Diagnostic Report on Advanced Physics Test

Global Model Test Bed (GMTB)

Final May 1, 2019 (first draft March 2019)

Points of Contact: Ligia Bernardet (ligia.Bernardet@noaa.gov) and Grant Firl (grantf@ucar.edu)

Additional GMTB staff that contributed to this activity: Lindsay Blank, Jimy Dudhia, Mike Ek, Grant Firl, Michelle Harrold, Judy Henderson, Tracy Hertneky, Weiwei Li, Linlin Pan, Jamie Wolff, and Man Zhang.

Table of Contents

Introduction

The GMTB tested four configurations of NOAA's Unified Forecast System (UFS) to support EMC in selecting an advanced physics suite for the GFS v16, due to be implemented operationally in 2021. A report on the configuration of the runs can be found here. Runs were initialized every five days between 1 January 2016 and 31 December 2017, alternating between 00 and 12 UTC. In addition, 16 cases studies were added under the recommendation of the EMC Model Evaluation Group (MEG). The runs were conducted by GMTB on the NOAA research and development platform *Jet* using a simplified workflow supplied by EMC.

Using model output from the runs conducted by the GMTB, the primary verification for this test was conducted by EMC. For examples of the available verification generated by EMC, see the following websites below as well as the report issued by the MEG:

- Verification of all four suites with ECMWF analyses as “truth” - 00 UTC initializations
- Verification of all four suites with ECMWF analyses as “truth” - 12 UTC initializations
- Verification of suite 1 and FV3GFSB with ECMWF analyses as “truth” - 00 UTC initializations

GMTB computed selected verification metrics and diagnostics to complement the results provided by EMC and to provide additional information to the independent evaluation panel. The work by GMTB has focused on the following areas: scorecards to summarize the verification results, precipitation, planetary boundary layer (PBL), energy budget, tropical cyclones intensity and track, and horizontal/vertical kinetic energy spectral decomposition. In this report we provide a description of observation datasets and/or benchmarks and methodology as well as summarize the main findings and/or provide sample results in each of these areas. Note that GMTB is not providing a comprehensive evaluation of the test results since that will be done by the independent evaluation panel.

It was not possible to fit all results in this report. Additional figures are provided for download through the DTC website (see Appendix A). A website to describe the methods and results of this test is available here.

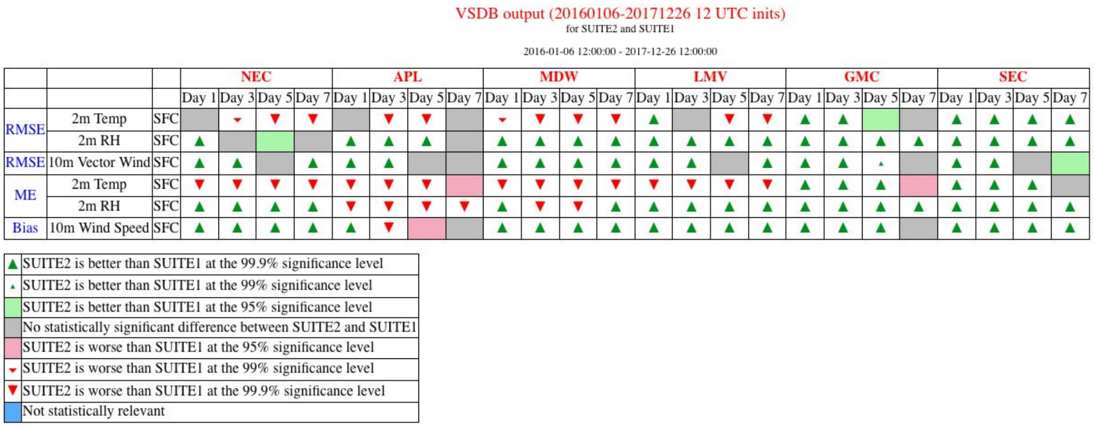
Scorecards: Temperature, RH, and wind speed

- **Observation data set:** NDAS PrepBUFR files for surface verification over CONUS and GDAS PrepBUFR files for upper-air verification over global regions.
- **Method:** With a copious amount of verification results being produced from this test, a “scorecard” is a straightforward way to identify patterns in the difference of performance between two configurations, including level of significance for specified metrics, variables, thresholds, regions, and times. Scorecards for surface and upper-air temperature, RH, and wind speed were produced using VSDB output from EMC's verification system that was then loaded into DTC’s

METviewer, a database and display system. To help identify diurnal signals, the scorecards are available separately for 00 and 12 UTC initializations. For the surface, results are available for the CONUS-East and CONUS-West regions and are further broken down by subregions. For upper air, results are available for the NH, SH, and Tropics. Subregion definitions and further information is provided here.

- **Results:** Due to the generation of a large number of verification scorecards, only a subset of example results are shown to provide detail on how to interpret the scorecard. A comprehensive set of results is available for 00 and 12 UTC initializations at: https://dtcenter.org/eval/gmtb/2019_advphystest/scorecard/

  The scorecards compare Suites 2, 3, and 4 against Suite 1. If an alternative suite is favored over Suite 1 with SS, the cell will be indicated with a green symbol or shading; the symbol or shading will be red if Suite 1 is favored with SS. Figures 1-2 are provided as examples of the types of scorecards created for this physics test with the full suite of results available at the website listed above.



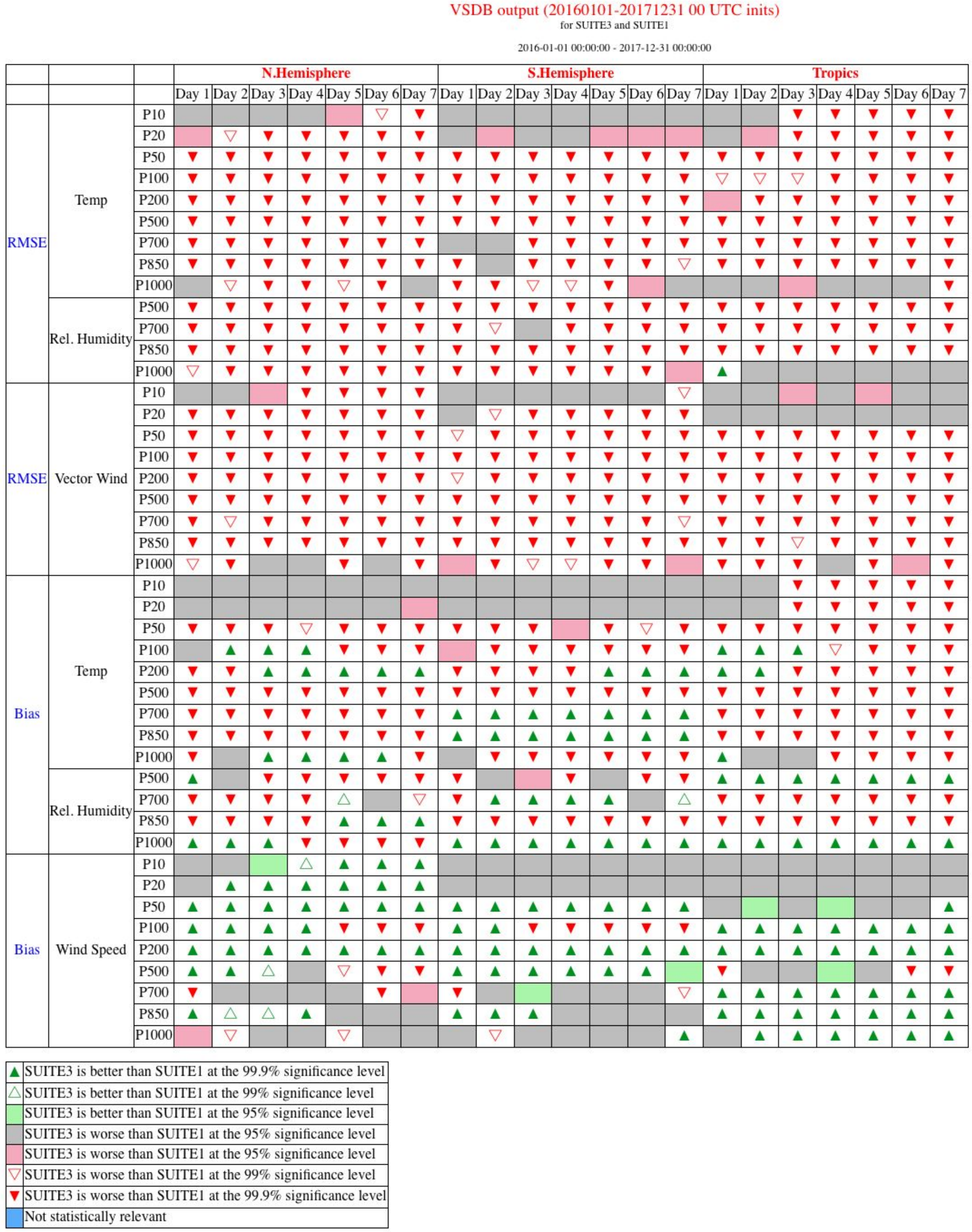
VSDB output (20160106-20171226 12 UTC inits)
for SUITE2 and SUITE1

2016-01-06 12:00:00 - 2017-12-26 12:00:00

			NEC				APL				MDW				LMV				GMC				SEC			
			Day 1	Day 3	Day 5	Day 7	Day 1	Day 3	Day 5	Day 7	Day 1	Day 3	Day 5	Day 7	Day 1	Day 3	Day 5	Day 7	Day 1	Day 3	Day 5	Day 7	Day 1	Day 3	Day 5	Day 7
RMSE	2m Temp	SFC	▨	▾	▼	▼	▨	▼	▼	▨	▾	▼	▼	▼	▲	▨	▼	▼	▲	▲	🟩	▨	▲	▲	▲	▲
	2m RH	SFC	▲	▨	🟩	▨	▲	▲	▲	▨	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
RMSE	10m Vector Wind	SFC	▲	▲	▨	▲	▲	▲	▨	▨	▲	▲	▲	▲	▲	▲	▨	▲	▲	▲	▴	▨	▲	▲	▨	🟩
ME	2m Temp	SFC	▼	▼	▼	▼	▼	▼	▼	🟥	▼	▼	▼	▼	▼	▼	▼	▼	▲	▲	▲	🟥	▲	▲	▲	▨
	2m RH	SFC	▲	▲	▲	▲	▼	▼	▼	▼	▲	▼	▼	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Bias	10m Wind Speed	SFC	▲	▲	▲	▲	▲	▼	🟥	▨	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▨	▲	▲	▲	▲

▲	SUITE2 is better than SUITE1 at the 99.9% significance level
▴	SUITE2 is better than SUITE1 at the 99% significance level
🟩	SUITE2 is better than SUITE1 at the 95% significance level
▨	No statistically significant difference between SUITE2 and SUITE1
🟥	SUITE2 is worse than SUITE1 at the 95% significance level
▾	SUITE2 is worse than SUITE1 at the 99% significance level
▼	SUITE2 is worse than SUITE1 at the 99.9% significance level
🟦	Not statistically relevant

Figure 1. Scorecard documenting performance of Suite 2 as compared to Suite 1 over the subregions across the eastern portions of the CONUS for RMSE and mean error (also referred to as bias) of 2-m temperature, 2-m RH, and 10-m wind speed by forecast lead time for all 12 UTC initializations during the entire test period (20160101-20171231). Green (red) marks indicate Suite 2 (Suite 1) is better than Suite 1 (Suite 2). Statistical significance is represented by the type of marks: shading, small arrows, and large arrows indicate 95%, 99%, and 99.9% significance, respectively.



VSDB output (20160101-20171231 00 UTC inits)
for SUITE3 and SUITE1

2016-01-01 00:00:00 - 2017-12-31 00:00:00

			N.Hemisphere							S.Hemisphere							Tropics						
			Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
RMSE	Temp	P10	▨	▨	▨	▨	🟥	▽	▼	▨	▨	▨	▨	▨	▨	▨	▨	▨	▼	▼	▼	▼	▼
		P20	🟥	▽	▼	▼	▼	▼	▼	▨	🟥	▨	▨	🟥	🟥	🟥	▨	🟥	▼	▼	▼	▼	▼
		P50	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼
		P100	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▽	▽	▽	▼	▼	▼	▼
		P200	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	🟥	▼	▼	▼	▼	▼	▼
		P500	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼
		P700	▼	▼	▼	▼	▼	▼	▼	▨	▨	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼
		P850	▼	▼	▼	▼	▼	▼	▼	▼	▨	▼	▼	▼	▼	▽	▼	▼	▼	▼	▼	▼	▼
		P1000	▨	▽	▼	▼	▽	▼	▨	▼	▼	▽	▽	▼	🟥	▨	▨	▨	🟥	▨	▨	▨	▼
	Rel. Humidity	P500	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼
		P700	▼	▼	▼	▼	▼	▼	▼	▼	▽	▨	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼
		P850	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼
		P1000	▽	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	🟥	▲	▨	▨	▨	▨	▨	▨
RMSE	Vector Wind	P10	▨	▨	🟥	▼	▼	▼	▼	▨	▨	▨	▨	▨	▨	▽	▨	▨	🟥	▨	🟥	▨	▨
		P20	▼	▼	▼	▼	▼	▼	▼	▨	▽	▼	▼	▼	▼	▼	▨	▨	▨	▨	▨	▨	▨
		P50	▼	▼	▼	▼	▼	▼	▼	▽	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼
		P100	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼
		P200	▼	▼	▼	▼	▼	▼	▼	▽	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼
		P500	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼
		P700	▼	▽	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▽	▼	▼	▼	▼	▼	▼	▼
		P850	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▽	▼	▼	▼	▼
		P1000	▽	▼	▨	▨	▼	▨	▼	🟥	▼	▽	▽	▼	▼	🟥	▼	▼	▼	▨	▼	🟥	▼
Bias	Temp	P10	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▼	▼	▼	▼	▼
		P20	▨	▨	▨	▨	▨	▨	🟥	▨	▨	▨	▨	▨	▨	▨	▨	▨	▼	▼	▼	▼	▼
		P50	▼	▼	▼	▽	▼	▼	▼	▼	▼	▼	🟥	▼	▽	▼	▼	▼	▼	▼	▼	▼	▼
		P100	▨	▲	▲	▲	▼	▼	▼	🟥	▼	▼	▼	▼	▼	▼	▲	▲	▲	▽	▼	▼	▼
		P200	▼	▼	▲	▲	▲	▲	▲	▼	▼	▼	▼	▲	▲	▲	▲	▲	▼	▼	▼	▼	▼
		P500	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼
		P700	▼	▼	▼	▼	▼	▼	▼	▲	▲	▲	▲	▲	▲	▲	▼	▼	▼	▼	▼	▼	▼
		P850	▼	▼	▼	▼	▼	▼	▼	▲	▲	▲	▲	▲	▲	▲	▼	▼	▼	▼	▼	▼	▼
		P1000	▼	▨	▲	▲	▲	▲	▼	▨	▼	▼	▼	▼	▼	▼	▲	▨	▨	▼	▼	▼	▼
	Rel. Humidity	P500	▲	▨	▼	▼	▼	▼	▼	▼	▨	🟥	▼	▨	▼	▼	▲	▲	▲	▲	▲	▲	▲
		P700	▼	▼	▼	▼	△	▨	▽	▼	▲	▲	▲	▲	▨	△	▼	▼	▼	▼	▼	▼	▼
		P850	▼	▼	▼	▼	▲	▲	▲	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼
		P1000	▲	▲	▲	▼	▼	▼	▼	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
Bias	Wind Speed	P10	▨	▨	🟩	△	▲	▲	▲	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
		P20	▨	▲	▲	▲	▲	▲	▲	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
		P50	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▨	🟩	▨	🟩	▨	▨	▲
		P100	▲	▲	▲	▲	▼	▼	▼	▲	▲	▼	▼	▼	▼	▼	▲	▲	▲	▲	▲	▲	▲
		P200	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
		P500	▲	▲	△	▨	▽	▼	▼	▲	▲	▲	▲	▲	▲	🟩	▼	▨	▨	🟩	▨	▼	▼
		P700	▼	▨	▨	▨	▨	▼	🟥	▼	▨	🟩	▨	▨	▨	▽	▲	▲	▲	▲	▲	▲	▲
		P850	▲	△	△	▲	▨	▨	▨	▲	▲	▲	▨	▨	▨	▨	▲	▲	▲	▲	▲	▲	▲
		P1000	🟥	▽	▨	▨	▽	▨	▨	▨	▽	▨	▨	▨	▨	▲	▨	▲	▲	▲	▲	▲	▲

▲	SUITE3 is better than SUITE1 at the 99.9% significance level
△	SUITE3 is better than SUITE1 at the 99% significance level
🟩	SUITE3 is better than SUITE1 at the 95% significance level
▨	SUITE3 is worse than SUITE1 at the 95% significance level
🟥	SUITE3 is worse than SUITE1 at the 95% significance level
▽	SUITE3 is worse than SUITE1 at the 99% significance level
▼	SUITE3 is worse than SUITE1 at the 99.9% significance level
🟦	Not statistically relevant

Figure 2. Scorecard documenting performance of Suite 3 as compared to Suite 1 over the NH, SH, and Tropics for RMSE and bias of temperature, RH, and wind speed at various pressure levels by forecast lead time for all 00 UTC initializations during the entire test period (20160101-20171231). Green (red) marks indicate Suite 3 (Suite 1) is better than Suite 1 (Suite 3). Statistical significance is represented by

the type of marks: shading, open arrows, and filled arrows indicate 95%, 99%, and 99.9% significance, respectively.

Grid-to-grid precipitation verification

- **Observation data set:** The CCPA QPE data set was used over the CONUS; CCPA has a resolution of ~4.8 km. For the global evaluation, CMORPH precipitation analyses (60°N-60°S) were used due to its high spatial resolution (8 km at the equator, ~0.07°).
- **Method:** Both the CCPA QPE analyses and the 0.25° post-processed model output were interpolated to G218. The CMORPH analyses were interpolated to a 0.25° global grid and compared to the forecast over the Northern Hemisphere (NH: 20°–60°N), Southern Hemisphere (SH: 20°–60°S), and Tropics (20°S–20°N) regions. Grid-to-grid precipitation verification focused on 6-h accumulations for the CONUS domain and 24-h accumulation period (valid from 12 UTC to 12 UTC) for the NH, SH, and Tropics and was computed using the MET *grid-stat* tool. Verification metrics computed for both the CONUS and global regions include the frequency bias and the FSS. For the precipitation statistics, the percentile bootstrap method (using 1000 replicates) was used to compute confidence intervals (CIs) at the 95% level. Verification data was loaded into METviewer, which was used to create the verification plots.
- **Results:** Example results are shown in Figs. 3-11, with a comprehensive set of results available for 00 and 12 UTC initializations, multiple seasons, lead times, accumulation periods, regions, and thresholds at https://dtcenter.org/eval/gmtb/2019_advphystest/FBias/ (for frequency bias) and at https://dtcenter.org/eval/gmtb/2019_advphystest/FSS/ (for FSS). Please note when interpreting the plots presented here and on the webpage that the y-axes frequently differ.

  Figures 3-7 provide examples of 6-h accumulated precipitation verification results. Figure 3 illustrates the breakdown of frequency bias by 00 and 12 UTC initializations for the full two-year test period. The breakdowns of frequency bias by season are also provided, as shown in Figure 4. FSS can be assessed by neighborhood size as a function of threshold for a variety of forecast lead times (Figure 5) or by threshold as a function of forecast lead time for a variety of neighborhood sizes (Figure 6). Shaded contour plots of FSS for each suite by neighborhood size as a function of threshold are also shown in Figure 7. These plots present similar information to Figures 5 and 6 but in a two-dimensional format with neighborhood size by lead time for 6-h accumulated precipitation.

  Figures 8-11 provide examples of 24-h accumulated precipitation verification results. Figure 8 shows the breakdown of frequency bias by region for the 12 UTC initializations over the full two-year test period. Seasonal breakdowns are also available (shown in Figure 9, NH only). FSS is provided in Figure 10 by neighborhood size as a function of threshold for a variety of forecast lead times for each of the three domains for the full year. Seasonal breakdowns are not shown for the sake of brevity. Figure 11 provides FSS by threshold as a function of forecast lead time for a variety of neighborhood sizes for two thresholds and all three domains.

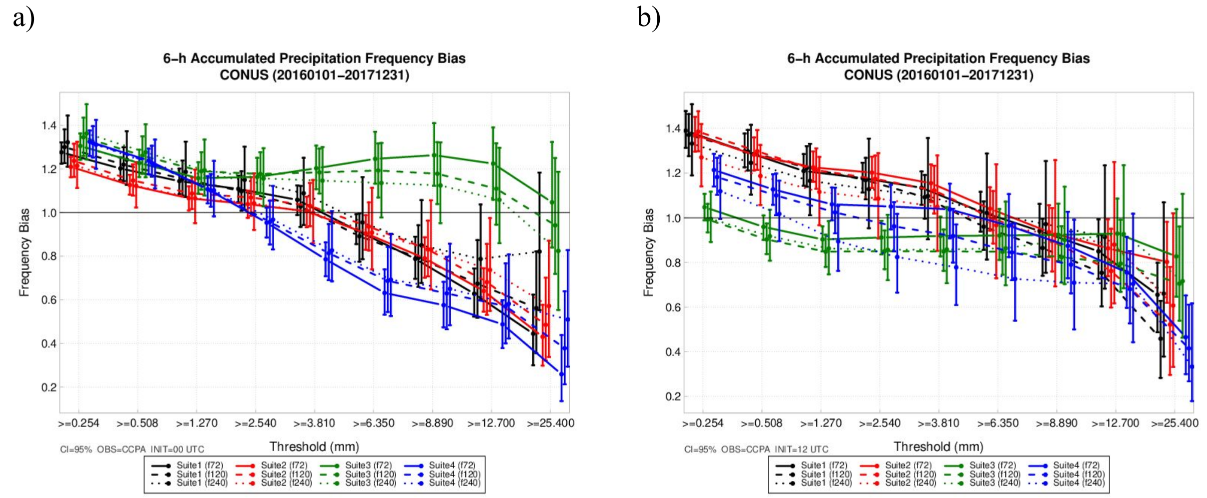


Figure 3: Frequency bias of 6-h accumulated precipitation (mm) for Suite 1 (black), Suite 2 (red), Suite 3 (green), and Suite 4 (blue) aggregated over the CONUS domain for the entire test period (20160101-20171231) for all a) 00 UTC and b) 12 UTC initializations. The 72-h, 120-h, and 240-h forecasts are represented by the solid, dashed, and dotted lines, respectively. The vertical bars surrounding the aggregate values represent the 95% CIs.

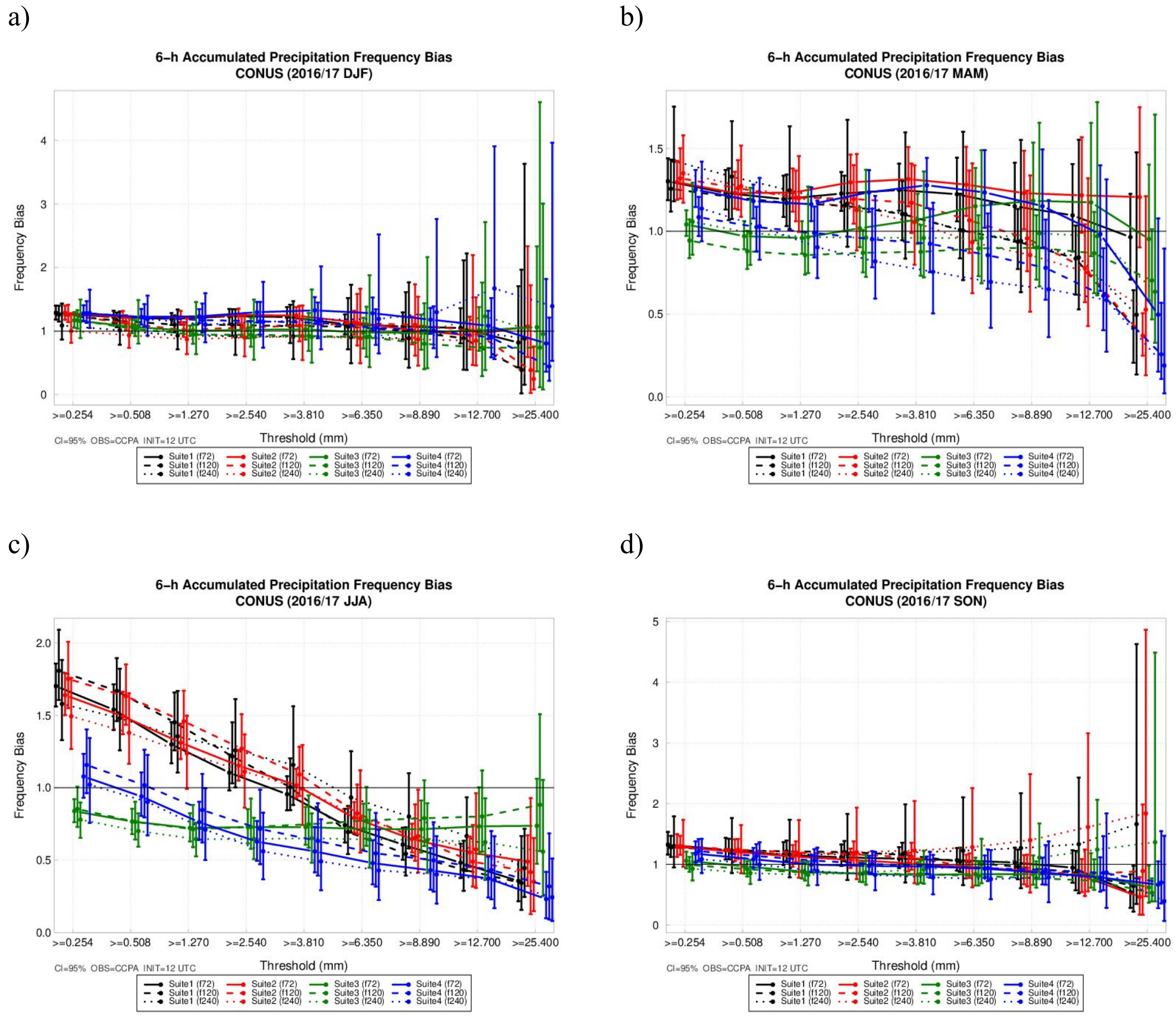


Figure 4. Same as Figure 1 (aggregated over 12 UTC initializations) but for a) DJF, b) MAM, c) JJA, and d) SON. Each plot encompasses the cases initialized during that particular season for both 2016 and 2017.

