The EMC Model Evaluation Group's Assessment of Advanced Physics Testing

22 March 2019

Geoff Manikin, Alicia Bentley, Logan Dawson, and Tracey Dorian

Point of Contact: geoffrey.manikin@noaa.gov

INTRODUCTION

To support NOAA's Environmental Modeling Center (EMC) in selecting an advanced physics suite for Version 16 of the Global Forecast System (GFS), the Global Model Test Bed (GMTB) tested four configurations of NOAA's Unified Forecast System (UFS). EMC's Model Evaluation Group (MEG), a central team within EMC's Verification, Post-Processing, and Product Generation (VPPG) Branch, is assisting with the effort of selecting a physics suite for GFSv16 by assessing statistics and forecasts from specific cases, within the framework of the MEG possessing deep knowledge of operational GFS and the GFS with the FV3 model core (FV3GFS or GFSv15) biases and overall performance. The GMTB initialized runs every five days between 1 January 2016 and 31 December 2017, alternating between 0000 and 1200 UTC initial times. The GMTB ran sixteen additional cases at the recommendation of the MEG, covering a wide variety of high-impact weather events including tropical cyclones, winter storms, significant rainfall events, severe weather, and extreme heat. Other cases were selected to assess performance of the suites with respect to well-documented GFS biases. Images from these cases can be found on the MEG GFS Physics Evaluation web site (which has not been publicized beyond those involved in this project.)

This test involved a collaboration between GMTB, EMC, NOAA Earth System Research Laboratory (ESRL) Global Systems Division (GSD), the Naval Research Laboratory (NRL), NOAA ESRL Physical Sciences Division (PSD), and the National Center for Atmospheric Research (NCAR). Full details of the model configurations can be found in the GMTB initial report, but Table 1 provides a quick overview.

Table 1. *Physics suite, dynamics namelist options (latter indicated by blue color fill), computational options (indicated by green color), and code base (indicated by yellow color) for preliminary advanced physics testing. Acronyms are defined in Appendix A. The definition of the dynamics namelist options can be found at https://www.gfdl.noaa.gov/wp-content/uploads/2017/09/fv3_namelist_Feb2017.pdf.*

	Suite 1 (GFS v15)	**Suite 2**	**Suite 3**	**Suite 4**
Deep convection	SA-SAS	SA-SAS	CS-AW	SA/AA-GF
Shallow convection	SA-MF	SA-MF	SA-MF	MYNN-EDMF and SA GF
Microphysics	GFDL	GFDL	AA-MG3	AA-Thompson
Saturation adjustment in dycore	True	True	True	False
PBL/Turbulence	K-EDMF	SA-TKE-EDMF	K-EDMF	MYNN-EDMF
Land Surface Model	Noah	Noah	Noah	RUC
Physics-Dynamics coupling	non-CCPP	non-CCPP	non-CCPP	CCPP
nord	2	2	2	3
dddmp	0.1	0.1	0.1	0.2
d4_bg	0.12	0.12	0.12	0.15
vtdm4	0.02	0.02	0.02	0.06
sponge	10	10	26	10
tau	10	10	5	10
hord_mt	5	5	6	5
hord_ct	5	5	6	5
hord_tm	5	5	6	5
hord_dp	-5	-5	-6	-5
Platform	xjet	xjet	xjet	vjet
Intel compiler	v15	v15	v15	v18
Nodes/PPN	72/12	72/12	72/12	108/16
Layout	8x16	8x16	8x16	16x16
Threads	2	2	2	72
Code base	Oct 2018	Nov 2018	Nov 2018	Nov 2018

STATISTICS

GMTB provided a complete diagnostic report, and some of those statistics and diagnostics are included in the next two sections. EMC also generated a comprehensive set of statistics, comparing forecasts to both ECMWF analyses and observations, and those are scrutinized in this section.

The 0000 UTC 500-hPa geopotential height anomaly correlation (AC) scores for Physics Suites 1–4 reveal notable differences in each suite's ability to capture the structure of the mid-level flow pattern (Fig. 1). Suites 1 and 2 have the highest 500-hPa AC scores of any physics suite through Day 10, remaining nearly identical to each other through Day 8. Suites 3 and 4 both have statistically significantly worse 500-hPa AC scores than Suites 1 and 2, with Suite 3 exhibiting the fastest reduction in 500-hPa AC scores of any physics suite. ECMWF 500-hPa AC scores remain statistically significantly better than those of all four physics suites through Day 6. While only 0000 UTC 500-hPa AC scores are shown, it should be noted that 1200 UTC 500-hPa AC scores look extremely similar. This extreme similarity is true of the majority of statistics examined in this summary. For this reason, only 0000 UTC statistics will be discussed.

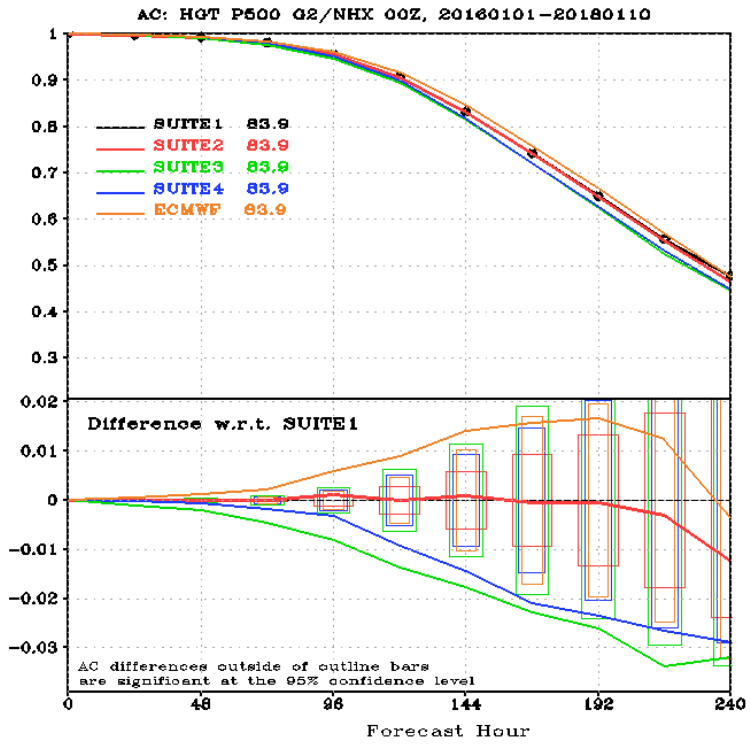


Fig. 1. Northern Hemisphere 500-hPa geopotential height AC score die-off curves for Physics Suites 1–4 and the ECMWF during 0000 UTC 1 January 2016–0000 UTC 1 January 2018.

Suites 1 and 2, which utilize the same GFDL microphysics scheme as the FV3GFS retrospectives and real-time parallel, exhibit a low 500-hPa geopotential height bias that increases with forecast lead time (Fig. 2). This low height bias grows at the same rate in both physics suites, consistent with their nearly identical configurations. Suite 3 differs considerably from Suites 1 and 2, exhibiting a high 500-hPa geopotential height bias that increases with forecast lead time. Suite 4 has the best mean 500-hPa geopotential height bias of the four physics suites, with a bias remaining relatively close to 0 at all forecast lead times.

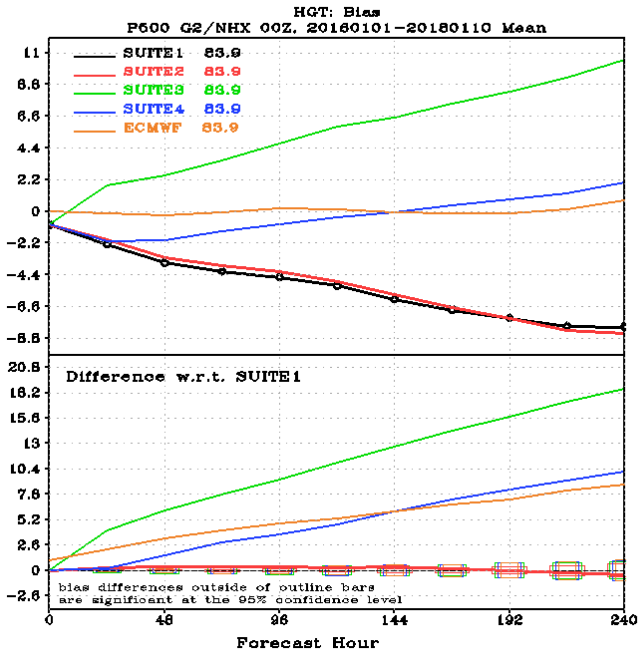


Fig. 2. Northern Hemisphere 500-hPa geopotential height bias from Physics Suites 1–4 and the ECMWF as a function of forecast lead time during 0000 UTC 1 January 2016–0000 UTC 1 January 2018.

The low 500-hPa geopotential height bias in Suites 1 and 2 is 1) worse than the low bias in the FV3GFS retrospectives and 2) better than the low bias in the real-time parallel (Fig. 2). These results can be explained by the inclusion of the radiation bug fix in the real-time parallel and Suites 1 and 2, which is not included in the FV3GFS retrospectives. The low height bias in Suites 1 and 2 did not get as bad as the low bias in the real-time parallel because Suites 1 and 2 were initialized using ECMWF analyses (i.e., no cycling).

Based on the principles of thermodynamics, the 850-hPa temperature bias patterns (Fig. 3) should be similar to the 500-hPa geopotential height bias patterns in all four physics suites. This is mostly true, with Suites 1 and 2 exhibiting a cold bias, Suite 3 exhibiting a warm bias, and Suite 4 remaining close to 0 at all forecast lead times. The biggest difference from the expected bias patterns occurs in Suite 2, which exhibits a 850-hPa temperature bias that is notably colder than the bias in Suite 1. The PBL scheme is the only difference between Suites 1 and 2, suggesting that the updated PBL scheme in Suite 2 (which includes TKE) may be causing an increase in the low-level cold bias.

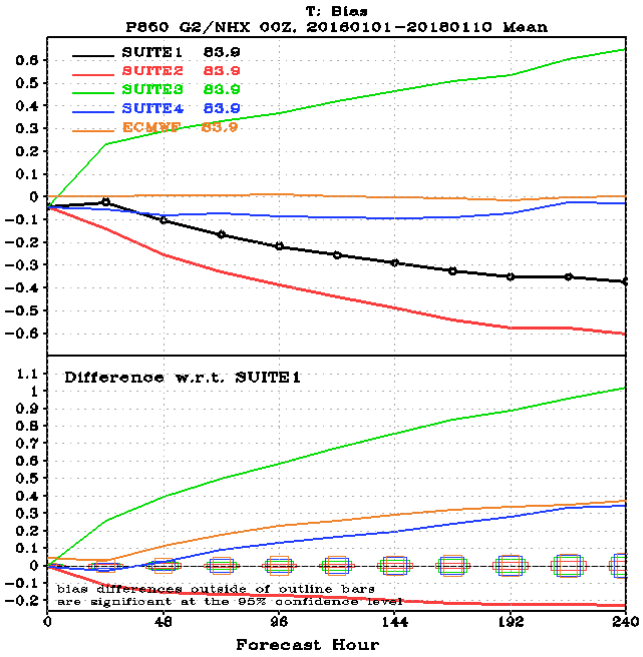


Fig. 3. Northern Hemisphere 850-hPa temperature bias from Physics Suites 1–4 and the ECMWF as a function of forecast lead time during 0000 UTC 1 January 2016–0000 UTC 1 January 2018.

Mitigating the low-level cold bias in the FV3GFS is a major goal of the GFSv15 implementation, so it is important to investigate the potential causes of this low-level cold bias in GFSv16. Further investigation into the 850-hPa temperature bias (Fig. 4) reveals that the inclusion of the radiation bug fix in the FV3GFS (i.e., Suite 1) results in an amplification of the seasonal fluctuations in the 850-hPa temperature bias seen in the GFS and FV3GFS retrospectives (warm bias in summer and cold bias in winter). Similar to Suite 1, the FV3GFS real-time parallel also includes the radiation bug fix, but this version of the model has not been run for a long enough period of time with the radiation bug fix for the amplified seasonal cycle of 850-hPa temperature bias to be seen.

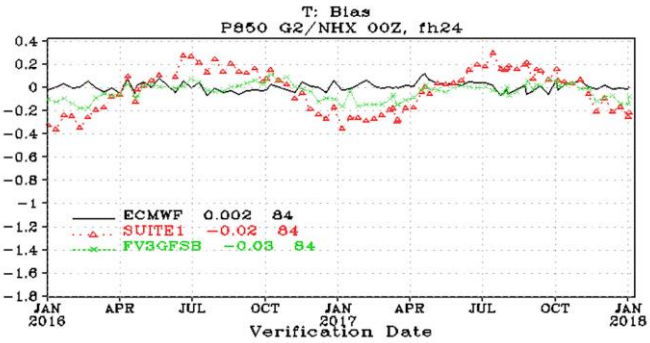


Fig. 4. Northern Hemisphere 850-hPa temperature bias from the ECMWF, Physics Suite 1, and FV3GFS retrospectives as a function of time of year during 0000 UTC 1 January 2016–0000 UTC 1 January 2018.

Large seasonal fluctuations in the 850-hPa temperature bias (larger than those seen in the FV3GFS retrospectives) can be seen in all four physics suites, regardless of the microphysics scheme used (Fig. 5). All four physics suites include the radiation bug fix. It is also important to note that, unlike the FV3GFS retrospectives, Physics Suites 1–4 are initialized from ECMWF analyses and are not fully cycled. While "spin up" issues could affect the 850-hPa temperature bias at 24 h, differences between the four physics suites resulting from each suite's own unique temperature biases can already be seen at this time, suggesting that the models have had time to acclimate.

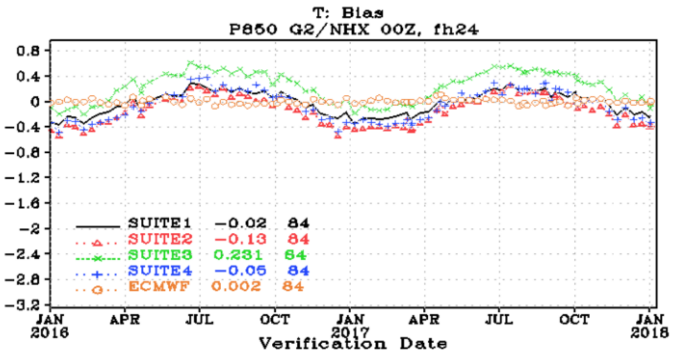


Fig. 5. Day-1 Northern Hemisphere 850-hPa temperature bias from Physics Suites 1–4 as a function of time of year during 0000 UTC 1 January 2016–0000 UTC 1 January 2018.

The seasonal cycle in 850-hPa temperature bias is nearly symmetric about 0 at Day 1, but can be shifted (colder or warmer) at longer forecast lead times (e.g., Day 5) by each physics suite's own unique biases (Fig. 6). For example, the cold bias in Suite 1 and 2 shifts both curves colder during all seasons by Day 5, whereas the warm bias in Suite 3 shifts the curve warmer during all seasons by Day 5. The amplified seasonal cycles of temperature bias in Physics Suites 1–4 are washed out and cannot be seen when averaging over forecast hour (Fig. 3).

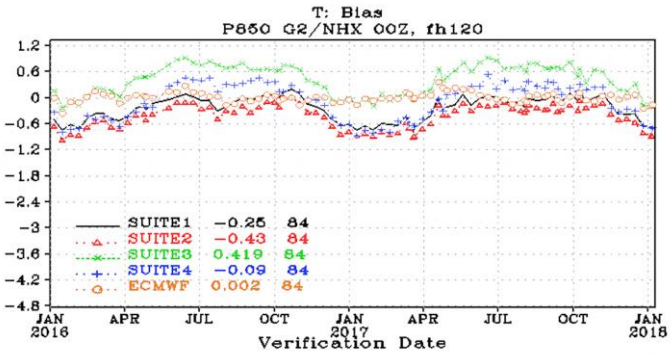


Fig. 6. Day-5 Northern Hemisphere 850-hPa temperature bias from Physics Suites 1–4 as a function of time of year during 0000 UTC 1 January 2016–0000 UTC 1 January 2018.

This winter's large 850-hPa cold bias is likely caused by the combination of a) the amplified seasonal cycle of low-level temperature bias being "cold" in winter and b) the cold bias associated with GFDL microphysics increasing with forecast lead time. Without mitigating the amplified seasonal cycle of low-level temperature bias caused by the radiation bug fix (i.e., an

imbalance in the energy budget), Suites 1 and 2 will continue to have a large cold bias during winter, resulting in inflated snowfall totals and inaccurate low-level temperature forecasts. In summer, low-level temperature forecasts from Suites 1 and 2 will have relatively little bias, as the cold bias associated with GFDL microphysics interacts with the seasonal cycle of low-level temperature bias (which is warm in summer). In contrast to Suites 1 and 2, Suite 3 will have a small cold bias in winter and large warm bias in summer as its own warm bias interacts with the "warm" seasonal cycle of low-level temperature bias. With little inherent bias of its own, Suite 4 will have a warm bias in the summer and a cold bias in the winter associated with the amplified seasonal cycle of low-level temperature bias alone. When averaged throughout the year, the seasonal changes in low-level temperature bias look like a near-zero low-level temperature bias at all forecast lead times in Suite 4.

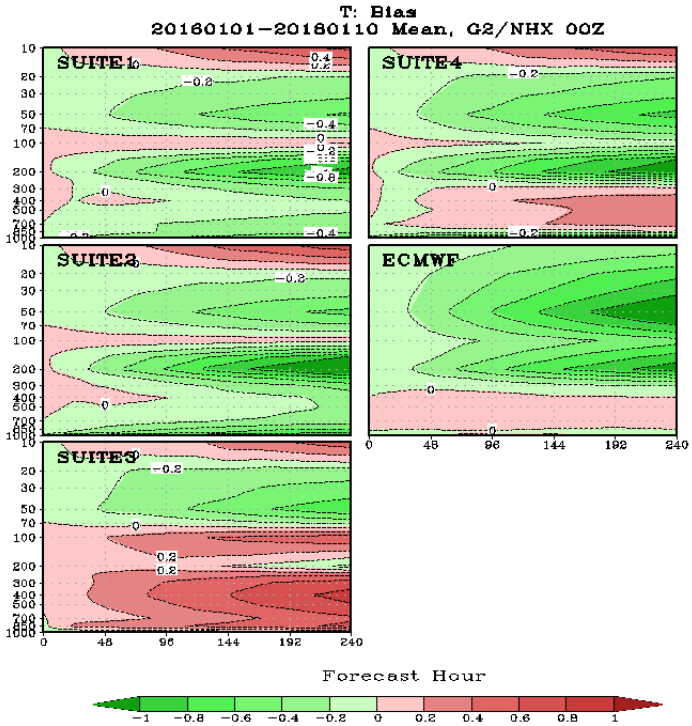


Fig. 7. Vertical profile of Northern Hemisphere temperature bias from Physics Suites 1–4 and the ECMWF as a function of forecast lead time during 0000 UTC 1 January 2016–0000 UTC 1 January 2018.

Vertical profiles of temperature bias as a function of forecast lead time (Fig. 7) highlight many of the key points previously discussed. Suites 1 and 2 exhibit a low-level cold bias that gets worse with forecast lead time. Interestingly, Suite 2 also exhibits a cold bias at 200 hPa that is colder than the 200-hPa cold bias in Suite 1. Suite 3 has a warm bias throughout the majority of the troposphere that increases with forecast lead time. Suite 4 has the least low-level and mid-level temperature bias of the four physics suites, but an intense cold bias near the surface when averaged over the Northern Hemisphere. Suite 2 also has an intense cold bias near the surface when averaged over the Northern Hemisphere. These intense near-surface cold biases are likely the results of the different (planetary boundary layer) PBL schemes used

in each suite. A comparison of low-level temperature forecasts to observations near the surface provides additional information.

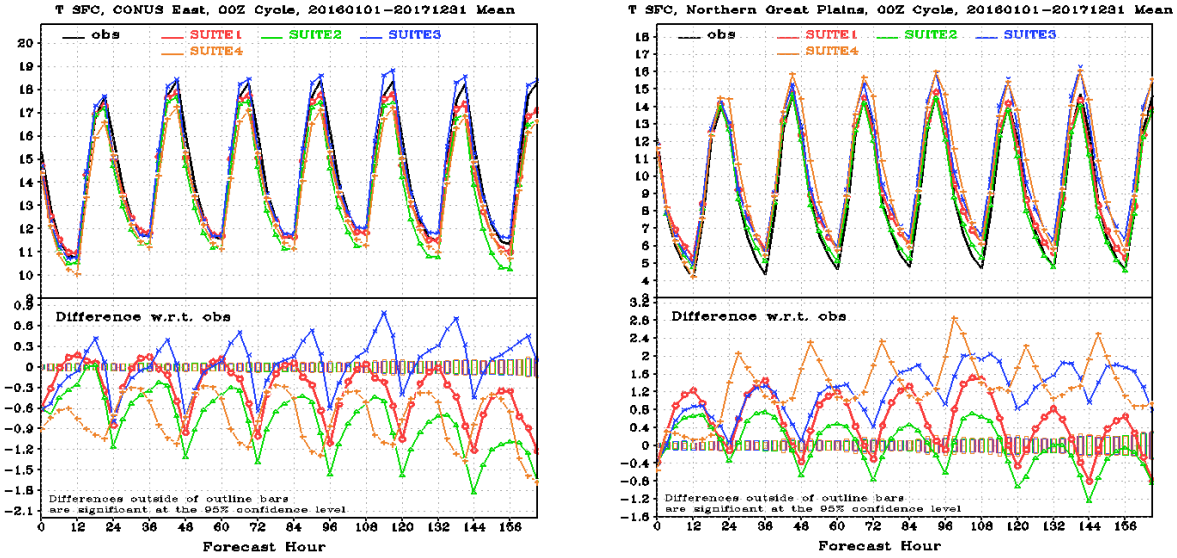


Fig. 8. Time series of a) eastern CONUS and b) Northern Great Plains 2-m temperature forecasts relative to observations from Physics Suites 1–4 as a function of forecast lead time during 0000 UTC 1 January 2016–0000 UTC 31 December 2017.

A comparison of 2-meter temperature forecasts relative to observations from Physics Suites 1–4 reveals that all physics suites except Suite 3 have a near-surface cold bias at the majority of forecast lead times over the eastern CONUS (Fig. 8a). The 2-m temperature cold bias over the eastern CONUS is the worst in Suite 2, followed closely by Suite 4. It is interesting to note, however, that Suite 4 can have a 2-m temperature bias of the opposite sign over different locations. For example, over the Northern Great Plains (Fig. 8b), Suite 4 can have a noteworthy warm bias (warmer than observations at peak heating *and* in the overnight hours). In fact, the temperature bias patterns seen over the eastern CONUS are all shifted warmer over the northern Great Plains, but Suite 4 shows the most dramatic shift, surpassing Suites 1 and 3.

In the northern and southern Mountain regions of the CONUS (Figs. 9a,b), Suite 2 consistently does the worst job with 2-m temperatures. Suite 4 arguably does the best job, with the smallest values and smallest diurnal cycle of differences with respect to observations. This is also reflected in the root mean square error (RMSE) of each suite over this region, with Suite 4 and Suite 2 producing the best and worst 2-m temperature forecasts, respectively (not shown). The same distribution is true along the northwest coast and southwest coast (not shown), with Suite 4 performing the best with 2-m temperature relative to observations and Suite 2 performing the worst relative to observations.

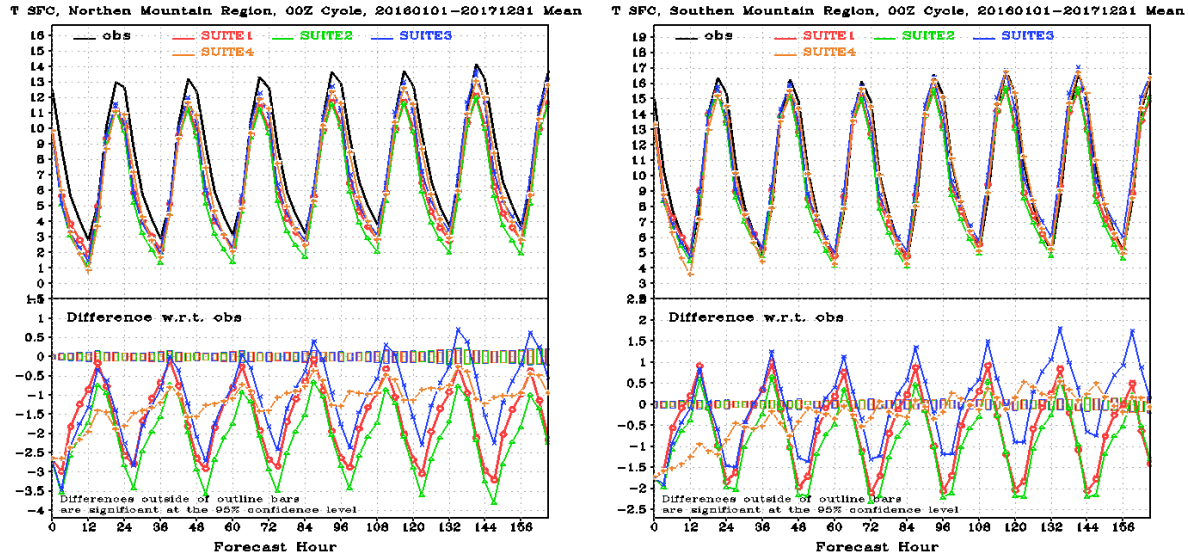


Fig. 9. Time series of a) Northern Mountain Region and b) Southern Mountain Region 2-m temperature forecasts relative to observations from Physics Suites 1–4 as a function of forecast lead time during 0000 UTC 1 January 2016–0000 UTC 31 December 2017.

