1. Introduction

The Model Evaluation for Research Innovation Transition (MERIT) project provides a critical framework for physics developers to test innovations within their schemes using selected meteorological cases that have been analyzed in depth. Comparing their results to baseline MERIT simulations will allow developers to determine whether their innovations address model shortcomings and improve operational numerical weather prediction.

For the DTC's AOP 2018, three high-impact global FV3 baseline cases were selected for in-depth analysis: the Mid-Atlantic blizzard of January 2016, Hurricane Matthew, and the May 2017 severe weather outbreak in the Southern Plains. These cases were chosen after consultation with the Model Evaluation Group (MEG) at NOAA's Environmental Modeling Center, as each case exhibits known deficiencies in the global GFS configuration of the Finite-Volume Cubed-Sphere (FV3) model. Multiple-day simulations were run using an end-to-end workflow developed to handle the pre-processing of initial conditions, the integration of the model, post-processing with the Unified Post Processor (UPP), and verification with the Model Evaluation Tools (MET). Also, in collaboration with the MEG, the MERIT team worked on applying unique verification techniques and metrics that can be used to assess the impact that physics innovations may have on these known FV3 biases. In particular, the progression of certain meteorological features were assessed through the MET Method for Object-Based Diagnostic Evaluation (MODE) time-domain (TD)/storm-relative feature analyses.

Providing the research and operational communities with an end-to-end framework will streamline the testing process, leading to more effective and efficient physics development. In addition, it will also encourage community engagement and provide an infrastructure that supports R2O and O2R.

2. Experimental Design/Methods

To conduct the three baseline cases selected through collaboration with the MEG, the fv3gfs workflow repository was checked out on the Theia NOAA supercomputer and a workflow was generated for each case using the python workflow generation scripts. The most recent version tag of the NEMSfv3gfs Vlab repository (nemsfv3gfs_beta_v1.0.12) was automatically checked out and built during the workflow generation step and was used for each case study.

The Mid-Atlantic blizzard case was initialized at 12Z on 18 January 2016 and run for 240 hours, ending at 12Z on 28 January 2016. Verification for that case extended to 168 hours, or 12Z on 25 January 2016. Similarly, a 240-hr simulation was generated for the May 2017 severe weather outbreak case study, beginning at 00Z on 12 May 2017 and ending at 00Z on 22 May 2017. Verification was conducted up to 168 hours, or 00Z on 19 May 2017. For the Hurricane Matthew case study, an initialization of 00Z on 29 September 2016 was used, matching the

MERIT initialization of the same case from AOP 2018. However, analyses from last year showed that a longer forecast was required to sufficiently incorporate the period of time when the hurricane was the closest to the continental United States. Therefore, forecasts were issued for 12 days (288 hours), to 00Z on 11 October 2016. Verification for this case was also conducted for the full 12 days of the simulation.

Each case was run using an end-to-end workflow generated through the fv3gfs workflow repository python scripts. The XMLs produced through these scripts included tasks to get and produce initial conditions from pre-existing GFS forecasts, run the global FV3 simulations, and run the post-processing of the output NetCDF files generated during the forecasts. Following the completion of the post tasks, a verification workflow was submitted to process the post data through MET. Traditional verification was conducted within this verification workflow, while individual diagnostics, such as MODE-TD, the storm-relative tool, and the hurricane tracker were run separately.

3. Verification Methods

In order to run both the traditional verification and some of the individual diagnostics, a pre-compiled version of MET v8.0 was used by issuing a module load command on Theia. The verification workflow was then submitted to conduct point verification (using point_stat) for 2-/10-m variables (temperature, relative humidity, specific humidity, u-/v-component of the wind, full-vector wind speed, and mean sea-level pressure) using NDAS station observations and upper-air variables at standard pressure levels (temperature, relative humidity, height, u-/v-component of the wind, and full-vector wind speed) using GDAS radiosonde observations. Precipitation accumulation verification was conducted for all three cases against the CCPA and CMORPH gridded data sets at 6-h and 24-h accumulation intervals (using grid_stat). Further verification was undertaken for the severe storm case, including forecasts of two-meter dew point using NDAS surface observations, and CAPE/CIN using NARR data.

Following the completion of the verification workflow, MET output was transferred to a local server at NCAR and loaded into a mysql database to be queried for batch plotting. METViewer XML scripts were then developed to produce all traditional metrics. Bias and RMSE were plotted as a function of forecast hour for all surface variables and for upper-air variables at all levels. Vertical profile plots were also produced for specific forecast hours. Finally, frequency bias and equitable threat score plots were generated for precipitation accumulation fields.

Verification was conducted over a number of regions, including the CONUS, northern and southern hemispheres, the tropics, and the full global domain. Within this report, results are presented for the CONUS and global domains.

3.X Diagnostic Tools

Feature relative

The new METplus feature relative use case was configured to run on Hurricane Matthew. While this use case was made available in earlier versions, METplus 2.1 is required for a successful run due to various bug fixes. This tool requires two gridded datasets, generally a forecast and some reference dataset, track files for each, which identifies the center point of your feature for each forecast, and a user defined configuration file. For this case, the FV3GFS output was used for the forecast, while the GFS analysis was used for the reference dataset. While the GFS is not the best dataset to use as truth for a hurricane, due to the coarse resolution not resolving the intensity of the hurricane well, it is used to demonstrate the capability of the feature relative diagnostic tool here. The GFDL vortex tracker was used to create the track files. The configuration file (feature_relative.conf) details the various paths/parameters necessary to run the case and defines the processes to be run. The first process involves running the MET tool tc-pairs, which matches the two track datasets and processes the track and intensity output. Once tc-pairs is complete, the process ‘extract tiles’ is run, which extracts a user-defined tile, generally centered on the track location of the feature, for each forecast and refencence file for each field and level requested. The last process uses the MET tool series analysis to aggregate statistics spatially over a specified time period for the forecast versus analysis tiles. For this case, series analysis was run over the entire 12-day forecast as well as aggregated by 1-day periods. As part of the feature relative tool, simple visualization using METs plot_data_plane tool is also provided, however; for this case, an NCL script wrapped by a user edited shell script was used to create 2D plots of the feature relative output,.with coordinates relative to the feature in degrees.

Vertical cross-sections

An NCL script was created for plotting 2D vertical cross sections along a transect between any two user-defined endpoints. The transect is defined by calculating the great circle distance (shortest distance) between the two points, and then interpolating over some number of points, determined by desired resolution, along the great circle path between the endpoint. For planes desired at a constant latitude, a separate script exists, since the great circle path between two points at constant latitude does not follow the latitude. Additionally, values that are at a pressure higher than the model surface pressure at that grid point are masked out, where there should be no data. The plots are used to look at cross sections across hurricane Matthew to examine the structure of variables such as wind speed and relative humidity.

MODE Tool

The Method for Object-based Diagnostic Evaluation (MODE) tool is a spatial diagnostic tool used to identify objects from a spatial field (Bullock et al. 2006). An object is identified by first defining a threshold of a field, for example, composite reflectivity greater than 30 dBZ. Other filters such as a convolution radius and attribute filters, like length, are applied to the field in question to produce objects. Attributes are calculated for the objects, both forecast and observed. These include: object area, object count, centroid displacement, and more. This tool is used to identify Convective Available Potential Energy (CAPE), Convective Inhibition (CIN), and surface specific humidity gradient objects to explore model deficiencies in the May 2017 severe storms outbreak case.

MODE Time Domain

The MODE time domain (MTD) tool in MET is an object based approach to verification similar to MODE, but in both space and time, which can be thought of as a 3D object or ‘volume’. MTD ingests a user provided series of forecast files and an accompanying reference file list for evaluating objects in space over the forecast time period as defined from the file lists. A configuration file used to run MTD lists the tunable parameters and fuzzy logic algorithms for identifying objects in space and time. Output includes output for 3D single and paired attributes of all simple and cluster objects identified. No graphics are currently output from running MTD, however; a few stand-alone scripts are available for visualization.

4. 2016 East Coast Blizzard

4.1 Synoptic Discussion

Heavy snowfall took place over parts of the Mid-Atlantic and Northeast US producing up to 36” of snow (Figure B1) from the 22nd to the 24th of January, 2016. More than 10,000 flights were cancelled in connection to the storm. At least 55 people were killed, and the total economic losses are estimated between $500 million and $3 billion. We are going to examine the deficiency which is the progressiveness of the upper level trough.

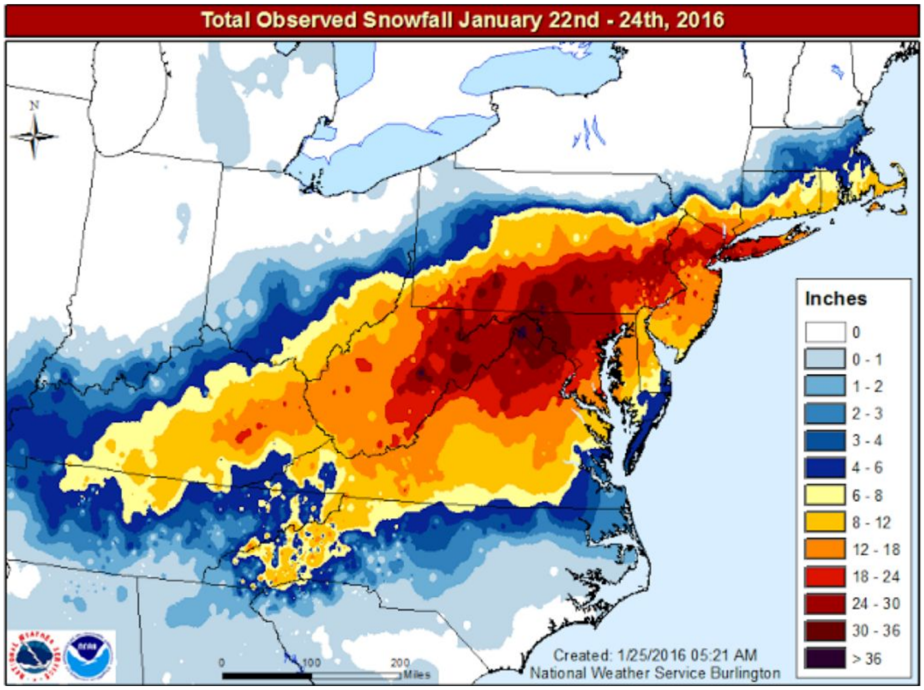

Total Observed Snowfall January 22nd - 24th, 2016
N
Inches
0
0 - 1
1 - 2
2 - 3
3 - 4
4 - 6
6 - 8
8 - 12
12 - 18
18 - 24
24 - 30
30 - 36
> 36
0
100
200
Miles
Created: 1/25/2016 05:21 AM
National Weather Service Burlington

Figure B1:The total observed snowfall on January 22nd-24th, 2016.

4.2 Verification Results

The peak of the event (20160123 12 UTC) is being chosen for the verification. The FV3 model accurately captured the magnitude of the upper-level low over Virginia/North Carolina (Figure B2b). However, the model has a positive tilted trough which means the model has not quite reached the maturity of the low pressure system yet, while the observations have almost reached maturity with the neutral tilted trough at 120 hours (Figure B2a). As for the deficiency, the model shows slightly more progressiveness compared to the observations. At the 540s height, the system in the model is slightly further east than observed.

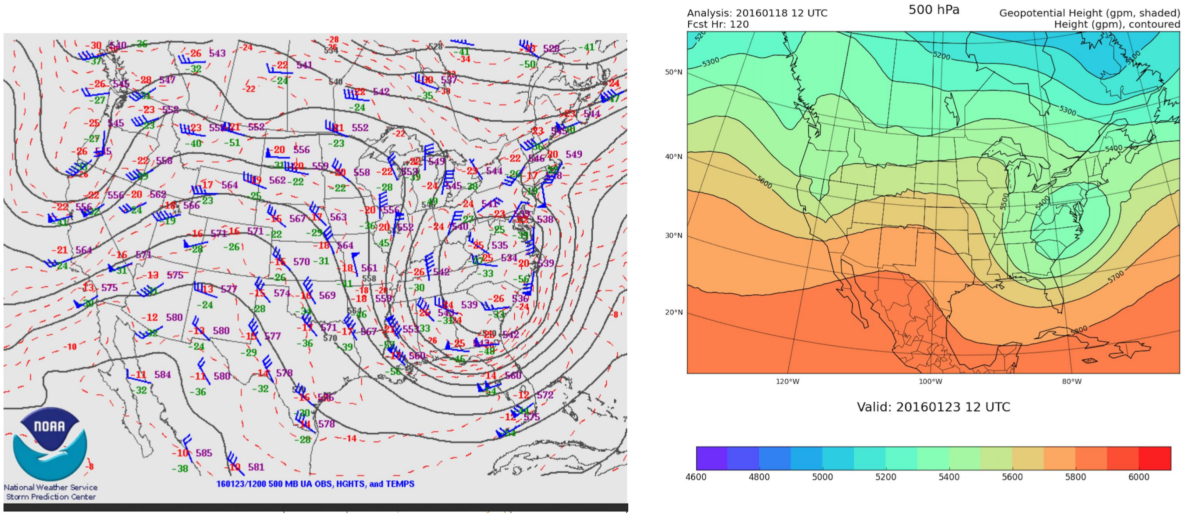


Figure B2: (a) Observed 500mb upper air observations, heights (solid black lines), temperatures (dashed red lines), and wind speed (wind barbs, in knots) valid at 20160123 12 UTC, and (b) FV3 500mb geopotential heights (shaded) and heights (contoured) valid 20160123 12 UTC.

Surface Verification (CONUS only)

Temperature

There is a noticeable diurnal variation in the temperature bias. The minima occur at 12Z valid times with the maxima appearing at 00Z or 06Z valid times. The minima have small negative values, -1.2 to 0 C while the maxima exhibit small positive values, 0.5-1.3 C. There is an overall shift of the bias toward positive values with forecast lead time. This behavior continues until 132 hours, with a sharp negative minimum appearing at 156 hours (Figure B3a). The temperature RMSE stays more or less steady between 2-3 C in the first 36 hours. It then rises in the range 3-5 C for the next 90 hours, demonstrating an approximate diurnal variation with maxima at or near 00Z valid times. After a sudden drop to 2.5 C at 132 hours, the RMSE moves up sharply, reaching above 6 C at the end of the forecast interval. (Figure B3b)

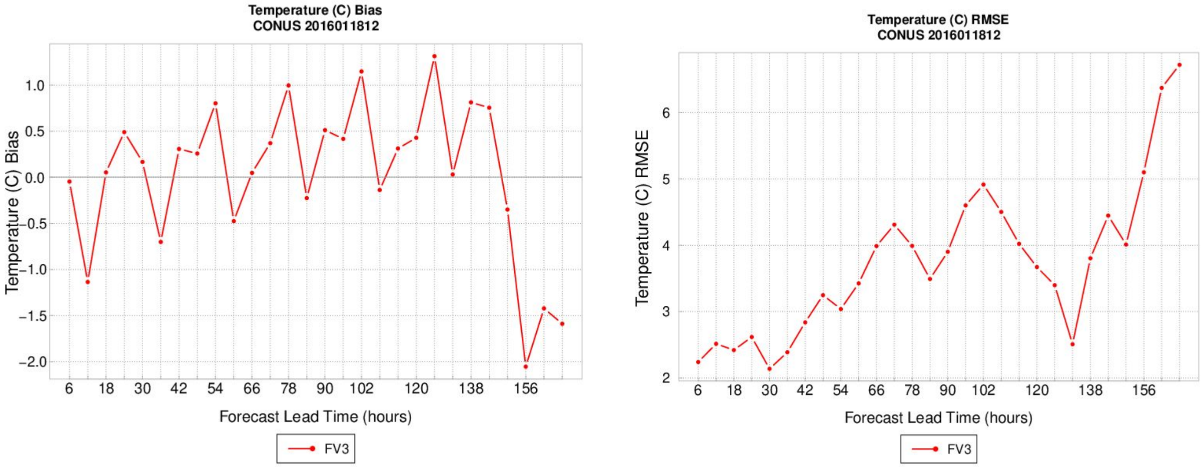


Figure B3: (a) Mean error for 2-m temperature and (b) root mean squared error for 2-m temperature.

Specific Humidity

The specific humidity bias has small negative values, <1.e-04 kg_kg, up to forecast hour 48, at which point the bias moves to positive territory with values of up to >4.e-04, and significant variation. A trace of a small diurnal variation is also discernible. (Figure B4a) The overall RMSE for the specific humidity has a steady increase until forecast hour 78. Two characteristic peaks appear at forecast hours 78 and 108, followed by a big drop around forecast hour 126, and further increase afterwards until the end of the forecast period. (Figure B4b)

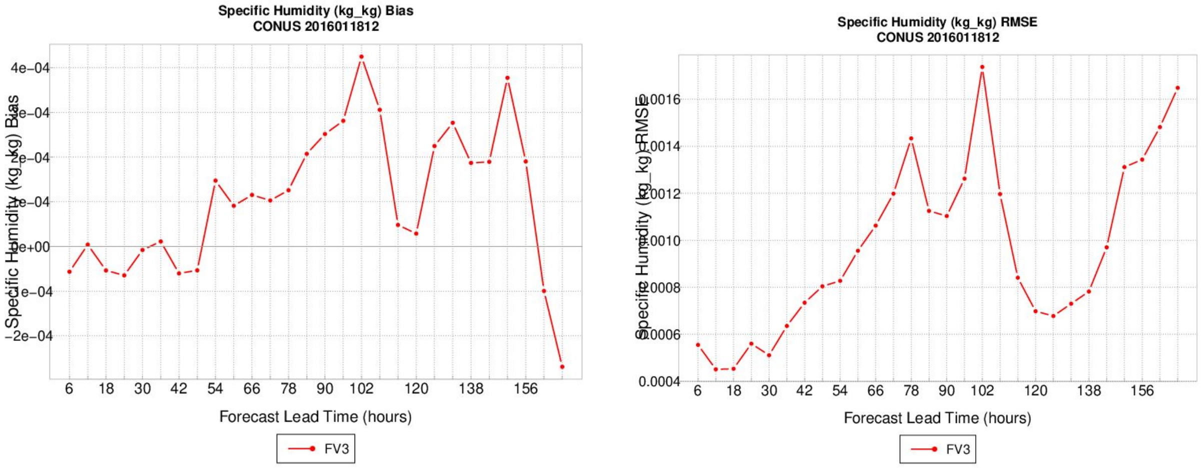


Figure B4: As in Figure B3, but for surface specific humidity.

Wind Speed

A fast bias with values ranging from 0.2 - 1.2 m/s is found in the Wind bias plot (Figure B5a). A diurnal signal is present with varying magnitude for the duration of the forecast period. In general, the bias values increase with forecast lead time, being above 0.6 m/s at the end of the forecast interval. We don't detect any considerable diurnal cycle trend in the Wind RMSE (Figure B5b). The RMSE values slowly rise with forecast lead time, varying between 1.7 and 2.9 m/s. The RMSE values remain at high levels, >2.4 m/s, after forecast hour 96.

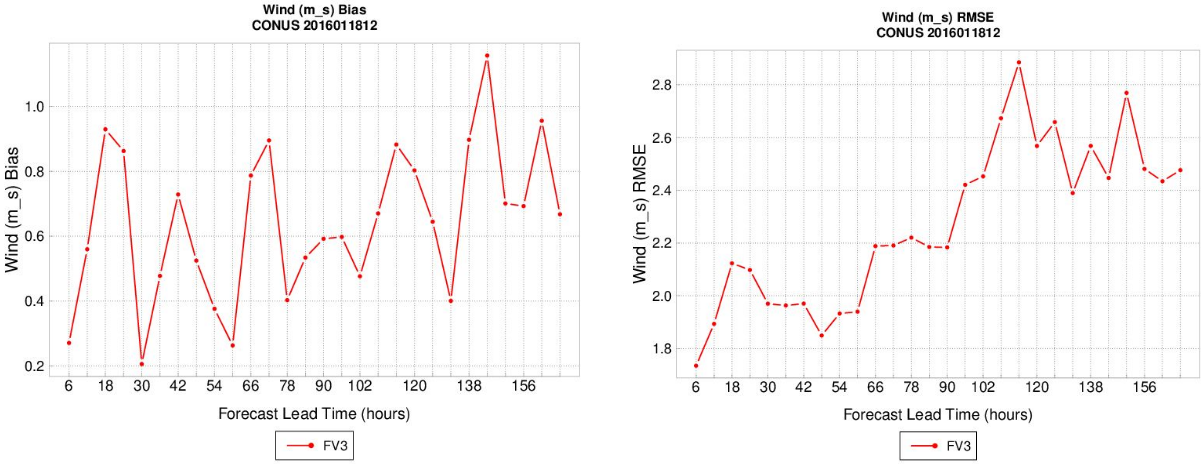


Figure B5: As in Figure B3, but for 10-m wind speed.

Upper Air Verification

Six forecast hours are examined for upper air analysis: 12, 24, 60, 96, 120, and 144. These six were chosen to represent model behavior at the short, medium, and long-range forecasts.

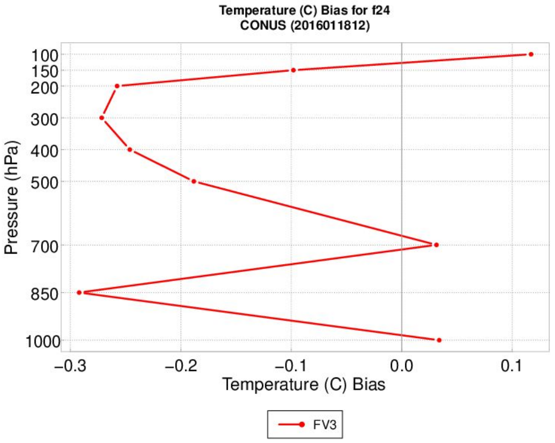
CONUS

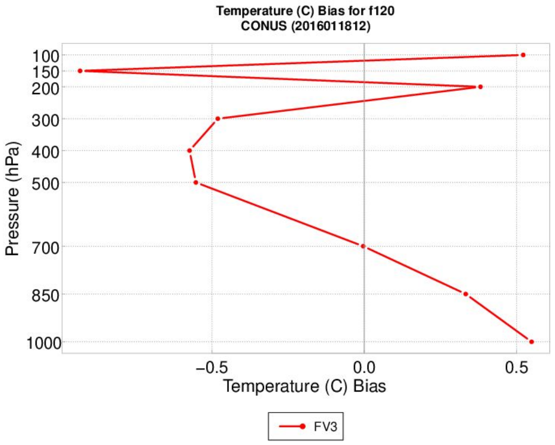
Temperature

The general trend for mean error for all forecast lead times is as follows:

For the short range forecast, cool bias generally exists throughout the pressure regime, with only a few small exceptions. (Figure B6a) For the medium and long range forecasts, the bias behaves differently from the short range forecast. A warm bias occurs from the surface to 850-700mb, which then switches to cold bias for most of the remaining pressure range, all the way to the upper troposphere (~125mb). (Figure B6b)

The overall short and medium-range forecasts have the RMSE decreasing with height from the surface until mid to upper levels (500 - 300mb), and then increasing again, exhibiting a distinctive peak at 200mb (Figure B6c). For the long-range forecasts, the RMSE appears to have significant variations, with higher values at lower pressures, and a characteristic secondary peak at ~200mb. (Figure B6d)





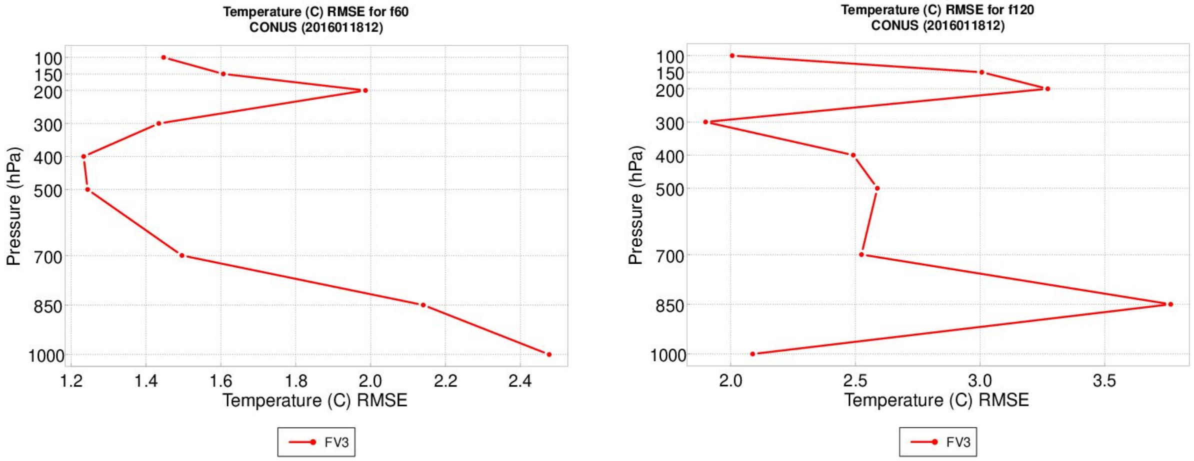


Figure B6: (a) mean error for temperature at pressure levels for forecast lead time 24, (b) as in (a) but for forecast lead time 120, (c) root mean squared error for temperature at pressure levels for forecast lead time 60 and (d) as in (c) but for forecast lead time 120.

Relative Humidity

A moist bias is present that increases steadily with pressure level for all forecasts (Figure B7a). The moist bias ranges from 0 - 35%. All RMSE plots show increases with pressure level (Figure B7b), with only a few small exceptions. The RMSE values range from 15 - 45%.

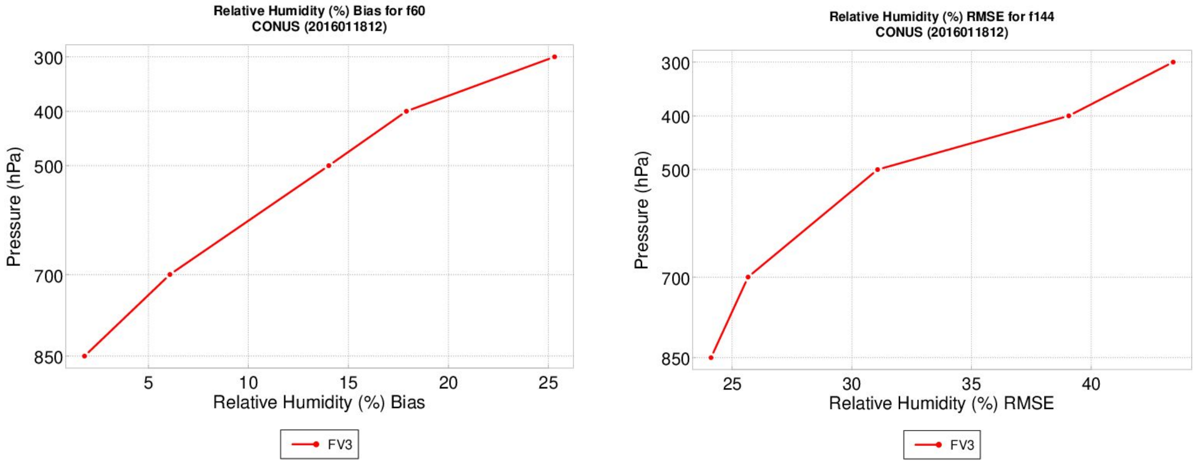


Figure B7: (a) mean error for relative humidity at pressure levels for forecast lead time 60 and (b) root mean squared error for temperature at pressure levels for forecast lead time 144.