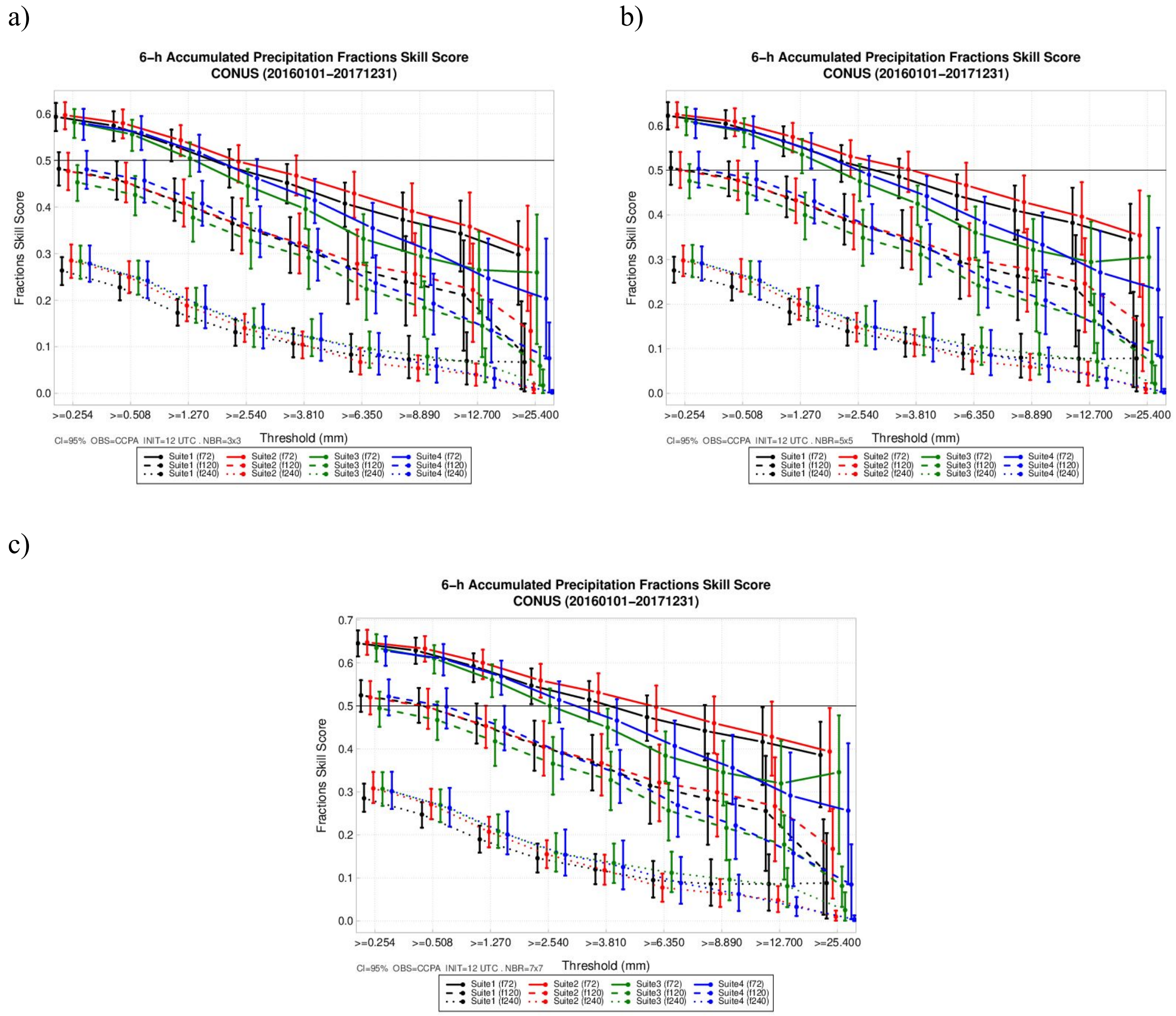


Figure 5. FSS of 6-h accumulated precipitation (mm) for Suite 1 (black), Suite 2 (red), Suite 3 (green), and Suite 4 (blue) aggregated over the CONUS domain for the entire test period (20160101-20171231) for all 12 UTC initializations using a neighborhood size of a) 9 (3×3 grid squares), b) 25 (5×5 grid squares), and c) 49 (7×7 grid squares). The 72-h, 120-h, and 240-h forecasts are represented by the solid, dashed, and dotted lines, respectively. The vertical bars surrounding the aggregate value represent the 95% CIs.

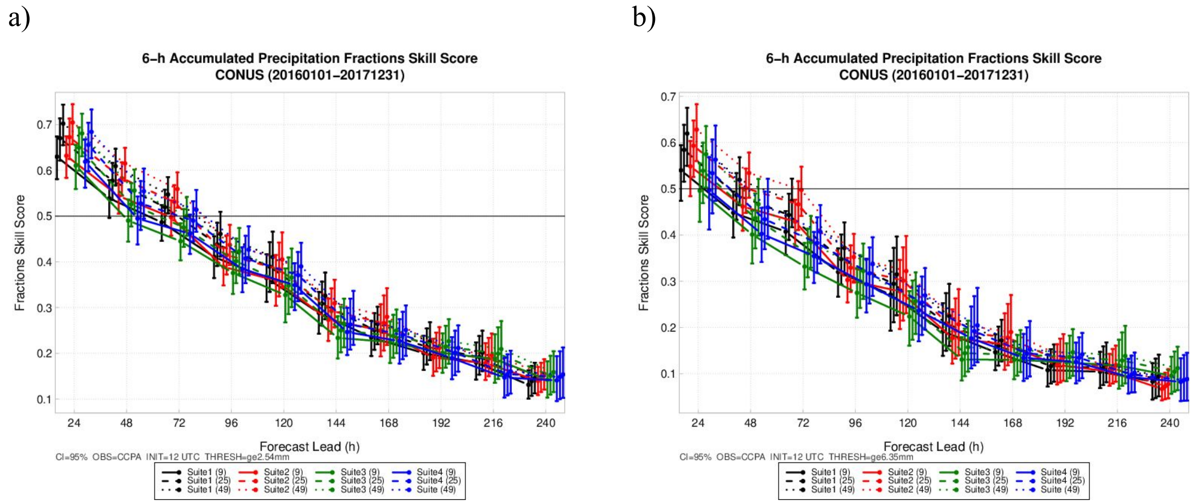


Figure 6. FSS of 6-h accumulated precipitation (mm) for Suite 1 (black), Suite 2 (red), Suite 3 (green), and Suite 4 (blue) aggregated over the CONUS domain for the entire test period (20160101-20171231) for all 12 UTC initializations at the precipitation threshold of a) ≥2.54 mm and b) ≥6.35 mm. The neighborhood size of 9 (3×3 grid squares) is represented by the solid lines, 25 (5×5 grid squares) in dashed lines, and 49 (7×7 grid squares) by dotted lines. The vertical bars surrounding the aggregate value represent the 95% CIs.

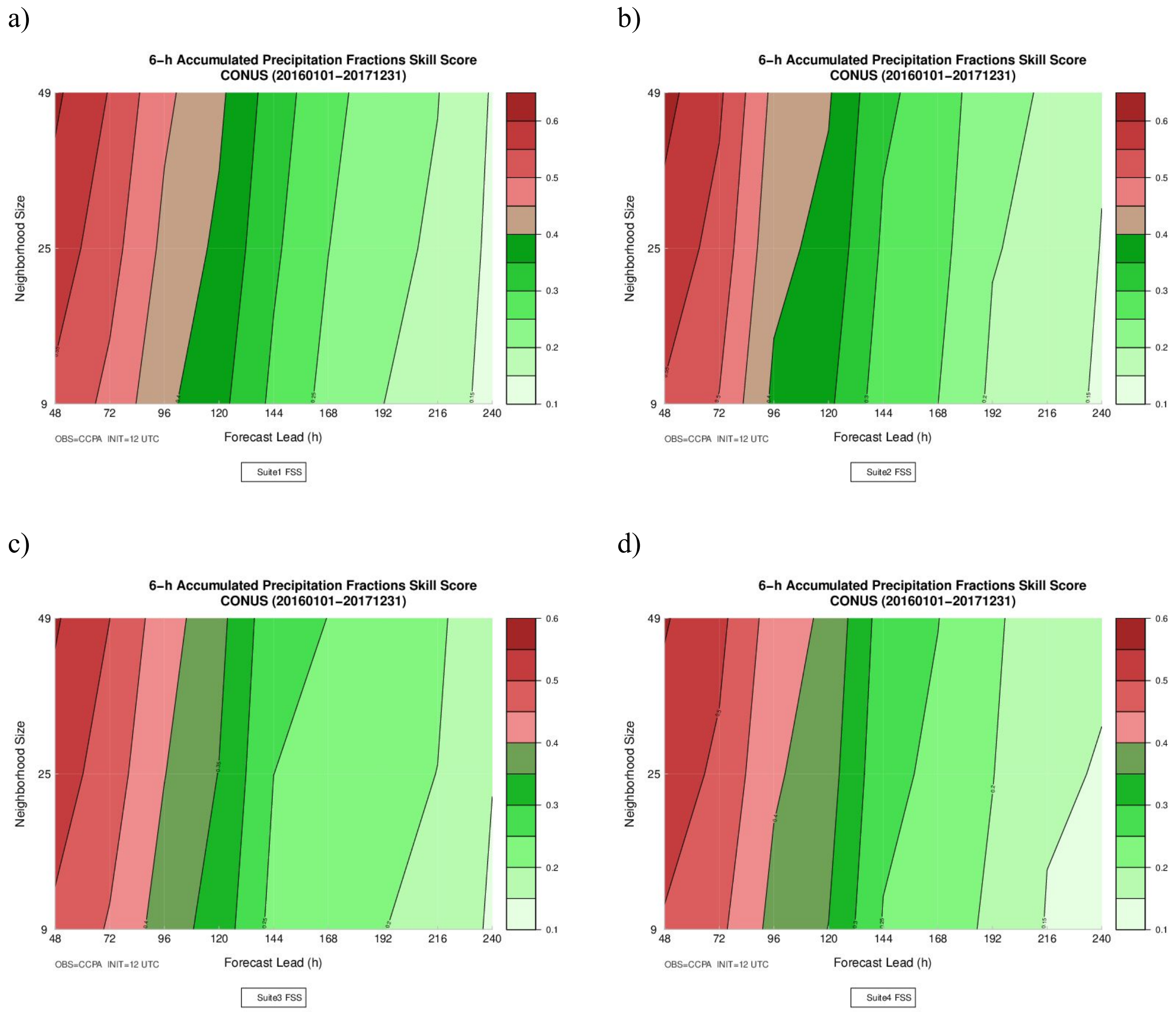


Figure 7. Contour plot of FSS of 6-h accumulated precipitation ≥2.54 mm for a) Suite 1, b) Suite 2, c) Suite 3, and d) Suite 4 aggregated over the CONUS domain for the entire test period (20160101-20171231) for all 12 UTC initializations as a function of neighborhood size, where 9 is 3×3 grid squares, 25 is 5×5 grid squares, and 49 is 7×7 grid squares, and forecast lead time.

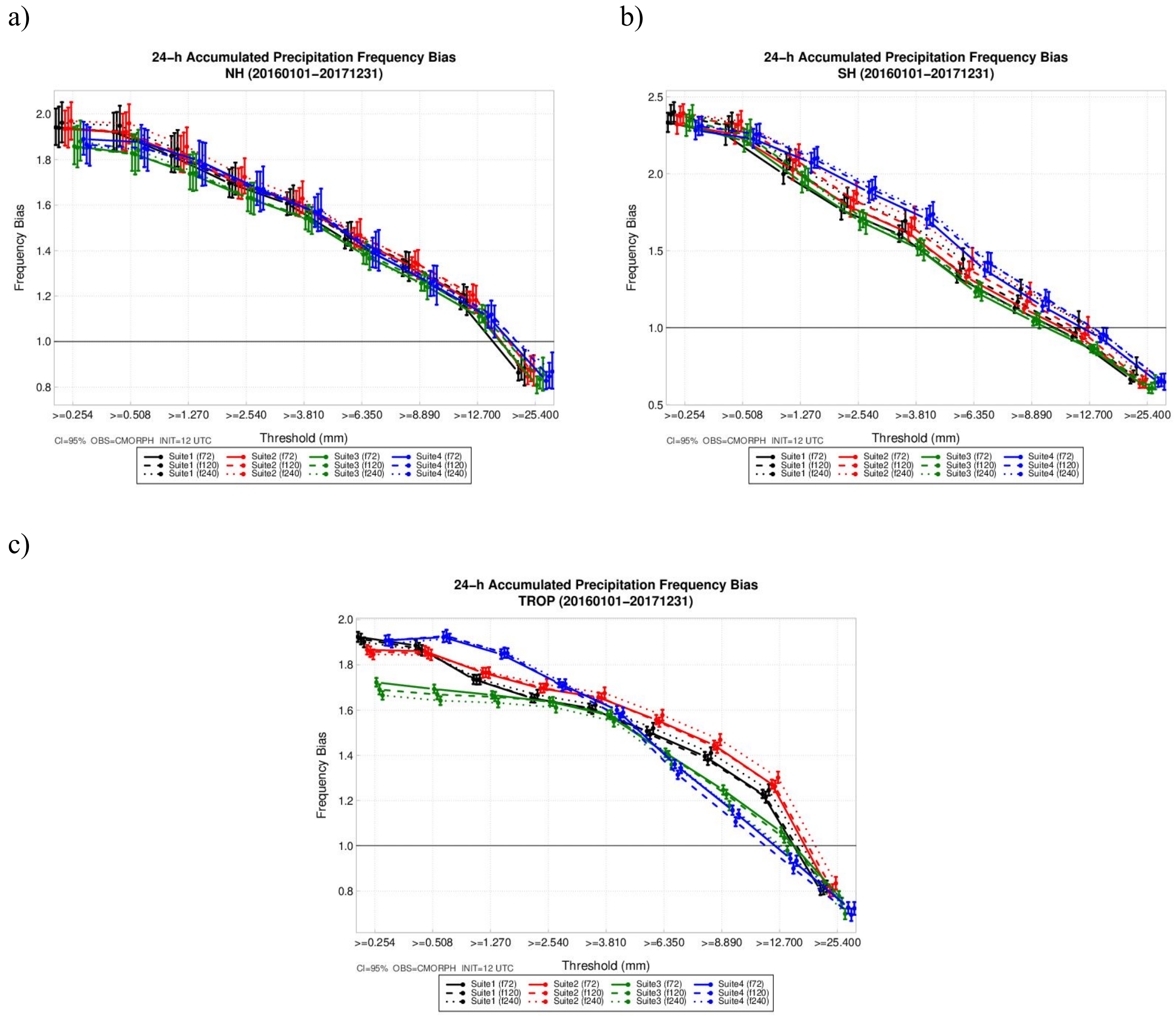


Figure 8. Frequency bias of 24-h accumulated precipitation (mm) for Suite 1 (black), Suite 2 (red), Suite 3 (green), and Suite 4 (blue) aggregated over the a) NH, b) SH, and c) Tropics domains for the entire test period (20160101-20171231) for all 12 UTC initializations. The 72-h, 120-h, and 240-h forecasts are represented by the solid, dashed, and dotted lines, respectively. The vertical bars surrounding the aggregate value represent the 95% CIs.

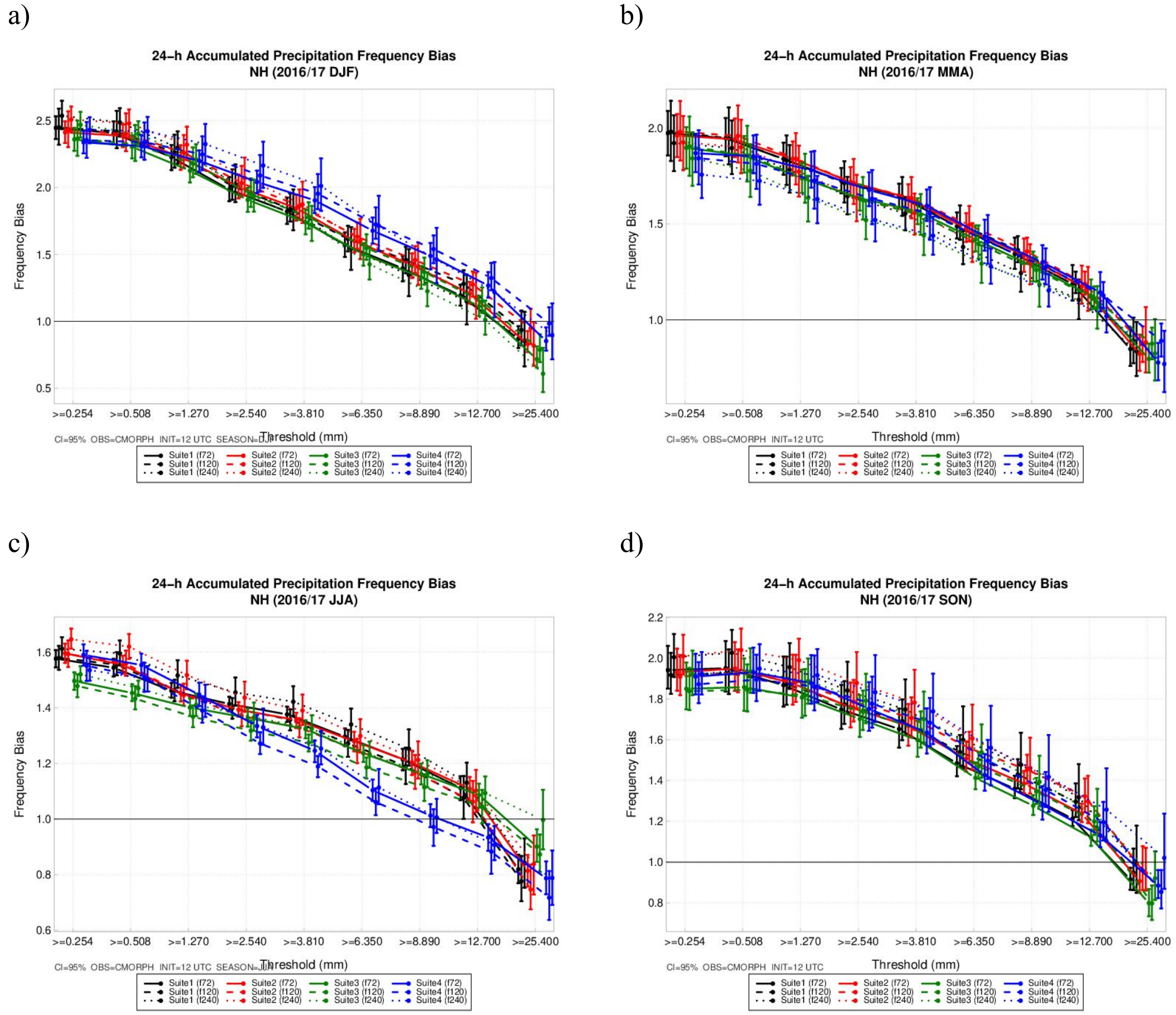


Figure 9. Frequency bias of 24-h accumulated precipitation (mm) for Suite 1 (black), Suite 2 (red), Suite 3 (green), and Suite 4 (blue) aggregated over the NH domain for a) DJF, b) MMA, c) JJA, d) SON for all the 12 UTC initializations. The 72-h, 120-h, and 240-h forecasts are represented by the solid, dashed, and dotted lines, respectively. The vertical bars surrounding the aggregate value represent the 95% CIs.

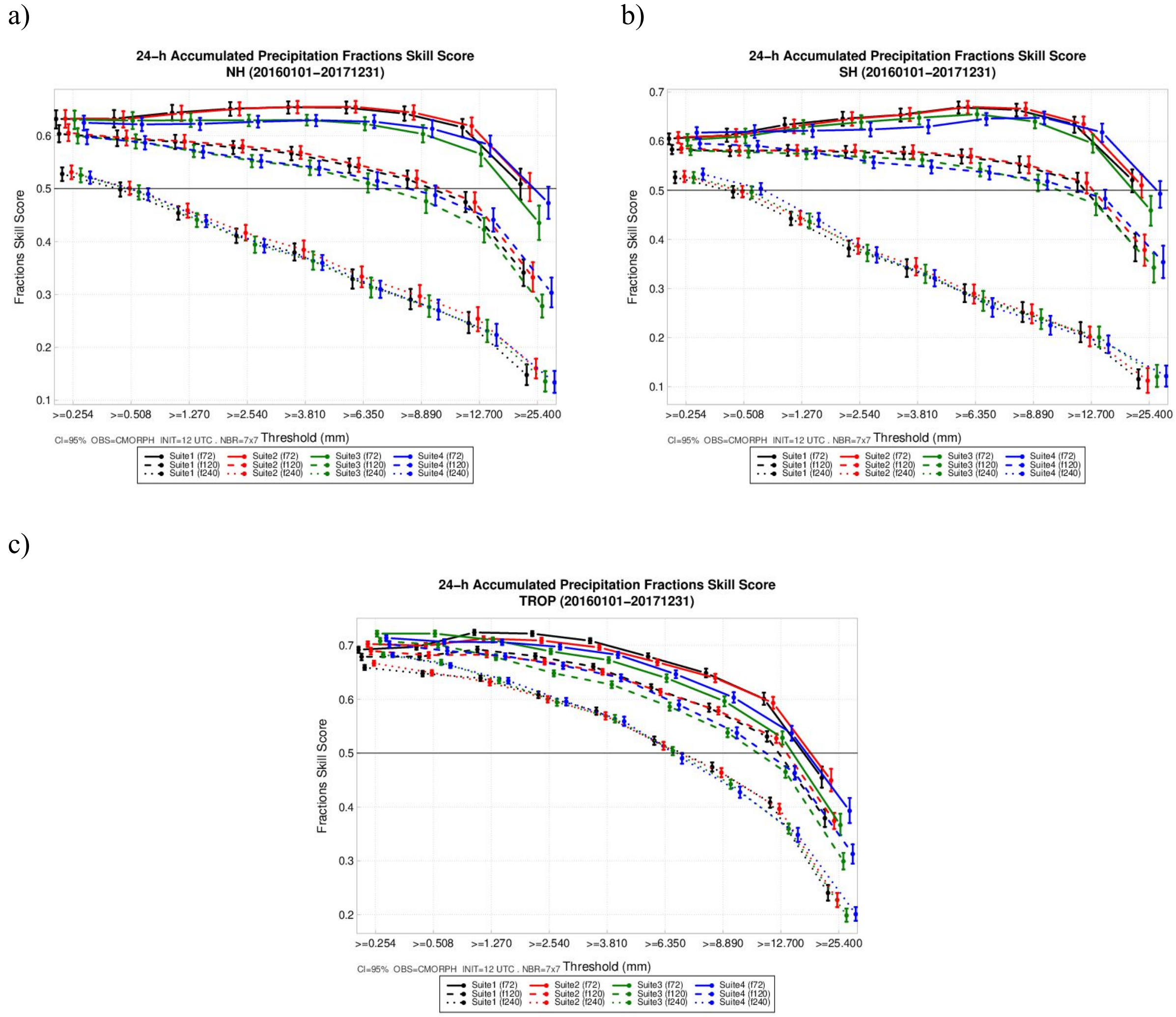


Figure 10. FSS of 24-h accumulated precipitation (mm) for Suite 1 (black), Suite 2 (red), Suite 3 (green), and Suite 4 (blue) aggregated over the a) NH, b) SH, c) Tropics domains for the entire test period (20160101-20171231) for all 12 UTC initializations using a neighborhood size of 49 (7×7 grid squares). The 72-h, 120-h, and 240-h forecasts are represented by the solid, dashed, and dotted lines, respectively. The vertical bars surrounding the aggregate value represent the 95% CIs.

