Configuration of Statistical Postprocessing Techniques for Improved Low-Level Wind Speed Forecasts in West Texas

MEGHAN J. MITCHELL AND BRIAN ANCELL
Texas Tech University, Lubbock, Texas

JARED A. LEE
Research Applications Laboratory, National Center for Atmospheric Research, Boulder, Colorado

NICHOLAS H. SMITH
Texas Tech University, Lubbock, Texas

(Manuscript received 6 November 2018, in final form 31 August 2019)

ABSTRACT

The wind energy industry needs accurate forecasts of wind speeds at turbine hub height and in the rotor layer to accurately predict power output from a wind farm. Current numerical weather prediction (NWP) models struggle to accurately predict low-level winds, partially due to systematic errors within the models due to deficiencies in physics parameterization schemes. These types of errors are addressed in this study with two statistical postprocessing techniques—model output statistics (MOS) and the analog ensemble (AnEn)—to understand the value of each technique in improving rotor-layer wind forecasts. This study is unique in that it compares the techniques using a sonic detection and ranging (SODAR) wind speed dataset that spans the entire turbine rotor layer. This study uses reforecasts from the Weather Research and Forecasting (WRF) Model and observations in west Texas over periods of up to two years to examine the skill added to forecasts when applying both MOS and the AnEn. Different aspects of the techniques are tested, including model horizontal and vertical resolution, number of predictors, and training set length. Both MOS and the AnEn are applied to several levels representing heights in the turbine rotor layer (40, 60, 80, 100, and 120 m). This study demonstrates the degree of improvement that different configurations of each technique provides to raw WRF forecasts, to help guide their use for low-level wind speed forecasts. It was found that both AnEn and MOS show significant improvement over the raw WRF forecasts, but the two methods do not differ significantly from each other.

1. Introduction

As the installed capacity of wind energy increases, it is becoming more important to make accurate wind forecasts. Wind power is a variable, weather-driven source of energy and therefore requires backup base-load and spinning reserve units during periods of low wind. These base-load and spinning reserve units are usually powered by other sources such as coal, nuclear, or natural gas (DOE 2008). There are high costs associated with starting up these units, and accurate wind power forecasts can provide important guidance for their more efficient operation. In turn, accurate turbine-layer wind forecasts are vital to the efficient operation and development of wind power resources desired to support a clean and reliable global power supply. In fact, the quality of numerical weather prediction (NWP) forecasts has recently been identified as a crucial science priority if meteorological support of the wind energy industry is to improve (Schreck et al. 2008; Shaw et al. 2009; Banta et al. 2013).

To accurately forecast wind power, it is critical to predict wind speeds well at turbine hub heights and throughout the rotor layer, where rotating blades convert kinetic wind energy into electricity. The typical hub height for a utility-sized modern wind turbine is between 80 and 120 m, with rotor layers generally extending from about 40 to 180 m. Underscoring the need for high-quality wind forecasts in this layer is the fact
that the power output from a wind turbine is proportional to the wind speed cubed (DOE 2008), and therefore a small error in the wind speed forecast can result in much larger errors in projected power output. Forecast errors in wind speed are also particularly important near turbine cut-out wind speed values, since the difference between a power prediction and the actual generated power can be very large at these thresholds.

Current NWP models struggle with forecasting low-level winds, such as those at turbine hub height and within the rotor swept area, partially due to systematic errors related to deficiencies in model physics parameterizations. These errors can be partially addressed with statistical postprocessing techniques that use statistical models over training data periods to relate model forecasts to observations. One common and established technique is model output statistics (MOS). MOS uses a multiple linear regression to correct systematic errors in a forecast model by using deterministic NWP forecasts of certain weather variables as predictors and fitting these predictors to observations (Glahn and Lowry 1972). The process of selecting the predictors from a pool of potential predictors to use in the MOS regression equation is called a screening regression, which is explained in Wilks (2011). One of the most common methods of screening regression is called forward selection. Forward selection is a method that aids in screening the potential predictors and selecting variables that will add the most value to the regression equation, avoiding redundant or mutually correlated predictors (Wilks 2011). There are also postprocessing techniques that use an ensemble framework to take advantage of the spread of solutions produced by the ensemble members. One of those techniques is the analog ensemble (AnEn). The AnEn method uses forecasts from a high-resolution deterministic NWP model run with archived observations over the same period to select the best “analogs” to serve as ensemble members to produce a probabilistic ensemble forecast (Delle Monache et al. 2013) with an improved deterministic prediction (the ensemble mean).

This study applies and compares two statistical postprocessing techniques, MOS and AnEn, to raw deterministic NWP forecasts to understand the degree of improvement the techniques may provide to low-level wind speed forecasts in the Texas panhandle. It also tests different configurations of the techniques (e.g., number of predictors) to understand how to optimize the techniques. Model configurations are chosen to mimic the real-time Texas Tech University (TTU) weather prediction system to foster a seamless transition of research results into improved operational wind forecasts.

This study is unique in that it applies the statistical techniques to sonic detection and ranging (SODAR) instrument observations existing throughout typical turbine rotor layers, instead of meteorological tower measurements and surface observations, providing an understanding of the potential forecast improvements with regard to the wind speeds that directly and more completely affect wind power prediction. This work also directly compares the deterministic predictions from two postprocessing techniques (MOS and AnEn), something that is rare in prior research on statistical postprocessing of NWP forecasts. In the literature, AnEn was mostly evaluated against other ensemble statistical postprocessing methods (Delle Monache et al. 2013; Alessandrinii et al. 2015; Junk et al. 2015a).