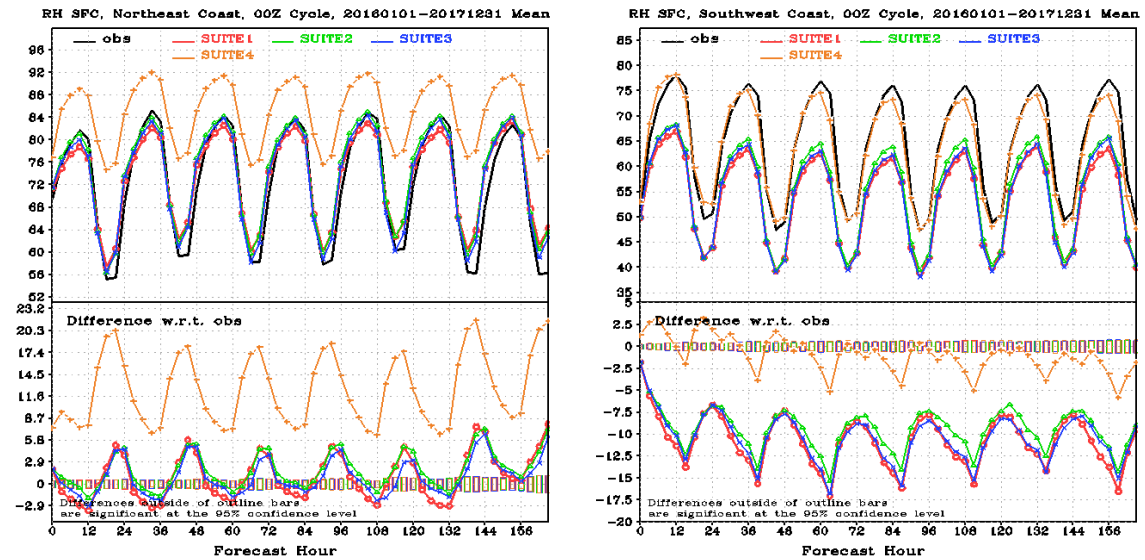


Fig. 10. Time series of a) Northeast Coast and b) Southwest Coast 2-m relative humidity forecasts relative to observations from Physics Suites 1–4 as a function of forecast lead time during 0000 UTC 1 January 2016–0000 UTC 31 December 2017.

A comparison of 2-m relative humidity forecasts relative to observations from Physics Suites 1–4 reveals that all physics suites are fairly close to observations along the Northeast Coast except for Suite 4, which is always too moist (Fig. 10a). Interestingly, the opposite is true along the Southwest Coast (Fig. 10b). Along the Southwest Coast, Suite 4 produces the only forecasts near observed relative humidity values. All other physics suites produce forecasts that are much too dry over the Southwest Coast. All four physics suites do a good job of matching 2-m relative humidity observations over the Northern and Southern Mountain Regions (not shown). Over the Northern and Southern Great Plains (Figs. 11a,b), Suite 2 does the best job accurately capturing 2-m relative humidity values, whereas Suite 4 does the worst job. It is interesting to note that Suites 1 and 2 seem to moisten near the surface at longer forecast lead times.

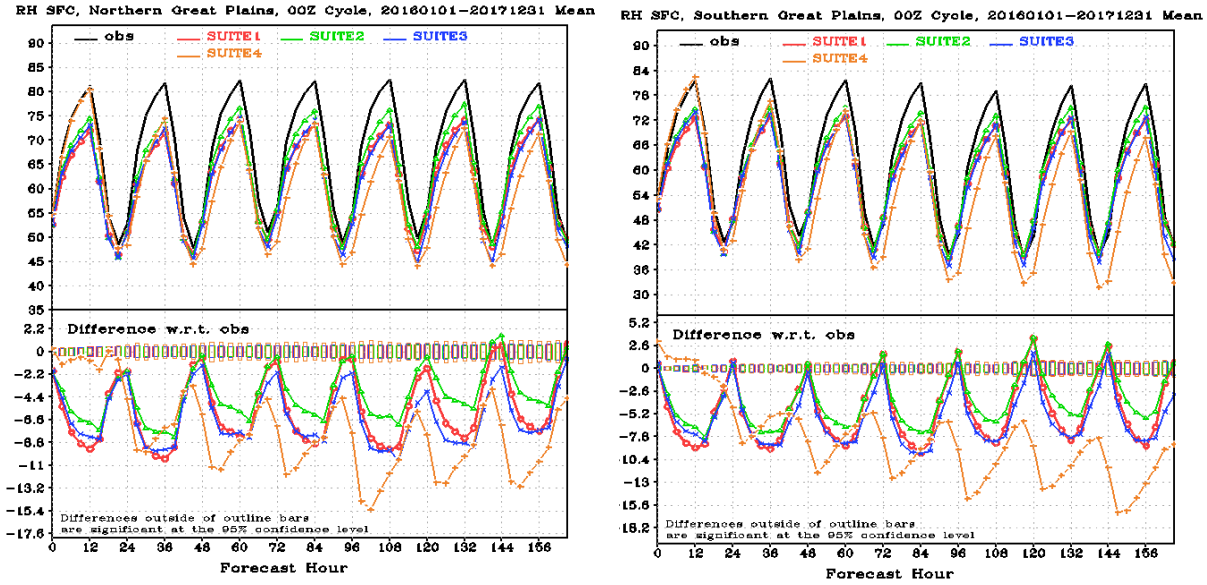


Fig. 11. Time series of a) Northern Great Plains and b) Southern Great Plains 2-m relative humidity forecasts relative to observations from Physics Suites 1–4 as a function of forecast lead time during 0000 UTC 1 January 2016–0000 UTC 31 December 2017.

A comparison of the vertical profiles of temperature bias relative to RAOBS from Physics Suites 1–4 (Figs. 12a,b) reveals that Suite 1 does the best job capturing the majority of the vertical temperature profile at Days 3 and 6. Suite 2 is similar to Suite 1, but has more of a cold bias near the tropopause and near the surface. As expected, the cold bias in Suites 1 and 2 gets worse at longer forecast lead times. Similarly, the warm bias of Suite 3 gets worse with increasing forecast lead time. Suite 4 has a cold bias near the surface, a mid-level warm bias, and a cold bias again near the tropopause, consistent with results previously shown in Fig. 7.

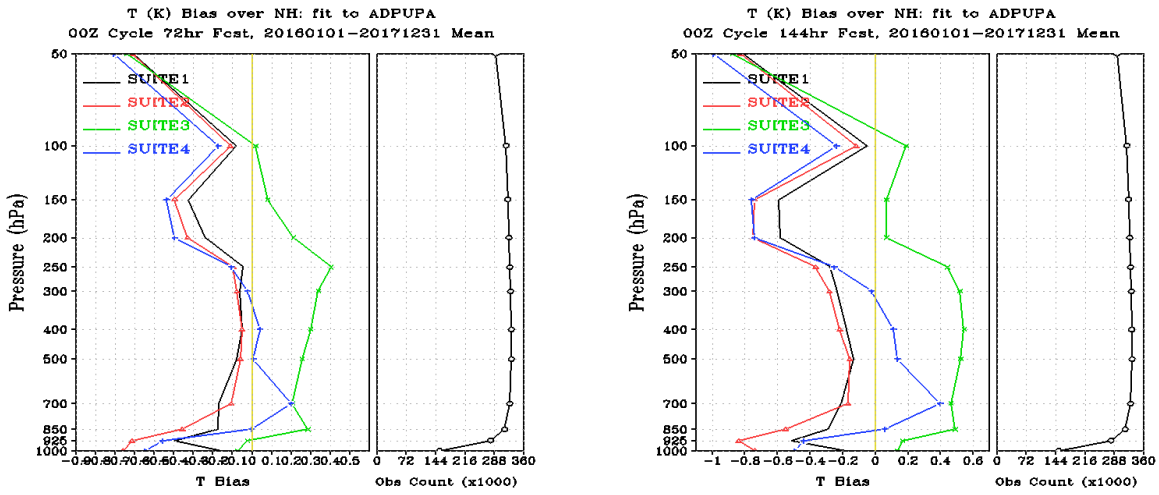


Fig. 12. Vertical profile of Northern Hemisphere temperature bias at a) Day 3 and b) Day 6 relative to RAOBS from Physics Suites 1–4 during 0000 UTC 1 January 2016–0000 UTC 31 December 2017.

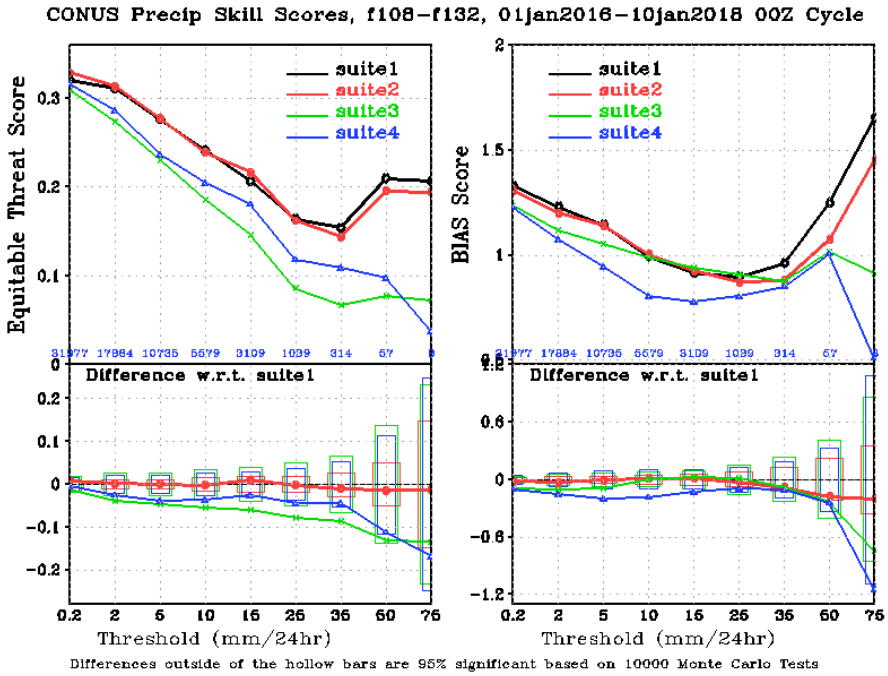


Fig. 13. Distribution of 0000 UTC CONUS precip. a) equitable threat and b) bias scores from Physics Suites 1–4 during 0000 UTC 1 January 2016–0000 UTC 10 January 2018 at Day 5.

A comparison of the 0000 UTC CONUS precipitation equitable threat scores (ETS) from Physics Suites 1–4 (Fig. 13a) reveals that Suites 1 and 2 do the best job capturing CONUS precipitation totals at all thresholds between F108 and F132 (Day 5). Suite 3 does the worst job capturing CONUS precipitation totals at almost all thresholds, with Suite 4 falling somewhere in the middle. A similar pattern in ETS can be seen at Days 4 and 6 (not shown). All four physics suites have a wet bias at low thresholds and dry precipitation bias at middle thresholds (Fig. 13b). Suite 4 notably has the driest CONUS precipitation bias of any suite. There are too few examples of the highest thresholds to draw any meaningful conclusions.

A comparison of the 1200 UTC CONUS precipitation equitable threat scores (ETS) from Physics Suites 1–4 (Fig. 14a) reveals similar results to those shown in Fig. 13. Suites 1 and 2 do the best job capturing CONUS precipitation totals at all thresholds between F96 and F120 (Day 5), whereas Suite 3 does the worst job. However, there are notable differences in CONUS precipitation bias between 0000 UTC and 1200 UTC. For 1200 UTC CONUS precipitation bias (Fig. 14b), Suites 1–3 have a wet bias at low thresholds that approaches unity at middle thresholds, but it does not become a dry bias like at 0000 UTC. Suite 4, however, continues to have the dry bias at 1200 UTC, similar to what exists at 0000 UTC.

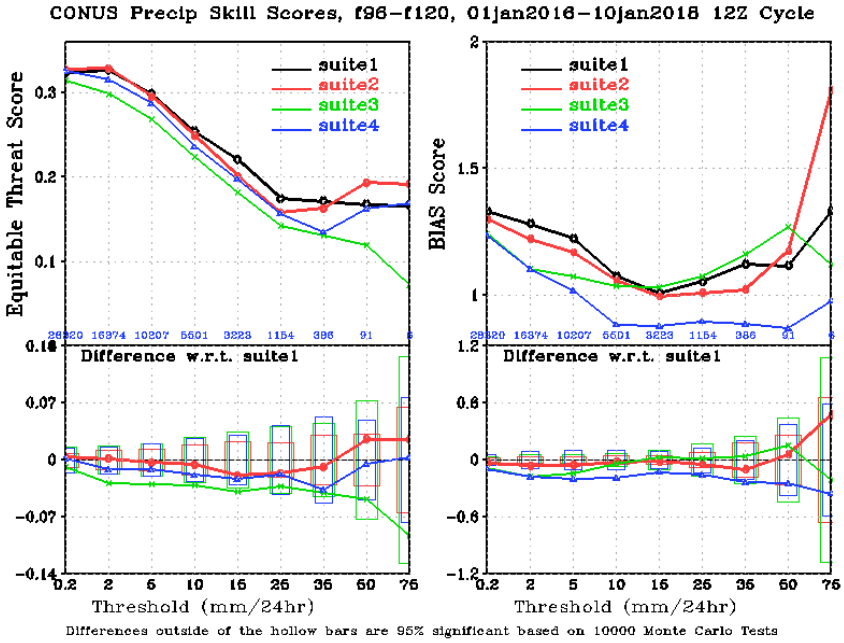


Fig. 14. Distribution of 0000 UTC CONUS precip. a) equitable threat and b) bias scores from Physics Suites 1–4 during 0000 UTC 1 January 2016–0000 UTC 10 January 2018 at Day 5.

An examination of CONUS precipitation bias as a function of forecast hour (Fig. 15) expands upon the themes from Figs. 13–14. At lower precipitation thresholds, a wet bias exists at all forecast lead times in all four suites. All four suites have an increasing dry bias at middle thresholds with time. Suite 4 notably displays the largest dry bias at middle and high thresholds.

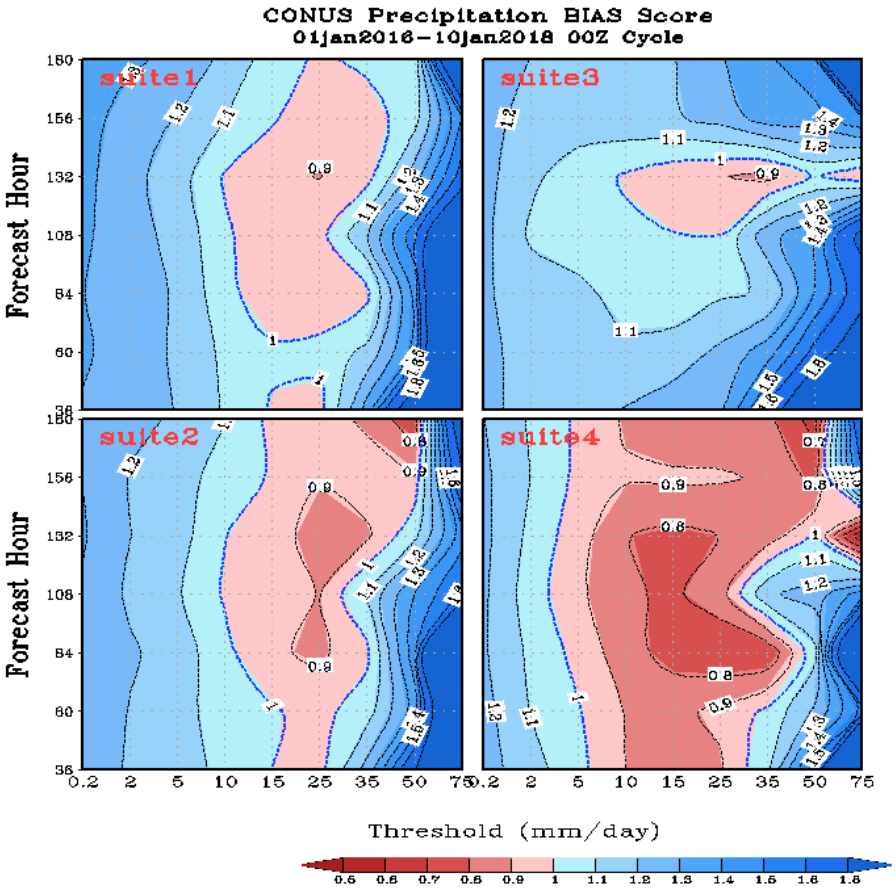


Fig. 15. Distribution of 0000 UTC CONUS precipitation bias as a function of forecast lead time from Physics Suites 1–4 during 0000 UTC 1 January 2016–0000 UTC 10 January 2018.

The precipitation score cards included in the GMTB Diagnostic Report on Advanced Physics Testing Report (not shown) indicate that Suites 2–4 typically have worse precipitation forecasts than Suite 1 across most regions and precipitation thresholds. Suite 2 is very similar to Suite 1 across most of the Northern and Southern Hemisphere, but worse than Suite 1 at the majority of precipitation thresholds in the Tropics. Suite 3 is worse than Suite 1 over much of the Northern Hemisphere, Southern Hemisphere, and Tropics. Suite 4 is worse than Suite 1 at middle thresholds in the Northern Hemisphere, Southern Hemisphere, and Tropics, but is better than Suite 1 at the smallest thresholds in the Southern Hemisphere and Tropics. Maps of accumulated precipitation bias across the globe (Fig. 15), also in the GMTB Diagnostic Report on Advanced Physics Testing Report, suggest that Suites 1–3 typically have a bias in the medium range across most of the Tropics. This dry bias is the smallest in Suite 1, slightly worse in Suite 2, and considerably worse in Suite 3. Suite 4 exhibits a wet bias across the majority of the Tropics in the medium range.

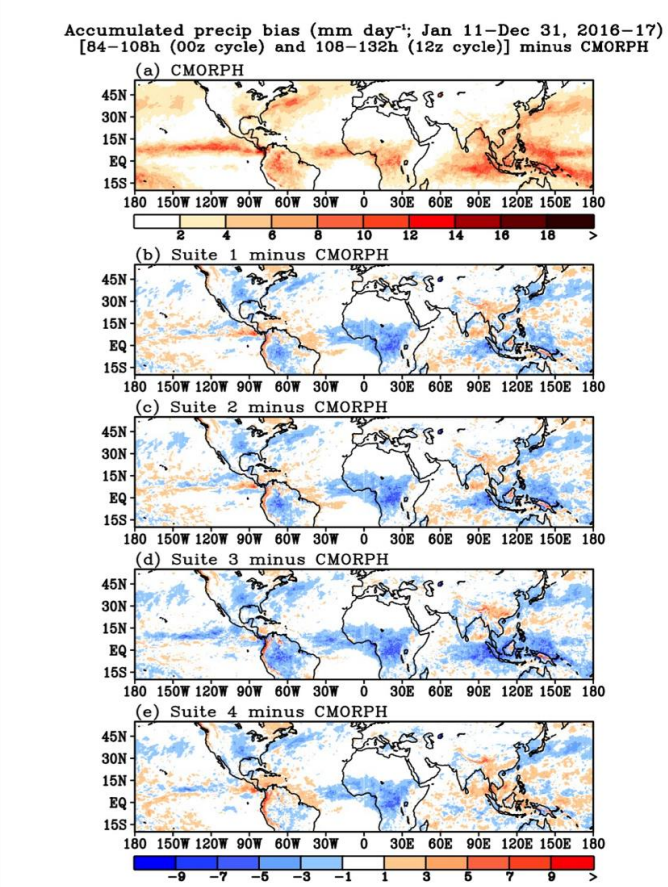


Fig. 16. Maps of accumulated precipitation bias (mm/day) relative to CMORPH observations from Physics Suites 1–4 during 11 January 2016–31 December 2017.

TROPICAL

The performance of FV3GFS forecasts of tropical cyclones (TCs) was assessed for each of the four suites included in the GFS physics testing. This evaluation focused on forecasts of the eight tropical cyclones listed in Table 2. The majority of the cases were Atlantic Basin tropical cyclones that threatened the contiguous United States (CONUS). One storm each was examined for the East Pacific and West Pacific basins. One forecast cycle was evaluated for each tropical cyclone. Generally, the specified forecast cycle was chosen based on past evidence of a major forecast challenge seen in the operational GFS and/or based on the significance of observed impacts from the tropical cyclone.

Table 2: List of tropical cyclones and associated forecast cycles evaluated by the NCEP/EMC Model Evaluation Group

Tropical Cyclone	Forecast Cycle
TC Joaquin (2015)	0000 UTC 1 October 2015
TC Matthew (2016)	0000 UTC 2 October 2016
TC Noru (2017)	0000 UTC 31 July 2017
TC Harvey (2017)	0000 UTC 26 August 2017
TC Irma (2017)	0000 UTC 7 September 2017
TC Nate (2017)	0000 UTC 4 September 2017
TC Lane (2018)	0000 UTC 19 August 2018
TC Florence (2018)	1200 UTC 11 September 2018

Various aspects of the forecast for each case were subjectively evaluated. This included assessment of track forecasts, intensity (maximum wind and minimum pressure) forecasts, and quantitative precipitation forecasts (QPFs). In this section, performance related to each case's specific forecast challenge(s) will be discussed, and general conclusions from this evaluation will be provided.

Tropical Case Summaries

TC Joaquin (2015)

TC Joaquin presented a significant challenge to the operational GFS, which continually produced guidance showing Joaquin making landfall in the Mid-Atlantic region while guidance from other global modeling centers showed Joaquin recurving safely away from the CONUS. During the official evaluation of GFSv15 (the rollout of the FV3GFS), the MEG noted that FV3GFS retrospective forecasts indicated that Joaquin would recurve into the Atlantic prior to the cycle when the operational GFS finally showed the correct forecast solution. For this present evaluation, the 0000 UTC 1 October 2015 forecast from Suites 1 and 2 (which both use a physics suite nearly identical to that of GFSv15) closely followed Joaquin's observed track out to sea (Fig. 17). Corresponding forecasts from Suites 3 and 4, however, showed Joaquin making landfall in the Carolinas before recurving back out to sea. The improved forecast guidance from GFSv15 was considered a significant "win" for the transition of NCEP's global model to the FV3 dynamical core. Based on this single forecast cycle, this current testing would suggest that some aspect(s) of the physics parameterizations in Suites 3 and 4 caused a degradation of the forecast for TC Joaquin.

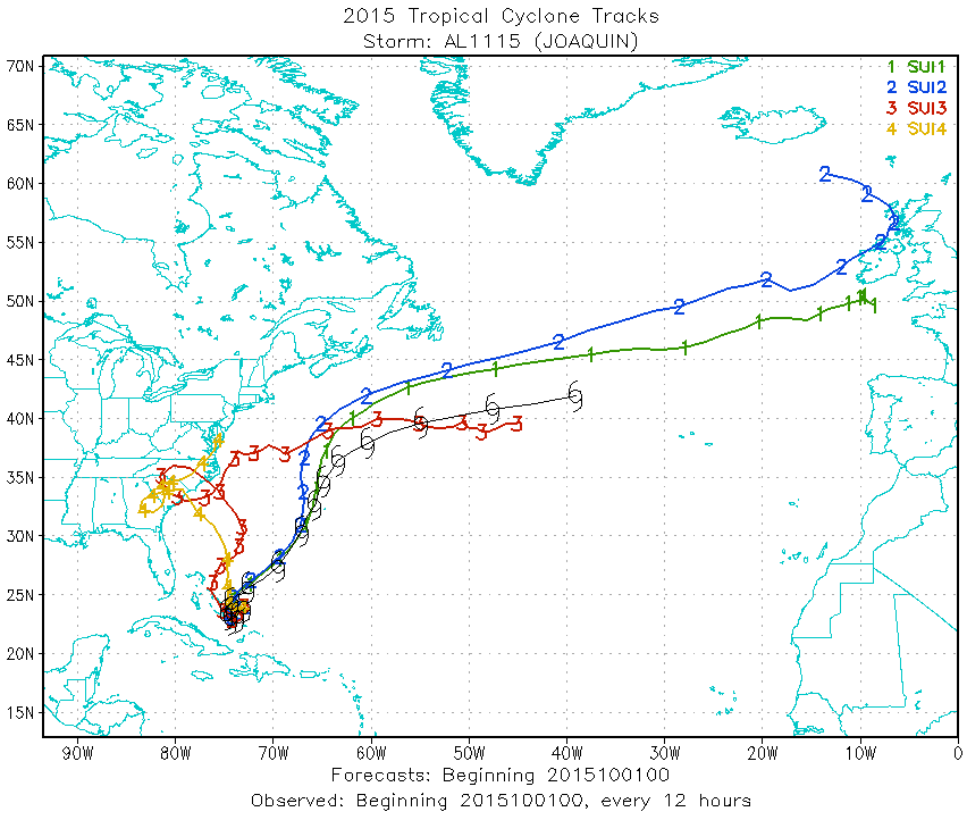


Fig. 17. Track forecasts for TC Joaquin from the 0000 UTC 1 October 2015 forecast cycle

TC Matthew (2016)

TC Matthew tracked north out of the Caribbean Sea as a major hurricane and made a near approach to the southeast US coast. Retrospective forecasts run during the official evaluation of GFSv15 found that Matthew's movement northward out of the Caribbean Sea was accelerated in the FV3GFS compared to the operational GFS. This fast latitude gain was also seen in forecasts run with Suites 1, 2, and 3. TC Matthew's central pressure was comparable to observations and analyses by Days 2-5, but Suite 3's forecast had a much weaker system. This suggests that the strength of a tropical cyclone is likely not the dominant factor causing the erroneously fast latitude gain in FV3GFS forecasts. Suite 4 had the weakest and slowest forecast solution overall, and it tracked Matthew much closer to the Southeast coast (Fig. 18).

