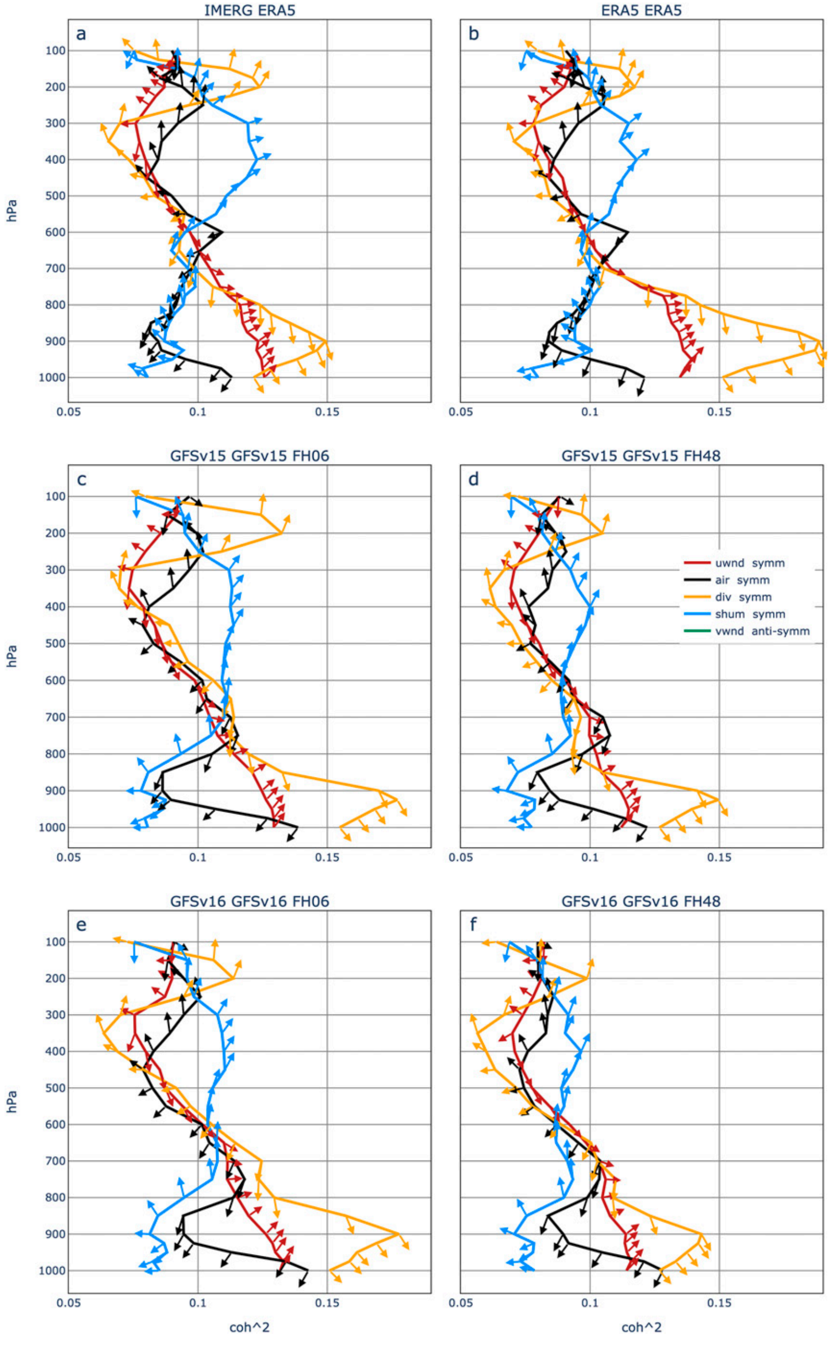

FIG. 9. Vertical structure of coherence squared from 15°S to 15°N between Kelvin-filtered 6-hourly precipitation and dynamical variables for April–October 2020. Wavenumber–frequency-averaged coherence squared for (a) IMERG precipitation and ERA5 dynamical fields, (b) ERA5 precipitation and ERA5 dynamical fields, (c) GFSv15 precipitation and dynamical fields at FH06, (d) GFSv15 precipitation and dynamical fields at FH48, (e) GFSv16 precipitation and dynamical fields at FH06, and (f) GFSv16 precipitation and dynamical fields at FH48. Vertical profiles are shown for zonal wind (red), temperature (black), divergence (yellow), and specific humidity (blue). Wavenumber–frequency cross spectra are computed between 15°S and 15°N and significant values of coherence squared are averaged at each vertical level. Arrows at each level describe the phase relationship between precipitation and the dynamical variables as follows: upward arrow—in phase, right arrow—precipitation leads, downward arrow—out of phase, and left arrow—precipitation lags.

