NOAA Technical Memorandum OAR GSL-69
https://doi.org/10.25923/7ccf-pk47

Assessment of the Convective Weather Avoidance Model Upgrade

Tanya R. Peevey, Geary J. Layne, Joan E. Hart, Kenneth R. Fenton, Xue Wei, and Matthew S. Wandishin

October 2024

National Oceanic and Atmospheric Administration
Office of Oceanic and Atmospheric Research
Global Systems Laboratory
Assimilation, Verification, and Innovation Division
Boulder, Colorado

NOAA Technical Memorandum OAR GSL-69
https://doi.org/10.25923/7ccf-pk47

Assessment of the Convective Weather Avoidance Model Upgrade

By Tanya R. Peevey[1,3], Geary J. Layne[2,3], Joan E. Hart[2,3], Kenneth R. Fenton[3], Xue Wei[1,3], and Matthew S. Wandishin[3]

[1]Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO

[2]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO

[3]NOAA/GSL

October 2024

National Oceanic and Atmospheric Administration
Office of Oceanic and Atmospheric Research
Global Systems Laboratory
Assimilation, Verification, and Innovation Division
Boulder, CO

U.S. Department of Commerce
Secretary of Commerce Gina Raimondo

Under Secretary of Commerce for Oceans and Atmosphere
Richard Spinrad, Ph.D.

Assistant Administrator Oceans and Atmospheric Research
Steven Thur, Ph.D.

Contents

List of Tables

List of Figures

Abstract

The Quality Assessment Product Development Team (QA PDT) was tasked with assessing the recent upgrade of the Convective Weather Avoidance Model (CWAM) Weather Avoidance Field (WAF). The purpose of this assessment was to measure the performance of this recent upgrade; the current version of the CWAM WAF is used operationally in aviation forecast systems (e.g., the Route Availability Planning Tool and the Dynamic Routes for Arrivals in Weather system). The period for the assessment was June through August 2019. The assessment made use of air traffic information from the Traffic Flow Management System (TFMS) data from the Federal Aviation Administration (FAA). These observations were used to generate a database of historic flight routes which was compared to the CWAM WAF to assess the quality of the latter. For simplicity the previous product is called 'Legacy' and the current product is called 'New'. Both versions were evaluated.

An examination of the field characteristics between the two products showed that they differed over land and water primarily due to the New product using Corridor Integrated Weather System (CIWS) observations versus the Legacy product's use of a one-hour CIWS forecast. This resulted in differences between the New WAF product and the Legacy WAF, particularly near the outer range of weather radars that resulted in rings which were most strongly evident over Tennessee, Kansas, Oklahoma, and Texas. Additionally, the New WAF had gaps in coverage over the contiguous United States (CONUS) that were not present in the Legacy WAF, resulting in the large differences in the Big Bend region of Texas and throughout the mountain west.

In terms of performance, the Legacy WAF generally had a higher Pierce Skill Score (PSS) during the convective and high-traffic hours of the day, while the New WAF had higher scores during the non-convective hours of the day. Overall, the New WAF was better calibrated in the high layer (all regions) and in the eastern CONUS. The Legacy WAF had a higher Area Under the Curve (AUC) in the Receiver Operating Characteristic (ROC) plot than the Legacy WAF in the high layer (all regions) and in the middle layer of the eastern CONUS.

1 Introduction

This document summarizes the Quality Assessment Product Development Team's (QA PDT) assessment results from the recent upgrade of the Convective Weather Avoidance Model (CWAM) Weather Avoidance Field (WAF). The purpose of this assessment was to measure the performance of this recent upgrade; the current version of the CWAM WAF is used operationally in aviation forecast systems (e.g., Route Availability Planning Tool and the Dynamic Routes for Arrivals in Weather). For simplicity the previous product is called 'Legacy' and the current product is called 'New.' Both the Legacy and New WAF products were evaluated in this assessment. The Legacy WAF output is available every 1 kft and the New WAF at three vertical layers.

Currently, automated routing advisories generated with CIWS are used by traffic managers to make tactical routing decisions for en route air traffic. Quantifying the performance of the CWAM product, and by extension the information feeding route advisories, could lead to greater acceptance by pilots, thus reducing advisory revisions and work load for pilots and controllers. CWAM translates convective weather information into a probability field that communicates pilot preference for deviations that can be used to create more effective weather reroutes. Product skill was assessed by comparing WAF to flight track information, specifically looking at the intersection of WAF and air traffic. Results from this work could impact downstream products that ingest CWAM in addition to CWAM itself.

The assessment made use of Traffic Flow Management System (TFMS) data from the Federal Aviation Administration (FAA) that contains air traffic information. These observations were used to generate quantitative assessments of the quality of the product.

The document is organized as follows. Section 2 provides details of the datasets used in this assessment. The methods and evaluations used in this assessment, including how the flight data was used, are described in Sections 3 and 4, respectively. Assessment results are presented in Section 5. Finally, findings are summarized in Section 6. There are also two appendices with supplemental information.

2 Data

This section describes all datasets included in the assessment. The period for the assessment was June through August 2019.

2.1 CWAM

The Convective Weather Avoidance Model (CWAM), first developed a decade ago by MIT Lincoln Laboratory in collaboration with NASA (DeLaura et al. 2008), uses Vertically Integrated Liquid (VIL) and echo top fields from the Corridor Integrated Weather System (CIWS) and aircraft positional and temporal data from the Traffic Flow Management System (TFMS) to

create a Weather Avoidance Field (WAF). Echo top and VIL from CIWS are used to establish convective height and intensity, respectively. The intent is to translate raw weather information into aviation impact products, specifically for decreasing the workload of pilots and controllers by producing a probability field of pilot avoidance that can be used to create more realistic reroutes.

The original version of CWAM, Legacy WAF, used manually identified cases of pilot avoidance due to weather to build a lookup table based on VIL and ET fields. This lookup table is then used to generate the WAF. Recently there have been efforts to upgrade CWAM using more data, additional CIWS products, increased computational capabilities, and advancements in computational techniques. In the new version of CWAM, New WAF, weather is combined with pilot avoidance to train a Convolutional Neural Network (CNN) to produce a WAF (Mattioli et al. 2020). To generate the WAF, first the aircraft's en-route position relative to its planned position is assumed to have a Gaussian distribution. Then the deviation distance is modeled as a Chi-squared random variable. More specifically, a Chi-square distribution with $\boldsymbol{k}$ degrees of freedom can be created that is the sum of the squares of $\boldsymbol{k}$ independent gaussian random variables. Each point along the trajectory is then assigned a $\boldsymbol{k}$ value that is fed into the CNN along with VIL and ET fields from CIWS (Mattioli et al. 2020). After the CNN is trained, a new dataset is created for the upgraded CWAM from an aircraft avoidance database merged with weather information. This approach supports automated avoidance detection using supervised machine learning with an expanded flight trajectory dataset.

WAF represents the probability (0-100%) a pilot will fly around hazardous weather and is generated every 5 minutes over the Contiguous United States (CONUS). It is on the CIWS grid, which is Lambert azimuthal equal-area, with a 1-km horizontal resolution. Both versions of CWAM are available in 5-minute timesteps, although the QA PDT only received four of those files per hour [e.g., 0-5, 15-20, 30-35, and 45-50 minutes]. The new algorithm was trained using weather information from the 2018 convective season. The Legacy WAF output is available every 1 kft from 29 to 39 kft plus the 40+ kft 'High' layer. The New WAF output contains three vertical layers:

Low	21 ≤ altitude < 32 kft
Middle	32 ≤ altitude < 38 kft
High	38+ kft

CWAM output is currently available and serves as an input to operational products. This assessment will use WAF data for the retrospective period of Summer 2019, a convectively active time of year. An example of the WAF output over the CONUS is shown in Figure 1.

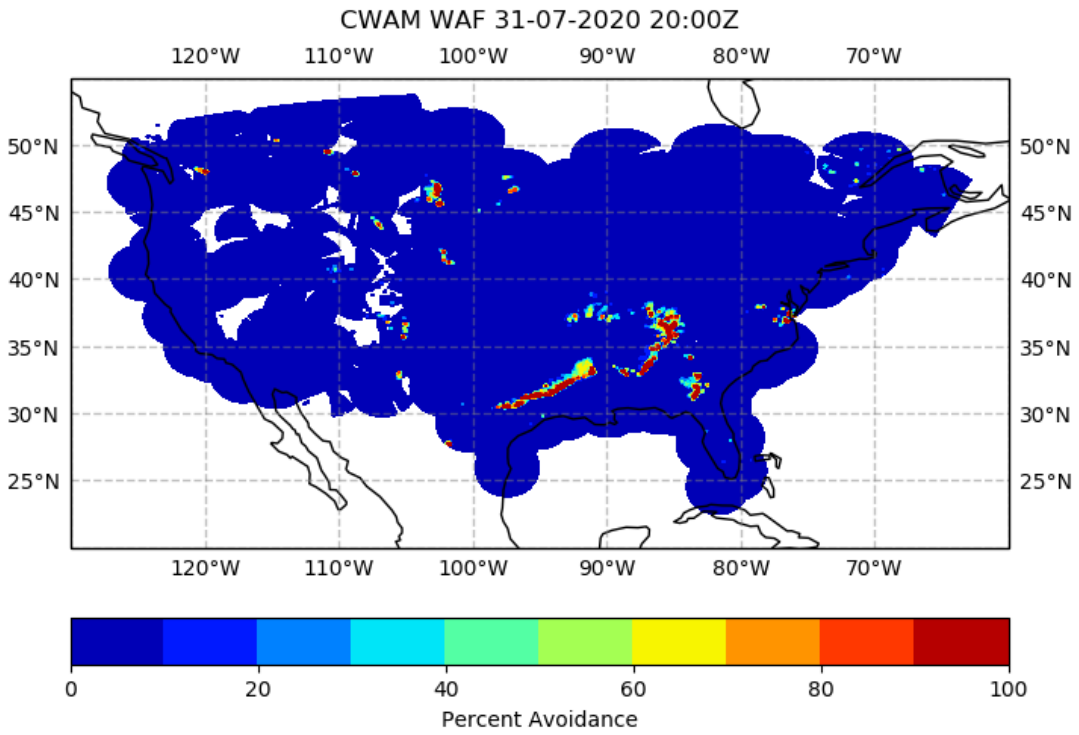

Figure 1. CWAM WAF probability or percent avoidance over the CONUS for June 2nd 2019 21:00Z. Data is from the middle layer of the New WAF product.

2.2 Air Traffic Information

Flight information was obtained from data originating from the Traffic Flow Management System (TFMS), containing both positional and temporal information of all flights regulated by the FAA. Originally three different baselines were proposed, but analyses during the assessment revealed that only the historic traffic would be an appropriate baseline and, as a result, that is the only baseline presented in this report. Supporting material for this decision can be found in Appendix 8.1. Both the historic traffic and actual flights contain latitude, longitude, altitude, and time information and so no airport mask was applied, as originally proposed. The historic traffic represents where planes were typically located for a given time-of-day and day-of-week (e.g., a time-of-day, day-of-week average) and attempts to incorporate typical non-weather deviations, such as flight delays and cutting corners. The historic traffic used in this assessment was generated with data from the summer of 2018 and 2019. An example of pilot deviation due to weather is shown in Figure 4. The TFMS data was transferred in real-time via a web service. Though information for all airports and airlines were available, this analysis focuses on airlines departing from or arriving at one of the 35 primary airports. There is no set temporal resolution since the messages come in as issued. Multiple messages are included in the data feed and these messages were used in this assessment to determine to the following:

1. Historic Traffic
2. Actual Flight or Track Data

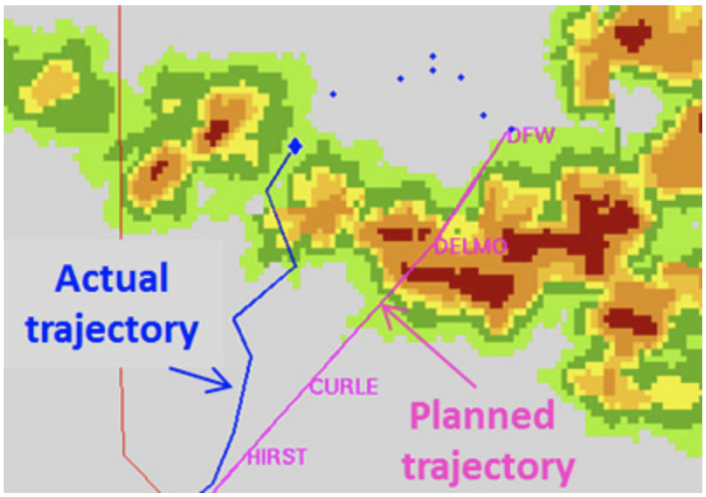

Figure 2.. An example of pilot deviation due to weather. Image obtained from https://www.faa.gov/nextgen/programs/weather/tfm_support/translation_products/.

