KMWN MOS Verification

Joshua Elms and Jacob Garside

August 23, 2024

Abstract

Model Output Statistics (MOS), a statistical post-processing layer on top of dynamical weather models, is integral to mountain weather forecasting. The goal of this study is to analyze the forecast skill of three MOS products (GFS MOS, NAM MOS, and NBM) in regard to five numerical variables (temperature, dewpoint, dewpoint depression, wind speed, and wind direction) at the summit of Mount Washington (station identifier: KMWN). By comparing archived MOS forecasts released at 00Z and 12Z out to 72 hours with official observations collected at the Mount Washington Observatory (MWO) between 11/01/2020 and 05/16/2024, we validate the implicit knowledge of their veteran forecasters at MWO: the GFS MOS, NAM MOS, and NBM are similarly skillful in forecasting temperature, dewpoint, and dewpoint depression, while the NBM lags significantly behind the other two models in forecasting both wind speed and direction. Analysis of the individual and aggregated cases where MOS forecasts had the highest error yields the conclusion that multiple distinct synoptic situations around KMWN are prevalent during the most errant forecasts. These results are intended to serve as guidelines for MWO's new forecasters, replacing some of the years of experience otherwise required to understand when and how to use each MOS variant.

1 Introduction

The Mount Washington Observatory (MWO) produces twice-daily operational 48-hour forecasts for the higher summits of the White Mountains. Model Output Statistics (MOS) plays a key part in informing this forecast. To understand why MOS is vital to mountain weather forecasting, we must first understand how MOS works.

Harry Glahn and Dale Lowry, two National Weather Service (NWS) meteorlogists, introduced the term "Model Output Statistics" in a seminal paper in 1972 [1]. They define it as "determining a statistical relationship between a predictand and variables forecast by a numerical model at some projection time(s)." In the 52 years since their publication, the same fundamental process has been iterated and improved upon to develop MOS products which can accurately forecast variables such as temperature, dewpoint, wind, etc. many days into the future. The gap between the start of the forecast, or "initialization time", and the time of the prediction, or "valid time", is known as the "lead time".

The three MOS products of focus in this study are the Global Forecast System (GFS) MOS, the North American Mesoscale Model (NAM) MOS, and the recently-developed National Blend of Models (NBM). The first two employ traditional MOS techniques: historical output solely from the GFS (or NAM) is related to historical weather observations at a station such as KMWN via linear regression. The process of calculating the optimal weights for the linear model such that its forecasts match the observed weather conditions is the training phase. Once the MOS model is trained (the weights are calculated), new outputs from the dynamical weather model can be fed into the model to produce forecasts for multiple lead times, variables, and stations [1]. The NBM is a more recent innovation by the National Oceanic and Atmospheric Administration's (NOAA) Model Development Laboratory (MDL). As the name implies, the NBM combines output from multiple (more than 30, as of NBM version 4.0) dynamical weather models, both global and regional, to produce forecasts for the continental United States (CONUS). These forecasts exist both for individual weather stations, as with the GFS MOS and NAM MOS, and on a grid covering the CONUS [7].

Regardless of whether the MOS takes one or many models as input, the same fundamental technique holds: learn how historical weather patterns around a station map to the observed weather at that point and apply that mapping to future weather model outputs. This process is especially important for mountain weather forecasting because of a quirk of the dynamical weather models that exist today: the terrain models they employ are grainy, with resolutions measured in dozens of kilometers [6] instead of the decameters that would allow a mountain to be properly captured. Because dynamical model computation time grows as a polynomial function of the model's horizontal resolution (how fine the model grid is), scaling up the model resolution to properly resolve the terrain would be prohibitively expensive and time-consuming for global or continental weather models. In short, our current weather models cannot forecast mountain weather on their own. The beauty of MOS is that it does not require explicit representation of terrain; in fact, it does not even attempt to simulate physical processes. As stated above, it simply relates the synoptic weather conditions to the weather at a point through a statistical relationship. This means that it can implicitly "learn" the effect of mountains on larger air masses by finding patterns in historical data, which can then be leveraged for forecasts in the near-present. As a result, MOS is an important final step in weather forecasting, especially here at Mount Washington Observatory (MWO).

2 Data

2.1 Data Sources

The GFS MOS and NAM MOS datasets are from the Iowa Environmental Mesonet MOS Archive (IEM) [3]. We downloaded the NBM data from the Amazon Web Service archive for the NBM [4]. The period of record for this study is 11/01/2020 through 5/16/2024, which roughly corresponds to the time that the NBM versions 4.0 and then 4.1 have been operational. The NBM wind formula was drastically altered in version 4.2, so that data is excluded from this study.

The observational data is proprietary and comes from the MWO database. The five variables of interest to this study, along with their abbreviations are shown in Table 1. Two of the variables are manually measured (T, T_d), some are measured by instrumentation (V, θ) and the dewpoint

Variable	Short name	Units
Temperature	T	°F
Dewpoint Temperature	T_d	°F
Dewpoint Depression	$T - T_d$	°F
Wind Speed	V	kts
Wind Direction	θ	degrees

Table 1: Variables included for analysis. These are either available in or possible to derive from MWO and MOS records.

depression is derived from a combination of the preceding variables ($T - T_d$). Dewpoint depression is important because it can be used to forecast foggy conditions on the summit. If the temperature and dewpoint are within around 4°F, it is likely that the summit is enshrouded in fog, or 'in the clouds'. This is notoriously hard to forecast on the summit, and therefore a reliance on the MOS and other models is crucial to making an accurate forecast.

2.2 Data Description

In the 1,293 days between 11/01/2020 and 5/16/2024, the three MOS variants each produced at least two daily forecasts (initialized at 00Z and 12Z), both of which include lead times from 6 to 72 hours at 3-hour increments. These forecasts can be directly compared to a subset of MWO's hourly observations valid at the same times. Each MOS variant has different forecast hours that could be used in a future study, but they are compared on the subset of shared initialization and forecast times here for simplicity. Neither the MOS nor MWO data contain any missing values for the period of this study. The NBM was updated during this period, though, operationally switching from version 4.0 to 4.1 on 1/17/2023 [5].

2.3 Preprocessing

The MOS data preprocessing stage comprises two steps: extracting the variables of interested for the study (see Table 1) and determining the forecast initialization and valid times which are shared by the three MOS products. An inner join of the GFS MOS, NAM MOS, and NBM records is performed on a unique ID column which is defined as the product of the hashes of the forecast time and the valid time for each (forecast time, valid time) pair. This ensures that the MOSes are compared at the same lead times and for the same ranges of dates.

The MWO data preprocessing also consists of two simple steps. First, the timestamps are rounded to the nearest hour to account for observations being taken up to 15 minutes before the top of the hour. Next, five hours are added to each timestamp to convert from EST to Zulu (Z) time. At this point, both the MOS and OBS datasets exist in the same time zone and the same forecast hours. This allows us to join them.

3 Methods

This study focuses on determining the skill of the MOS products at the summit of Mount Washington with quantitative and qualitative methods. The former consists of metrics and plots that describe and illustrate the nature of errors in the MOS forecasts relative to observed conditions at the summit. The latter is a set of case studies, each of which focuses on the synoptic conditions occurring during particularly poor forecasts of one or more variables of interest by one or more of the MOS variants. These case studies are chosen by selecting a subset of highest-error events derived from the quantitative analysis.