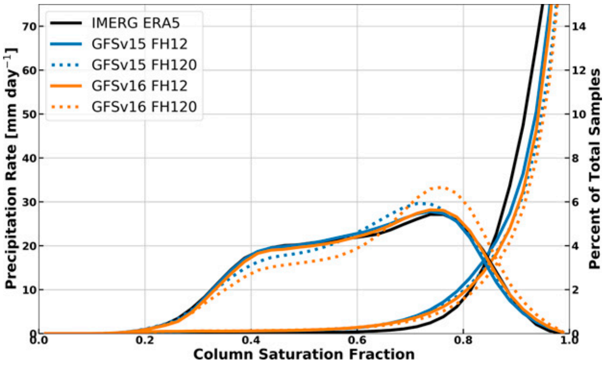

FIG. 10. Distribution of column saturation fraction (CSF) on the right y axis and CSF conditionally averaged precipitation rates on the left y axis for IMERG precipitation and ERA5 CSF (black), GFSv15 FH12 (solid blue), GFSv16 FH12 (solid orange), GFSv15 FH120 (dotted blue), and GFSv16 FH120 (dotted orange). Distributions are computed for all dates during the verification period. The CSF distribution is the CSF probability density distribution given as percentage of total samples, whereas the conditional precipitation rates are precipitation averages for each CSF bin.

d. Moisture–convection coupling

Related to circulation-convection coupling, moisture-convection coupling is another well-known source of model errors in the tropics. Observations indicate that conditionally averaged precipitation rate increases rapidly as column saturation fraction (CSF), a measure of column integrated moisture, increases beyond a "critical point" (Bretherton et al. 2004; Peters and Neelin 2006; Rushley et al. 2018). Yet several studies have documented models exhibiting a "too early and too gradual" increase of conditionally averaged precipitation rate with increasing CSF. This has been shown to be related to the common model bias of over precipitating in dry regimes, which is, in part, a result of insufficient sensitivity of model convection to variations in tropospheric moisture (e.g., Kuo et al. 2017).

Figure 10 shows that this precipitation pick-up is less rapid for GFSv15 and GFSv16 at 12-h lead time, but particularly for GFSv15 suggesting that GFSv15 convection is insufficiently sensitive to variations in tropospheric humidity. Model precipitation rates for CSF values above 0.8 are too small compared to IMERG and are too large for moderate CSF values. The model CSF distribution shifts away from the initial bimodal distribution by losing density at CSF values between 0.3 and 0.6 (below the "critical point" for precipitation pick-up). While the GFSv16 precipitation pick-up is improved over GFSv15, especially at later lead times, the CSF distribution shifts further from verification.

Examining the temporal tendencies of CSF and precipitation rate, Wolding et al. (2020) found that column moisture and convection coevolve in a cyclical fashion that traces a clockwise evolution through CSF-precipitation space, a process that is illustrated in Fig. 11a. In observations, this cyclical evolution corresponds to a transition from predominantly shallow, to convective, to stratiform type precipitation, a progression characteristic of several types of CCEWs and the MJO. This evolution is robust across time and spatial scales.

Modeling and observational studies indicate that the progressive deepening of convective heating, and the associated transition to increasingly top-heavy large-scale circulations, play a crucial role in driving the coupled evolution of column moisture and convection (Schumacher et al. 2004; Wolding and Maloney 2015; Ruppert and Johnson 2015; Inoue and Back 2017). Wolding et al. (2020) found that several climate models produced erroneous moisture-convection coupling, and suggested that these models may have difficulties reproducing the progressive deepening of convective heating seen in real-world convection.

Figure 11a shows the moisture-convection coupling diagnostic applied to ERA5 CSF and IMERG precipitation. The magnitude and direction of the vectors indicate the bin-mean evolution of precipitation and CSF over the time period considered. Arrow centers are located at bin centers and are specific to observations in that bin. See Wolding et al. (2020) for more details. We note that, while the clockwise coupled evolution seen in Fig. 11a is a robust feature seen in several combinations of observational precipitation and reanalysis CSF datasets examined (not shown), the strength of the coevolution is sensitive to the specific datasets and time periods used. As discussed in Wolding et al. (2022), the sources of these sensitivities are still under investigation. Nevertheless, Figs. 11b and 11c indicate that GFSv15 shows slightly weaker than observed coevolution initially and that coevolution loses strength with lead time. By 120-h lead time GFSv15 only has very small moistening and drying tendencies. Results for GFSv16 are similar, but slightly improved. The coevolution at 12-h lead time is stronger than in GFSv15, although still weaker than observed and it is weakened, but still distinguishable, at 120-h lead time. The weakening of the precipitation–CSF coevolution with increasing lead time in both model versions warrants further investigation and has the potential to lead the way toward future model improvements.

The results presented here suggest that changes made between GFSv15 and GFSv16 lead to somewhat more realistic moisture precipitation relationship, although this translates to only modest improvements in the representation of the MJO and CCEW activity and some improvement in the space-time coherence between precipitation and dynamics. We also considered the convective adjustment time scale (Jiang et al. 2016) and found no significant difference between the model versions. Modest improvements in moisture–convection coupling seen here could be related to the improvements in dynamics–convection coupling represented by the coherence spectra in Fig. 8.

e. Python package: Tropical diagnostics

A python package including all the above diagnostics is available for download at https://github.com/mgehne/tropical_diagnostics. The package includes examples on how to compute the diagnostics as well as the underlying computational routines. This package is easily portable, and the authors encourage users to apply it to output from other NWP models, as well as to provide feedback regarding other tropical diagnostics of interest. Distribution of the python package includes all underlying computational routines and users have