3 Methods

3.1 CWAM Mask

Both versions of the CWAM product covered the CONUS, with the Legacy product extending further out over the ocean than the New product. This is highlighted by the magenta regions in Figure 3. The difference in spatial coverage between the two products could be due to either the version/type of CIWS used (i.e., Legacy uses the CIWS forecast and New uses CIWS observations) and/or the CWAM algorithm that produces the WAF product. The mask, shown in green, represents the overlapping coverage area between the new and legacy product and was applied to both products before processing the data and comparing their statistical performance.

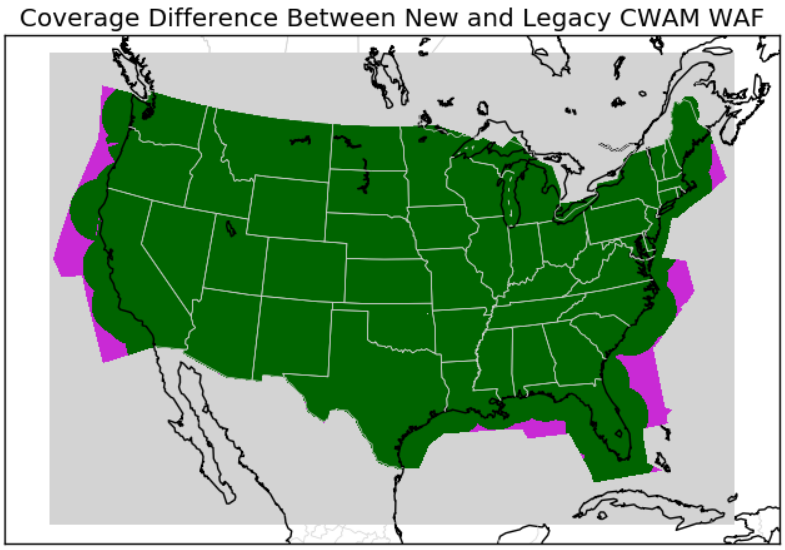

Figure 3. Difference in coverage between New and Legacy CWAM WAF products. The filled contours represent the following: Grey is the CIWS grid, Magenta is where the Legacy WAF product extends beyond the New WAF product, and Green is the mask that is applied to both products such that everything outside is not included in the analysis.

3.2 Stratifications

3.2.1 Probability Stratifications

For this assessment of WAF performance was evaluated across a set of thresholds that encompassed the full field of available probabilities. The specific values are 0, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100%. Each probability threshold corresponds to the number of aircraft that should have avoided that area, e.g., a WAF of 50% means that 5 out of every 10 aircraft should have been diverted around that area.

3.2.2 Vertical Stratifications

Flight data was placed into altitude bins corresponding to the vertical dimension of the New WAF data.

3.2.3 Temporal Stratifications

Product performance was evaluated by time of day since both convection and flight activity vary at that scale. Specifically, convection peaks between 1800 and 0000 UTC and air traffic peaks between 1500 and 2200 UTC.

3.2.4 Geographic Stratifications

Performance was also evaluated over the CONUS and five sub regions (Figure 4).

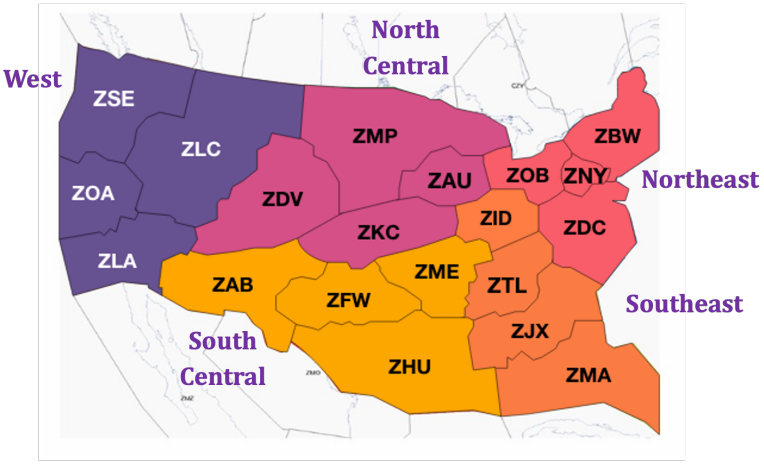

Figure 4. Map of geographic regions.

3.3 Techniques for Evaluating Product Output against Flight Observations

3.3.1 Flight Avoidance

To create the flight avoidance, first the flight density grids were generated by isolating the flight data within each 5-minute WAF period and then matching it spatially onto the horizontal and vertical dimensions of the WAF grid. Each time a flight passed through a pixel and layer that grid box was incremented by 1. This was repeated for the actual flights and historic traffic. The next step was to overlay WAF values onto all traffic density grids. A schematic of this process is shown in Figure 5.

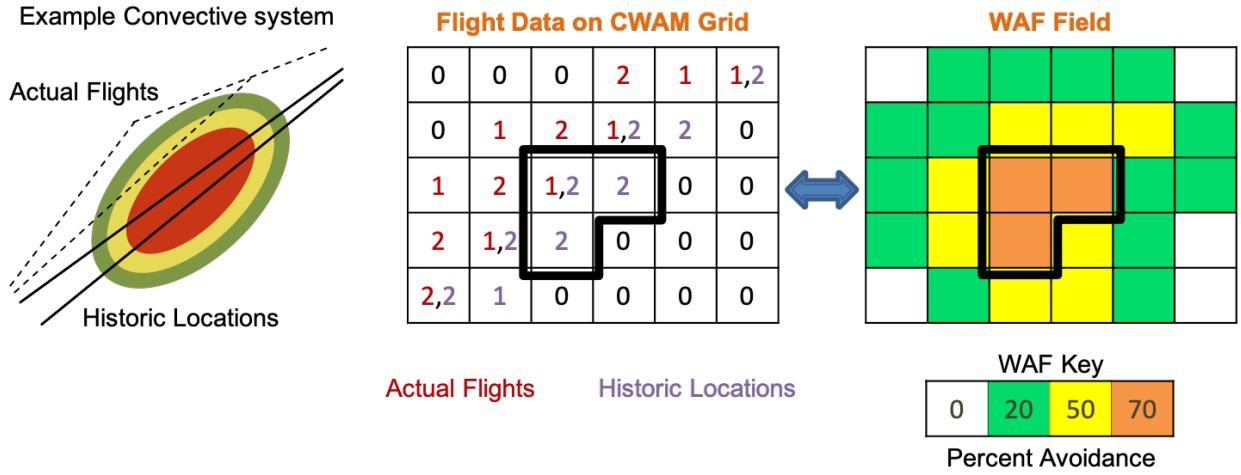

Figure 5. Schematic of Methodology used for assessment. Process is from left to right, with the left being the raw flight data and paths relative to a convective system. Center grid represents the flight density field and the rightmost grid the WAF field for the same pixels.

Finally, where traffic density is nonzero, the distribution of WAF probability values for the actual and baseline data were compared. To accomplish this, for each given WAF value, flight density counts were aggregated further, i.e., the flight data illustrated in Figure 5 (middle) were

aggregated for all active and historic flights, separately, to produce flight density values as shown in Figure 6 (numbers) overlaid on a WAF grid (color fill). Specifically, flight data was aggregated over the same 5-minute periods as the CWAM data and then again to hourly increments to ensure a large sample size for the analysis. This process was repeated for all probabilities listed in Section 3.2.1, creating data like the example shown in Table 1. With those numbers, flight avoidance was then calculated using Equation 1, where the number of actual flights was compared to historic traffic for each WAF value. Results from this illustrative example are shown in the two rightmost columns of Table 1. By repeating this process for the proposed stratifications, paired datasets were generated and used to create distributions to evaluate product performance.

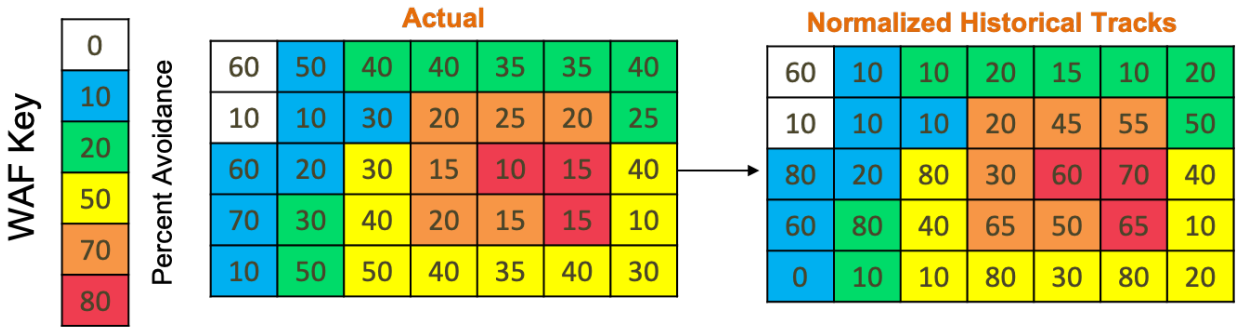

Figure 6. Example of what flight density fields could look like for the actual flights (left) and historic traffic (right). The numbers represent the flight densities while the colors represent WAF values.

Table 1.. Example of a flight density table used to calculate flight avoidance. This is illustrative only; the real table has two more columns containing flight density values for the other flight plans and the corresponding avoidance field.

WAF	Actual	Historic	Historic avoidance
0	70	60	-17%
10	250	195	-28%
20	295	240	-23%
50	315	395	20%
70	115	300	62%
80	40	230	82%

$$Flight\ Avoidance = \left(1 - \frac{actual}{baseline}\right)\ x\ 100 \qquad (1)$$

3.3.2 Comparing Legacy and New WAF Fields

The New and Legacy WAF products have different vertical dimensions, the Legacy WAF with multiple levels (every 1 kft) and the New WAF with three vertical layers. To compare these products, the New WAF field was evaluated against the maximum of the Legacy WAF field within each New WAF layer (Figure 7).

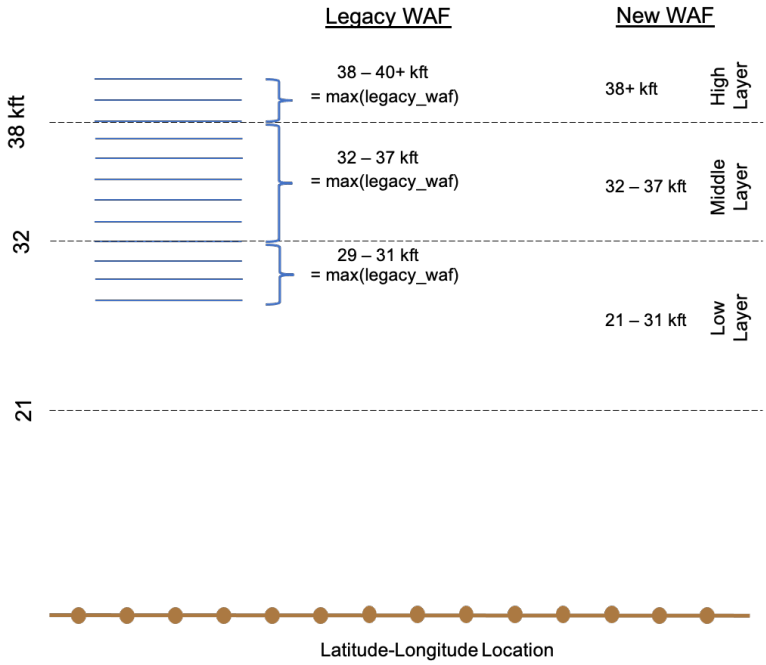

Figure 7. Graphic illustrating how the 1 kft layers from the Legacy WAF product were grouped for comparison to the New WAF product. Vertical layers or levels for each CWAM WAF product are also annotated on the schematic for quick reference.

To confirm that taking the maximum of the Legacy product within a layer was a reasonable representation of that layer, the variability within the layer was examined along with the alternative approach of taking the mean rather than the maximum. The differences between distributions of the New WAF and that of the original the Legacy WAF, the maximum of the Legacy WAF, and the mean of the Legacy WAF are shown in Figure 8. Note that the lowest layer is not shown since the Legacy product was not available at all levels between 21 and 31 kft and, as a result, there was no clean way to compare the products in that layer. Only results from the middle and high layers are presented in this report.

In Figure 8, taking the mean of the Legacy product resulted in the most swings in the distributions due to the mean being sensitive to outliers. The standard deviation within each layer

for the Legacy product (not shown) was also examined and the vast majority of values within a layer were equal to the maximum. The few that did deviate from the maximum value in the layer were evident in the mean (Figure 8). Taking the maximum produced a distribution that was similar to that of the New WAF product and was closest to the distribution of the original Legacy WAF product, especially for the high layer.