3.1 Analysis Software

This study exclusively uses Python v3.11 for data preparation and analysis/visualization. The data are represented as Pandas dataframes and the plots are produced using Matplotlib and Seaborn. Pathlib is used for cross-platform path management. Conda is used for package management.

3.2 Metrics

The two primary metrics reported are Mean Absolute Error (MAE) and bias. The formulas for each of these are shown in Figures 1 and 2 below.

$$MAE = \frac{\sum_{i=1}^{N} |x_i - y_i|}{N}$$

Figure 1: Mean Absolute Error (MAE), the average magnitude of error between the MOS forecast (x) and the observed value (y) of a variable at some time i given N observations. Lower values are better.

$$bias = \frac{\sum_{i=1}^{N} (x_i - y_i)}{N}$$

Figure 2: Bias, the difference between the means of the two distributions. The MOS forecast is x, the observed value is y, the time is given by i, and N is the number of observations.

3.3 Case Study Selection

To select what would be used for our case studies, we first selected the top 20 errors for each each variable from each MOS variant, ending with a total of 300 times. To limit the pool of potential cases, only lead times [12, 24, 36, 48, 72]h were examined for high errors. Then, they were ranked by largest error, and the top five errors were pulled into their own respective files, again by model and variable for a total of 75 separate dates. These were again ranked by the top five largest errors for each model and variable, leaving us with 15 errors for each variable due to the use of three models, the GFS MOS, the NAM MOS, and the NBM. After separating all of them out, each time was pulled into separate files, and plots were made for different variables, such as relative humidity at 1000 hPa,

wind speeds and geopotential heights at 500 hPa and 800 hPa, temperatures at 800 hPa, and vorticity and geopotential height at 500 hPa. These parameters were derived from the European Center for Medium Range Weather Forecasting (ECMWF) Renalysis v5 (ERA5).This produced a comprehensive look at the atmosphere for our analysis and allowed us to discern common synoptic conditions and weather patterns that lead to these extreme errors. This resulted in 450 plots. Five representative plots of selected parameters were then chosen to graphically show the causes of these errors.

As another verification of these findings, we used the top 20 errors for each variable and model and made composite images of sea level pressure, 850 hPa and 250 hPa vector winds, 850 hPa air temperatures, and 500 hPa geopotential heights.

4 Results

4.1 Plot Categories

We use a variety of plots to explore the performance and characteristics of the MOS on Mt. Washington. The first category is verification plots. These show Mean Absolute Error (MAE) and bias trends for each model (GFS MOS, NAM MOS, NBM) at 3-hour intervals for lead times from 6-72 hours. For both MAE and bias, values closer to 0 indicate more skillful forecasts. These plots and short discussions of their significance are shown in Figures 3-7. Figure 8 is a visualization of the error distribution of each of the MOS variants for each of the five variables at all lead times.

The secondary category directly compare the observations on the x-axis to MOS forecasts on the y-axis, either as individual points (scatterplots) or distribution-to-distribution (quantile-quantile (Q-Q) plots). The scatter plots are constructed using the "jointplot" function from Seaborn, which produces a central plot of two-dimensional data and two distribution plots, either histograms or kernel density estimates (KDE), on the margins. By explicitly showing the density of points in a region, these marginal plots address one of the primary issues of scatter plots: overlapping points can make it difficult to tell how dense the data is.

The final category of plot is spatial. These plots comprise one or more of the following: contoured fields, filled fields, or wind barbs, which are placed on top of a map of the Continental United States (CONUS) with the location of Mt. Washington indicated by a star. The actual parameters being plotted are individual times extracted from the European Center for Medium Range Weather Forecasting (ECMWF) Reanalysis v5 (ERA5) [2] and are intended to represent the synoptic conditions at that time.

4.2 Verification

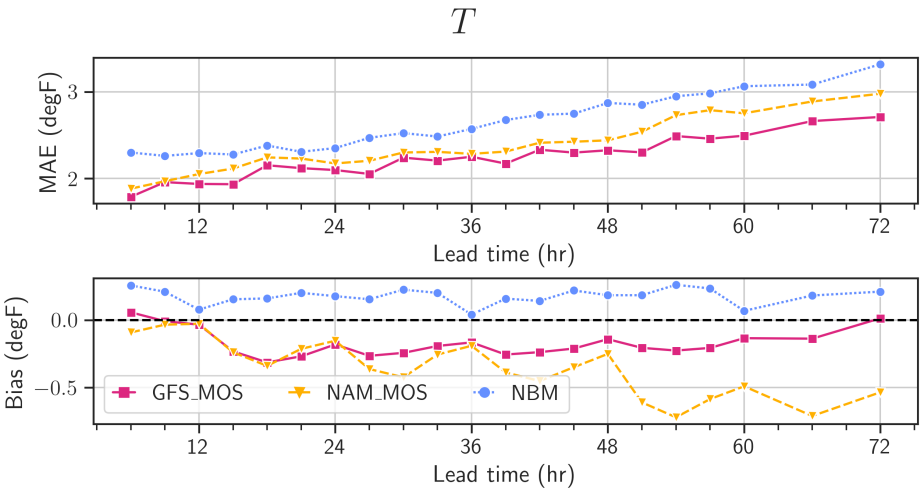

Figure 3: MAE: The GFS MOS and NAM MOS are close, with the NBM trailing by around 0.5 °F in many cases. Spikes and dips are observed at similar lead times (eg. 16 and 54 hours) for all models. Bias: The GFS MOS and NAM MOS tend to perform well until 12 hours, after which they are similar until the NAM MOS becomes much worse after 48 hours. Surprisingly, the GFS MOS appears to become unbiased around hour 72. The NBM overestimates temperatures by around 0.2 °F for the entire 72 hour forecast.

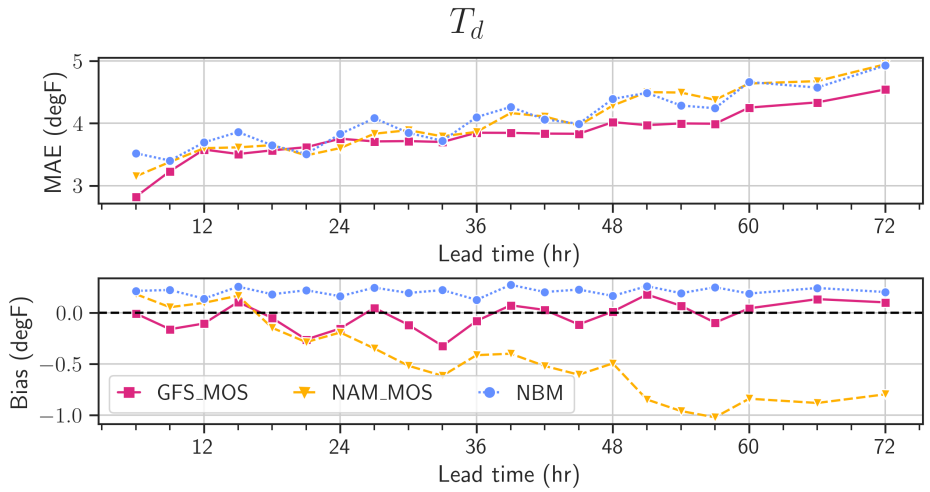

Figure 4: MAE: All 3 models are similar in regards to dewpoint accuracy. The GFS MOS has the lowest error at all lead times outside of 21 and 24 hours, though it only consistently outperforms the other two after 36 hours. The NBM and eventually the NAM MOS exhibit a strong diurnal signal, evidenced by the spikes at 12-hour intervals which arise from the averaged diurnal cycles in the 00z and 12z cycles which compose the dataset. The two offset 24-hour diurnal cycles, when combined, make a visible 12-hour cycle. Bias: The GFS MOS is unbiased at many lead times, with either a low or high bias evident at times over the 12-hour cycle. The NBM consistently overshoots dewpoints by around 0.25 °F . The NAM MOS tracks the GFS MOS until hour 24, at which point it exhibits a low bias that increases until an extrema of -1.0 °F at hour 57.