Wind Speed

The short range forecasts behave opposite to the medium and long range forecasts. For the short range forecasts, fast bias occurs at the lower pressure level while the slow bias occurs at the higher pressure level. The bias range is between -2.0m/s and 1.0m/s. (Figure B8a) For the medium and long range forecasts, we found slow bias at the lower pressure level while the fast bias is found at the higher pressure level. The bias range is between -2.0m/s and 8.0m/s. (Figure B8b)

For all forecast lead times, the RMSE increases from the surface to 300 - 200mb, and then decreases all the way to the top. The RMSE values are between 2m/s to 12m/s with the larger magnitude for the later forecast lead times. (Figure B8c-d)

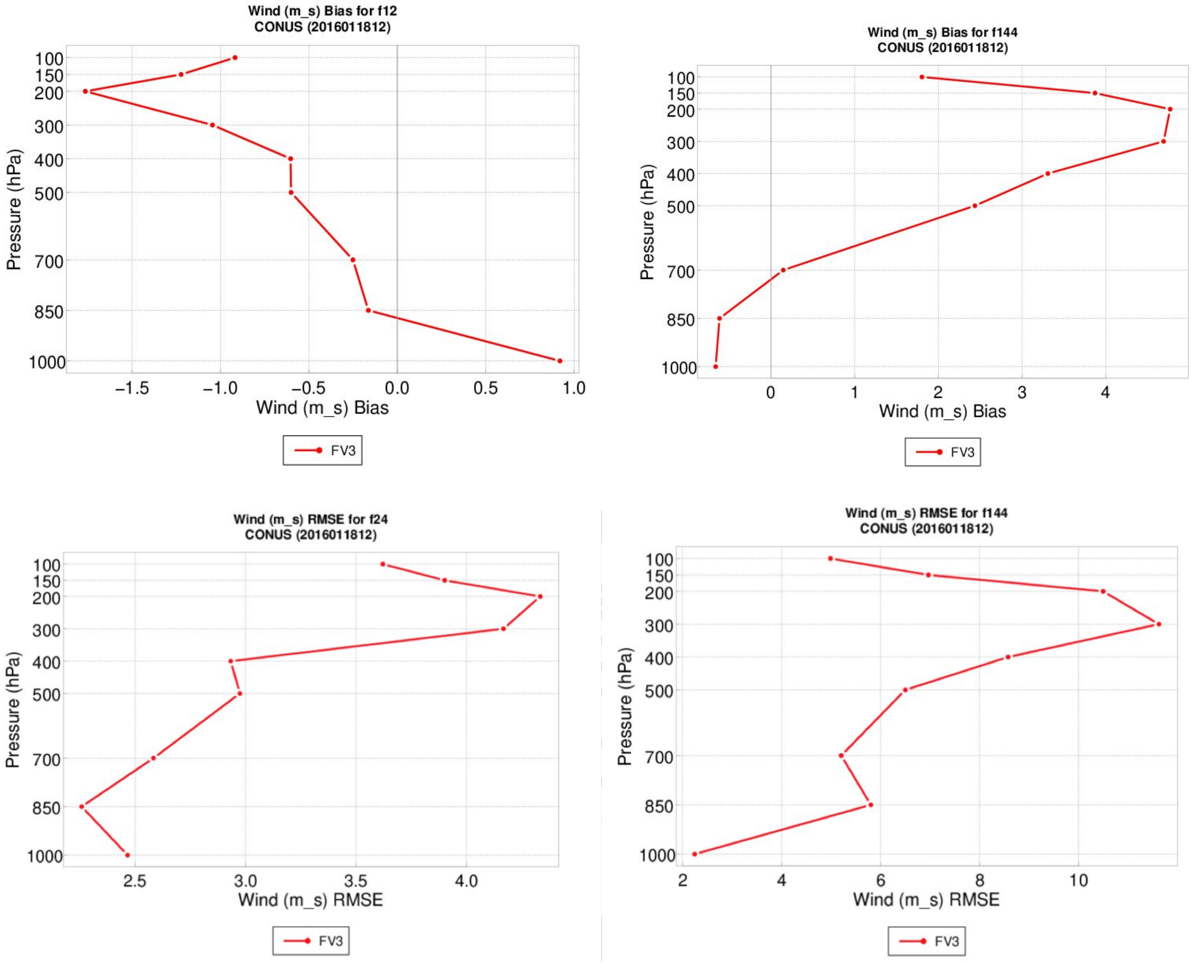


Figure B8: As in Figure B6, but for wind speed at (a) forecast lead time 12, (b) forecast lead time 144, (c) forecast lead time 24 and (d) forecast lead time 144.

Global

Temperature

For all range forecasts, cool bias is observed throughout the pressure regime, with only very few exceptions. The cold bias increases with the forecast time. The shorter range forecasts have smaller the cold bias. The short range forecasts have maxima at -0.35C (Figure B9a), the medium range forecasts have maxima at -0.7--0.8, and the long range forecasts have maxima at larger than -1.0C (Figure B9b).

All forecasts have the RMSE decreasing with height from the surface until mid to upper levels (700 - 300mb) (Figure B9c). For the long-range forecasts, the RMSE appears to have significant variations, with a characteristic peak at ~200mb, much stronger than in the medium and short range forecasts. (Figure B9d).

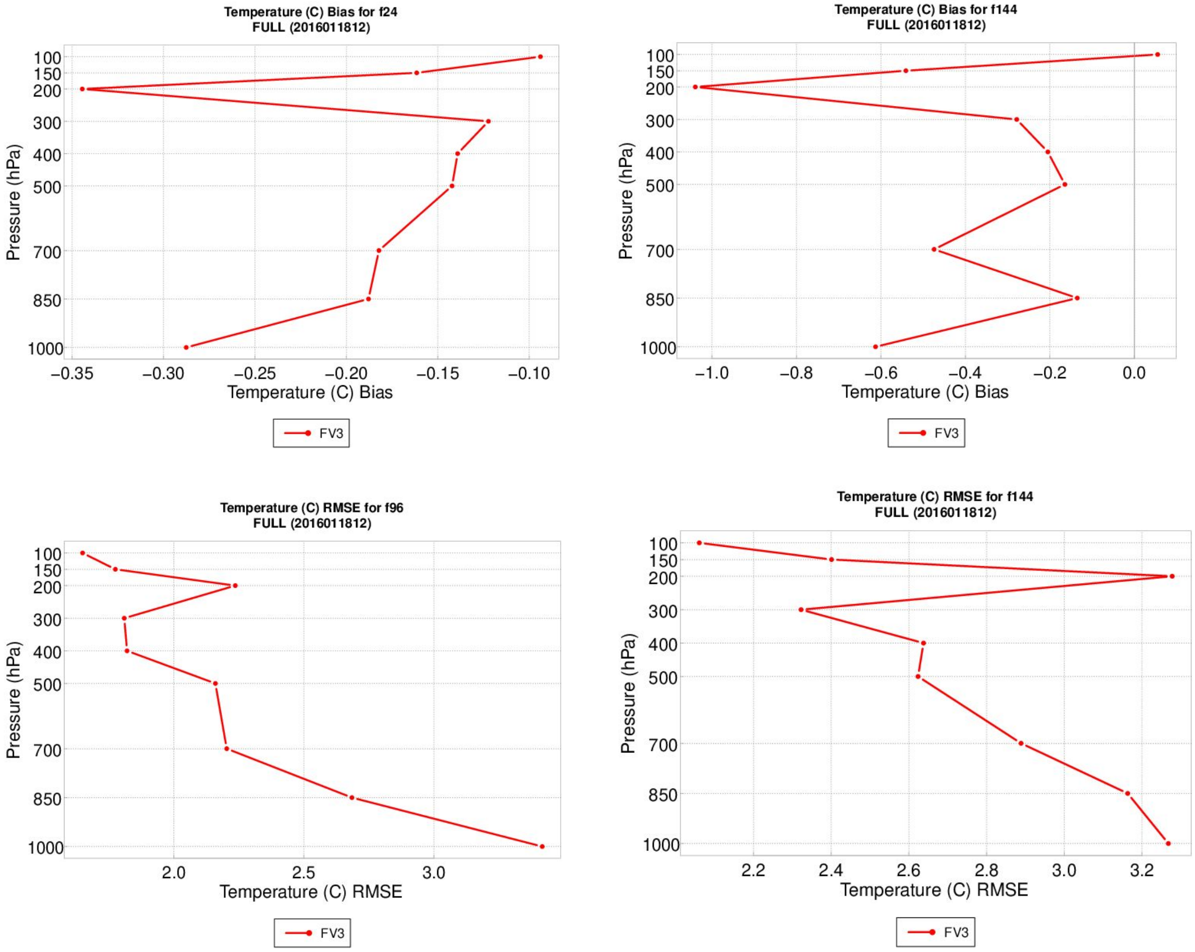


Figure B9: Global temperature (a) mean error at pressure levels for forecast lead time 24, (a) mean error at pressure levels for forecast lead time 144, (d) root mean squared error at pressure levels for forecast lead time 96 and (d) root mean squared error at pressure levels for forecast lead time 144.

Relative Humidity

Like in the CONUS case, a moist bias is present that increases monotonically with pressure level for the short and medium range forecasts (Figure B10a). For the long range forecasts, the moist bias reaches a peak at 400mb (Figure B10b). The moist bias ranges from 0 - 20%. Likewise, the short and medium range forecasts exhibit monotonic RMSE increases with pressure level (Figure B10c), while the long range forecasts have again a peak at 400mb (Figure B10d). The RMSE values range from 15 to 36%.

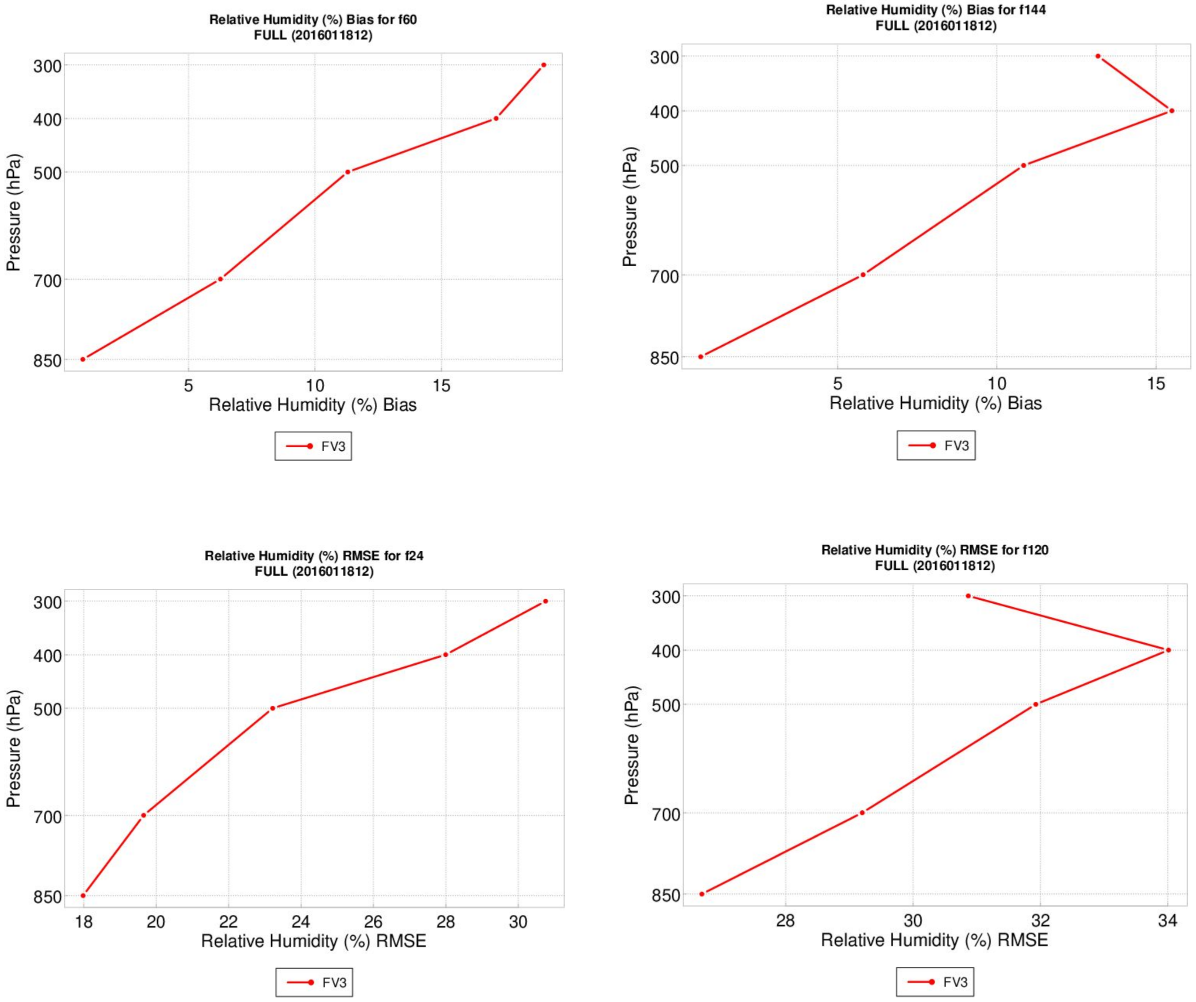


Figure B10: As in Figure B9, but for relative humidity at forecast lead time (a) 60, (b) 144, (c) 24 and (b) 120.

Wind Speed

For all forecast lead times (Figure B11a-b), fast bias occur at lower pressure levels, turning into slow bias above 900mb, which increase significantly at higher pressure levels, occasionally exhibiting a peak near 200mb. The bias range is between -2.5m/s and 0.5m/s.

For all forecast lead times (except for hour 12), the RMSE increases from the surface to 300mb, and decreases at higher altitudes. The RMSE values are between 2m/s to 13m/s, with larger magnitude for the later forecast lead times. Forecast hour 12 has a constant RMSE value wit height, 2-3m/s, except for a characteristic peak at 300mb. (Figure B11c-d)

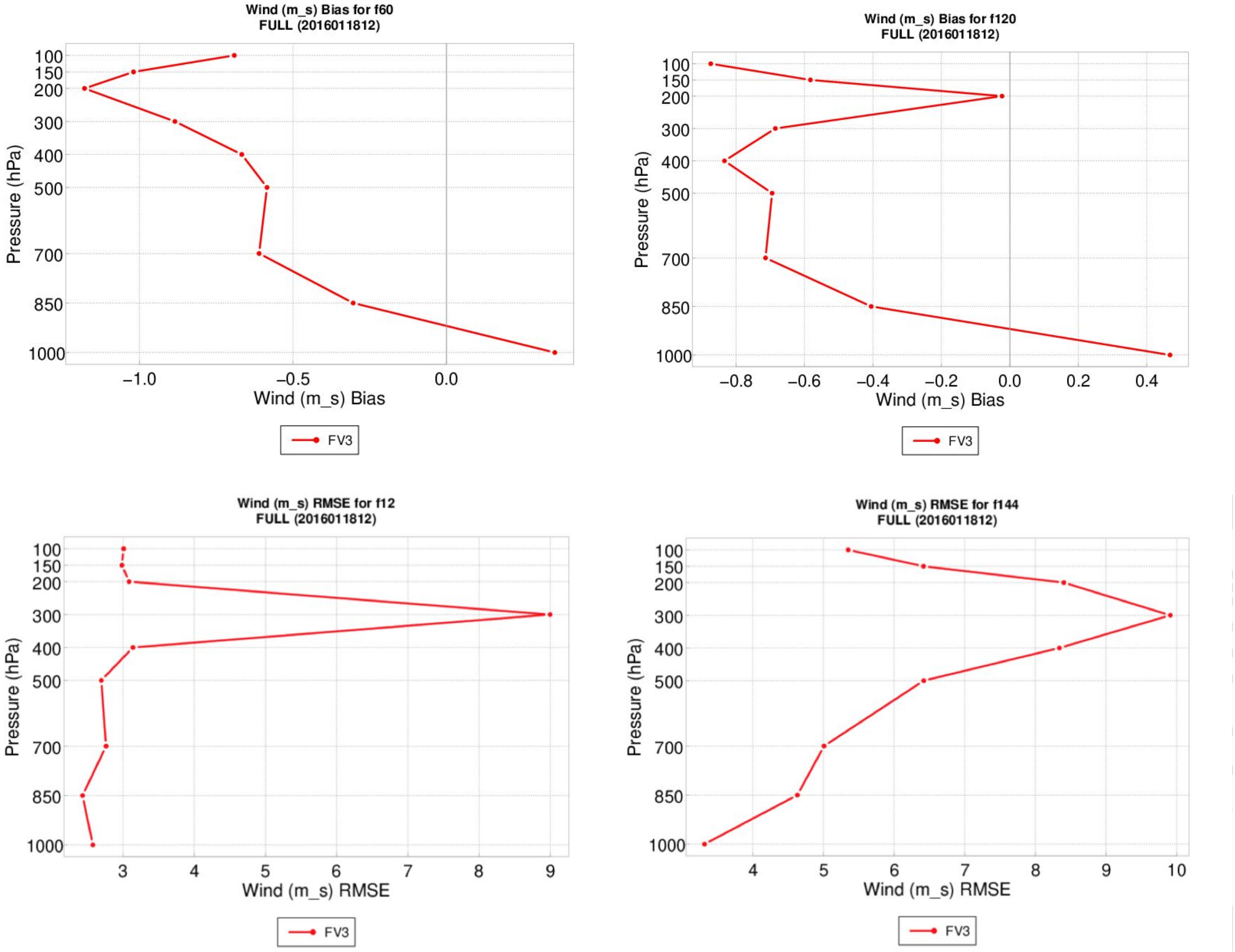


Figure B11: As in Figure B9, but for relative humidity at forecast lead time (a) 60, (b) 120, (c) 12 and (b) 144.

Precipitation Verification

Accumulated 24 hour precipitation is examined in this section. Three forecast lead times were chosen to examine model performance, hours 96, 120 and 144. Three precipitation thresholds were chosen to examine model performance for light (>=6.35mm/0.25 inches), moderate (>=12.7mm/0.5 inches), and heavy (>=25.4mm/1 inch) accumulations.

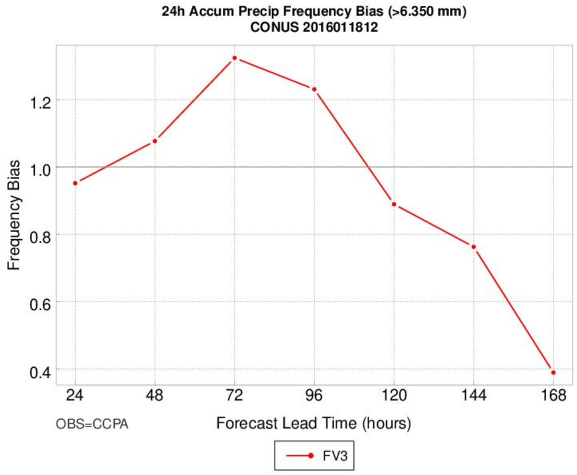
CONUS

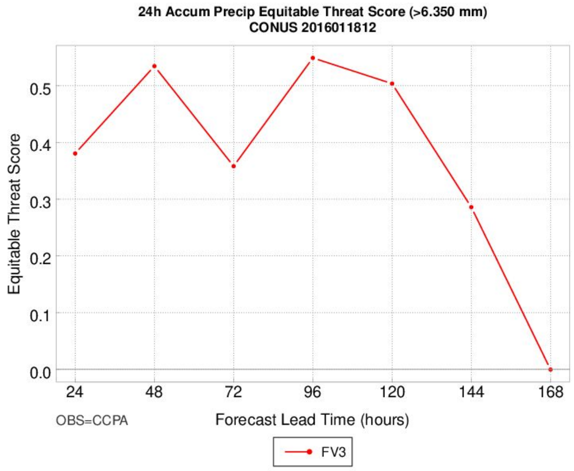
Model performance over CONUS is compared to the gridded CCPA dataset.

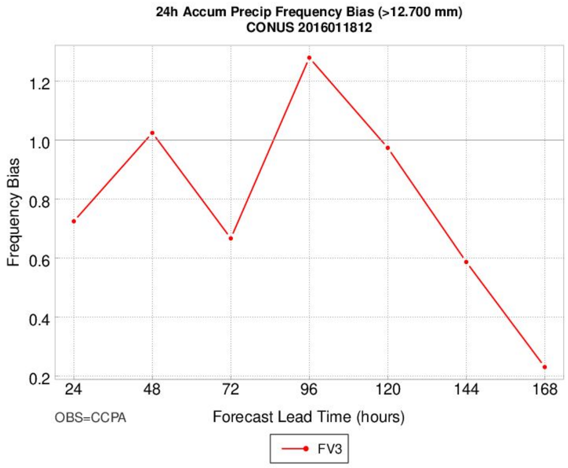
Analysis by Lead Time

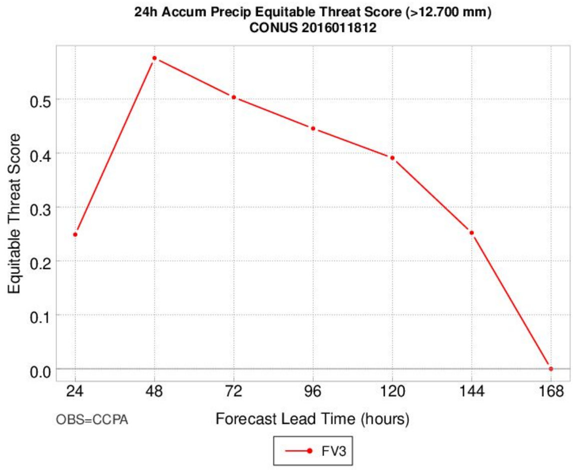
Frequency bias is the ratio of the frequency of forecast events to observed events. Values <= 1 indicate an under-forecast, values >= 1 indicate an over-forecast, and a value of 1 is a perfect forecast. All three accumulations have very different trends: For the light accumulation (Figure B12a), over-forecast occurs until forecast hour 96, with under-forecasts manifesting for the rest of the forecast period. For the moderate accumulation (Figure B12b), under-forecasts are prevailing for the entire forecast interval with the exception of an over-forecast at hour 96. Notice also the two near-perfect forecasts at hours 48 and 120. For the heavy accumulations (Figure B12c), significant under-forecasts dominate the entire forecast interval, with the exception of forecast hours 96 and 120, when a small over-forecast occurs.

Equitable Threat Score (ETS) measures the fraction of the observed events that were correctly predicted. ETS values range from 0 to 1. A perfect score is 1, so the higher the ETS value the better the model performed. Moderate to high ETS values (0.35 - 0.5) prevail at the beginning of the forecast period and up to hour 120 for the light accumulations, followed by a sharp decrease with forecast lead time afterwards (Figure B12d). For the moderate accumulations (Figure B12e), a very high ETS value (~0.6) appears at forecast hour 48, followed by a slow decrease to ETS 0.4 by hour 120, and a more rapid decline for the remaining of the forecast interval, reaching ETS value of 0 at hour 168. For the heavy accumulations (Figure B12f), a significantly lower ETS value (<0.35) is observed throughout the forecast interval, demonstrating significant variations, with peaks of higher than 0.3 ETS at hours 48 and 120.


24h Accum Precip Frequency Bias (>6.350 mm)
CONUS 2016011812


24h Accum Precip Equitable Threat Score (>6.350 mm)
CONUS 2016011812


24h Accum Precip Frequency Bias (>12.700 mm)
CONUS 2016011812


24h Accum Precip Equitable Threat Score (>12.700 mm)
CONUS 2016011812

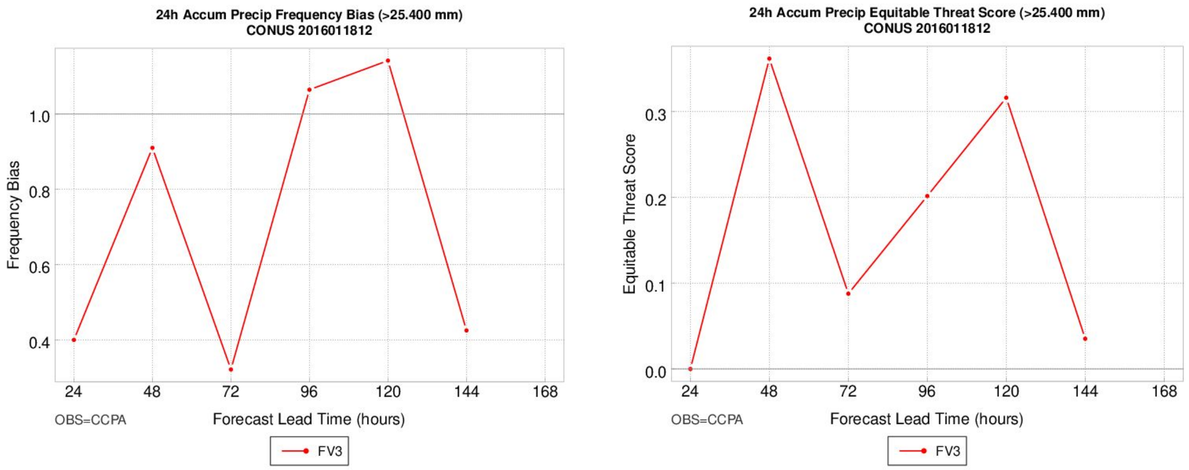


Figure B12: Frequency bias for 24 hour accumulated precipitation thresholds of (a) >6.35mm, (b) >12.7mm, and (c) >25.4mm throughout the forecast period. Equitable threat score for 24 hour accumulated precipitation thresholds of (d) >6.35mm, (e) >12.7mm, and (f) >25.4mm throughout the forecast period.