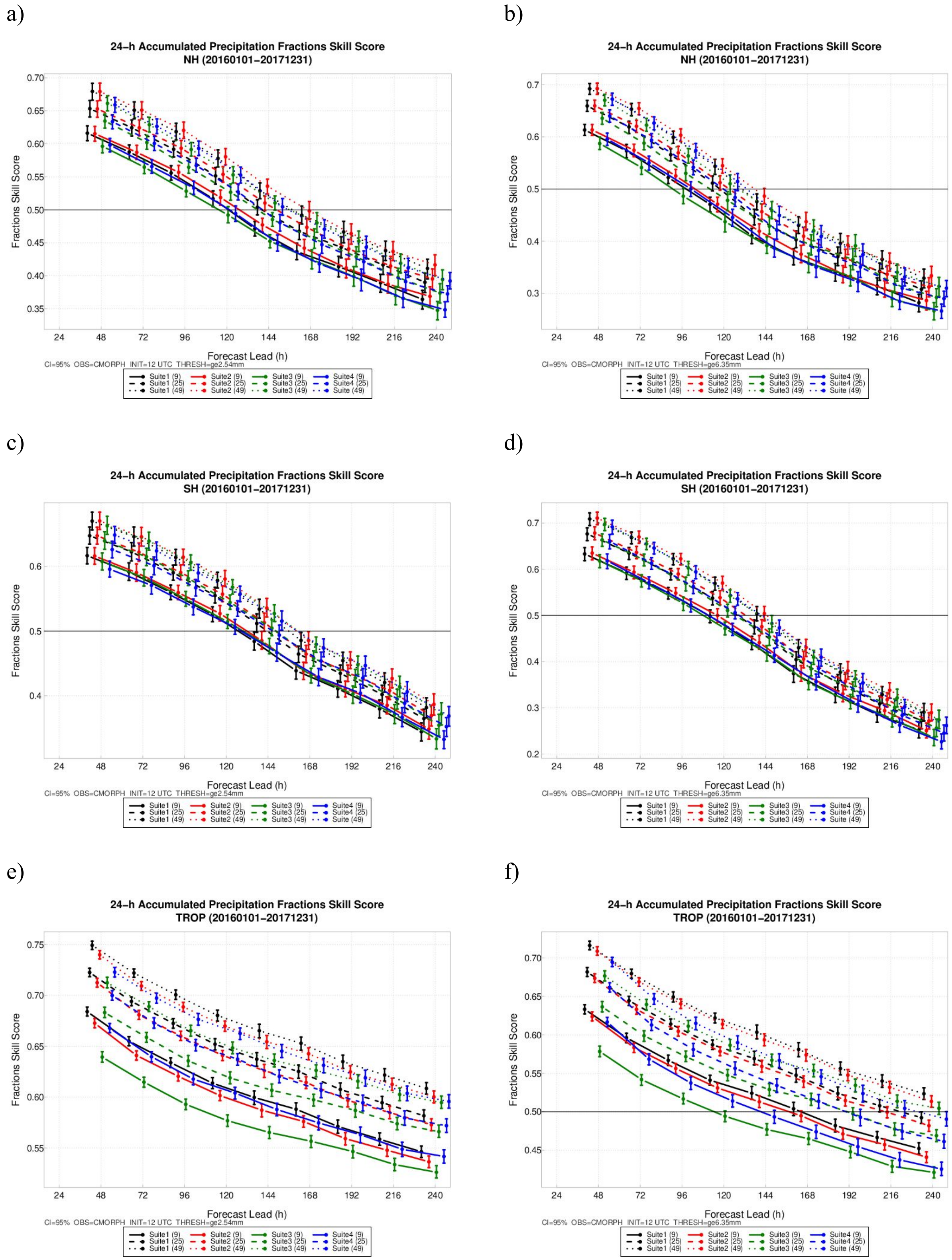
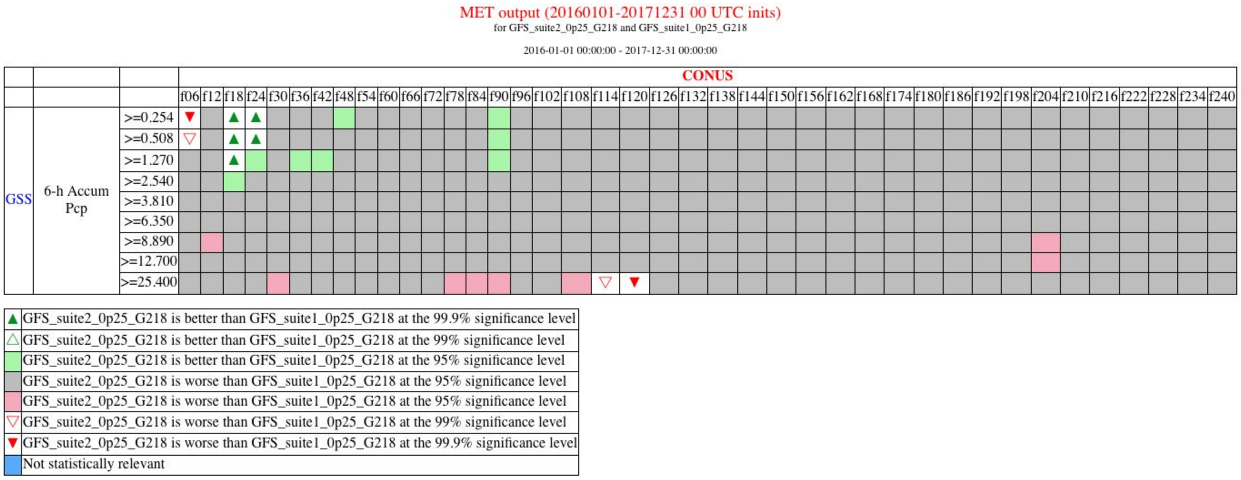


Figure 11. FSS of 24-h accumulated precipitation (mm) for Suite 1 (black), Suite 2 (red), Suite 3 (green), and Suite 4 (blue) aggregated over the NH (top row), SH (middle row), and Tropics (bottom row)

domains for the entire test period (20160101-20171231) for all 12 UTC initializations at the precipitation threshold of ≥2.54 mm (left column) and ≥6.35 mm (right column). The neighborhood size of 9 (3×3 grid squares) is represented by the solid lines, the 25 (5×5 grid squares) in dashed lines, and the 49 (7×7 grid squares) by dotted lines. The vertical bars surrounding the aggregate value represent the 95% CIs.

Scorecards: Precipitation

- **Observation data set:** Same as listed in Grid-to-grid precipitation verification section above.
- **Method:** Verification methodology same as listed in Grid-to-grid precipitation verification section above. Verification data was loaded into METviewer, which was used to create scorecards. The scorecards help identify overall patterns in the difference of performance between each of the alternative suites (Suite 2, 3, and 4) compared to Suite 1. If an alternative suite is favored over Suite 1 with SS, the cell will be indicated with a green symbol or shading; the symbol or shading will be red if Suite 1 is favored with SS.
- **Results:** Example results are shown in Figs. 12-17, with a comprehensive set of results available for 00 and 12 UTC initializations at https://dtcenter.org/eval/gmtb/2019_advphystest/scorecard/.



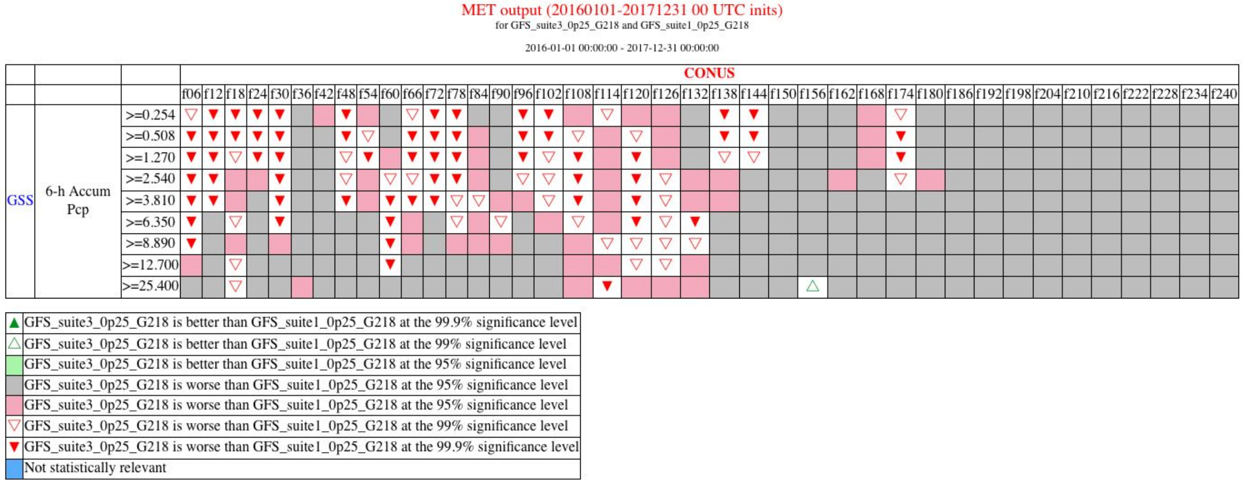
MET output (20160101-20171231 00 UTC inits)
for GFS_suite2_0p25_G218 and GFS_suite1_0p25_G218

2016-01-01 00:00:00 - 2017-12-31 00:00:00

			CONUS																																					
			f06	f12	f18	f24	f30	f36	f42	f48	f54	f60	f66	f72	f78	f84	f90	f96	f102	f108	f114	f120	f126	f132	f138	f144	f150	f156	f162	f168	f174	f180	f186	f192	f198	f204	f210	f216	f222	f228	f234	f240
GSS	6-h Accum Pcp	>=0.254	▼	▨	▲	▲	▨	▨	▨	🟩	▨	▨	▨	▨	▨	▨	🟩	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
		>=0.508	▽	▨	▲	▲	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	🟩	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
		>=1.270	▨	▨	▲	🟩	▨	🟩	🟩	▨	▨	▨	▨	▨	▨	▨	🟩	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
		>=2.540	▨	▨	🟩	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
		>=3.810	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
		>=6.350	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
		>=8.890	▨	🟥	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	🟥	▨	▨	▨	▨	▨	▨
		>=12.700	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	🟥	▨	▨	▨	▨	▨	▨
		>=25.400	▨	▨	▨	▨	🟥	▨	▨	▨	▨	▨	▨	▨	🟥	🟥	🟥	▨	▨	🟥	▽	▼	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨

▲	GFS_suite2_0p25_G218 is better than GFS_suite1_0p25_G218 at the 99.9% significance level
△	GFS_suite2_0p25_G218 is better than GFS_suite1_0p25_G218 at the 99% significance level
🟩	GFS_suite2_0p25_G218 is better than GFS_suite1_0p25_G218 at the 95% significance level
▨	GFS_suite2_0p25_G218 is worse than GFS_suite1_0p25_G218 at the 95% significance level
🟥	GFS_suite2_0p25_G218 is worse than GFS_suite1_0p25_G218 at the 95% significance level
▽	GFS_suite2_0p25_G218 is worse than GFS_suite1_0p25_G218 at the 99% significance level
▼	GFS_suite2_0p25_G218 is worse than GFS_suite1_0p25_G218 at the 99.9% significance level
🟦	Not statistically relevant

Figure 12. Scorecard documenting performance of Suite 1 and Suite 2 over the CONUS of aggregate ETS (also referred to as GSS) for 6-h accumulated precipitation by forecast lead time and precipitation threshold for 00 UTC initializations during the entire test period (20160101-20171231). Green (red) marks indicate Suite 2 (Suite 1) is better than Suite 1 (Suite 2). Statistical significance is represented by the type of marks: shading, unfilled arrows, and filled arrows indicate 95%, 99%, and 99.9% significance, respectively.



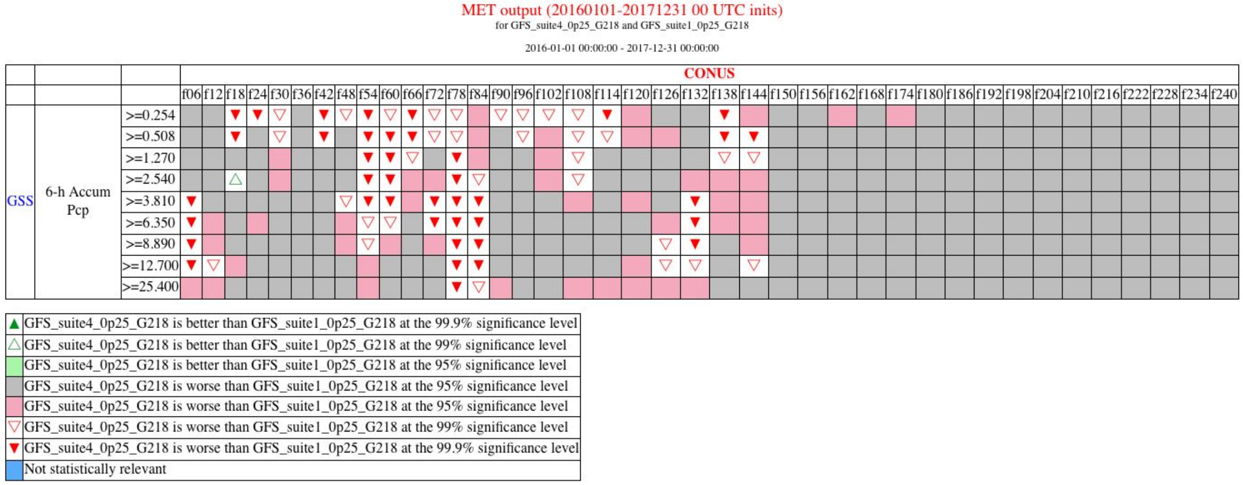
MET output (20160101-20171231 00 UTC inits)
for GFS_suite3_0p25_G218 and GFS_suite1_0p25_G218

2016-01-01 00:00:00 - 2017-12-31 00:00:00

			CONUS																																							
			f06	f12	f18	f24	f30	f36	f42	f48	f54	f60	f66	f72	f78	f84	f90	f96	f102	f108	f114	f120	f126	f132	f138	f144	f150	f156	f162	f168	f174	f180	f186	f192	f198	f204	f210	f216	f222	f228	f234	f240
GSS	6-h Accum Pcp	>=0.254	▽	▼	▼	▼	▼	▒	🟥	▼	🟥	▒	▽	▼	▼	▒	▒	▼	▼	🟥	▽	🟥	🟥	▒	▼	▼	▒	▒	▒	🟥	▽	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
		>=0.508	▼	▼	▼	▼	▼	▒	▒	▼	▽	▒	▼	▼	▼	🟥	▒	▼	▼	▽	🟥	▽	🟥	▒	▼	▼	▒	▒	▒	🟥	▼	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
		>=1.270	▼	▼	▽	▼	▼	▒	▒	▽	▼	🟥	▼	▼	▼	🟥	▒	▼	▽	▼	🟥	▼	🟥	▒	▽	▽	▒	▒	▒	🟥	▼	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
		>=2.540	▼	▼	🟥	🟥	▼	▒	▒	▽	🟥	▽	▽	▼	▼	🟥	▒	▽	▽	▼	🟥	▼	▽	🟥	🟥	▒	▒	▒	🟥	▒	▽	🟥	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
		>=3.810	▼	▼	🟥	▒	▼	▒	▒	▼	🟥	▼	▼	▼	▽	▽	🟥	🟥	▽	▼	🟥	▼	▽	🟥	🟥	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
		>=6.350	▼	▒	▽	▒	▼	▒	▒	▒	▒	▼	🟥	▒	▽	🟥	▽	▒	🟥	▽	🟥	▼	▽	▼	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
		>=8.890	▼	▒	🟥	▒	🟥	▒	▒	▒	▒	▼	🟥	▒	🟥	🟥	🟥	▒	▒	🟥	▽	▽	▽	▽	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
		>=12.700	🟥	▒	▽	▒	▒	▒	▒	▒	▒	▼	▒	▒	▒	▒	▒	▒	▒	🟥	🟥	▽	▽	🟥	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
		>=25.400	▒	▒	▽	▒	▒	🟥	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	🟥	▼	🟥	🟥	🟥	▒	▒	▒	△	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒

▲	GFS_suite3_0p25_G218 is better than GFS_suite1_0p25_G218 at the 99.9% significance level
△	GFS_suite3_0p25_G218 is better than GFS_suite1_0p25_G218 at the 99% significance level
🟩	GFS_suite3_0p25_G218 is better than GFS_suite1_0p25_G218 at the 95% significance level
▒	GFS_suite3_0p25_G218 is worse than GFS_suite1_0p25_G218 at the 95% significance level
🟥	GFS_suite3_0p25_G218 is worse than GFS_suite1_0p25_G218 at the 95% significance level
▽	GFS_suite3_0p25_G218 is worse than GFS_suite1_0p25_G218 at the 99% significance level
▼	GFS_suite3_0p25_G218 is worse than GFS_suite1_0p25_G218 at the 99.9% significance level
🟦	Not statistically relevant

Figure 13. Same as Figure 12, except for Suite 1 and Suite 3.



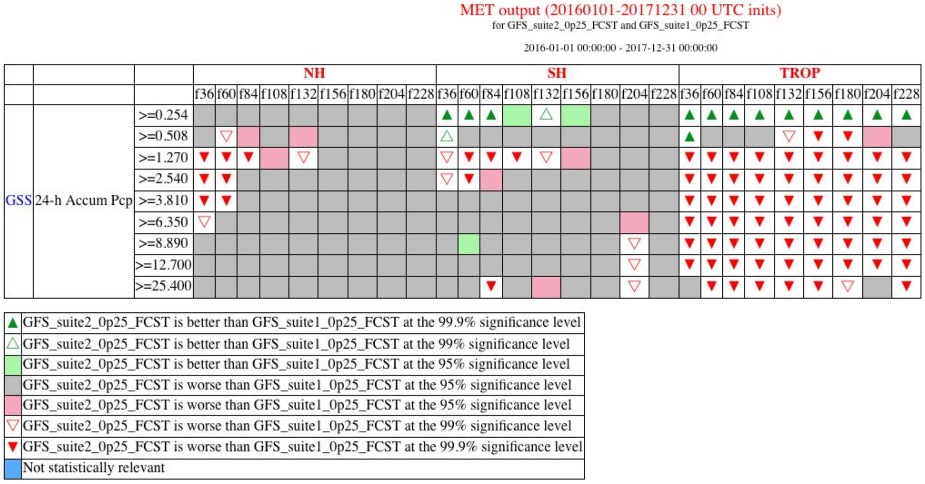
MET output (20160101-20171231 00 UTC inits)
for GFS_suite4_0p25_G218 and GFS_suite1_0p25_G218

2016-01-01 00:00:00 - 2017-12-31 00:00:00

			CONUS																																							
			f06	f12	f18	f24	f30	f36	f42	f48	f54	f60	f66	f72	f78	f84	f90	f96	f102	f108	f114	f120	f126	f132	f138	f144	f150	f156	f162	f168	f174	f180	f186	f192	f198	f204	f210	f216	f222	f228	f234	f240
GSS	6-h Accum Pcp	>=0.254	▒	▒	▼	▼	▽	▒	▼	▽	▼	▽	▼	▽	▽	🟥	▽	▽	▽	▽	▼	🟥	▒	▒	▼	🟥	▒	▒	🟥	▒	🟥	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
		>=0.508	▒	▒	▼	▒	▽	▒	▼	▒	▼	▼	▼	▽	▽	🟥	▒	▽	🟥	▽	▽	🟥	🟥	▒	▼	▼	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
		>=1.270	▒	▒	▒	▒	🟥	▒	▒	▒	▼	▼	▽	▒	▼	🟥	▒	▒	🟥	▽	▒	▒	▒	▒	▽	▽	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
		>=2.540	▒	▒	△	▒	🟥	▒	▒	▒	▼	▼	🟥	🟥	▼	▽	▒	▒	🟥	▽	▒	▒	▒	🟥	🟥	🟥	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
		>=3.810	▼	▒	▒	▒	▒	▒	▒	▽	▼	▼	🟥	▼	▼	▼	▒	▒	▒	🟥	▒	🟥	▒	▼	🟥	🟥	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
		>=6.350	▼	🟥	▒	🟥	▒	▒	▒	🟥	▽	▽	▒	▼	▼	▼	▒	▒	▒	▒	▒	▒	🟥	▼	🟥	🟥	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
		>=8.890	▼	🟥	▒	▒	▒	▒	▒	🟥	▽	🟥	▒	🟥	▼	▼	▒	▒	▒	▒	▒	▒	▽	▼	▒	🟥	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
		>=12.700	▼	▽	🟥	▒	▒	▒	▒	▒	🟥	▒	▒	▒	▼	▼	▒	▒	▒	▒	▒	🟥	▽	▽	▒	▽	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
		>=25.400	🟥	🟥	▒	▒	▒	▒	▒	▒	🟥	▒	▒	▒	▼	▽	🟥	▒	▒	🟥	🟥	🟥	🟥	🟥	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒

▲	GFS_suite4_0p25_G218 is better than GFS_suite1_0p25_G218 at the 99.9% significance level
△	GFS_suite4_0p25_G218 is better than GFS_suite1_0p25_G218 at the 99% significance level
🟩	GFS_suite4_0p25_G218 is better than GFS_suite1_0p25_G218 at the 95% significance level
▒	GFS_suite4_0p25_G218 is worse than GFS_suite1_0p25_G218 at the 95% significance level
🟥	GFS_suite4_0p25_G218 is worse than GFS_suite1_0p25_G218 at the 95% significance level
▽	GFS_suite4_0p25_G218 is worse than GFS_suite1_0p25_G218 at the 99% significance level
▼	GFS_suite4_0p25_G218 is worse than GFS_suite1_0p25_G218 at the 99.9% significance level
🟦	Not statistically relevant

Figure 14. Same as Figure 12, except for Suite 1 and Suite 4.



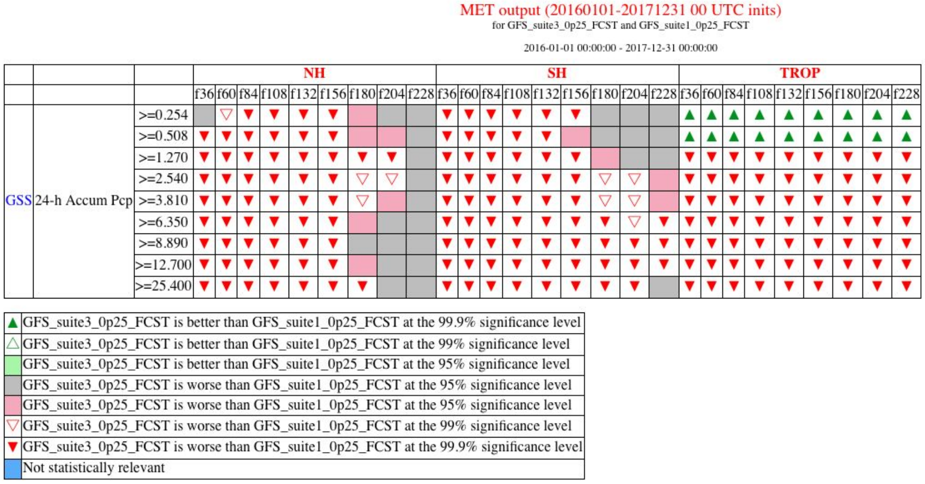
MET output (20160101-20171231 00 UTC inits)
for GFS_suite2_0p25_FCST and GFS_suite1_0p25_FCST

2016-01-01 00:00:00 - 2017-12-31 00:00:00

			NH									SH									TROP								
			f36	f60	f84	f108	f132	f156	f180	f204	f228	f36	f60	f84	f108	f132	f156	f180	f204	f228	f36	f60	f84	f108	f132	f156	f180	f204	f228
GSS	24-h Accum Pcp	>=0.254	▒	▒	▒	▒	▒	▒	▒	▒	▒	▲	▲	▲	🟩	△	🟩	▒	▒	▒	▲	▲	▲	▲	▲	▲	▲	▲	▲
		>=0.508	▒	▽	🟥	▒	🟥	▒	▒	▒	▒	△	▒	▒	▒	▒	▒	▒	▒	▒	▲	▒	▒	▒	▽	▼	▼	🟥	▒
		>=1.270	▼	▼	▼	🟥	▽	▒	▒	▒	▒	▽	▼	▼	▼	▽	🟥	▒	▒	▒	▼	▼	▼	▼	▼	▼	▼	▼	▼
		>=2.540	▼	▼	▒	▒	▒	▒	▒	▒	▒	▽	▼	🟥	▒	▒	▒	▒	▒	▒	▼	▼	▼	▼	▼	▼	▼	▼	▼
		>=3.810	▼	▼	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▼	▼	▼	▼	▼	▼	▼	▼	▼
		>=6.350	▽	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	🟥	▒	▼	▼	▼	▼	▼	▼	▼	▼	▼
		>=8.890	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	🟩	▒	▒	▒	▒	▒	▽	▒	▼	▼	▼	▼	▼	▼	▼	▼	▼
		>=12.700	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▽	▒	▼	▼	▼	▼	▼	▼	▼	▼	▼
		>=25.400	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▼	▒	🟥	▒	▒	▽	▒	▒	▼	▼	▼	▼	▼	▽	▒	▼

▲	GFS_suite2_0p25_FCST is better than GFS_suite1_0p25_FCST at the 99.9% significance level
△	GFS_suite2_0p25_FCST is better than GFS_suite1_0p25_FCST at the 99% significance level
🟩	GFS_suite2_0p25_FCST is better than GFS_suite1_0p25_FCST at the 95% significance level
▒	GFS_suite2_0p25_FCST is worse than GFS_suite1_0p25_FCST at the 95% significance level
🟥	GFS_suite2_0p25_FCST is worse than GFS_suite1_0p25_FCST at the 95% significance level
▽	GFS_suite2_0p25_FCST is worse than GFS_suite1_0p25_FCST at the 99% significance level
▼	GFS_suite2_0p25_FCST is worse than GFS_suite1_0p25_FCST at the 99.9% significance level
🟦	Not statistically relevant

Figure 15. Scorecard documenting performance of Suite 1 and Suite 2 over the NH, SH, and Tropics of aggregate ETS for 24-h accumulated precipitation by forecast lead time and precipitation threshold for 00 UTC initializations during the entire test period (20160101-20171231). Green (red) marks indicate Suite 2 (Suite 1) is better than Suite 1 (Suite 2). Statistical significance is represented by the type of marks: shading, unfilled arrows, and filled arrows indicate 95%, 99%, and 99.9% significance, respectively.