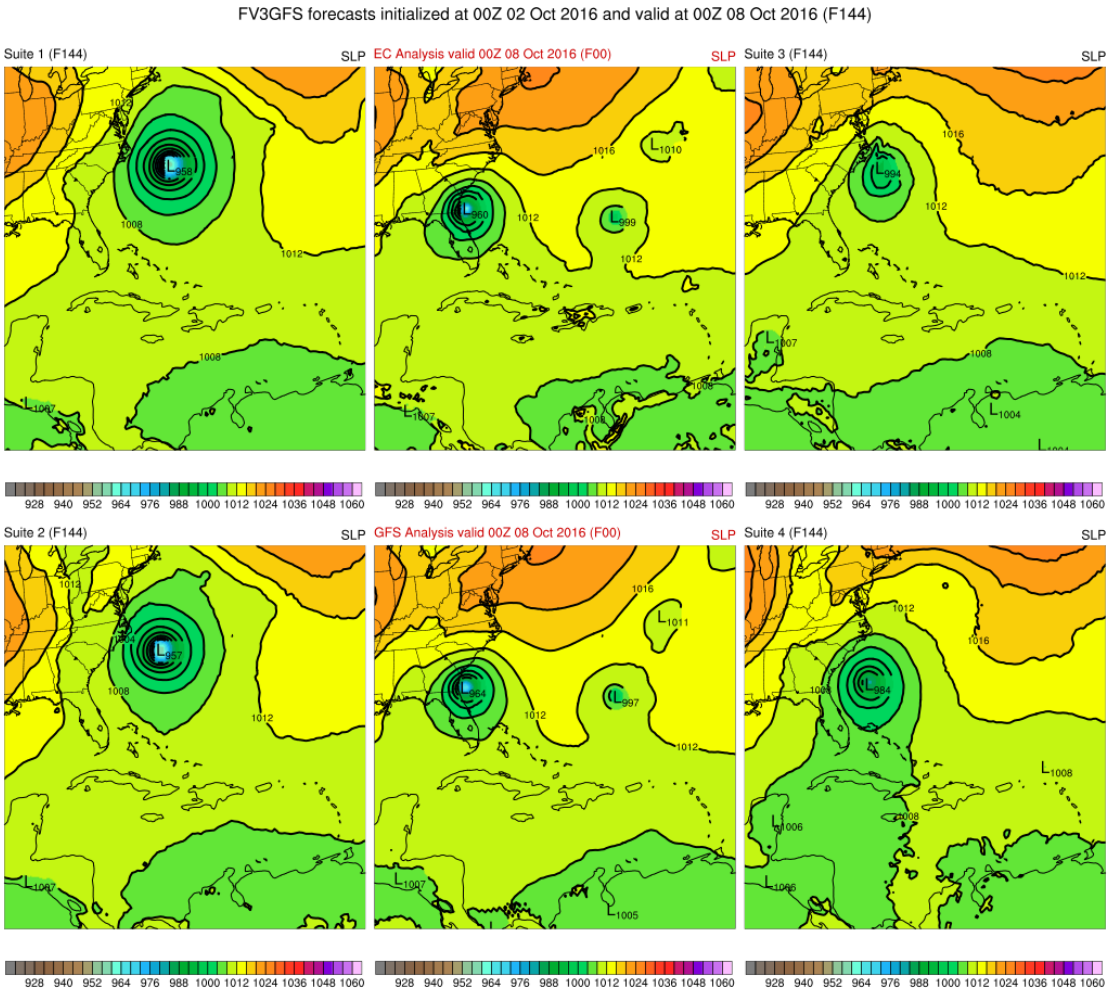


Fig. 18. 144-h forecasts of mean sea level pressure during TC Matthew (left and right columns). Forecasts were initialized at 0000 UTC 2 October 2016 and are valid at 0000 UTC 8 October 2016. The ECMWF analysis (top middle) and the GFS analysis (bottom middle) valid at 0000 UTC 8 October 2016 are also shown.

TC Noru (2017)

TC Noru was an intense super typhoon that had a long, winding track ending with a landfall in Japan. All FV3GFS forecasts from Suites 1-4 did a good job at predicting reasonable central pressures unlike the operational GFS, which continually forecasted Noru's central pressure to fall below 900 mb. Overall, for the cycle of interest, Suite 4 had the weakest forecast solution with Suites 1-3 predicting similar central pressures (Fig. 19). With the complexities of Noru's observed track, none of the suites did a great job at predicting the timing of landfall in Japan.

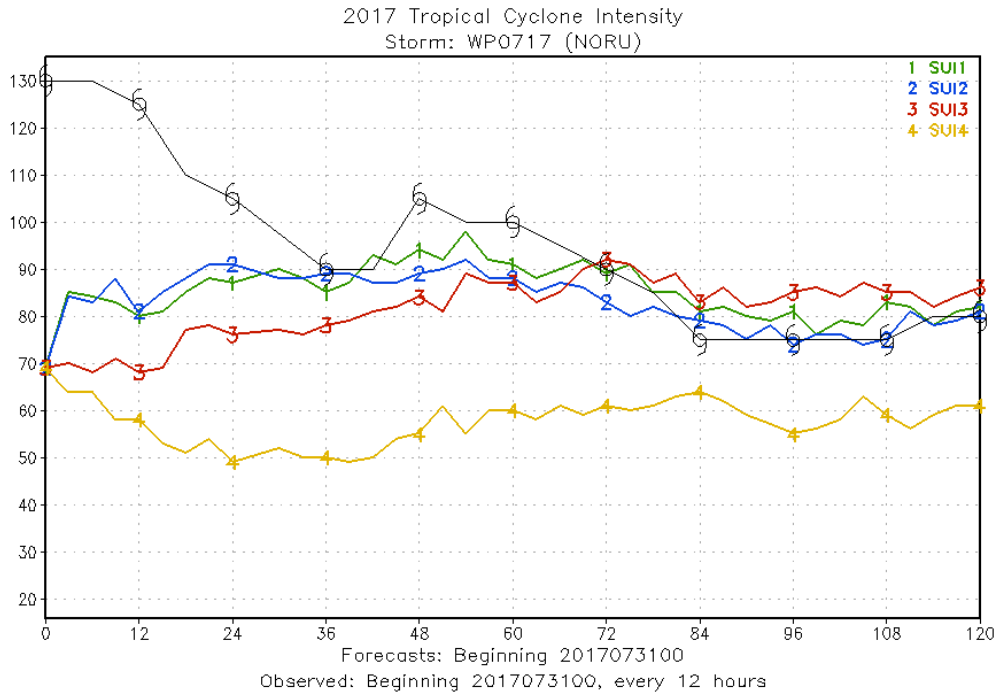


Fig. 19. Maximum wind speed forecasts for TC Noru from the 0000 UTC 31 July 2017 forecast cycle

TC Harvey (2017)

The available forecast cycle was initialized just as TC Harvey was making landfall along the Texas Gulf Coast. None of the suites did well with forecasting Harvey's loop back into the Gulf of Mexico and the subsequent ejection to the northeast. Suite 1, 3, and 4 all forecasted westward movement after stalling and/or looping in southern Texas. With none of the suites forecasting Harvey's complex track very well, the heaviest precipitation amounts were not correctly located near the Houston metropolitan area (Fig. 20). Like the operational global models, all four suites placed the heaviest precipitation close to the center of Harvey's circulation. Suites 2 and 4 were the best at capturing heavier precipitation further northeast along the Gulf Coast at various times during the forecast period.

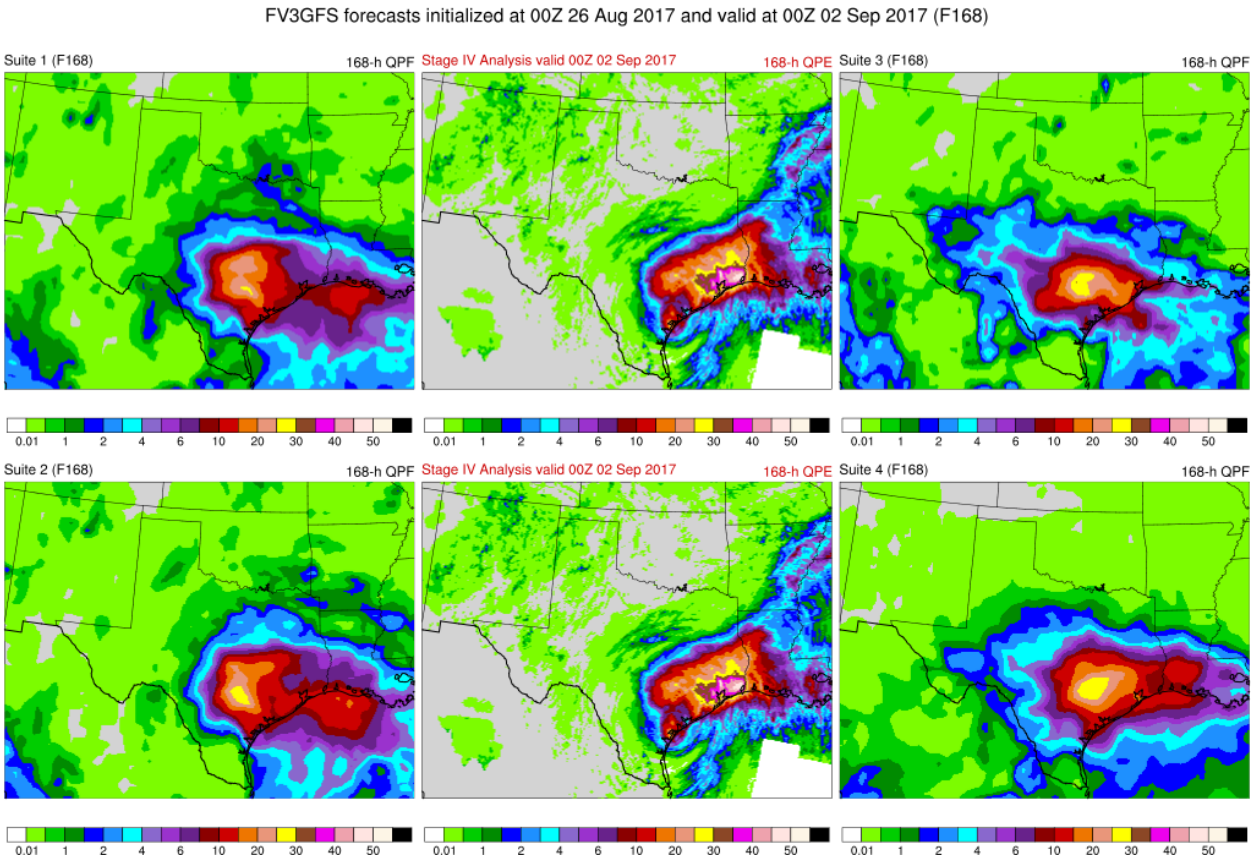


Fig. 20. 168-h forecasts of total accumulated precipitation during TC Harvey (left and right columns). Forecasts were initialized at 0000 UTC 26 August 2017 and are valid at 0000 UTC 2 September 2017. The 168-h Stage IV quantitative precipitation estimate analysis (top middle and bottom middle) valid at 0000 UTC 2 September 2017 is also shown.

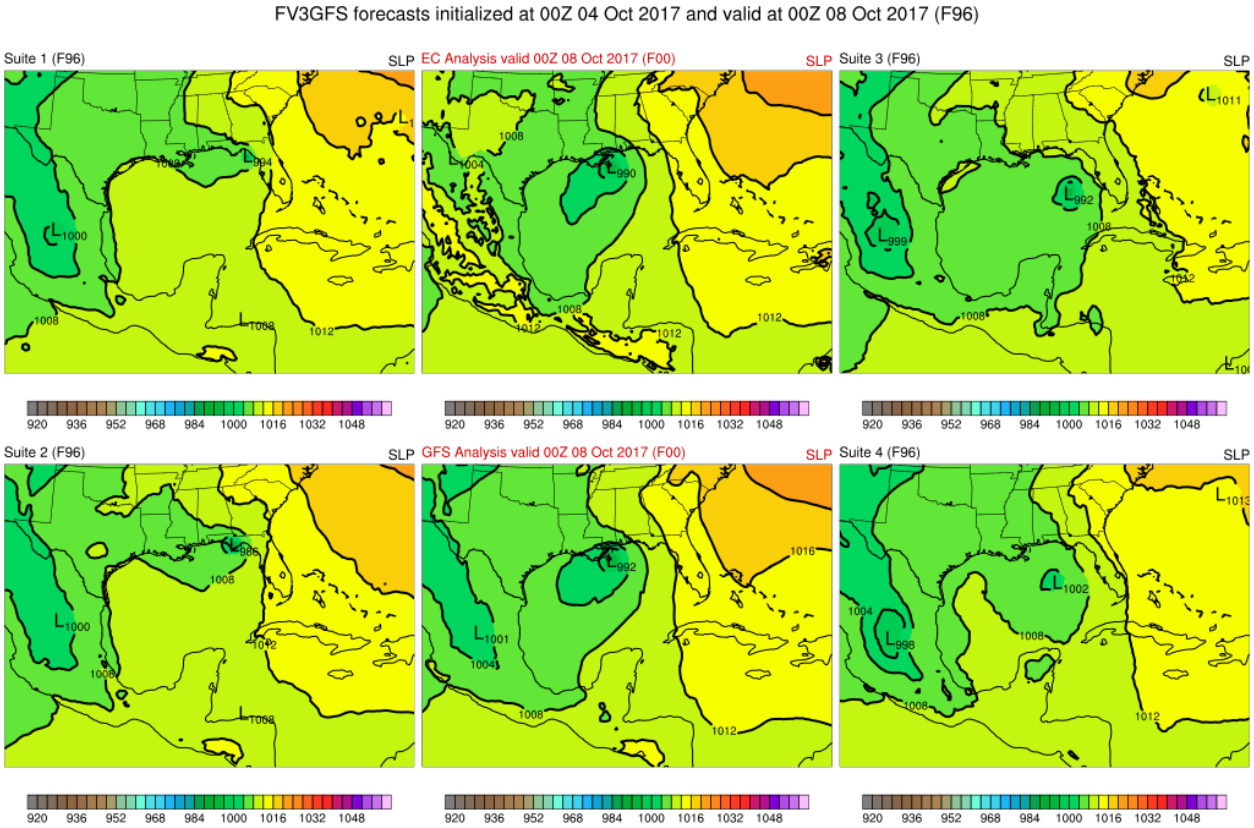
TC Irma (2017)

The major forecast challenge with TC Irma was the gradual shift with early forecast tracks showing landfall along Florida's Atlantic coast and later forecast tracks showing landfall on the west side of the Florida Peninsula. For the available cycle, all four suites had fairly good track forecasts through the first three days. Suites 1 and 4 showed the north turn toward Florida too soon and had a right-of-track bias at landfall and along the Florida Peninsula. Conversely, Suites 2 and 3 turned north too slowly and had a left-of-track bias at and after landfall. The most notable result was the extremely weak intensity predicted by Suite 4. This will be discussed further in the summary of intensity forecast performance.

TC Nate (2017)

Forecasting the formation and evolution of TC Nate was challenging for both the operational GFS and retrospective FV3GFS guidance run during the official evaluation of GFSv15. The retrospective forecasts were found to track Nate northward too quickly through the Gulf of Mexico. This fast latitude gain was also seen in the forecasts from Suites 1 and 2, which had physics settings nearly identical to those used for GFSv15. Although Suites 1 and 2 appear

to have the correct landfall time, the fast latitude gain caused Nate to make landfall too far east along the Florida Panhandle (Fig. 21). On the other hand, Suites 3 and 4 had forecasts that tracked Nate northward too slowly. Moreover, TC Nate actually never made landfall in the Suite 3 forecast; the tropical cyclone stagnated in the eastern Gulf of Mexico and dissipated over time.



]Fig. 21. 96-h forecasts of mean sea level pressure during TC Nate (left and right columns). Forecasts were initialized at 0000 UTC 4 October 2017 and are valid at 0000 UTC 8 October 2017. The ECMWF analysis (top middle) and the GFS analysis (bottom middle) valid at 0000 UTC 8 October 2017 are also shown.

TC Lane (2018)

Five-day forecasts from Suites 1, 2, and 4 all provided good guidance for TC Lane's track, including the beginning of the northward turn toward the Hawaiian Islands (Fig. 22). Suite 3's forecast was an example of an incorrect outlier solution with its forecast showing Lane continuing a WNW motion that would not have impacted Hawaii. Forecasts from all four suites were too weak compared to Lane's observed strength, but Suites 1 and 2 had decidedly stronger maximum winds and lower minimum pressures compared to Suites 3 and 4.

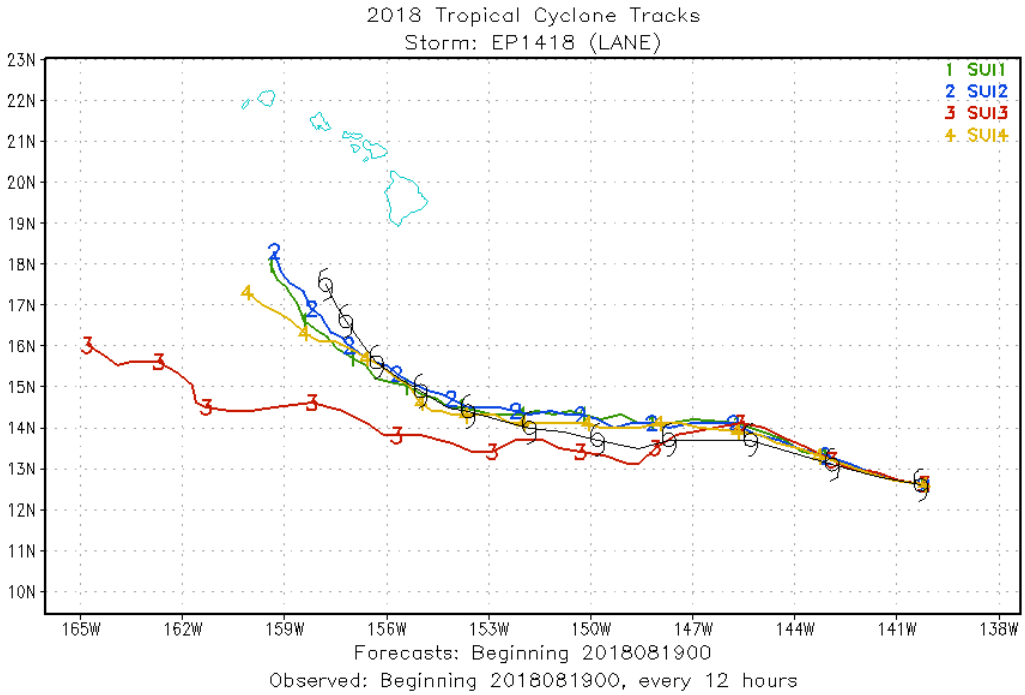


Fig. 22. Track forecasts for TC Lane from the 0000 UTC 19 August 2018 forecast cycle.

TC Florence (2018)

TC Florence's slow motion through the Carolinas was a major forecast challenge given the stagnating large-scale steering flow and the scale of the heavy precipitation across the region. Differing from many of the other cases, Suites 3 and 4 actually produced forecast solutions with tracks closest to Florence's observed path (Fig. 23). Forecasts from Suites 1 and 2 tracked Florence too far south well into Georgia. Because of these differences in the position of the low at Day 5, the remnants of Florence also tracked differently. In Suites 1 and 2, the remnants remained relatively stationary in southern Georgia before dissipating over time. With more correct forecast positions in the Suite 3 and 4 forecasts, Florence's remnants correctly tracked along the western spine of the Appalachian Mountains. In terms of intensity and precipitation, Suites 3 and 4 had weaker forecast solutions than Suites 1 and 2 at landfall. As such, Suites 3 and 4 also had lower precipitation maxima compared to the Suite 1 and Suite 2 forecasts. Suites 3 and 4, however, showed heavier precipitation moving northeast through the Appalachians because of the improved track forecasts beyond Day 5.

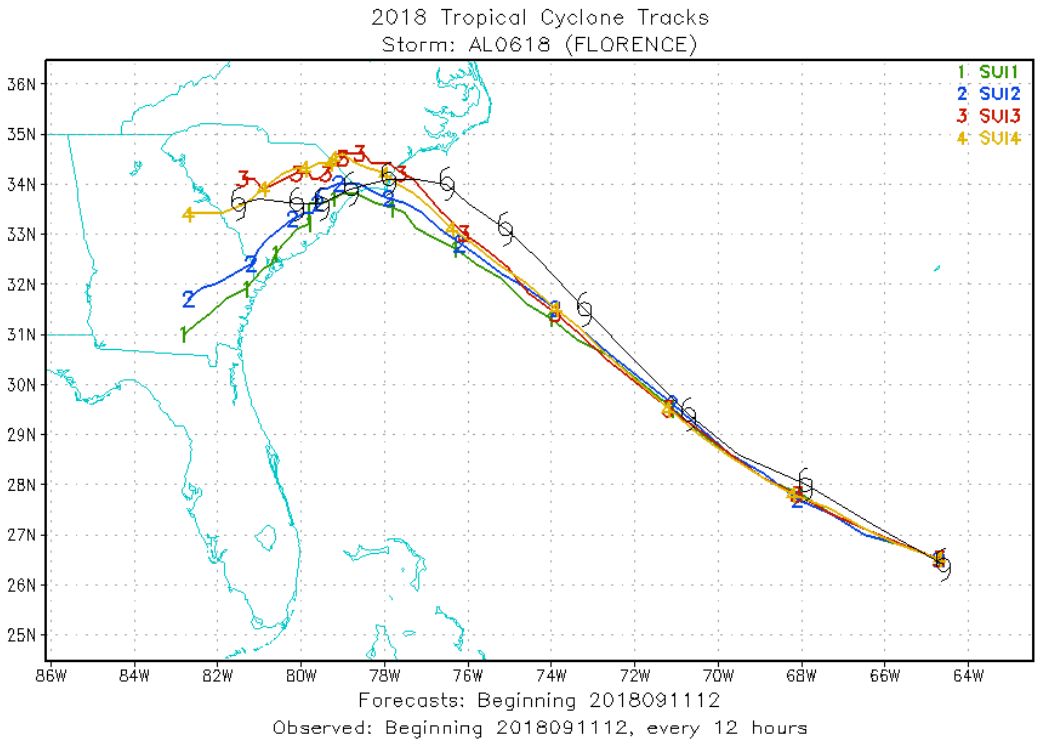


Fig. 23. Track forecasts for TC Florence from the 1200 UTC 11 September 2018 forecast cycle

Track Forecasts

For many cases, the track forecasts appeared fairly comparable between all four suites. Suites 1 and 2 were often nearly identical with only slight differences; there were generally no significant changes in the forecast solution when a forecast was generated with Suite 2 instead of Suite 1. Suites 3 and 4, however, appeared to have an increased likelihood of producing poor forecast solutions. The track forecasts overall were often quite good, but the incorrect outliers were limited to Suites 3 and 4. For example, in the case for TC Lane, Suite 3 produced a forecast showing the tropical cyclone continuing its WNW motion without a more distinct turn toward the Hawaiian Islands that was observed and forecasted by the other suites (Fig. 22). However, in the case of TC Noru, Suite 3 produced an outlier solution that appears to have lower track errors than the other suites. The most egregious track degradation in Suites 3 and 4 was noted in the case of TC Joaquin (Fig. 17).