Analysis by Threshold

Each forecast range shows different frequency bias with threshold. At forecast hour 96, the model over-forecasts for all precipitation thresholds. It has the lowest bias of a little more than 1 for the >25.4 and >31.75mm thresholds, with the highest values at the high end of the threshold range. At hour 120 lead time, the model slightly under-forecasts for the smaller precipitation thresholds (>6.35, >8.89, and >12.7mm), switching to over-forecasts for all thresholds greater than >25.4mm, monotonically increasing to a maximum of >2 at the highest threshold. At forecast hour 144, the model significantly under-forecasts all the thresholds, with a minimum of ~0.4 at >25.4mm, demonstrating a single over-forecast at the >38.1mm threshold, with no bias reported for any threshold above. (Figure B13a-c)

All examined forecast lead times show a decrease of the ETS with precipitation threshold, albeit at different rates and/or initial values. For forecast hour 96, the ETS value is at >0.5 at the lowest threshold, >6.35mm, decreasing slowly throughout the precipitation threshold range, with the exception of a sharp drop of ~0.25 at threshold >25.4mm, reaching below 0.1 ETS value at the highest threshold. The 120 hours forecast begins a little lower, ETS at 0.5, and declines with a steady rate for all thresholds concluding at ~0.15 ETS values at the heaviest accumulation considered. The long range forecast, 144 hours lead time, starts at below 0.3 ETS, diminishing slowly for the light/moderate accumulation thresholds, up to >12.7mm, at which point it suddenly falls to near-zero ETS value for the heavy accumulation of >25.4mm, and then to 0 for the heavier accumulations. (Figure B13d-f)

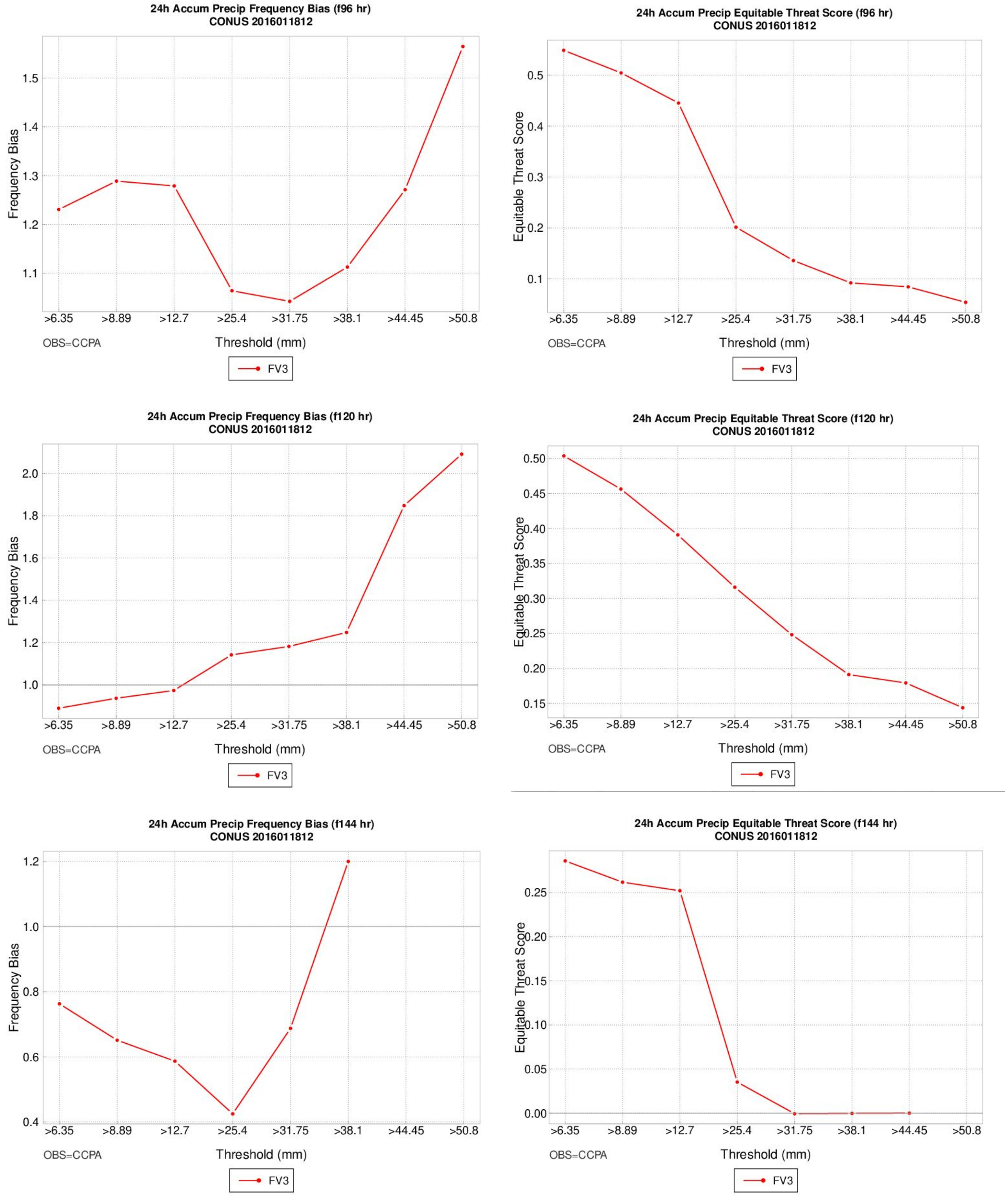


Figure B13: Frequency bias for all thresholds at forecast hours (a) 96, (b) 120, and (c) 144. Equitable threat score for all thresholds at forecast hours (d) 96, (e) 120, and (f) 144.

Global

Global model performance is compared to the gridded CMORPH dataset.

Analysis by Lead Time

The light accumulation frequency bias exhibits an over-forecast for the entire forecast range (Figure b14a). A decreasing trend with lead time is also present, with bias registering above 1.2 at lower lead times, <96 hours, moving clearly below 1.2 for higher lead times. For the moderate accumulation conditions, a variation around 1 is evident, within 0.9 - 1.1 bias values (Figure B14b). At first we see mostly small over-forecasts, lead time <96 hours, which then turn to predominantly under-forecasts at later forecast lead times. Finally, the heavy accumulation case has frequency biases exclusively in the 0.7 - 0.9 range (Figure B14c). The clear under-forecast is at its highest point, ~0.85, at forecast hours 72 and 96, but dips below 0.75 at lead times 120 and 144 hours.

All three precipitation threshold cases have generally decreasing ETS trends with forecast lead time. The highest initial ETS value (~0.32) occurs for the light accumulation conditions, compared to the moderate (ETS~0.26), and heavy (ETS~0.18) accumulation conditions. The declining rate is relatively steady for the light accumulations (Figure B14d), while the moderate and heavy accumulations exhibit a local peak around forecast hours 72 and 96 (Figures B14e and B14f).

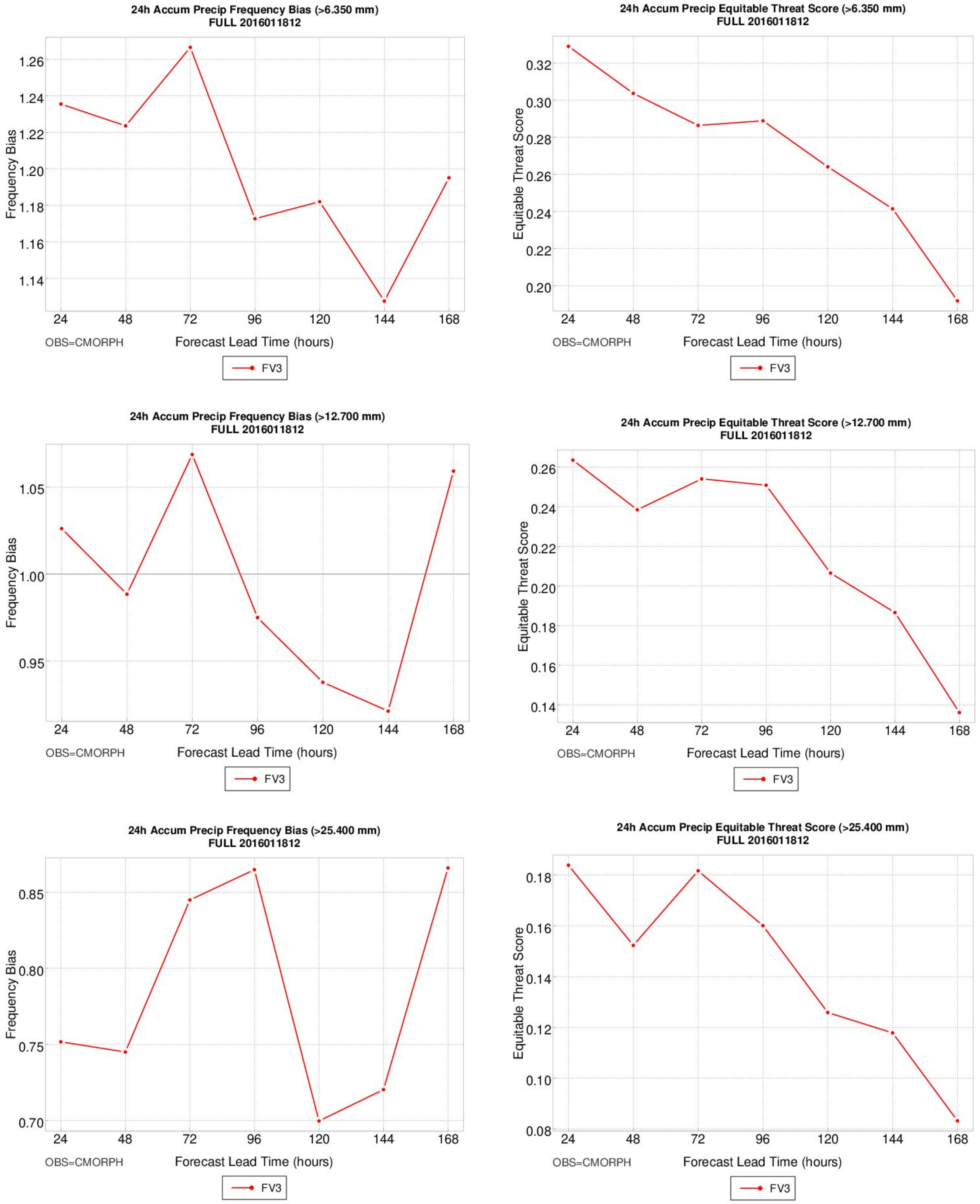


Figure B14: Global frequency bias for 24 hour accumulated precipitation thresholds of (a) >6.35mm, (b) >12.7mm, and (c) >25.4mm throughout the forecast period. Global equitable threat score for 24 hour accumulated precipitation thresholds of (d) >6.35mm, (e) >12.7mm, and (f) >25.4mm throughout the forecast period.

Analysis by Threshold

For the global model performance, all range forecasts manifest over-forecasts for the 24-hour accumulated precipitation thresholds >6.35mm and >8.89mm, and under-forecasts for the precipitation thresholds >12.7mm and >25.4mm (Figures B15a-c). The 96 hour forecast varies from bias >1.15 to ~0.85, while the two higher forecast lead times have similar initial bias but drop further to as low as 0.7 frequency bias. No bias is reported for heavier accumulation conditions.

Similar behavior is observed with the ETS values for all three forecast hours, decreasing as the precipitation threshold increases (Figures B15d-f). The model shows better performance at the lowest accumulation threshold, with ETS values rapidly decreasing with increasing threshold. The overall magnitude of the ETS shows a small decreasing trend with increasing forecast hour. There is no ETS measurement for accumulated precipitations of 31.75mm and above.

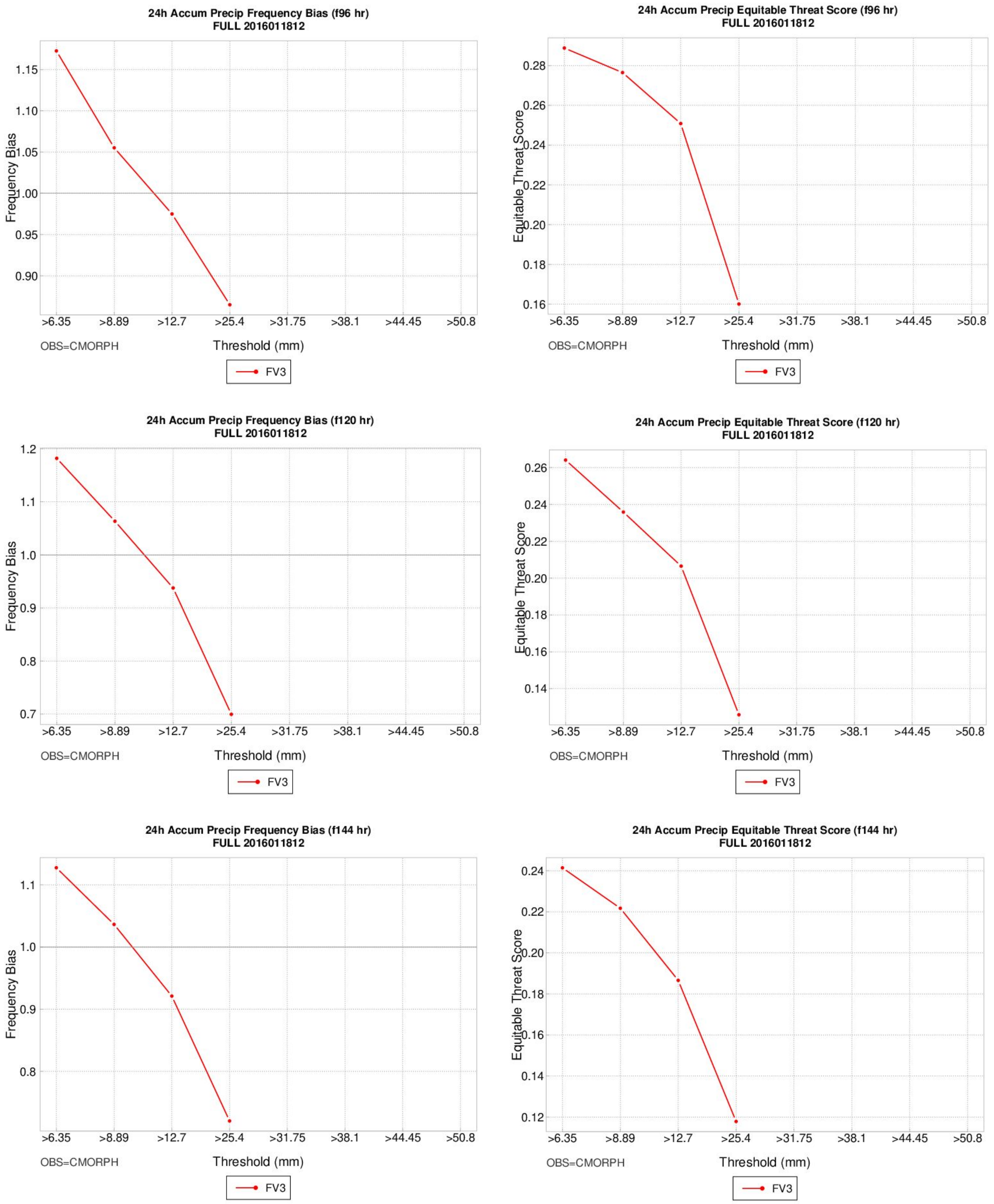


Figure B15: Global 24 hour accumulated precipitation frequency bias for all thresholds at forecast hours (a) 96, (b) 120, and (c) 144. Global equitable threat score for all thresholds at forecast hours (d) 96, (e) 120, and (f) 144.

5. Hurricane Matthew

5.1 Synoptic discussion

Hurricane Matthew was a category 5 hurricane that developed east of the Lesser Antilles as a tropical cyclone on September 28th, 2016, becoming a hurricane the next day. It became a category 5 storm on October 1st, then headed north, making landfall over extreme western Haiti and eastern Cuba on October 4th. Widespread damage, costing billions of US dollars, was reported in both countries. Despite weakening, Hurricane Matthew began impacting the United States on October 6th and brought wind damage and power outages to coastal areas of Florida and Georgia. The storm also caused massive freshwater flooding to areas of the Carolinas and Virginia.

Hurricane Matthew underwent rapid intensification west of the Lesser Antilles, increasing from a category 1 storm to category 5 in just two days. This explosive development was possible due to the storm moving into an area dominated by a mid-level ridge, combined with a decrease in southwesterly wind shear. Matthew moved north around the ridge and eventually weakened some due to cooler sea surface temperatures and interactions with the mountainous terrain of Cuba and Haiti. Significant weakening was then seen as the storm paralleled the Florida Georgia coastlines.

In addition to upper-level troughs in the FV3 being shown to progress eastward too quickly, hurricane forecasts have been found to suffer from a similar bias, including Hurricane Matthew. In the parallel FV3GFS simulation, Matthew was shown to gain latitude too quickly as it moved north, generating significant sea-level pressure and track errors. Tropical cyclones in the FV3 have been found to extend deeper into the troposphere/stratosphere, which could potentially expose them to stronger steering winds, providing a possible explanation for this progressiveness

5.2 Verification results

Based on qualitative comparisons of a number of basic meteorological fields such as sea level pressure, geopotential height, and surface/upper-level winds, the FV3 simulation brings Matthew westward too quickly after passing the Lesser Antilles, placing the storm too far west prior to turning north. This progressiveness results in Hurricane Matthew moving north over Jamaica, while in reality, it was hundreds of miles to the east. For example, at 120 hours into the forecast initialized at 00Z on 29 September 2016, the FV3 places the lowest mean sea level pressure between Jamaica and Cuba, whereas the storm at this time was making landfall in extreme western Haiti, indicating a northwestward progressiveness bias (Fig Xa). Even at 108 hours into the forecast, when analyzing the 10-m wind speed forecast, the FV3 simulation already has Matthew impacting Jamaica (Fig Xb), while at this time, the true location of the storm was still south of Haiti. These findings bolster the conclusions found within a number of parallel FV3GFS simulations, showing a progressiveness bias for tropical cyclones.

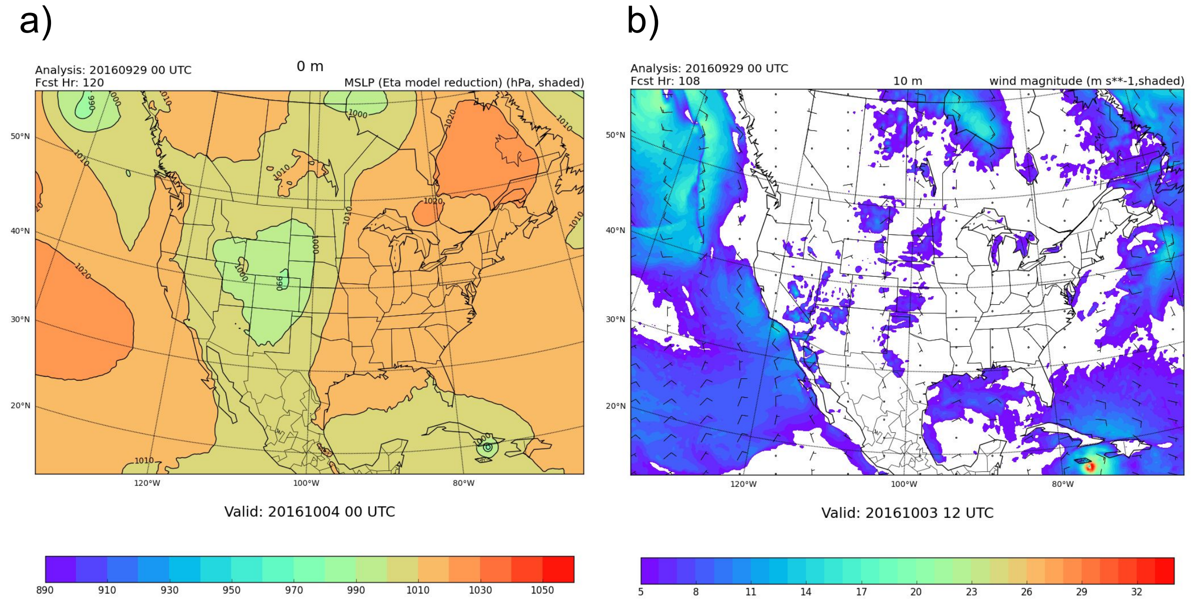


Figure H1. a) Mean sea level pressure at 120 hours and b) 10-m wind speed at 108 hours into the forecast, initialized at 00Z on 29 September 2016.

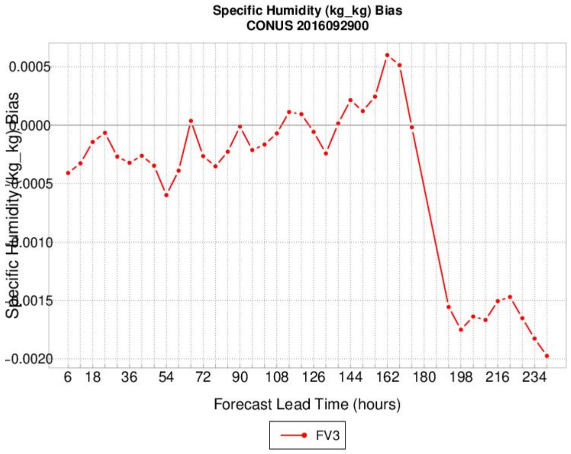
5.3 Surface and Upper Air Verification

Surface Verification (CONUS)

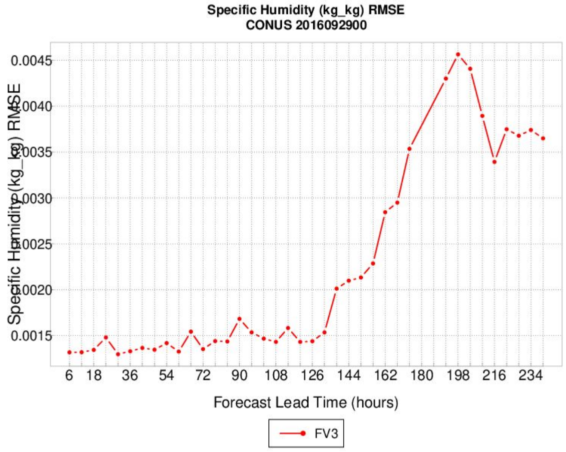
While surface verification was only available over the continental United States, it is still worthwhile to evaluate bias and RMSE for surface variables in the FV3 simulation as a function of lead time over the course of the full 10-day forecast. It is important to note that Hurricane Matthew began impacting the southeastern United States during day seven of the forecast (~168 hours); however, it's not possible to isolate its impact on the full CONUS verification. Separate, regional verification over the southeastern United States would be one option in the future to better understand the impact of Hurricane Matthew on surface CONUS verification.

Figure H2 shows both bias and RMSE for 2-m specific humidity, 2-m temperature, and 10-m wind speed over the CONUS as a function of lead time out to 240 hours. Diurnal variation is clearly apparent in the bias plots, particularly for 2-m temperature (Fig. Xc) and 10-m wind speed (Fig. Xe), and in the RMSE plots to a lesser extent. One general finding that holds true across all three surface verification variables is a more or less constant bias and RMSE out to forecast hour 126. After this time, error increases both through larger variability in bias and steadily increasing RMSE for each variable. More analysis would be necessary to pinpoint the cause of this change in error characteristics.

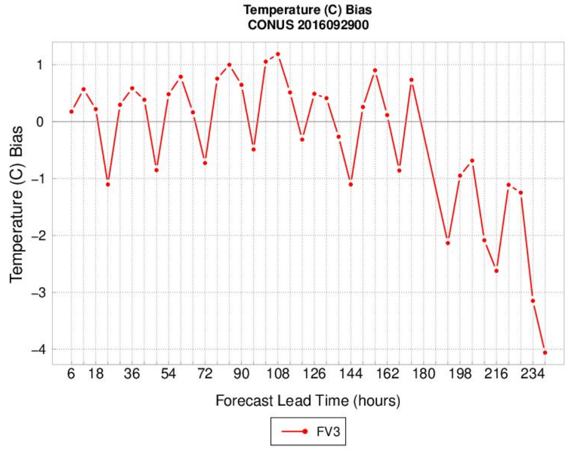
a)



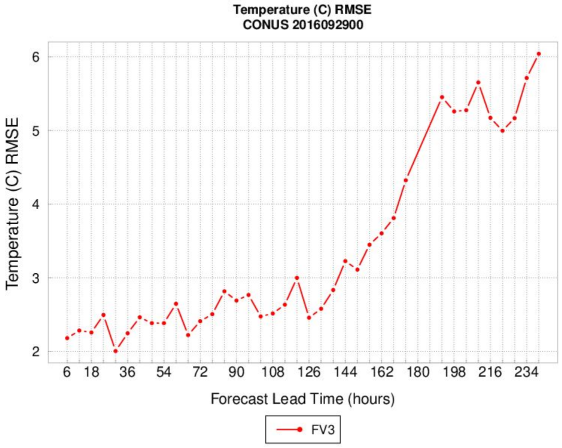
b)



c)



d)



e)

f)

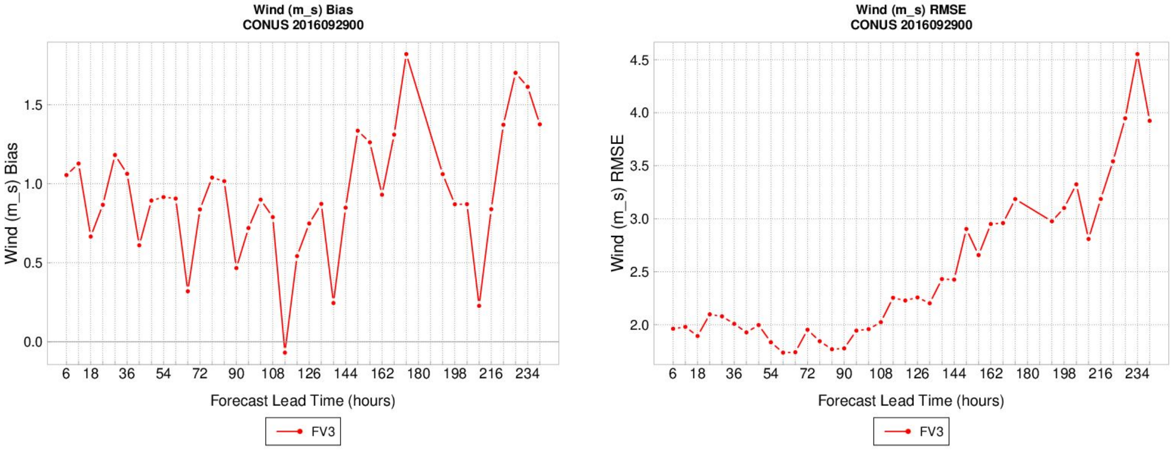
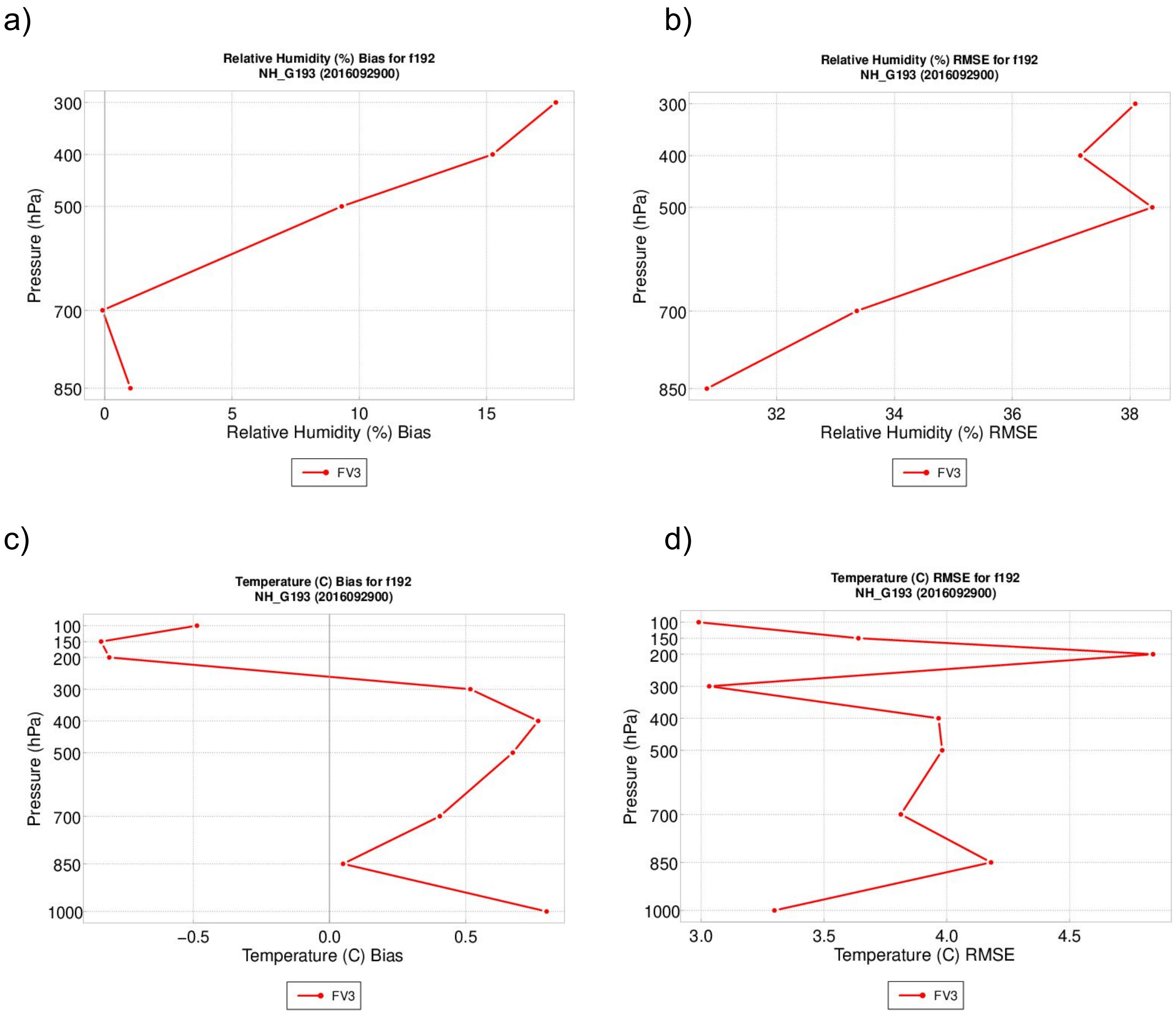


Figure H2. Bias (left) and RMSE (right) plots for 2-m specific humidity (a,b), 2-m temperature (c,d), and 10-m wind speed (e,f) over the CONUS for the 10-day FV3 forecast initialized at 00Z on 29 September 2016.