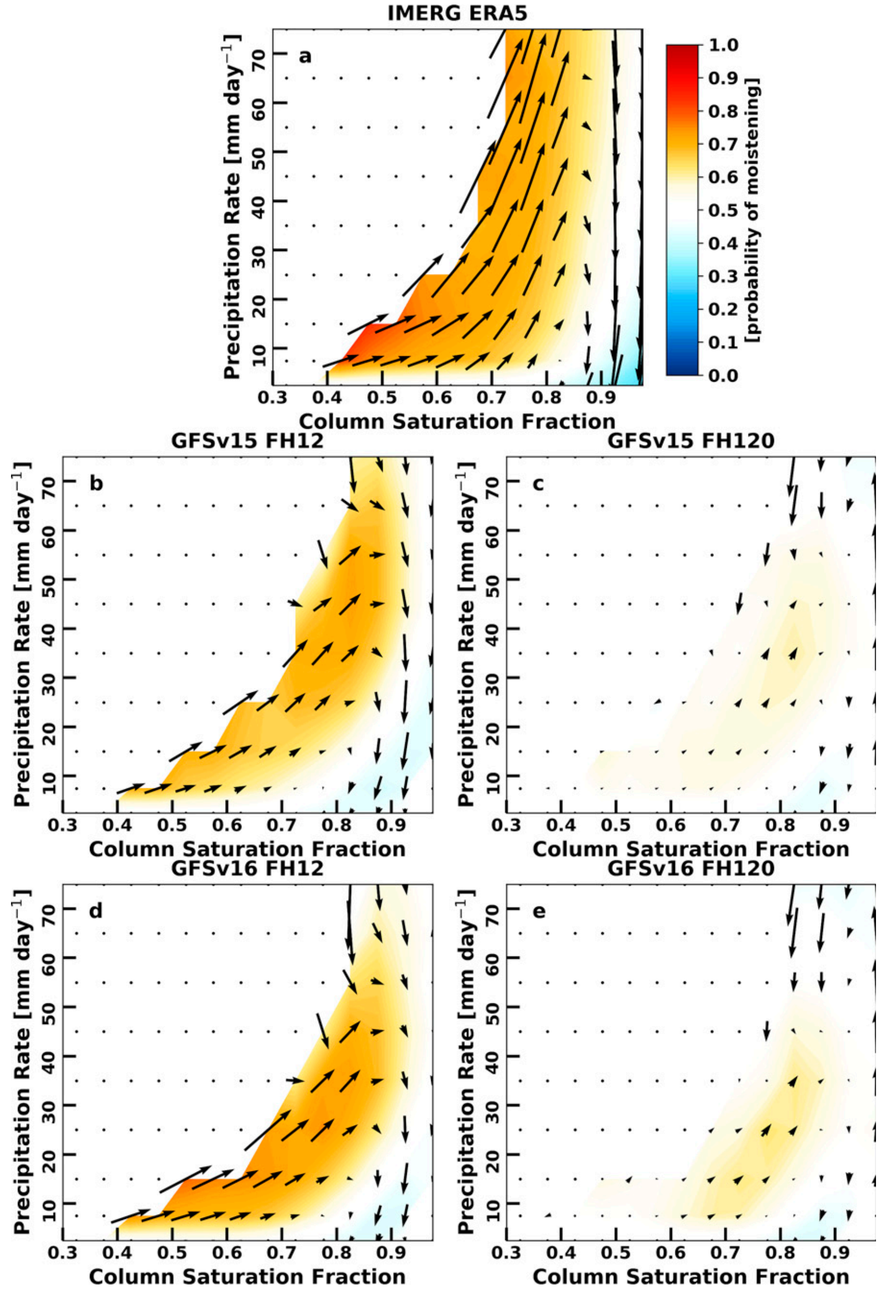

FIG. 11. Coevolution of binned precipitation and column saturation fraction for (a) IMERG–ERA5, (b) GFSv15 FH06, (c) GFSv15 FH120, (d) GFSv16 FH06, and (e) GFSv16 FH120. Vectors represent the bin-mean temporal difference of precipitation and CSF, and color shading indicates the fraction of observations having a positive CSF difference within each bin. The magnitude and direction of the vector indicates the net evolution of precipitation and CSF over 12 h across the lead time, with arrow centers located at bin centers. Bins containing less than 200 observations are marked with stippling.

the ability to edit those to better fit their needs if necessary. The package also includes netcdf data files of the EOFs used in the CCEW activity diagnostic.

In addition to the stand-alone python package several of the diagnostics have been included in the recent release of METcalcpy and METplotpy (Win-Gildenmeister et al. 2021). There are also ongoing efforts to add these diagnostics to the MDTF diagnostics code repository, as we anticipate that they will be useful for climate modeling as well.

5. Summary and discussion

In this work we introduce a set of diagnostics designed to evaluate tropical NWP forecasts. The utility of the diagnostics is demonstrated by applying them to 6 months of operational GFSv15 and retrospective GFSv16 forecasts. The diagnostics assess model skill by lead time of zonal propagation of large-scale precipitation features, coherence-squared spectra between precipitation and dynamical variables and the structure of this coherence in the vertical, and CCEW activity skill and moisture–convection coupling. We tested the sensitivity of the diagnostics to the length of evaluation period by using as little as 3 months and observed only minor changes, demonstrating the value of these diagnostics during model development. Being able to evaluate tropical variability and moisture–convection coupling during model development is advantageous when testing convective parameterization changes and other model improvements as shown in, e.g., Bengtsson et al. (2019, 2021). By having the diagnostics available as a function of lead time, users can distinguish between model error and initialization error impacts. The diagnostics introduced here address different aspects of tropical forecast variability. One aspect concerns the realism of the model representation of fundamental processes/relationships of tropical convection. The other is the question of whether the model is performing better in a forecast setting. Taken together these metrics speak to a third aspect: Does improving model representation of fundamental processes directly and immediately translate to improved forecast capabilities?