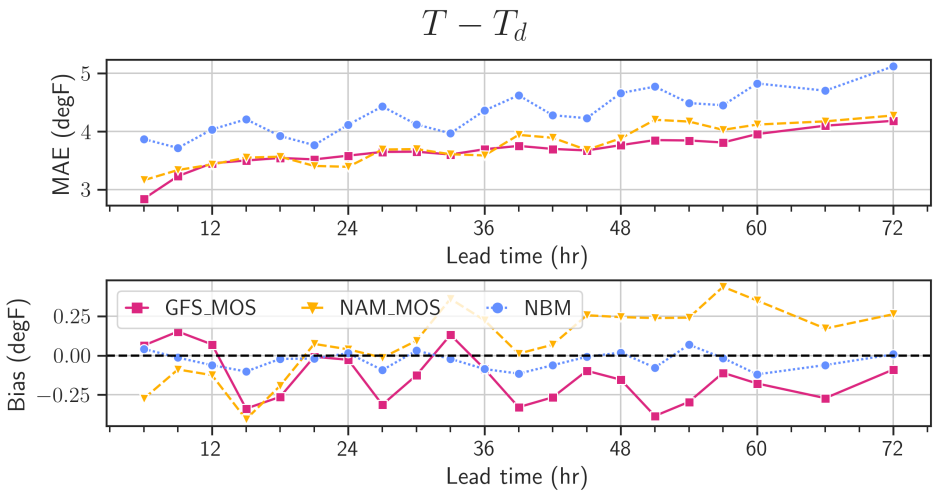

Figure 5: MAE: The GFS MOS and NAM MOS are quite similar, with errors consistently 0.5-1.2 °F lower than the NBM error. The NBM also exhibits a strong diurnal cycle. Bias: The NBM has the lowest bias for the duration of the forecast. The GFS MOS and NAM MOS both fluctuate throughout, with biases ranging from -0.4 to 0.15 °F for the GFS MOS and -0.4 to 0.4 °F for the NAM MOS.

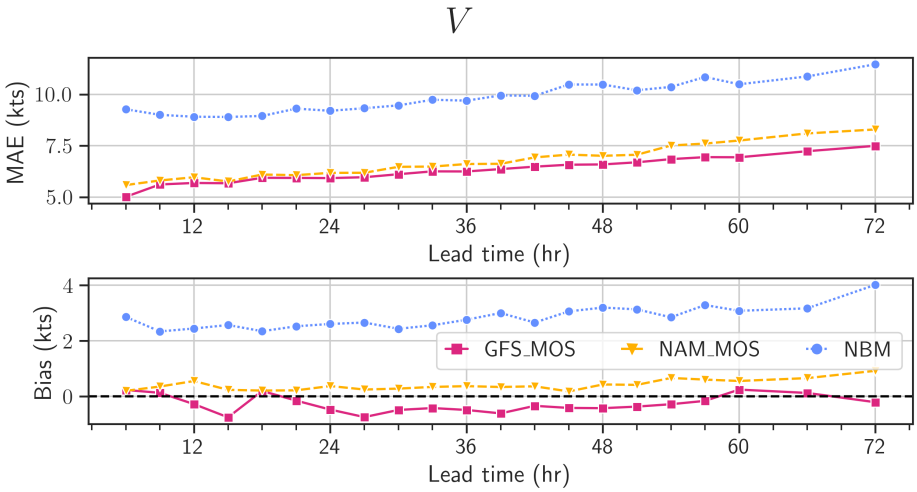

Figure 6: MAE: The GFS MOS errors rise from around 5 to 7 kts, while the NAM MOS starts around the same error but rises slightly more steeply than the GFS MOS, with errors over 8 kts by hour 72. The NBM errors are almost 50% higher across the forecast, starting at 9 kts and rising to more than 11 kts. Bias: Similarly, the GFS MOS and NAM MOS have biases between -1 and 1 kts, while the NBM is hovering around a 3 kt bias.

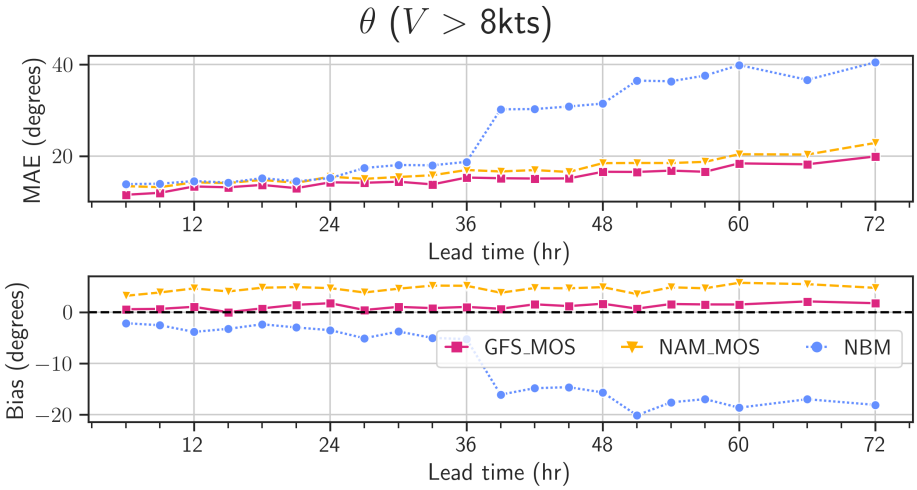

Figure 7: MAE: Until hour 36, all three models are similar, with the GFS MOS only a few degrees below the others in error. At hour 36, the NBM errors rise precipitously, first to more than 30 degrees at hour 39 and eventually more than 40 degrees by hour 72. The GFS and NAM MOS continue along the same trajectory after hour 36, with their errors increasing to a maximum at or just above 20 degrees. Bias: All three models are steady at their initial biases until hour 36, with the NAM MOS strongly positive (around 5 degrees), the GFS MOS mildly positive (around 1 degree) and the NBM strongly negative (3-5 degrees). While the other two continue at those levels, the NBM negative bias drops significantly at hour 36, decreasing to -15 to -20 degrees for the remainder of the forecast period.