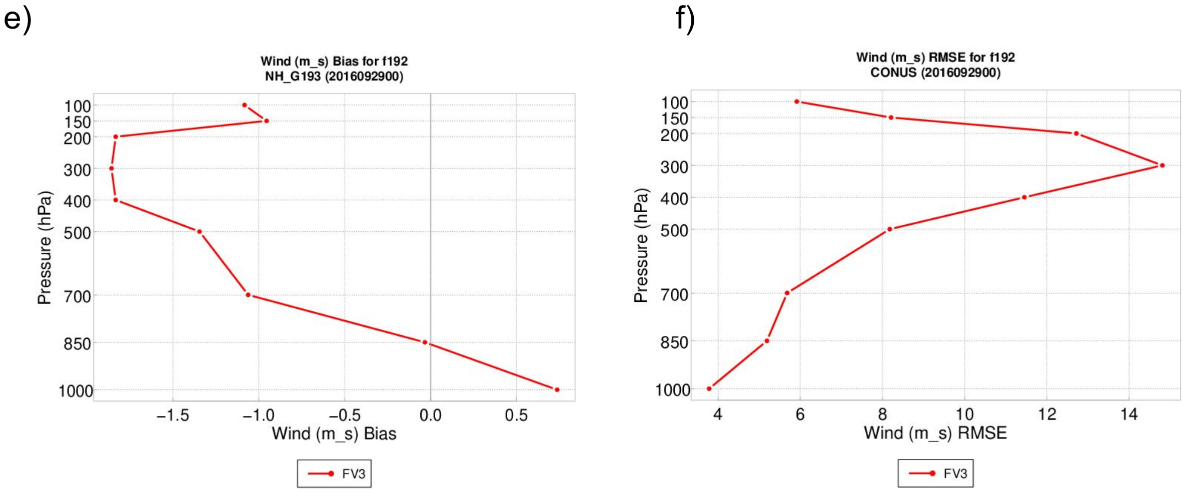


Figure H3. Vertical profiles of upper-air bias (left) and RMSE (right) for relative humidity (a,b), temperature (c,d), and wind speed (e,f) over the Northern Hemisphere for the 10-day FV3 forecast initialized at 00Z on 29 September 2016.

Upper-Air Verification (Northern Hemisphere)

Similar to with surface verification, it is difficult to highlight the impact of one phenomena (Hurricane Matthew, in this case) on vertical profile verification over a large domain. However, when analyzing upper-air verification, the northern hemisphere domain was chosen to at least include the hurricane in the verification domain itself. In addition, forecast hour 192 was used for these analyses, since this is the period when Hurricane Matthew made its largest impact on the US mainland.

Figure H3 shows vertical profiles of bias and RMSE for upper-air relative humidity, temperature, and wind speed at 192 hours into the FV3 simulation. Relative humidity shows a large moist bias, increasing with height up to the tropopause, and RMSE values as high as 35% at 500 and 300 hPa. Temperatures has a much lower (warm) bias of only ~0.5 degrees in the lower troposphere, becoming negative at pressure levels above 300 hPa. Apart from the surface, wind speeds are seen to be too low, with a peak of nearly 2 ms^{-1} at pressure levels between 400 and 200 hPa. RMSE for wind speeds increase with height up to 300 hPa where errors reach values over 14 ms^{-1} before decreasing again.

Overall, there are a number of high surface and upper-air biases/RMSE, particularly at long forecast lead times and at pressure levels around the tropopause. Further investigation of these errors would be worthwhile to identify potential causes and outline areas that could benefit from future model improvements. To this end, physics changes and other dynamics modifications could be tested through the MERIT framework using the Hurricane Matthew test case.

Precipitation Verification

Analysis of 6-h and 24-h precipitation accumulation was conducted over the CONUS region, similar to the surface verification, therefore, analysis of later lead times is more relevant, since Hurricane Matthew didn't affect the United States until late into the forecast. A threshold of > 12.7 mm was chosen to provide a moderate precipitation accumulation amount in the 24-h accumulation period. For the purpose of this report, the 168-hr forecast lead time was chosen for frequency bias analyses and for Gilbert Skill Score (GSS).

Analysis by Lead Time

Figure H4 shows 24-h precipitation accumulation frequency bias (a) and GSS (b) as a function of lead time out to 168 hours, for a threshold of > 12.7 mm. Frequency bias is nearly one at early forecast lead times, indicating that the model is correctly forecasting domain wide precipitation accumulations greater than 12.7 mm for 24 hour periods. However, after 72 hours into the forecast, a sharp drop in frequency bias occurs, indicating under-forecasting of 24-hr precipitation accumulations of > 12.7 mm until forecast hour 144, at which time the FV3 simulation over-forecasts up to 168 hours.

GSS for 24-hr precipitation accumulations of greater than 12.7 mm indicate that fewer than half of the forecast events at the 24-hr forecast lead time corresponded to observed events of 24-h precipitation accumulations greater than 12.7 mm (value of 0.4). The GSS value then drops as forecast lead time increases, with later forecast hours around 0.1, potentially indicating that a large number of misses or false alarms at certain grid points are contributing to a poor FV3 forecast in terms of GSS for precipitation accumulation.

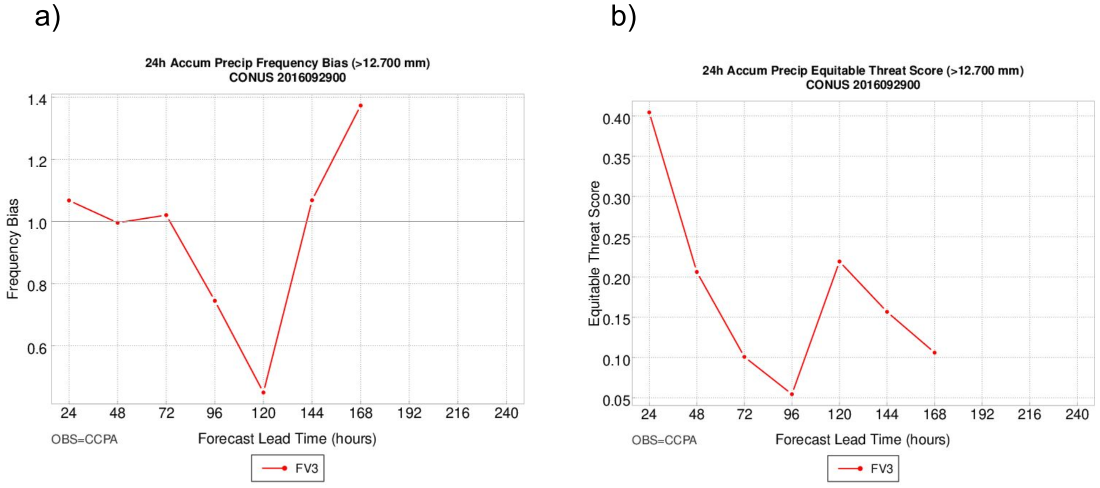


Figure H4. Frequency bias a) and Gilbert Skill Score b) as a function of lead time for a precipitation threshold of > 12.7 mm, verified over the CONUS.

Analysis by Threshold

Figure H5 shows the 24-hr precipitation accumulation CONUS frequency bias (a) and GSS (b) as a function of threshold for the 168-hr forecast lead time from the FV3 simulation initialized at 00Z 29 September 2016. Frequency bias for this lead time remains near one for threshold up to > 12.7 mm, after which an exponential increase in frequency bias can be seen for very heavy 24-hr precipitation accumulations. The FV3 is therefore producing too much heavy precipitation over the CONUS domain. The GSS for 24-hr precipitation accumulation as a function of threshold shows a sharp decrease in GSS (from an already low 0.15 for the lowest threshold of 24-hr precipitation accumulation) to zero at thresholds over > 25.4 mm. Even for low precipitation accumulation thresholds, the FV3 has little skill, however, at higher thresholds, misses and false alarms completely outweigh any hits and the model forecasts are completely unreliable for 24-hr precipitation accumulation at the 168-hr forecast lead time.

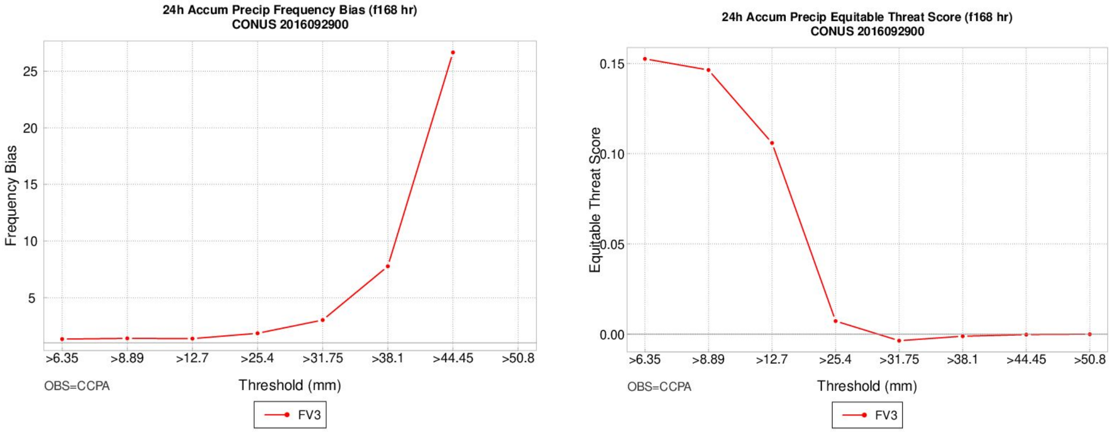


Figure H5. Frequency bias a) and Gilbert Skill Score b) as a function of threshold for the 168-hr forecast lead time, verified over the CONUS.

Diagnostics

Vortex Tracks and Cyclone Intensity

Hurricane track information was produced for FV3 by running the NOAA Geophysical Fluid Dynamics Laboratory (GFDL) vortex tracker starting at the initialization time out to 254 hours. Figure H6 shows the FV3 hurricane tracks every 6 hours compared to the actual track, marked every 12 hours. Up to 48 hours lead time, FV3 follows the actual track closely as it moves westward, diverging soon after, where it moves further west before tracking northward at a faster rate than the actual track. Along with continuing to progress too quickly, FV3 maintains a due north path instead of veering towards Florida and hugging the southeast coastline with the actual track.

Cyclone intensity is examined in terms of mean sea level pressure (Fig. Xb) and max wind speeds (Fig. Xc). While Hurricane Matthew intensified rapidly, gaining category 5 hurricane status on October 1st (day 3 of the FV3 forecast), FV3 produced only a weak low pressure system during the first few days of the forecast and beginning to intensify after day 3. For a majority of the forecast, especially the first 152 hours of the forecast, FV3 mean sea level pressure is underpredicted compared to observed. Similarly, maximum wind speed is also largely underpredicted, especially from day 2 to day 7 of the forecast lead time.

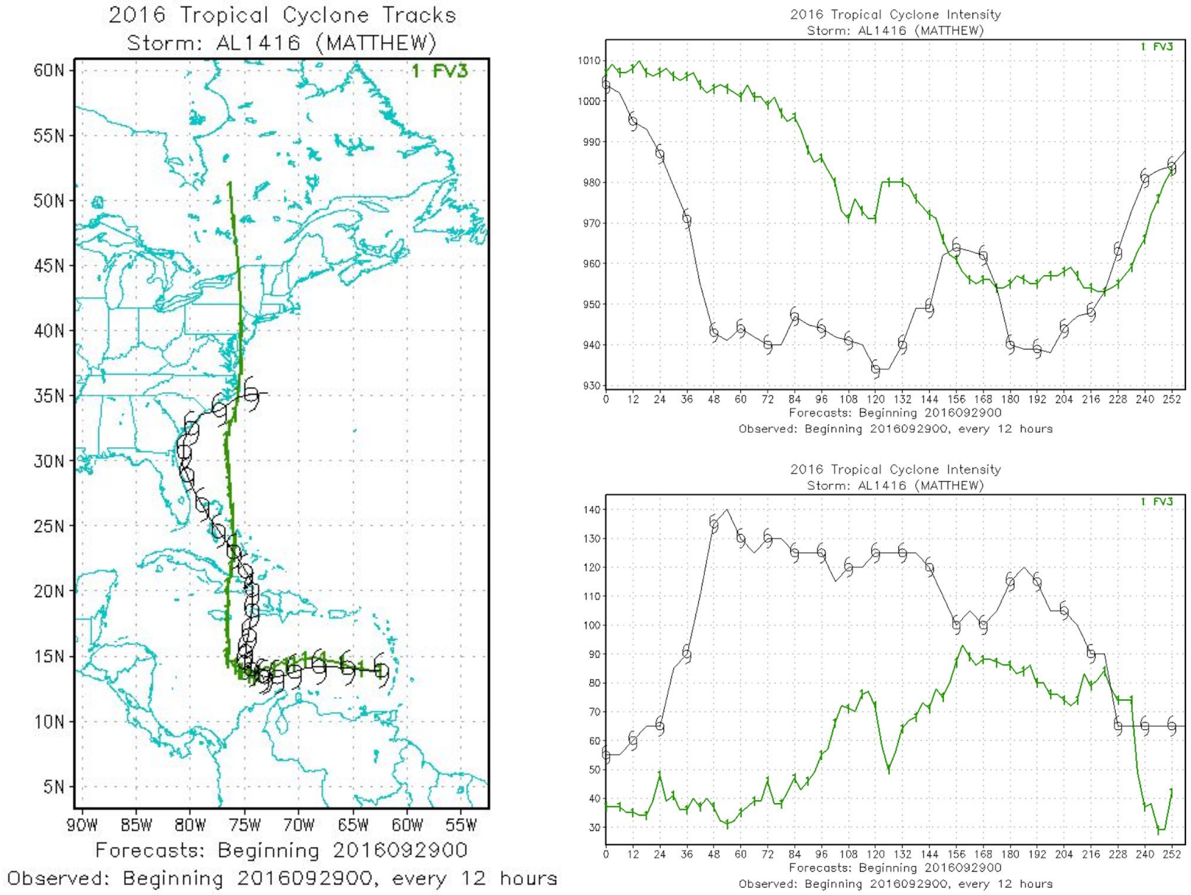
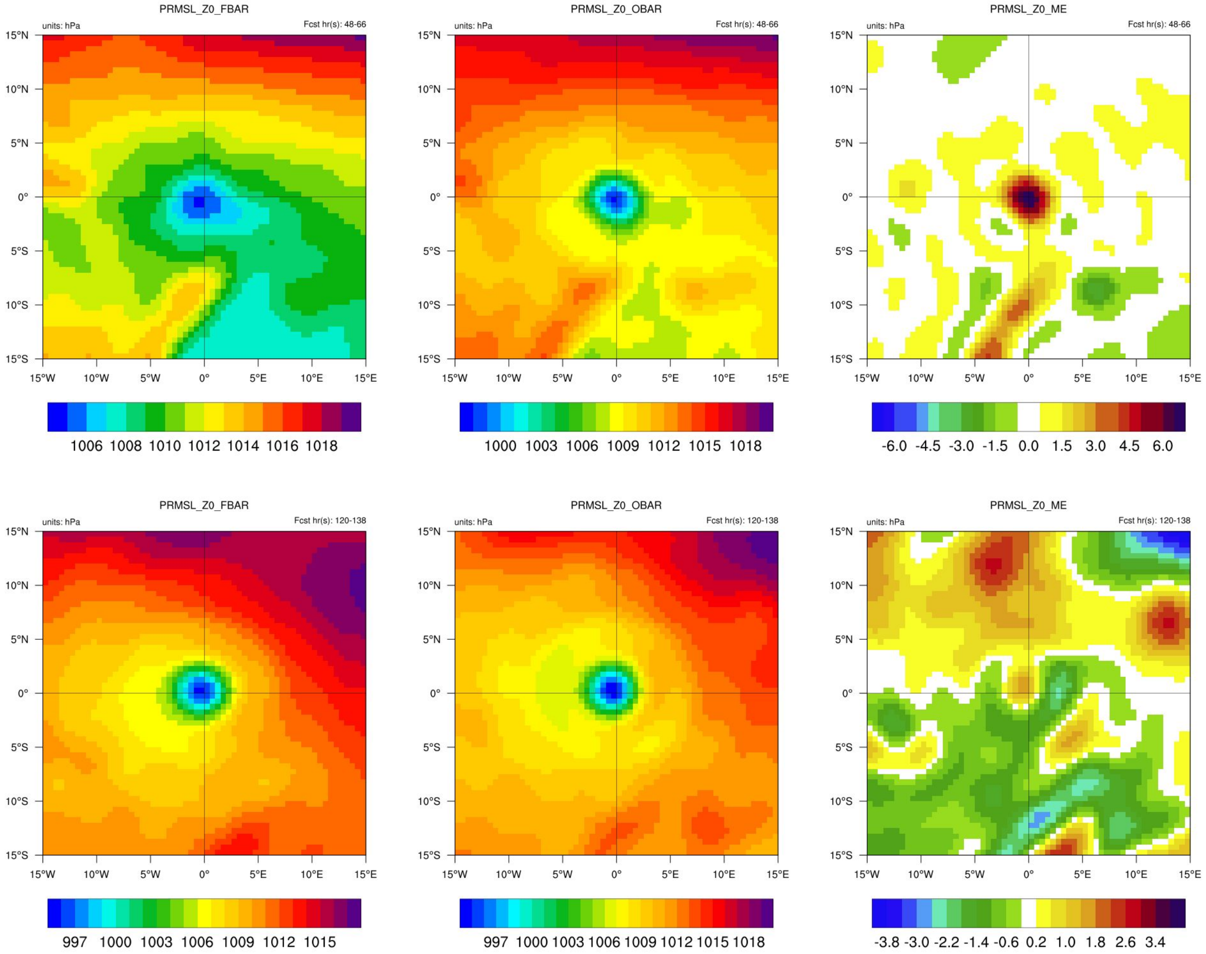


Figure H6. NOAA GFDL hurricane track information for initialization 2016092900, including (a) cyclone tracks, (b) mean sea level pressure intensity, and (c) maximum wind speed intensity for FV3 in green and actual in black. FV3 is every 6 hours and actual is every 12 hours out to 254 hours.

Feature Relative

Figure H7 shows mean sea level pressure for day 3, day 6 and day 9 with the forecast average, observation average and mean error. At the beginning of the forecast, during the rapid intensification of the observed hurricane and through to day 5 of the forecast, the FV3 model had a higher/weaker mean sea level pressure than the GFS analysis. On day 3, when the hurricane reached category 5, mean error (forecast-reference) at the center of the hurricane was the greatest with a difference of at least 7 hPa (Fig. H7c). Regions surrounding the hurricane center generally saw low biases in sea level pressure with exception to the south of the hurricane center where a tongue of high bias is observed along with a bullseye of low bias in the southeast quadrant, which is persistent during the early forecast hours. On day 6, when the observed hurricane begins to weaken as it moves northward, a bullseye of high bias at the hurricane center no longer stands out, with areas of positive biases (<3.0 hPa) observed in the northern quadrants and negative biases (>-3.0 hPa) in the southern quadrants (Fig. H7f). From day 7 through the remainder of the forecast, as the observed hurricane continues to weaken, biases at the hurricane center have transitioned to a low bias (Fig. H7i).



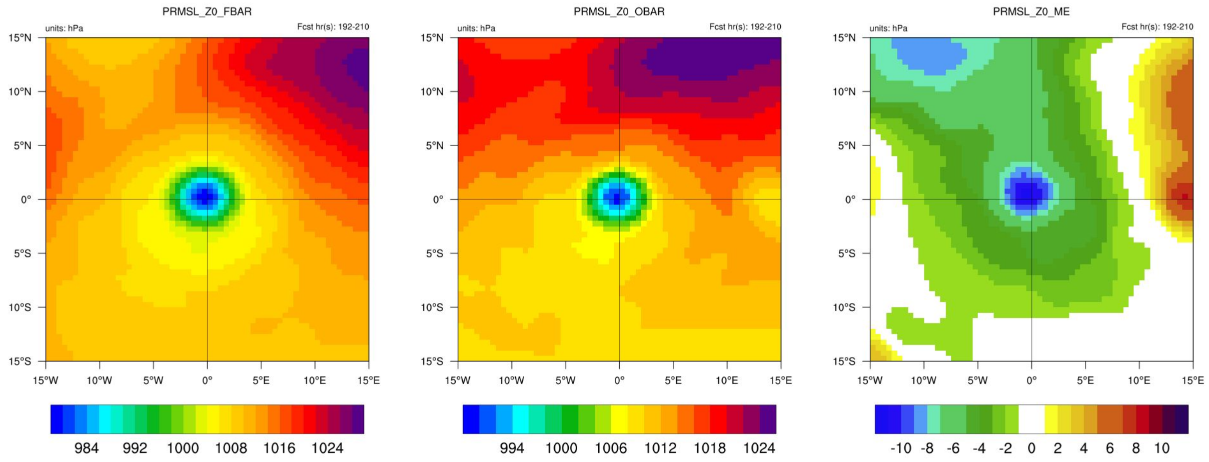
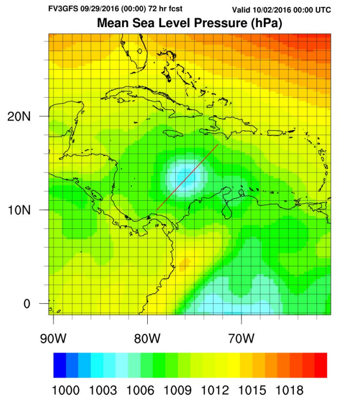
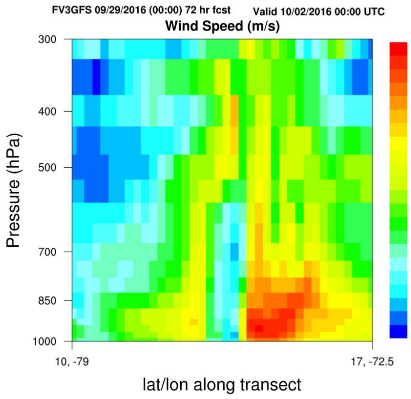
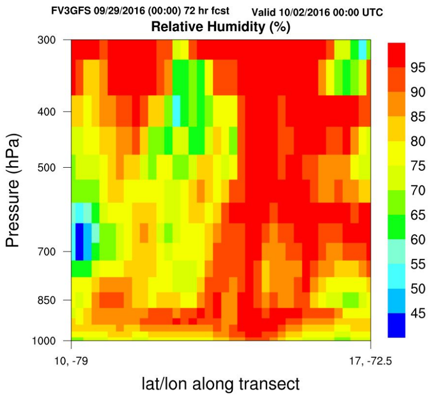


Figure H7. Mean sea level pressure in hPa relative to the hurricane center for day 3 (row 1), day 6 (row 2) and day 9 (row 3), for forecast average (column 1), observation average (column 2), and mean error (column 3).