The forecast performance metric Hovmöller pattern correlation assesses the zonal propagation skill of precipitation events and shows that there is loss of potential skill during the first hours of the forecast for both model versions. Focusing on each CCEW type shows that the GFS has useful skill for less than 24 h for CCE Kelvin waves and MRGs, less than 48 h for ERs and for about 144 h for the MJO, and that differences between model versions are not significant for this aspect.

For process-level diagnostics, coherence-squared zonal wavenumber–frequency spectra indicate at which scales model forecasts couple convection and large-scale circulation and that those are similar to observed scales. The coherence-squared spectra also show that models tend to propagate CCEWs too fast and that coupling strength decreases with lead time. Here we show that GFSv16 has improved compared to GFSv15. In addition, statistical evaluation of the moisture–convection coupling shows that GFSv16 has modest improvement over GFSv15.

We demonstrate that somewhat better performance in the tropical process-level diagnostics presented here does not necessarily translate to better performance in tropical forecast metrics and in traditional skill metrics. For example, while the CSF–precipitation relationship shows some improvement in GFSv16, q2m bias in the tropics is increased in GFSv16. Because some of the process-level improvements are modest, dramatic changes in precipitation forecast skill may not be expected. This interesting discrepancy between performance with respect to statistical metrics and the physically based diagnostics begs the question: Which statistical metrics are appropriate for model evaluation in the tropics? Given the results presented above it may be worth taking a closer look in future work at the connection between the statistical metrics and the diagnostics introduced here. A starting hypothesis would be that this is scale dependent, with the statistical metrics more influenced by skill on smaller scales and the CCEW diagnostics reflecting larger scales. We show that, while there is a lot of room for improvement, the tropical diagnostics can potentially provide practical additional information about model performance when used in conjunction with more traditional skill metrics.

The results from diagnostics introduced here can be used to reexamine the questions posed in the introduction. It appears that based on the coherence spectra results that GFSv15 model convection does not correctly respond to the dynamic environment. This is somewhat mitigated in GFSv16 with improvement in dynamic–convection coupling, but only modest and limited improvements in moisture-convection coupling. There also appear to be errors in how model convection feeds back onto the large-scale circulation, leading to the incorrect speed of propagation for CCEWs. Significant errors in precipitation during the first few hours of the model forecast as seen in the Hovmöller pattern correlation, along with recent tropical predictability studies (Judt 2020; Li and Stechmann 2020) indicate that there is potential skill the models lack at the beginning of the forecast due to deficiencies in initialization.

While the dynamic–convection coupling exhibits notable improvement from GFSv15 to GFSv16, improvements in moisture–convection coupling are limited and modest, leaving a lot of room for improvement in future development of the GFS.

Based on the forecasts evaluated here it is not clear which particular model changes (e.g., increased vertical resolution, boundary layer turbulence scheme or changes in data assimilation) are responsible for the convection–dynamics coupling improvement. Application of these diagnostics throughout the next model version development phase will help identify which changes lead to improved model performance in the tropics. Current efforts in developing improved stochastic convection parameterizations are already benefiting from usage of these diagnostics (e.g., Bengtsson et al. 2019, 2021).

Acknowledgments. This research was partially supported by the Physical Science Laboratory of NOAA's Earth System Research Laboratory, the California Department for Water Resources (CDWR), and the NOAA Cooperative Agreement with CIRES (NA17OAR4320101).

Data availability statement. Model output from the GFS runs is archived on the National Centers for Environmental Information's (NCEI) High Performance Storage System (HPSS) https://www.ncdc.noaa.gov/nomads/documentation/user-guide/introduction. ERA5 reanalysis at hourly resolution is available for download at Copernicus https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels?tab=form. IMERG v6 precipitation at 30-min, 0.1° resolution is available for download at https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGHH_06/summary?keywords=IMERG. PERSIANN

TABLE A1. Wavenumber–frequency regions used for filtering for CCEWs, zonal wavenumber limits, frequency limits (days), and equivalent height limits. Latitude–longitude regions used for the subsequent EOF analysis.

	Kelvin	ER	MRG	MJO
Wavenumber	$k = 1$ to 14	$k = -1$ to -10		Eastward
Frequency	2.5–20 days	10–40 days	2–6 days	30–96 days
Equivalent height	8–90 m	8–90 m		
Latitude	15°S–15°N	20°S–20°N	20°S–20°N	20°S–20°N
Longitude	130°–270°E	60°–220°E	150°–210°E	30°–240°E
Variance explained	25% (4 EOFs)	22.3% (4 EOFs)	5.4% (2 EOFs)	86.5% (4 EOFs)

CDR at daily, 0.25° resolution is available for download at NCEI https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00854/html.

APPENDIX A

CCEW Projection

First, observed daily precipitation is filtered for the study period using the respective wavenumber frequency regions defined in Wheeler and Kiladis (1999); Kiladis et al. (2014, 2016), then the first 40 EOFs of that filtered data are computed (von Storch and Zwiers 1999) for tropical latitude bands and different longitude bands, depending on the CCEW. For propagating signals this generally results in quadrature EOF pairs explaining nearly equal amounts of variance. Based on the explained variance and clear separation from the following EOFs the first four EOFs (two pairs) are kept except for MRGs where only the first two EOFs are kept for the projection. The EOFs are robust in the sense that using PERSIANNCDR or TRMM3B42 gives very similar results. Details on the filtering and spatial regions used for the EOF analysis are given in Table A1.