The subjective evaluation of track forecasts is supported by objective verification statistics generated for all tropical cyclones in all basins during the forecasts generated for this evaluation. All four suites produced forecasts that were largely similar during the first two days of the forecast (Fig. 24). This is not surprising given the fact that all forecasts were initialized with the same ECMWF analyses. Through the next three days of the forecasts (i.e., Days 3-5), there were no statistical differences between the track forecasts from Suites 1 and 2. However, Suites 3 and 4's forecasts grew increasingly worse after Day 2. Suite 3 appeared to be the

worst performer at longer forecast lead times. The track degradation noted in Suites 3 and 4 was not statistically significant.

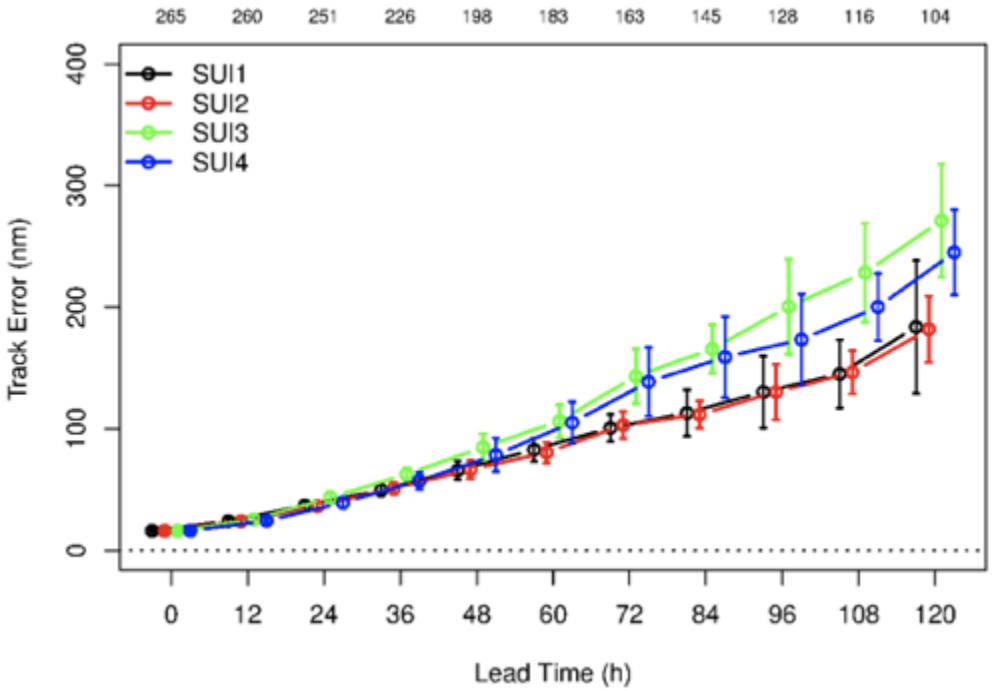


Fig. 24. Track errors (nm) for each suite averaged for all 163 tropical cyclones cases in all basins (courtesy: GMTB)

Intensity Forecasts

While there was not always large variability in the track forecast performance, there was a clear and consistent distinction between the quality of the intensity forecasts provided by Suites 1 and 2 as compared to those provided by Suites 3 and 4. Forecasts from all four suites had a weak intensity bias compared to observations (i.e., minimum pressures were not deep enough and maximum wind speeds were too low), but this result is not surprising given the resolution of the FV3GFS. However, the tropical cyclones simulated by Suites 3 and 4 were often markedly weaker than those simulated by Suites 1 and 2 (Fig. 25). This was reflected in the eight cases examined by the MEG and in bulk statistics generated for all tropical cyclones in all basins during the forecasts generated for this evaluation. Intensity verification statistics show that there was a statistically significant worsening of the weak intensity bias when forecasts were generated with Suites 3 and 4 while the weak intensity bias decreased with forecast lead-time in Suites 1 and 2.

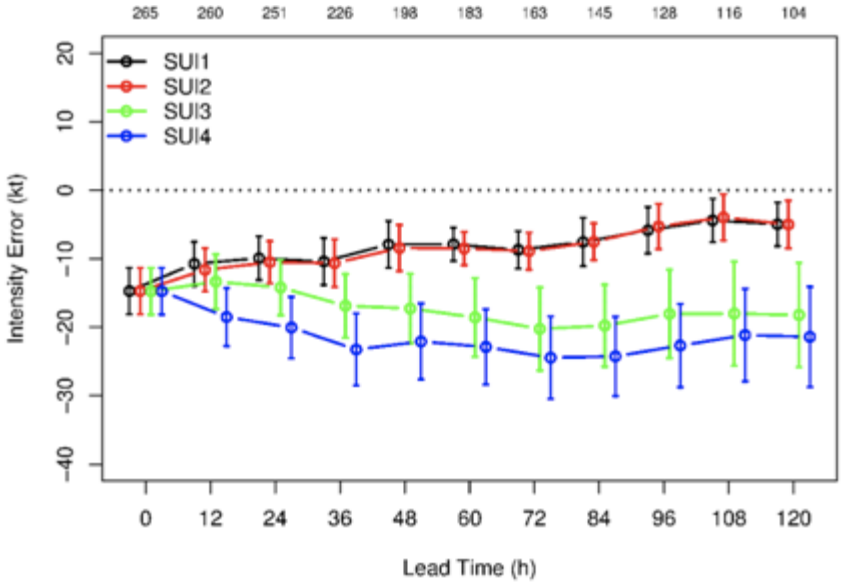


Fig. 25. Maximum wind speed errors (kt) for each suite averaged for all 163 tropical cyclones cases in all basins (courtesy: GMTB).

Although forecasts from all four suites were initialized with the same ECMWF analyses, it appears that tropical cyclones in Suite 3 and Suite 4 forecasts responded especially poorly during model spin-up. In those two suites, the initial tropical cyclone typically weakened during the first 6-12 hours of the forecast integration; this was seen in minimum pressure traces (not shown) and maximum wind speed traces (Fig. 26). This in part contributed to the weak intensity bias noted with Suites 3 and 4. In other cases like TC Matthew, Suites 3 and 4 simply demonstrated an inability to simulate strengthening that was observed and successfully predicted by Suites 1 and 2 (Fig. 27). These factors combined to result in Suite 3 and Suite 4 forecasts that were significantly weaker than the observed storms and the Suite 1 and Suite 2 forecasts. That said, the tropical cyclone intensity performance for each suite might be somewhat different in an experimental setup including fully-cycled data assimilation for each suite. Forecasts from all four suites clearly went through an adjustment period after being cold-started with tropical cyclones brought in from an external initial state, and it is possible that Suite 3 and 4 forecasts simply did not adjust as well as forecasts from Suites 1 and 2.

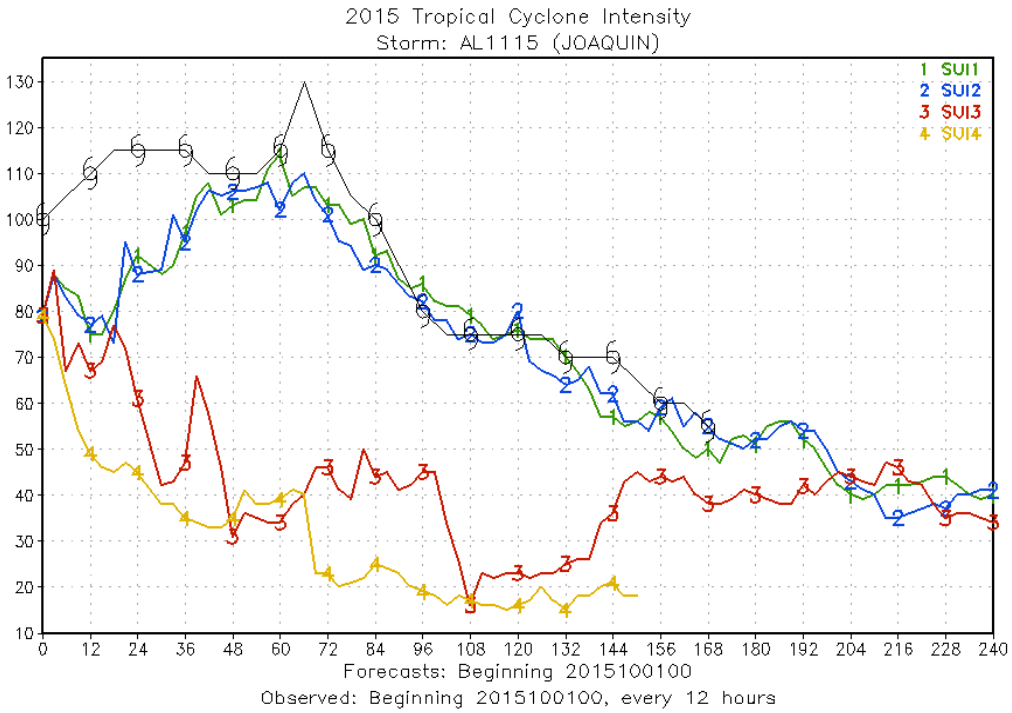


Fig. 26. Maximum wind speed forecasts for TC Joaquin from the 0000 UTC 1 October 2015 forecast cycle

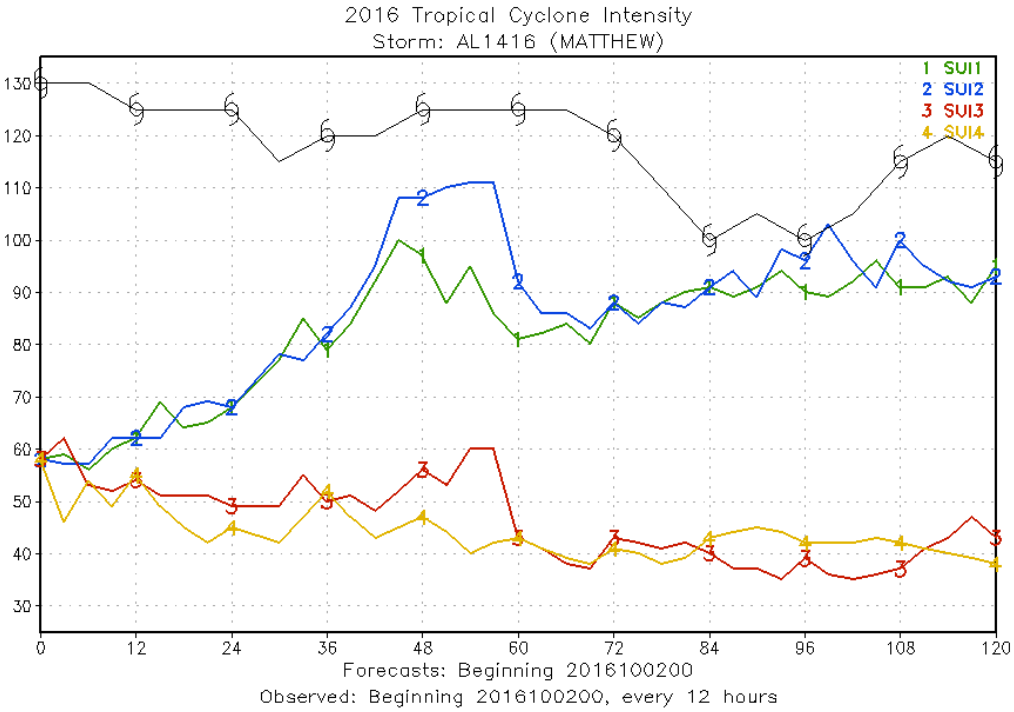


Fig. 27. Maximum wind speed forecasts for TC Matthew from the 0000 UTC 2 October 2016 forecast cycle

The TC Irma forecast cycle provided an excellent example of the especially degraded forecast performance in forecasts generated using Suite 3 or Suite 4. Based on global model analyses valid at 1200 UTC 10 September 2017, TC Irma had a minimum pressure around 940 hPa, which was fairly well forecasted by Suites 1, 2, and 3 (Fig. 28). However, Suite 4's forecast showed a 972 hPa minimum pressure at this valid time. Based on the associated maximum wind speed forecast (not shown), Suite 4 forecasted TC Irma to be a strong tropical cyclone at landfall, when in reality TC Irma made landfall in Florida as a major hurricane.

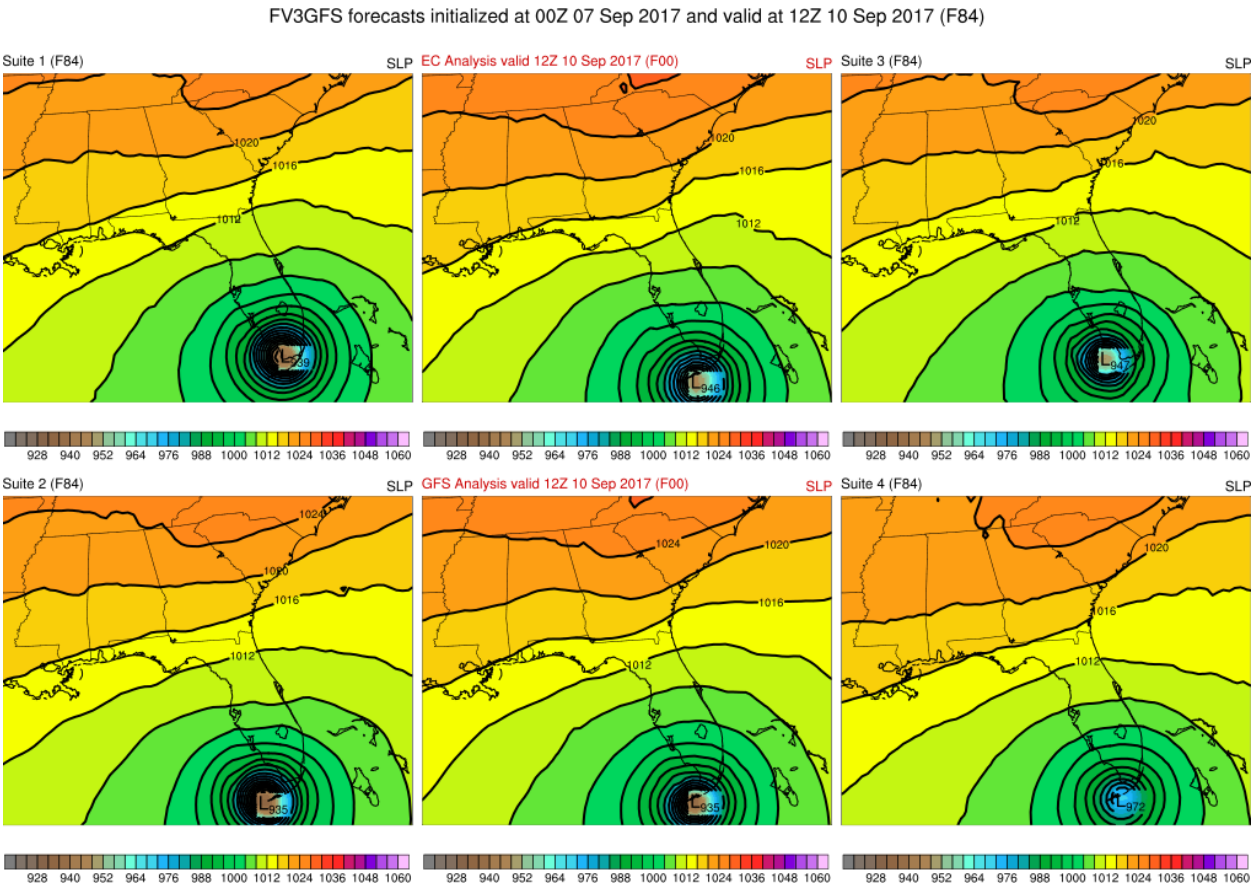


Fig. 28. 84-h forecasts of mean sea level pressure during TC Irma (left and right columns). Forecasts were initialized at 0000 UTC 7 September 2017 and are valid at 1200 UTC 10 September 2017. The ECMWF analysis (top middle) and the GFS analysis (bottom middle) valid at 1200 UTC 11 September 2017 are also shown.

Finally, it is worth noting that Suite 2 appeared to do a better job at representing stronger low-level wind speeds over landmasses. An example can be seen in the Suite 2 forecast for TC Irma (Fig. 29). Rapid weakening of surface winds over landmasses (but not over water) is a shortcoming that has been noted in forecasts from several NCEP models. Given that the addition of prognostic TKE to the PBL/turbulence parameterization scheme is the only change between Suite 1 and Suite 2, it appears that this change may produce more desirable low-level tropical cyclone winds over land surfaces in FV3GFS forecasts.

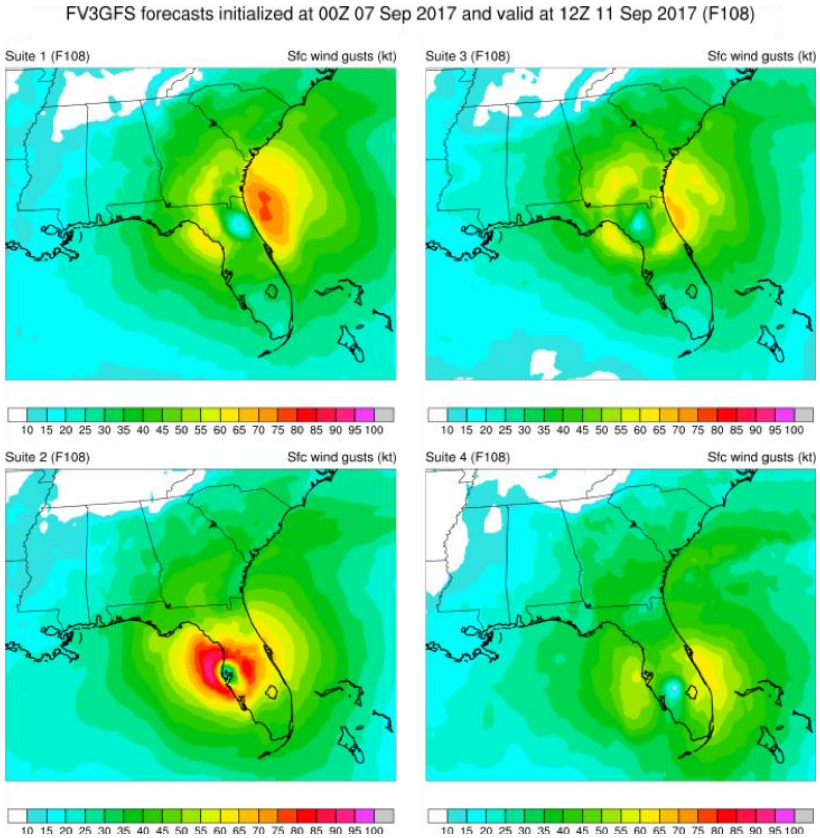


Fig. 29. 108-h forecasts of surface wind gusts during TC Irma. Forecasts were initialized at 0000 UTC 7 September 2017 and are valid at 1200 UTC 11 September 2017.

Quantitative Precipitation Forecasts

Not surprisingly, quantitative precipitation forecasts from these eight cases show that precipitation associated with tropical cyclones is closely tied to the skill of the track and intensity forecasts. In cases like TC Irma and TC Nate, the primary axis of QPF was displaced in forecasts where the there was a left- or right-of-track bias. Moreover, the weak intensity bias in Suites 3 and 4 was also reflected in generally lower accumulated precipitation amounts. When examining the total precipitation fields at extended forecast lengths, this low QPF bias was sometimes masked by relatively slower storm motions in the weaker forecast solutions (e.g., Suite 4 forecast for TC Irma). However, examination of the shorter 6-h and 24-h accumulation periods revealed that the weaker Suites 3 and 4 were unable to simulate the peak rainfall rates and accumulations that were observed and simulated by Suites 1 and 2. For example, a 24-h accumulated precipitation forecast associated with TC Florence is shown in Fig. 30, and Suites 3 and 4's peak accumulations are only in the 6-10" range while the observations and Suites 1 and 2 have peak amounts in the 10-15"+ range.

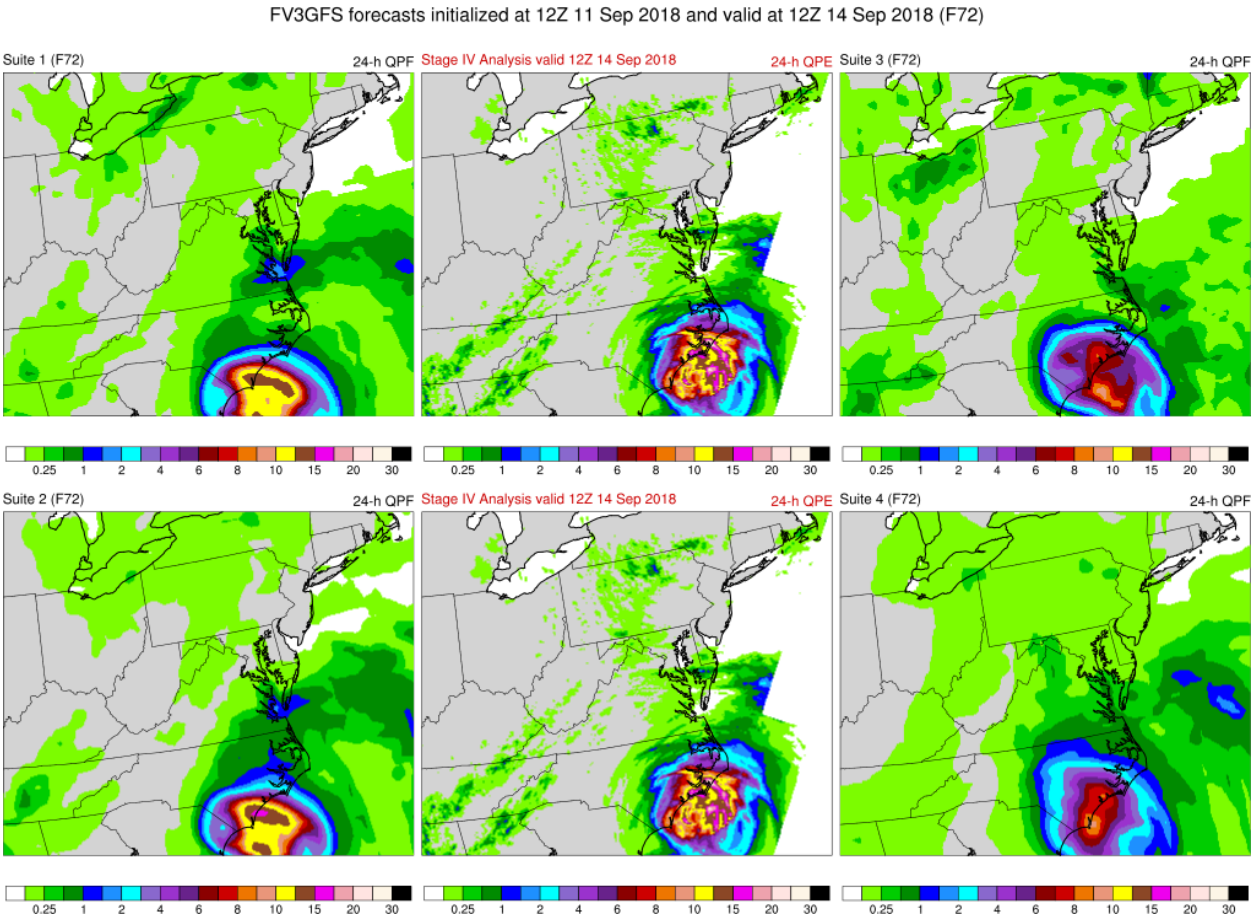


Fig. 30. 72-h forecasts of 24-h accumulated precipitation during TC Florence (left and right columns). Forecasts were initialized at 1200 UTC 11 September 2018 and are valid at 1200 UTC 14 September 2018. The 24-h Stage IV quantitative precipitation estimate analysis (top middle and bottom middle) valid at 1200 UTC 11 September 2018 is also shown.

Summary

By and large, Suites 1 and 2 clearly provide more useful guidance than Suites 3 and 4 for the eight tropical cyclones included in this evaluation. In some cases, track forecasts were somewhat similar for all four suites, but in cases like TC Joaquin, TC Nate, and TC Lane, track forecasts from Suite 3 and/or Suite 4 were notably degraded in comparison with track forecasts from Suites 1 and 2. Overall, objective verification statistics indicated that the poorer Suite 3 and Suite 4 track performance largely occurred in the Days 3–5 time range.

The performance distinction between Suites 1 and 2 and Suites 3 and 4 was even more drastic when assessing intensity forecasts. Forecasts from Suites 3 and 4 consistently forecasted tropical cyclones that were weaker than observations and the forecasts from Suites 1 and 2. This was reflected in both the maximum wind speed forecast and the minimum pressure forecasts. The weak intensity bias was attributed to rapid weakening in the first 12 hours of the forecast in some cases, and in others, the bias was owed to an apparent inability of these suites

to simulate substantial strengthening during the forecast integration. This overarching result is somewhat clouded by the fact that, with all four suites, the model went through an adjustment period after being cold-started with an external initial state. In a setup with fully-cycled data assimilation for each suite, the initialized tropical cyclones might behave somewhat differently in the ensuing forecasts.