Vertical cross-sections


FV3GFS 09/29/2016 (00:00) 72 hr fcst
Valid 10/02/2016 00:00 UTC
Mean Sea Level Pressure (hPa)
20N
10N
0
90W
80W
70W
1000
1003
1006
1009
1012
1015
1018





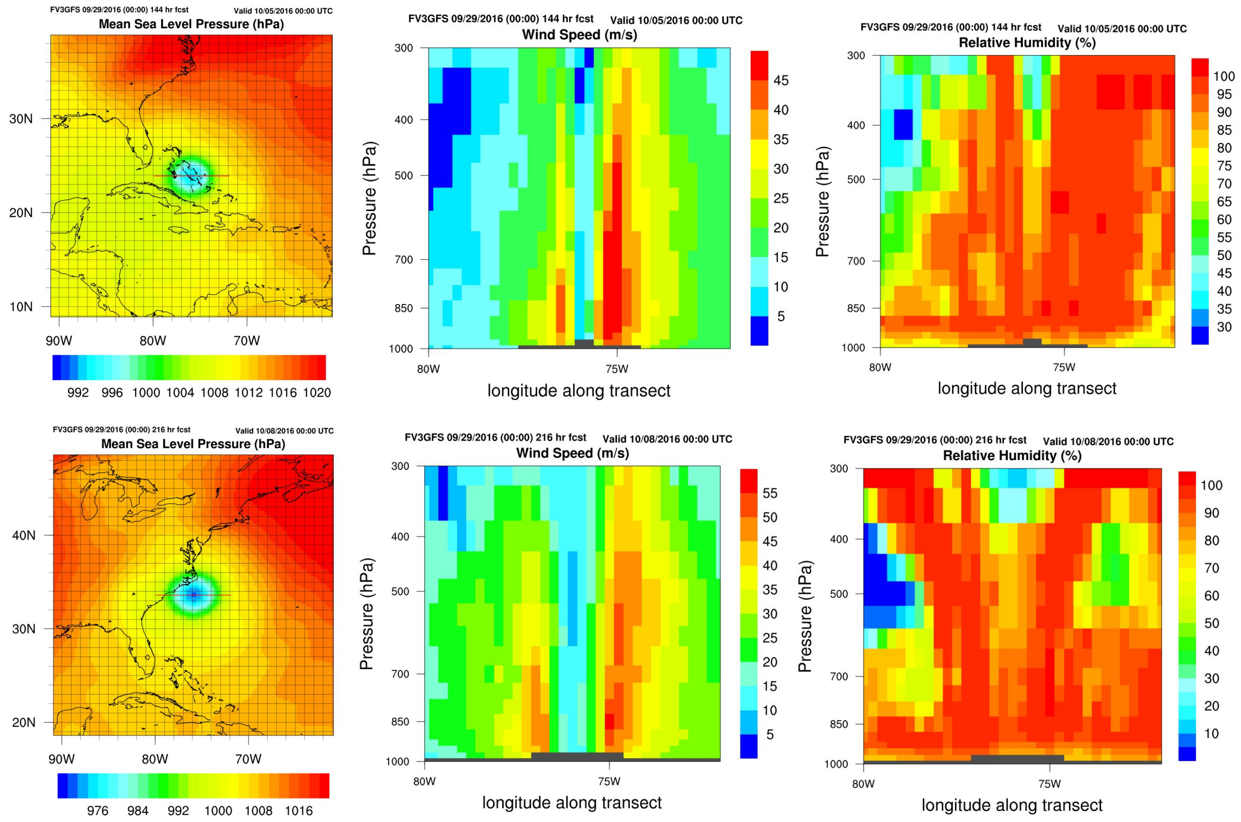


Figure H8. Vertical cross-sections of FV3 output through Hurricane Matthew of wind speed (column 2) and relative humidity (column 3) and accompanying maps of mean sea level pressure identifying the location of the transect (red line) for each specific time (column 1) for the 72 hour forecast (row 1), 144 hour forecast (row 2), and 216 hour forecast (row 3). For the vertical cross-sections, data at locations where surface pressure was lower than the pressure at that point were masked out (dark gray).

6. 2017 May Severe Weather Outbreak

6.1 Synoptic Discussion

This high impact event occurred May 18 - 19, 2017, causing an estimated $2.12 million in damages (https://www.ncdc.noaa.gov/stormevents/choosedates.jsp?statefips=-999%2CALL). The severe weather most affected Texas, Oklahoma, Kansas, Missouri, and Arkansas and included tornadoes, damaging winds, and hail (Figure S1). The synoptic set-up for this event involved a few different components including: an upper level low, an upper level jet, a dry line, and a warm and moist air mass.

a.

b.

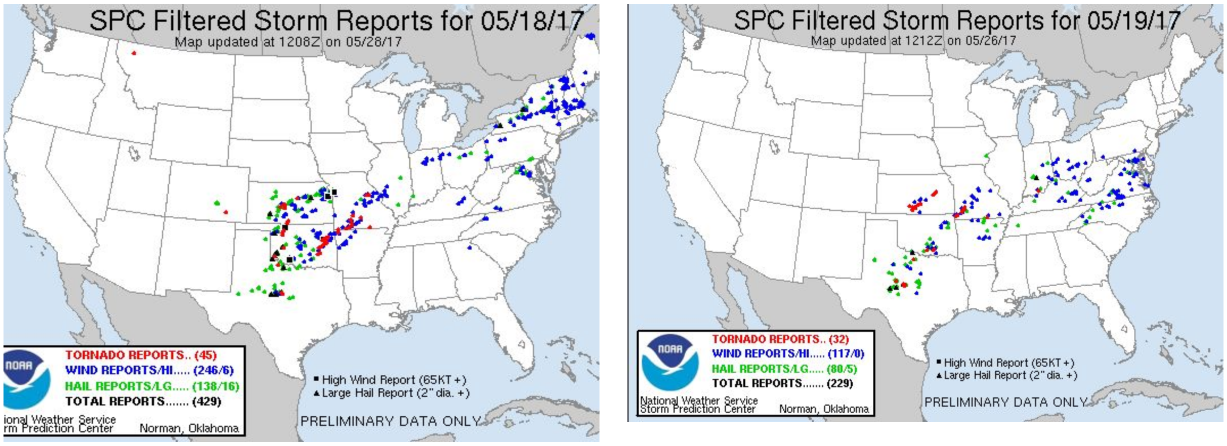

SPC Filtered Storm Reports for 05/18/17
Map updated at 1208Z on 05/26/17
TORNADO REPORTS.. (45)
WIND REPORTS/HI..... (246/6)
HAIL REPORTS/LG..... (138/16)
TOTAL REPORTS....... (429)
■ High Wind Report (65KT+)
▲ Large Hail Report (2" dia. +)
National Weather Service
Storm Prediction Center
Norman, Oklahoma
PRELIMINARY DATA ONLY
SPC Filtered Storm Reports for 05/19/17
Map updated at 1212Z on 05/26/17
TORNADO REPORTS.. (32)
WIND REPORTS/HI..... (117/0)
HAIL REPORTS/LG..... (80/5)
TOTAL REPORTS....... (229)
■ High Wind Report (65KT+)
▲ Large Hail Report (2" dia. +)
National Weather Service
Storm Prediction Center
Norman, Oklahoma
PRELIMINARY DATA ONLY

Figure S1: Filtered Storm Prediction Center (SPC) storm reports for (a) 05/18/17 and (b) 05/19/17.

The Model Evaluation for Research and Innovation Transition (MERIT) team communicated with the National Oceanic and Atmospheric Administration's (NOAA) Model Evaluation Group (MEG) about deficiencies the MEG team discovered while running FV3 retrospective case studies. The deficiencies specific to this case are: a dryline that is too progressive, a low CAPE bias, and a dry bias. Tools that can be used to investigate these deficiencies are described in the last section.

6.2 Verification results

Synoptic Comparison

A strong upper level low coinciding with the exit region of a 500-mb jet provided strong mid-level forcing for ascent. The FV3 model decently captured the 500-mb low's position and strength (Figure S2a-b). The model produced the minimum height in the 550s as was observed, but did not produce a closed low as was observed. FV3 mostly captured the center of the low, forecasting it over Utah and Colorado; however the model upper level low is slightly shifted to the north of what was observed. The FV3 model accurately captured the magnitude of the 500-mb jet (Figure S2c). The observed right exit region, located over western Kansas and the Oklahoma panhandle, has 500-mb wind speeds of approximately 50 - 60 knots (Figure S2a). The forecast right exit region has 500-mb wind speeds of approximately 30 m/s (~58 kts) which matches the observed values. The forecast jet is oriented more north-south than observed.

a.

b.

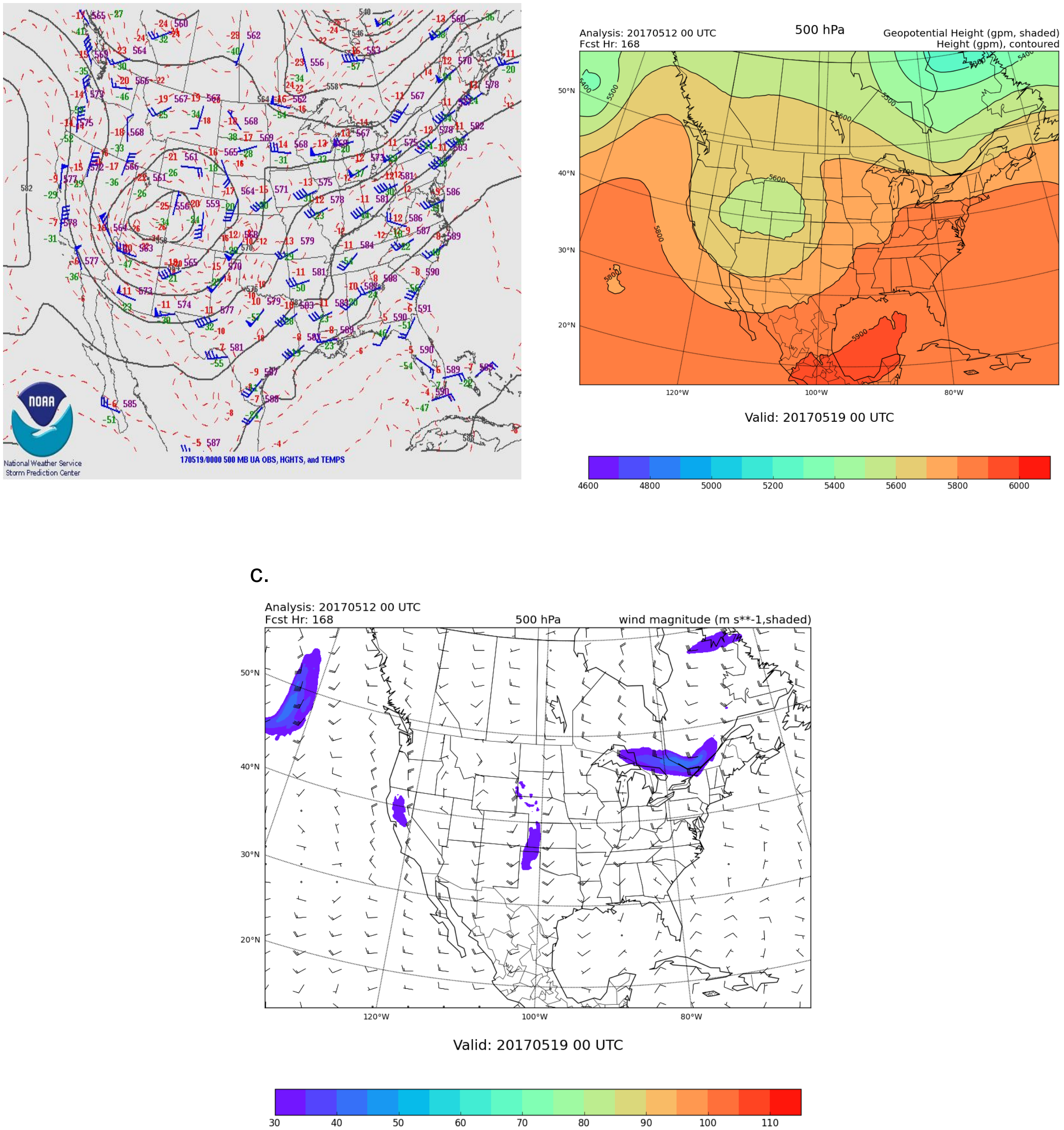


Figure S2: (a) Observed 500mb upper air observations, heights (solid black lines), temperatures (dashed red lines), and wind speed (wind barbs, in knots) valid at 20170519 00 UTC, (b) FV3 500mb geopotential heights (shaded) and heights (contoured) valid 20170519 00 UTC, and (c) FV3 500mb wind speeds in m/s.

Surface Verification (CONUS only)

Temperature

A strong diurnal cycle is present in the mean error, with 00Z valid times exhibiting the minima, resulting in a cool bias ranging from -1.5 - -2.5C (Figure S3a). The peaks occur at the 12Z valid times and exhibit a very small bias to no bias, ranging from 0.5 - -0.5C. For root mean squared error (RMSE), a strong diurnal cycle is present, with 00Z valid times exhibiting the peak error values and the 06Z valid times exhibiting the minima errors (Figure S3b). There is an overall increase in error value throughout the forecast time period, with a minimum of 2C and a maximum of 5C.

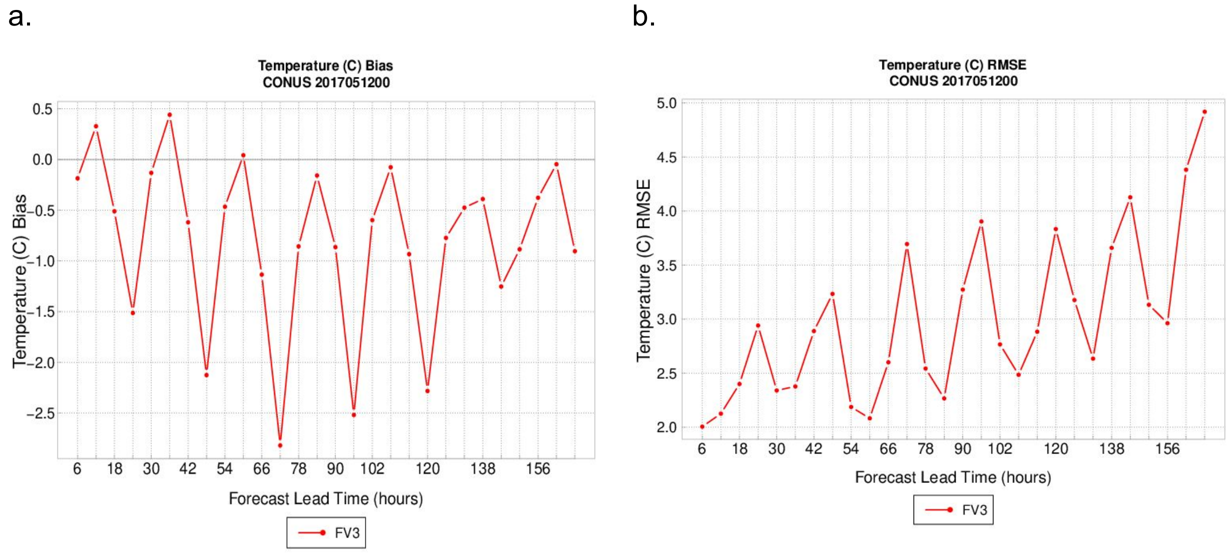


Figure S3: (a) Mean error for 2-m temperature and (b) root mean squared error for 2-m temperature.

Specific Humidity

As in temperature, a strong diurnal cycle is present for bias, with 00Z valid times exhibiting the peak error values around 0.001 kg/kg (Figure S4a). The 12Z valid times exhibit the minima error values that are generally between -0.0005 - 0.0005 kg/kg. For RMSE, there is a strong diurnal cycle present and an overall increase in magnitude across the forecast period (Figure S4b). The peak values occur at 00Z valid times and the minima values occur at the 12Z valid times. RMSE ranges from 0.0010 - 0.0030 kg/kg.

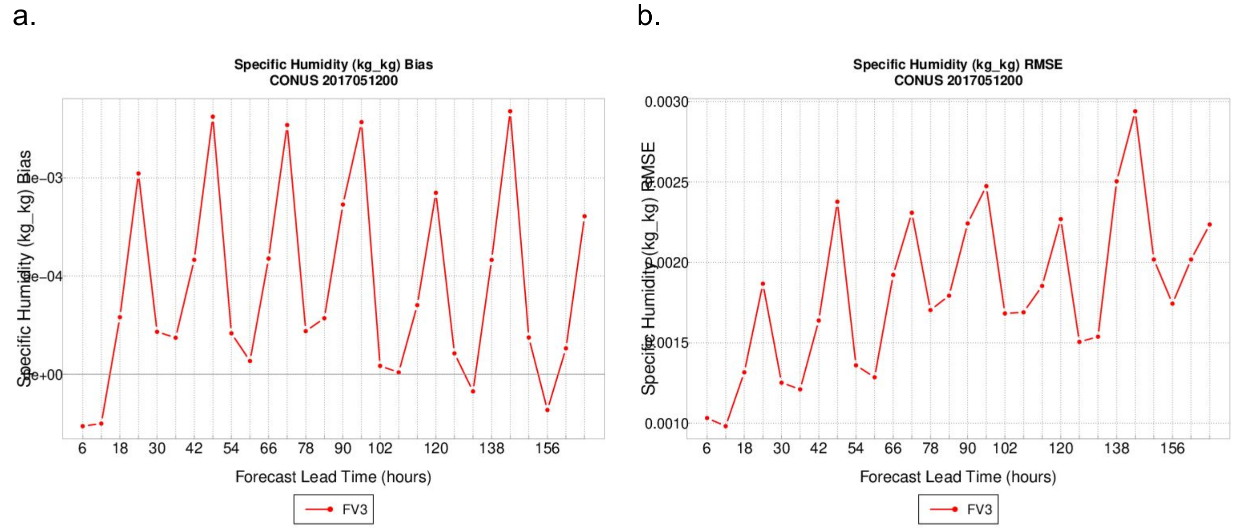


Figure S4: As in Figure S3, but for surface specific humidity.

Wind Speed

There is a fast bias with values ranging from 0.4 - 1 m/s (Figure S5a). A strong diurnal cycle is present with peak values occurring at 06Z valid times and minima occurring at 00Z valid times. The overall magnitude of error decreases with forecast lead time. There is no diurnal cycle present in RMSE (Figure S5b). Values range between approximately 1.7 - 2.9 m/s and remain somewhat steady around 2.0 m/s until forecast hour 108. There is an increase in value that occurs at forecast hour 120.

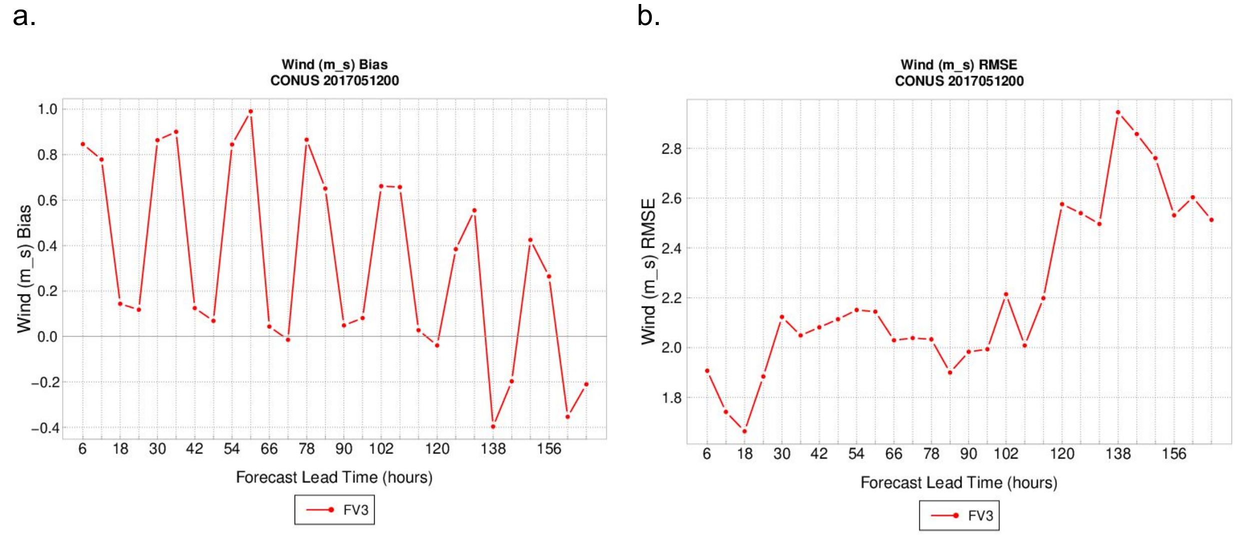


Figure S5: As in Figure S3, but for 10-m wind speed.

Dew Point Temperature

As with all previous variables, there is a strong diurnal cycle in bias, with peak values occurring at 00Z valid times and minima occurring at 06Z valid times (Figure S6a). Values range from just under 0.0 C to approximately 2.5 C. For RMSE, there is a strong diurnal cycle with peak values occurring at 00Z valid times and minima occurring at 12Z valid times (Figure S6b). Values range from approximately 2.5 - 5.0C.

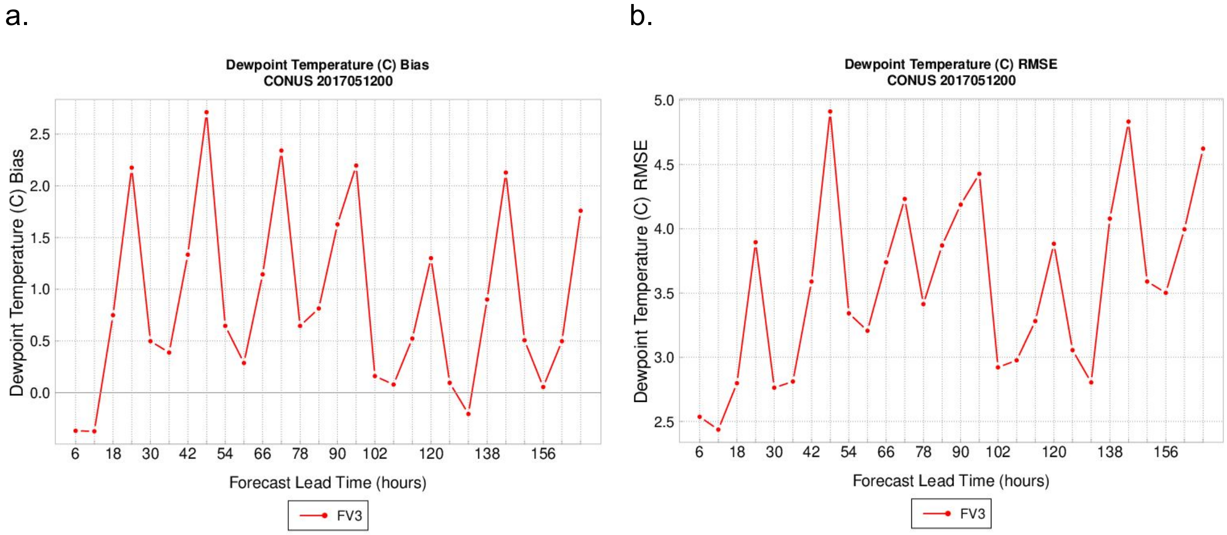


Figure S6: Same as in Figure S3, but for 2-m dew point temperature.

Convective Available Potential Energy (CAPE)

There is an overall low CAPE bias that increases in magnitude with forecast lead time (Figure S7a). A diurnal signal is apparent between forecast hours 6 - 102, with mean error maxima occurring at 06Z valid times and error minima occurring at 00Z valid times. At forecast hour 102, the trend changes and there is a majority low CAPE bias for the rest of the forecast period. The only exception to this is at forecast hour 142 where there is a 100 J/kg positive CAPE bias. In terms of RMSE, there is a strong diurnal signal apparent beginning at forecast hour 36 (Figure S7b). RMSE minima occur at 12Z valid times and maxima occur at 00Z valid times. The magnitude of RMSE increases throughout the forecast period.

a. b.

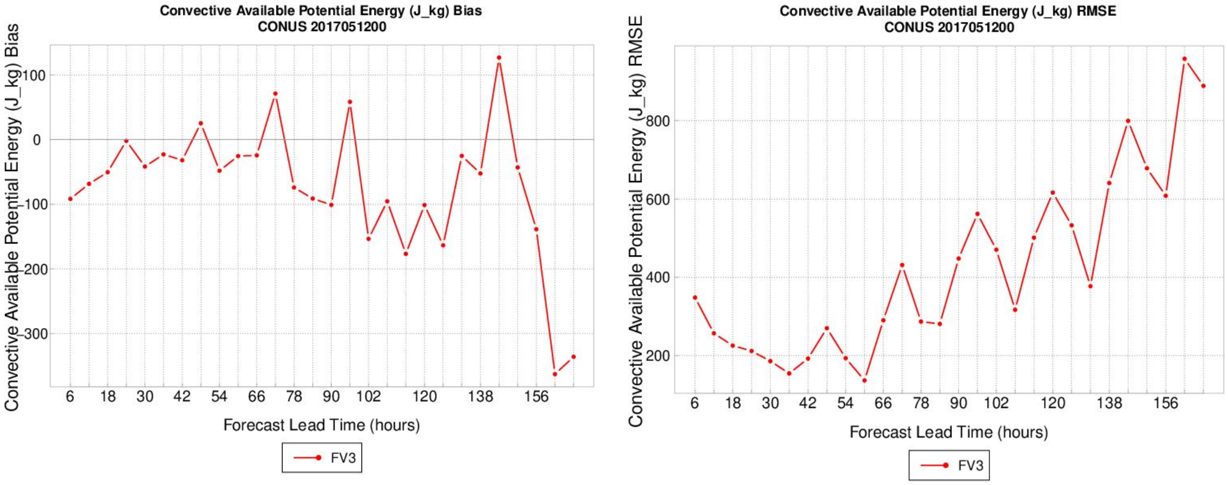


Figure S7: As in Figure S3, but for surface based CAPE.

Convective Inhibition (CIN)

There is an overall high CIN bias throughout the forecast period, which corresponds with the previously mentioned overall low CAPE bias (Figure S8a). The maximum CIN bias occurs at forecast hour 126. CIN RMSE shows a similar temporal trend as the CIN mean error (Figure S8b).