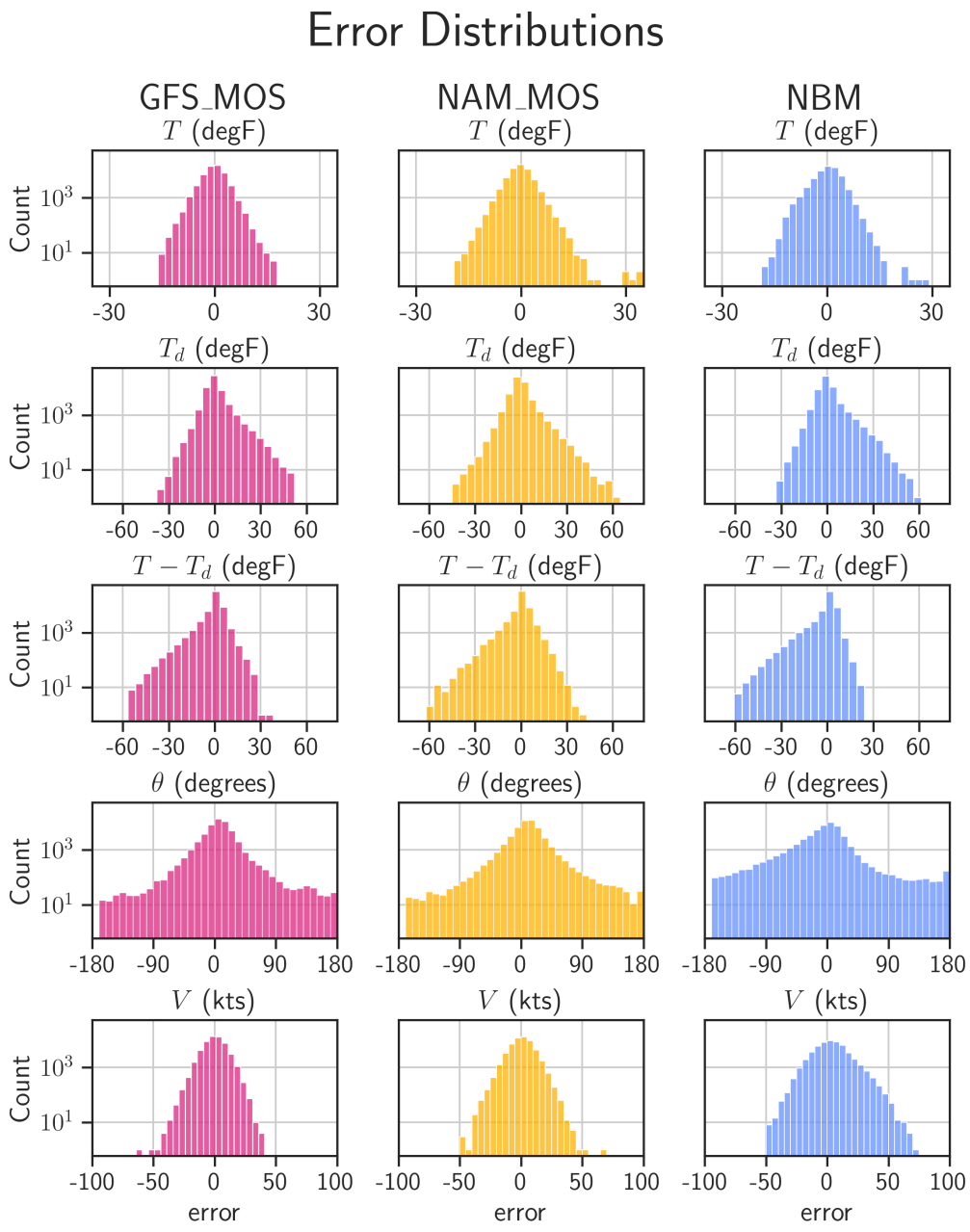

Figure 8: Error is calculated as the MOS minus the observed value for each model and each variable. These subsets of errors are displayed in histograms where plots of errors from the same model share the same column and color, and plots of errors for the same variable share the same row. The height of each bar indicates the log of the observed frequency of error shown on the x axis. For T, θ, and V, errors for the GFS MOS and NAM MOS are noticeably more symmetrical than the NBM. All models appear to exhibit a slight low bias for moisture variables(T_d and $T - T_d$), especially for extreme errors, though the log scale exaggerates this.

4.3 Case Studies

Shown in Table 2 below are the largest errors for each variable analyzed in this study; temperature, dewpoint, dewpoint depression, wind speed and wind direction. In the following figures (9-13), representative plots characterizing the synoptic conditions present at the time similarly large errors are shown.

The synoptic conditions present in Figure 9 commonly occur on days of large temperature errors. This particular figure shows the temperature field at roughly summit-level (800hPa) on February 4th, 2022 at 1200Z, showing a clear warm frontal boundary that could be a cause of the large error. This example shows a common theme between days with these high errors, with some sort of frontal boundary moving over the area when these errors occur. Again, this plot shows clearly what is shown in a plethora of other plots with temperature errors, leading to the hypothesis that these frontal boundaries may be causing the errors that have been seen.

Figure 10 depicts a common synoptic patterns associated with large wind speed errors. This particular figure shows the geopotential heights and winds at summit-level (800hPa) on December 23rd, 2022 at 1200Z, showing a large low pressure system with winds hitting the summit from the southeast which seems to be a key feature on days with the biggest errors in wind speed.

Figure 11 depicts synoptic conditions associated with high errors in wind direction. These cases usually had light synoptic winds in the vicinity of the summit. While the error in wind directions were only calculated for observed speeds above 8 knots, based on using another threshold of 15 knots, most of the largest errors were concentrated toward the set threshold. This figure shows a weak low pressure system moving over the summit with winds from the south, which could account for error.

Figures 12 and 13 depict the synoptic conditions associated with high errors in dewpoint and dewpoint depression, respectively. Since errors in dewpoint tended to be larger according to their MAE than errors in dry bulb temperature, it therefore follows that there was a large degree of overlap in times that had large errors in both parameters. These two figures show moderate synoptic flow at summit-level with a ridge axis near New England. Both of these figures show the base ofthe ridge located over the southeastern US.

Table 2: Maximum errors for each of the five variables analyzed for all forecasted lead times.

Variable (Units)	GFS MOS Abs. Error	NAM MOS Abs. Error	NBM Abs. Error
T (°F)	16	19.1	18.1
T_d (°F)	43.1	43.9	47.1
$T - T_d$ (°F)	49.8	50.4	52.4
V (kts)	64	70	74
θ (degrees)	180	180	180