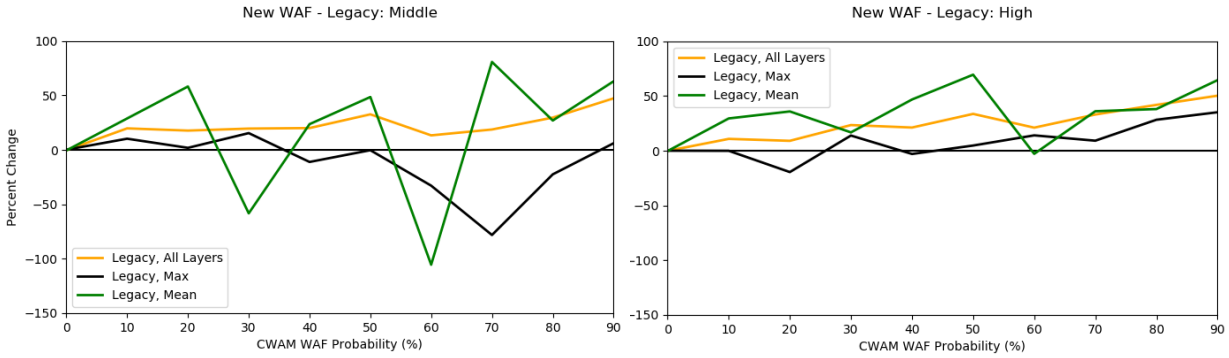

Figure 8. Difference, as percent change, between the New WAF and Legacy WAF distributions for the middle (left) and high (right) layer. Here the last bin is inclusive, so it represents counts of when WAF equals 90 or 100. The difference between the original (yell ow), maximum (black), and mean (green) of the Legacy WAF and New WAF are shown. A line below/above zero means the Legacy/New product has a high occurrence in that bin.

4 Evaluations

4.1 Field Characteristics

The makeup of the CWAM WAF avoidance fields were evaluated using value-based distributions and climatological maps. These visualizations utilize the probability values listed in Section 3.2.1 such that the avoidance fields were binned to align with these values. Both raw counts and percent change were evaluated. Additionally, similar graphics were created and analyzed for flight data.

4.2 Performance Statistics

Performance was evaluated using both reliability diagrams and ROC curves. The reliability diagram measured the agreement between the forecast and observation values, conveying how well the class of probabilities were estimated, e.g., for this storm did planes deviate when and where they were predicted to do so. The y-axis of the diagram is observed avoidance instead of observed frequency and was calculated using Equation 2. This approach was taken since observed avoidances tend toward 0 or 100% due to the binary nature of the dataset. These data characteristics only affect the reliability diagram and not the ROC curves since the ROC was calculated using thresholds that are greater than or equal to rather than equal to. The reliability

score (distance from the diagonal line) was measured using the reliability term of the Brier Score as shown in Equation 3.

$$\underline{o_k} = 1 - \frac{actual_k}{historic_k} \quad ; \quad for\ k = 0.1, 0.2, 0.3, \ldots, 1.0 \tag{2}$$

$$Reliability = \frac{1}{N}\sum_{k=1}^{K} n_k\left(p_k - \underline{o_k}\right)^2 \quad ; \quad n = number\ of\ pixels. \tag{3}$$

The reliability diagram was conditioned on the forecast and makes a good companion to the ROC curve which was conditioned on the observations. The values in the ROC curve were computed from the same distributions using the thresholds listed in Section 3.2.1. The Area Under the Curve (AUC) was calculated using the trapezoid rule. Definitions of Probability of Detection (POD) and Probability of False Detection (POFD) used to generate the ROC curves are shown in Tables 2 and 3. Additionally, the Pierce Skill Score (PSS) was also evaluated when stratifying by the time of day. Performance was considered overall and for the stratifications described in Section 3.2.

Table 2. 2x2 contingency table.

		Observation	
		Yes	**No**
Product Output	**Yes**	Hit	False Alarm
	No	Miss	Correct No

Table 3. List of statistics along with corresponding mathematical definition and description.

Statistic	Definition	Description
POD:	$\frac{Hits}{Hits + Misses}$	Probability of Detection, proportion of all observed events that are correctly forecast to occur. (Range: 0 to 1. Perfect Score: 1)
POFD:	$\frac{False\ Alarms}{False\ Alarms + Correct\ Nos}$	Probability of False Detection, proportion of all observed events that are incorrectly forecasted. (Range: 0 to 1. Perfect Score: 0)
PSS:	POD - POFD	Peirce's Skill Score, includes all elements of the contingency table and addresses how well the forecast separates the 'yes' from the 'no' events. (Range: -1 to 1, 0 = no skill. Perfect Score: 1)

4.3 Case Studies

Observations and product output were also examined in detail for a few cases to better understand and highlight differences between the New and Legacy WAF as seen in their field characteristics or statistical results.

5 Results

5.1 Field Distributions

Distributions and climatological maps of flight data, both historic traffic and tracks, are shown in Table 4 and Figure 9. For historic traffic, the total number of counts and the distribution of those counts across the layers was well matched against the track data. This indicates that historic traffic was a good dataset for comparison against actual tracks and for use in calculating the flight avoidance. To make sure their behavior was as expected, historic traffic avoidance was compared to actual tracks avoidance by overlaying each on the WAF, resulting in Figure 10. This figure shows the density of traffic as the number of flights per km^2 per hour, which were small numbers because planes avoided one another by large distances. Flights were normalized beforehand to remove any bias due to historic traffic and track data having a different number of total flights. Plots were created for both the New and Legacy WAF. Here the trend in historical traffic line (light grey) was almost unchanged as the CWAM WAF probability values increase, whereas the track data showed a decreasing trend. This difference highlights that the historic traffic was a good baseline for comparison to the track data. Additionally, the decrease in traffic density was correlated with increasing WAF percentages in both versions of the CWAM. The decreasing trends in the traffic density for the New and Legacy WAF were slightly different because the track data was conditioned on the WAF field and, from the distributions in Figure 10, these fields were not exactly the same.

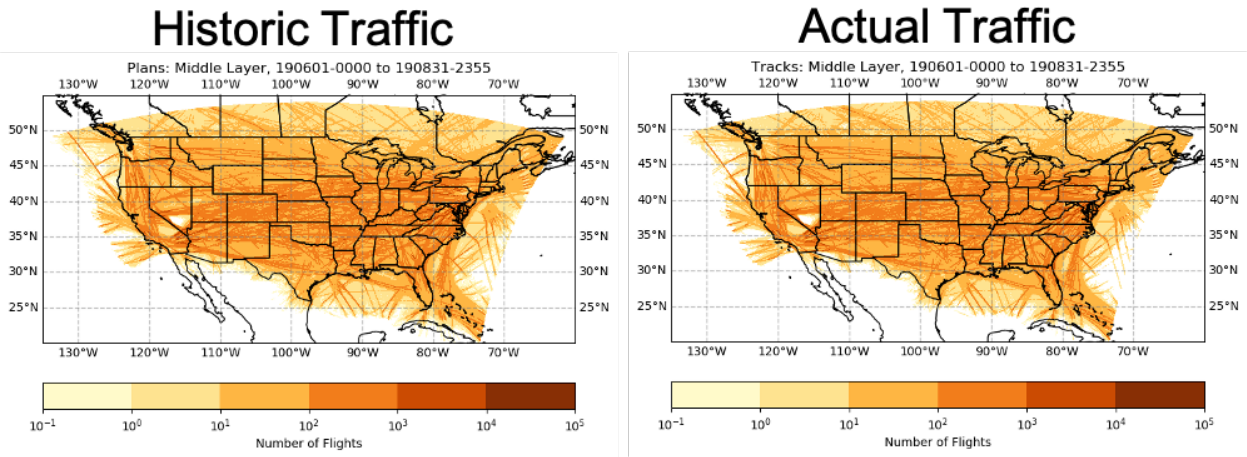

Figure 9. Distribution of flight data on the CIWS grid for the middle layer for historic traffic (left) and track data (right).

Table 4. Flight data counts by vertical layer (rows) for the historic traffic and the track data (columns). For each layer, raw counts and the percentages are shown, e.g., for the first row and first column (275,287,328/1,034,246,810) = 27%.

Layer	Historic	Tracks
Low: 21-31	275,287,328 (27%)	258,375,605 (27%)
Middle: 32-37	490,561,469 (47%)	450,296,467 (47%)
High: 38+	268,398,012 (26%)	247,481,438 (26%)
Total	1,034,246,810	956,153,510

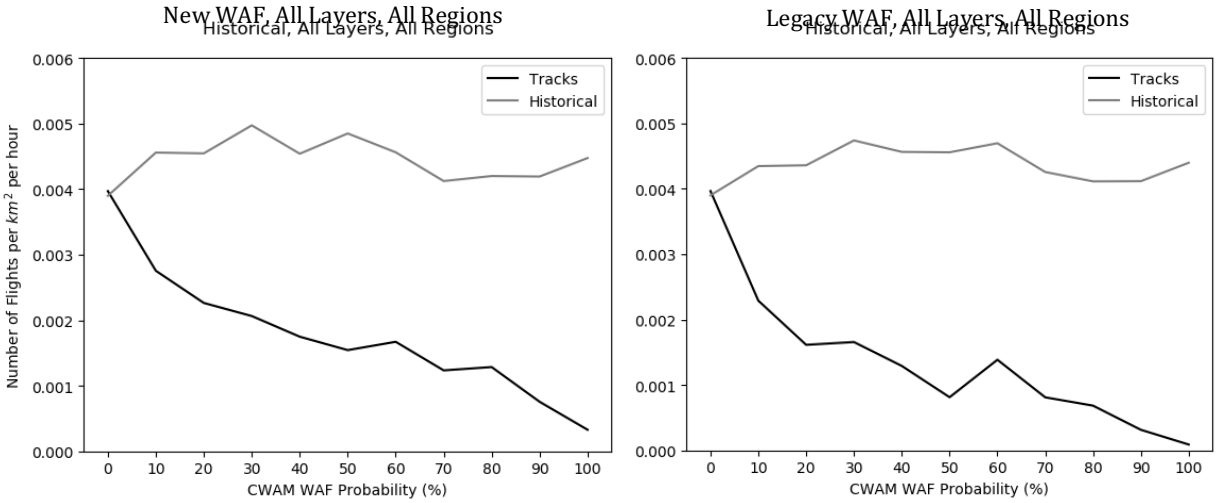

Figure 10. Number of flights per km^2 per hour conditioned on the CWAM WAF probability, showing both New (left) and Legacy (right) WAF. The grey line represents the results with the historic traffic and the black line the track data.

Climatological maps are shown in Figures 11 and 12 with differences in the marginal distributions also shown in Figure 12. The first set of maps in Figure 11 shows frequency of occurrence, e.g., how often the New WAF equaled 50 in this pixel, for both the middle and high layers combined. In this figure a couple features stand out. First, both products had their highest occurrence over the Great Plains and Southeast; the Legacy WAF had a higher event rate than the new WAF over the Great Plains, while the opposite was true in the Southeast. Second, radar rings were apparent in the New WAF and the difference plots. The radar rings over land could be due to how the machine learning algorithm processes the CIWS data, producing visible over- and under-sampling. Radar gaps over land and radar rings over the coastal waters were in part due to the products ingesting different CIWS products into their algorithms. The newer product uses the

CIWS mosaic product and the Legacy product the CIWS forecast. Specific examples of these characteristics are shown and discussed further in Section 4.3.

When examining all CWAM avoidance values (Figure 12), the New WAF tended to have higher event rates, specifically in the high layer when WAF was greater than or equal to 30. In the middle layer the Legacy product had a higher event rate when probabilities equaled 60 and 70%, specifically over the Great Plains and where there were issues with radar coverage (i.e., coastal areas and in the mountains). Note that the maps show the climatology when WAF is greater than and equal to 50%, and so can only be compared to the marginal distributions where they also were greater than or equal to 50%.