Finally, performance of quantitative precipitation forecasts associated with these eight tropical cyclones was very clearly tied to the quality of the track and intensity guidance for each storm. When the track forecasts were degraded, the accumulated precipitation fields were degraded as well. Most notably, the weak intensity bias noted in Suites 3 and 4 often resulted in a low QPF bias, largely due to the weaker tropical cyclones resulting in lower rainfall rates.

Overall, forecasts from Suite 1 and Suite 2 were largely similar, and Suite 2 appeared to provide some marginal improvements over Suite 1 in these specific cases. Based on the subjective evaluation of these eight tropical cyclone forecasts and the bulk track and intensity statistics, Suites 3 and 4 would require further development and testing before they would be capable of providing tropical cyclone guidance that is comparable to that provided by GFSv15.

SOUNDINGS

The operational GFS has long had several significant biases with regards to vertical profiles; these have been discussed in many of the weekly webinars conducted by the MEG. The most serious issue concerns the model's ability to correctly handle inversions, with too much mixing resulting in temperatures that are too warm at the surface and too cold just above the ground. This primarily involves late night radiation inversions, with the GFS often too warm with 2m temperature forecasts by amounts over 6 C in very short-range forecasts, but it also occurs in warm advection precipitation type events in which a "warm" layer is established above a cold surface, leading to sleet or freezing rain. With its muted inversions, the GFS often significantly underforecasts sleet and freezing rain. Another issue concerns instability, with the GFS tending to predict lapse rates that are too weak, leading to forecasts of convective available potential energy (CAPE) that are too low and forecasts of convective inhibition (CIN) that are too high. Another contributing factor to low CAPE / high CIN forecasts is an occasional tendency to overmix the boundary layer during peak heating, leading to low levels that are too dry (along with too hot). This problem has been mitigated with some model changes in recent years but has not been eliminated.

To evaluate sounding structure in this testing, forecast soundings for each suite were plotted for each case at all RAOB sites across the United States every 6 hours out to forecast hour 144 and scrutinized. The forecast soundings were compared to RAOB values whenever they were available (usually 0000 and 1200 UTC). It has been the MEG's general impression that the current GFSv15 running in parallel continues the same biases with vertical profiles shown in the GFS, so we expected to see similar problems in Suite 1 but hoped to see signs of improvement in Suites 2-4. An effort was made to focus on examples for which the synoptic

details of the forecast were generally handled well so that the true influence of the physics suites on the profiles can be assessed. The general findings are summarized in this section.

The most common finding with regard to the handling of the strength of nocturnal radiation inversions is that Suite 1 was consistently too weak, Suite 3 was a modest improvement over Suite1, Suite 2 was a more significant improvement over Suite 1, and Suite 4 was consistently the best at showing stronger inversions. Representative examples are shown in Figures 31-33.

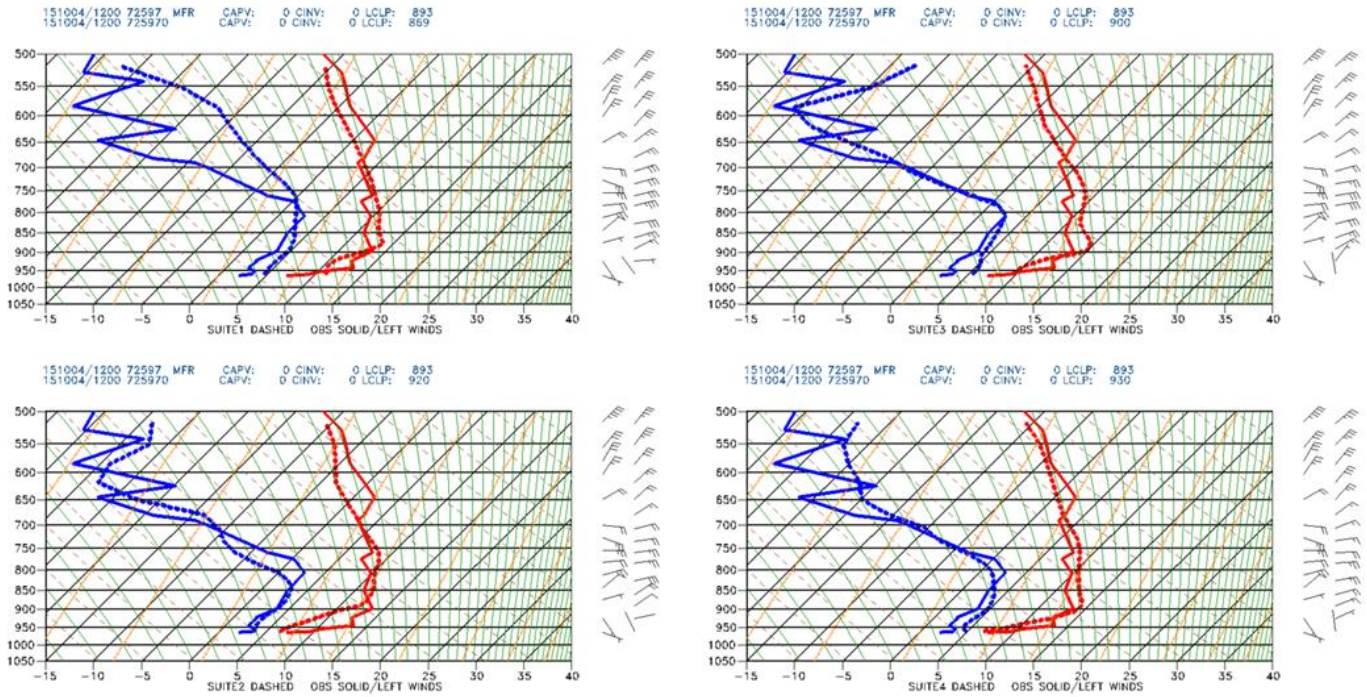


Fig 31. 84-hr forecasts for Medford, OR (MFR) valid 1200 UTC 4 October 2015. Temperatures are in red, and dew points are in blue. The solid line represents the observed RAOB data, while the dashed line is the model forecast. The left column winds and first line of index values at the top are from the RAOB data, while the right column winds and second line of index values are the model forecasts. The upper left panel shows the forecast from Suite 1, the lower left panel shows the forecast from Suite 2, the upper right panel shows the forecast from Suite 3, and the lower right panel shows the forecast from Suite 4. The organization of this plot is used for all images in this section, except for Figs. 43 and 44.

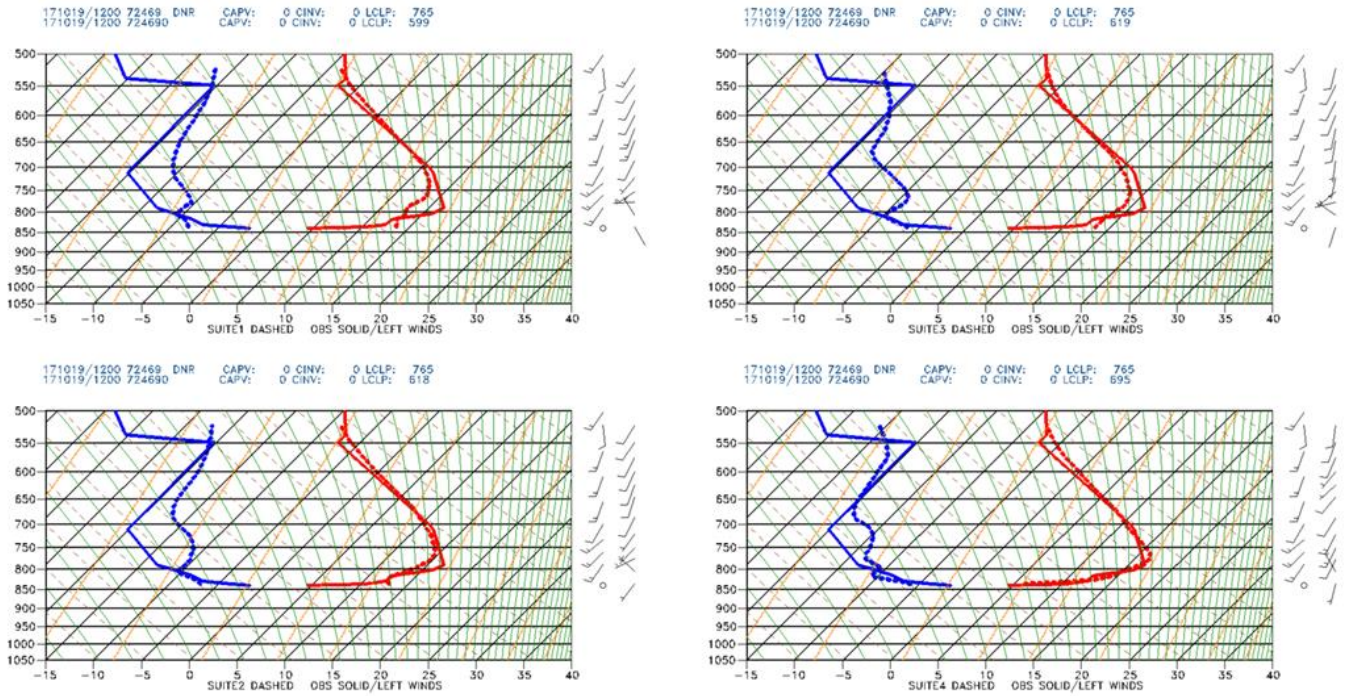


Fig. 32. 72-hr forecasts for Denver, Colorado valid 1200 UTC 19 October 2017.

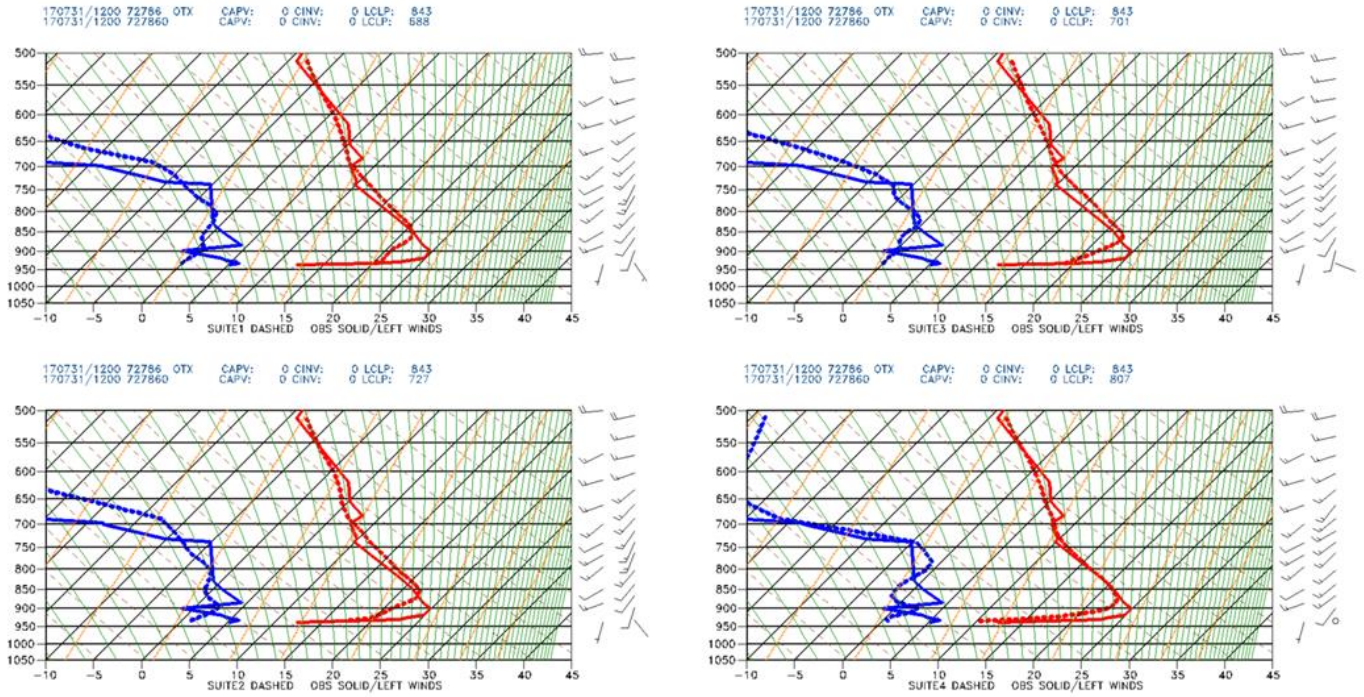


Fig. 33. 60-hr forecasts for Spokane, WA valid 1200 UTC 31 July 2017.

While Suite 4 is overall the best at capturing low-level inversion structure, there are examples in which it actually overstrengthens the low-level inversion, making the lowest level temperatures too cold. In some of these examples, are still too weak the inversion strength

(Fig. 34), but there are some instances in which Suite 4 is too strong with the inversion, but the other suites perform well. (Fig. 35). Finally, there were also examples in which Suite 4 did extremely well with cooling the lowest model level in an inversion but was way too warm at the top of the inversion (Figs. 36 and 37).

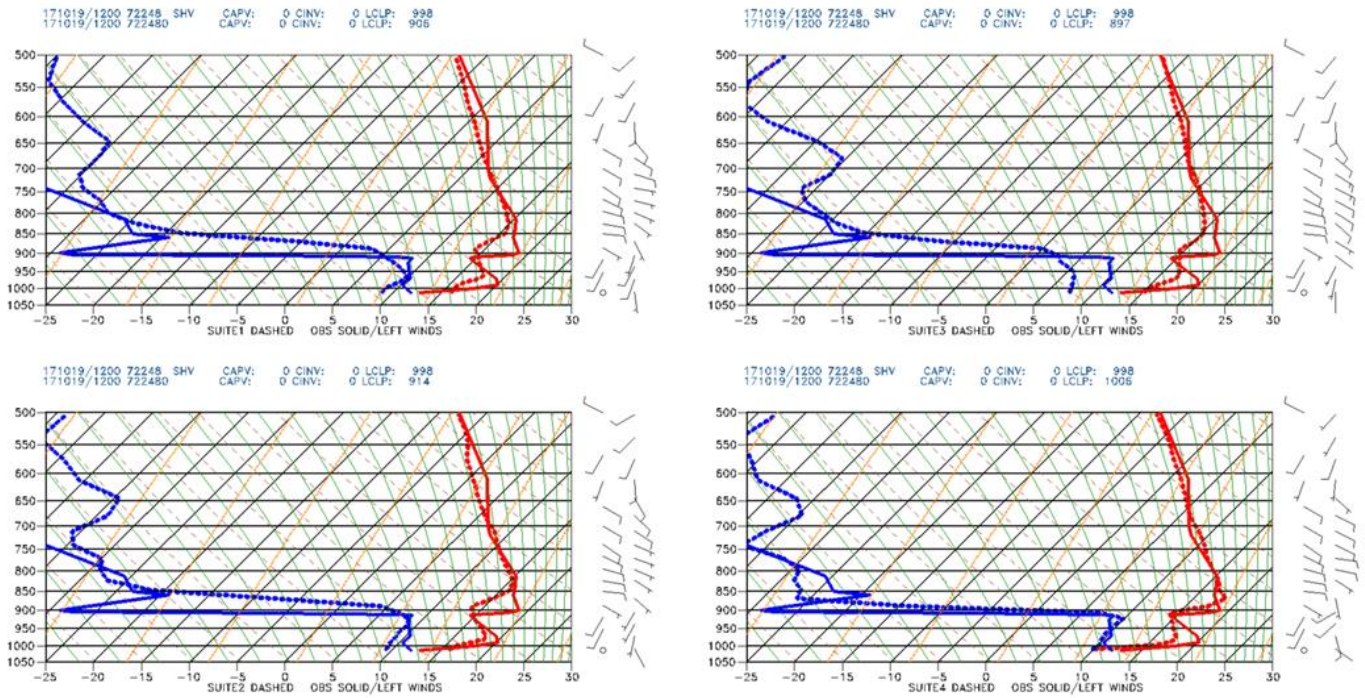


Fig. 34. 72-hr forecasts for Shreveport, LA valid 1200 UTC 19 October 2017.

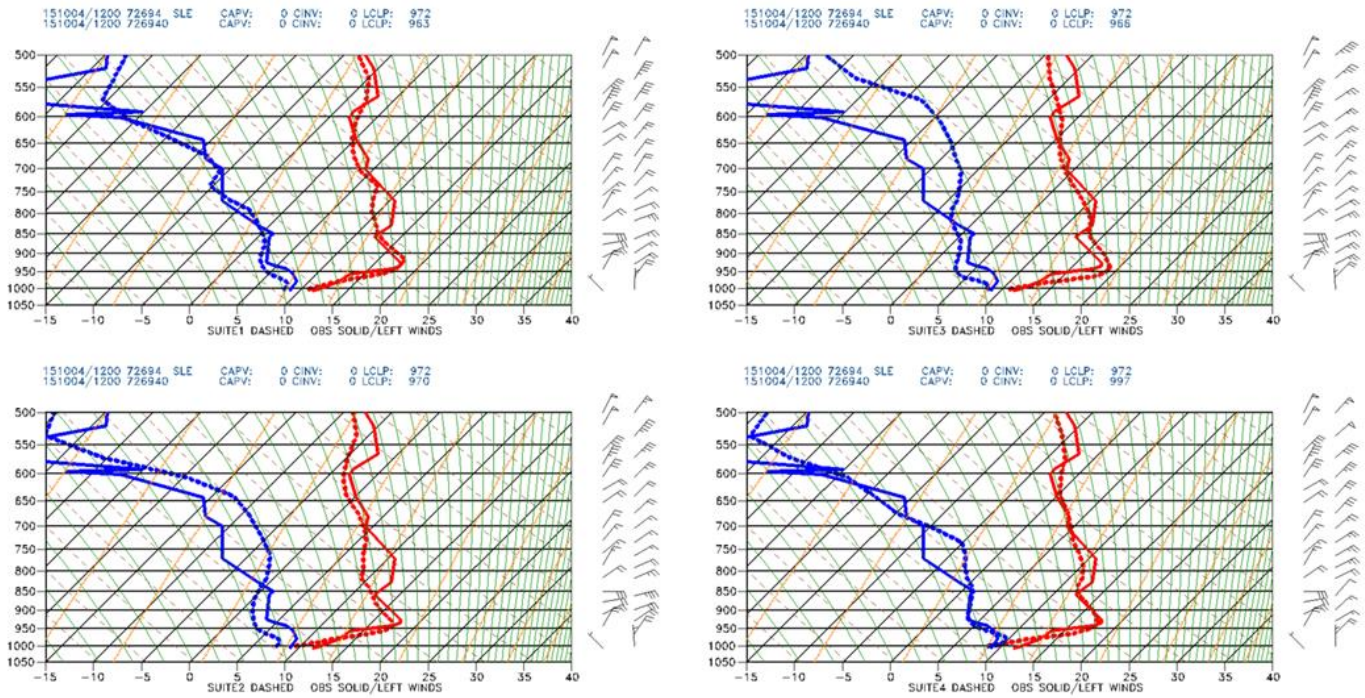


Fig. 35. 84-hr forecasts for Salem, OR valid 1200 UTC 4 October 2015.

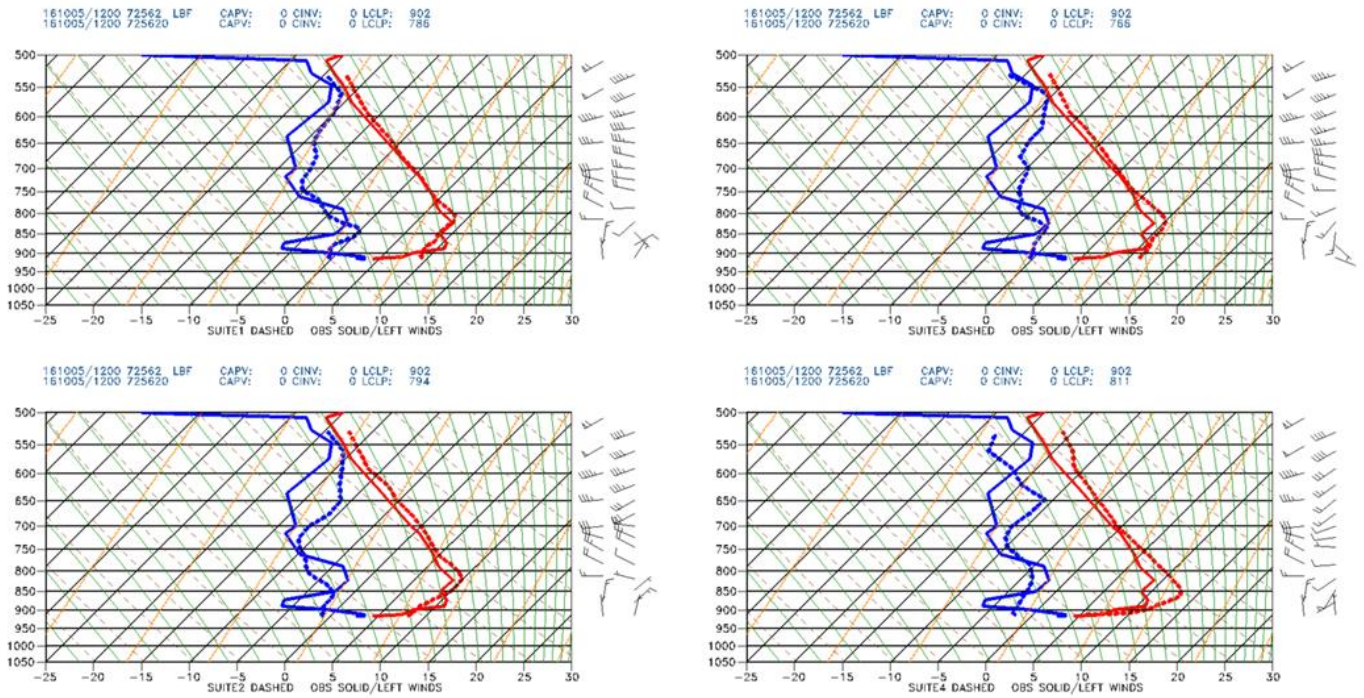


Fig. 36. 84-hr forecasts for North Platte, NE valid 1200 UTC 5 October 2016.

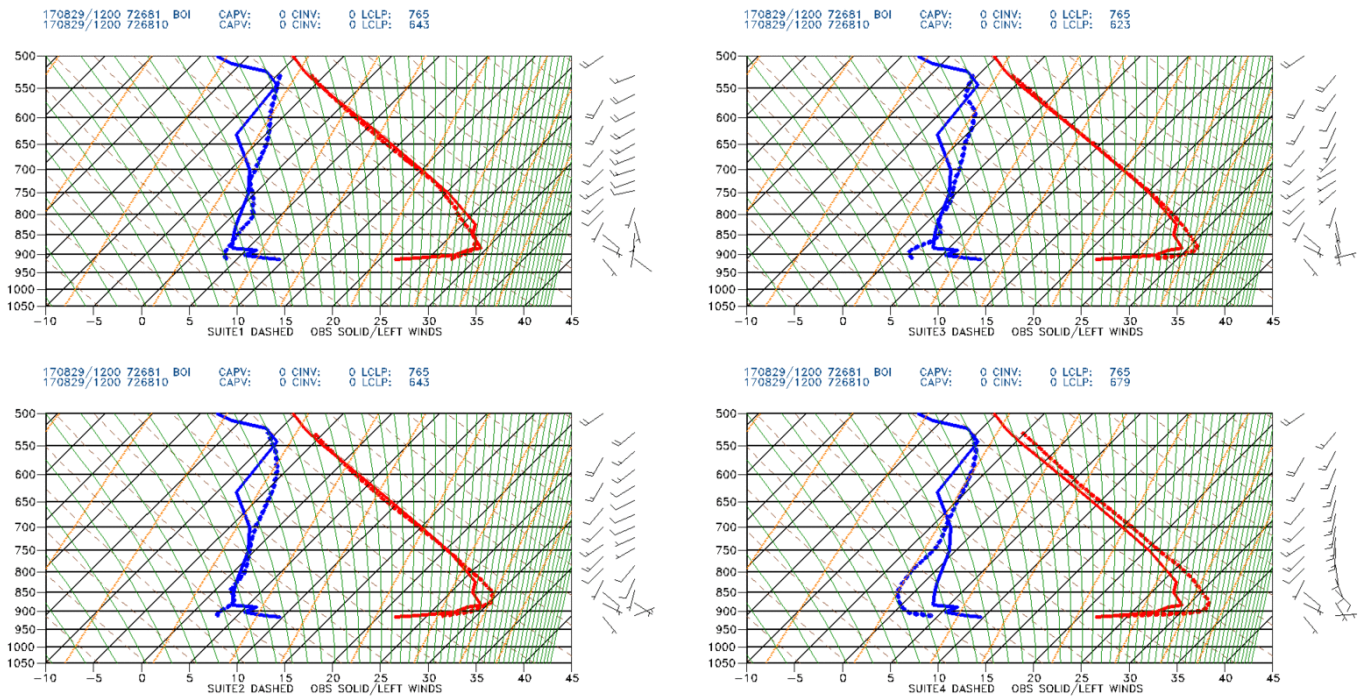


Fig. 37. 84-hr forecasts for Boise, ID valid 1200 UTC 29 August 2017.