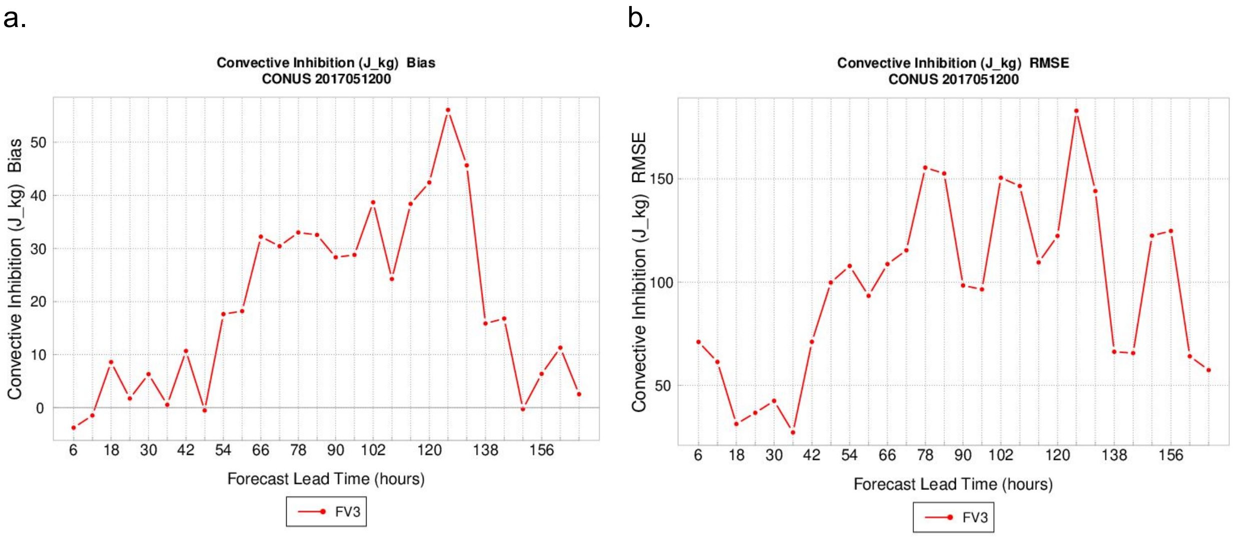


Figure S8: As in Figure S3, but for surface based CIN.

Precipitable Water

There is a high precipitable water bias throughout the forecast period (Figure S9a). The overall magnitude of this bias increases with forecast lead time. There is a diurnal signal present with error minima occurring at 12Z valid times and error maxima occurring at 06Z valid times. RMSE for precipitable water displays a strong temporal signal with RMSE minima occurring at 18Z valid times and RMSE maxima occurring at 00Z valid times. RMSE values increase overall with forecast lead time (Figure S9b).

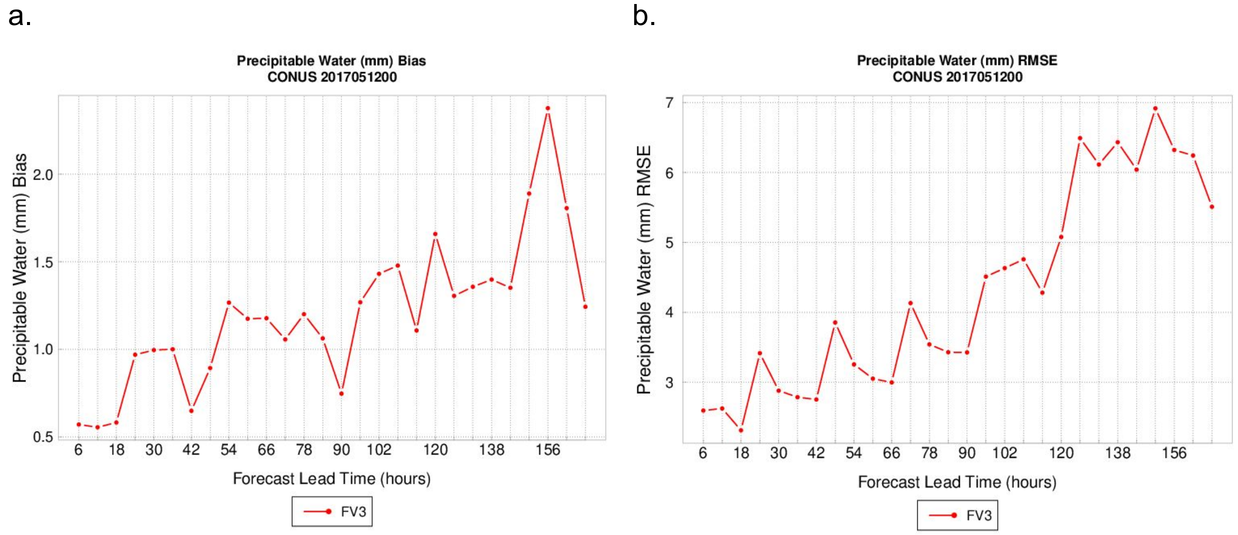


Figure S9: Same as in Figure S3, but for precipitable water.

Upper Air Verification

Six forecast hours are examined for upper air analysis: 12, 24, 60, 96, 120, and 144. These six were chosen to represent model behavior at the short, medium, and long-range forecasts.

CONUS

Temperature

The general trend for mean error throughout all forecast lead times is as follows: there is a slight cool bias that increases in magnitude with forecast lead time (Figure S10a). This bias never exceeds -1C. The cool bias generally exists between 1000 and 700mb with some exceptions. After 200mb, the bias generally switches from cool to warm every 50mb from 200 - 100mb.

For RMSE in the short and medium-range forecasts, RMSE decreases from the surface until mid to upper levels (500 - 300mb), after which, there is an overall increase until the upper troposphere (approximately 200-150mb) when there is a decrease (Figure S10b). For the long-range forecasts, RMSE increases with height until 400mb (120hr) and 850mb (144hr) after which the values switch from increasing to decreasing every 50 - 100mb.

a.

b.

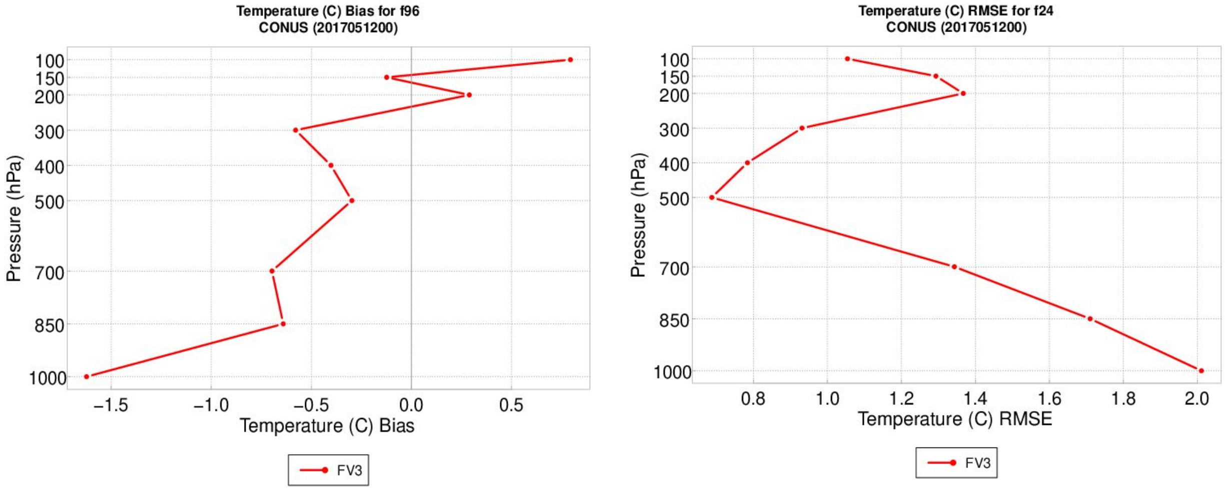


Figure S10: (a) mean error for temperature at pressure levels for forecast lead time 96 and (b) root mean squared error for temperature at pressure levels for forecast lead time 24.

Relative Humidity

For short and medium-range forecast hours, a moist bias is present that increases with pressure level (Figure S11a). For the long-range forecast hours, there is a moist bias that decreases from 850 to 700mb, then increases with pressure level. The moist bias ranges from 5 - 30%. For all forecast hours except 60 and 144, RMSE increases with pressure level. For forecast hours 60 and 144, RMSE switches back and forth between increasing and decreasing with each pressure level (Figure S11b). RMSE values range from 15 - 45%.

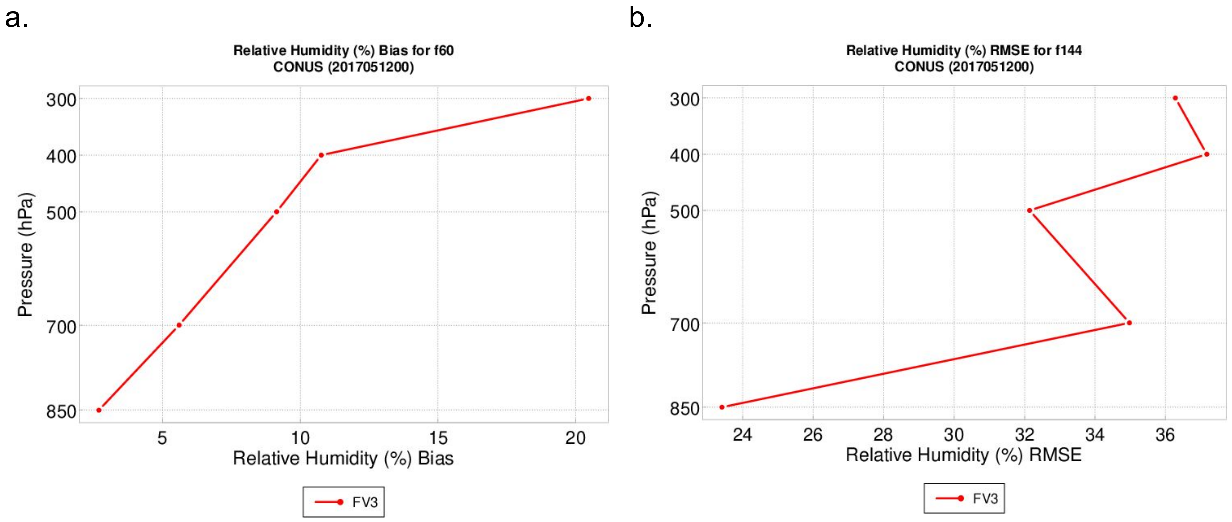


Figure S11: Same as in Figure S10 except for relative humidity at (a) forecast lead time 60 and (b) forecast lead time 144.

Wind Speed

In terms of mean error, there is a slow bias in the upper levels. Other than that, there is no universal trend followed by the selected forecast hours. Bias ranges from -2.5 - 1.0 m/s. Forecast hour 12 exhibits a slow bias at all levels except 1000mb. Maximum error occurs at 200mb (Figure S12a). Forecast hour 24 exhibits a fast bias from 1000 - 700mb that decreases in magnitude. From 500mb on, all bias is slow. Forecast hour 60 exhibited a fast bias at 1000mb, a slow bias from 850 - 500mb, a fast bias from 400 - 300 mb, and then a slow bias from 200 - 100mb. Forecast 96 exhibits the same pattern as seen in forecast hour 24. Forecast hour 120 exhibits a slow bias at all levels with the maximum error occurring at 400mb. Finally, in forecast hour 144, no bias is exhibited at 1000mb. There is a slow bias at all other pressure levels.

For all forecast lead times except 12, RMSE increases from 1000 - 400 or 300mb, then decreases through 100mb (Figure S12b). RMSE increases overall with pressure level for forecast hour 12. RMSE values range from 1.5 - 10 m/s with larger values occurring at later forecast lead times.

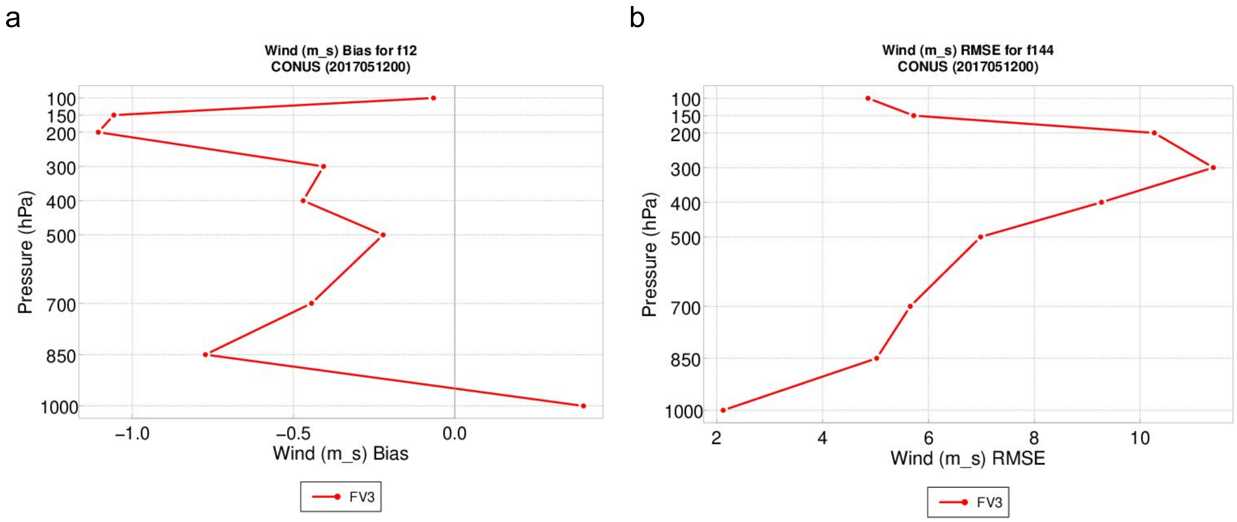


Figure S12: As in Figure S10, but for wind speed at (a) forecast lead time 12 and (b) forecast lead time 144.

Global

Temperature

Globally, the short-range forecast lead times exhibit a slight cool bias throughout the atmosphere which shifts to a slight warm bias at 100mb. Short range mean error values range from approximately -0.3 - 0.1 C. The medium-range forecast lead times exhibit a slight cool bias that decreases to near-neutral at 500 - 400mb, followed by a cool bias and a warm bias at

100mb (Figure S13a). Medium-range forecast mean error values range from approximately -0.6 - 0.4 C. The long-range forecast lead times exhibit two different trends. At forecast hour 120, there is slight cool bias at 1000mb that decreases to near-neutral at 400mb, followed by a warm bias at 100mb. Forecast hour 144 exhibits a slight cool bias that increases until 200mb, followed by a warm bias at 100mb. Long-range forecast mean error values range from approximately -0.6 - 0.4C.

Temperature RMSE for global short-range forecast lead times decreases from 1000mb to 500mb after which the value increases through the rest of the atmosphere. Short range forecast RMSE values range from approximately 0.8 - 1.8 C. Medium-range forecast lead time RMSE decreases from 1000mb until 500mb, after which the values increase until 200 or 150mb, then there is a sharp decrease at 100mb (Figure S13b). Medium-range forecast RMSE values range from approximately 1.2 - 2.2C. The long-range forecast RMSE switched from increasing to decreasing every 100 - 200mb and exhibited the max RMSE at 200mb. Long range forecast RMSE values range from approximately 1.8 - 3.2 C.

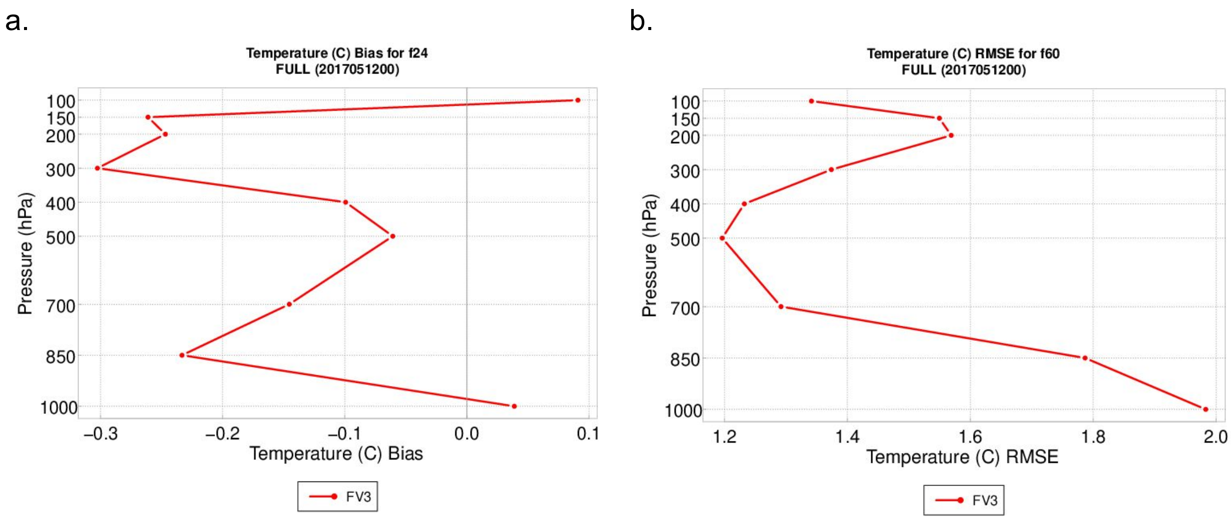


Figure S13: Global temperature (a) mean error at pressure levels for forecast lead time 24 and (b) root mean squared error at pressure levels for forecast lead time 60.

Relative Humidity

All forecast lead times follow the same general trend for global relative humidity mean error. Mean error increases with pressure level for short-range forecast lead times and has values ranging from approximately 5 - 25% (Figure S14a). Forecast hour 60 displays a slight decrease from 850 - 700mb. After that, mean error increases with pressure level. Forecast hour 96 behaves like the short-range forecast lead times. Values range from approximately 5 - 25%. Mean error decreases slightly from 850mb to 700mb, then increases with pressure level. Long range mean error values are slightly lower, ranging from approximately 2.5 - 25%. All forecast lead times exhibit RMSE that increases with pressure level and range in value from approximately 15 - 35% (Figure S14b).

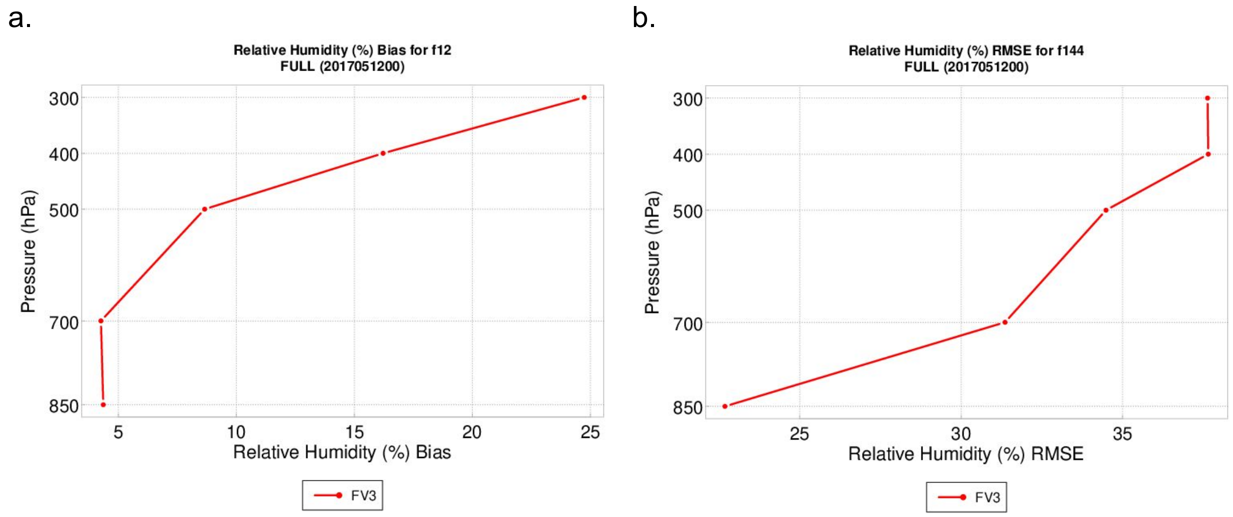


Figure S14: As in Figure S13, but for relative humidity at forecast lead time (a) 12 and (b) 144.

Wind Speed

Short-range forecast lead times exhibit a slight fast bias in global wind speed that changes over to a slow bias with height. Values range from -0.8 - 0.6 m/s. Medium range forecast lead times exhibit approximately the same trend as the short-range forecast lead times with maximum error occurring at 100mb (Figure S15a). Values range from -1.5 - 0.5 m/s. The long range forecast lead times display the same trend as the medium range hours with values ranging from -1.5 - 0.5 m/s.

For all forecast lead times, except for 120, global wind speed RMSE increases with pressure level and the maximum RMSE occurs at 300mb (Figure S15b). The short range RMSE values range from approximately 2.4 - 3.6 m/s. The medium range values range from approximately 2.5 - 7 m/s and the long range values range from 3.5 - 12.5 m/s. Maximum RMSE occurs at 700mb for forecast hour 120.

a. b.

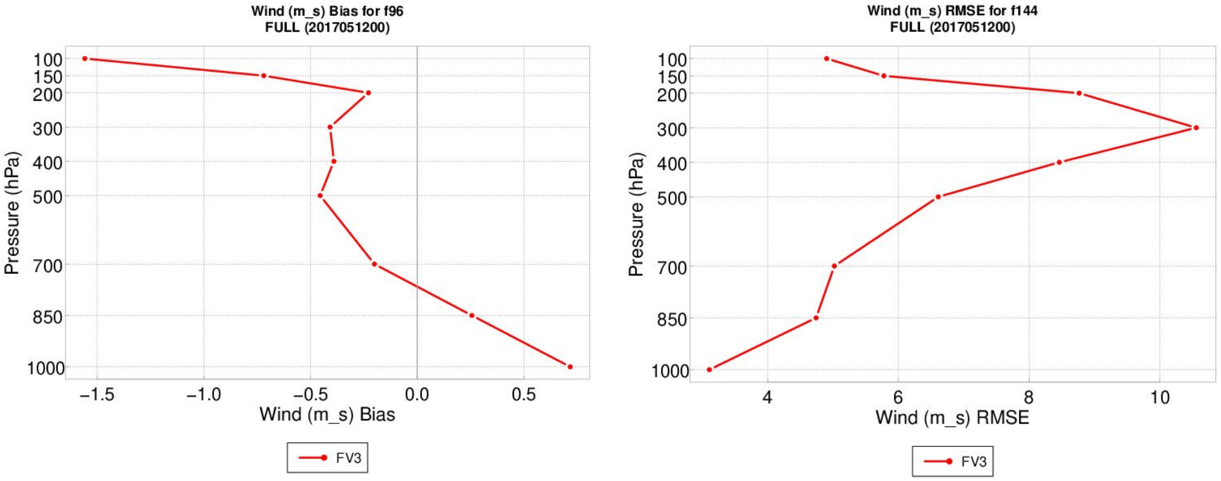


Figure S15: As in Figure S13, but for global wind speed at forecast lead time (a) 96 and (b) 144.

Precipitation Verification

Accumulated 24 hour precipitation is examined in this section. Three forecast lead times were chosen to examine model performance at the short (f24), medium (f72), and long range (f120). Three precipitation thresholds were chosen to examine model performance for light (>=6.35mm/0.25 inches), moderate (>=12.7mm/0.5 inches), and heavy (>=25.4mm/1 inch) accumulations.

CONUS

Model performance over CONUS is compared to the gridded CCPA dataset.

Analysis by Lead Time

Equitable Threat Score (ETS) measures the fraction of the observed events that were correctly predicted. ETS values range from -⅓ - 1. A perfect score is 1 so the higher the ETS score, the better the model performed. The highest ETS score occurs at the beginning of the forecast period for the light precipitation, then decreases overall with forecast lead time (Figure S16a). There is a sharp decrease at 96 hours which then recovers some by forecast hour 120. The highest ETS score occurs at forecast hour 72 and then displays a sharp decrease at forecast 96 (Figure S16b). The heavy precipitation follows a similar trend as the moderate forecast (Figure S16c).

a.

b.

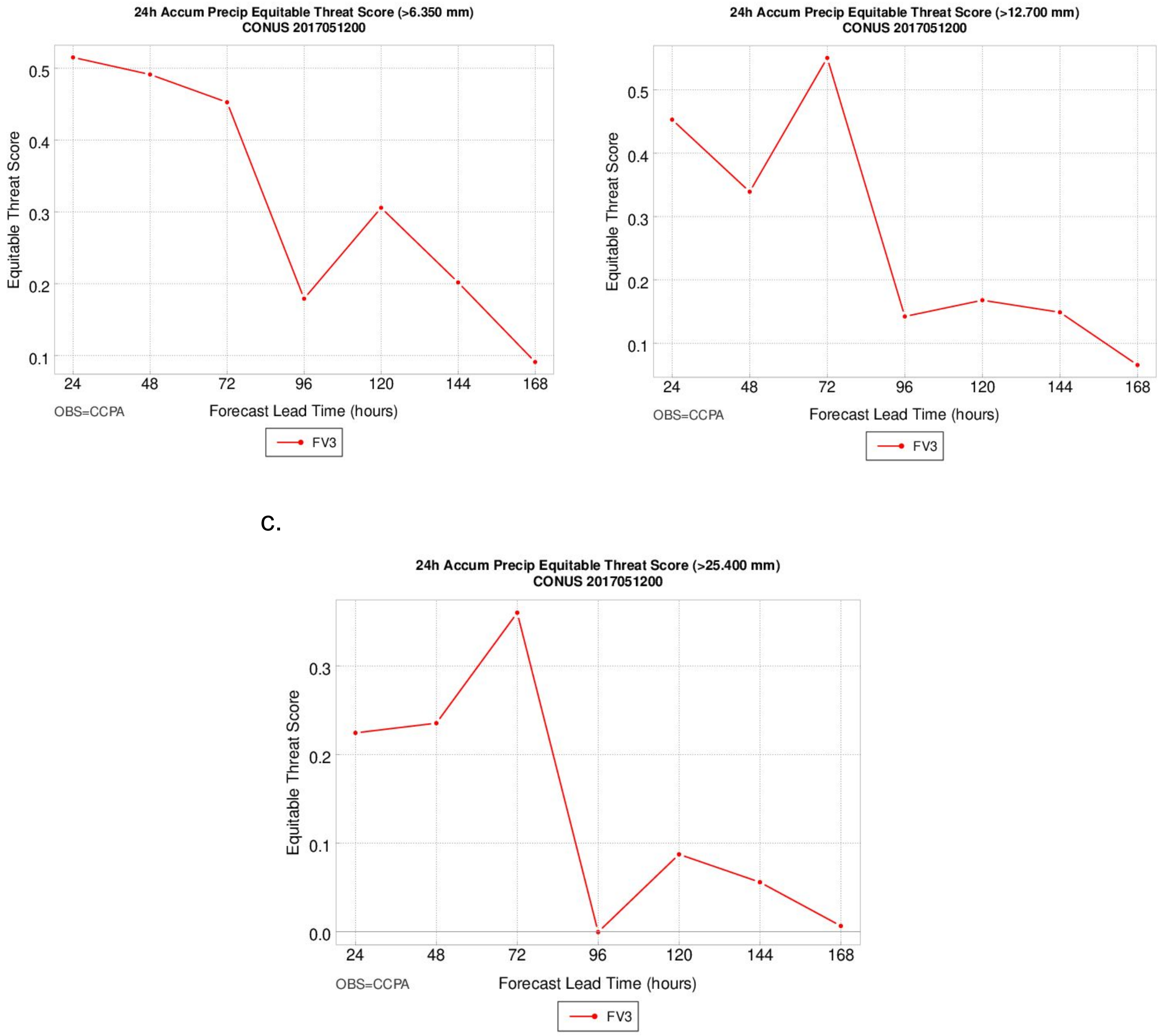


Figure S16: Equitable threat score for 24 hour accumulated precipitation thresholds of (a) >6.35mm, (b) >12.7mm, and >25.4mm throughout the forecast period.