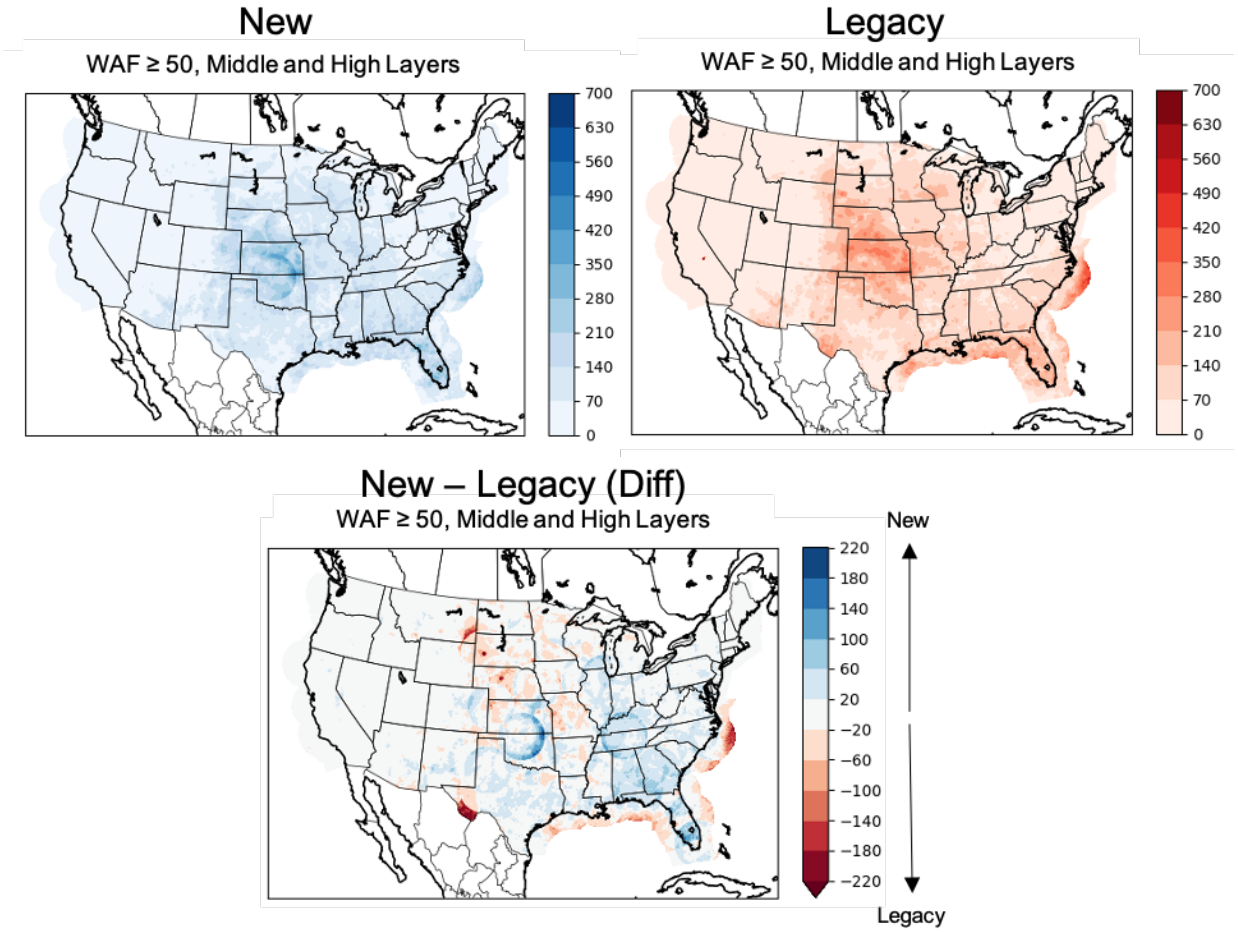

Figure 11. Climatological maps of frequency of occurrence for the New (upper left) and Legacy (upper right) WAF products along with their difference (lower center). All plots are showing the middle and high layers combined and when the WAF field is greater than or equal to 50. In the difference plot, blue/red indicates that the New/Legacy WAF has a higher event rate.

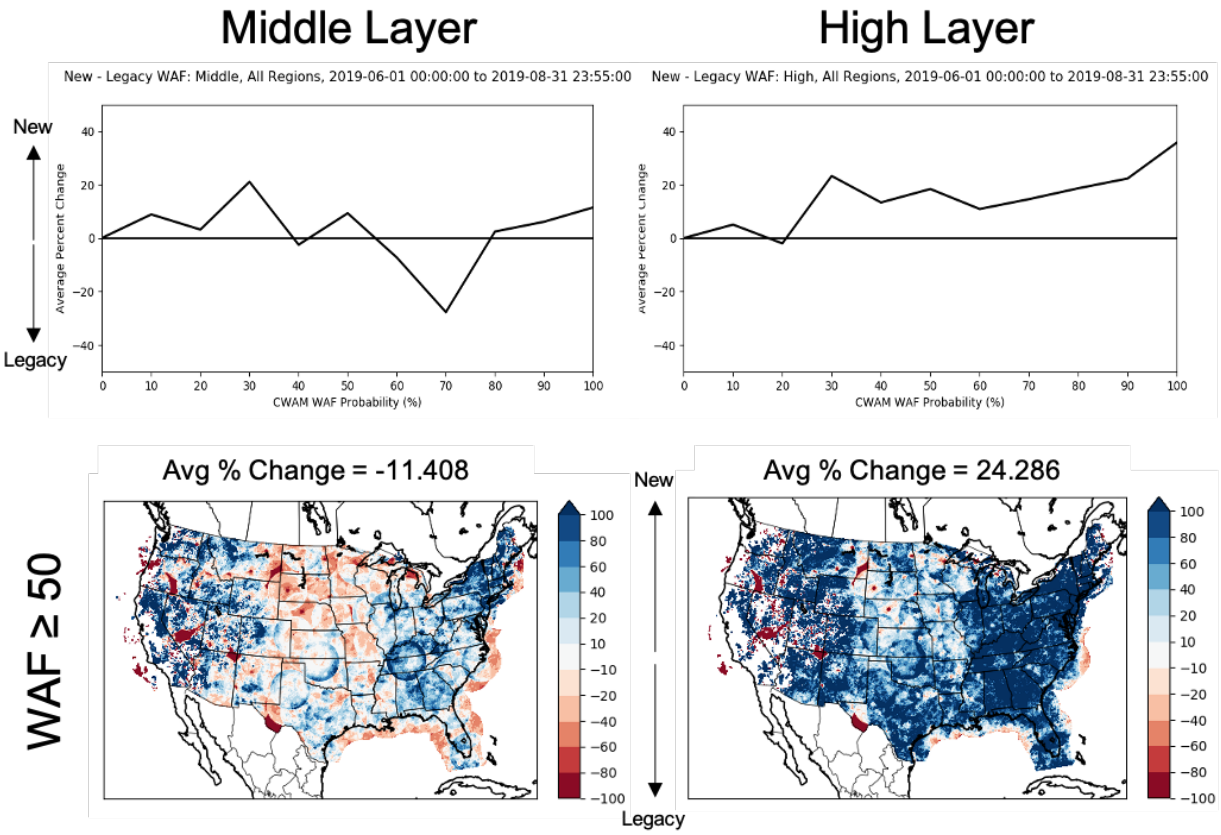

Figure 12. Difference in the average percent change distributions (upper row) and climatological maps of frequency of occurrence (lower row) for the middle (left column) and high (right column) layer. For all plots the average percent change is obtained by calculating the percent change for CWAM WAF probability value for each geographic pixel, then summing all of those values and dividing by the number of pixels. In the climatological maps the percent change values are represented by the filled contours, where blue/red means the New/Legacy WAF has a higher event rate. The average percent change is in the title.

5.2 Performance

The relative performance of the two CWAM WAF products was evaluated using dichotomous statistics, producing POD, POFD, and PSS scores, and diagrams, specifically reliability diagrams and ROC curves. The scores and diagrams are described in Section 4.2. Scores were stratified by time-of-day and by region.

Statistical performance by the time-of-day for a threshold of 60 and for both layers combined was evaluated using POD, POFD and PSS as shown in Figure 13. In these plots blue indicates that the New WAF performed better and red that the Legacy product performed better. The New WAF product had a higher POD for almost all hours of the day and a low POFD for the majority of the non-convective hours of the day (~0-16 UTC). The Legacy product had a better POFD during the convective hours of the day (~16-23 UTC), thus opening up more airspace during

periods of high air traffic (Figure 13, gray bars). Combining these values resulted in a PSS that was better for the New WAF during the non-convective hours and the Legacy WAF during convective hours primarily due to larger changes in the POFD. This tendency corresponded with the period when the Legacy product had a lower event rate over the CONUS (Figure 15, lower right). However, this was weighted more by the middle layer.

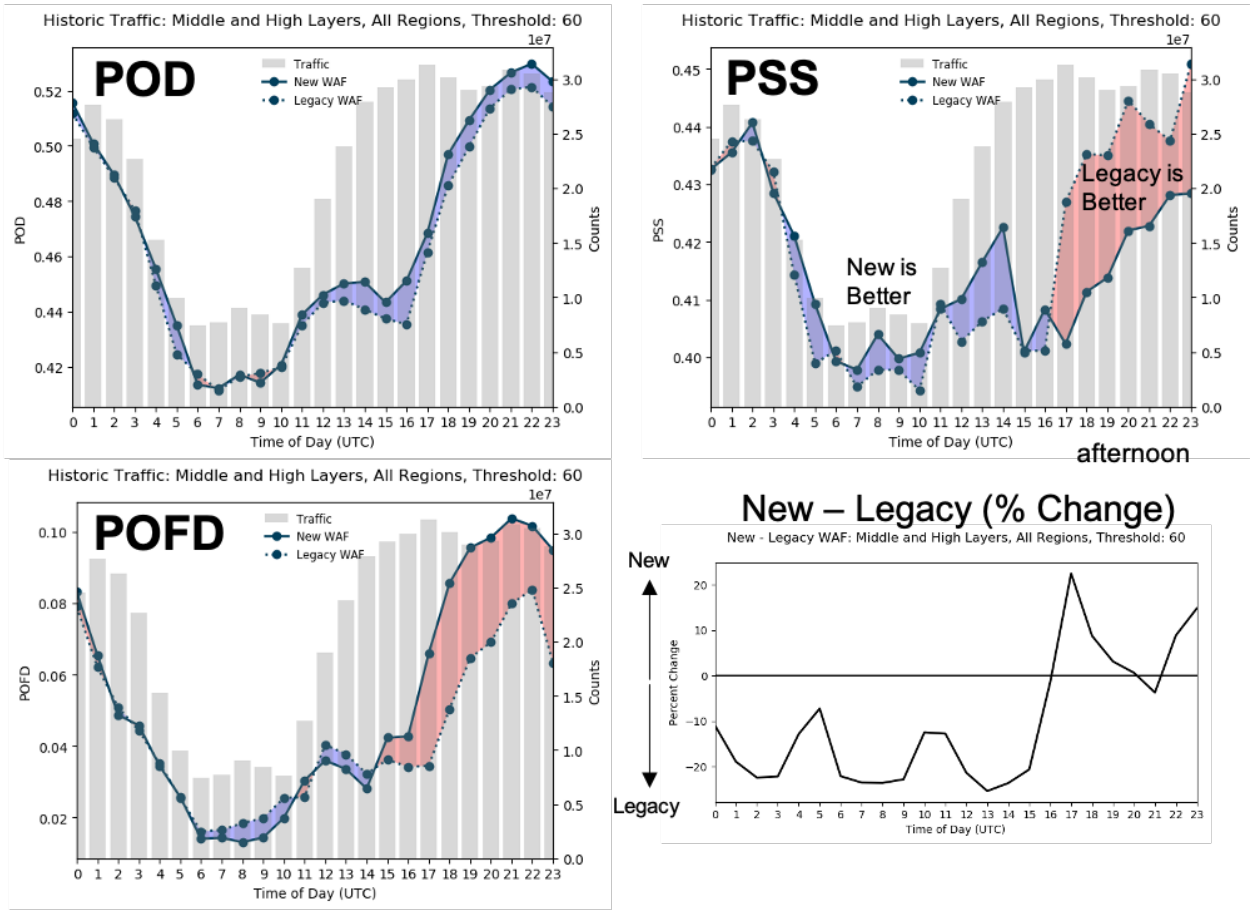

Figure 13. Statistical performance by time of day in UTC for a threshold greater than or equal to 60%, including POD (upper left), POFD (lower left), and PSS (upper right). Both layers are combined. Blue shading indicates times that the New product has a better score; red shading indicates times the Legacy product has a better score. The percent change [((New-Legacy)/Legacy) x 100] in the products distributions (lower right) shows at what time of the day each product has a higher frequency of occurrence. Gray bars show hourly traffic volume.

These results were also broken out by layer and evaluated for two thresholds, 20 and 60%. Here only PSS is shown, but POD and POFD graphs are available in Appendix 8.2. For these PSS plots the scales are different but the difference between the maximum and minimum values is the same so that the scale of the change can be compared. In the middle layer (Figure 14), the Legacy product had higher scores for both thresholds during the convective and high air traffic

hours due to a lower POFD. The New WAF had a higher PSS score for the lower thresholds during non-convective and low air traffic hours due to a high POD and low POFD. At the higher thresholds Legacy had higher PSS scores due to its POD surpassing the POD of the newer product. In the high layer (Figure 15), the Legacy product still had a higher PSS during the convective hours, again due to the consistently low POFD. The New WAF did not start to outperform the Legacy product until around a threshold of 60%, where the New WAF had a higher POD, partially the result of a higher event rate in the New WAF. Generally, the Legacy product had a higher PSS during the hours of the day that it had a lower frequency of occurrence. This was true in both layers (Figures 14 and 15), whereas the New WAF tended to require a higher event rate to obtain a higher PSS.