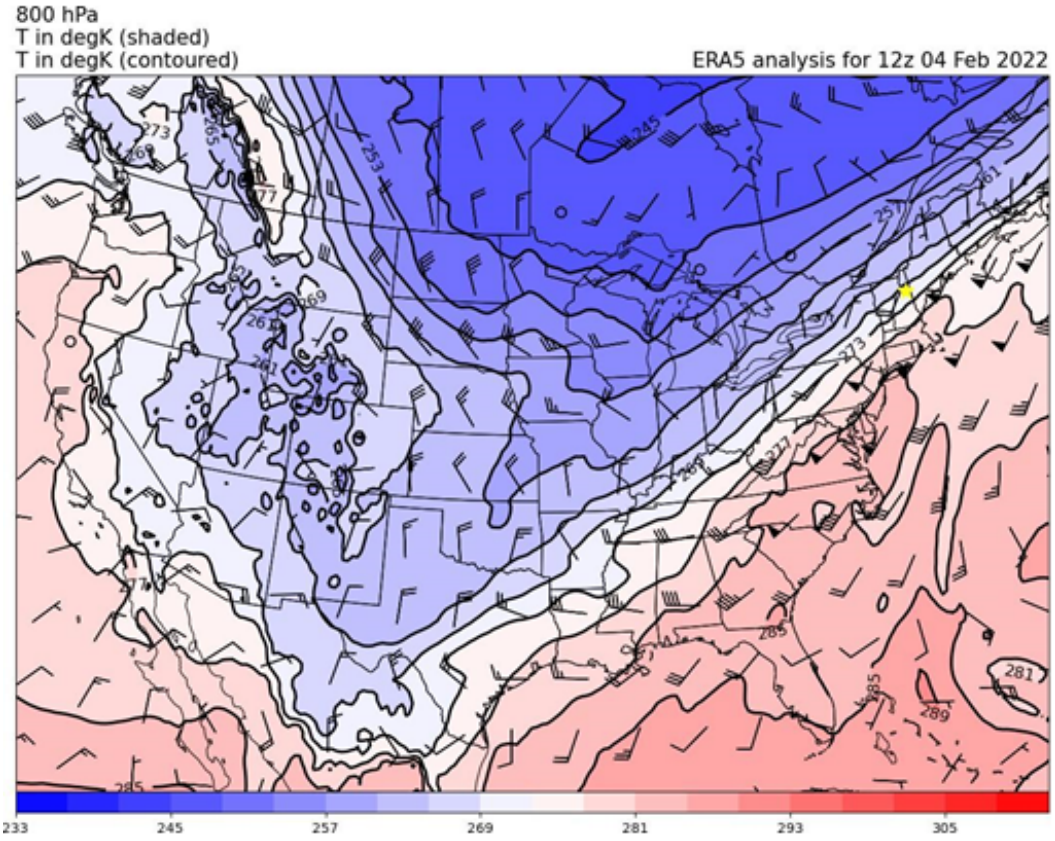

Figure 9: Depicted above is a plot of 800hPa temperatures (filled contours) and wind speeds and directions (wind barbs) on 2/04/2022 at 12Z. At this time, the GFS MOS experienced its 3rd largest error in temperature (15.5°F) and the NAM MOS experienced its 11th largest error (14.5 °F). This shows a common synoptic scenario of a warm frontal boundary over the summit, with a clear distinction between relatively warm and cold air in the vicinity of the summit.

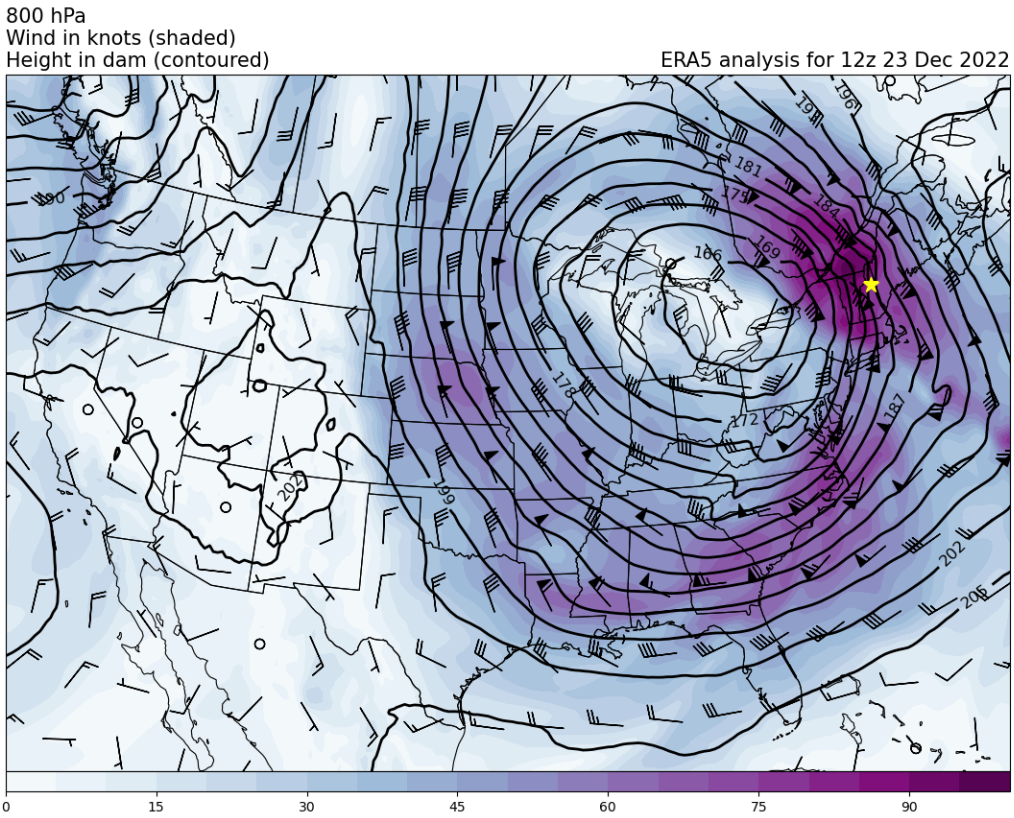

Figure 10: Depicted above is a plot of 800hPa geopotential heights (contours), wind speeds (filled contours and wind barbs), and wind directions (wind barbs) on 12/23/2022 at 12Z. At this time the GFS MOS experienced its 2nd largest wind speed error (44 kts) and hte NAM MOS had its 12th largest error (35 kts). Large wind speed errors were usually associated with a large low-pressure system located to the west causing southerly winds on the summit, accompanied by southeasterly oriented low-level jet over the summit .