Switching from early morning inversions to late day mixed PBLs, Suite 4 overall tended to be overly-mixed and therefore too hot and too dry in the boundary layer. Figures 38 and 39 show examples in which Suite 4 has a very deep PBL, leading to forecasts at the surface that are too hot and too dry and clearly inferior to the forecasts from the other three suites.

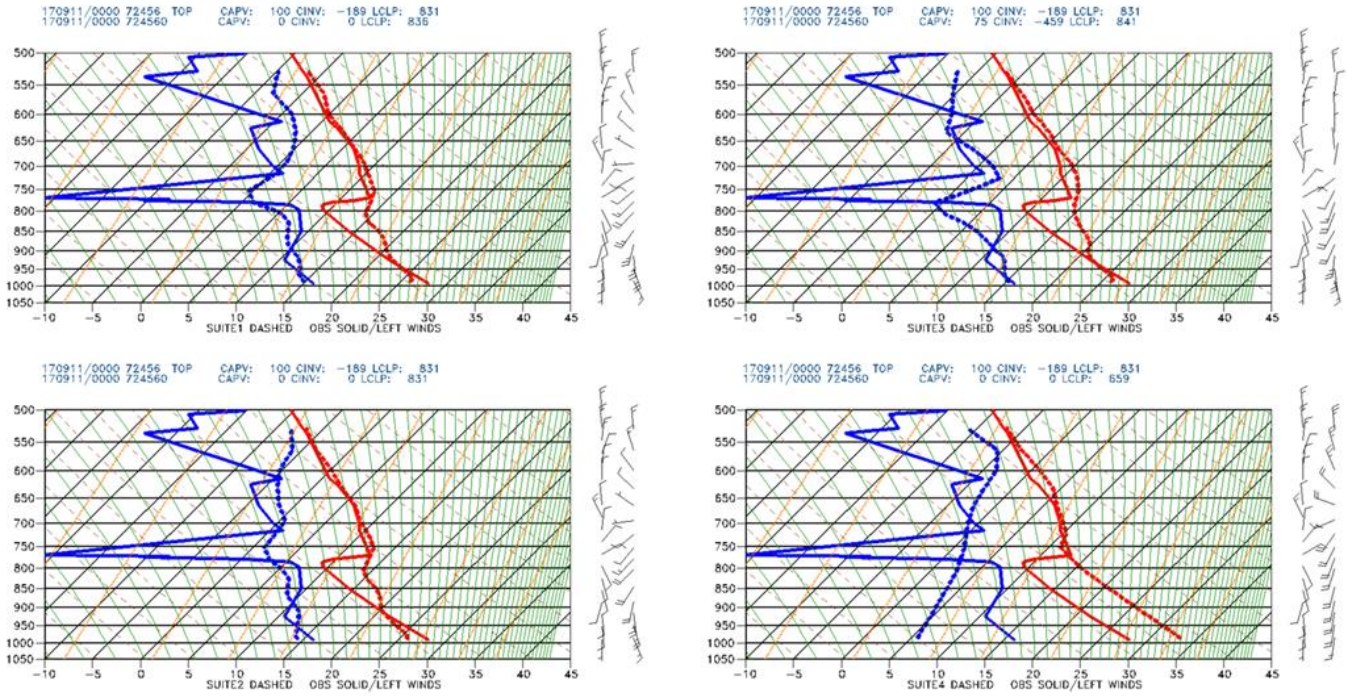


Fig. 38. 96-hr forecasts for Topeka, KS valid 0000 UTC 11 September 2017.

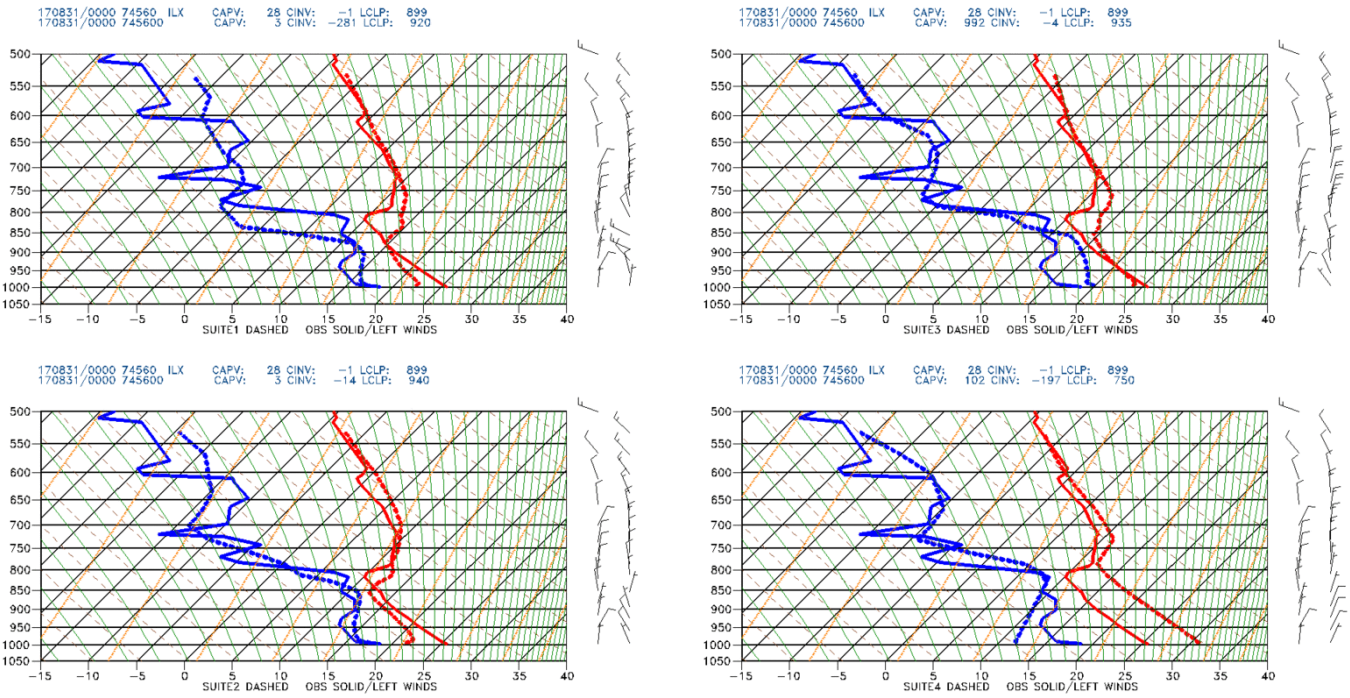


Fig. 39. 120-hr forecasts for Lincoln, IL valid 0000 UTC 31 August 2017.

It is also worth noting that while all of the Suites struggle to maintain capping inversions in pre-convective environments, Suite 4 seems prone to mixing them out the most aggressively. A 6-hr evolution is displayed in Figures 40 and 41; Fig. 40 shows forecasts valid at an 1800 UTC valid time. While we have no observations at that time, Suites 1-3 are generally similar with overall structure, with a PBL of modest depth and a fairly stout capping inversion above. But Suite 4 has clearly already removed most of the capping inversion (in addition to drying the boundary layer). At 0000 (Fig. 41), Suite 4 has completely eliminated the inversion and has a deep mixed layer and a boundary layer that is too hot and too dry. Suite 3 has a bit of an odd temperature profile, as it seems to have cooled the top of the inversion and warmed just below.. Suites 1 and 2 do the best job of maintaining an inversion.

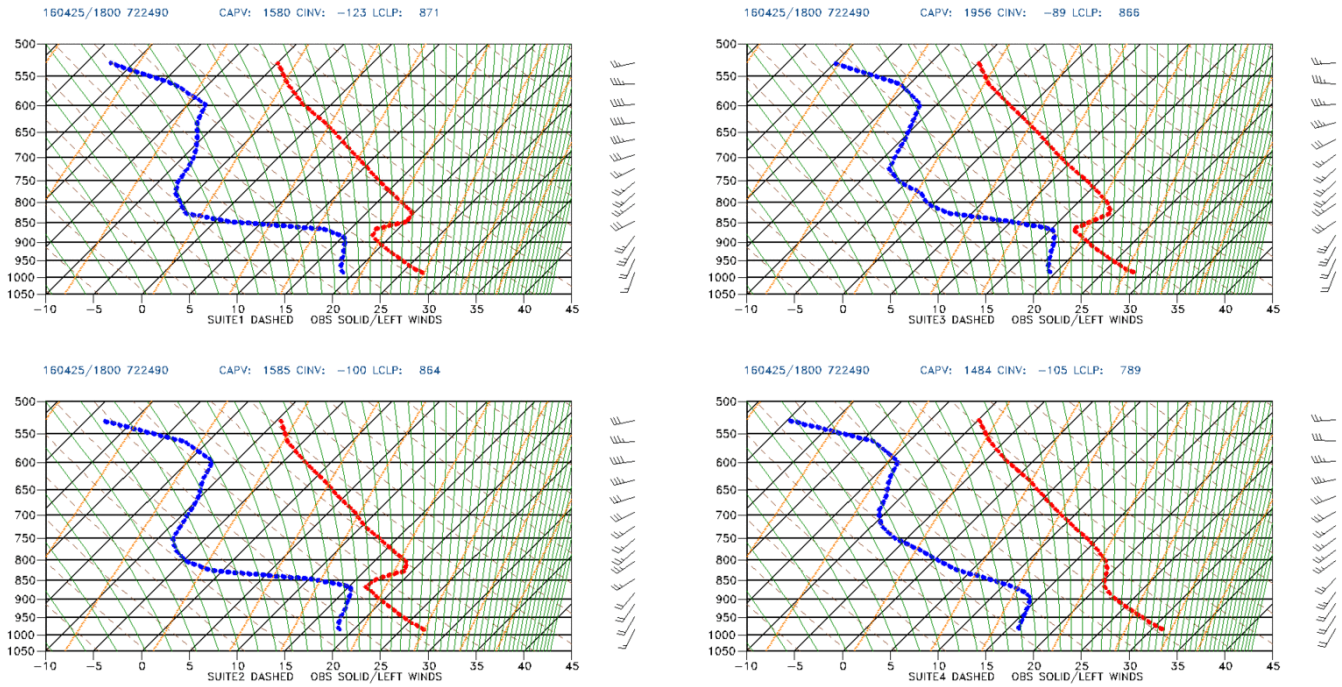


Fig. 40. 90-hr forecast for Fort Worth, TX valid 1800 UTC 25 April 2018. RAOB data is not available at this time.

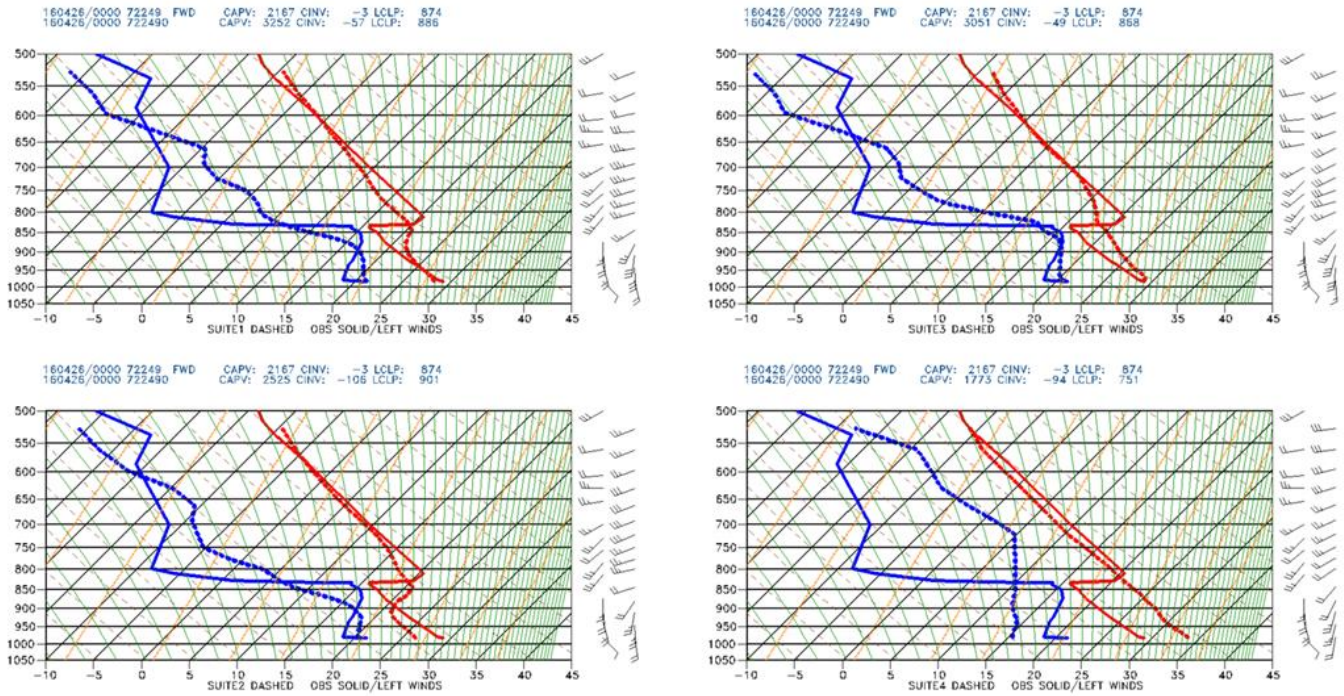


Fig. 41. 96-hr forecast for Fort Worth, TX valid 0000 UTC 26 April 2018.

As shown in the statistics section, Suite 2 has a noticeable cold bias at 0000 UTC valid times. A representative example is shown in Fig. 42. The observed sounding remains well-mixed through the lower levels at 0000 UTC, but all suites have started to decouple the lowest levels. The effect, however, is greatest in Suite 2. The Suite 2 cold bias appears throughout the day, and no clear cause was evident in either the soundings or the maps from the individual cases.

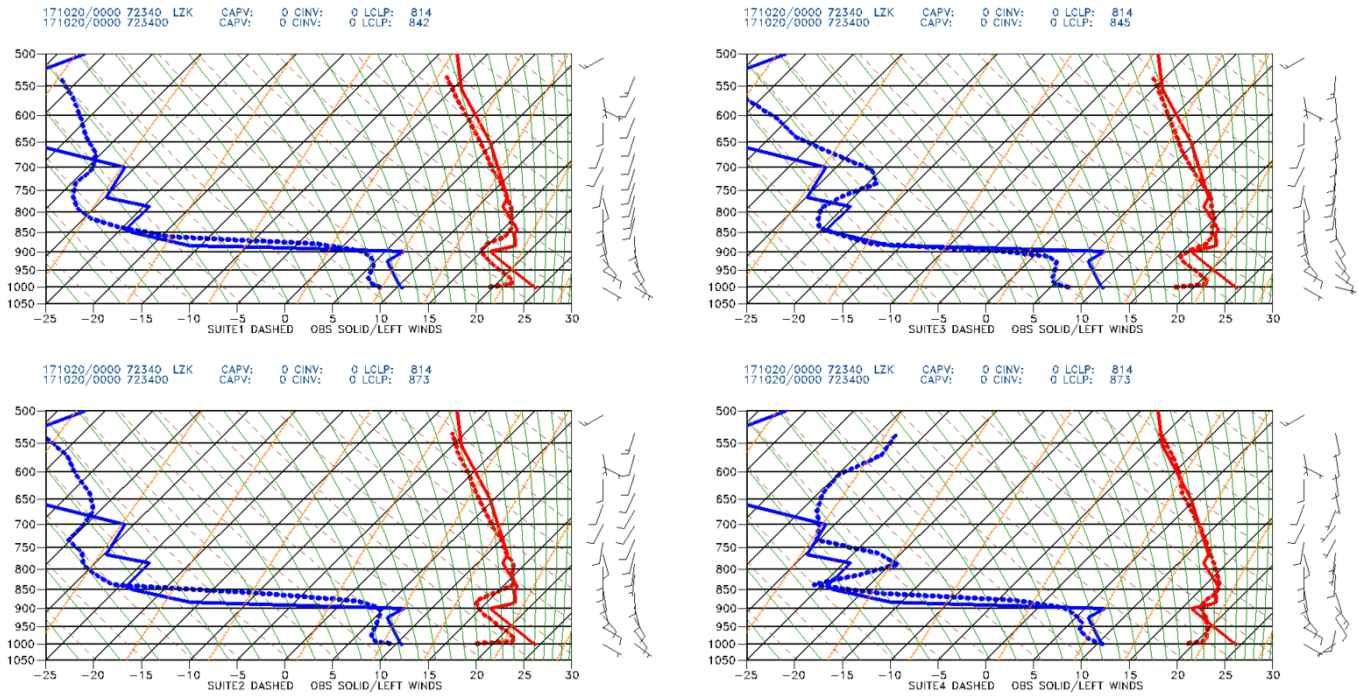


Fig. 42. 84-hr forecast for Little Rock, AR valid 0000 UTC 20 October 2017.

Suite 3 occasionally showed some very unrealistic structures. In Fig. 43, Suite 3 shows an odd mid-level inversion with a superadiabatic layer above. While we do not have verification at this time, it is fairly safe to assume that the structure in Suite 3 would not be possible.

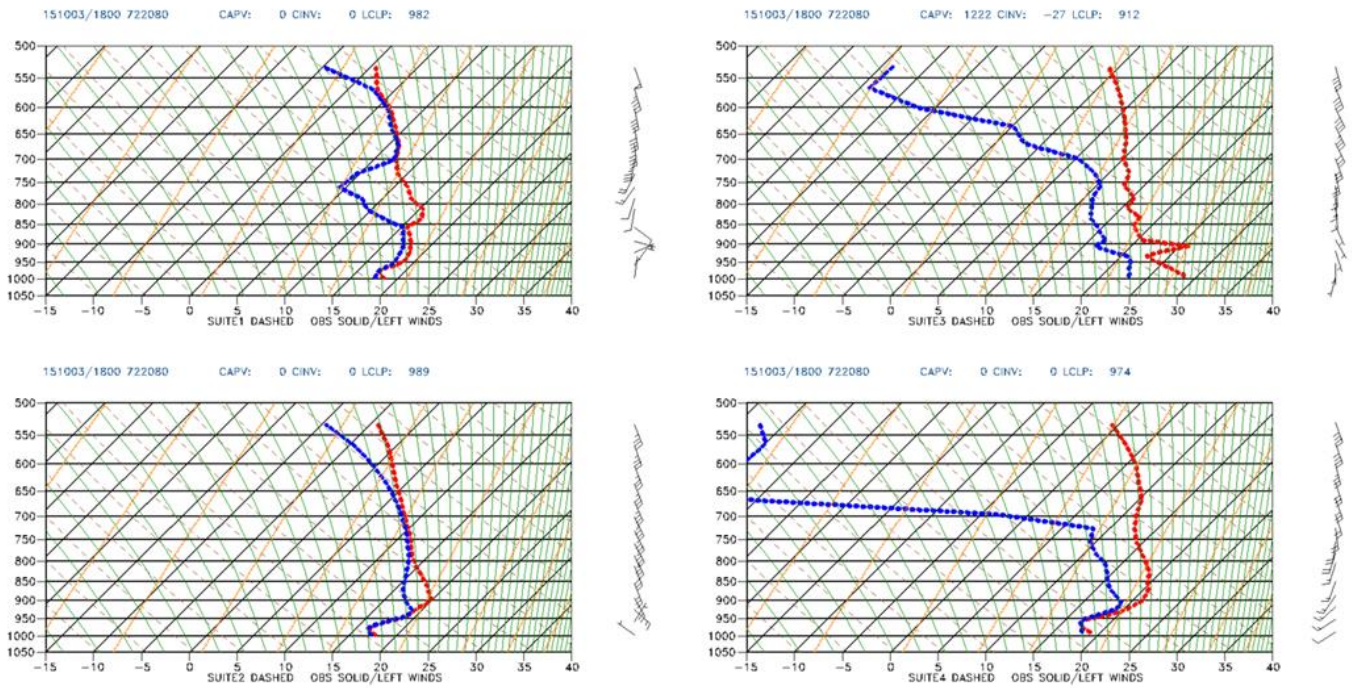


Fig. 43. XX-hr forecast for Charleston, SC valid 1800 UTC 3 October 2015.

Another example is shown in Fig. 44. Suite 3 has a low-level moisture profile that looks somewhat unrealistic, and while we do not have verification, its moisture profile is clearly unlike the other suites.

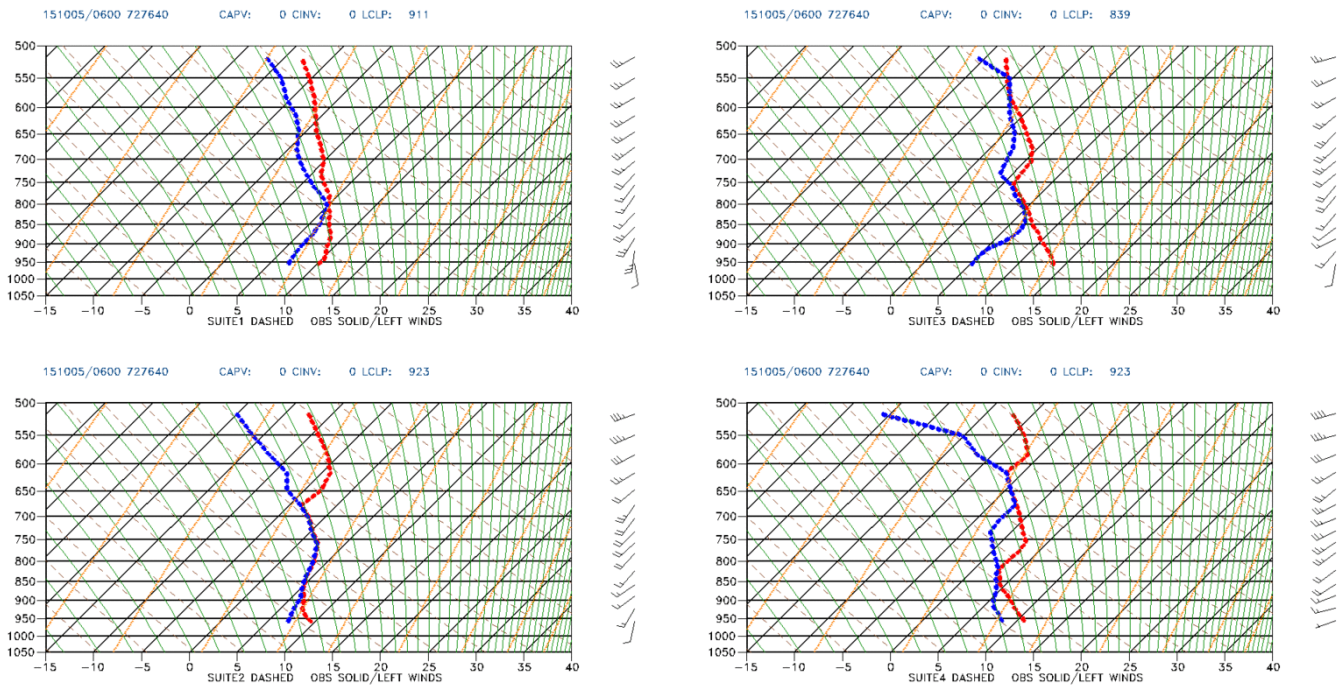


Fig. 44. XX-hr forecast for Bismarck, ND valid 0600 UTC 5 October 2015.

In summary, the impacts of the different suites on vertical profiles appear to be very mixed. Suite 4 clearly offers the best hope towards resolving the serious GFS problem of forecasting low-level radiation inversions that are clearly too weak, but it does have the occasional issue of forecasting low-level inversions that are too strong. More concerning, Suite 4 is prone to overmixing the boundary layer on warm days, leading to overly deep boundary layers that are too hot and too dry in the low levels. Consistent with this strong mixing, Suite 4 seems to have a tendency to mix out capping inversions.

Suite 2 offers a clear improvement over Suite 1 with regards to forecasting low-level inversions, but its overall handling of them is still not as good as Suite 4. Suite 2 also shows instances of being too cold during the daytime hours, which subjectively offset some of the gains from the handling of low-level inversions.

Suite 3 did not show handling of inversions or afternoon mixing that was any worse that any of the other suites, but benefits were also limited. Suite 3 also displayed some odd temperature and moisture structures.

NON-TROPICAL CASE SUMMARIES

For the non-tropical cases, the MEG was asked to focus on events occurring the medium range. The MEG realizes the limitations of drawing conclusions from single model runs, but it is critical for global model test suites to be able to capture significant events in the

medium range. Short recaps for a few of the cases listed in Table 3 are presented, covering some key details from the events or common themes.