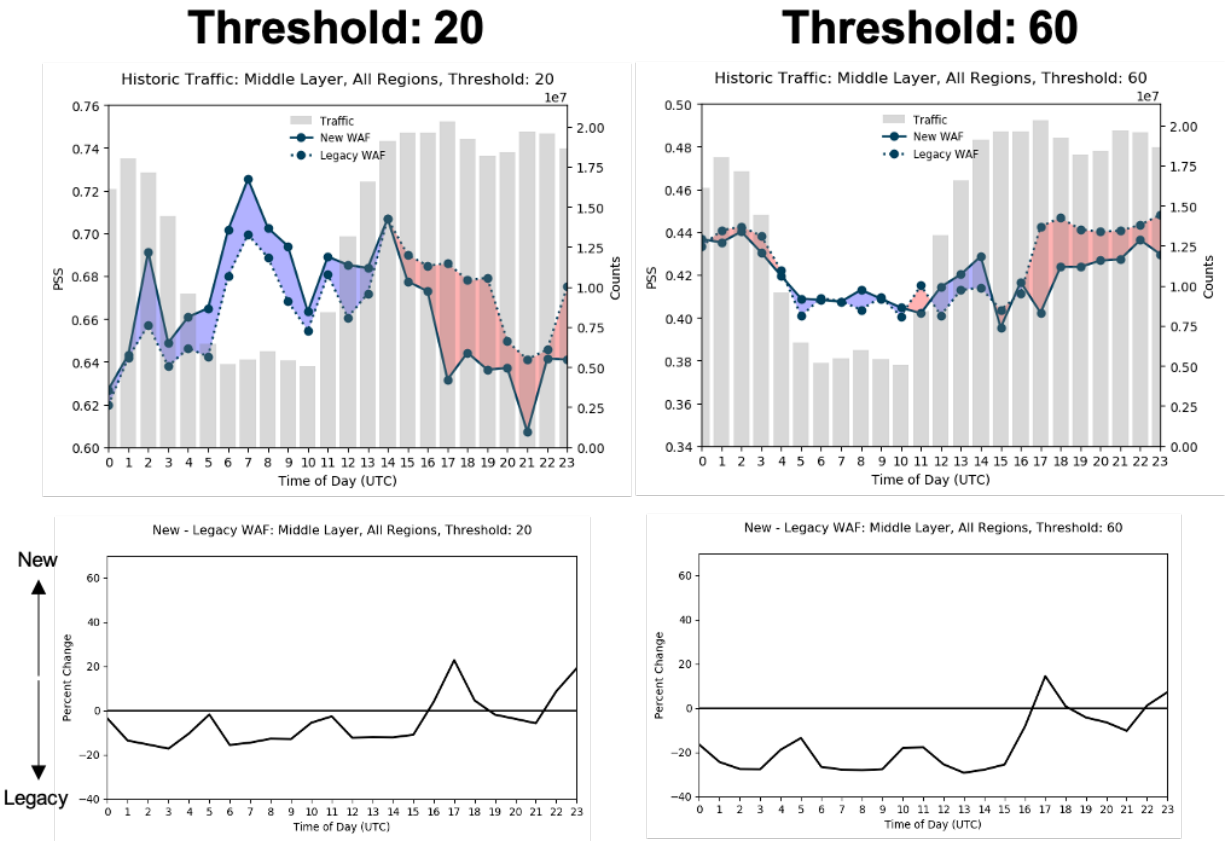

Figure 14. PSS (top rows) by time of day in UTC for the middle layer and for thresholds greater than or equal to 20% (left column) and 60% (right column). Blue shading indicates times that the New product has a better score; red shading times that the Legacy product has a better score. The percent change [((New-Legacy)/Legacy) x 100] in the products distributions (bottom row) shows at what time of the day each product has a higher frequency of occurrence. Gray bars show hourly traffic volume.

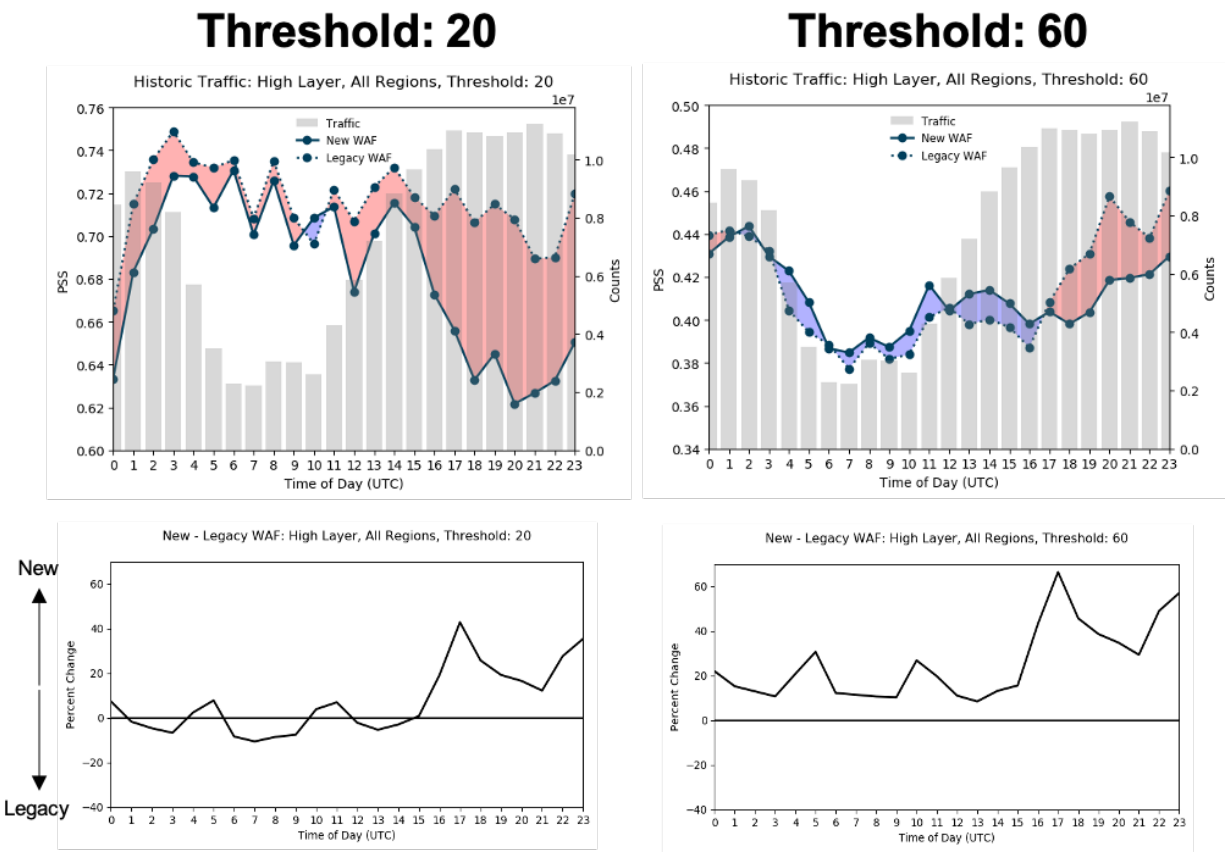

Figure 15. Same as Figure 14 but for the high layer.

Overall performance was evaluated using ROC curves as shown in Figure 16. This evaluates if the product was a good classifier, e.g., can discriminate between observed events and non-events. Only the observation threshold of 40% is shown but other thresholds produced similar relative results. This threshold was applied to the calculated observed flight avoidances (see Section 3.3.1). When looking at the overall performance, the Legacy product had a slightly higher AUC because the line is closest to the upper left corner of the plot. When breaking the results out by layer the Legacy WAF outperformed the New WAF in the high layer and they were about equivalent in the middle layer. The interplay between POD and POFD can be seen here, e.g., in the high layer the POD for the New WAF was higher than that for the Legacy WAF, but the lower POFD for the Legacy WAF was more significant. However, both products had AUC values greater than 0.9. When stratifying each layer by region the Legacy product had a higher AUC for all regions in the high layer, with a similar value from region to region (Figure 18). In the middle layer the Legacy WAF outperformed the New WAF in the Southcentral, Southeast and Northeast regions, all regions with high air traffic (Figure 17). Additionally, the New WAF performed best in the West, where the AUC equaled ~0.94, and worst in the Southeast, where the

AUC equaled ~0.85. Only four observation thresholds are shown in Figure 19 since the other thresholds had similar results.

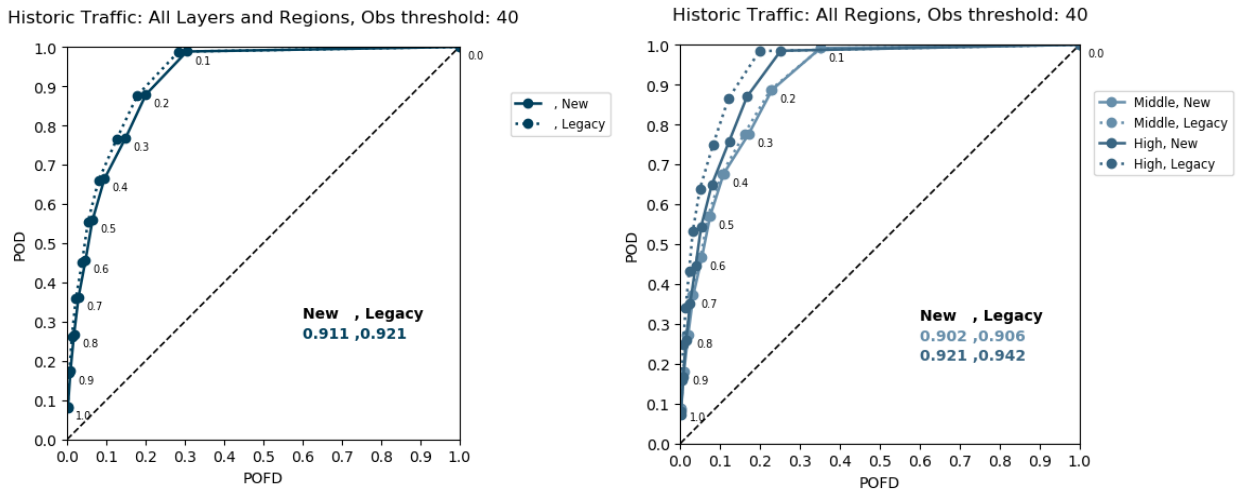

Figure 16. ROC Curves for the New and Legacy WAF with combined layers (left plot) and stratified by layer (right plot). Solid lines represent the New WAF and dotted lines the Legacy WAF. Observed flight avoidances are stratified using a threshold of 40% and AUC values, calculated using the trapezoid rule, are listed on each plot.

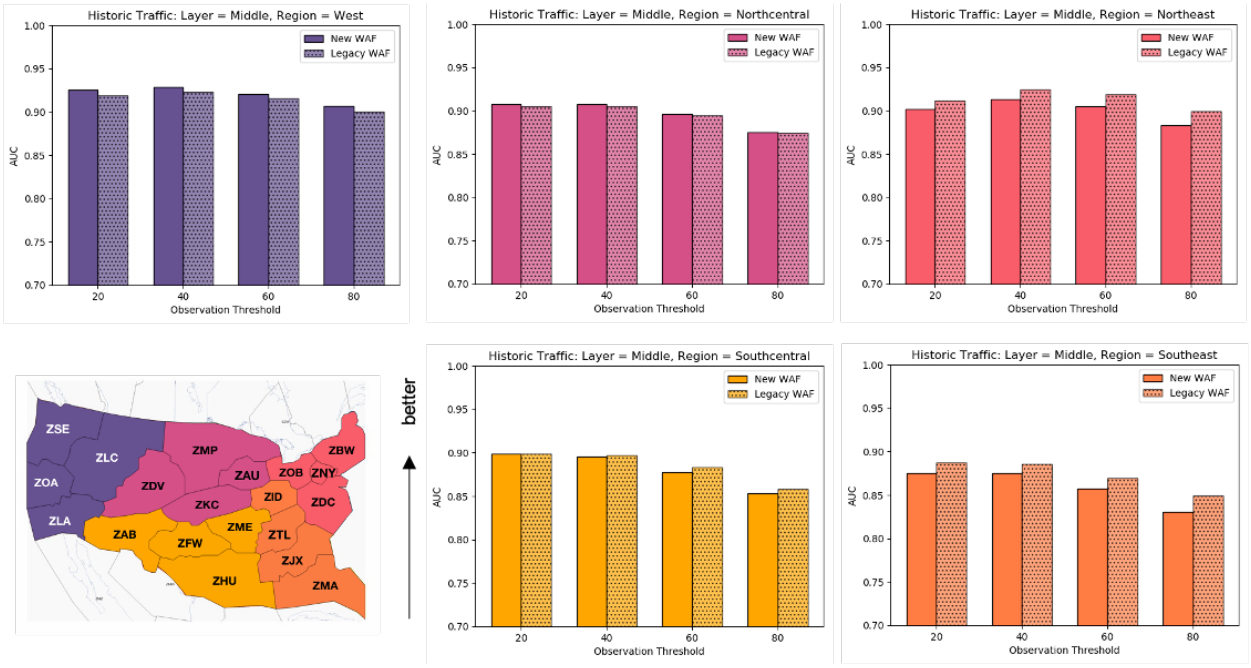

Figure 17. AUC values by region, for the middle layer and for four observation thresholds, 20, 40, 60, and 80%. The colors used for the bar charts match the colors used for each region (lower left; Section 3.2.4). From left to right and top to bottom the order is West, Northcentral, Northeast, Southcentral, and Southeast.