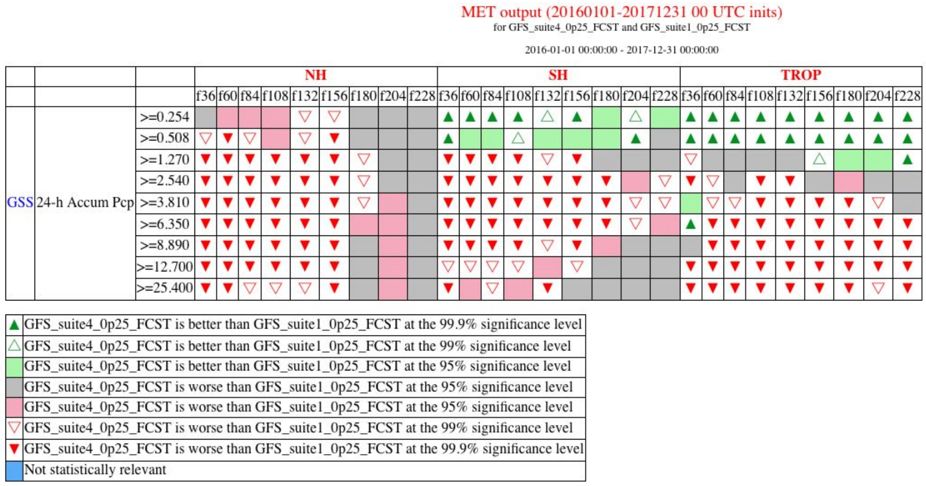
MET output (20160101-20171231 00 UTC inits)
for GFS_suite3_0p25_FCST and GFS_suite1_0p25_FCST

2016-01-01 00:00:00 - 2017-12-31 00:00:00

			NH									SH									TROP								
			f36	f60	f84	f108	f132	f156	f180	f204	f228	f36	f60	f84	f108	f132	f156	f180	f204	f228	f36	f60	f84	f108	f132	f156	f180	f204	f228
GSS	24-h Accum Pcp	>=0.254	▒	▽	▼	▼	▼	▼	🟥	▒	▒	▼	▼	▼	▼	▼	▼	▒	▒	▒	▲	▲	▲	▲	▲	▲	▲	▲	▲
		>=0.508	▼	▼	▼	▼	▼	▼	🟥	🟥	▒	▼	▼	▼	▼	▼	🟥	▒	▒	▒	▲	▲	▲	▲	▲	▲	▲	▲	▲
		>=1.270	▼	▼	▼	▼	▼	▼	▼	▼	▒	▼	▼	▼	▼	▼	▼	🟥	▒	▒	▼	▼	▼	▼	▼	▼	▼	▼	▼
		>=2.540	▼	▼	▼	▼	▼	▼	▽	▽	▒	▼	▼	▼	▼	▼	▼	▽	▽	🟥	▼	▼	▼	▼	▼	▼	▼	▼	▼
		>=3.810	▼	▼	▼	▼	▼	▼	▽	🟥	▒	▼	▼	▼	▼	▼	▼	▽	▽	🟥	▼	▼	▼	▼	▼	▼	▼	▼	▼
		>=6.350	▼	▼	▼	▼	▼	▼	🟥	▒	▒	▼	▼	▼	▼	▼	▼	▼	▽	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼
		>=8.890	▼	▼	▼	▼	▼	▼	▒	▒	▒	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼
		>=12.700	▼	▼	▼	▼	▼	▼	🟥	▒	▒	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼
		>=25.400	▼	▼	▼	▼	▼	▼	▼	▒	▒	▼	▼	▼	▼	▼	▼	▼	▼	▒	▼	▼	▼	▼	▼	▼	▼	▼	▼

▲	GFS_suite3_0p25_FCST is better than GFS_suite1_0p25_FCST at the 99.9% significance level
△	GFS_suite3_0p25_FCST is better than GFS_suite1_0p25_FCST at the 99% significance level
🟩	GFS_suite3_0p25_FCST is better than GFS_suite1_0p25_FCST at the 95% significance level
▒	GFS_suite3_0p25_FCST is worse than GFS_suite1_0p25_FCST at the 95% significance level
🟥	GFS_suite3_0p25_FCST is worse than GFS_suite1_0p25_FCST at the 95% significance level
▽	GFS_suite3_0p25_FCST is worse than GFS_suite1_0p25_FCST at the 99% significance level
▼	GFS_suite3_0p25_FCST is worse than GFS_suite1_0p25_FCST at the 99.9% significance level
🟦	Not statistically relevant

Figure 16. Same as Figure 15, except for Suite 3 and Suite 1.



MET output (20160101-20171231 00 UTC inits)

for GFS_suite4_0p25_FCST and GFS_suite1_0p25_FCST

2016-01-01 00:00:00 - 2017-12-31 00:00:00

			NH									SH									TROP								
			f36	f60	f84	f108	f132	f156	f180	f204	f228	f36	f60	f84	f108	f132	f156	f180	f204	f228	f36	f60	f84	f108	f132	f156	f180	f204	f228
GSS	24-h Accum Pcp	>=0.254	▨	🟥	🟥	🟥	▽	▽	▨	▨	▨	▲	▲	▲	▲	△	▲	🟩	△	🟩	▲	▲	▲	▲	▲	▲	▲	▲	▲
		>=0.508	▽	▼	▽	🟥	▽	▼	▨	▨	▨	▲	🟩	🟩	△	🟩	🟩	🟩	▲	▨	▲	▲	▲	▲	▲	▲	▲	▲	▲
		>=1.270	▼	▼	▼	▼	▼	▼	▽	▨	▨	▼	▼	▼	▼	▽	▼	▨	▨	▨	▽	▨	▨	▨	▨	△	🟩	🟩	▲
		>=2.540	▼	▼	▼	▼	▼	▼	▽	▨	▨	▼	▼	▼	▼	▼	▼	▼	🟥	▽	▼	▽	▨	▼	▼	▨	🟥	▨	▨
		>=3.810	▼	▼	▼	▼	▼	▼	▽	🟥	▨	▼	▼	▼	▼	▼	▼	▼	▽	▽	🟩	▽	▽	▼	▼	▼	▼	▽	▨
		>=6.350	▼	▼	▼	▼	▼	▼	🟥	🟥	▨	▼	▼	▼	▼	▼	▼	▼	▽	🟥	▲	▼	▼	▼	▼	▼	▼	▼	▼
		>=8.890	▼	▼	▼	▼	▼	▼	▨	🟥	▨	▼	▼	▼	▼	▽	▼	🟥	▨	▨	▨	▼	▼	▼	▼	▼	▼	▼	▼
		>=12.700	▼	▼	▼	▼	▼	▼	▨	🟥	▨	▽	▽	▽	▽	🟥	▽	▨	▨	▨	▼	▼	▼	▼	▼	▼	▼	▼	▼
		>=25.400	▼	▼	▽	▽	▽	▼	▨	🟥	▨	▼	🟥	▽	🟥	▼	▨	▨	▨	▨	▼	▼	▼	▼	▼	▼	▼	▽	▼

▲	GFS_suite4_0p25_FCST is better than GFS_suite1_0p25_FCST at the 99.9% significance level
△	GFS_suite4_0p25_FCST is better than GFS_suite1_0p25_FCST at the 99% significance level
🟩	GFS_suite4_0p25_FCST is better than GFS_suite1_0p25_FCST at the 95% significance level
▨	GFS_suite4_0p25_FCST is worse than GFS_suite1_0p25_FCST at the 95% significance level
🟥	GFS_suite4_0p25_FCST is worse than GFS_suite1_0p25_FCST at the 95% significance level
▽	GFS_suite4_0p25_FCST is worse than GFS_suite1_0p25_FCST at the 99% significance level
▼	GFS_suite4_0p25_FCST is worse than GFS_suite1_0p25_FCST at the 99.9% significance level
🟦	Not statistically relevant

Figure 17. Same as Figure 15, except for Suite 1 and Suite 4.

Mean state bias of precipitation

- **Benchmark:** The daily averaged 0.25º x 0.25º CMORPH precipitation analyses.
- **Method:** Given that the model was initialized every 5 days, to make the verification periods consistent for all the forecast lead times, we referred the day-5 precipitation forecast as the total precipitation during 84-108h for 00 UTC cycle and 108-132h for 12 UTC cycle, respectively. The verification days were 2016011100, 2016011612, 2016012100,..., and 2017123100 (145 days in total).
- **Results:** Figure 18 shows that all four suites are associated with dry biases over the African monsoon region and the Central U.S. Suites 1, 2, and 3 also underpredict precipitation over the Indo-Pacific warm pool and the Amazon tropical rainforest. On the other hand, Suites 1, 2, and 4 slightly overestimate rainfall over the Caribbean Sea and the equatorial Atlantic Ocean, and Suite 4 has a wet bias over the Indo-Pacific warm pool.

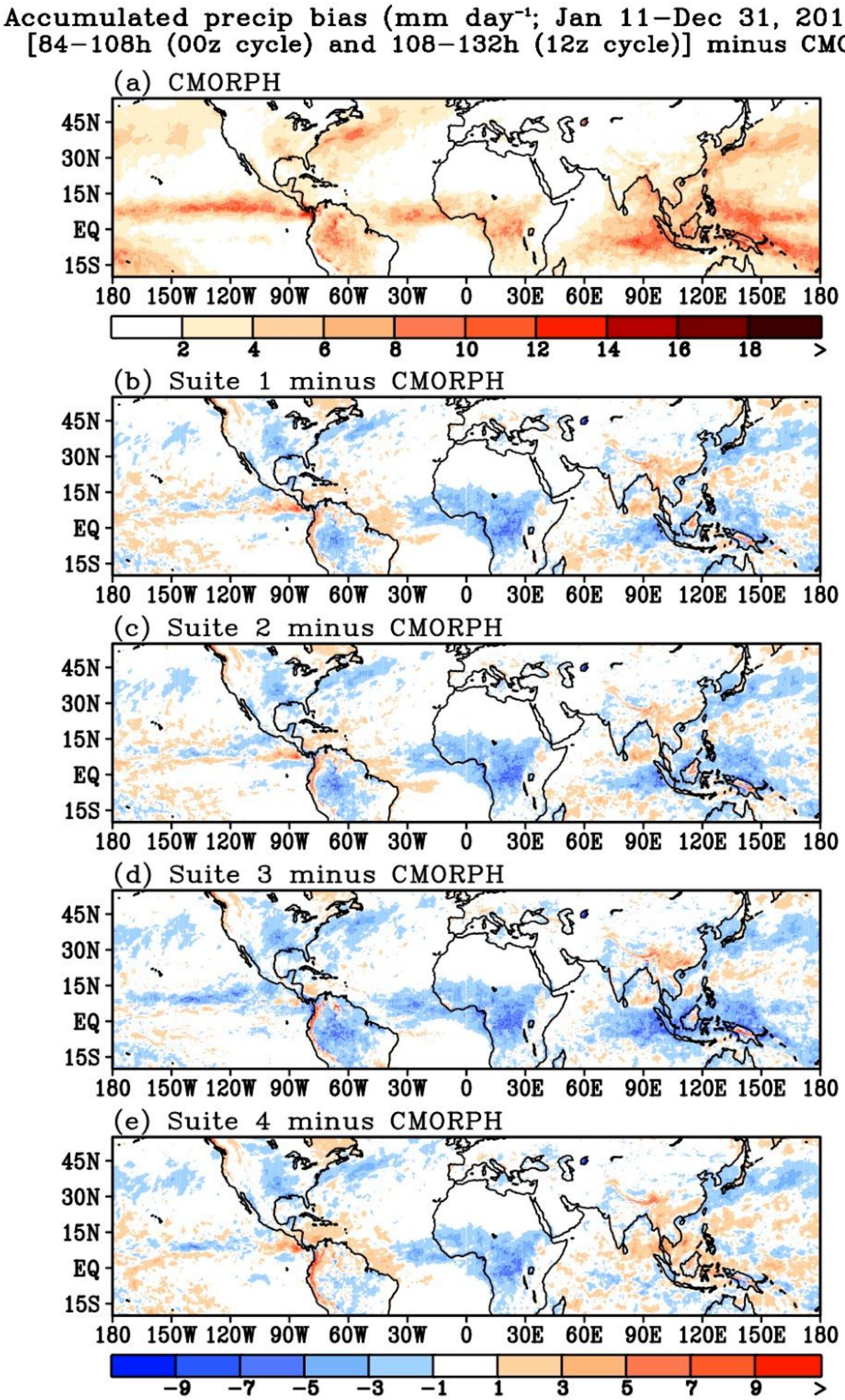

Accumulated precip bias (mm day⁻¹; Jan 11–Dec 31, 2016–17)
[84–108h (00z cycle) and 108–132h (12z cycle)] minus CMORPH
(a) CMORPH
45N
30N
15N
EQ
15S
180 150W 120W 90W 60W 30W 0 30E 60E 90E 120E 150E 180
2 4 6 8 10 12 14 16 18 >
(b) Suite 1 minus CMORPH
(c) Suite 2 minus CMORPH
(d) Suite 3 minus CMORPH
(e) Suite 4 minus CMORPH
−9 −7 −5 −3 −1 1 3 5 7 9 >

Figure 18. The average daily precipitation rate (mm day^{-1}) in the tropics for (a) the CMORPH analyses and (b)-(e) differences between Suite 1-4 and the CMORPH anaylses.

Relationship between precipitation and precipitable water

- **Benchmark:** The observed relationship between precipitation and PW was examined using two datasets: PW from the SSMIS v7+v8 (f16+f17+f18 satellites) and precipitation from the CMORPH analyses. This was due to very limited sample size of precipitation during 2016-17 in the SSMIS, so the CMORPH precipitation analyses were used instead.
- **Method:** The analysis is ocean-only because the SSMIS does not cover land, and the relationships examined are quite different between ocean and land.
- **Results:** Figure 19a shows that all four suites can capture the nonlinear relationship between precipitation and PW. But Suites 1, 2, and 3 generate too much precipitation for a given amount of PW, implying that precipitation may be triggered too early in the model in terms of the PW accumulation. This is likely a typical issue of simplified Arakawa-Schubert (SAS) scheme. Suites 1 and 2 used scale aware (SA) SAS, which suggests that SA may not help alleviate the precipitation early triggering problem. Suite 4 used SAS/aerosol-aware (AA)-Grell-Freitas (GF) deep convection scheme and had the most realistic curve. Figure 19b shows a dry bias in the model initialization (day 0) compared to the observations. Such initialized dry bias leads to larger dry biases in the forecasts for all four suites, which is consistent with the dry bias shown in Figure 18. Suite 4 has the mildest shift of the PW distribution. It is reasonable to expect that the PW distribution may be more realistic in Suite 4 given data assimilation and improved model initialization.

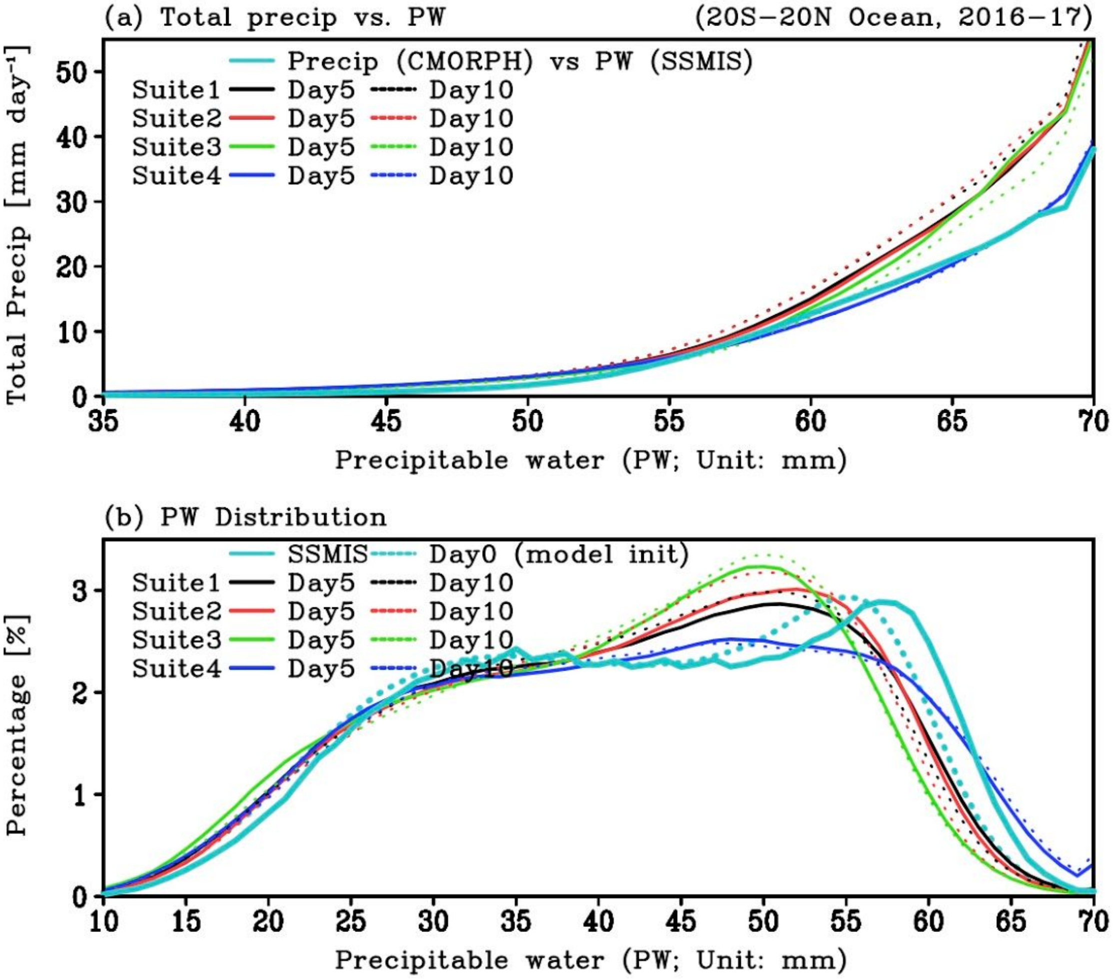


Figure 19. (a) The average daily precipitation rate (mm day^{-1}) stratified with 1-mm-wide bins of the PW (mm) and (b) the PW probability distribution (%) over the tropical ocean basins (20ºS-20ºN around the globe).

Vertical profiles of relative humidity (RH) and diabatic heating rate (Q1) for different PW thresholds

- **Benchmark:** The fifth generation of ECMWF atmospheric reanalyses of the global climate (ERA5), which has much higher spatial and temporal resolution and much improved troposphere compared to ERA-interim.
- **Method:** RH and diabatic heating rate (Q1) were stratified with 1-mm-wide bins of the PW. The purpose of this analysis is to provide information on the vertical distribution of moisture and heating. The diabatic heating rates were calculated for both ERA5 and model forecasts using five pressure-level variables including temperature, mixing ratio, zonal, meridional and vertical wind components (Yanai 1973; Yanai and Tomita 1998). For RH, ERA5 likely provides the best available three dimensional moisture field. For diabatic heating, the one calculated from ERA5 was also used as the truth mostly due to limited observational data over the ocean. We, therefore, treated the differences between the forecasts and ERA5 as the biases. The tropical ocean and land will be examined separately to shed light on possible model deficiency in the cumulus schemes.
- **Results:**
    1) In ERA5, large RH extends to the upper troposphere for large PW values over the ocean, which is associated with tropical convection (Fig. 20a). Figure 20b shows that the model initialization resembles ERA5 with only 2-5% less RH in the middle to upper troposphere than ERA5. On day 5, all suites exhibit a dry bias in the lower troposphere (lower than 700 hPa) and a moist bias above. Since the lower troposphere contains most of the water vapor in column, the dry bias largely contributes to the dry bias shown in the PDF of PW (Fig. 19a). When PW is greater than 55 mm, the dry bias within the marine boundary layer and the moist bias in the free atmosphere suggest an issue of convection development in the cumulus schemes used in Suites 1, 2, and 3. Turbulence, thermals, and/or convection may hyperactively transport moisture upward to the free atmosphere, which leads to insufficient moisture in the boundary layer and convection triggered too early, as shown in Fig. 19a. Suite 4 has a relatively smaller moisture bias for PW greater than 55 mm than the other suites. Near the surface, a dry bias can be found for a wide range of PW thresholds for all suites. In contrast to the surface dry bias, all suites show different degrees of wet biases near 900 hPa, which may suggest an air-sea coupling issue and/or the models being overmixed at lower levels (see Fig. 30).
    2) A trimodal structure of diabatic heating for PW larger than 50 mm was shown in ERA5, with one maximum around 750 hPa, one around 500 hPa, and another around 300 hPa (Fig. 21a). They correspond to shallow convection, deep convection, and stratiform process, respectively. The positive heating rate for PW less than 45 mm is associated with marine stratus clouds. Compared to ERA5, day-5 forecast using Suite 1 shows similar trimodal structure but with larger magnitudes. Figures 21c and 21d show that all the heating modes are overestimated in Suites 1 and 2 when PW is greater than 55 mm, which is consistent with the early triggered and hyperactive convection found in these

two suites (Fig. 19a). On the contrary, Suite 3 underestimates all the heating modes. It is possible that less shallow and congestus clouds may lead to less deep convection and stratiform clouds. Suite 4 has a weaker stratiform process than ERA5. Since a weaker stratiform process has a positive heating bias at the lower troposphere, the negative bias shown in Suite 4 is likely due to a lack of shallow convection. All suites underpredict the marine stratus clouds (PW < 50 mm around 900 hPa), although a wet bias was found in the moisture field (Fig. 20).

3) Over the land (Fig. 22), the pattern of dry bias below and moist bias above is similar to the one over the ocean (Fig. 20) but has smaller amplitude. The bias is the mildest in Suite 4 compared to the other suites. The dry bias near the surface is associated with the land-atmosphere interaction and the land-surface models. The heating structure over the land in ERA5 is more top-heavy compared to that over the ocean (Fig. 23). Bimodal heating structure is dominant for PW less than 55 mm. At lower troposphere (below 800 hPa), sensible heating is prevailing for PW less than 35 mm. Large discrepancies exist in the forecasts compared to ERA5. All suites overproduce heating at lower levels when PW is greater than 50 mm, which may be due to less shallow convection or a lack of stratiform process. All suites except Suite 4 have less sensible heating and more radiative cooling above boundary layer for PW less than 35 mm, which may result from underpredicted low clouds and imply an issue with the land-surface models and the radiative schemes.

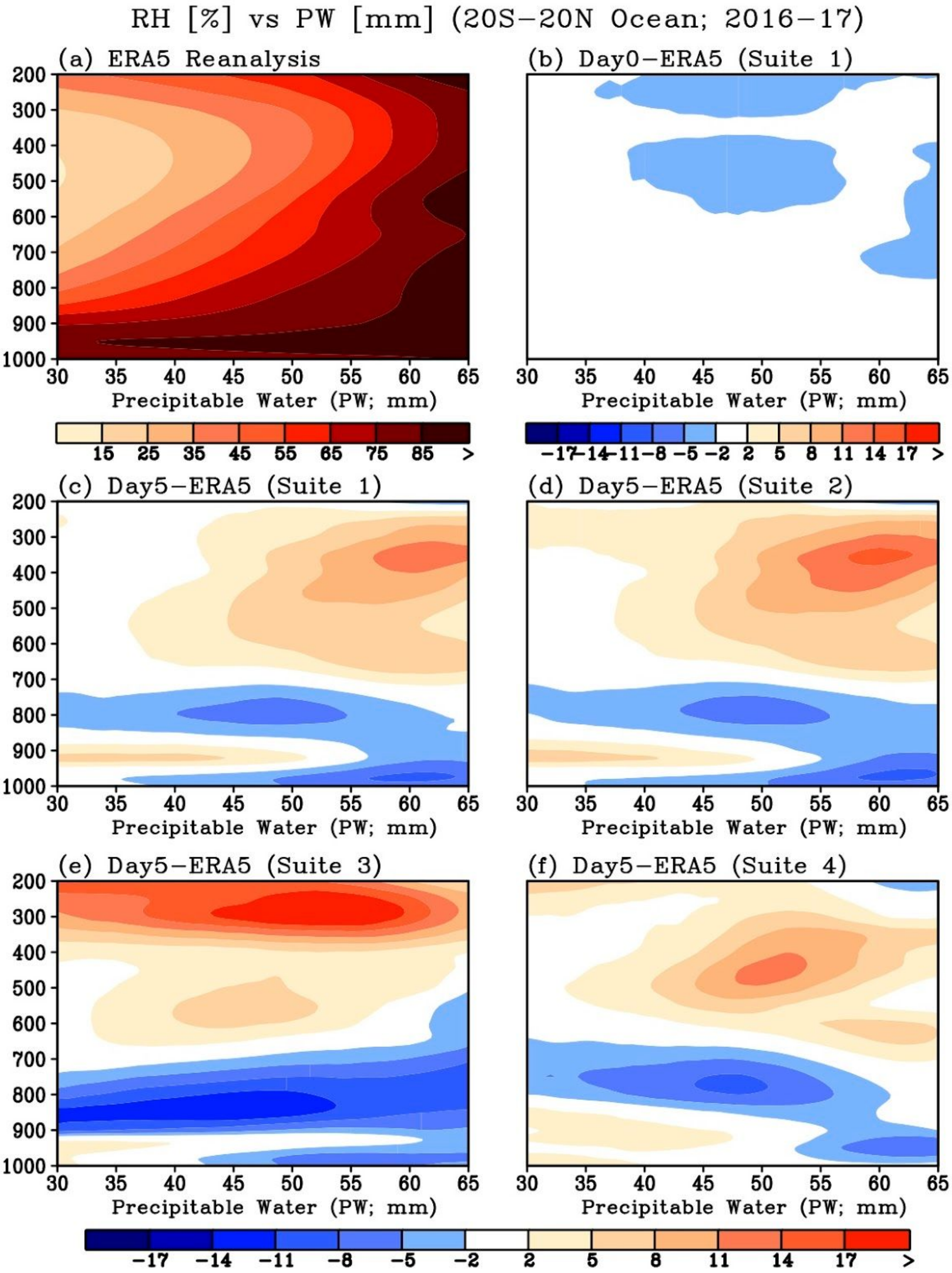


Figure 20. Vertical profiles of RH (%) stratified with 1-mm-wide bins of the PW (mm) over the tropical ocean basins for a) ERA5 Reanalysis, b) Day0-ERA5 for Suite 1, c) Day5-ERA5 for Suite 1, d) Day5-ERA5 for Suite 2, e) Day5-ERA5 for Suite 3, and f) Day5-ERA5 for Suite4.

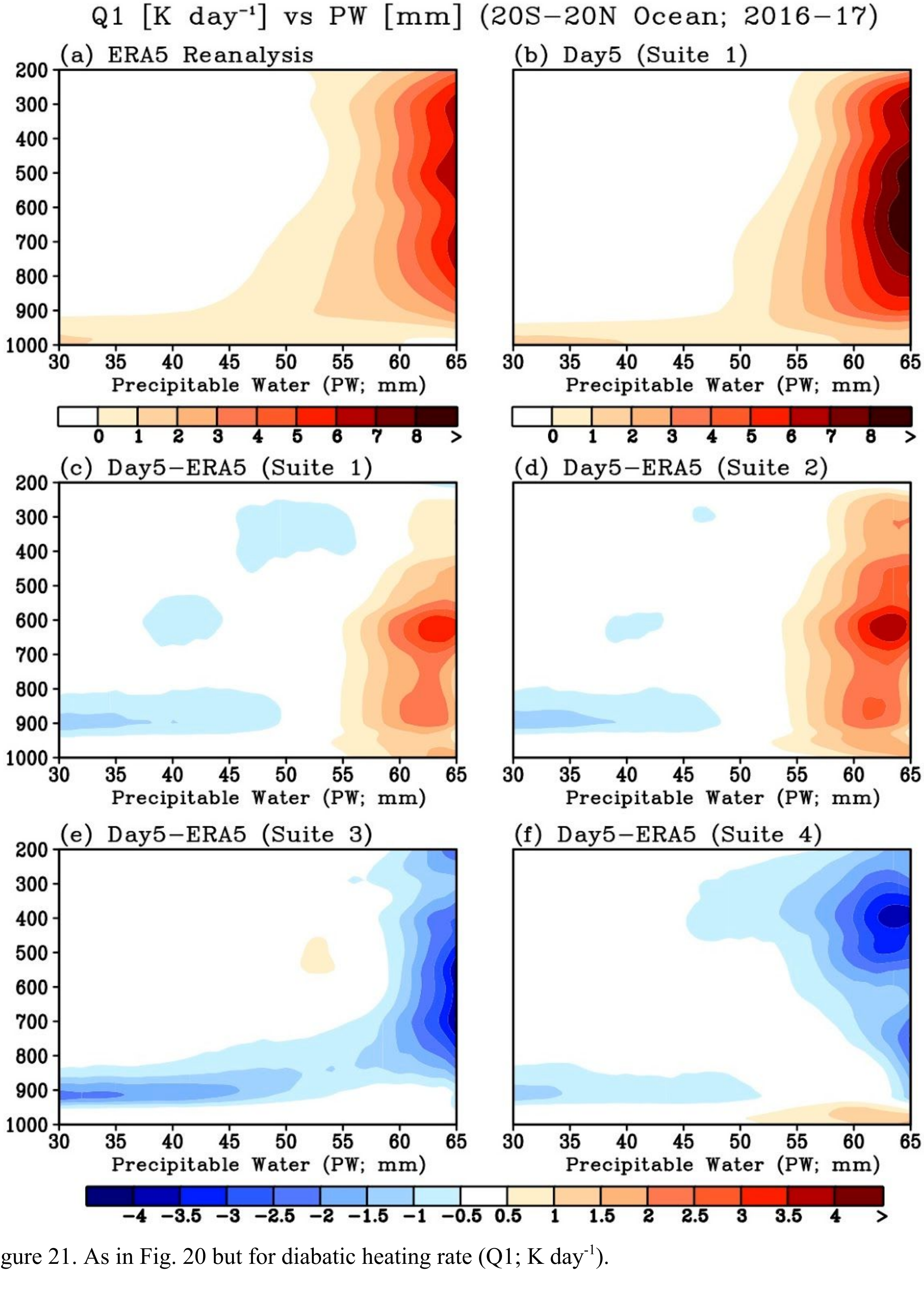


Figure 21. As in Fig. 20 but for diabatic heating rate (Q1; K day^{-1}).