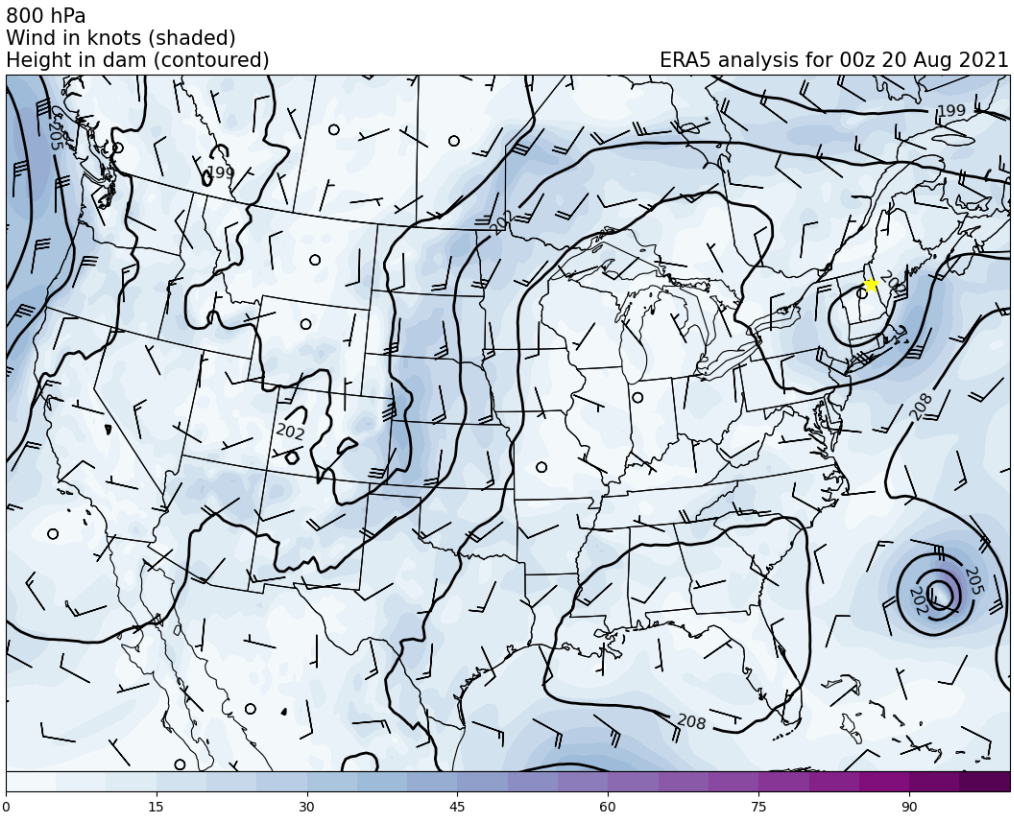

Figure 11: Depicted above is a plot of 800hPa geopotential heights (contours), wind speeds (filled contours and wind barbs) and wind direction (wind barbs) valid at 8/20/2021 at 00Z. This was one of the few times that the GFS MOS and NAM MOS had 180 degree errors. This example shows a small, weak low-pressure system centered nearly over the summit with nearly calm winds (filled circle) at summit-level. It is important to note that at this time the NBM's error was below 140 degrees and did not appear in its top 32 errors across all times.

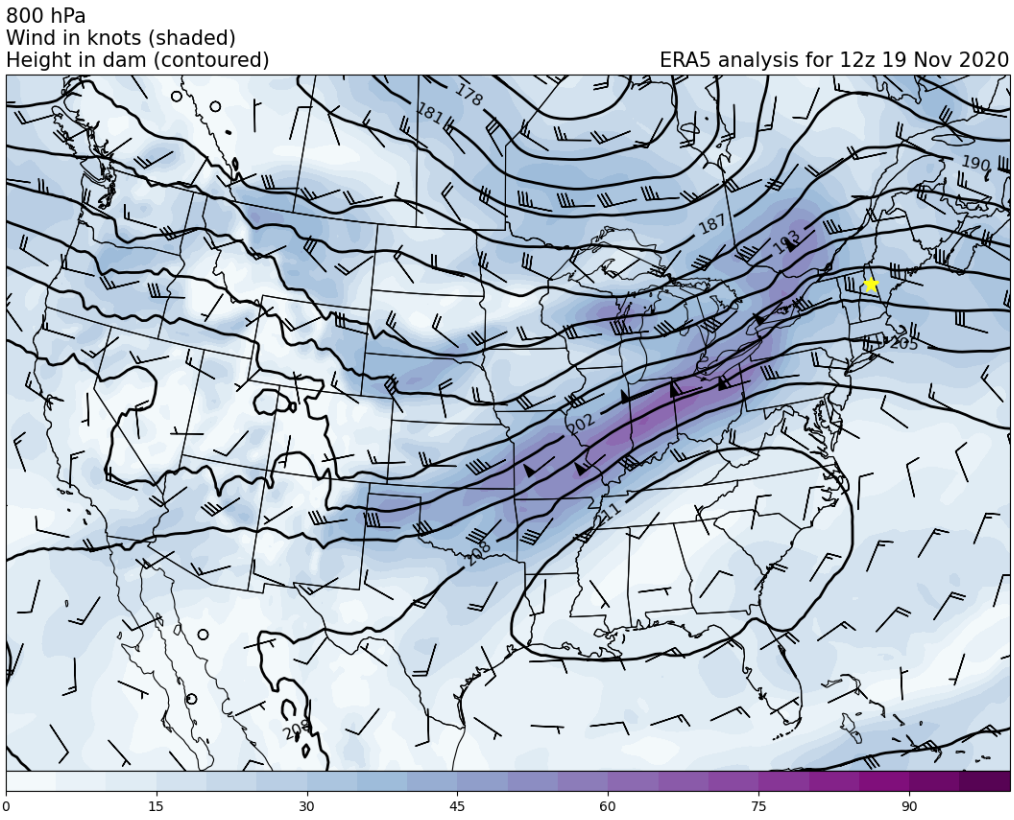

Figure 12: Depicted above is a plot of 800hPa geopotential heights (contours), wind speeds (filled contours and wind barbs) and wind directions (wind barbs) valid at 11/19/2020 at 12Z. At this time, the NAM, GFS, and NBM MOS experienced their 3rd, 10th, and 21st largest errors in dewpoint temperatures. Note the presence of elevated southwest winds somewhat west of New England with a weak ridge axis over the summit extending down to the southeastern US.

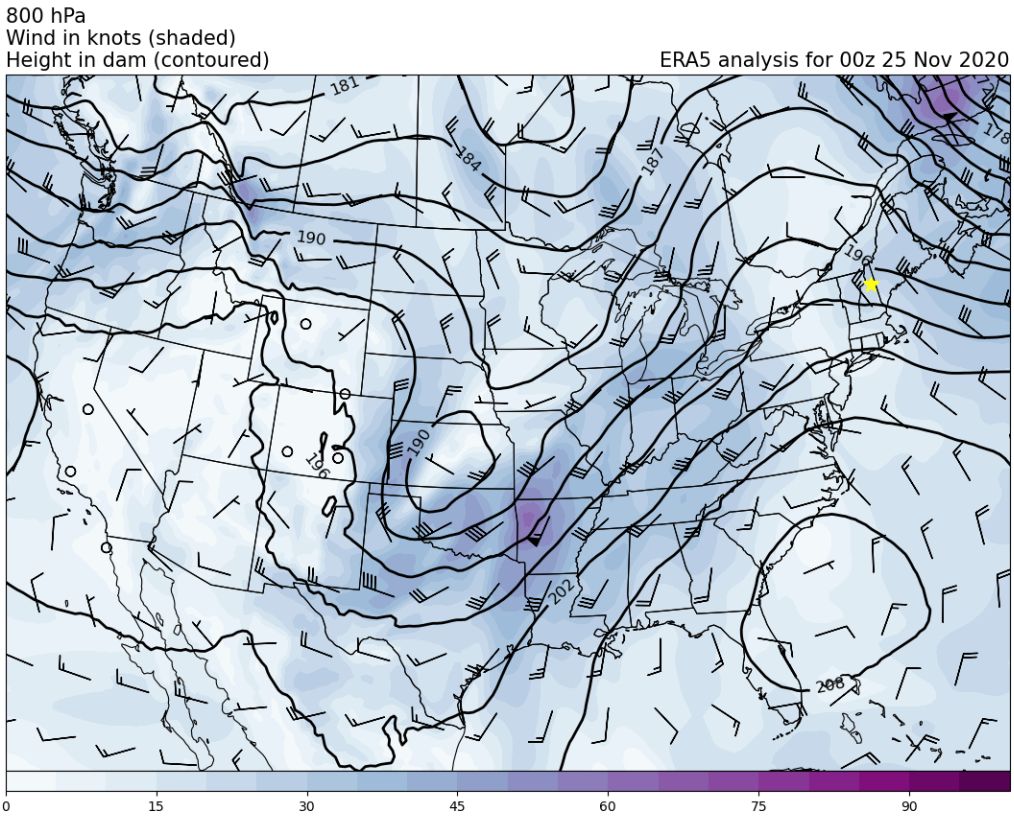

Figure 13: Depicted above is a plot of 800hPa geopotential heights (contours), wind speeds (filled contours and wind barbs) and wind direction (wind barbs) valid at 11/25/2020 at 00Z. At this time, the NBM MOS (52.4 °F) and NAM MOS (50.4 °F) experienced their largest forecast errors in dewpoint depression, with the GFS MOS (47.4 °F) experiencing its 3rd largest error. Note the ridge of high pressure over New York state, with relatively low winds over the summit. These are common synoptic conditions for these large errors at the summit.