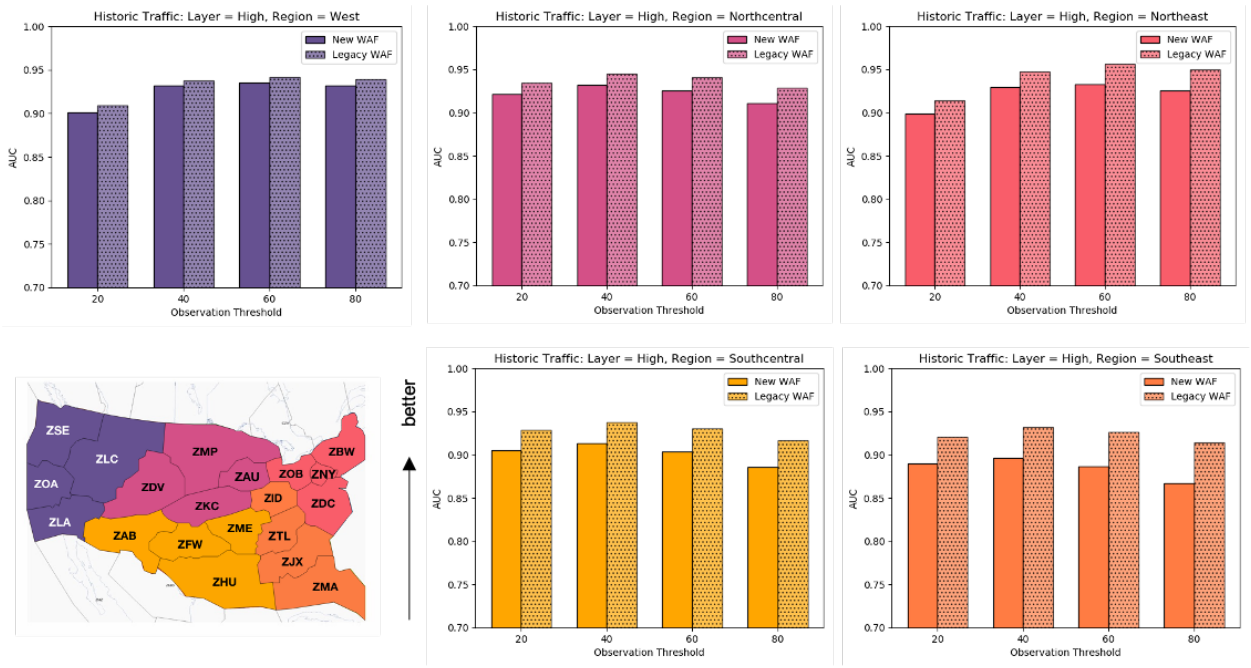

Figure 18. Same as Figure 17 but for the high layer.

Reliability diagrams were also used to evaluate overall performance since they were conditioned on the forecast rather than the observation and thus complement ROC curves. Results are shown when no stratifications were applied and when the products were stratified by layer (Figure 19). When both the middle and high layers were combined, the products had nearly equal reliability; however, when broken down by layer the products had similar reliability in the middle layer and the New product was more reliable (closer to the diagonal line) in the high layer. When stratified by region, the New WAF was more reliable in the eastern USA, a region of high air traffic, whether evaluating the layers individually or combined (Figures 20 and 21). However, the difference between the two products was largest in the middle layer. In the middle layer, the Legacy product was more reliable in the three western regions (Figure 21; left plot). In the high layer, the New WAF was more reliable in all regions (Figure 21; right plot). Overall, both products were more reliable in the middle layer and less reliable in the higher layer.

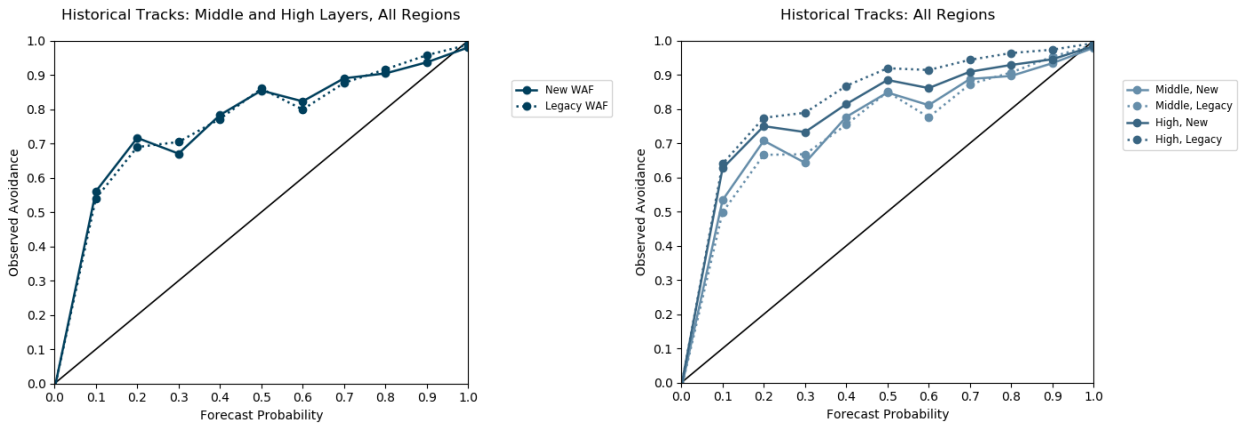

Figure 19. Reliability diagrams when not applying stratifications (left plot) and when stratifying by layer (right plot). Solid lines represent the New WAF and dotted lines the Legacy WAF.

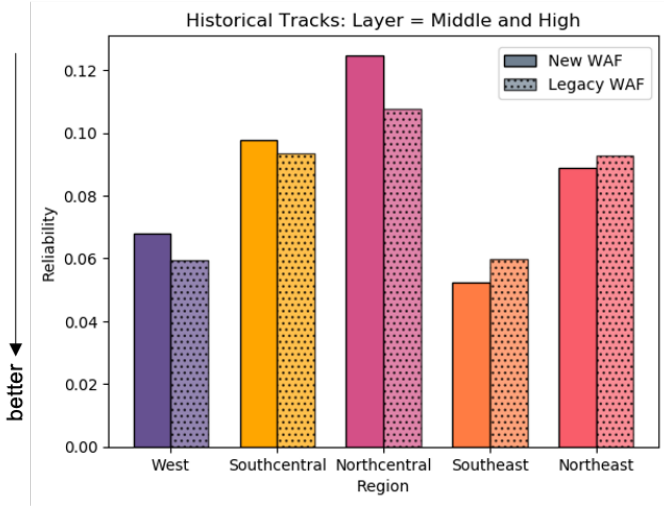

Figure 20. Reliability for each region when the middle and high layer are combined. Each region's color corresponds with the colors used in the geographic map in Section 3.2.4.

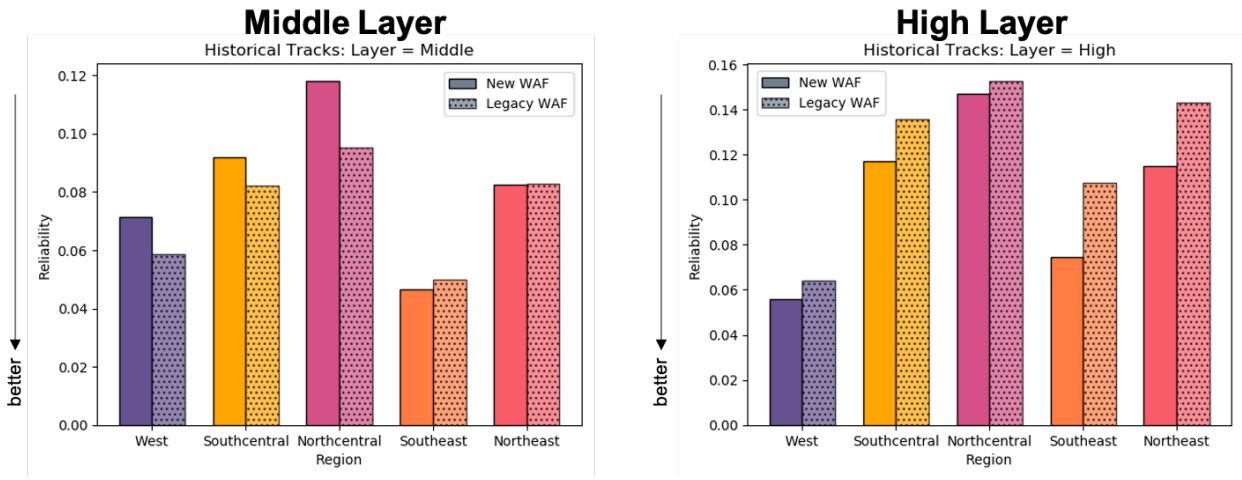

Figure 21. Same as Figure 20 but separated into each layer, middle (left) and high (right).

5.3 Case Studies

Multiple cases were identified to illustrate differences between the two CWAM WAF products. The first case occurred on 29 June 2019 over Pennsylvania where there was a strong line of thunderstorms moving across the state that impacted flight routes in ZOB and ZNY (Figure 22). The east-west jet routes can be seen in the historic traffic and, for comparison, in the initial flight plans, highlighting how frequently these routes are used and preferred (Figure 23; upper row). As the line of thunderstorms passed through these jet routes the air traffic deviated, avoiding the thunderstorms and areas where either version of CWAM WAF was greater than zero (Figure 23; lower row). Moreover, the higher the WAF values, the higher the likelihood of pilot deviation. The horizontal extents and locations of the thunderstorms were similar for both products (Figure 24). Additionally, the New WAF had more low values and fewer high values compared to the Legacy WAF, consistent with Figure 12.

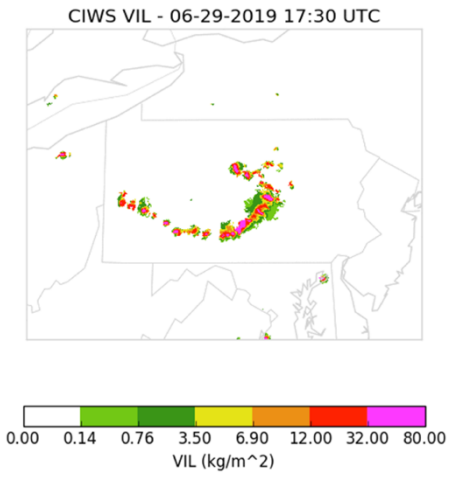

Figure 22. CIWS vertically integrated liquid and echo tops above 32 kft over Pennsylvania on 29 June 2019 1730 UTC.