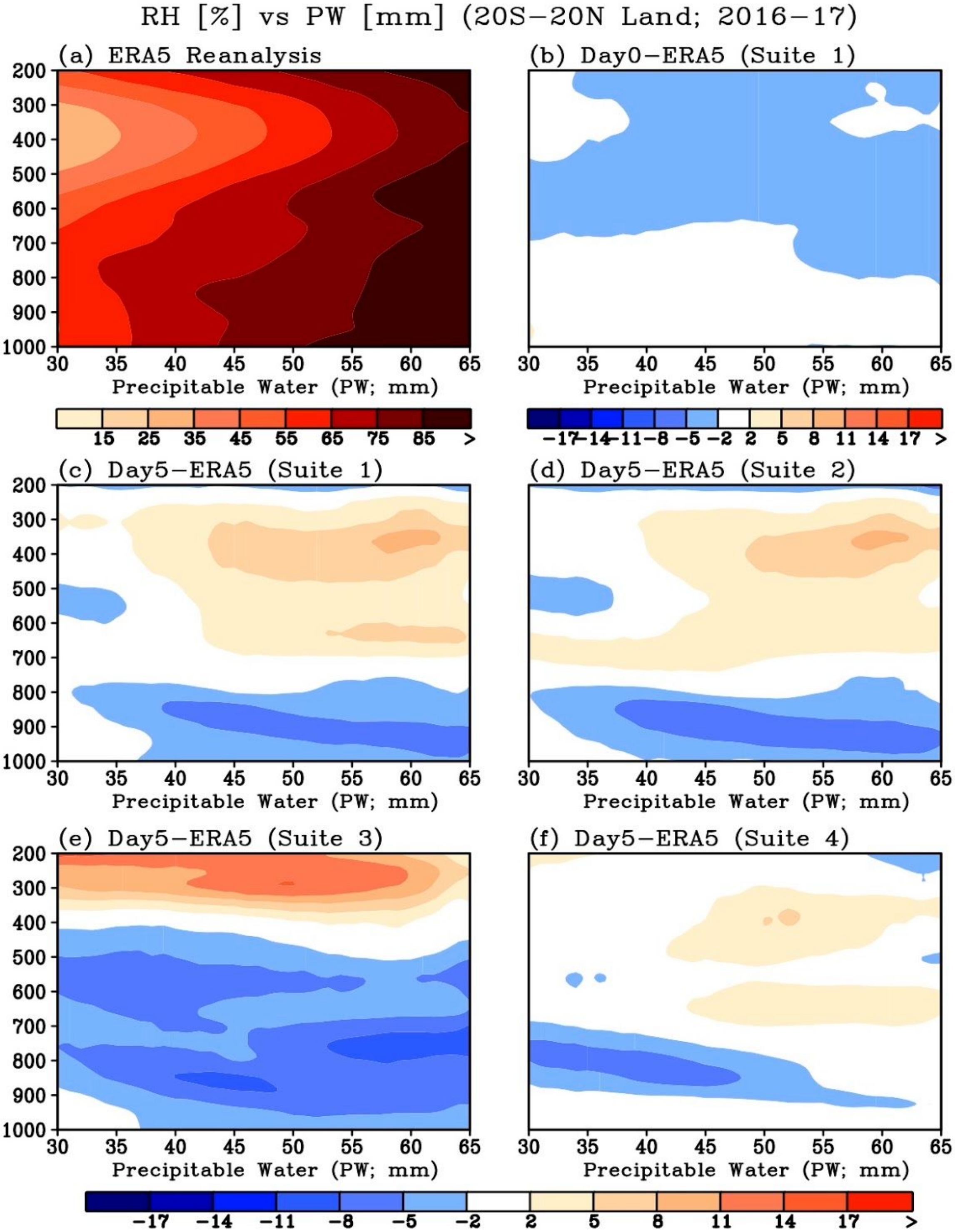


Figure 22. As in Fig. 20 but for the tropical land areas.

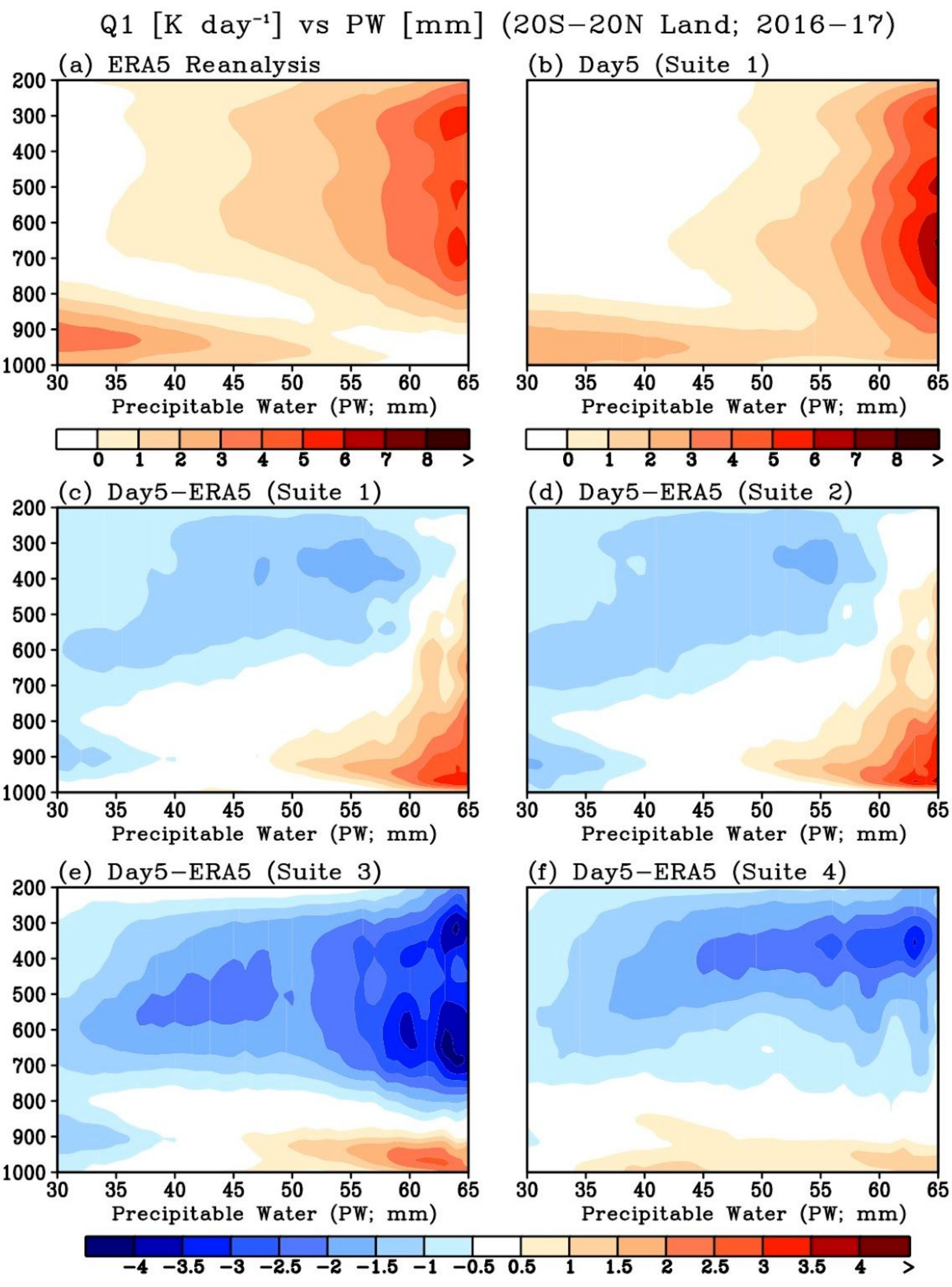


Figure 23. As in Fig. 21 but for the tropical land areas.

Skew-T log-P diagrams

- **Observation data set:** IGRA version 2. Data from 81 stations over CONUS (Figure 24), at 00 and 12 UTC, were used for this analysis.
- **Method:** The skew-T log-P diagrams were created to plot the soundings to compare the thermodynamic properties of the suites against the observations. The vertical distribution of wind speed and wind direction were plotted with wind barbs, and atmospheric characteristics, such as saturation, atmospheric instability, wind shear, and CAPE, were also analyzed. Diagrams were created for forecast hours 0 - 240 h, in 12 hour intervals.
- **Results:** The CAPE, Showalter Index (stability), and PLCL were also analyzed in Figs. 26-29. Figure 26 gives the station average of CAPE over CONUS at 12 UTC and 00 UTC for different forecast lead times. All the suites have smaller CAPE than the observations. Suites 1 and 4 are slightly closer to the observations than Suites 2 and 3. All four suites have lifting condensation levels lower in pressure than the observations at 12 UTC (Fig. 27a) but higher in pressure than the observations at 00 UTC (Fig. 27b). The lifting condensation level of Suite 4 is slightly closer to the observations than that of Suites 1, 2 and 3 at 12 UTC, but the difference is small at 00 UTC. The Showalter stability index for all suites is similar to the observed value (Fig. 28), and they are all stable on average at both 12 UTC and 00 UTC. CAPE for summer at 12 and 00 UTC are shown in Fig. 29.

  For reference, the skew-T log-P diagram figures are available on Jet, Theia and Cheyenne. Below is the directory structure on the different platforms for the 163 cases (e.g., 2015100100):

  Jet: /lfs3/projects/hfv3gfs/lpan/sounding/$case/
  Theia: /scratch4/BMC/gmtb/Linlin.Pan/sounding/$case
  Cheyenne: /glade/scratch/lpan/sounding/$case

  The files can also be found at:
  https://dtcenter.org/eval/gmtb/phytest2019/2019_advphystest/SKEWT/skewt.tar.gz

  The file name was constructed with station number, case and forecast hour, for example: Skewt_723180_2015100100_240.png, where the station number is 723180, the case is 2015100100, and for the forecast hour is 240.

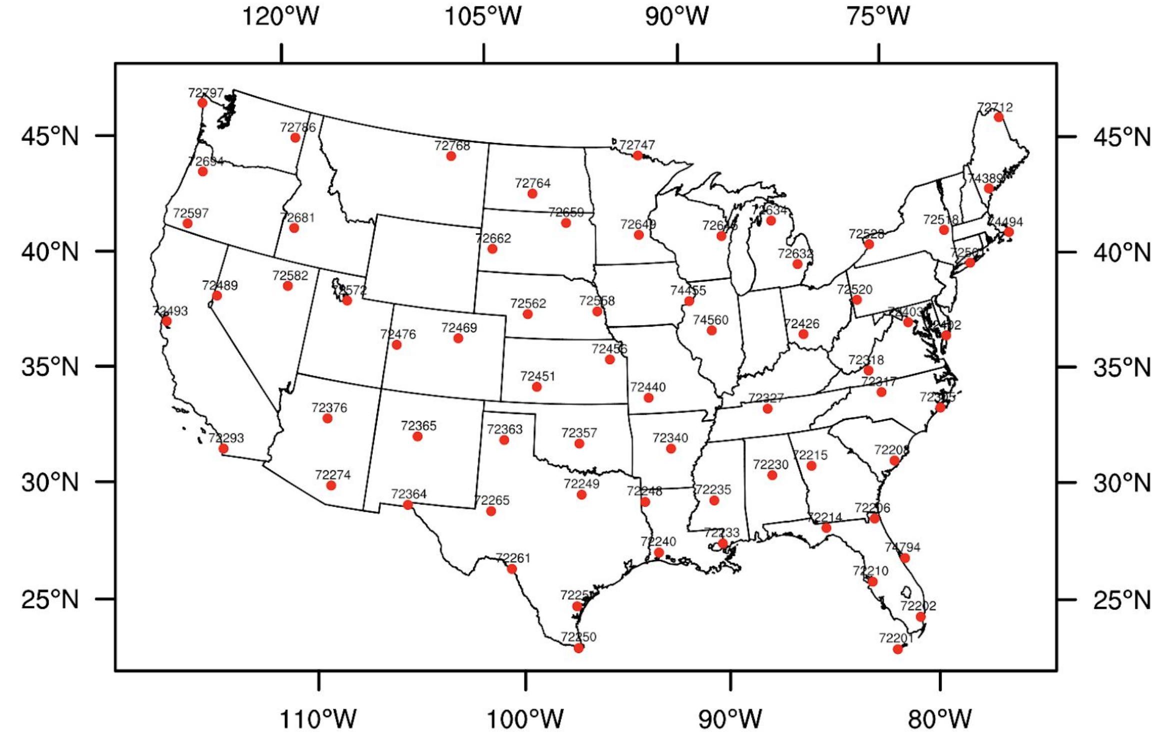


Figure 24. Station distribution with sounding station location denoted by the red dot with the station number above,

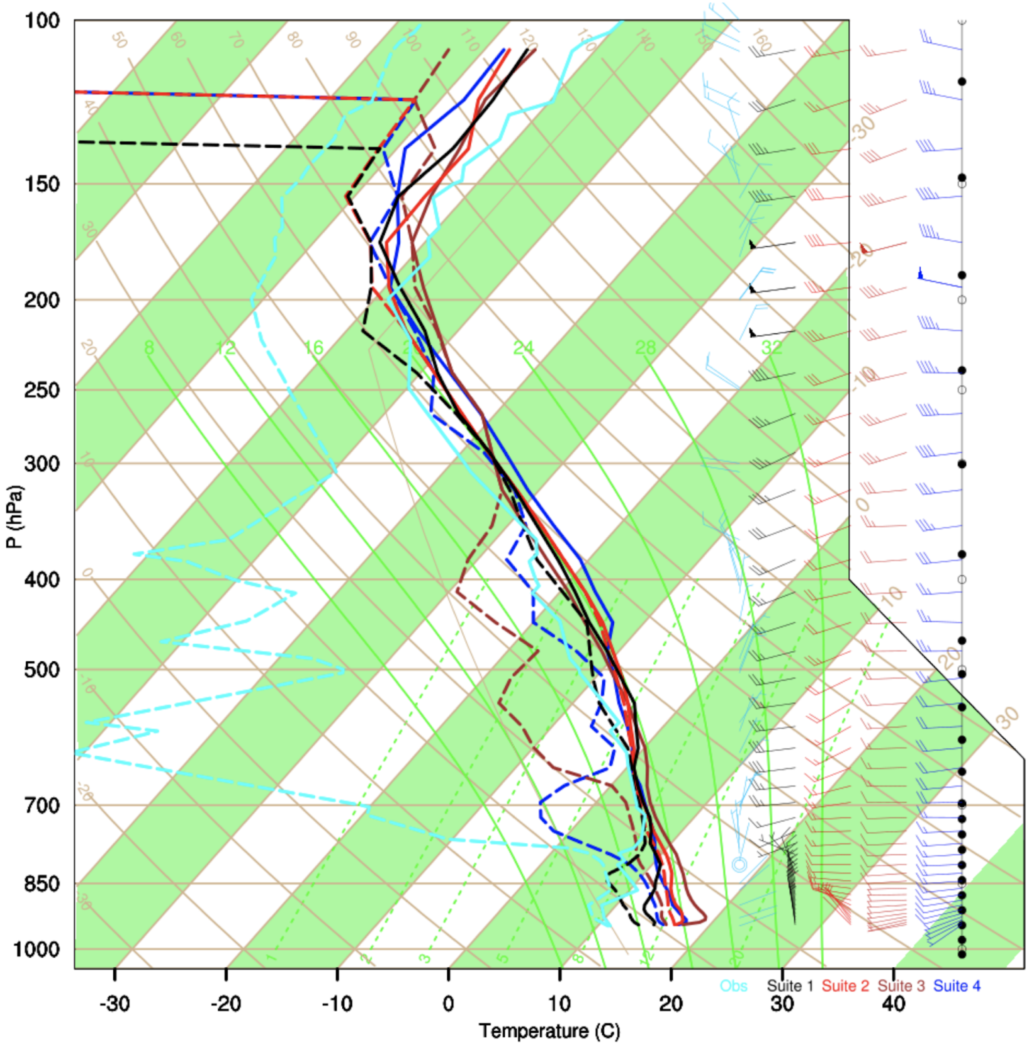


Figure 25. Skew-T log-P diagram of station 723180 (Blackburg, VA) for case 2015100100 at the 240 hour forecast. The temperature and the dewpoint temperature are displayed using solid and dashed lines, respectively. The vertical representation of the wind speed and direction is given with wind barbs at the right side of the figure. Suites 1, 2, 3, and 4 are in black, red, brown, and blue, respectively, while the observation is in cyan.

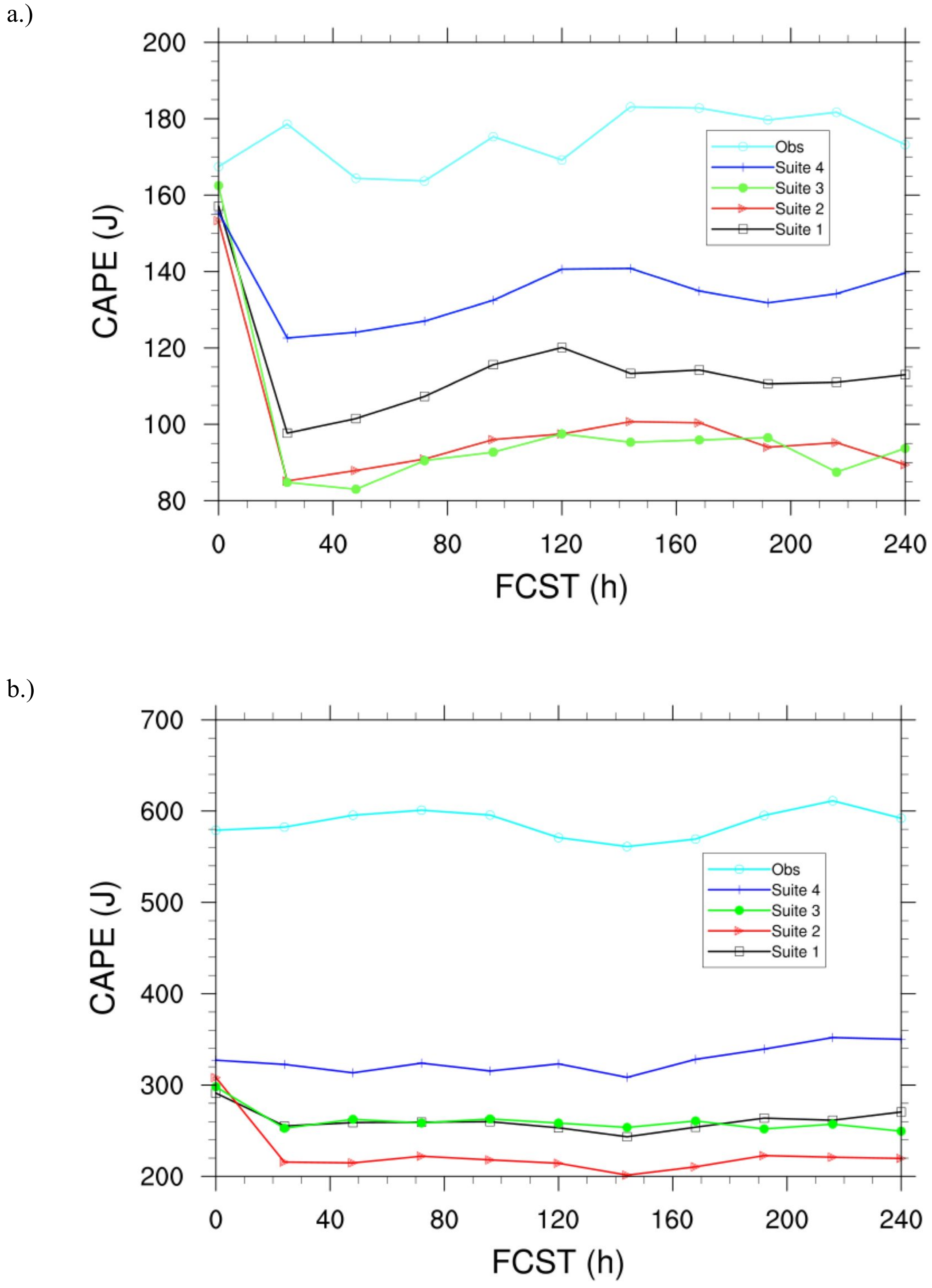


Figure 26. CAPE (J kg^{-1}) averaged for all CONUS stations by forecast lead time (h) for a) 12 UTC initializations and b) 00 UTC initializations. Suites 1, 2, 3, and 4 are in black, red, green, and blue respectively, while the observation is in cyan.

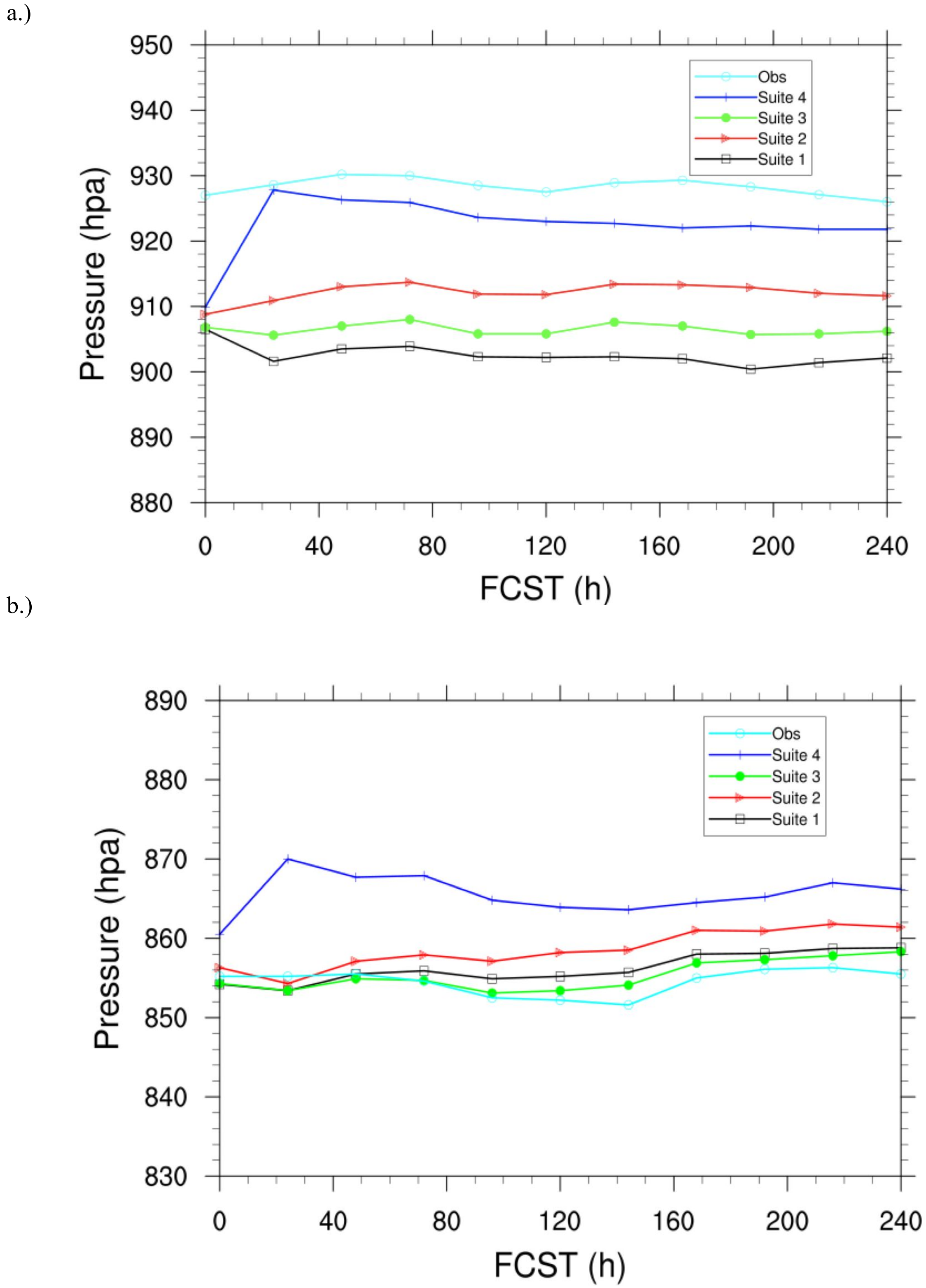


Figure 27. Same as Fig. 26, but for lifting condensation level (hPa).