Wave activity is defined by the projection of observed or forecast precipitation anomalies onto the EOFs, squaring the result, averaging over the EOFs, and then taking the square root. The result is a time series of CCEW activity for both the observed and modeled precipitation. Model precipitation anomalies are computed by lead time by removing the time mean over the 4 months of data available. We note that for longer time series computing the lead time dependent climatology for each day of the year and computing anomalies from that would be preferable.

To assess the ability of the model to maintain CCEW activity as a function of lead time, the CCEW skill is defined as the correlation between forecast and observed CCEW activity at a given lead time (τ):

$$\mathrm{AC}(\tau) = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} o_k(t_i) f_k(t_i, \tau)}{\sqrt{\sum_{i=1}^{N}\sum_{k=1}^{K} o_k^2(t_i)}\sqrt{\sum_{i=1}^{N}\sum_{k=1}^{K} f_k^2(t_i, \tau)}}. \qquad \text{(A1)}$$

Here, $f_k(t_i, \tau)$ is the forecast precipitation projection onto EOF k at lead time τ and time t_i; $o_k(t_i)$ is the projection of observed precipitation onto EOF k at time t_i; and N is the total number of times and $K = 4$ or 2 for CCEWs and the MJO, respectively.

APPENDIX B

Vertical Profiles of Coherence Squared

To compute coherence squared for a particular CCEW band at all vertical levels, precipitation is first filtered for the CCEW band of interest (see Table A1 for filter regions) and the space–time cross-spectrum with a dynamical field (e.g., winds, temperature, or specific humidity) is computed at each vertical level. When filtering model precipitation for CCEW wavenumber–frequency regions we pad the time series with IMERG precipitation starting in 2001 prior to the start of the forecast period (Janiga et al. 2018).

Because precipitation is filtered, the coherence-squared is not significant outside the filtered wavenumber–frequency region. The significant coherence-squared values are averaged in wavenumber and frequency, resulting in a single number per vertical level. All other cross-spectral components are also averaged where the coherence-squared is significant. The averaged co and quadrature spectra are used to estimate a phase angle for each vertical level.

Statistical significance of the coherence-squared is estimated by computing a background coherence-squared that consists of two components. The first is estimated by running one time series backward and computing the coherence-squared spectrum (C_{BG}). This approach retains the autocorrelation of both time series, but destroys any temporal relationship the time series had. The second is a distribution of random coherence-squared spectra (C_{rand}), where both samples are drawn from a random normal distribution with zero mean and standard deviation matching the filtered precipitation and dynamical variable time series, respectively. The sum $C_{\mathrm{BG}} + C_{\mathrm{rand}}$ is used to determine the significance level for coherence-squared at each wavenumber–frequency. Here we choose a significance level of 0.95, which means that coherence-squared values larger than the 95th percentile of the background distribution are considered statistically different from zero.

REFERENCES

Alpert, J. C., M. Kanamitsu, P. Caplan, J. Sela, G. H. White, and E. Kalnay, 1988: Mountain induced gravity wave drag parameterization in the NMC medium-range forecast model. *Eighth*

Conf. on Numerical Weather Prediction, Baltimore, MD, Amer. Meteor. Soc., 726–733.

Arakawa, A., and W. H. Schubert, 1974: Interaction of a cumulus cloud ensemble with the large-scale environment. Part I. *J. Atmos. Sci.*, **31**, 674–701, https://doi.org/10.1175/1520-0469(1974)031<0674:IOACCE>2.0.CO;2.

Ashouri, H., K.-L. Hsu, S. Sorooshian, D. K. Braithwaite, K. R. Knapp, L. D. Cecil, B. R. Nelson, and O. P. Prat, 2015: PERSIANN-CDR: Daily precipitation climate data record from multisatellite observations for hydrological and climate studies. *Bull. Amer. Meteor. Soc.*, **96**, 69–83, https://doi.org/10.1175/BAMS-D-13-00068.1.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Bengtsson, L., and Coauthors, 2019: Convectively coupled equatorial wave simulations using the ECMWF IFS and the NOAA GFS cumulus convection schemes in the NOAA GFS model. *Mon. Wea. Rev.*, **147**, 4005–4025, https://doi.org/10.1175/MWR-D-19-0195.1.

——, J. Dias, S. Tulich, M. Gehne, and J.-W. Bao, 2021: A stochastic parameterization of organized tropical convection using cellular automata for global forecasts in NOAA's unified forecast system. *J. Adv. Model. Earth Syst.*, **13**, e2020MS002260, https://doi.org/10.1029/2020MS002260.

Branstator, G., 2014: Long-lived response of the midlatitude circulation and storm tracks to pulses of tropical heating. *J. Climate*, **27**, 8809–8826, https://doi.org/10.1175/JCLI-D-14-00312.1.

Bretherton, C. S., M. E. Peters, and L. E. Back, 2004: Relationships between water vapor path and precipitation over the tropical oceans. *J. Climate*, **17**, 1517–1528, https://doi.org/10.1175/1520-0442(2004)017<1517:RBWVPA>2.0.CO;2.

Brown, B., and Coauthors, 2021: The Model Evaluation Tools (MET): More than a decade of community-supported forecast verification. *Bull. Amer. Meteor. Soc.*, **102**, E782–E807, https://doi.org/10.1175/BAMS-D-19-0093.1.

Chen, J.-H., and S.-J. Lin, 2011: The remarkable predictability of inter-annual variability of Atlantic hurricanes during the past decade. *Geophys. Res. Lett.*, **38**, L11804, https://doi.org/10.1029/2011GL047629.

——, and ——, 2013: Seasonal predictions of tropical cyclones using a 25-km-resolution general circulation model. *J. Climate*, **26**, 380–398, https://doi.org/10.1175/JCLI-D-12-00061.1.