5 Discussion

5.1 Ranking the MOS Varieties

The table below summarizes our results in the form of best (#1) to worst (#3) rankings of the three MOS varieties for each variable. "GFS" and "NAM" in the table refer to the GFS MOS and NAM MOS, respectively. Rankings were subjectively determined by weighing the MAE against the bias for each MOS variety. MAE is given a higher relative weight because, unlike systematic bias, it cannot be adjusted for by the forecaster, but a particularly high bias will lower the ranking of that model.

Variable	#1	#2	#3
T	GFS	NAM	NBM
T_d	GFS	NBM	NAM
$T - T_d$	GFS	NAM	NBM
V	GFS	NAM	NBM
θ	GFS	NAM	NBM

Table 3: The three MOS variants ranked in order of performance, with 1 being the best, and 3 being the worst, for each of the five variables of interest.

The table supports one of the primary conclusions of this study: generally, the GFS MOS is the most skilled at forecasting each of the five variables, followed by the NAM MOS and then the NBM. This pattern is clearer for some variables than others. For temperature and moisture variables (T, T_d, $T - T_d$), the differences between the three MOS varieties are small. The largest 6-hour forecast MAE gaps between the GFS MOS and NBM in each of the preceding variables were 0.5-1.0 °F . While a single degree can be important, those differences (2.0 vs. 2.5 °F) were minor when considering the 2-5 °F MAE ranges over the entire forecast period. For the wind variables (V, θ), though, the GFS MOS and NAM MOS outperform the NBM significantly. In V, the NBM MAE is approximately double that of the others, while in θ it is comparable until hour 36, at which point its error increases dramatically. The NBM's high errors in forecasting wind are well-known by its developers at NOAA's Modeling Development Laboratory (MDL). The period of study for this paper (01/11/2020 - 16/05/2024) covers versions 4.0 and 4.1 of the NBM, but the most recent release as of writing (v4.2) employs a method called Quantile Mapping to improve wind forecasting. It remains to be seen whether updates to the NBM have bridged the performance gap between it and the NAM and GFS for winds.

5.2 Case Studies

As mentioned previously, each depicted case is representative of many other extreme errors, enabling us to show just one plot instead of hundreds.

5.2.1 Justification

As explained in the above rankings, all three MOS products are skilled at forecasting the study variables. The vast majority of their forecasts showcase their skill well, with them predicting the temperature within 3 °F or the wind speed within 7 kts even 72 hours before the event. Figure 8 confirms this; each error histogram shows a peak around errors of 0 and rapidly falling counts of forecasts with higher errors. MOS has a markedly positive impact on MWO's Higher Summits Forecast. Despite this, there are also extreme cases in which one or more of the MOSes make predictions with high errors; these are the tails of the distributions in the aforementioned plot. The goal of performing these case studies is to understand what causes each of the MOSes to have high forecast errors. In the short-term, this will help forecasters make more accurate predictions. In the long-term, this will enable the MOS' developers to improve their algorithms, thereby strengthening statistically post-processed weather modeling as a field.

5.2.2 The Cases

Temperature, as seen in Figure 9, had a distinct reason behind a lot of the largest errors. The synoptic condition that seemed to be causing the largest errors was the passage of frontal boundaries, usually with high pressure on one side of the summit and low pressure on the other side of the summit. There is also usually a sharp contrast or steep gradient between high and low pressure. The synoptic conditions hint at cold and warm fronts that can these cause large errors. If a storm is barreling towards Mt. Washington, the forecasted temperature can often be off by a large margin due to the timing of the passage and the extreme contrast between cold and warm temperatures. This can be

due to the MOS forecasting in three hour segments. Temperatures can change drastically on Mount Washington in three hours, and if the MOSes forecasted the front to arrive in three hours and it arrived in two hours, these errors can be exaggerated on paper but can "verify" later in the forecast period. These errors are real errors, but only because the passage of fronts was off by a few hours. Upon manual inspection of our Meteorological Aerodrome Report (METAR) records, we saw that this was true in some cases; sometimes the MOSes were just off by a few hours and the temperature they forecasted would pan out later or earlier than expected. Synoptic conditions like large fronts moving over the summit on the shoulder of a forecast period where it could be teetering on which three hour chunk it would fall under should be a red flag for observers at the mountain when using the MOSes to forecast for the summit. Another red flag would be a strong temperature gradient, where the summit may be changing by tens of degrees in a matter of hours, which can also throw the MOSes off in their forecasts.

Dewpoint, as seen in Figure 12 was harder to find a reason behind its errors for our observers to look out for. There could be a slew of reasons, one of them being the models just have a hard time with dewpoint on the summit which could be true. With our unique weather on the summit, specifically the statistical average of being in the fog for 60% of the year. This sets the trend of our dewpoint being close to our temperatures a majority of the year, but under certain circumstances the fog can drop below the summit or adiabatically mix out briefly during an observation as the air dries out, causing large errors. within the top 20 errors per model for dewpoint, totaling 60 errors, not one dewpoint was undershot; all 60 errors were times when the MOSes overshot the dewpoint by a large margin. In the case that we decided was representative of the majority of the other errors you can see somewhat below normal winds impacting the summit, with high pressure building over the northeast. This was what most of the other plots looked like, leading to to the hypothesis that high pressure may have had something to do with the anomalies. Using the archived data from the summit, specifically METARs, we were able to decipher exactly why this may be happening. When high pressure moves into our area, it can sometimes push the fog that we are usually shrouded with down below the summit, bringing drier air and making the dewpoint drop suddenly. This, accompanied with actual METAR data showing a "V" shape in the rapidly falling and rising dewpoints led us to believe this may be the case for a lot of these large errors found in the dewpoint data. This leads to a need for caution when looking at forecasted dewpoints to forecast cloud cover on the summit. Using a model with a higher time resolution to determine cloud cover on the summit is more useful in some cases than any of the MOSes. It can also be important to forecast for an event, such as frontal passage or a clearing period instead of an exact time.