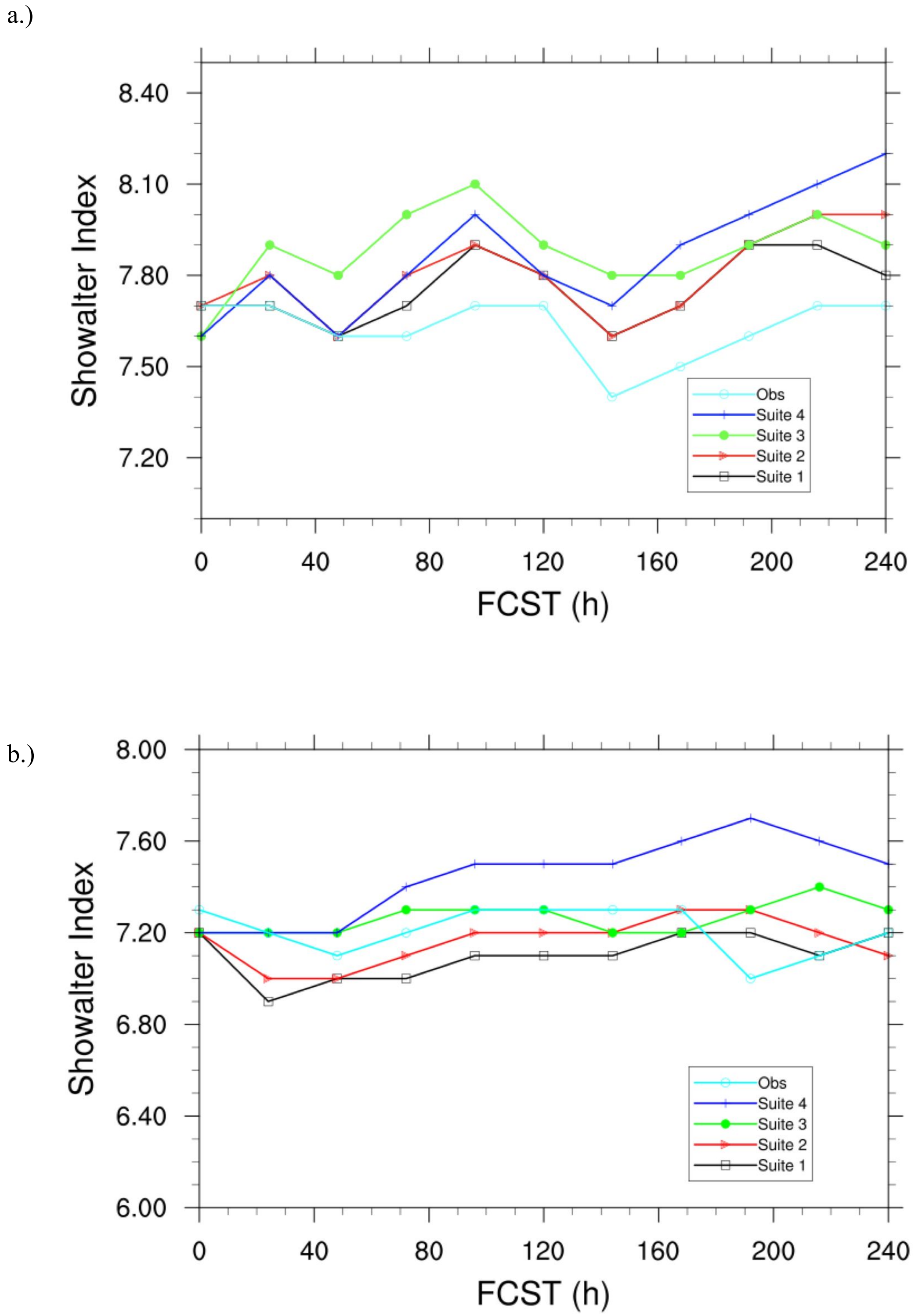


Figure 28. Same as Fig. 26, but for Showalter Index.

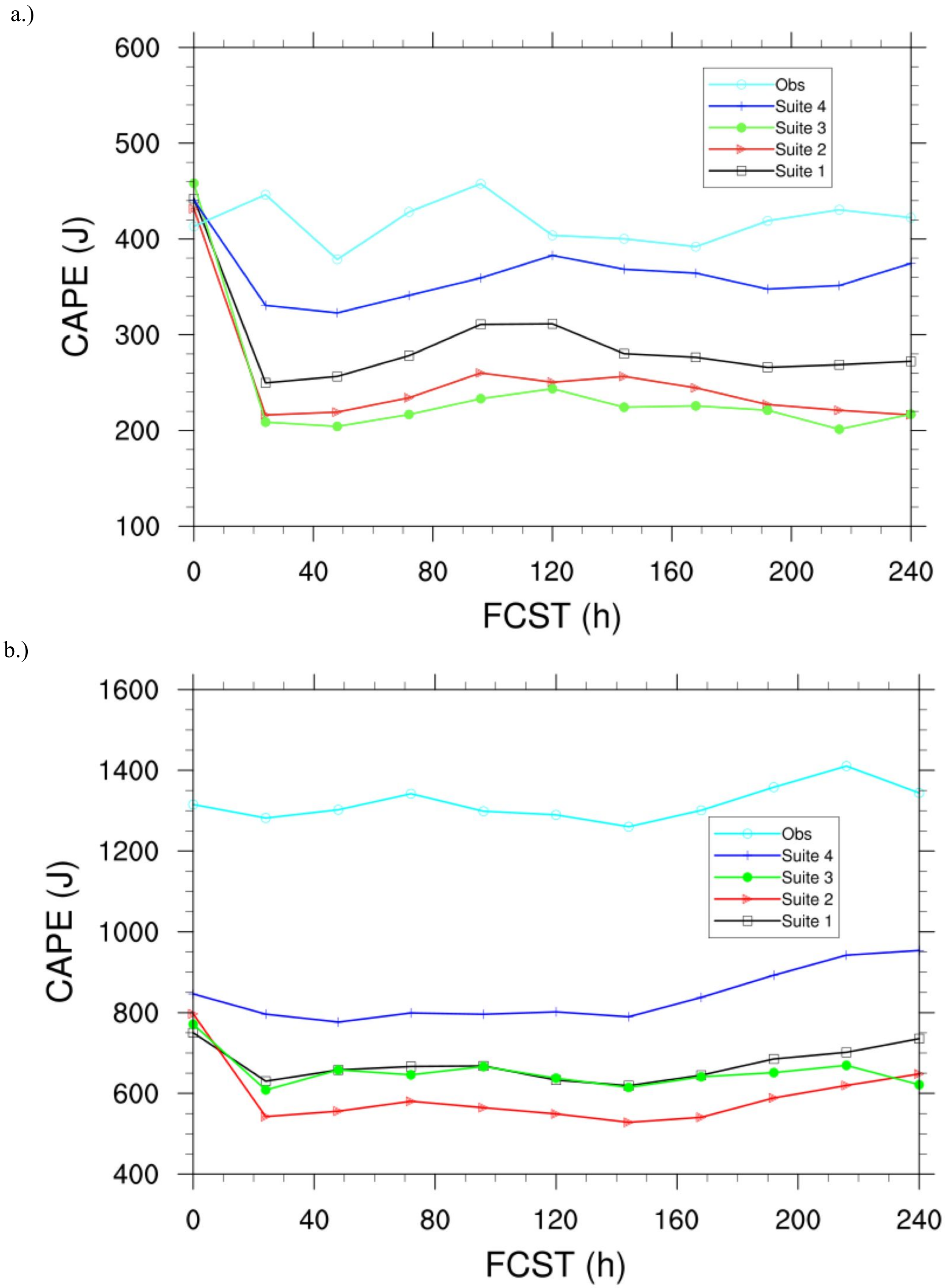


Figure 29. Same as the Figure 26, but for forecasts initialized during Summer (June, July, August).

Vertical profiles of wind and thermodynamic variables

- **Benchmark and method:** 65 sites were matched between the "sounding" profiles generated from the forecasts (in PrepBUFR format) and the observational soundings from IGRAv2 over the CONUS (excluding Alaska) during January-December in 2016-17. IGRAv2 is NCDC's baseline upper-air dataset. Both observational and forecast profiles were interpolated in height with 50 m interval to make them quantitatively comparable. The following analysis was only based on one 12-h forecast validated on 21 January 2016 at 00 UTC over the CONUS.
- **Results:** After converting to local time, 12 UTC is associated with early morning over the CONUS. In other words, the boundary layer can likely be categorized as stable boundary layer (SBL with colder surface and warmer air above and very weak turbulence mixing). Figure 30 shows the average bias over all the 65 stations. All four suites exhibit cold and dry biases in the 12-h forecast from surface to about 600 m. Suite 4 is associated with the smallest biases below 400 m for both temperature and moisture fields. The bias of wind speed suggests an overmixing issue within the PBLs for all four suites, and Suite 4 was the worst in terms of overmixing. The smaller biases of temperature and moisture in Suite 4 may be attributable to the overmixing.

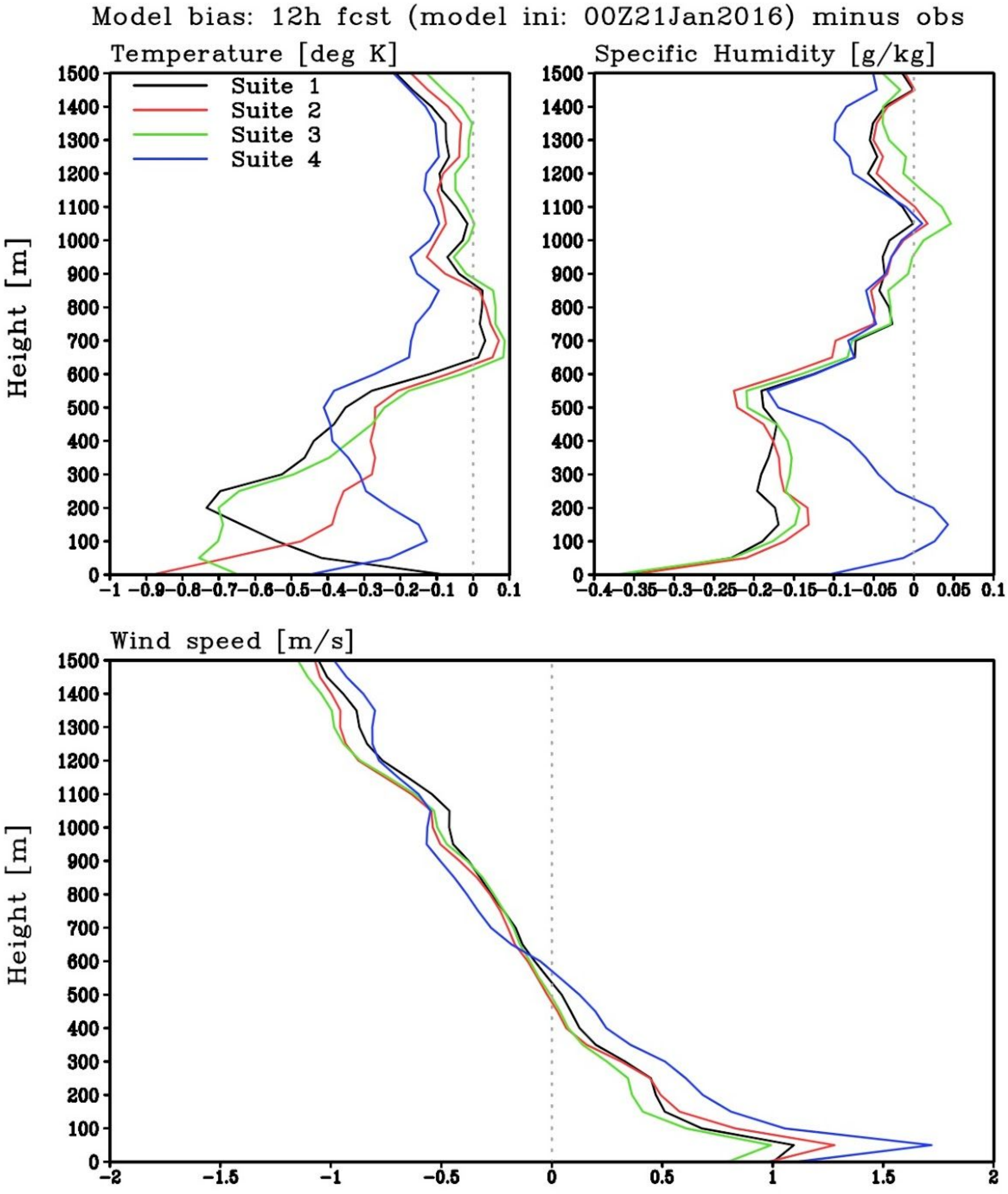


Figure 30. Vertical profiles of temperature (top left), specific humidity (top right), and horizontal wind speed biases (lower) in all four suites compared to the observations.

Energy Budget

The model data included both instantaneous and 6-hourly averages of the radiation flux variables. For these studies, only the 6-hourly averages were used. For the surface-based diagnostics, outgoing and incoming shortwave and longwave radiation fluxes at the surface were compared to SURFRAD data for individual cases and aggregated over the different seasons for statistical comparisons. For the satellite-based diagnostics, radiation fluxes at the top of the atmosphere were also evaluated.

Surface-based diagnostics (SURFRAD)

- **Observation data set:** The SURFRAD data for 7 stations (Fig. 31) was used for point observation comparisons for both case studies and aggregated statistics using MET.
- **Methods:** Though the SURFRAD data has high temporal resolution, it was necessary to process it for comparison against the 6-hourly averages from the model output. The median and 10th and 90th percentiles of the observations were calculated over the same 6-hour time window as the forecasts using the MET *ascii2nc* tool. Using MET output, time series of SURFRAD observations and the four suites at each site for a number of cases were plotted (Fig. 32). Supplemental time series plots of cloud cover percent were also created for select cases (Fig. 33).

  In addition to the case-by-case approach, results were also broken down by 00 and 12 UTC initializations and aggregated across the seasons for each of the four suites. This was accomplished by employing the MET *point-stat* tool to calculate verification statistics at the seven observation points against the model forecast output.
- **Results:** For the case-by-case approach, only one case at a single site is shown to provide an example of the type of plots that were created (Fig. 32). A comprehensive set of plots for all sites, cases, and both initializations are available at: https://dtcenter.org/eval/gmtb/2019_advphystest/SURFRAD/SURFRAD_CASE.tar.gz. A number of case studies during the winter months exhibited differences between Suite 4 when compared to the other three suites at a number of SURFRAD sites. Figure 32 provides an example of this finding by displaying this difference as observed at the Bondville, IL site for the 2016010100 UTC case. When looking at downward shortwave radiation, Suite 4 displayed lower values on most days compared to the other models and observations (Fig. 32a). This behavior is reflected in upward shortwave radiation (Fig. 32c), where all suites tend to be lower than observations for most days, with Suite 4 having the lowest values. When evaluating longwave radiation, Suite 4 has higher values compared to the other suites at all forecast lead times. Compared to observations, Suite 4 often over forecasts downward longwave radiation while the others often under forecast (Fig. 32b). For upward longwave radiation, while Suite 4 exhibits a diurnal signal, the magnitude is smaller than observations for most days compared to the other configurations (Fig. 32d). Supplemental plots of total cloud cover were also created to investigate whether differences in cloud cover amounts helped explain the large differences between Suite 4 compared to the other suites. Figure 33 shows total cloud cover for the 2016010100 UTC case at

Bondville, IL, where Suite 4 exhibits higher cloud cover percent compared to the other suites for the first 7 days of the forecast.

Plots of seasonally aggregated statistics were also evaluated for all the 00 and 12 UTC initializations. While only downward shortwave radiation is shown here (Fig. 34), all plots for incoming and outgoing longwave and shortwave radiation for all aggregations, sites, seasons, and both initializations are available at https://dtcenter.org/eval/gmtb/2019_advphystest/SURFRAD/SURFRAD_MET.tar.gz. Similarly to the behavior exhibited in the 2016010100 UTC case shown above, Suite 4 also shows lower downward shortwave radiation values compared to the other suites when the results are aggregated over the entire winter (DJF). While this behavior was observed at a number of sites, it was particularly pronounced at the Bondville, IL site (Fig. 34a). When looking at mean error over winter (Fig. 34b), Suite 4 typically underestimates downward shortwave radiation, while the other suites tend to overestimate it. During the spring season, Suite 4 also displays lower downward shortwave radiation than other suites, but to a lesser extent than in winter. Suite 3 generally had a high bias at the 3 easternmost sites (Goodwin Creek, MS, Bondville, IL, and Penn State, PA). Figure 35 illustrates this behavior at the Goodwin Creek site, where Suite 3 displays higher forecast mean for downward shortwave radiation at nearly all forecast lead times (Fig. 35a) and higher positive bias for most days (Fig. 35b). For the summer season, the four configurations show more agreement in surface shortwave fluxes. For downward shortwave radiation at the Table Mountain Boulder, CO site, all suites typically have higher values compared to observations, with Suite 1 and Suite 2 forecast means lower and closer to observations (Fig. 36a). This behavior is reflected in bias (Fig. 36b), with a positive bias observed at most forecast lead times for all suites, with Suites 1 and 2 often having a lower bias. Similar to the spring aggregation, for the fall aggregation Suite 3 displays higher forecast means in downward shortwave radiation at a number of sites such, e.g., Penn State, PA (Fig. 37a). While other suites transition from positive to negative bias, depending on the day, Suite 3 has a positive bias for all days (Fig. 37b).

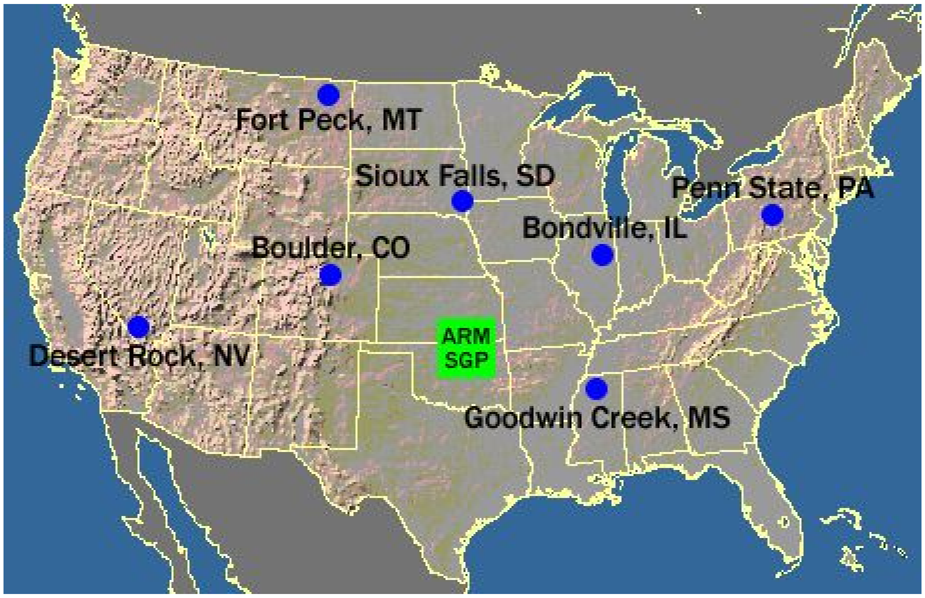


Figure 31. The current network of SURFRAD stations. The blue dot indicates the established stations used in the verification.

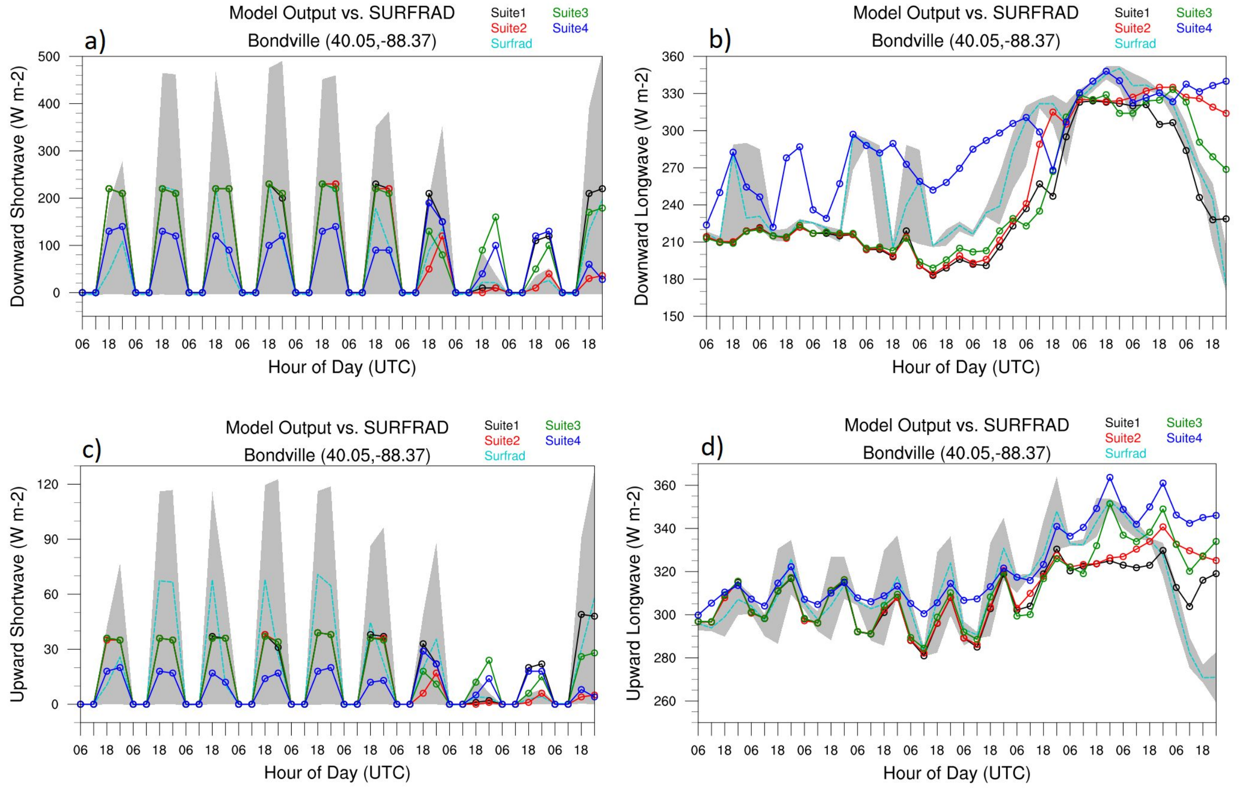


Figure 32. Case study for January 1, 2016 at the Bondville, IL site for surface 6 hour averages of a) downward shortwave, b) downward longwave, c) upward shortwave and d) upward longwave. Suite 1 is black, Suite 2 is red, Suite 3 is green, Suite 4 is blue and median SURFRAD is cyan. The grey shaded area represents the 6 hour 10th-90th percentile (p10-p90) range from the SURFRAD observations.

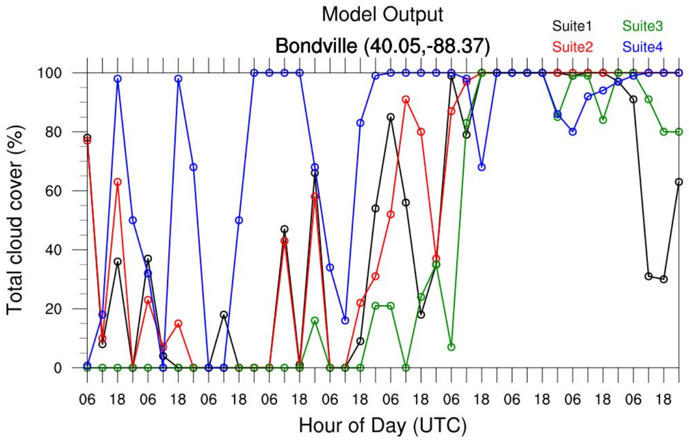


Figure 33. Case study for January 1, 2016 at the Bondville, IL site for 6 hour average of total cloud cover percent. Suite 1 is black, Suite 2 is red, Suite 3 is green, Suite 4 is blue.

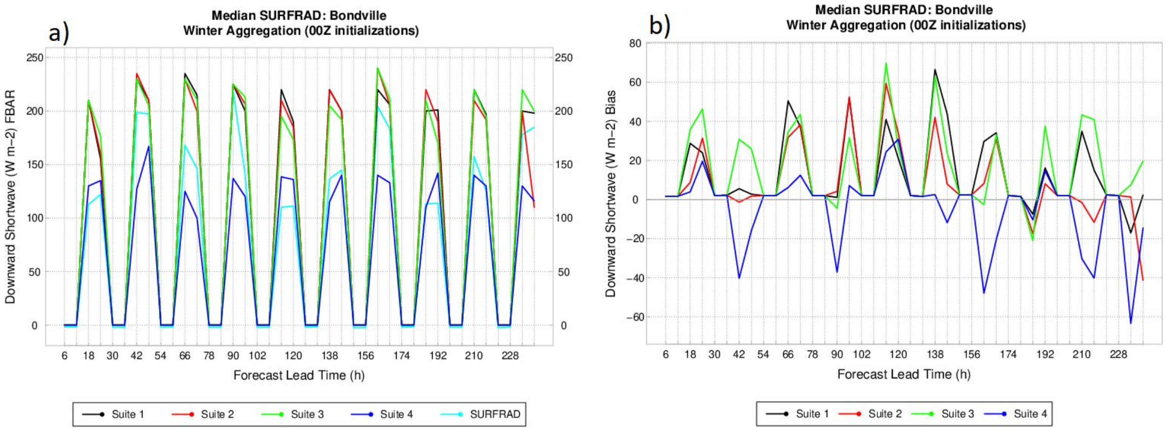


Figure 34. Downward shortwave radiation flux at the Bondville, IL site aggregated across all 00 UTC initializations for the winter season for a) forecast mean b) mean error. Suite 1 is black, Suite 2 is red, Suite 3 is green, Suite 4 is blue and SURFRAD is cyan.