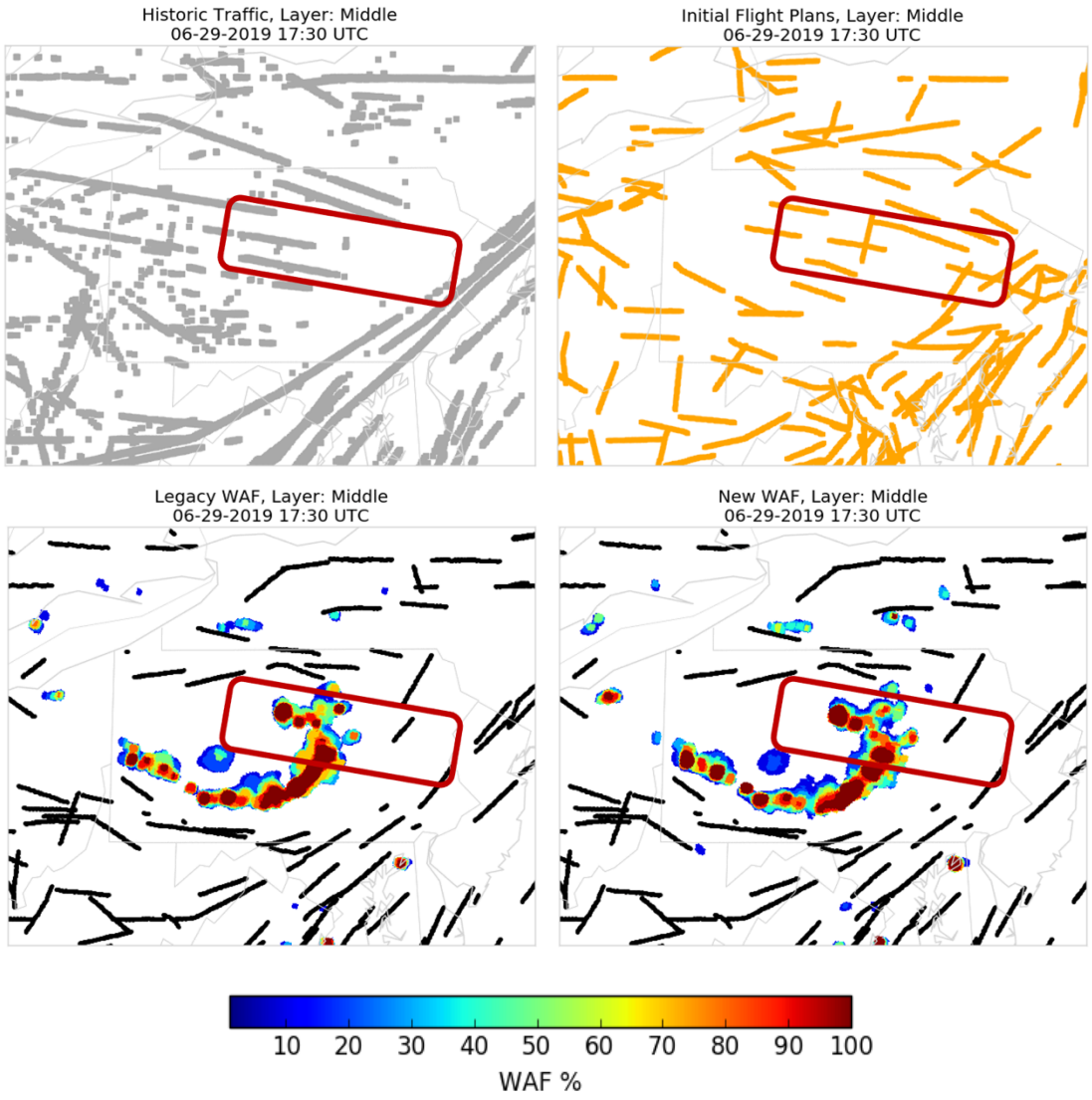

Figure 23. Initial flight plans (upper left), historic traffic (upper right), Legacy WAF (lower left), and New WAF (lower right) fields in the middle layer on 29 June 2019 at 1730 UTC. In the upper row, locations of initial plans are shown in grey and historic traffic in yellow. In the lower row, WAF fields are represented by the filled contours and the actual tracks by the black caterpillar-like lines. The dark red rectangle on all diagrams identifies the region or frequently used jet routes that were avoided during this event.

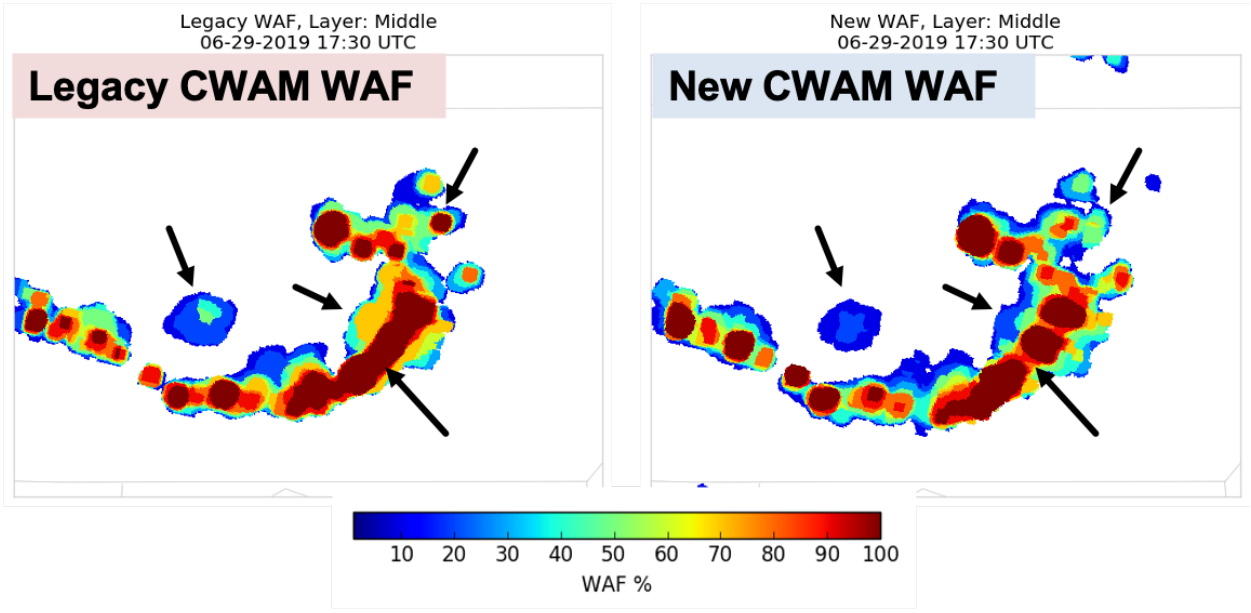

Figure 24. Legacy (left) and New (right) WAF fields over Pennsylvania on 29 June 2019 at 1730 UTC. Arrows indicate areas where the products differ.

The next few case studies highlight differences in the products, likely resulting from the different CIWS products used by the Legacy and New CWAM algorithms. The first example shows the impact of radar on the newer product, resulting in higher WAF values that correspond with the structure of the CIWS VIL fields (Figure 25). These values in the New WAF were higher than what was seen in the Legacy WAF, producing the rings seen in the climatological maps. Additionally, this case study highlights that this was a strong signal seen both in individual cases and in aggregate. The next two are examples of regions where the new product had no WAF value where the Legacy product did. The first shows a radar gap over Montana (Figure 26) and the next where the Legacy WAF continued further off the Virginia and North Carolina coast than the New WAF (Figure 27). These differences in field characteristics could be due to the Legacy WAF using the CIWS forecast and the New WAF using the CIWS mosaic in their algorithms.

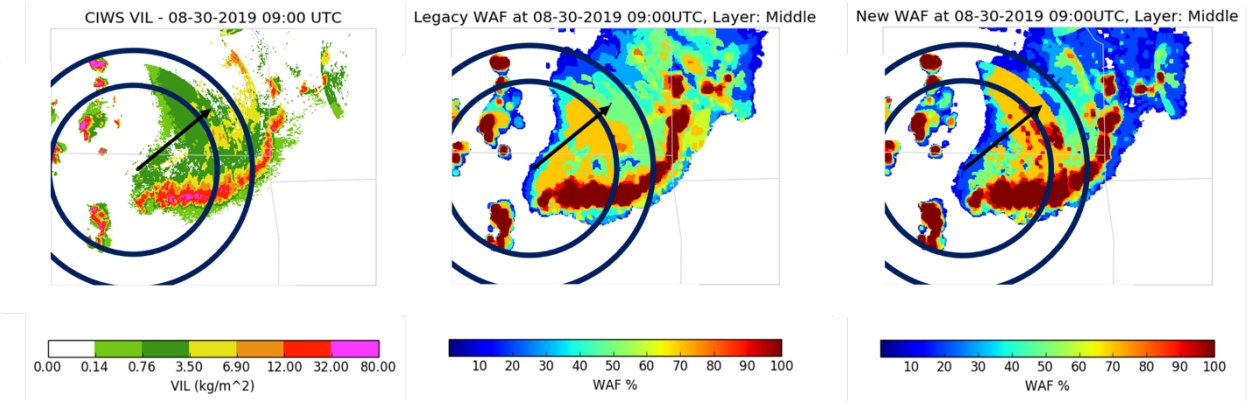

Figure 25. CIWS VIL (left), Legacy WAF (middle) and New WAF (right) fields over Oklahoma and Kansas on 30 August 2019 at 0900 UTC. Arrows indicate areas where the products differ and the rings highlight the circular shape of this difference.

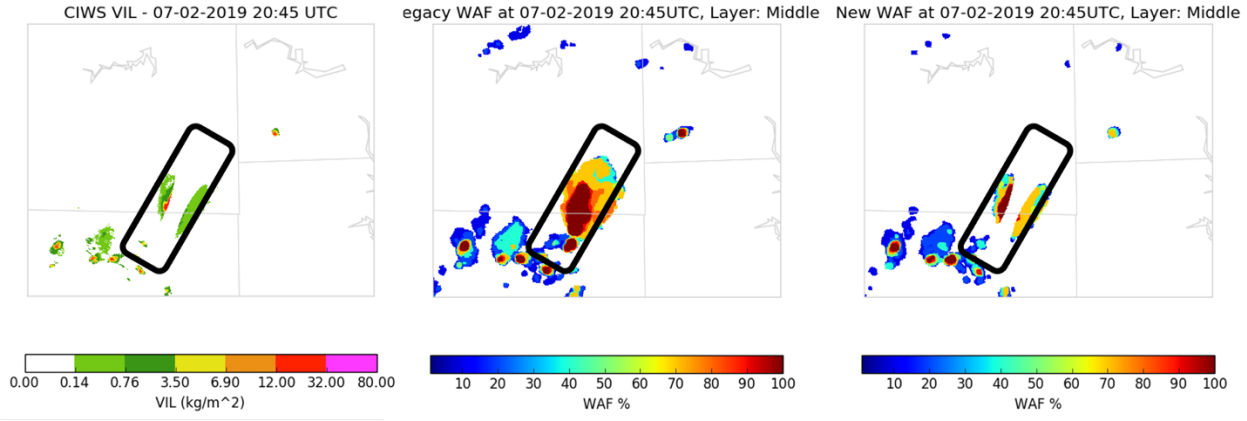

Figure 26. Same as Figure 25, but over Montana on 2 July 2019 at 2045 UTC. The black box highlights the radar gap, where the products differ.

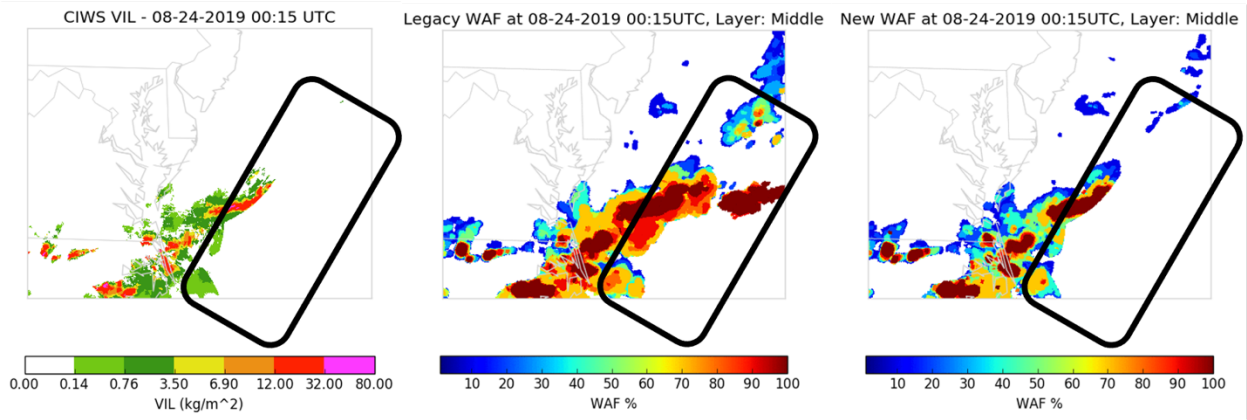

Figure 27. Same as Figure 25, but off the Virginia and North Carolina coast on 2 July 2019 at 2045 UTC. The black box highlights where the Legacy product extends further off the coast than the New product.

6 Conclusions

The purpose of this investigation was to measure the performance of the recent upgrade of the CWAM WAF. Currently, automated routing advisories generated with CIWS are used by traffic managers to make tactical routing decisions for en route air traffic. Quantifying the performance of the CWAM product, and by extension the information feeding route advisories, could lead to greater acceptance by pilots, thus reducing advisory revisions and work load for pilots and controllers.

There were significant differences in the characteristics of each data set that came through in both the case studies and the climatological maps. This includes radar gaps and radar rings over land and different off-shore coverage, resulting in differences in the spatial coverage of the products and the inclusion of a spatial mask to mitigate some of these issues. Additionally, the lowest layer was not evaluated in this assessment since there was no clean way to match the lowest layer in the New product with the available levels in the Legacy product.