Table 3: List of non-tropical cases evaluated by the NCEP/EMC Model Evaluation Group

Case	Forecast Cycle
Blizzard of 2016 (2016)	1200 UTC 18 January 2016
Plains Severe Weather (2016)	0000 UTC 22 April 2016
"Pi Day" Winter Storm (2017)	0000 UTC 10 March 2017
West Coast Atmospheric River (2017)	0000 UTC 15 March 2017
Flooding in the Mississippi Valley (2017)	0000 UTC 20 April 2017
Extreme California Heat (2017)	0000 UTC 29 July 2017
Inversions (2018)	1200 UTC 16 October 2018
East Coast "Bomb" Cyclone (2018)	0000 UTC 1 January 2018

Blizzard of 2016

The Blizzard of 2016 was noted for its remarkable predictability in the medium range with operational models. While the retrospective FV3GFS correctly captured many of the details, it was notably too progressive. All four suites appear to be too progressive (Fig. 45), with Suites 1 and 3 being the most progressive, but overall not as progressive as the FV3GFS retrospective runs. Suites 2 and 4 incorrectly brought in too much low-level warm air which resulted in those forecasts changing from snow to rain in parts of the Mid-Atlantic.

The precipitation forecasts were generally good, but Suite 3 did not extend precipitation far enough to the north into New York City and Boston. Oddly, the snowfall maps from Suite 4 (Figs. 46 and 47) look fine when computed from the snow water equivalent (10:1 ratio), but are remarkably low when examining the change in snow depth. It is possible that there is some sort of disconnect between snow accumulating inside the model and the land-surface scheme.

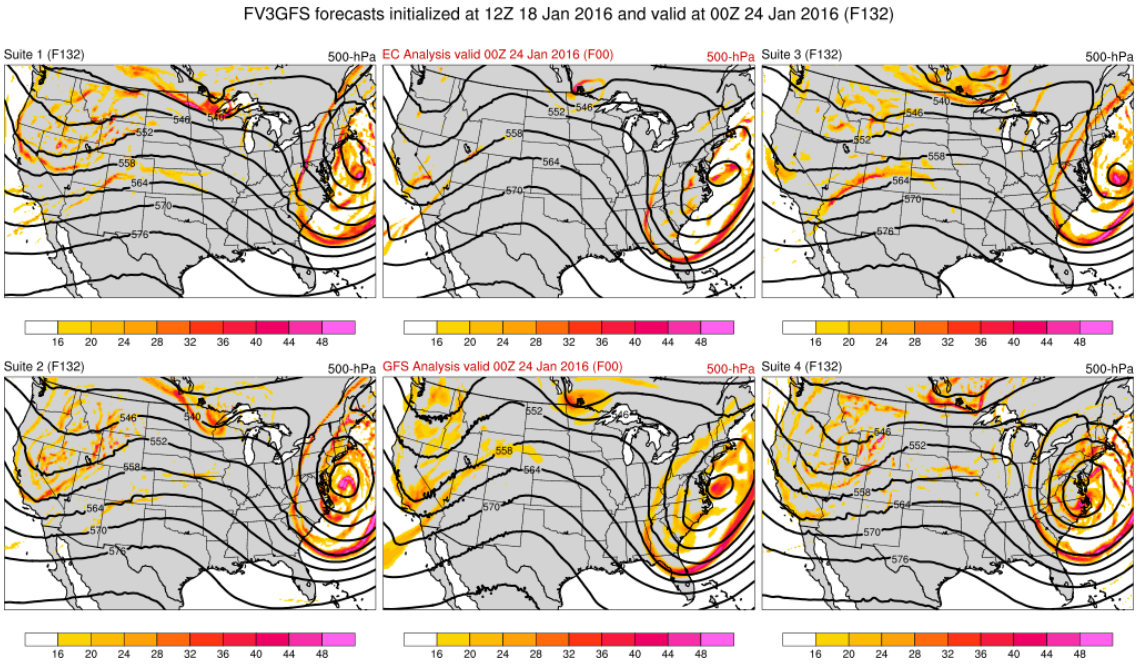


Fig. 45. 132-h forecasts of mean sea level pressure (left and right columns), initialized at 1200 UTC 18 January 2016 and valid at 0000 UTC 24 January 2016. The ECMWF analysis (top middle) and the GFS analysis (bottom middle) for this valid time are also shown.

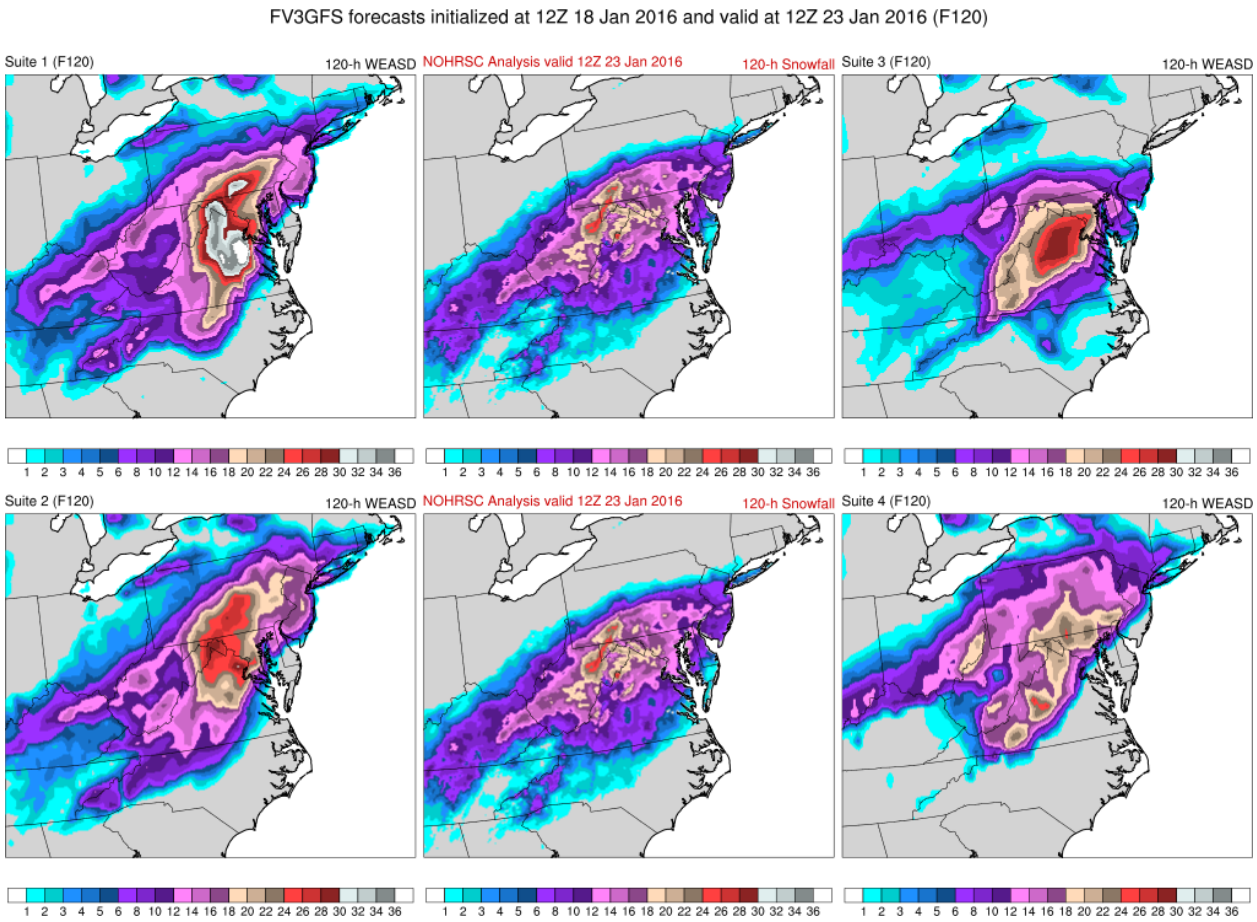


Fig. 46. 1200 UTC 18 January 2016 cycle 120-hour forecasted water equivalent snow accumulations (with a 10:1 SLR applied) valid 1200 UTC 23 January 2016.

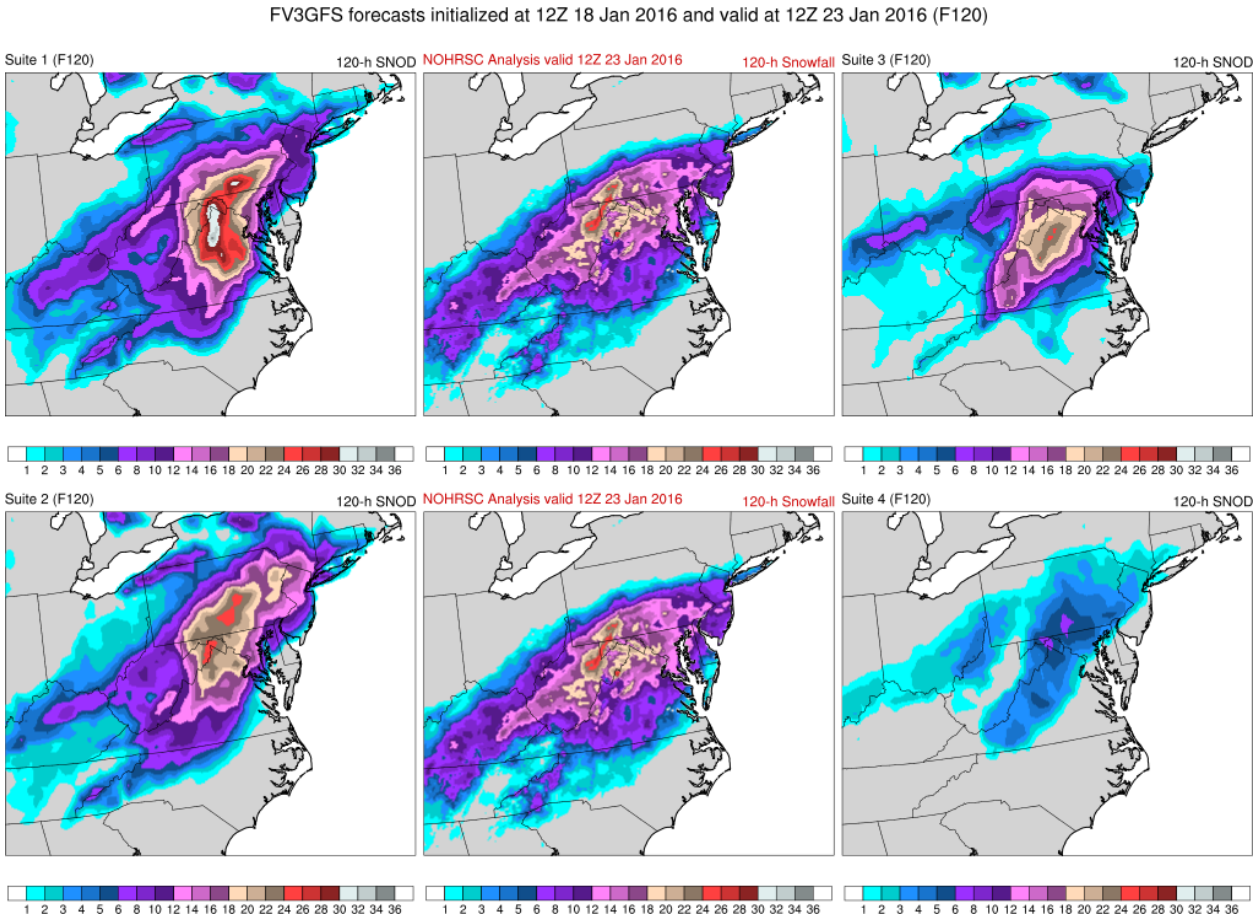


Fig. 47. Same as in Fig. 46, except for accumulated snow depth.

Great Plains Severe Weather

This event featured multiple shortwave troughs moving across the plains and provided an opportunity to examine instability forecasts and the handling of a sharp dryline. Overall, all suites struggled to some extent with synoptic details, making a true assessment of convective details difficult. That said, Suite 1 generally had higher CAPE values than Suite 2 (Fig.), which is concerning due to the documented issues of the GFS and FV3GFS being too weak with instability. Suite 3 overall had the largest coverage of higher values, with suite 4 close behind. Both overall did well with warm-sector instability relative to suites 1 and 2, but both tended to pool moisture along boundaries, leading to narrow corridors of erroneous extreme instability. All four suites overmixed the boundary layer in this case (consistent with the operational GFS) and pushed the dryline too far to the east (inferred in Fig. from the north-south CAPE gradient, although this could partially be explained by synoptic errors since all of the suites were too progressive with the cutoff low. Suite 4 by far overmixed the most of all of the suites, and pushed the dryline the furthest east. Suite 4 also seemed to forecast a much tighter gradient in the dryline than the other suites.

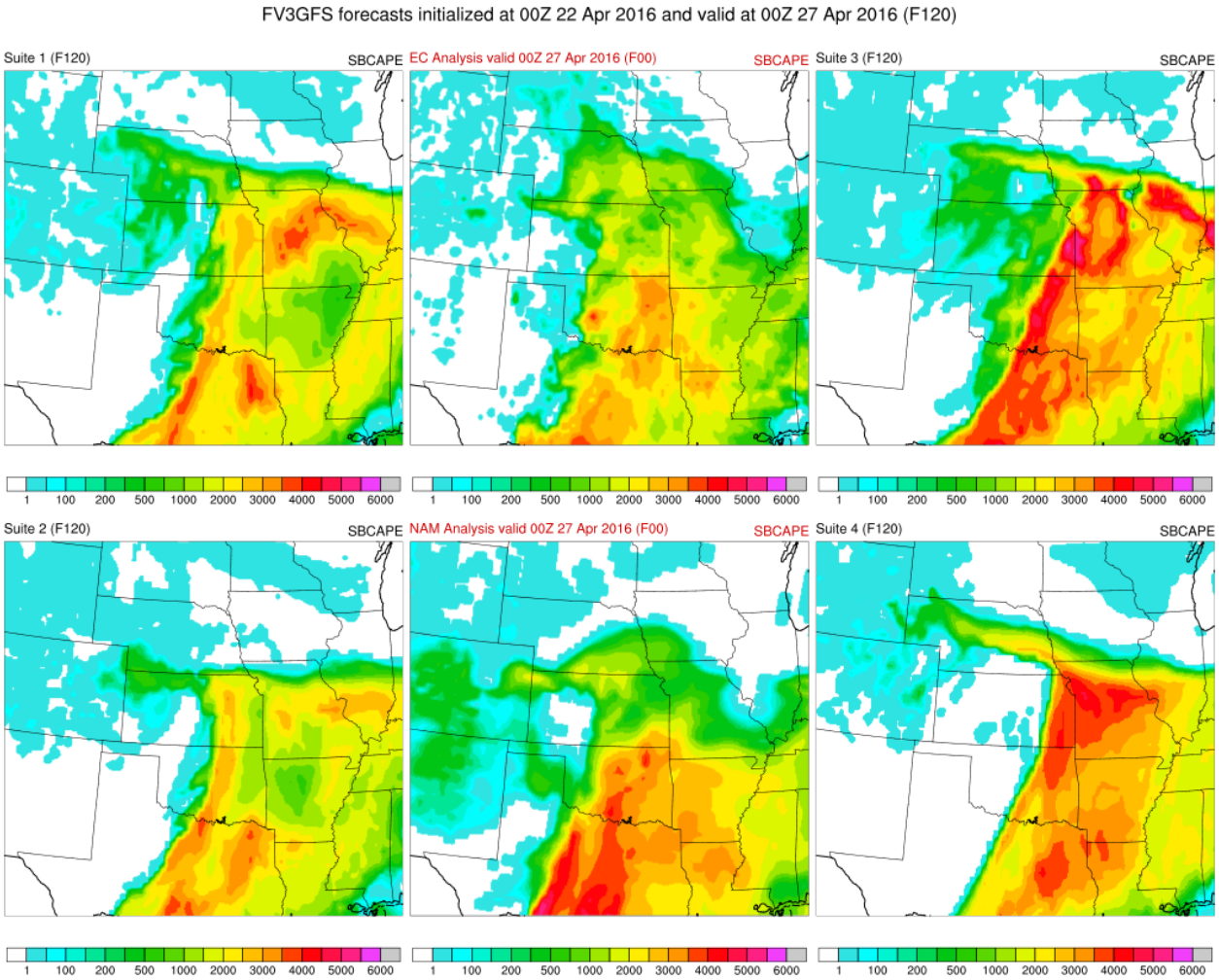


Fig. 48. 132-h forecasts of mean sea level pressure (left and right columns), initialized at 1200 UTC 18 January 2016 and valid at 0000 UTC 24 January 2016. The ECMWF analysis (top middle) and the GFS analysis (bottom middle) for this valid time are also shown.

"Pi Day" Winter Storm

Suites 1-4 all have acceptable 132-h forecasts of the "Pi Day" Winter Storm, with each suite producing a cyclone located over/near Maine by 1200 UTC 15 March 2017 (Fig. 49) with a central pressure between 980–988 hPa (983 hPa in GFS analysis). Suite 3 produces the weakest cyclone of the four physics suites, likely due to its westward-shift track that takes the center of the cyclone slightly over land. The 850-hPa winds in all four suites are weaker than those in the ECMWF and NAM analyses, likely the result of each suite producing a slightly weaker cyclone (with a slightly weaker pressure gradient) than was observed. All four physics suites are too fast with the eastward progression of the 500-hPa shortwave trough associated with cyclogenesis. This results in the formation of the surface cyclone 6-h earlier in all four physics suites than in the ECMWF and GFS analyses. All four suites do a reasonable job with 24-h precipitation totals, capturing the western extent and maximum values well. All four suites also do a reasonable job with 24-h snowfall totals, with all four suites indicating the potential for >20" in their 120-h forecasts.

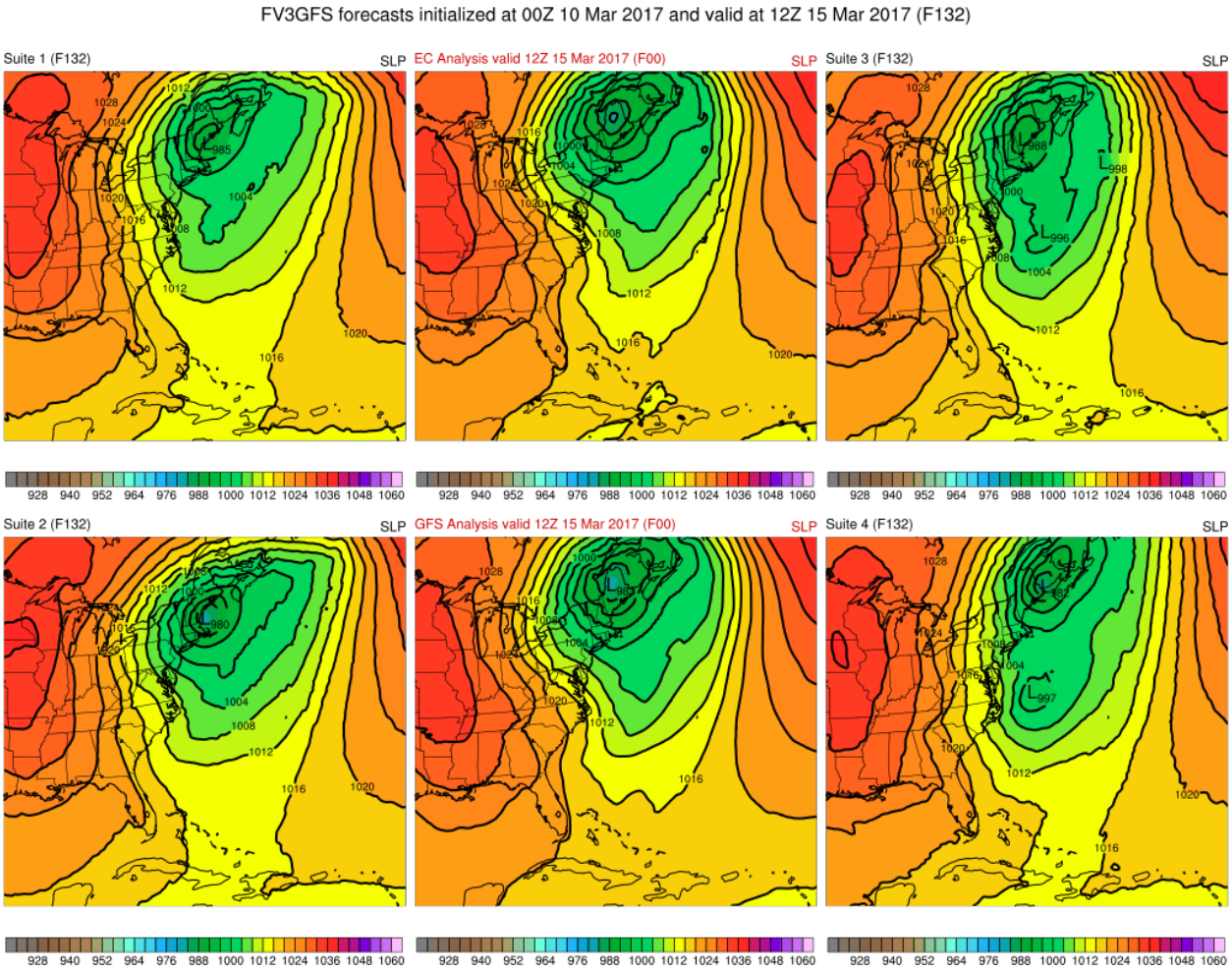


Fig. 49. 132-h forecasts of mean sea level pressure (left and right columns), initialized at 0000 UTC 10 March 2017 and valid at 1200 UTC 15 March 2017. The ECMWF analysis (top middle) and the GFS analysis (bottom middle) for this valid time are also shown.

Atmospheric River

All four physics suites were able to capture the location/orientation of the atmospheric river axis extending from Hawaii to the coasts of Washington and Oregon at 1200 UTC 18 March 2017. The 24-h precipitation totals during this period (0000 UTC 18–19 March 2017) are fairly similar to the STAGE IV analysis (Fig. 50), with higher totals in the Olympic and Cascade Mountains. Suite 4 seems to do a slightly better job with orographic precipitation and higher precipitation totals along the coasts of Washington and Oregon than the other suites, although all four suites are generally too low with precipitation totals along the West Coast.

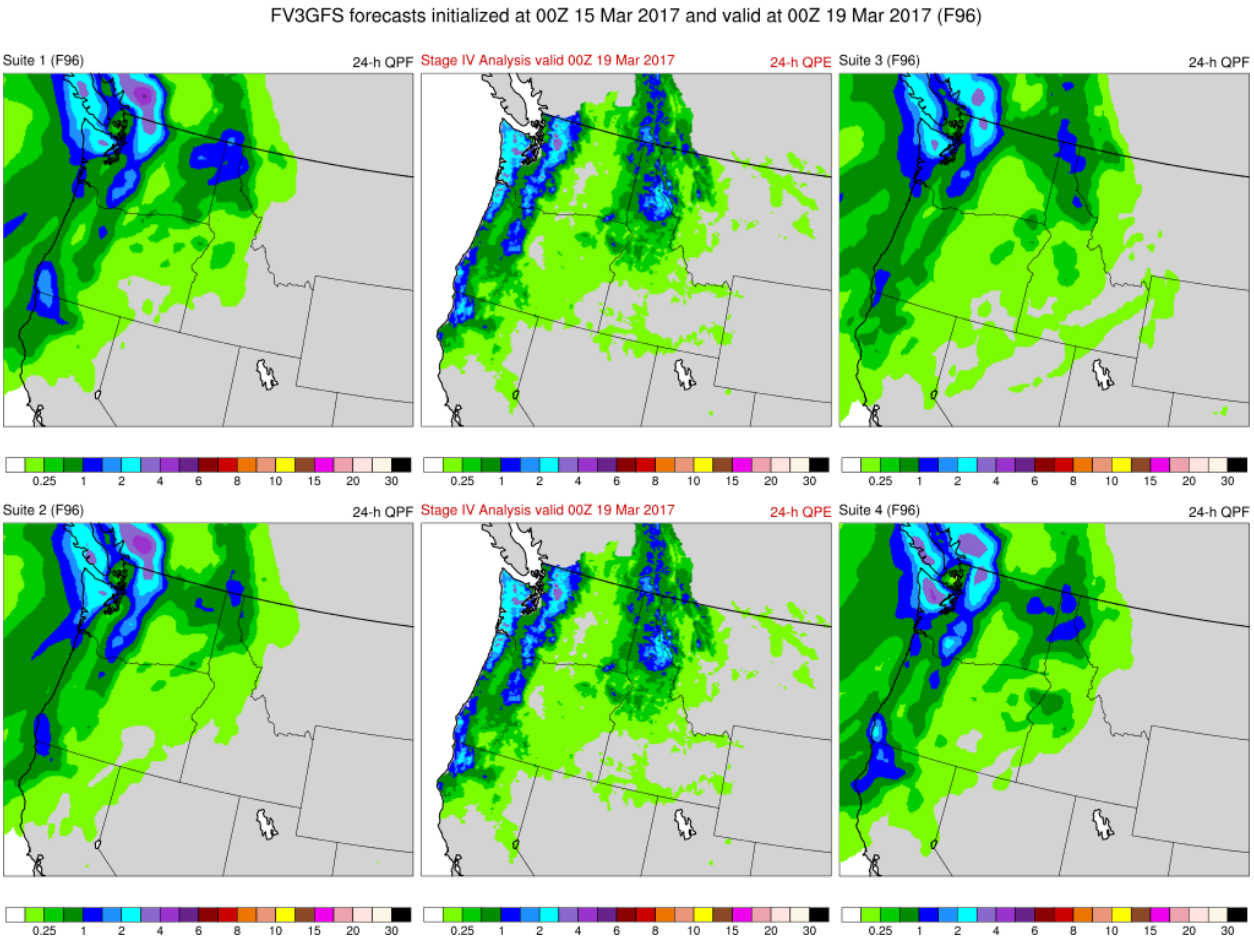


Fig. 50. 96-h forecasts of 24-h accumulated precipitation (left and right columns). Forecasts were initialized at 0000 UTC 15 March 2017 and are valid at 0000 UTC 19 March 2017. The 24-h Stage IV quantitative precipitation estimate analysis (top middle and bottom middle) valid at 0000 UTC 19 March 2017 is also shown.