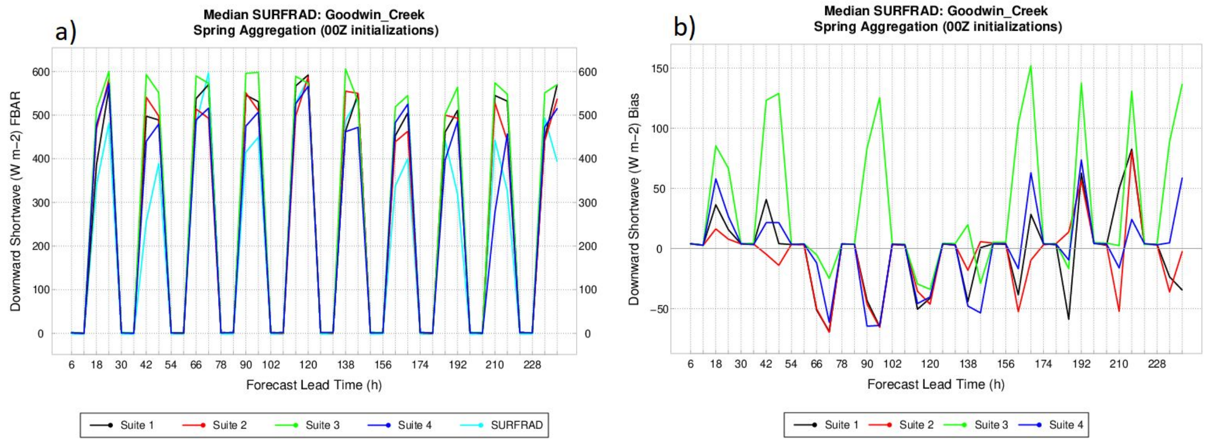


Figure 35. Same as 34, but for the Goodwin Creek, MS site for the spring aggregation.

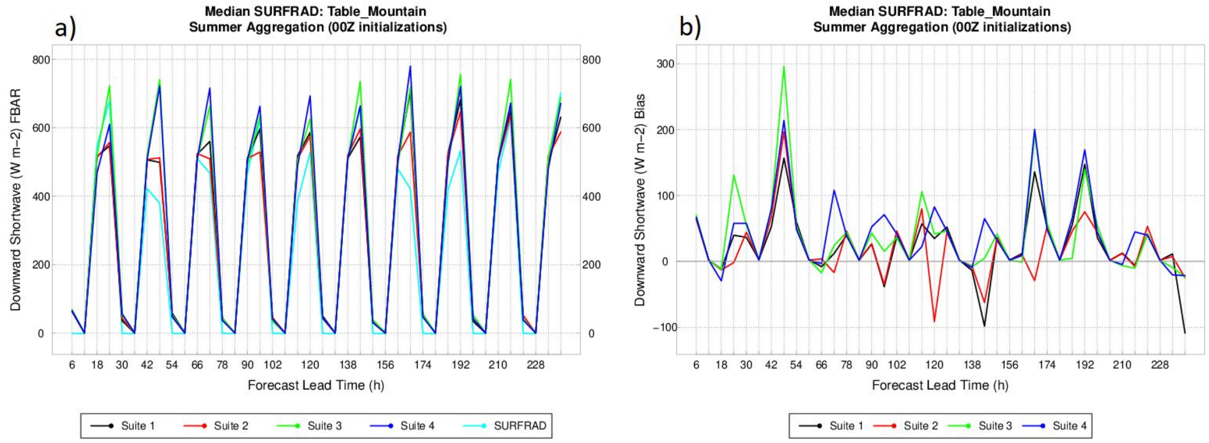


Figure 36. Same as 34 but for the Table Mountain, CO site during the summer aggregation.

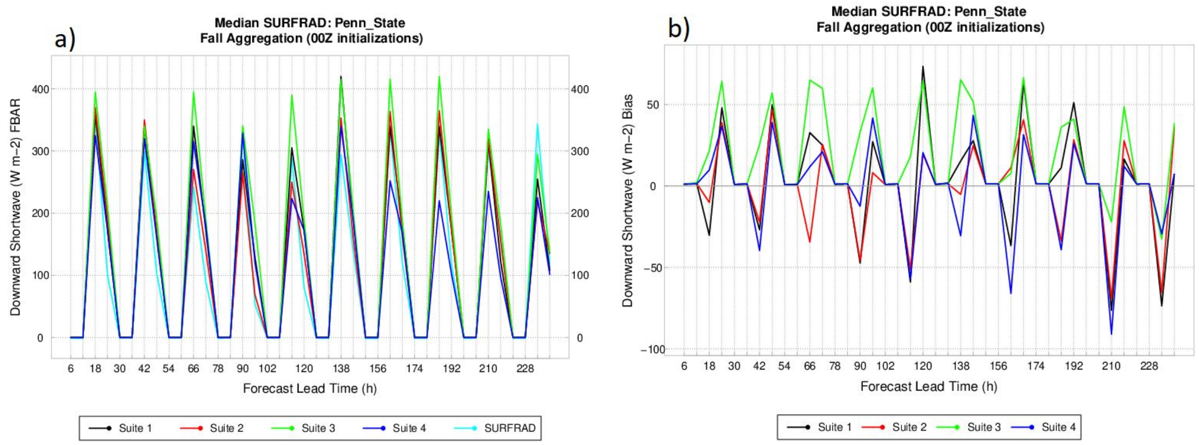


Figure 37. Same as 34 but for the Penn State, PA site during the fall aggregation.

Satellite-based diagnostics (CERES)

- **Observation data set:** The Clouds and the Earth Radiant Energy System (CERES) satellite products of both longwave and shortwave radiation at the surface and top of the atmosphere were used for comparison against the four model configurations. The CERES monthly mean data was obtained from here. These means are calculated by spatially averaging the instantaneous fluxes on a 1-degree grid, temporally interpolating between observed values at 1-hour increments, then averaging all hours in a month.

- **Method:** Seasonal means were computed from CERES monthly averages using MET for spring (March/April/May), summer (June/July/August), fall (September/October/November), and winter (December/January/February) for the combined 2016-2017 period. Similarly, seasonal means were also computed using MET for each of the FV3GFS configurations for day 1, 3, 5, and 10 using forecast valid time to parse each forecast into the correct season. Since the CERES monthly means are derived using all hours of the day, the 1, 3, 5, and 10 day forecast means were also averaged over the full 24-hour period. For example, the day 10 mean included the 6-hourly averaged fluxes from the model of forecast hour 222, 228, 234, and 240.

- **Results:** Spatial panel plots of CERES, FV3GFS and difference fields of the seasonal means were created for each suite, season, and 24-hour period using NCL. Sample plots and discussion of shortwave and longwave radiation at the top of the atmosphere can be found in Figures 38 and 39 (suite 1 only). All plots are available at https://dtcenter.org/eval/gmtb/2019_advphystest/CERES, and include TOA upward SW and LW, surface downward and upward SW.

  Seasonal mean differences of shortwave radiation at the top of the atmosphere during the winter are negative over much of the tropics and southern ocean for Suites 1 (Fig. 38c; day 10 only), 2 and 3 for each of the 24 hour periods. Suite 4 differences are negative in the tropics, with much of the differences observed in the southern hemisphere being positive for all 24 hour periods.

  An evaluation of longwave radiation at the top of the atmosphere in winter for Suites 1 and 2 show positive seasonal mean differences with the largest difference observed in the tropics for each of the 24 hour periods (Fig. 39c; suite 1 only). Suites 3 and 4 have a similar positive day 1 difference field to Suites 1 and 2, but fewer differences for day 3, 5, and 10 generally showing a mix of positive and negative differences in the tropics.

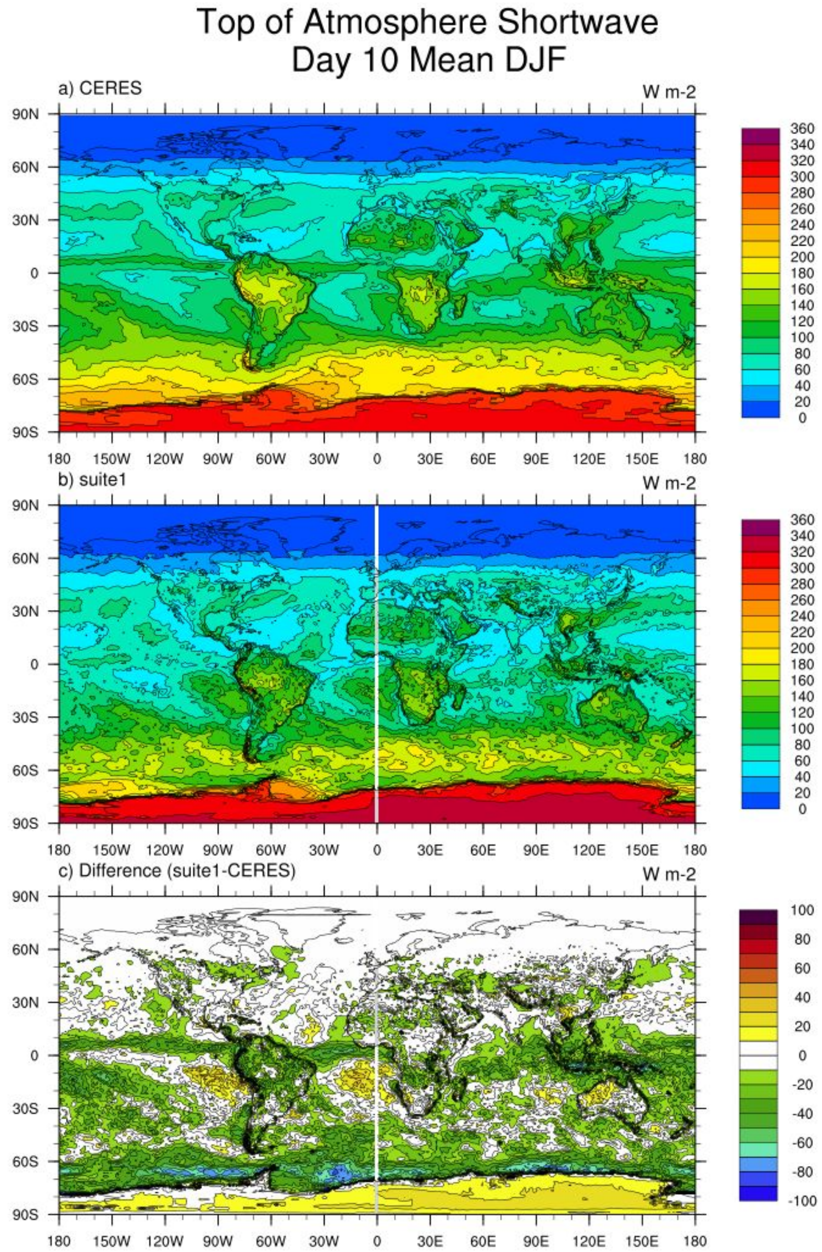


Figure 38.Top of the atmosphere mean shortwave radiation for day 10 forecast for the winter aggregation for a) CERES, b) Suite 1, and c) difference (Suite 1 - CERES).

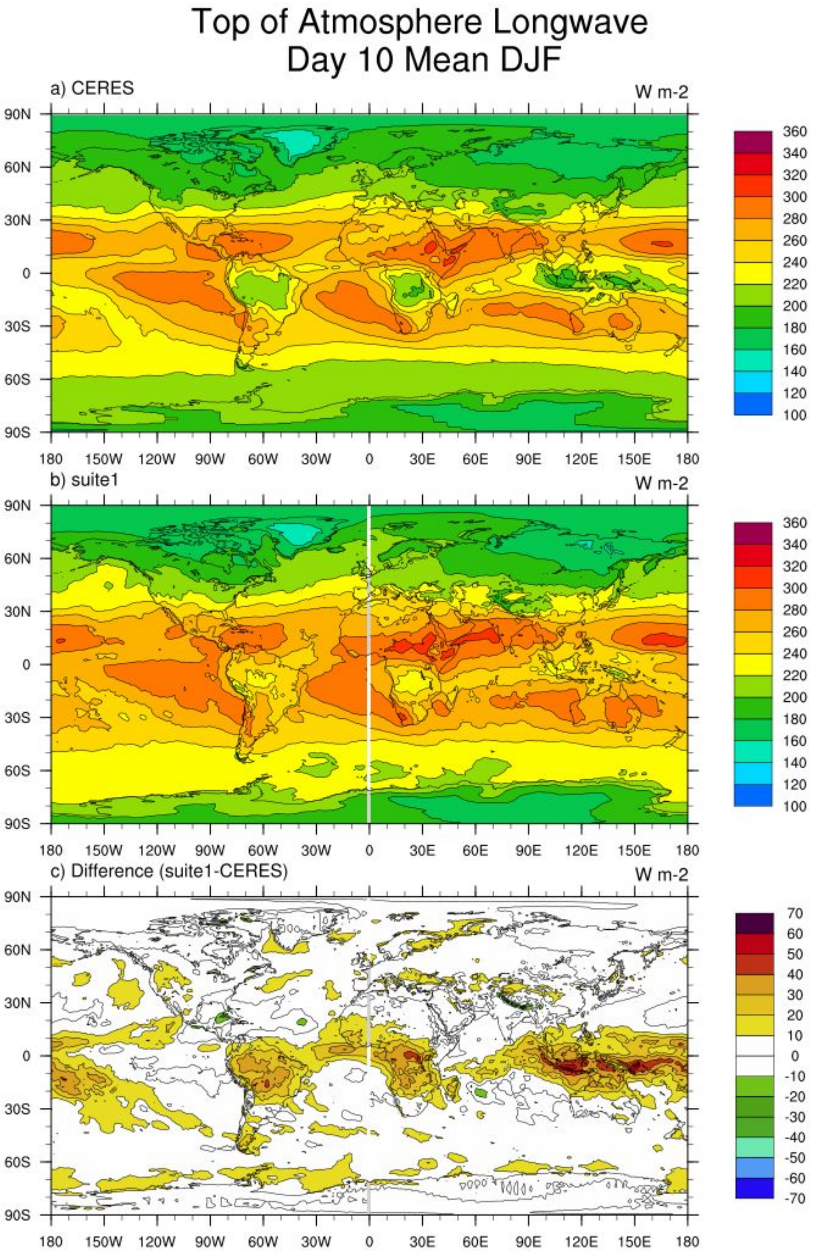


Figure 39. Same as 38, but for top of the atmosphere longwave radiation.

Tropical Cyclone

Tracking and verification of existing cyclones

- **Observation data set:** Forecasts from all four suites were verified against the Best Tracks.

Method: Existing tropical cyclones (TCs) were tracked for all basins using the public release of the GFDL Vortex Tracker v3.9a supported by the DTC (Biswas et al. 2018). The script and fix files used for running the tracker can be found here (refer to script run_tracker.ksh). The resulting 5-day forecasts of track, intensity, and size of storms were verified against the Best Tracks using the MET-Tropical Cyclone (TC; Halley Gotway et al. 2018) verification tools.

- **Results:** Results indicate that the track errors from Suites 1 and 2 are similar and have the lowest errors of all suites (Fig. 40). Suite 3 has the largest errors, especially later in the forecast. These results are representative of the Atlantic and Eastern Pacific basins. The number of cases in the Western Pacific and Central Pacific is not enough to make a clear distinction among the suites.

  All suites have negative intensity error (Fig. 41), which is not uncommon for models run at the resolution used in this test. However, Suites 1 and 2 have the least negative intensity bias and Suite 4 has the largest biases.

  Storm size was verified using the maximum radial extent of the 34-kt winds. Figure 42, a sample for the Atlantic basin, indicates that storms are initialized small, but the model corrects their size two days into the forecast. Suite 3 has a tendency to shrink the storms as forecast lead time increases.

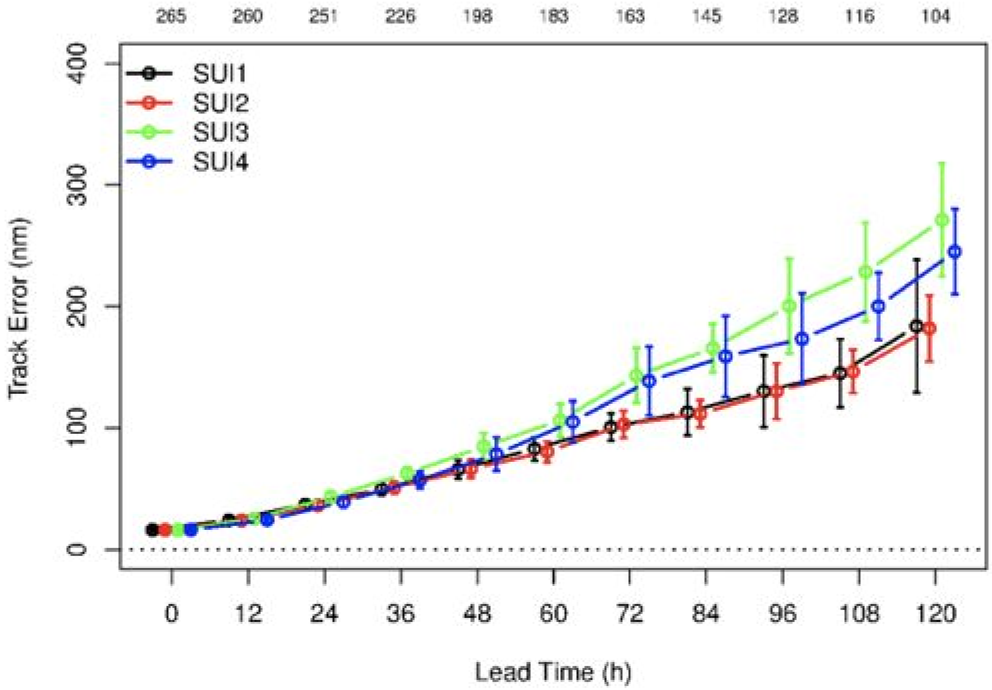


Figure 40. Track error (nm) for all basins combined for each suite averaged for all 163 cases of the test. Suite 1 is in black, Suite 2 is in red, Suite 3 is in green, and Suite 4 is in blue.

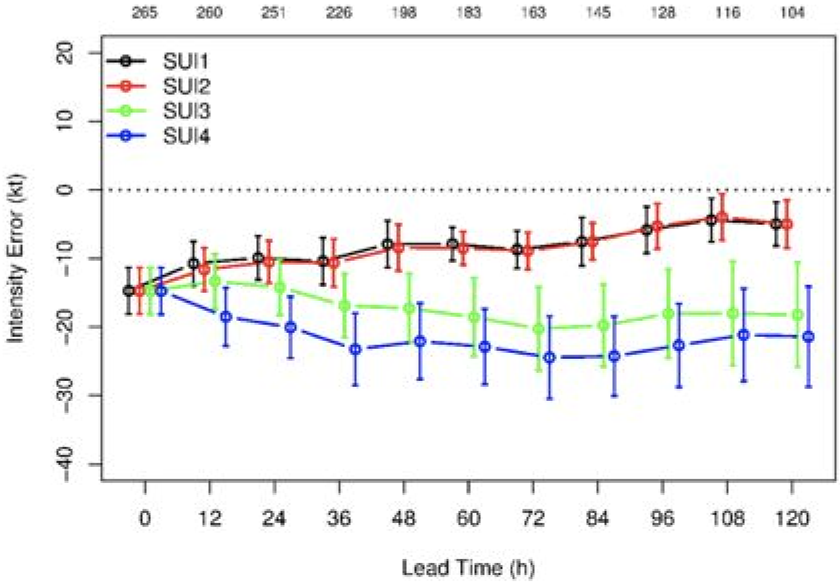


Figure 41. Intensity error (kt) for all basins combined for each suite averaged for all 163 cases of the test. Suite 1 is in black, Suite 2 is in red, Suite 3 is in green, and Suite 4 is in blue.

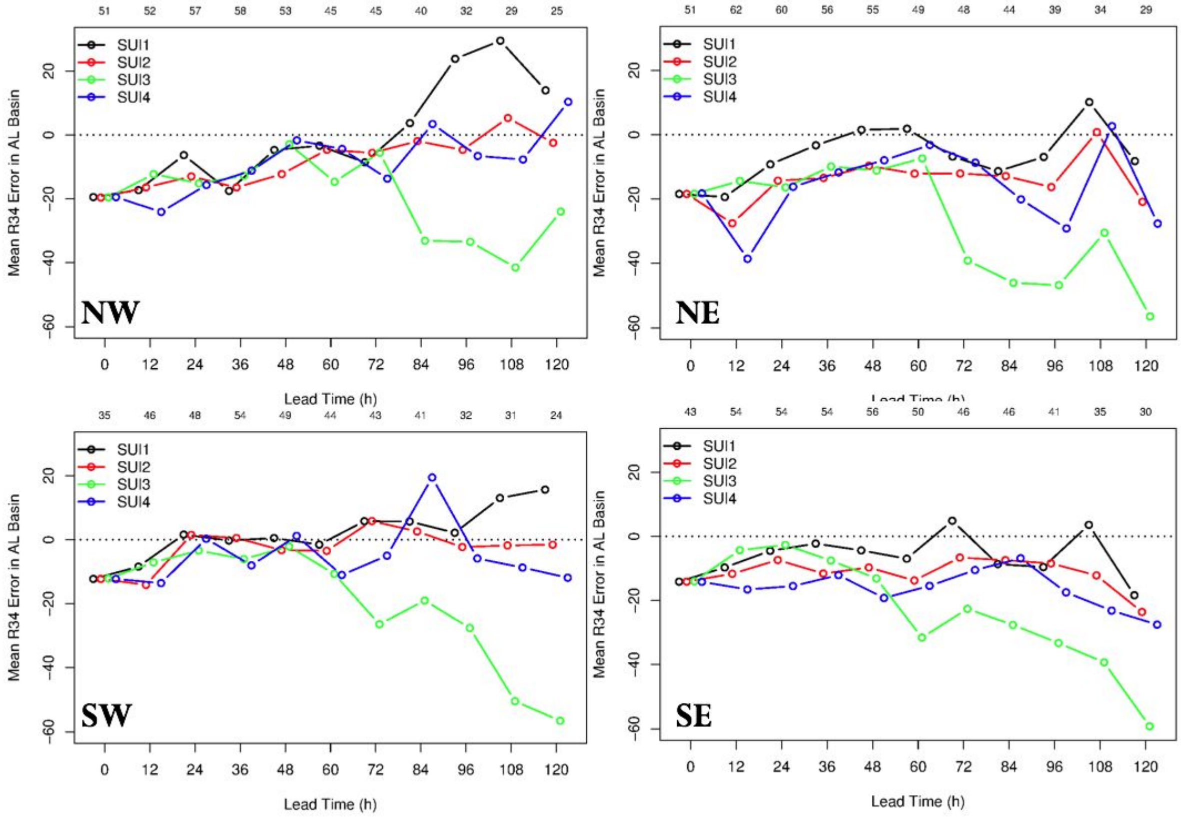


Figure 42. Tropical storm size as shown by the radius of 34 kt winds (nm) for the Atlantic basin in the NW (top left), NE (top right), SW (bottom left), and SE (bottom right) quadrants. Suite 1 is in black, Suite 2 is in red, Suite 3 is in green, and Suite 4 is in blue.

Tropical cyclogenesis identification and verification

The ability to accurately predict TC genesis is an important operational need. The purpose of this study is to develop reliable TC genesis forecasts based on global model output to serve as skillful guidance for forecasters.

- **Observation data set:** Forecasts from all four suites were verified against the Best Tracks.
- **Method:** TC genesis was identified using for all basins using the public release of the GFDL Vortex Tracker v3.9a supported by the DTC. The script and fix files used for running the genesis tracker can be found here (refer to script loop_genesis.ksh).

  The default configuration of the GFDL Vortex Tracker for genesis was applied to the 0.25° post-processed model output for the band of latitudes between 30°S and 30°N. It was noticed that the output files contained a large and unrealistic number of forecast genesis. A postprocessing step was then applied to eliminate spurious cyclogenesis by:

a) Retaining only storms that lasted for at least 24 h with the maximum sustained surface wind speed of 34 kt or more.

b) Removing any model genesis forecasts whose forecast genesis time is at forecast hour 000. We assume that these are existing TCs and not forecasts of TC genesis.

c) Removing any model genesis forecasts whose forecast genesis time is greater than forecast hour 120. In this way, we examine the model performance in the first 5 days as most of the operational hurricane model does.

This filtering process eliminated many spurious storms. In the future, we believe additional criteria should be considered to assess the thermal structure of the cyclone, such as the presence of warm core in the mid-troposphere and the Cyclone Phase Space (CPS) diagnostics (Hart 2003). This additional criteria could eliminate some of the false alarms produced by the models, especially at higher latitudes. In spite of these caveats, we decided to include the genesis verification in this report as it provides a measure of the differences among the suites and a benchmark for future assessments.

The model TC genesis events were then verified with the observed using a script adapted from Halperin (Halperin et. al. 2013). The method is currently done to match as closely as possible the NHC Tropical Weather Outlook (TWO), a product that provides categorical and probabilistic forecasts of TC genesis. In this method, for each model genesis forecast, find all entries in the best track that (i) have a time stamp that matches the valid time of the model genesis forecasts and (ii) have a latitude and longitude within 5° of the model indicated TC latitude and longitude at the valid time of the model genesis forecast.

a. If the initialization time is 0-120 h before the best-track genesis time (defined as first entry of tropical depression (TD) or tropical storm (TS) in the b-decks), then we have a "hit."
b. If the initialization time is >120 h before the best-track genesis time, then we have a false alarm (FA).
c. If the initialization time is after the best-track genesis time, but still within 72 h temporal and 5° spatial tolerance, then we have a late genesis (LG).

Forecast genesis that does not meet the criteria for hit or late genesis is considered as a false alarm.

- **Results:** Figure 43 shows the verification for the Northern Hemisphere. The number of hits for all suites, even considering late genesis, is quite low compared to the number of observed genesis. It is noted that Suite 2 and Suite 4 produce the most hits in 2016, while Suite 1 and Suite 2 produce the most hits in 2017. However, Suite 3 produces the least hits in both years. Conversely, all the suites have a large number of false alarms, with Suite 1 having the largest number. As mentioned above, it is likely that some of the false alarms correspond to extratropical storms.