Clough, S., M. Shephard, E. Mlawer, J. Delamere, M. Iacono, K. Cady-Pereira, S. Boukabara, and P. Brown, 2005: Atmospheric radiative transfer modeling: A summary of the AER codes. *J. Quant. Spectrosc. Radiat. Transfer*, **91**, 233–244, https://doi.org/10.1016/j.jqsrt.2004.05.058.

Dias, J., and G. N. Kiladis, 2019: The influence of tropical forecast errors on higher latitude predictions. *Geophys. Res. Lett.*, **46**, 4450–4459, https://doi.org/10.1029/2019GL082812.

——, M. Gehne, G. N. Kiladis, N. Sakaeda, P. Bechtold, and T. Haiden, 2018: Equatorial waves and the skill of NCEP and ECMWF numerical weather prediction systems. *Mon. Wea. Rev.*, **146**, 1763–1784, https://doi.org/10.1175/MWR-D-17-0362.1.

——, S. N. Tulich, M. Gehne, and G. Kiladis, 2021: Tropical origins of weeks 2–4 forecast errors during the Northern Hemisphere cool season. *Mon. Wea. Rev.*, **149**, 2975–2991, https://doi.org/10.1175/MWR-D-21-0020.1.

Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarpley, 2003: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *J. Geophys. Res.*, **108**, 8851, https://doi.org/10.1029/2002JD003296.

Farrar, M., 2021: Upgrade NCEP Global Forecast Systems (GFS) to v16. NOAA/NWS/National Centers for Environmental Prediction, 13 pp., https://www.weather.gov/media/notification/pdf2/scn21-20_gfsv16.0_aac.pdf.

Ferranti, L., T. N. Palmer, and F. Molteni, 1990: Tropical–extratropical interaction associated with the 30–60 day oscillation and its impact on medium and extended range prediction. *J. Atmos. Sci.*, **47**, 2177–2199, https://doi.org/10.1175/1520-0469(1990)047<2177:TEIAWT>2.0.CO;2.

Garfinkel, C., O. Shamir, I. Fouxon, and N. Paldor, 2021: Tropical background and wave spectra: Contribution of wave–wave interactions in a moderately nonlinear turbulent flow. *J. Atmos. Sci.*, **78**, 1773–1789, https://doi.org/10.1175/JAS-D-20-0284.1.

Gehne, M., T. Hamill, G. Kiladis, and K. Trenberth, 2016: Comparison of global precipitation estimates across a range of temporal and spatial scales. *J. Climate*, **29**, 7773–7795, https://doi.org/10.1175/JCLI-D-15-0618.1.

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

Hamill, T. M., and G. N. Kiladis, 2014: Skill of the MJO and Northern Hemisphere blocking in GEFS medium-range reforecasts. *Mon. Wea. Rev.*, **142**, 868–885, https://doi.org/10.1175/MWR-D-13-00199.1.

Han, J., and H. 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.

——, and C. S. Bretherton, 2019: TKE-based moist eddy-diffusivity mass-flux (EDMF) parameterization for vertical turbulent mixing. *Wea. Forecasting*, **34**, 869–886, https://doi.org/10.1175/WAF-D-18-0146.1.

——, W. Wang, Y. C. Kwon, S. 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.

Harris, L., 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.

Hendon, H. H., and M. C. Wheeler, 2008: Some space–time spectral analyses of tropical convection and planetary-scale waves. *J. Atmos. Sci.*, **65**, 2936–2948, https://doi.org/10.1175/2008JAS2675.1.

Hersbach, H., and Coauthors, 2018: ERA5 hourly data on pressure levels from 1979 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), accessed August 2021, https://doi.org/10.24381/cds.bd0915c6.

——, and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hovmöller, E., 1949: The trough-and-ridge diagram. *Tellus*, **1**, 62–66, https://doi.org/10.3402/tellusa.v1i2.8498.

Huffman, G. J., D. T. Bolvin, E. J. Nelkin, and D. B. Wolff, 2007: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *J. Hydrometeor.*, **8**, 38–55, https://doi.org/10.1175/JHM560.1.

——, E. Stocker, D. Bolvin, E. Nelkin, and J. Tan, 2019: GPM IMERG final precipitation L3 half hourly 0.1 degree × 0.1 degree V06 (GPM_3IMERGHH). Goddard Earth Sciences Data and Information Services Center (GES DISC), accessed October 2021, https://doi.org/10.5067/GPM/IMERG/3B-HH/06.

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.

Inoue, K., and L. Back, 2017: Gross moist stability analysis: Assessment of satellite-based products in the GMS plane. *J. Atmos. Sci.*, **74**, 1819–1837, https://doi.org/10.1175/JAS-D-16-0218.1.

Janiga, M. A., C. Schreck, J. A. Ridout, M. Flatau, N. Barton, E. J. Metzger, and C. Reynolds, 2018: Subseasonal forecasts of convectively coupled equatorial waves and the MJO: Activity and predictive skill. *Mon. Wea. Rev.*, **146**, 2337–2360, https://doi.org/10.1175/MWR-D-17-0261.1.

Jiang, X., M. Zhao, E. D. Maloney, and D. E. Waliser, 2016: Convective moisture adjustment time scale as a key factor in regulating model amplitude of the Madden–Julian Oscillation. *Geophys. Res. Lett.*, **43**, 10 412–10 419, https://doi.org/10.1002/2016GL070898.