For performance, flights generally avoided where either product was non-zero and the higher the WAF value the more likely the avoidance. The New WAF was better calibrated than the Legacy WAF in the higher layer and they were about equivalent in the middle layer. When using ROC curves as a measure of performance, the Legacy WAF had a higher AUC in the high layer when compared to the New WAF and their performance was roughly equivalent in the middle layer. Regionally, the New WAF was better calibrated and the Legacy WAF had a higher AUC in all regions for the high layer. In the middle layer, the New WAF was slightly better calibrated than the Legacy WAF in the two eastern regions and the reverse for the other three regions. For AUC in the middle layer, the Legacy WAF outperformed the New WAF in the Northeast and Southeast regions and slightly in the Southcentral region. The New WAF had higher AUC values in the West and slightly lower values in the Northcentral region. When considering the time of day, the Legacy WAF had a higher PSS during the convective and high traffic hours of the day and the New WAF had a higher PSS during the non-convective hours of the day.

7 References

DeLaura, R.A., B.A. Crowe, R.F. Ferris, J.F. Love and W.N. Chan, 2008. Comparing Convective Weather Avoidance Models and Aircraft-Based Data, AMS Annual Mtg., Conf. Paper P.15, 89th ARAM Symp., 4 August. https://ams.confex.com/ams/89annual/techprogram/paper_145829.htm

Mattioli, C.J., M. Matthews, H. Iskendarian, and M.S. Veillette, 2020: Improvements to Convective Weather Avoidance Modeling Using Supervised Learning, AMS Annual Mtg., 19th Conference on Artificial Intelligence for Environmental Science, 16 January. https://ams.confex.com/ams/2020Annual/meetingapp.cgi/Paper/362623.

8 Appendix

8.1 Flight Plan Characteristics

Characteristics of all three baselines, Initial Flight Plans (12 hours before flight), Predeparture Flight Plans (last filed flight plan), and Historic Traffic (time of day, day of week average), were examined within each vertical layer. An analysis found that the total number of initial flight plans closely matched the total number for historic traffic and actual tracks, but the distribution by flight level did not match (Table 5). Specifically, almost all of the flight plans were concentrated in the middle layer, while only 50% of the historical traffic and historical tracks were in the middle layer (Table 5). This was apparent in both Table 5 and Figure 30. The predepature flight plans were too few in total number and had a similar mismatch in the flight level distribution. These flight plan distributions would then produce more avoidances in the middle layer than was true in reality, resulting from the presence of more flight plans than actual flight tracks. The opposite occurred in the other layers.

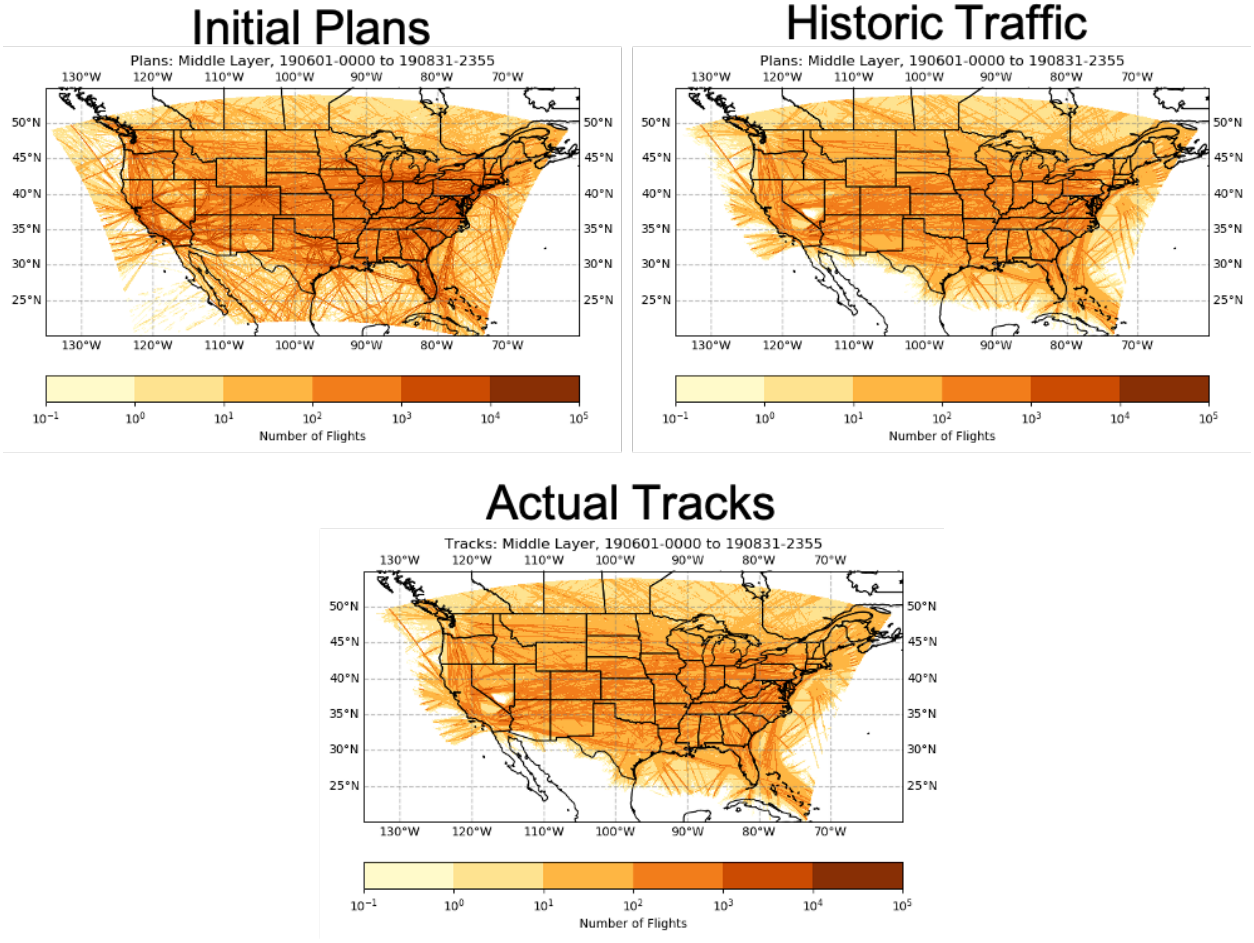

Figure 28. Distribution of flight data on the CIWS grid for the middle layer for initial flight plans (upper left), historic traffic (upper right), and track data (lower middle).

Table 5. Flight data counts for each vertical layer (rows) and all three baselines plus the track data (columns). For each layer, raw counts and the percentages are shown, e.g., for the first row and first column (106,427,654/991,500,672) = 11%.

Layer	Initial	Predeparture	Historic	Tracks
Low	106,427,654 (11%)	143,155,911 (22%)	275,287,328 (27%)	258,375,605 (27%)
Middle	856,075,775 (86%)	481,387,204 (74%)	490,561,469 (47%)	450,296,467 (47%)
High	28,997,243 (3%)	27,223,121 (4%)	268,398,012 (26%)	247,481,438 (26%)
Total	991,500,672	651,766,236	1,034,246,810	956,153,510

8.2 Performance by Time of Day

The components of the PSS score shown in Section 5.2 are presented there. Both thresholds (20 and 60%) are shown and each layer. In the middle layer (Figure 31), the influence of the POD during non-convective hours and POFD during convective hours on the final PSS score was apparent. As a result, the higher PSS for New WAF during non-convective hours diminishes as the threshold increases. Additionally, even though the expected diurnal cycle was not found in the PSS score, it was seen in the components that produced that score. In the high layer (Figure 32), the very low POFD during convective hours was still there and occurred throughout the day. As before, during non-convective hours changes in POD had a greater impact than changes in POFD, producing a higher PSS for New WAF at the higher thresholds.

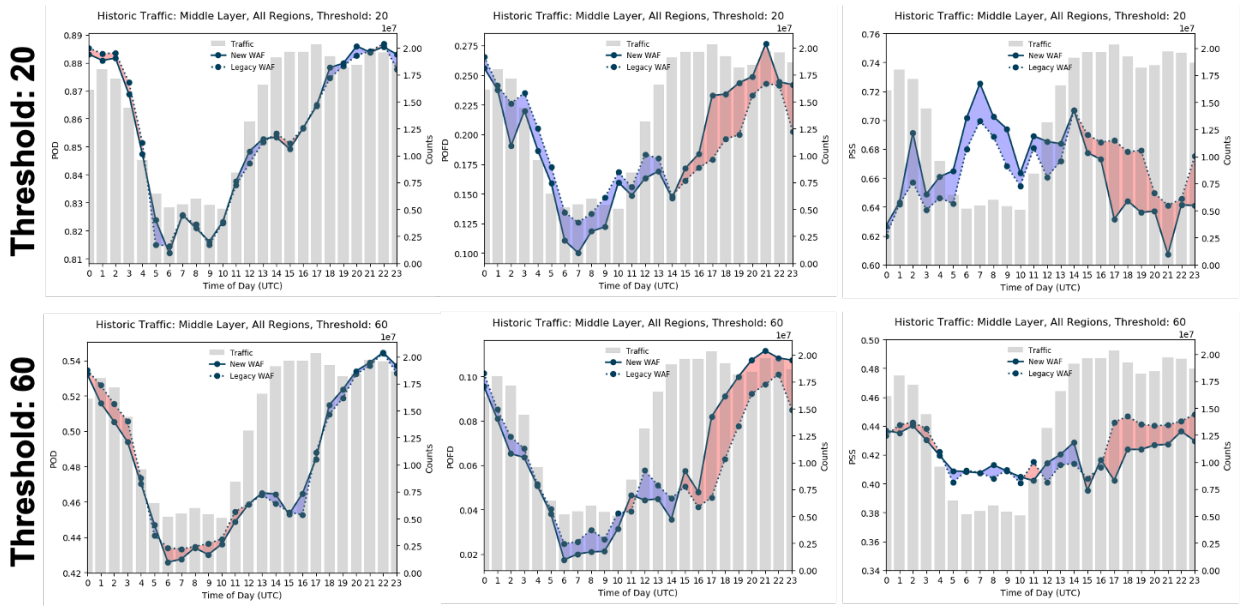

Figure 29. Statistical performance by time of day in UTC for the middle layer for two thresholds, 20% and 60%. Plots include POD (left column), POFD (middle column), and PSS (right column). Blue/red means that the New/Legacy product has a better score. Gray bars show hourly traffic volume.

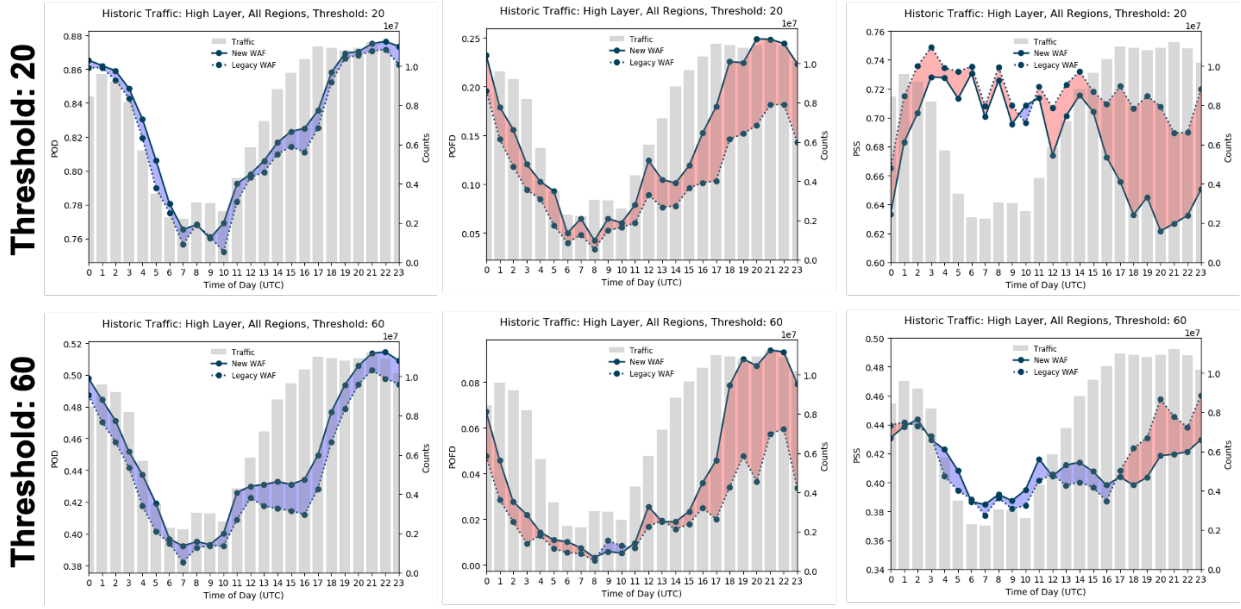

Figure 30. Same as Figure 29 but for the high layer.