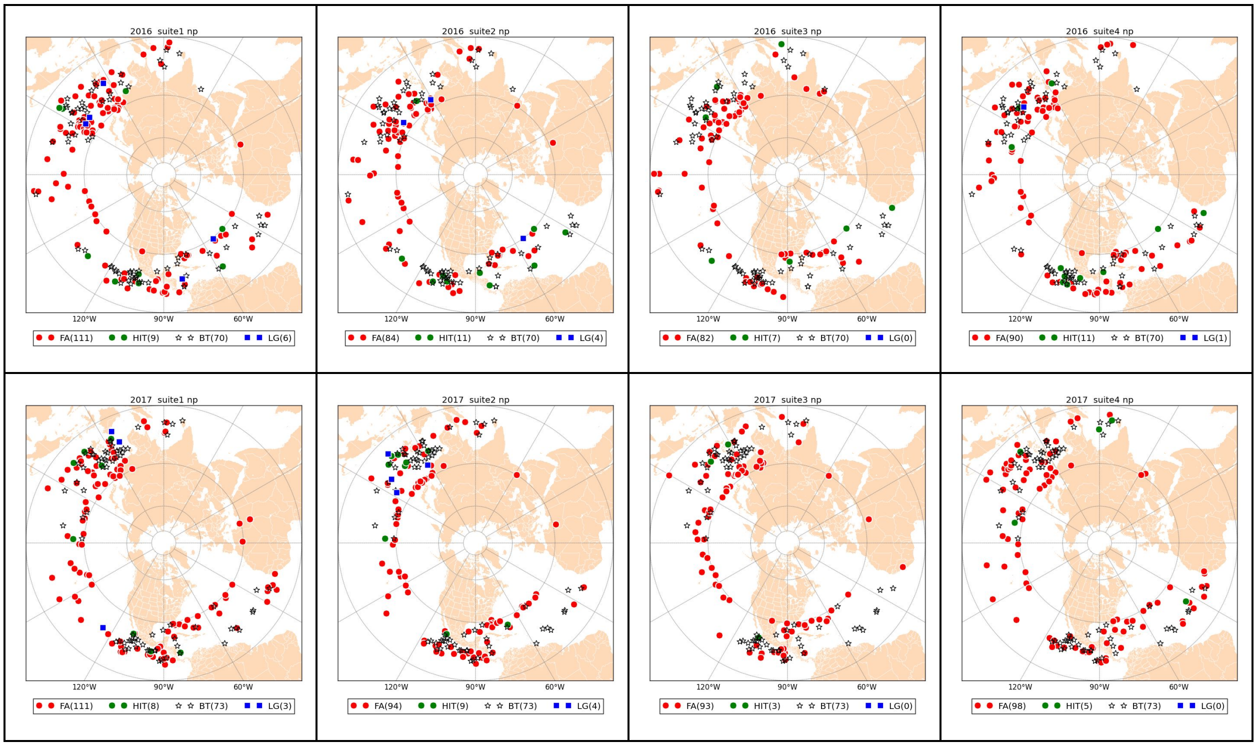

2016 suite1 np
120°W
90°W
60°W
● ● FA(111) ● ● HIT(9) ☆ ☆ BT(70) ■ ■ LG(6)
2016 suite2 np
120°W
90°W
60°W
● ● FA(84) ● ● HIT(11) ☆ ☆ BT(70) ■ ■ LG(4)
2016 suite3 np
120°W
90°W
60°W
● ● FA(82) ● ● HIT(7) ☆ ☆ BT(70) ■ ■ LG(0)
2016 suite4 np
120°W
90°W
60°W
● ● FA(90) ● ● HIT(11) ☆ ☆ BT(70) ■ ■ LG(1)
2017 suite1 np
120°W
90°W
60°W
● ● FA(111) ● ● HIT(8) ☆ ☆ BT(73) ■ ■ LG(3)
2017 suite2 np
120°W
90°W
60°W
● ● FA(94) ● ● HIT(9) ☆ ☆ BT(73) ■ ■ LG(4)
2017 suite3 np
120°W
90°W
60°W
● ● FA(93) ● ● HIT(3) ☆ ☆ BT(73) ■ ■ LG(0)
2017 suite4 np
120°W
90°W
60°W
● ● FA(98) ● ● HIT(5) ☆ ☆ BT(73) ■ ■ LG(0)

Figure 43. Tropical cyclogenesis verification for the NH for 2016 (top row) and 2017 (bottom row) for Suites 1-4 (left to right). Symbols represent the Best Track (black), hits (green), late Genesis (blue) and false alarms (red).

Spectral Decomposition

This section describes the kinetic energy spectrum from the horizontal and vertical wind components.

- **Method:** The cubed-sphere grid does not directly lend itself to the computation of KE spectra given its unstructured nature and lack of a global coordinate system. To compute spectra, we used the UPP postprocessor to interpolate the cubed-sphere fields to an equal latitude-longitude grid (0.125°) having a mesh spacing similar to that of the models tested. A spherical harmonics transform is applied to these latitude-longitude fields and the resulting 2D wavenumber decomposition is then summed over spherical harmonics with the same total spherical wavenumber to produce the one-dimensional spectrum. All the spectra are truncated at the minimum wavelength resolvable on the FV3 mesh (i.e., the 2Δ wavelength).

  Kinetic energy spectra for the forecasts are computed using a spectral decomposition of the velocities (*u*,*v*) and *w* along west-east horizontal grid lines on equally-spacing latitude-longitude grid. The energy densities are time averaged over a 5-day period, at 24-h intervals over globe. We begin the time-averaging 120 h into a forecast in order to avoid model spin-up issues. Additionally, the energy densities are spatially averaged (i.e., the energy densities from the west-east grid lines are averaged over the south-north extent of the domain). We have computed spectra on constant pressure surfaces. On NCEP UPP 0.125° grid, the computation of the final spectrum uses an average of more than 1440 individual one-dimensional spectra (from the west-east grid lines). The one-dimensional kinetic energy spectrum, based on the spectral decomposition of the velocity fields, represents a concise measure of energy as a function of length scale.

Horizontal KE spectrum

- **Benchmark:** For reference, when investigating horizontal kinetic energy spectra, the -3 and -5/3 power-law spectra (characteristic of two-dimensional and three-dimensional turbulence) are shown along with 5 times the nominal grid spacing (5Δ=65km).
- **Method:** Model spectra reveals important information concerning resolved and under-resolved modes in a simulation. We can define the effective resolution of a model as the wavelength where a model's spectrum begins to decay relative to the observed spectrum. The scale is estimated using kinetic energy spectra of winds near the tropopause (200 hPa). The variance has been plotted as compensated variances (variance × $k^{5/3}$) to more easily identify features (Skamarock et al. 2014).
- **Results:** The 2016011812 case was chosen to compute kinetic energy spectra, using the 5-10 days forecasts of the case (Fig. 44). Both the IFS/ECMWF initial condition and the Suite 3 spectra start to fall off sharply due to diffusion at approximately 10Δx. Suites 1, 2 and 4, however, fall off at a scale closer to 5Δ, indicating that suite Suites 1, 2 and 4 have a higher effective resolution. Suites 1, 2 and 4 show the expected flattening of the spectrum in the mesoscale indicating a transition from two to three-dimensional turbulence, while the Suite 3 does not.

It should be noted that the damping options specified in the Suite 3 namelist are different from the other suites, i.e., in Suite 3 namelist:

- sponge = 26 # 10 is used in other suites
- tau= 5 # 10 is used in other suites
- rf_cutoff = 750 # pressure below which no Rayleigh damping is applied if tau >0

where ***sponge*** controls the number of layers at the upper boundary on which the 2Δ filter is applied; ***tau*** is the time scale (in days) for Rayleigh friction applied to horizontal and vertical winds (lost kinetic energy is converted to heat, so larger values yield less damping), and ***rf_cutoff*** defines the pressure below which no Rayleigh damping is applied if tau>0.

Compared with IFS/ECMWF initial condition, it seems that Suite 2 is losing energy in the large-to-mesoscale transition scale. The fact that the mesoscale spectra of Suites 1-4 is so different means that the physics suite and namelist options can be reflected in KE structure in mesoscale region.

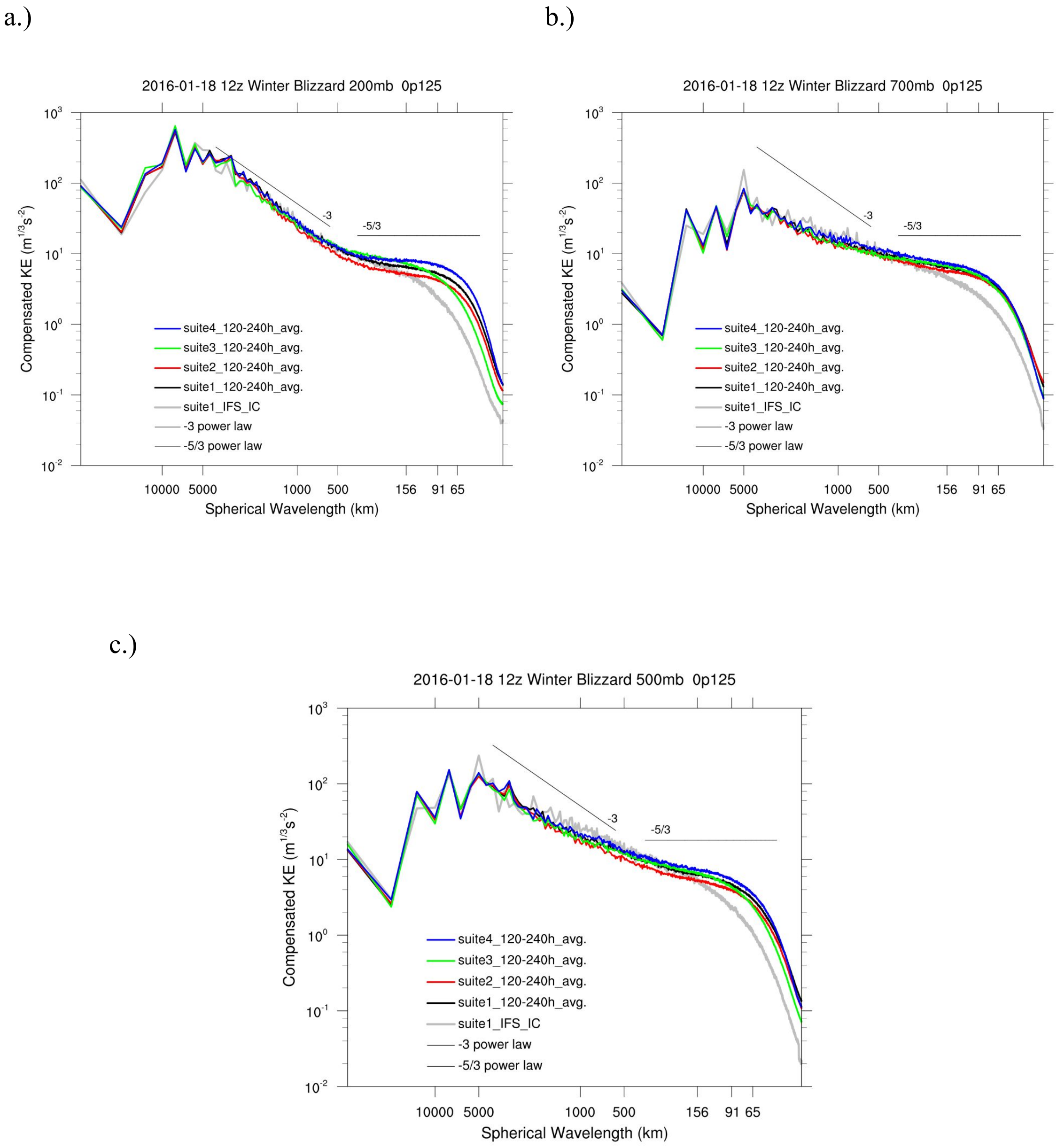


Figure 44. Horizontal KE spectra for 5-10 day forecast at (a) 200 hPa, (b) 500 hPa, and (c) 700 hPa, averaged at 24-h intervals over the globe for the 20160118 12 UTC case. Power-law spectra corresponding to powers of -3 and -5/3 are shown for reference.

Vertical KE spectral

- **Results:** Figure 45 shows the vertical velocity KE spectra from the 2016011812 case. Compared with the horizontal counterpart for the same case, the vertical velocity possesses much less energy than the horizontal velocity spectra, and there is less variation as a function of horizontal wave number. Note the KE scale differs dramatically from the horizontal counterpart.

  At 200 hPa (Fig. 45a), vertical velocity KE, averaged over the globe shows evidence of two peaks in Suite 3 and Suite 4. One occurs at the synoptic scale at a few thousand kilometers, and a second occurs at the mesoscale regime between 5Δ and 9Δ. The mesoscale peak is likely associated with grid-scale convection, waves generated by convection, and other under-resolved small-scale processes. Suite 2 has less energy at the synoptic scales, which presumably is associated with vertical motions associated with large-scale waves.

  At 500 and 700 hPa (Fig. 45b,c), the most significant feature is that Suite 3 exhibits higher vertical velocity KE in the mesoscale regime on the middle and lower troposphere.

  Model robustness consideration: For Suite 3, of 163 total forecasts, 29 became computational unstable and did not complete when using the same 225 second timestep as other suites. However, all of incomplete runs ran to completion with a second job submission or smaller timestep. The higher vertical wind fields in the middle and lower troposphere in Suite 3 may be related the less model robustness.

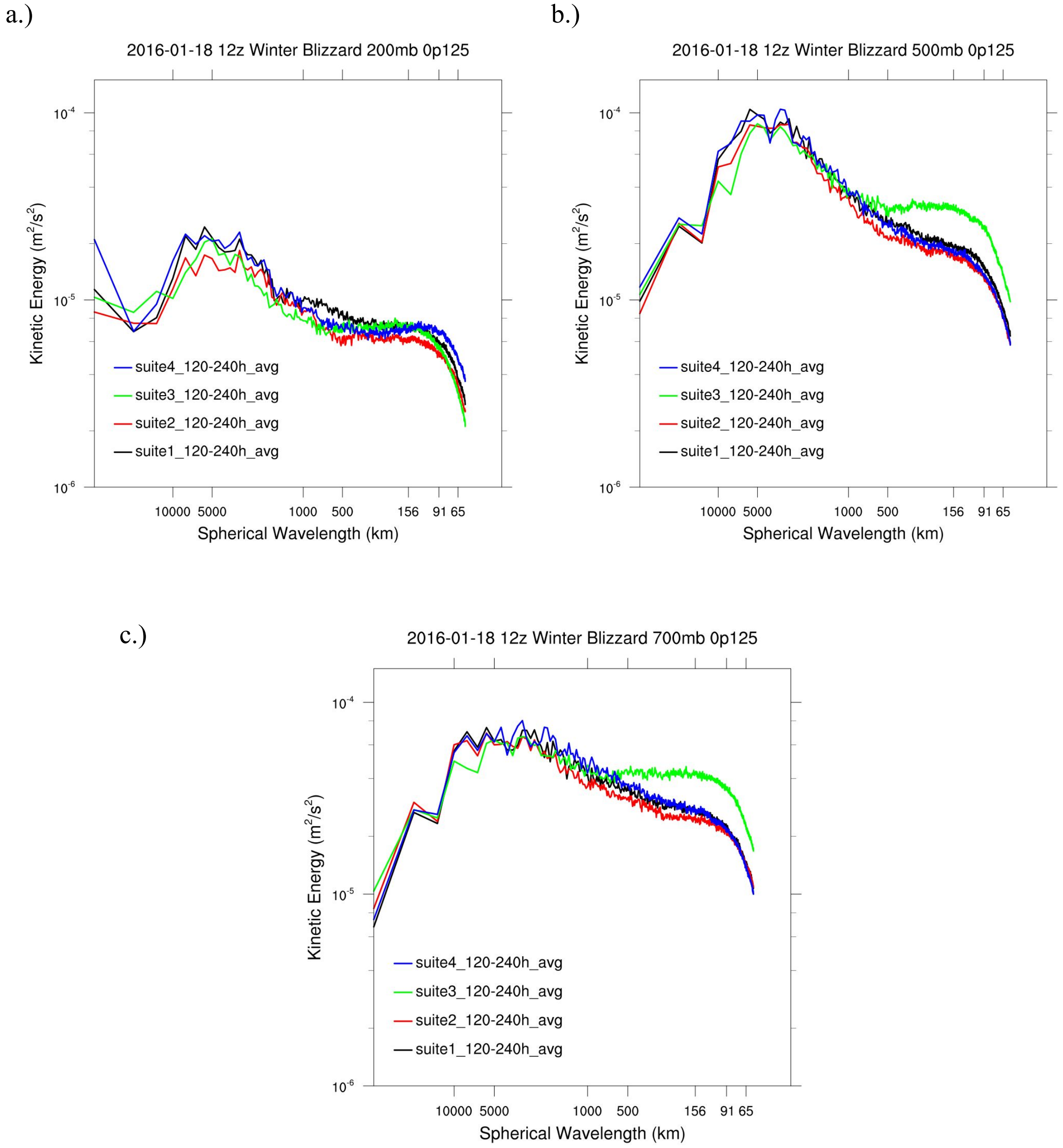


Figure 45. Vertical velocity KE spectra for 5-10 day forecast at (a) 200 hPa; (b) 500 hPa and (c) 700 hPa, averaged at 24-h intervals over the globe for the 20160118 12 UTC case.

Sensitivity to dynamics options and CCPP

- **Methods and Benchmark:** As noted in the initial report (link), the suites differ in multiple aspects beyond physics, including dynamics options, physics-dynamics interface, and age of code base. A limited test was conducted to evaluate the impact of these differences. A configuration was created using the same physical parameterizations as Suite 1, but using the dynamics options, physics-dynamics interface, and code base of Suite 4. Therefore, in this test, Suite 1 was run using the CCPP interface. The test was run for a single case of the test, a Hurricane Irma case initialized on 20170907 at 00 UTC. The results can be intercompared between the original runs conducted for the physics test (Phys-Suite1 and CCPP-Suite4) and the new run conducted for purposes of this sensitivity test (CCPP-Suite1).

- **Results:** Comparison between the physics test runs (Phys-Suite1 and CCPP-Suite4) indicate that the Suite 4 configuration makes much weaker storms (see section on Tropical Cyclone Verification above). Figure 40 exemplifies this for Hurricane Irma: Phys-Suite1 and CCPP-Suite4 produce similar tracks for Hurricane Irma, but the intensity is much weaker for CCPP-Suite4.

  The results for CCPP-Suite1, shown in Figure 46, show a storm of similar track and intensity to Phys-Suite1, suggesting that the weak TCs in the Suite 4 runs are caused by the physics suite and not by the dynamics options or by the use of the CCPP.

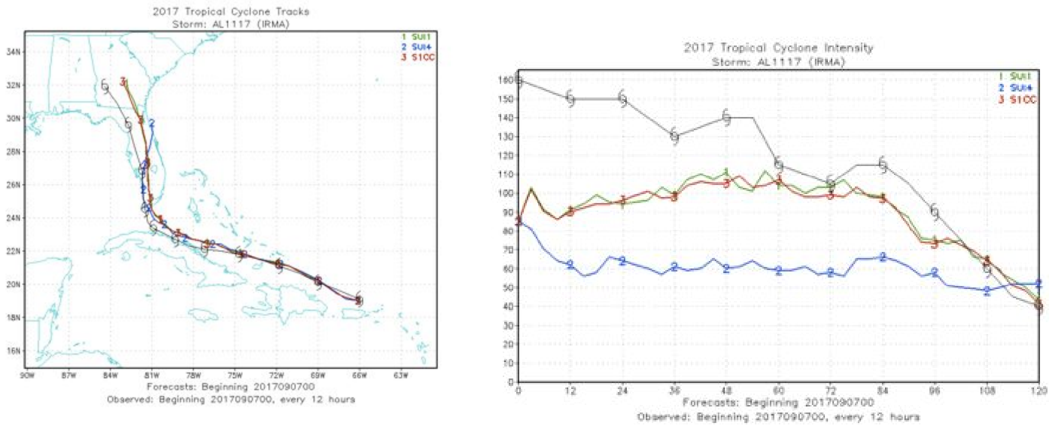


Figure 46. Track (left) and intensity (right; maximum 10-m winds; kt) for Hurricane Irma initialized on 20170907 00 UTC. The original runs for the physics test using Suite 1 and Suite 4 are shown in green and blue, respectively. The sensitivity test CCPP-Suite 1 is shown in red. The real-time guidance from NHC (CARQ) is shown in gray.

References

Biswas, M. K., D. Stark and L. Carson, 2018. GFDL Vortex Tracker Users Guide Version 3.9a. Available at https://dtcenter.org/HurrWRF/users/docs/users_guide/standalone_tracker_UG_v3.9a.pdf.

Halley Gotway, J., K. Newman, T. Jensen, Brown, B., R. Bullock, and T. Fowler, 2018: The Model Evaluation Tools v8.0 (METv8.0) User's Guide. Developmental Testbed Center. Available at: http://www.dtcenter.org/met/users/docs/users_guide/MET_Users_Guide_v8.0.pdf. 407 pp.

Halperin, D.J., E.F. Henry, R.E. Hart, J.H. Cossuth, and Philip Sura, 2013: An Evaluation of Tropical Cyclone Genesis Forecasts from Global Numerical Models. *Weather and Forecasting*, **28**, 1423-1445.

Hart, R.E, 2003: A Cyclone Phase Space Derived from Thermal Wind and Thermal Asymmetry. *Mon. Wea. Rev.* **131**, 585-615.

Skamarock, W.C., S. Park, J.B. Klemp, and C. Snyder, 2014: Atmospheric Kinetic Energy Spectra from Global High-Resolution Nonhydrostatic Simulations. *J. Atmos. Sci.,* **71**, 4369–4381, https://doi.org/10.1175/JAS-D-14-0114.1

Yanai, M., S. Esbensen, and J. Chu, 1973: Determination of Bulk Properties of Tropical Cloud Clusters from Large-Scale Heat and Moisture Budgets. *J. Atmos. Sci.,* **30**, 611–627, https://doi.org/10.1175/1520-0469(1973)030<0611:DOBPOT>2.0.CO;2

Yanai, M. and T. Tomita, 1998: Seasonal and Interannual Variability of Atmospheric Heat Sources and Moisture Sinks as Determined from NCEP–NCAR Reanalysis. *J. Climate,* **11**, 463–482, https://doi.org/10.1175/1520-0442(1998)011<0463:SAIVOA>2.0.CO;2

Acknowledgments

EMC workflow design, testing, collaboration, and execution: Ratko Vasic
EMC VSDB, coordination, and collaboration: Fanglin Yang
EMC technical development: Anning Chang, Shrinivas Moorthi, Jongil Han, Ruiyu Sun, Fanglin Yang, George Gayno, Guang-Ping Lou
ECMWF data sharing: Linus Magnusson, Florian Pappenberger
GSD technical Development: Georg Grell, Joe Olson, Tanya Smirnova, Haiqin Li, Dom Heinzeller
NCAR METplus team: Tara Jensen, John Halley Gotway, Minna Win-Gildenmeister, George McCabe, and Julie Prestopnik
Independent expert panel: Jian-Wen Bao (NOAA/ESRL/PSD), Lisa Bengtsson (CIRES at NOAA/ESRL/PSD), Jim Doyle (Naval Research Laboratory), and Jimy Dudhia (NCAR)
Test coordination: Jack Kain (EMC)

Appendix A: Location of additional materials and figures in DTC website

- Scorecards for precipitation and for T, RH, winds against observation (upper air and surface): https://dtcenter.org/eval/gmtb/2019_advphystest/scorecard/
- Precipitation bias: https://dtcenter.org/eval/gmtb/2019_advphystest/FBias/.
- Precipitation FSS: https://dtcenter.org/eval/gmtb/2019_advphystest/FSS/
- Verification of energy budget components against CERES (TOA upward SW and LW, Surface downward and upward SW: https://dtcenter.org/eval/gmtb/2019_advphystest/CERES/
- Verification of energy budget components against SURFRAD: https://dtcenter.org/eval/gmtb/2019_advphystest/SURFRAD/

Appendix B: Acronyms

AA: Aerosol Aware
BUFR: Binary Universal Form for Representation of Meteorological Data
CAPE: Convective available potential energy
CCPA: Climatology-Calibrated Precipitation Analysis
CCPP: Common Community Physics Package
CERES: Clouds and the Earth Radiant Energy System
CPC: Climate Prediction Center
CMORPH: CPC Morphing technique
DTC: Developmental Testbed Center
ECMWF: European Centre for Medium-Range Weather Forecasts
EMC: Environmental Modeling Center
FSS: Fractions Skill Score
FV3: Finite-Volume Cubed-Sphere dynamical core
FV3GFS: Version of the GFS that employs the FV3 dynamical core
IC: Initial Conditions
K-EDMF: Hybrid Eddy-Diffusivity Mass-Flux PBL parameterization that employs K-theory
ECMWF: European Centre for Medium-Range Weather Forecasts
EMC: NOAA Environmental Modeling Center
ESRL: NOAA Earth System Research Laboratory
ETS: Equitable Threat Score
GDAS: Global Data Assimilation System
GF: Grell-Freitas cumulus parameterization
GFDL: NOAA Geophysical Fluid Dynamics Laboratory
GFS: Global Forecast System
GMTB: Global Model Test Bed
GSD: Global Systems Division
GSS: Gilbert Skill Score
HPSS: High Performance Storage System
IGRA: Integrated Global Radiosonde Archive
MEG: Model Evaluation Group
MET: Model Evaluation Tools
MF: Mass Flux
MG3: Morrison-Gettelman microphysics parameterization version 3
MYNN: Mellor-Yamada-Nakanishi-Niino POBL parameterization
NCAR: National Center for Atmospheric Research
NCDC: National CLimatic Data Center
NCEP: National Centers for Environmental Prediction
NDAS: North American Data Assimilation System
NEMS: NOAA Environmental Modeling System
NH: Northern Hemisphere
NRL: Navy Research Laboratory
NOAA: National Oceanic and Atmospheric Administration
PBL: Planetary Boundary Layer
PLCL: Pressure of the lifting condensation level
PDF: Probability density function
PPN: Processor Per Node
PW: Precipitable water
QPE: Quantitative Precipitation Estimate
RRTM: Rapid Radiative Transfer Model
RRTMG: RRTM for General Circulation Models
SA: Scale Aware
SAS: Simplified Arakawa-Schubert cumulus parameterization
SSMIS: Special Sensor Microwave Imager / Sounder
SURFRAD: Surface Radiation Network
SH: Southern Hemisphere

SS: Statistical significance
TKE: Turbulent Kinetic Energy
TKE-EDMF: EDMF PBL parameterization based on TKE
Tropical Cyclone: TC
UPP: Unified Post Processor
UFS: Unified Forecast System
UTC: Coordinated Universal Time
VLab: NOAA’s Virtual Laboratory
VSDB: Verification Statistics Database