Jolliffe, I. T., and D. B. Stephenson, 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* 2nd ed. Wiley-Blackwell, 274 pp.

Jones, C., D. E. Waliser, K. M. Lau, and W. Stern, 2004: Global occurrences of extreme precipitation and the Madden–Julian Oscillation: Observations and predictability. *J. Climate*, **17**, 4575–4589, https://doi.org/10.1175/3238.1.

Judt, F., 2020: Atmospheric predictability of the tropics, middle latitudes, and polar regions explored through global storm-resolving simulations. *J. Atmos. Sci.*, **77**, 257–276, https://doi.org/10.1175/JAS-D-19-0116.1.

Jung, T., M. J. Miller, and T. N. Palmer, 2010: Diagnosing the origin of extended-range forecast errors. *Mon. Wea. Rev.*, **138**, 2434–2446, https://doi.org/10.1175/2010MWR3255.1.

Kiladis, G. N., M. C. Wheeler, P. T. Haertel, K. H. Straub, and P. E. Roundy, 2009: Convectively coupled equatorial waves. *Rev. Geophys.*, **47**, RG2003, https://doi.org/10.1029/2008RG000266.

——, J. Dias, K. H. Straub, M. C. Wheeler, S. N. Tulich, K. Kikuchi, K. M. Weickmann, and M. J. Ventrice, 2014: A comparison of OLR and circulation-based indices for tracking the MJO. *Mon. Wea. Rev.*, **142**, 1697–1715, https://doi.org/10.1175/MWR-D-13-00301.1.

——, ——, and M. Gehne, 2016: The relationship between equatorial mixed Rossby-gravity and eastward inertio-gravity waves. Part I. *J. Atmos. Sci.*, **73**, 2123–2145, https://doi.org/10.1175/JAS-D-15-0230.1.

Kleist, D. T., and K. Ide, 2015: An OSSE-based evaluation of hybrid variational–ensemble data assimilation for the NCEP GFS. Part II: 4DEnVar and hybrid variants. *Mon. Wea. Rev.*, **143**, 452–470, https://doi.org/10.1175/MWR-D-13-00350.1.

Kuo, Y.-H., J. D. Neelin, and C. R. Mechoso, 2017: Tropical convective transition statistics and causality in the water vapor–precipitation relation. *J. Atmos. Sci.*, **74**, 915–931, https://doi.org/10.1175/JAS-D-16-0182.1.

——, and Coauthors, 2020: Convective transition statistics over tropical oceans for climate model diagnostics: GCM evaluation. *J. Atmos. Sci.*, **77**, 379–403, https://doi.org/10.1175/JAS-D-19-0132.1.

Lei, L., and J. S. Whitaker, 2016: A four-dimensional incremental analysis update for the ensemble Kalman filter. *Mon. Wea. Rev.*, **144**, 2605–2621, https://doi.org/10.1175/MWR-D-15-0246.1.

Li, Y., and S. N. Stechmann, 2020: Predictability of tropical rainfall and waves: Estimates from observational data. *Quart. J. Roy. Meteor. Soc.*, **146**, 1668–1684, https://doi.org/10.1002/qj.3759.

Lin, H., G. Brunet, and J. Derome, 2009: An observed connection between the North Atlantic Oscillation and the Madden–Julian oscillation. *J. Climate*, **22**, 364–380, https://doi.org/10.1175/2008JCLI2515.1.

Maloney, E. D., and Coauthors, 2019: Process-oriented evaluation of climate and weather forecasting models. *Bull. Amer. Meteor. Soc.*, **100**, 1665–1686, https://doi.org/10.1175/BAMS-D-18-0042.1.

Maxson, B., 2019: Upgrade NCEP Global Forecast Systems (GFS) to V15.1. NOAA/NWS/National Centers for Environmental Prediction, 8 pp., https://www.weather.gov/media/notification/pdf2/scn19-40gfs_v15_1.pdf.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.*, **102**, 16 663–16 682, https://doi.org/10.1029/97JD00237.

Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424, https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2.

Pan, H.-L., and W.-S. Wu, 1995: Implementing a mass flux convective parameterization package for the NMC medium-range forecast model. NMC Office Note 409, NOAA, 40 pp., https://repository.library.noaa.gov/view/noaa/11429.

Pendergrass, A., and D. Hartmann, 2014: The atmospheric energy constraint on global-mean precipitation change. *J. Climate*, **27**, 757–768, https://doi.org/10.1175/JCLI-D-13-00163.1.

Persson, A., 2017: The story of the Hovmöller diagram: An (almost) eyewitness account. *Bull. Amer. Meteor. Soc.*, **98**, 949–957, https://doi.org/10.1175/BAMS-D-15-00234.1.

Peters, O., and J. Neelin, 2006: Critical phenomena in atmospheric precipitation. *Nat. Phys.*, **2**, 393–396, https://doi.org/10.1038/nphys314.

Ruppert, J., and R. Johnson, 2015: Diurnally modulated cumulus moistening in the preonset stage of the Madden–Julian oscillation during DYNAMO. *J. Atmos. Sci.*, **72**, 1622–1647, https://doi.org/10.1175/JAS-D-14-0218.1.

Rushley, S., D. Kim, C. Bretherton, and M.-S. Ahn, 2018: Reexamining the nonlinear moisture-precipitation relationship over the tropical oceans. *Geophys. Res. Lett.*, **45**, 1133–1140, https://doi.org/10.1002/2017GL076296.