Frequency bias is the ratio of the frequency of forecast events to observed events. Values <= 1 indicate an under-forecast, values >= 1 indicate an over-forecast, and a value of 1 is a perfect forecast. All three accumulations exhibit the same temporal trend: general over-forecast until forecast hour 96 where FV3 under-forecasts (Figure S17). Then the forecast switches from over- and under-forecasting for the remainder of the forecast period.

a. b.

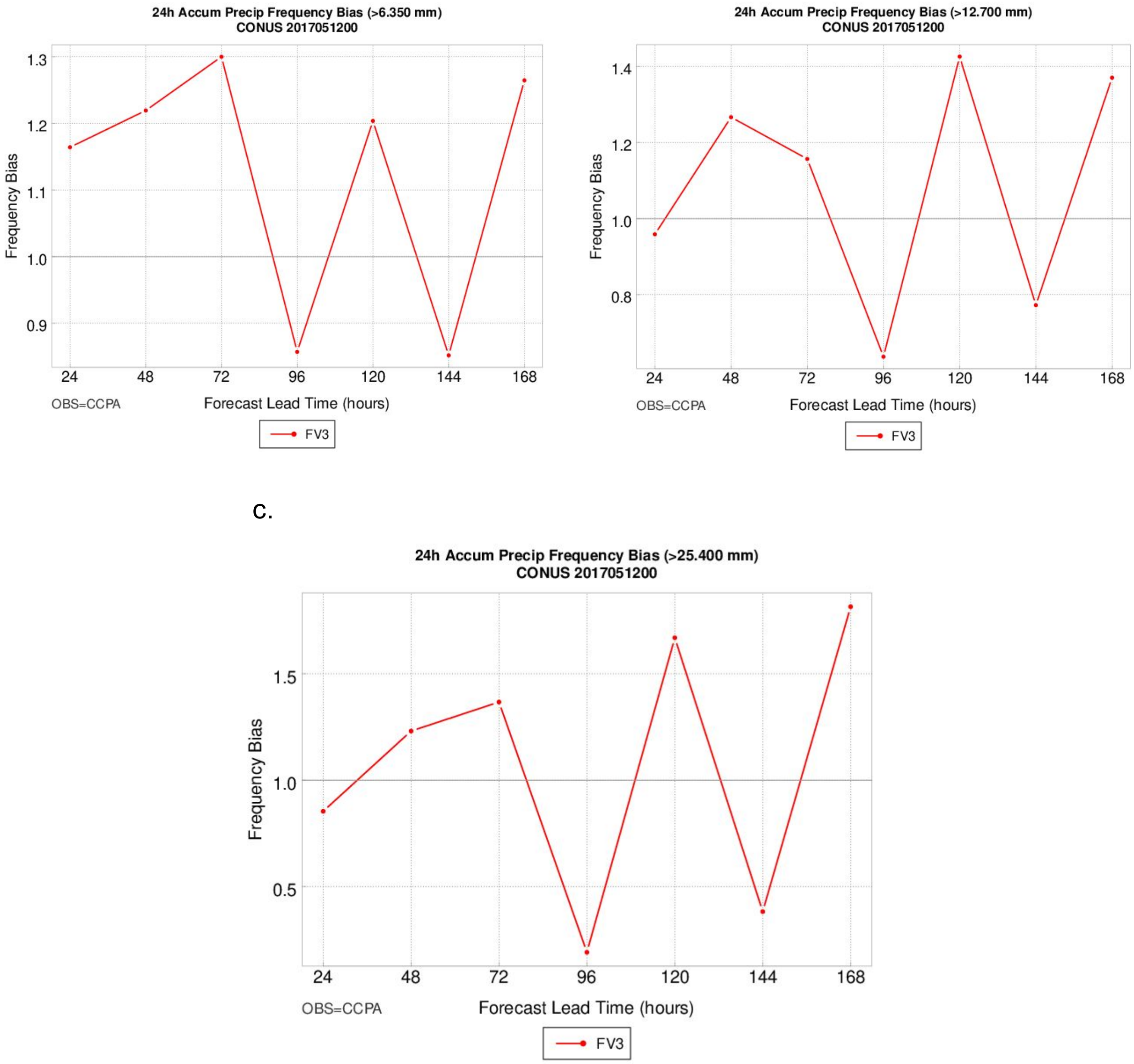


Figure S17: As in Figure S16, but for frequency bias.

Analysis by Threshold

ETS is proportionally higher with lower accumulation thresholds which makes sense due to the higher rate of occurrence of these values (Figure S18). There were few instances of 24-hour accumulation values >38.1mm which is contributes to the ETS values near-0 at those thresholds.

a.

b.

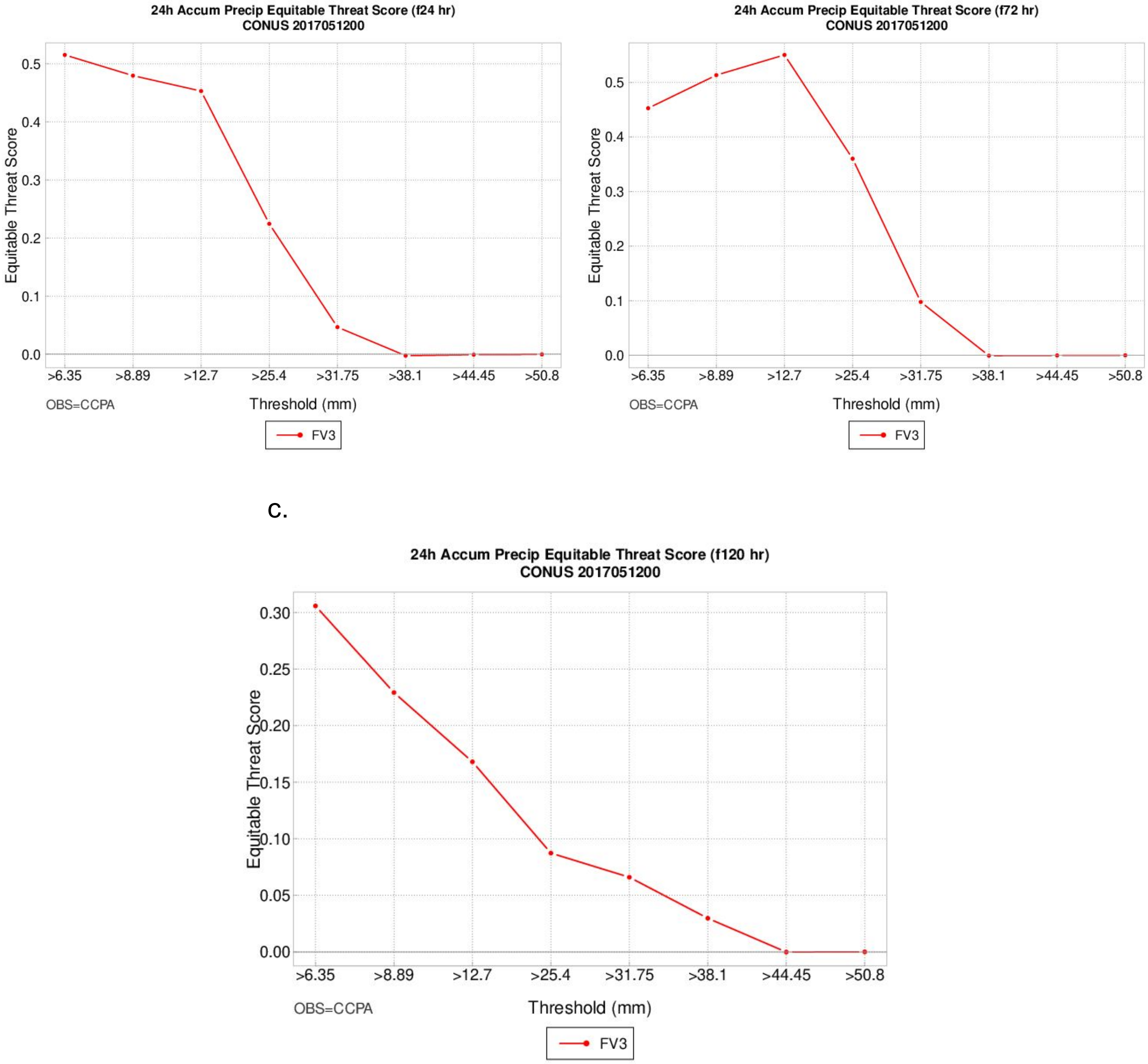
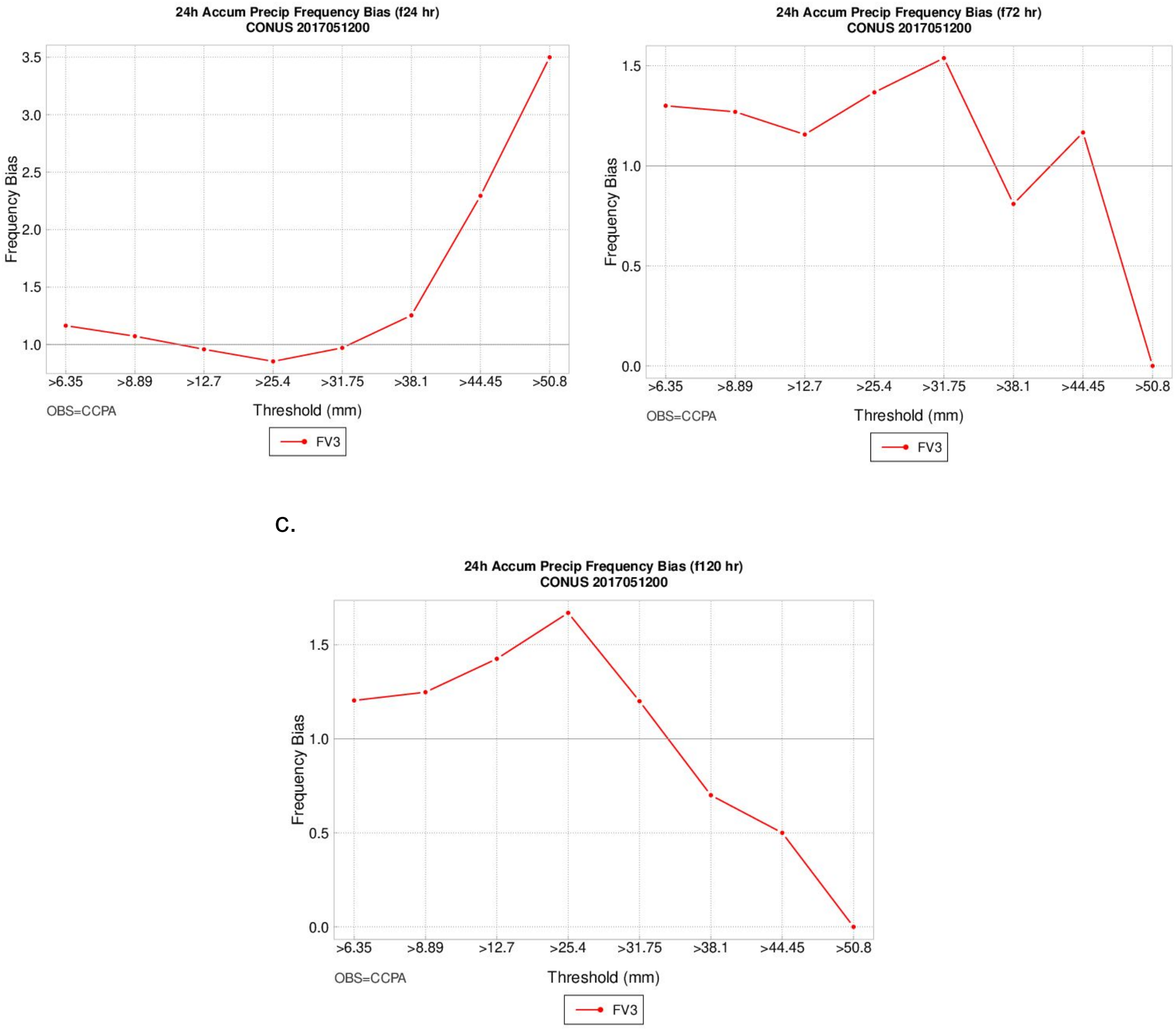


Figure S18: Equitable threat score for all thresholds at forecast hours (a) 24, (b) 72, and (c) 120.

Each forecast range exhibits different frequency bias behavior with threshold. The short-range forecast is near-neutral (around 1) until accumulation values >38.1mm, after which the model over-forecasts (Figure S19a). The medium-range forecast over-forecasts accumulations until >38.1mm, where there is a slight under-forecast, then a slight over-forecast at >44.5mm and no-skill at >50.8mm (Figure S19b). Lastly, the long-range forecast over-forecasts for all thresholds less than 38.1mm (Figure S19c). FV3 then under-forecasts for all thresholds greater than >38.1mm.

a.

b.



c.

Figure S19: As in Figure S18, but showing frequency bias.

Global

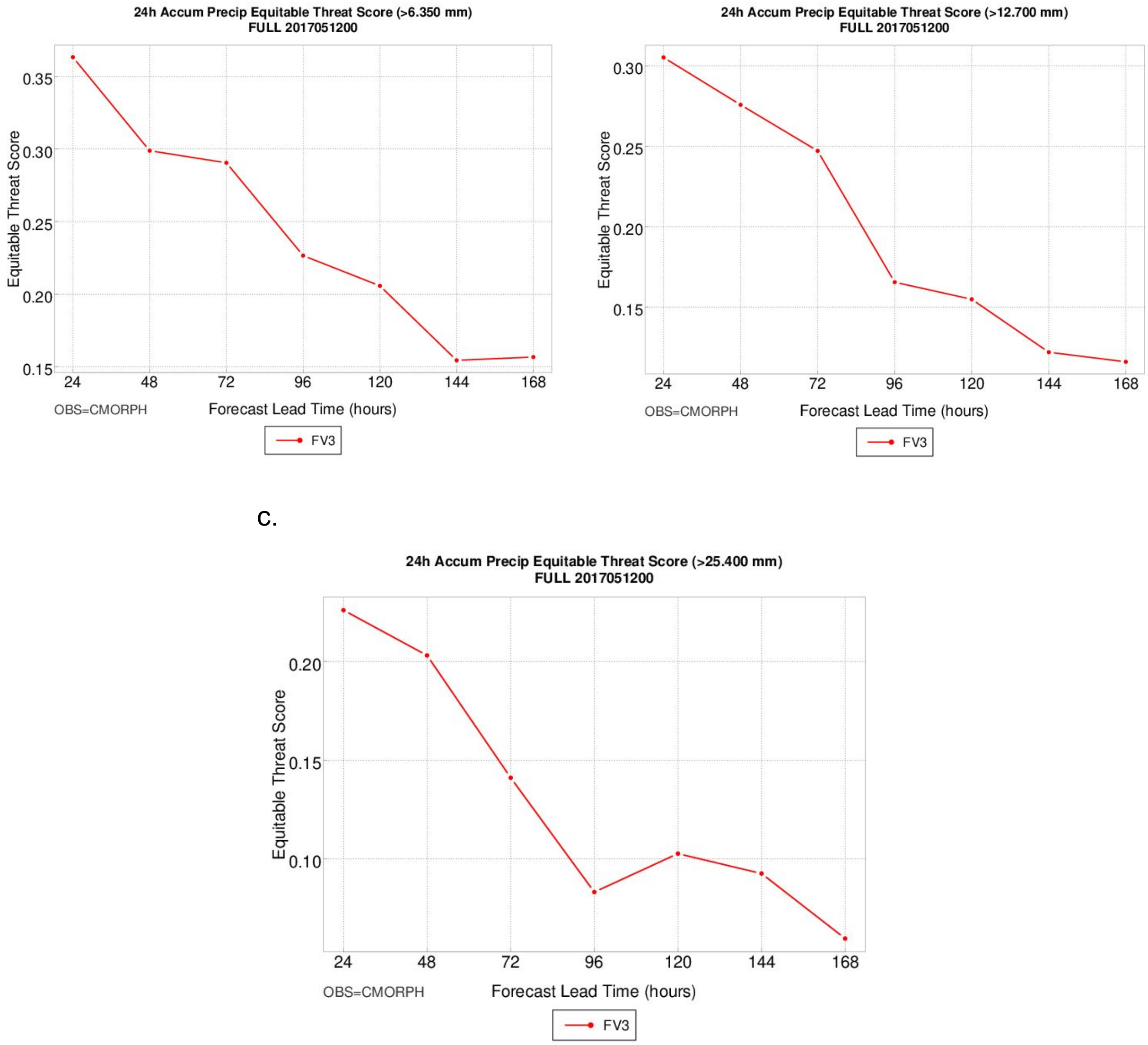
Global model performance is compared to the gridded CMORPH dataset.

Analysis by Lead Time

Globally, the light and moderate precipitation accumulation thresholds exhibit the same pattern. ETS decreases with forecast lead time and the best score occurs at forecast hour 24 (Figure S20a-b). The heavy precipitation threshold exhibits the same trend as light and moderate until forecast hour 120, after which there is a slight recovery and then a subsequent decrease throughout the rest of the forecast period (Figure S20c).

a.

b.



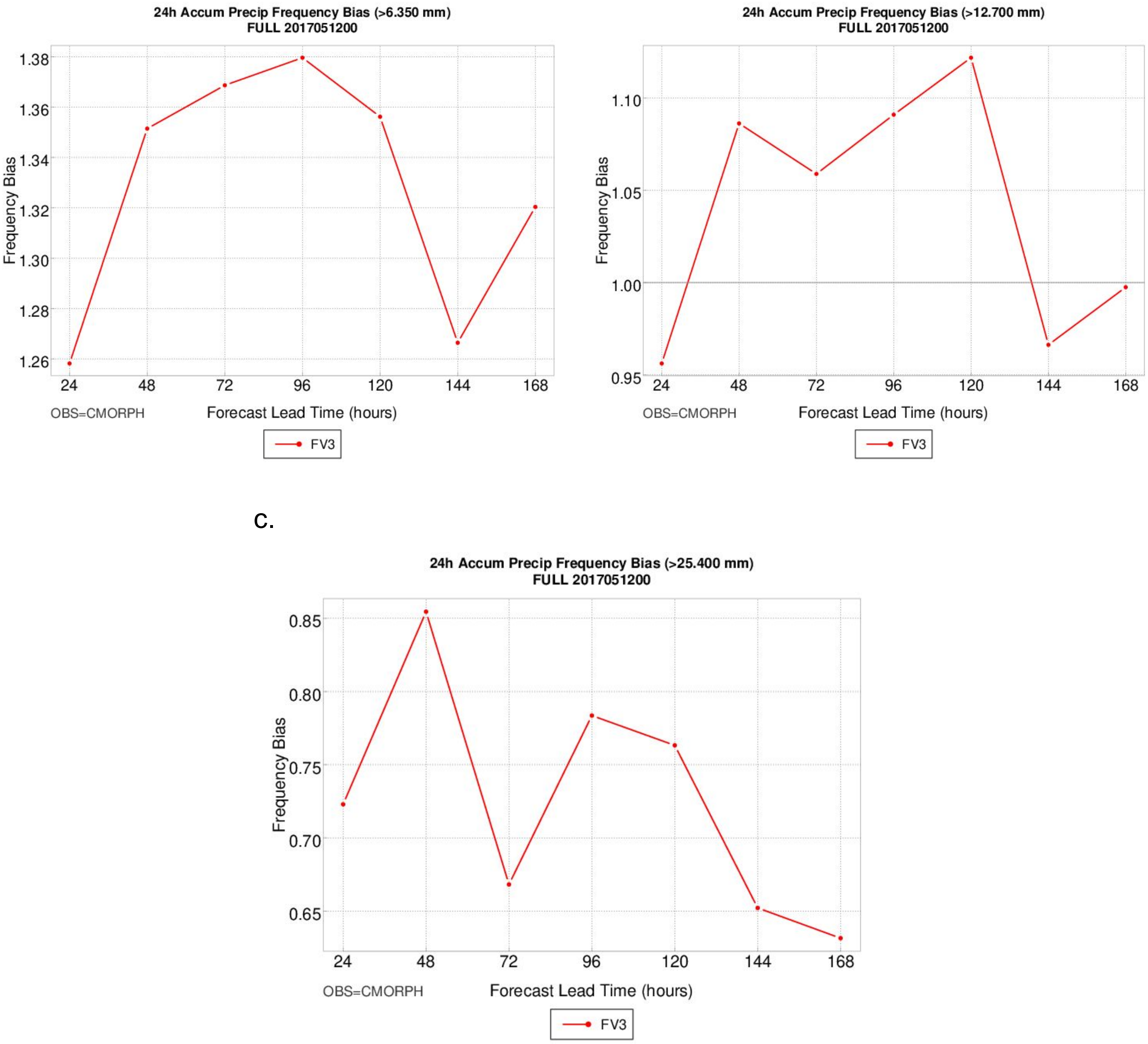
c.

Figure S20: As in Figure S18, but for global 24 hour accumulated precipitation.

FV3 over-forecasts the global light precipitation accumulation threshold throughout the forecast period (Figure S21a). FV3 begins with a slight under-forecast at f24 for the moderate precipitation threshold, then over-forecasts in general, then returning to an under-forecast for the 144 hour lead time on (Figure S21b). Finally, FV3 under-forecasts the heavy precipitation accumulation threshold globally throughout the forecast period (Figure S21c).

a.

b.



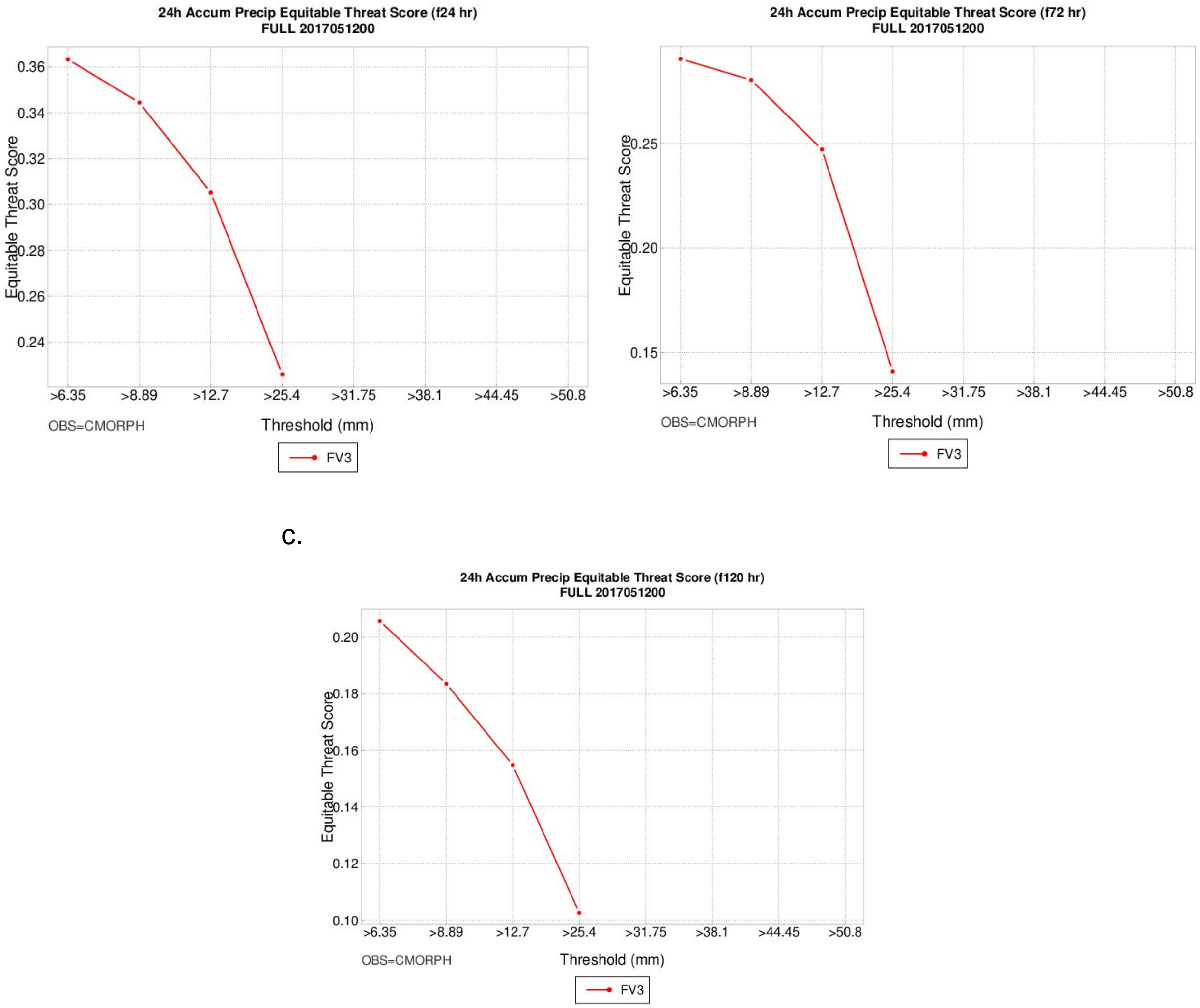
c.

Figure S21: As in Figure S20, but for frequency bias.

Analysis by Threshold

Globally, the short-, medium-, and long-range forecast hours behaved the same (Figure S22). FV3 displays the most skill at the lowest accumulated precipitation threshold, then ETS score decreases with increased threshold. There are no occurrences of 24-hour accumulated precipitation greater than 31.75mm according to the CMORPH dataset, therefore there is no ETS. The maximum score value decreases with forecast lead time.

a. b.