Dewpoint depression is related to dewpoint, as seen in 13. Dewpoint depression commonly experienced large forecast errors under the same conditions as dewpoint which helped cause some of the errors. Dewpoint depression is important on the summit because it can be used to forecast foggy conditions on the summit. If the temperature and dewpoint on the summit are within around 4°F it is likely that the summit is encased in fog, or "in the clouds". This is notoriously hard to forecast on the summit, and therefore reliance on MOSes and other models is crucial to making an accurate forecast. Because dewpoint depression is so closely linked to dewpoint, most of the errors were ranked in the same place or close to it, with the top five errors for dewpoint and dewpoint depression being

populated with almost all of the same dates. This leads to the conclusion that when forecasting dewpoint depression it is important to follow the same cautions you would use with dewpoint; most of the synoptic conditions are the same. If it looks like there might be a large error in the dewpoint, there will likely be a large error in the dewpoint depression and vice versa. Even looking at temperatures and where they can go wrong would help make cloud cover forecasts as temperature plays a role in dewpoint depression.

Wind speed was interesting to look at because each observer had already determined their own bias correction factor for the MOSes based on how each had performed in the past and in their experience. In 10 there is a large low pressure system spinning cyclonically with a low over the Great Lakes and a jet streak over the summit. The jet streak, consisting of winds around 90 kts from the southeast, is common for this type of error, which is why this day was chosen to represent the largest wind speed errors. In almost every case looked at, there was a low pressure system or trough centered just west of the summit. This usually leads to a jet streak or strong winds, with a strong contrast in wind speeds close by. This strong contrast of wind speeds can manifest in the form of 100 kts winds over the summit, and 20 kts 50 miles to the west of the summit. This can cause the timing of the system moving through to play a large factor when forecasting on the summit. In Figure 10, you can see just to our east the winds are around half of what they are on the summit, meaning that if the forecast is off by an hour or two it could impact the scale of the error dramatically. There are a few reasons why these synoptic conditions cause the largest errors, and the first is the wind direction. Mount Washinton's prevailing wind direction is from the northwest, with southwest, south, and southeast winds being much more uncommon. Our observational record is fed into the MOSes for KMWN meaning if we don't have a extensive record for these wind directions, each scenario when we get winds from those directions is less flushed out than if it was a northwest wind. This can lead to errors, which observers have known for years, but this study can finally prove and quantify that. The plot associated with wind speed errors was one of many errors like this; this example showed a 44 kts error made by the GFS MOS, which is tame compared to the 74 kts error forecasted by the NBM.

Forecasting wind direction had some of the same errors as forecasting wind speed, in that there were often southwest, south or southeast winds, but with the difference being low winds. These large forecast errors We first tried collecting the worst errors for each model for wind direction; this yielded almost entirely 180-degree errors, which was cause for concern. After deliberating we raised the wind speed threshold to 15 kts before it would register as an error, as very low wind could cause the instruments to swing in a direction, and then stay there if there wasn't any subsequent wind. This helped our data drastically, with only a few 180-degree errors to be found. We then decided to bring that wind speed threshold down to 8 kts, as that is what the NBM developers used when doing verifications on their product. This yielded more 180-degree errors than the 8 kts, but was still more diverse than no threshold. A point could be made that a 15 kts threshold is better for Mount Washington; as our winds are statistically higher than most other places in the US, but for this study we used an 8 kts threshold. Depicted in Figure11 is what was seen in most wind errors, with low winds from the southwest, south, or southeast. This combination of lower than average winds that the MOSes have an inability to forecast (see fig ???), and southerly winds that the MOSes have a

hard time forecasting causes wind directions to be off by their maximum value of 180 degrees over and over again. Being less critical, these MOSes can forecast correct wind directions most of the time which is a feat in itself, but these maximum errors are more prevalent and more populated than any other variable in this study. The GFS MOS had three of these errors, the NAM MOS had one of these errors, and the NBM had 15 180-degree errors out of its 32 worst error days data set, making up the large majority of these errors. It is key to look out for low and/or southerly winds when making a forecast for Mount Washington, as these seem to bring about the largest forecasted errors.

6 Conclusion

6.1 Operational Impacts

All three MOS products are nearly equally skillful in forecasting temperature, dewpoint, and dewpoint depression, though the GFS MOS is slightly better than the NAM MOS, which is slightly better than the NBM. When it comes to the wind variables, the GFS MOS has only marginally lower errors than the NAM MOS, but both are significantly more skilled than previous versions of the NBM. When forecasting for temperature, be cautious about using the MOS products when a warm or cold front is moving through as this is when the largest errors tended to occur. When forecasting dewpoint and dewpoint depression, it is important to watch out for a brief clearing of skies with high pressure as this was where the largest errors occurred, and it may be beneficial to look at map representations of model output rather than only MOS products. When forecasting wind speed and wind direction it is important to be wary of southerly winds, and a high wind speed gradient near the summit, as most large errors happen under these conditions.

6.2 Future Work

Even with the broad ambit of this verification study, much work remains. Most notable is an extension of this study to include data derived from the NBM v4.2, which released in May 2024, the end of this study's period of interest. This update was intended to address the high wind errors seen in previous versions of the NBM. More developments in the statistical methods used by the NBM could merit another verification, especially as the GFS MOS and NAM MOS are slated for decommissioning in the next few years, to be replaced by the NBM. Even within our focus period, more information remains to be analyzed: the MOS products produce probabilistic precipitation products, visibility, and temperature extrema, all of which could be analyzed to better understand the role MOS could play in developing MWO forecasts.

The case studies presented here are selected from among only [12, 24, 36, 48, 72]h lead times, which could be expanded to all of the 3-hour increments between 6 and 72 hours of lead time in order to include all of the most extreme forecast errors. Along the same lines, more rigorous statistical analysis of the synoptic conditions that produce high forecast errors is certainly possible. This could be performed by manually classifying the synoptic conditions of each high-error case and then producing summary statistics of the synoptic categories that were most likely to yield high errors.

References

[1] Harry R. Glahn and Dale A. Lowry. "The Use of Model Output Statistics (MOS) in Objective Weather Forecasting". In: *Journal of Applied Meteorology* 11.8 (Dec. 1972), pp. 1203–1211. DOI: 10.1175/1520-0450(1972)011<1203:tuomos>2.0.co;2. URL: https://journals.ametsoc.org/view/journals/apme/11/8/1520-0450_1972_011_1203_tuomos_2_0_co_2.xml.

[2] H. Hersbach et al. *(2023): ERA5 hourly data on single levels from 1940 to present.* DOI: 10.24381/cds.adbb2d47.

[3] URL: https:%20//mesonet.agron.iastate.edu/mos/.

[4] URL: https://noaa-nbm-grib2-pds.s3.amazonaws.com/index.html.

[5] URL: https://vlab.noaa.gov/web/mdl/nbm-versions.

[6] URL: https://www.ncei.noaa.gov/products/weather-climate-models/global-forecast#:~:text=GFS%20is%20a%20global%20model,one%20week%20and%20two%20weeks..

[7] David E. Rudack. "An Historical Overview of NOAA's National Blend of Models (NBM)". In: *An Historical Overview of NOAA's National Blend of Models (NBM) (Invited Presentation)* (2020). URL: https://ams.confex.com/ams/2020Annual/webprogram/Manuscript/Paper364390/AMS%20Extended%20Abstract%20_2020_NBM-History-11-25-19._Figures-atEnd_DB.pdf.