Schreck, C. J., III, M. A. Janiga, and S. Baxter, 2020: Sources of tropical subseasonal skill in the CFSv2. *Mon. Wea. Rev.*, **148**, 1553–1565, https://doi.org/10.1175/MWR-D-19-0289.1.

Schumacher, C., R. A. Houze, and I. Kraucunas, 2004: The tropical dynamical response to latent heating estimates derived from the TRMM precipitation radar. *J. Atmos. Sci.*, **61**, 1341–1358, https://doi.org/10.1175/1520-0469(2004)061<1341:TTDRTL>2.0.CO;2.

Selz, T., and G. C. Craig, 2015: Simulation of upscale error growth with a stochastic convection scheme. *Geophys. Res. Lett.*, **42**, 3056–3062, https://doi.org/10.1002/2015GL063525.

Sorooshian, S., K. Hsu, D. Braithwaite, H. Ashouri, and NOAA CDR Program, 2014: NOAA Climate Data Record (CDR) of precipitation estimation from remotely sensed information using artificial neural networks (PERSIANN-CDR), version 1 revision 1. NOAA, accessed December 2021, https://doi.org/10.7289/V51V5BWQ.

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

Stephens, G. L., and Coauthors, 2010: Dreary state of precipitation in global models. *J. Geophys. Res.*, **115**, https://doi.org/10.1029/2010JD014532.

Tropical Rainfall Measuring Mission (TRMM), 2011: TRMM (TMPA) rainfall estimate L3 3 hour 0.25 degree × 0.25 degree v7. Goddard Earth Sciences Data and Information Services Center (GES DISC), accessed September 2021, https://doi.org/10.5067/TRMM/TMPA/3H/7.

von Storch, H., and F. Zwiers, 1999: *Statistical Analysis in Climate Research.* Cambridge University Press, 484 pp.

Weber, N. J., D. Kim, and C. F. Mass, 2021: Convection–Kelvin wave coupling in a global convection-permitting model. *J. Atmos. Sci.*, **78**, 1039–1055, https://doi.org/10.1175/JAS-D-20-0243.1.

Wheeler, M. C., and G. N. Kiladis, 1999: Convectively coupled equatorial waves: Analysis of clouds and temperature in the wavenumber–frequency domain. *J. Atmos. Sci.*, **56**, 374–399, https://doi.org/10.1175/1520-0469(1999)056<0374:CCEWAO>2.0.CO;2.

——, 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.

Wilks, D., 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Win-Gildenmeister, M., T. Burek, H. Fisher, C. Kalb, D. Adriaansen, D. Fillmore, and T. Jensen, 2021: The METcalcpy version develop user's guide. Developmental Testbed Center, accessed 17 May 2021, https://github.com/dtcenter/METcalcpy/releases, https://metcalcpy.readthedocs.io/en/main_v1.0/Users_Guide/release-notes.html.

Wolding, B., and E. D. Maloney, 2015: Objective diagnostics and the Madden–Julian oscillation. Part II: Application to moist static energy and moisture budgets. *J. Climate*, **28**, 7786–7808, https://doi.org/10.1175/JCLI-D-14-00689.1.

——, J. Dias, G. Kiladis, F. Ahmed, S. W. Powell, E. Maloney, and M. Branson, 2020: Interactions between moisture and tropical convection. Part I: The coevolution of moisture and convection. *J. Atmos. Sci.*, **77**, 1783–1799, https://doi.org/10.1175/JAS-D-19-0225.1.

——, S. W. Powell, F. Ahmed, J. Dias, M. Gehne, G. Kiladis, and J. D. Neelin, 2022: Tropical thermodynamic-convection coupling in observations and reanalyses. *J. Atmos. Sci.*, **79**, 1781–1803, https://doi.org/10.1175/JAS-D-21-0256.1.

Yang, F., and V. Tallapragada, 2018: Implementation and evaluation of the NOAA next generation global prediction system with FV3 dynamical core and advanced physics. *Eighth Conf. on Transition of Research to Operations*, Austin, TX, Amer. Meteor. Soc., 1.4, https://ams.confex.com/ams/98Annual/webprogram/Paper329963.html.

——, and Coauthors, 2020: Model upgrade plan and initial results from a prototype NCEP Global Forecast System version 16. *10th Conf. on Transition of Research to Operations,* Boston, MA, Amer. Meteor. Soc., https://ams.confex.com/ams/2020Annual/meetingapp.cgi/Paper/362797.

Yang, G.-Y., S. Ferrett, S. Woolnough, J. Methven, and C. Holloway, 2021: Real-time identification of equatorial waves and evaluation of waves in global forecasts. *Wea. Forecasting*, **36**, 171–193, https://doi.org/10.1175/WAF-D-20-0144.1.

Yudin, V., R. Akmaev, T. Fuller-Rowell, and J. Alpert, 2016: Gravity wave physics in the NOAA environmental modeling system: Improving predictions of Whole Atmosphere Model across the stratosphere. *Int. SPARC Gravity Wave Symp.*, State College, PA, WCRP, http://adapt.psu.edu/2016SPARCGWSymposium/ABSTRACTS/POSTER/Valery_Yudin.html.

——, R. A. Akmaev, J. C. Alpert, T. J. Fuller-Rowell, and S. I. Karol, 2018: Gravity wave physics and dynamics in the FV3-based atmosphere models extended into the mesosphere. *25th Conf. on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., 4B.6, https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345706.html.

Zhou, L., S.-J. Lin, J.-H. Chen, L. Harris, X. Chen, and S. 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.

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.