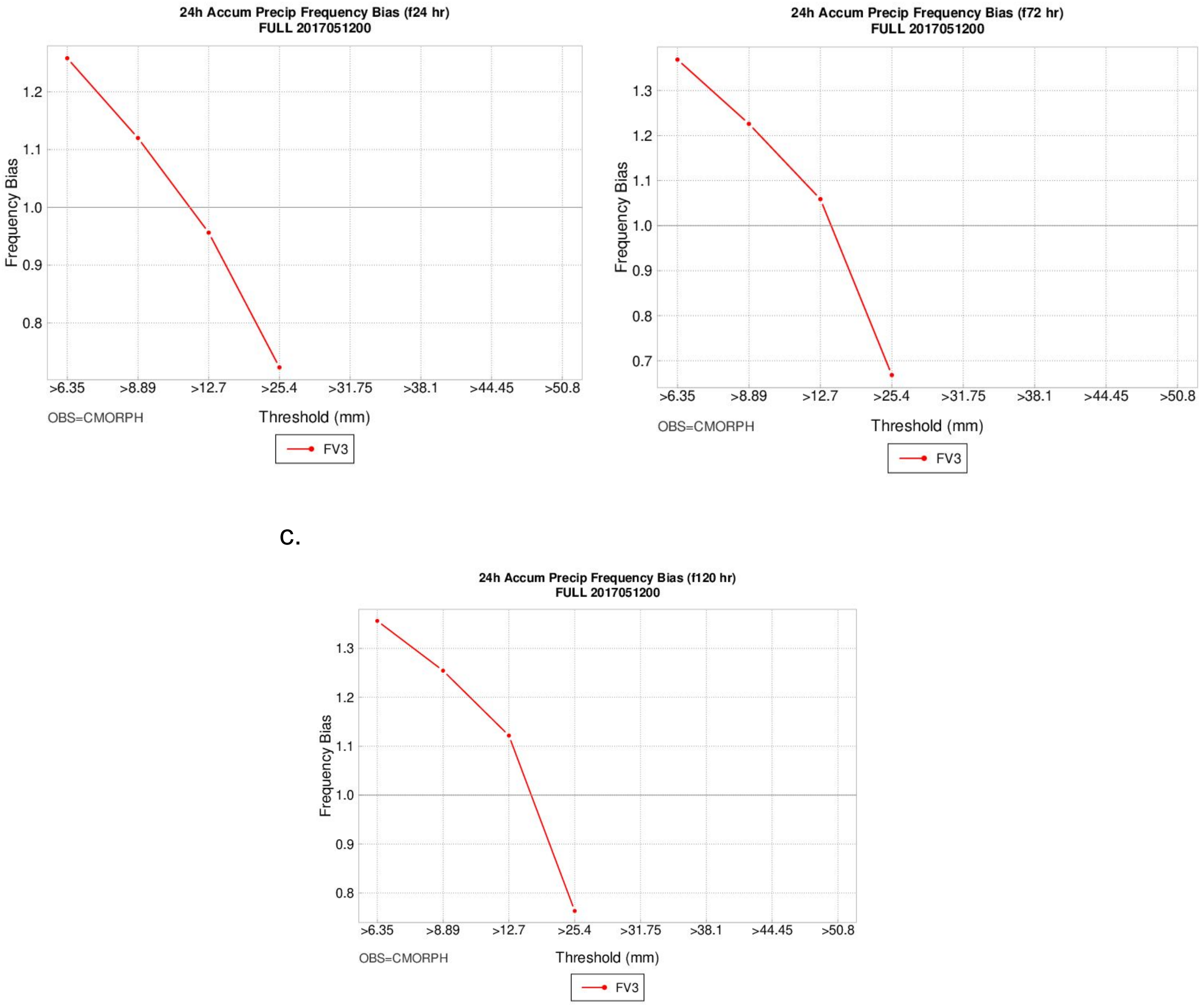
c.

Figure S22: Equitable threat score for global 24-hour accumulated precipitation thresholds at forecast hour (a) 24, (b) 72, and (c) 120.

At the short-range lead time, FV3 over-forecasts global 24-hour accumulated precipitation thresholds >6.35 and >8.89mm and under-forecasts thresholds >12.7 and >25.4mm (Figure S23a). Frequency bias values decrease with precipitation threshold. At the medium- and long-range forecast lead times, frequency bias also decreases with precipitation threshold (Figure S23b-c). FV3 over-forecasts for all thresholds except for >25.4mm, where it under-forecasts.

a.

b.



c.

Figure S23: As in Figure S22, but for frequency bias.

Diagnostic Tools

As previously mentioned, the MEG identified some deficiencies present in the FV3 model. The three deficiencies displayed in this case study are: a dry bias, low CAPE bias, and an overly progressive dryline. A number of diagnostic tools were created and employed to investigate these deficiencies. The following sections are categorized by deficiency and detail the tools as well as an evaluation of the results of these tools.

Dry Bias

In addition to accumulated precipitation information (see previous section), two new fields were investigated this year to look at the FV3 dry bias: 2-m dew point temperature (Figure S6) and precipitable water (Figure S9). Mean error and RMSE were calculated and evaluated for both metrics.

Low CAPE Bias

A number of statistics were gathered to investigate the low CAPE bias deficiency identified by the MEG. Mean error and root mean squared error were calculated for both CAPE (Figure S7) and CIN (Figure S8). A low CAPE bias and high CIN bias were identified for the forecast lead times when the convective event occurred. The gridded NARR dataset was used as the observation dataset.

MODE was performed on both fields. Two sets of CAPE objects were created by using the following filtering thresholds: >=500 J/kg and >=1000 J/kg. An example of the outcome of this object identification is detailed in Figure S24. This is output from MODE for CAPE objects >=1000 J/kg. Figure S24a shows the forecast objects (red filled) overlayed with the observation objects (blue outlines) and Figure S24b displays vice versa. These figures illustrate that the forecast (FV3) produced a smaller region of CAPE >=1000 J/kg and that CAPE did not extend as far east as was observed in the NARR.

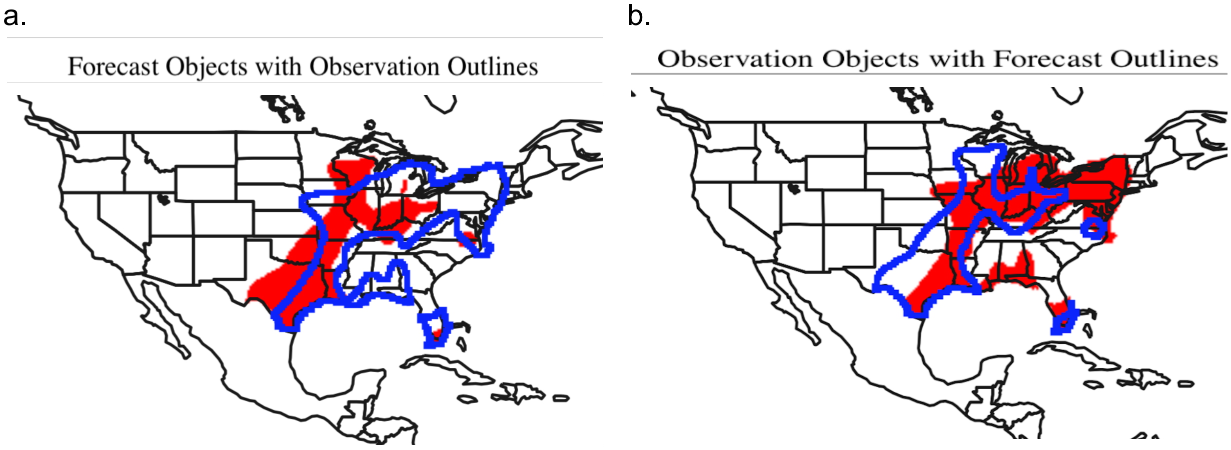

a.
Forecast Objects with Observation Outlines
b.
Observation Objects with Forecast Outlines

Figure S24: MODE output for CAPE objects >=1000 J/kg for forecast valid time 2018-05-17 at 18:00 UTC. (a) forecast CAPE objects (red filled) overlayed with observation objects (blue outlines) and (b) observation CAPE objects (red filled) overlayed with forecast objects (blue outlines).

In addition to creating these sets of forecast and observed CAPE objects, the following statistics were calculated: total object area (Figure S25a), total object count (Figure S25b), east/west centroid displacement (Figure S25c), and north/south centroid displacement (Figure S25d). Object area was calculated by summing up the number of grid points in each simple CAPE object (no clusters) for the entire forecast period. The red line represents the forecast object area and the blue line represents the observed object area. A diurnal cycle is apparent starting at forecast hour 72. Peak object areas occur at 00Z valid times. Starting at 120, forecast generally produces smaller objects than observation, except for 132 and 144.

Object count was calculated by summing up the number of simple objects at each forecast lead time. FV3 tends to produce fewer objects than observed for the first 4 days of the forecast period. After day 4, a phase shift is apparent with FV3 displaying maxima when the NARR displays minima and vice versa.

Centroid displacement is calculated by taking the difference of the observed centroid location and forecast centroid location. Centroid displacement can only be calculated on matched object clusters. The median centroid displacement is displayed in both Figure S25c and d to illustrate the general behavior of the objects. For east/west displacement, values greater than (less than) 0 indicate an easterly (westerly) displacement. Except for three distinct peaks at 24, 42, and 64, the forecasted CAPE objects are displaced to the west. For north/south displacement, values greater than (less than) 0 indicate a northerly (southerly) displacement. The forecasted CAPE objects display a diurnal signal, with maximum southerly displacement occurring at 06Z - 12Z valid times and maximum northerly displacement occurring at 18 - 24Z valid times.

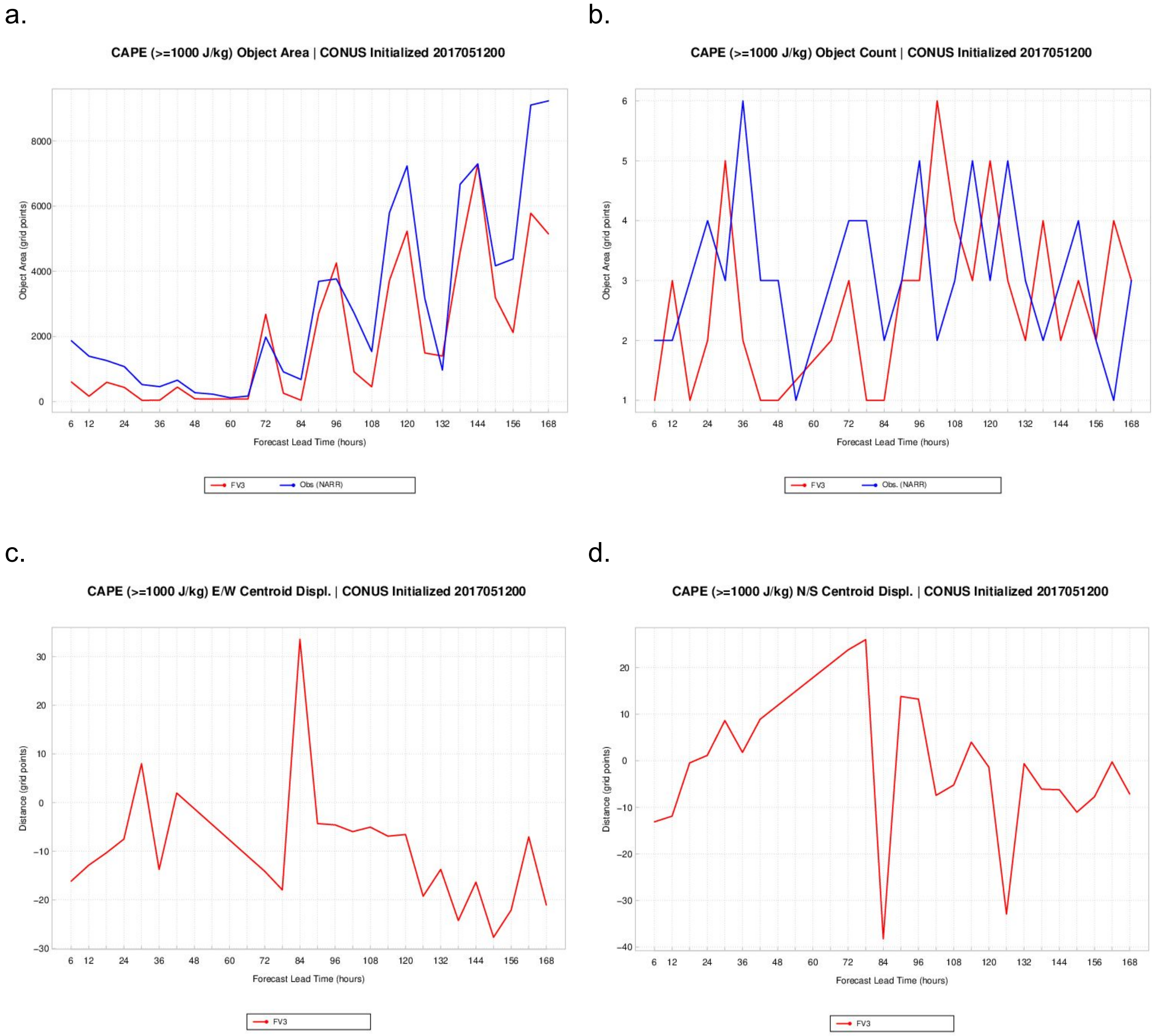


Figure S25: (a) sum of the FV3 (red) and observed (blue) simple object areas for objects of CAPE >=1000 J/kg, (b) total number of forecast (red) and observed (blue) simple objects at the same threshold in (a), (c) the east-west centroid displacement of forecast objects, and (d) north-south centroid displacement of forecast objects.

Two sets of CIN objects were created by using the following thresholds: <= -250 J/kg and <= -100 J/kg. CIN was investigated as it is directly proportional to CAPE. An example of CIN <= -100 J/kg object identification is detailed in Figure S26. Figure S26a shows the forecast objects (green filled) overlayed with the observation objects (blue outlines) and Figure S26b displays vice versa. These figures illustrate that the forecast (FV3) produced a smaller region of CIN <= -100 J/kg than observed and did not create objects in areas that were observed in the NARR.

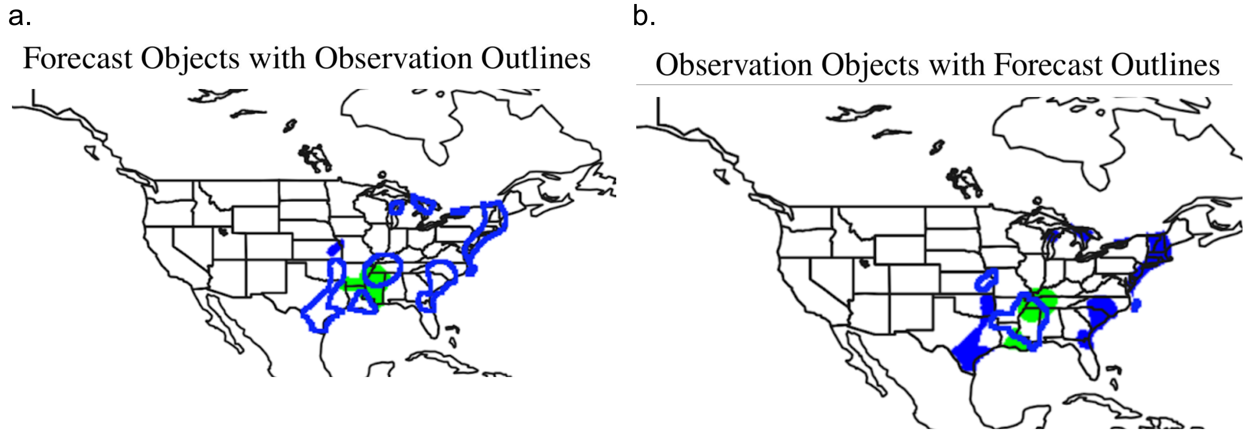

a.
Forecast Objects with Observation Outlines
b.
Observation Objects with Forecast Outlines

Figure S26: MODE output for CIN objects <= -100 J/kg for forecast valid time 2018-05-17 at 18:00 UTC. (a) forecast CIN objects (green filled) overlayed with observation objects (blue outlines) and (b) observation CIN objects (green filled) overlayed with forecast objects (blue outlines).

The same object attribute statistics calculated for CAPE were calculated for CIN. FV3 produced smaller objects than observed throughout the majority of the forecast period (Figure S27a). After forecast 132, during the main convection, FV3 sometimes produces close to the observed area of CIN, but generally produces smaller areas of CIN objects <= -100 J/kg. In terms of object count, FV3 again displays a phase shift as seen in the CAPE object counts (Figure S27b).

For centroid displacement, CIN objects <= -100 J/kg exhibit as diurnal signal for east-west displacement (Figure S27c). The maximum easterly displacement occurs at 18 - 00Z and the maximum westerly displacement occurs at 00-06Z. The magnitude of the displacement increases with forecast lead time. For north-south displacement, a diurnal signal is again apparent with maximum northerly displacement generally occurring at 00 - 06Z and maximum southerly displacement occurs at 18 - 00Z (Figure S27d).

a. b.

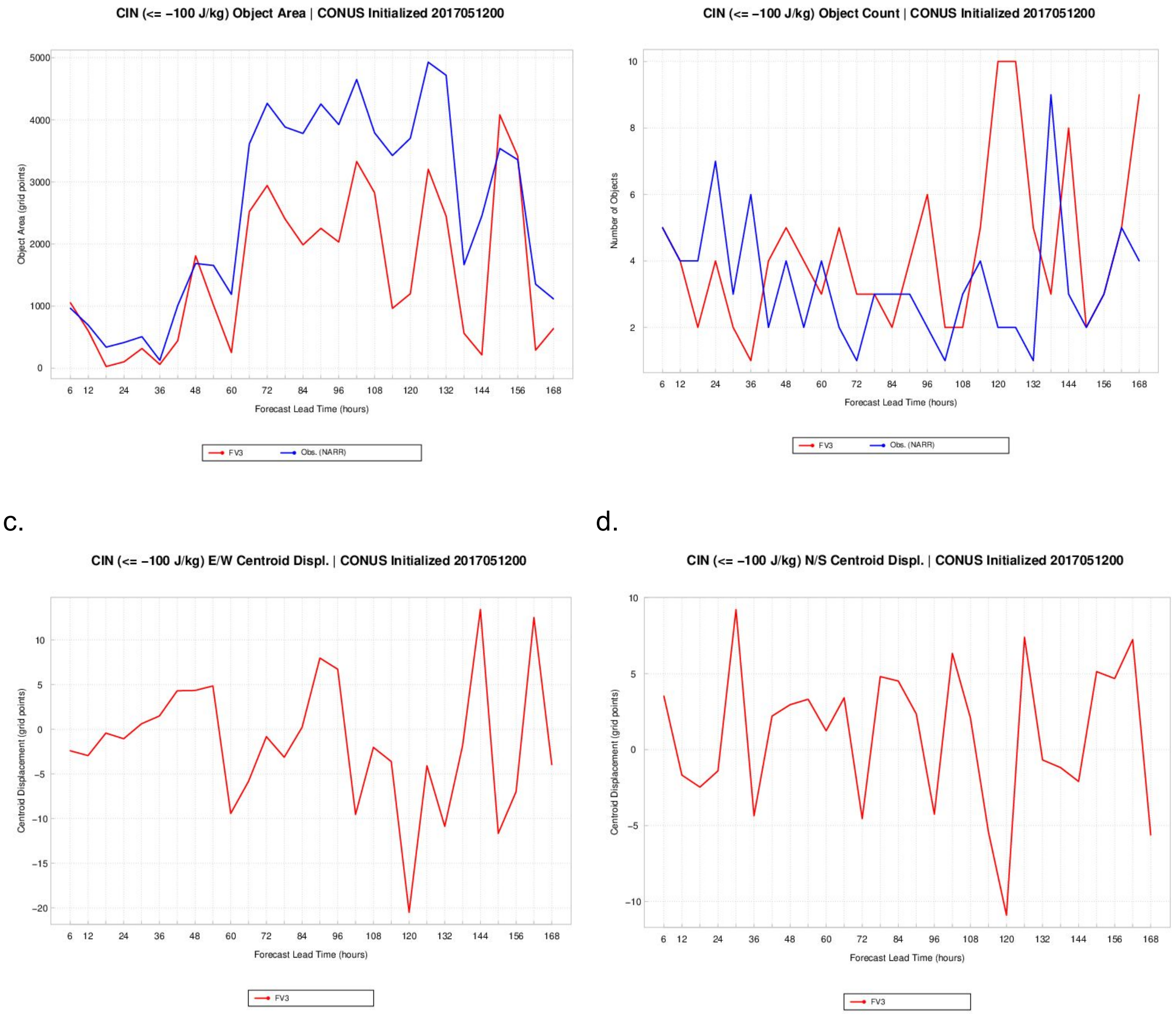


Figure S27: As in Figure S25, but for CIN objects <= -100 J/kg.

Overly-Progressive Dryline

An experimental and preliminary method to identify and evaluate drylines was developed based on previous studies. Drylines are air mass boundaries that separate dry continental air from moist maritime tropical air. Drylines are an important mesoscale feature because of their role in convective initiation. In this case, convection formed along and east of the dry line.

Drylines are characterized by a large moisture gradient (Hoch and Markowski 2005). Previous studies found that using a specific humidity gradient instead of a dew point temperature gradient was desirable because specific humidity does not have a proportional relationship with elevation like dew point does. Dew point temperature decreases with height and this change can contribute to the erroneous identification a dryline (Clark et al. 2015, Coffer et al. 2013, Hoch and Markowski 2005). Hoch and Markowski defined drylines as continuous areas of horizontal specific humidity gradient values greater than or equal to 3 g/kg over a distance of O(100km).

Coffer et al. 2013 and Clark et al. 2015 demonstrate that this definition, along with a few other requirements, is a good way to identify drylines.

MODE has never been used to identify drylines until this project. The object identification process was limited to horizontal specific gradient values as MODE can only identify objects using a single field (at present). The 2-m specific humidity gradient for both the forecast and observations were calculated using the gradient statistics options in the MET grid_stat tool (Figure S28a-b). The gradient was calculated over a distance of 91km. After the gradient data is produced, the gradient data was then put into MODE. Objects were filtered based on the >= 3 g/kg specific humidity gradient.

This method does a decent job of identifying the dryline qualicatically. Some objects that were not drylines were also identified, but this is partially due to the restriction of MODE to a single field. Figure S28c shows the forecast specific humidity gradient objects in red overlaid with the observed objects in blue. Figure S28d shows the opposite. From these images, it is apparent that the dryline produced by FV3 has a further eastward progression than the observed.

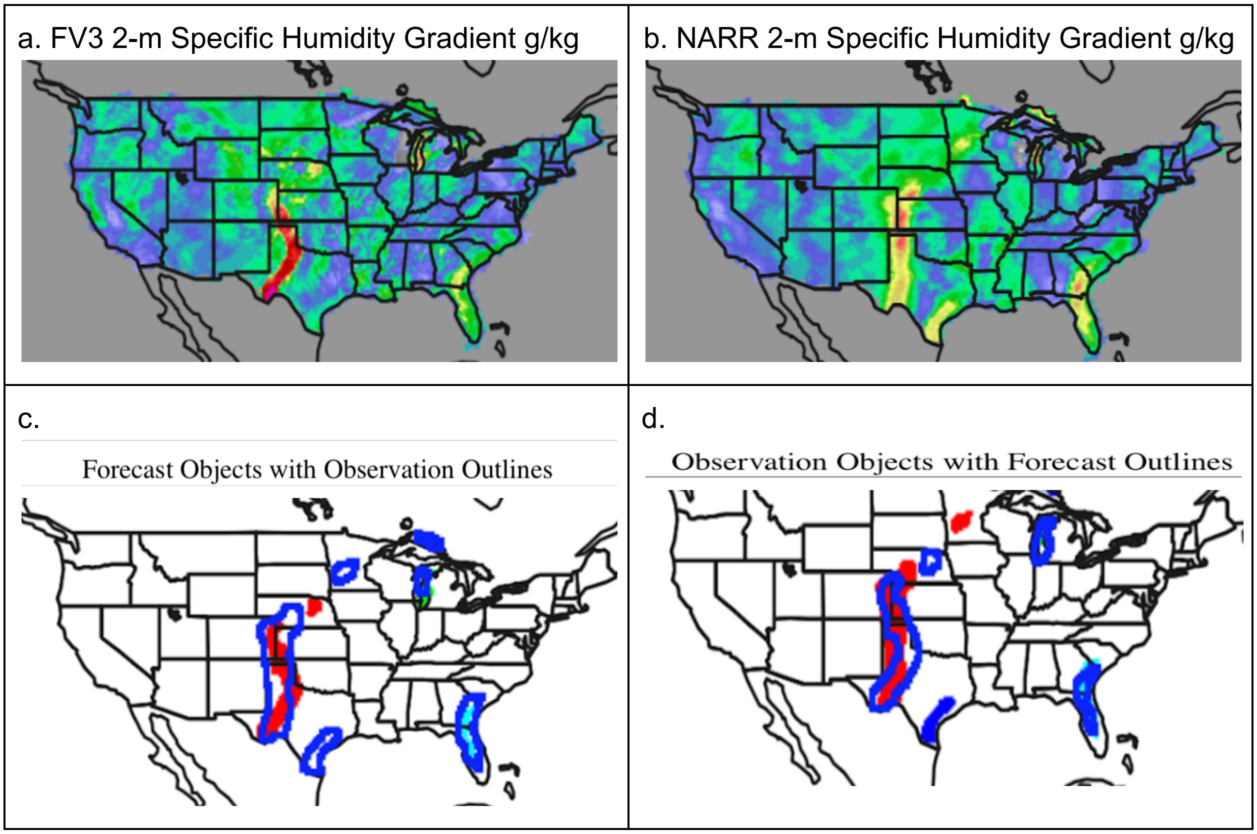

a. FV3 2-m Specific Humidity Gradient g/kg
b. NARR 2-m Specific Humidity Gradient g/kg
c.
Forecast Objects with Observation Outlines
d.
Observation Objects with Forecast Outlines

Figure S28: (a) the FV3 horizontal 2-m specific humidity gradient over a distance of 91km fields valid at 20170517 00Z, (b) as in (a) but for the NARR observations, (c) the forecast specific humidity gradient objects (red filled) overlaid with the observed specific humidity gradient objects (blue outlines), and (d) as in (c) but vice versa.

The forecast and observed objects that match the dryline were identified and the following statistics were calculated: object area, east-west centroid displacement, and north-south centroid displacement. The forecast dryline had a smaller area than the observed dryline by 124 grid points. The centroid of the model dryline was very close to the observed dryline, displaced about 0.56 grid points to the east and approximately 5.17 grid points to the south.

7. Summary

References

Bullock, R. G., B. G. Brown, and T. L. Fowler, 2016: Method for Object-Based Diagnostic Evaluation. NCAR Technical Note NCAR/TN-532+STR, 84 pp, doi:10.5065/D61V5CBS.

Clark, A. J., A. MacKenzie, A. McGovern, V. Lakshmanan, and R. Brown, 2015: An Automated, Multiparameter Dryline Identification Algorithm. Wea. Forecasting, 30, 1781-1794.

Coffer, B. E., L. C. Maudlin, P. G. Veals, and A. J. Clark, 2013: Dryline position errors in experimental convection-allowing NSSL-WRF Model forecasts and the operational NAM. Wea. Forecasting, 28, 746–761.

Hoch, J., and P. Markowski, 2005: A climatology of springtime dryline position in the U.S. Great Plains region. J. Climate, 18, 2132–2137.