Mississippi Valley flooding

The high-impact 2017 Mississippi Valley flooding event in 2017 had multiple stages, but the suites did show the potential for a significant flooding in longer-range forecasts covering the entire period, even though there were some significant errors associated with some of the individual events. The total precipitation for the entire 240 hour forecast (Fig. 51) shows that Suite 1 handled the magnitude the best and does well with the location of heaviest rainfall, even though it overpredicted in some areas. Suite 4 grossly underestimated the amounts, consistent with the overall dry bias shown in Fig. 15.

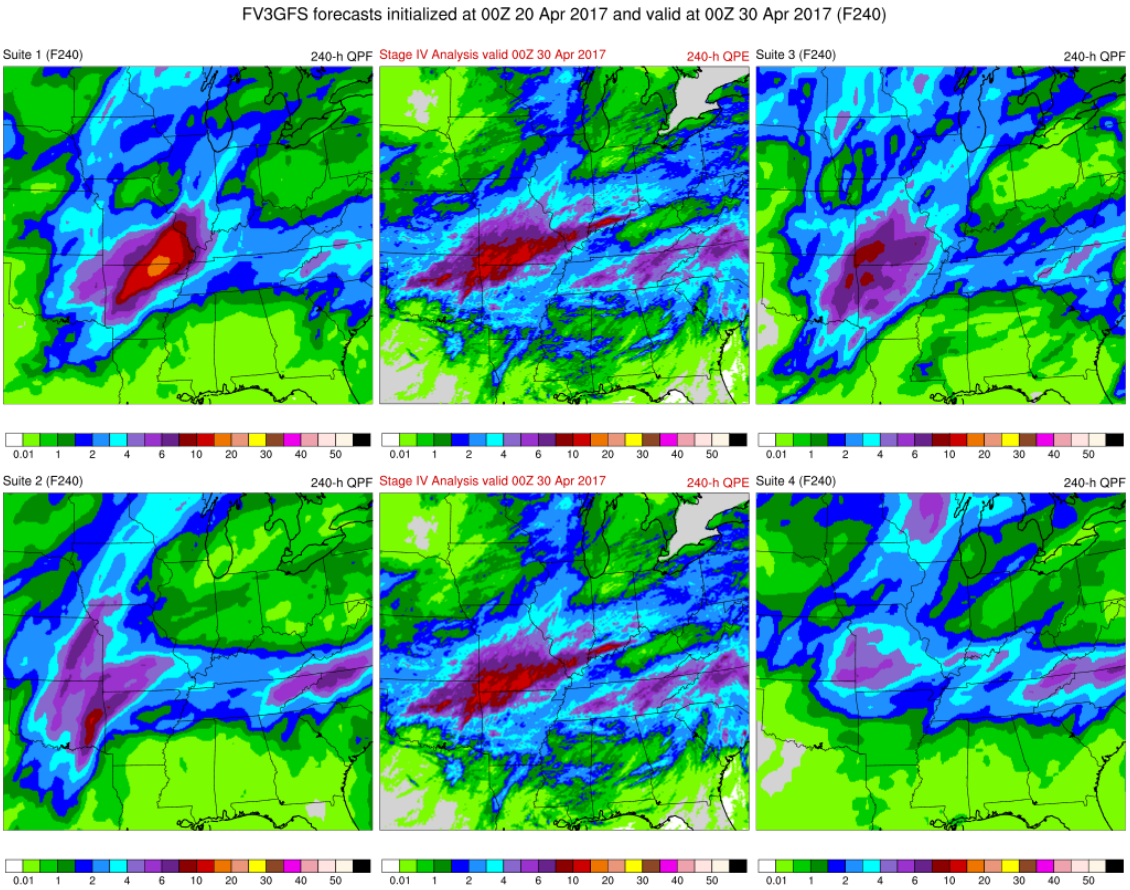


Fig. 51. 240-h forecasts of total accumulated precipitation (left and right columns). Forecasts were initialized at 0000 UTC 20 April 2017 and are valid at 0000 UTC 30 April2017. The Stage IV quantitative precipitation estimate analysis (top middle and bottom middle) covering the entire period is also shown.

Interior California Heat

In late July 2017, the GFS and FV3GFS incorrectly forecasted very warm 2 meter temperatures over interior California. The four suites showed no improvement over the FV3GFS or the GFS (Fig. 52), as all suites were still too warm. While examining this case, a side detail was noted in the cloud fields (Fig. 52). Forecasting the evolution of marine stratocumulus cloud cover along the West Coast is a common model struggle, but Suite 4 consistently forecasted the marine clouds when all of the other suites completely missed them (Fig. 53). There were indications throughout the period that Suite 4 forecasted the occurrence of these marine clouds too frequently, but no other suite was able to generate these clouds in throughout this model cycle.

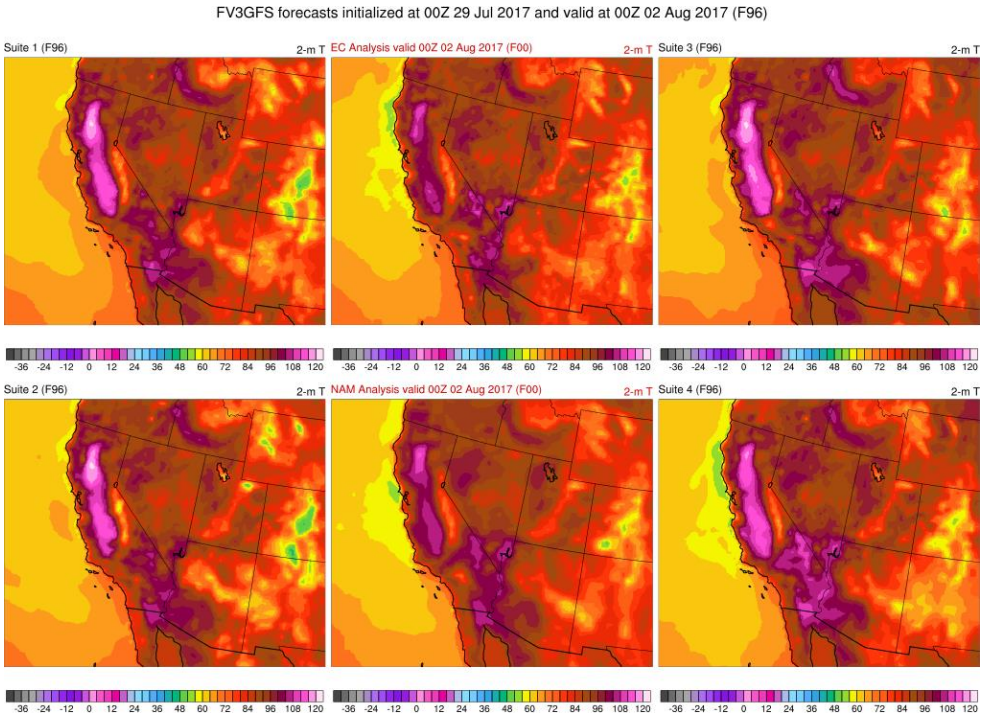


Fig. 52. 96-h forecasts of 2-m temperatures (left and right columns). Forecasts were initialized at 0000 UTC 29 July 2017 and are valid at 0000 UTC 2 August 2017. Corresponding analyses from the ECMWF model (top middle) and NAM (lower middle) valid at 0000 UTC 2 August 2017 are also shown.

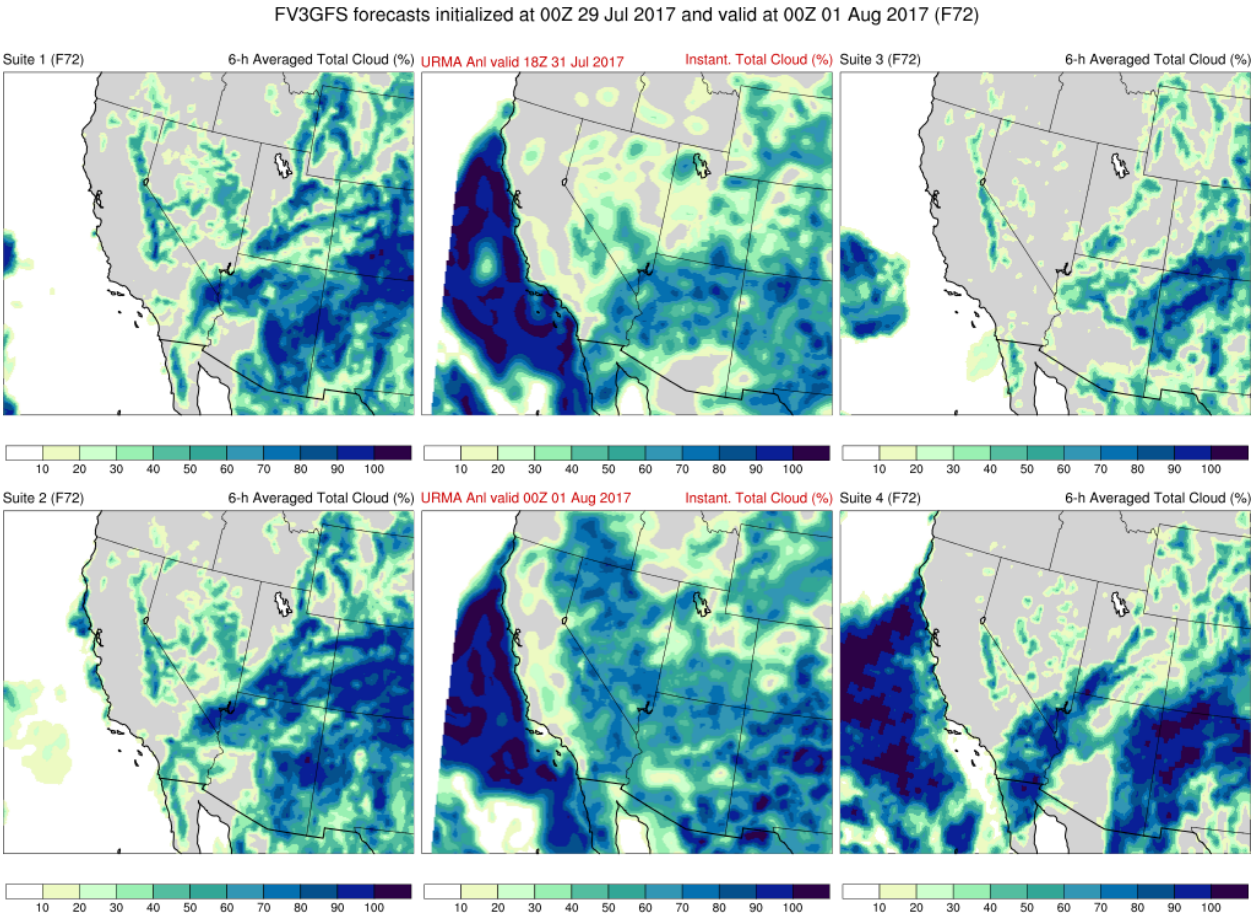


Fig. 53. 72-h forecasts of 6-hour averaged total cloud fraction (left and right columns). Forecasts were initialized at 0000 UTC 29 July 2017 and are valid at 0000 UTC 1 August 2017. The instantaneous URMA analyses valid at the start (top middle) and end (lower middle) of the 6-hour period are also displayed.

Inversion Case

The GFS and the FV3GFS struggle with low-level radiations inversions in the early morning hours, and this leads to morning 2-m temperature forecasts that are much too warm. This case provided multiple examples to examine the handling of inversions, and all four suites were consistently too warm in the early morning as shown in Fig. 54 with Suite 4 the coolest, consistent with the findings in the soundings section. The soundings section also noted a tendency for Suite 4 to significantly overmix during afternoon hours on the plains, leading to too hot and too dry conditions, and this case provided multiple examples. Fig. 55 shows one example, with Suite 4 showing very large warm errors that are not seen in the other suites.

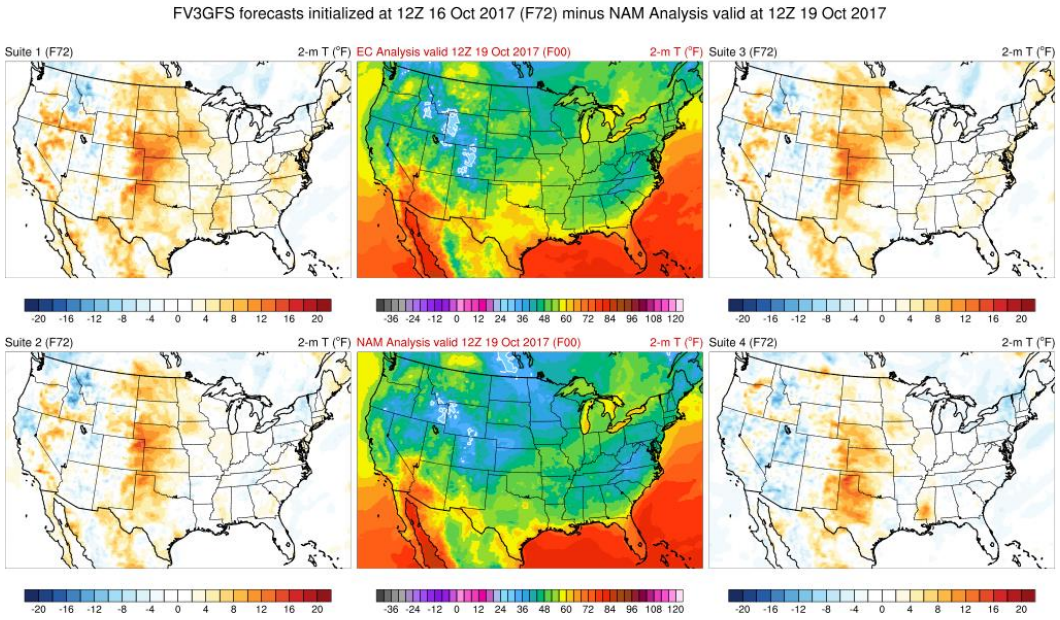


Fig. 54: 72-h forecasts of 2-m temperature differences between the NAM analysis and the forecast (left and right columns). Forecasts were initialized at 1200 UTC 16 October 2017 and are valid at 1200 UTC 19 October 2017. Corresponding analyses from the ECMWF model (top middle) and NAM (lower middle) valid at 1200 UTC 19 October 2017 are also shown.

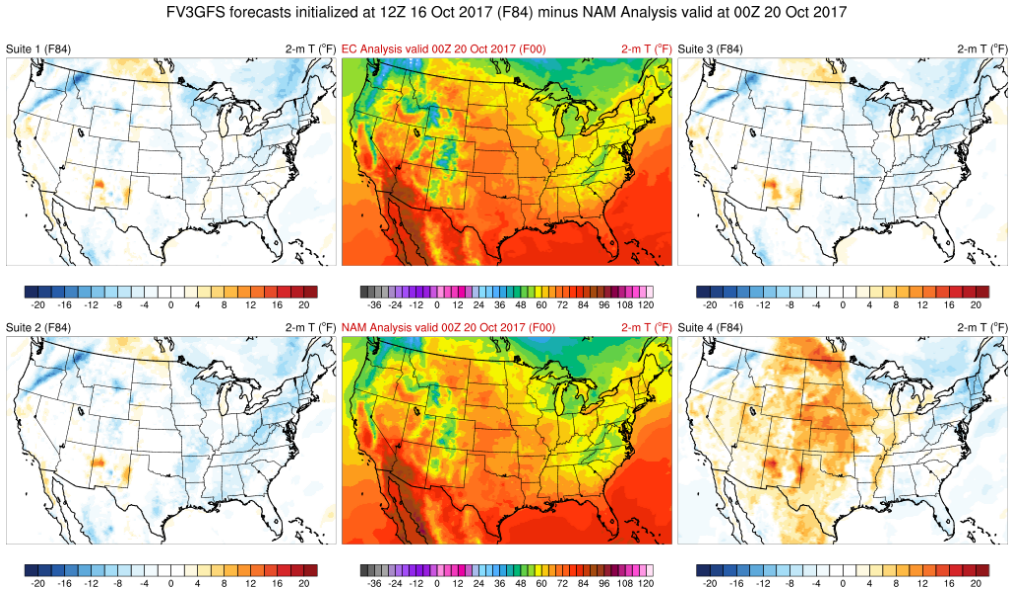


Fig. 55. 84-h forecasts of 2-m temperature differences between the NAM analysis and the forecast (left and right columns). Forecasts were initialized at 1200 UTC 16 October 2017 and are valid at 0000 UTC 20 October 2017. Corresponding analyses from the ECMWF model (top middle) and NAM (lower middle) valid at 0000 UTC 20 October 2017 are also shown.

Bomb Cyclone

The GFS and FV3GFS handled the 2018 bomb cyclone similarly, and all four suites were able to forecast the deepening rate analyzed by the EC and GFS. During the early development period from 66-h to 84-h forecasts, all suites had the low too far offshore. Further north, Suites 1 and 2 did the best job with forecasting the timing and position of the low tracking over the Bay of Fundy. Suite 3's track however was too far east over Nova Scotia while Suite 4's track was too far west over Maine (Fig. 56).

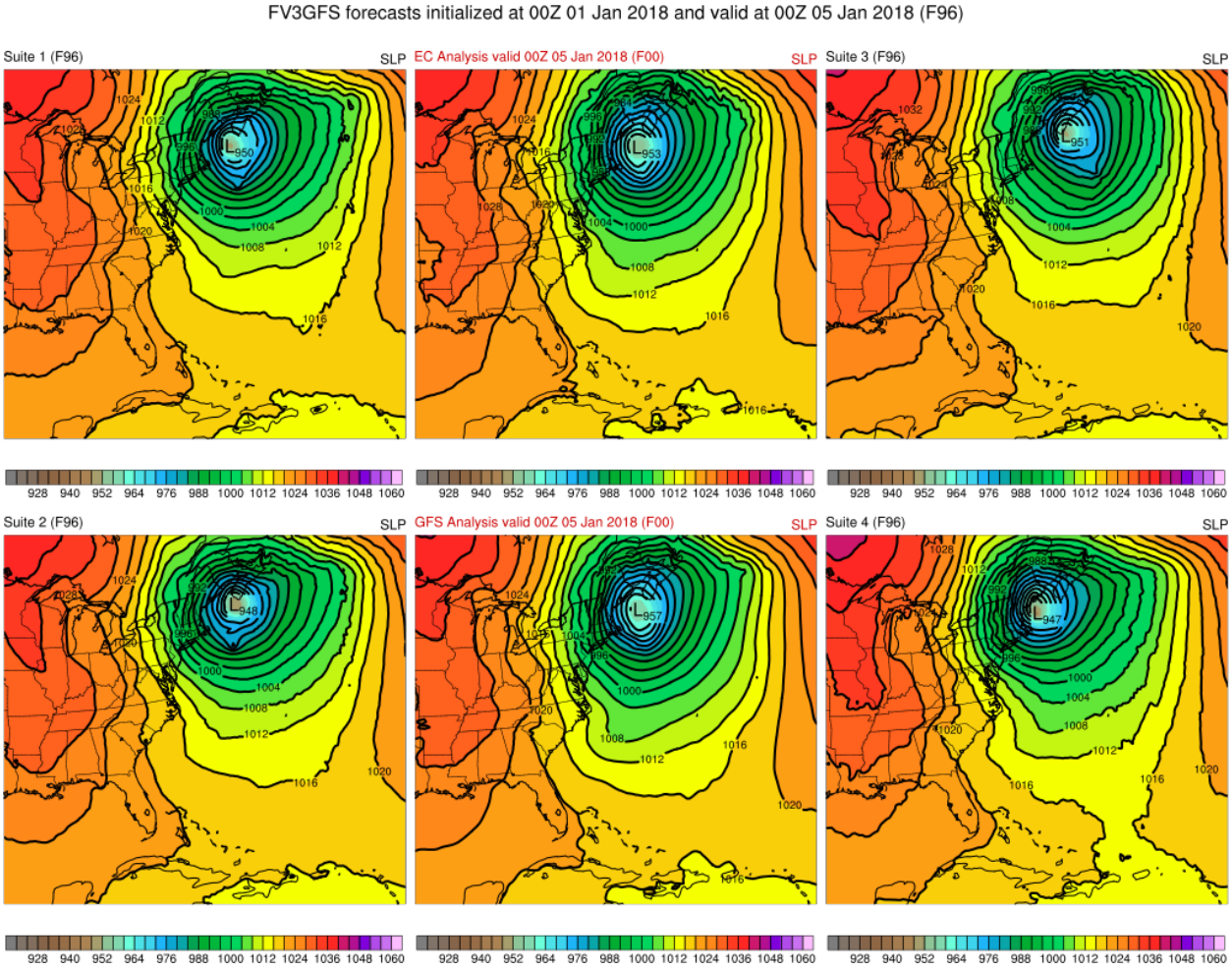


Fig. 56. 96-h forecasts of sea level pressure (left and right columns). Forecasts were initialized at 0000 UTC 1 January 2018 and are valid at 000 UTC 5 January 2018. Corresponding analyses from the ECMWF model (top middle) and NAM (lower middle) valid at 0000 UTC 5 January 2018 are also shown.

CONCERNS

The MEG wishes to express some concerns it has about the testing process and the bigger picture of updates to the physics within the GFS:

1) The radiation-microphysics interaction in the FV3GFS has recently been modified in GFSv15, and these changes were not included in these tests. How do we reconcile these test results with the fact that they were derived using a different treatment of radiation and the GFDL microphysics scheme?
2) These tests were run with 64 vertical levels, but GFSv16 will contain 96 (or possibly 128) vertical levels. The MEG has concerns about not testing physics using the vertical resolution at which the model be run.
3) The cold bias that increases with time is limited to suites that contain the GFDL microphysics. The causes of this cold bias need to be determined.
4) The fix to the radiation driver bug seems to have either introduced or worsened a seasonal cycle to temperature bias (Figs. 5 and 6). The MEG emphasizes that the impacts of that model change need to be better understood before finalizing GFSv16.
5) NCEP has longer-range plans to turn off the NAM model, but it cannot be done until the GFS is able to perform well with the NAM strengths. Two of those key strengths are the handling of inversions and instability, and the MEG is concerned that insufficient progress is being made in those areas to justify turning off the NAM anytime soon. In that regard, while the choice to investigate medium range cases makes sense from the perspectives of eliminating "spin-up" issues and focusing on key aspects of forecasts in the medium range, thought will need to be given to constructing a testing framework that allows for inspection of short-term forecasts so that issues such as boundary layer structure, precipitation type, and instability can be assessed without interference from medium-range synoptic errors.

FINAL COMMENTS

The MEG is grateful for having had the opportunity to participate in this process, and we welcome questions and comments related to this specific report and our evaluation activities in general. The MEG believes that there are significant positives from each suite, and it encourages development on all of them. With that said, this report concludes with Table 4, listing the MEG's assessment of the pros and cons of each suite.

Table 4. The MEG's assessment of the positive and negative characteristics of each of the four physics suites that were tested.

Suite 1	
Pros	**Cons**
Overall the best synoptic scores	Really struggles with inversions
Overall good tropical cyclone track/intensity	Underdoes instability
	Low-level cold bias that increases with time

Suite 2	
Pros	**Cons**
Synoptic scores very similar to Suite 1	Underdoes instability (more than Suite 1)
Overall good tropical cyclone track/intensity	Larger low-level cold bias than Suite 1
Improved handling of inversions	Even drier than Suite 1 with tropical precip.

Suite 3	
Pros	**Cons**
Some improvement handling inversions	Synoptic scores not as good as Suite 1
More representative instability magnitudes	Increasing low-level warm bias with time
	Struggles with tropical cyclone track/intensity

Suite 4	
Pros	**Cons**
Shows the most promise handling inversions	Synoptic scores not as good as Suite 1
Shows some promise for improving instability	Overmixes PBL, leading to hot/dry Plains
Reduction of low precipitation bias in tropics and ability to predict marine stratus	Too light in extreme precipitation events (largest low bias in precipitation overall)
Smallest low-level temperature bias	Struggles with tropical cyclone track/intensity