Specifically, this study addresses the following research questions:

1) How do the different aspects of MOS (model horizontal and vertical resolution, size of area from which predictors are selected, number of predictors, and training dataset length) affect the skill added to raw deterministic forecasts?

2) How do different aspects of AnEn (model horizontal and vertical resolution, predictor weighting, and training dataset length) affect the skill added to the raw deterministic forecasts?

3) How do the deterministic forecast improvements compare between MOS and the AnEn?

The outline of this paper is as follows: section 2 presents the methodology used in this work along with some background on the MOS and AnEn techniques, section 3 includes results and discussion pertaining to the experiments performed, and section 4 summarizes the research and includes the conclusions of this work.

2. Methodology

a. Study location and observations

Wind turbine rotor height (~40–180 m) wind speed observations are needed in order to develop the statistical models used in this study for the postprocessing of low-level winds. These observations are obtained from the west Texas Mesonet’s SODAR instruments (National Wind Institute 2019). Two years of observations from these instruments are used in this study. Figure 1 displays the different SODAR locations and the local topography of west Texas. The locations marked with an “HERE” and “SANG” in Fig. 1 indicate the Hereford and San Angelo SODARs, respectively, that are used in this study. While there are seven
SODARs in the network, the two that are selected for use in this study offer results based on different local areas with different topographic characteristics, something likely to play a role in systematic low-level wind speed errors. The two selected SODAR datasets also had similar training data available over the majority of the study period, unlike the other stations that had a significant amount of missing observations. Wind speed observations from these instruments are recorded in 10-min averages and sampled at heights of 30–320 m.

b. Model setup

The NWP reforecasts that are used in this study are made with version 3.5.1 of the Advanced Research version of WRF (WRF-ARW) Model (Skamarock et al. 2008; Powers et al. 2017). A 12-km domain is used to develop a 2-yr training dataset (2014–15) with a 4-km...
nest (both with 38 vertical levels) during all of 2015 to test
different horizontal model resolutions with the MOS and
AnEn techniques. Figure 2 depicts the 12-km domain
area and 4-km nest area (denoted with a box labeled d02)
used in the study. One-way nesting was used to provide
the nested domain with boundary conditions. The 2015
year is also rerun with 12-km grid spacing only, but
at higher vertical resolution (51 levels) to test different
vertical model resolutions. The 38 vertical levels are the
TTU real-time forecast system levels and the 51 vertical
levels are the operational levels used by NOAA’s High-
Resolution Rapid Refresh (HRRR) model (Benjamin
et al. 2016). Figure 3 shows the comparison between the
two models’ (TTU versus HRRR) vertical levels setup
below 1 km at the Hereford location. The verification for
all experiments was done using forecasts from 2016.

Each will be a deterministic WRF Model forecast
initialized from GFS model forecasts and run four
times a day to 24-h forecast time. GFS lateral boundary
conditions are provided to the 12-km domain, and
model output is saved hourly. The model physics con-
figuration was chosen to match the real-time TTU pre-
diction system, with the eventual goal of implementing
beneficial postprocessing corrections in that system.
These physics schemes include the Thompson micro-
physics (Thompson et al. 2008), Rapid Radiative Transfer
Model (RRTM) longwave radiation (Mlawer et al.
1997), Dudhia shortwave radiation (Dudhia 1989), Noah
land surface model (Tewari et al. 2004), Yonsei
University (YSU) planetary boundary layer (Hong et al.
2006), and the Tiedtke cumulus schemes (Tiedtke 1989)
on the 12-km domain only.

c. MOS setup

To find the optimal predictors to use for the MOS
multivariate regression equations, a screening regression
forward selection process is used as described in Wilks
(2011). The screening regression uses WRF-modeled $U$,
$V$, and $W$ wind components, wind speed (wspd), tem-
perature $T$, and pressure $P$ at every grid point within a
20 $\times$ 20 square (10 grid points in each direction) around
the SODAR of interest and at each of the vertical
levels in the WRF Model as potential predictors for the
4-km resolution WRF runs. Lapse rate and Richardson
number (RI) are calculated over the first few raw model
levels (from $\sim$20 to $\sim$150 m AGL) for each grid point
within the 20-gridpoint box to also be used as potential
predictors. For the 12-km runs, the same area within the
domain around the SODAR location is used to select
the pool of potential predictors as the 4-km runs, which
is about one-third of the number of grid points (3–4 grid
points in each direction) compared to the 4-km runs due
to the lower resolution. On the 12-km grid, an additional
comparison is made to determine if the amount of area
around a specific SODAR used to select potential pre-
dictors for the screening regression has an impact. To
make this comparison, MOS is trained using potential
predictors from three different areas: 1) the 12-km area
defined by the 6 $\times$ 6 square gridpoint box (base
screening area), 2) a 3 times larger 20-gridpoint area
around the SODAR location using grid points on the
12-km domain (large screening area), and 3) a single
point using horizontally interpolated model variables at
the SODAR station location (this is equivalent to the
MOS technique used in Glahn and Lowry (1972)). Thus,
these comparisons are able to highlight any increase in
MOS skill from using more potential predictors over
larger areas. Though many studies in the literature use
predictors at the single station location, others in other

---

FIG. 3. The number of vertical levels below 1 km AGL at the
Hereford location for the TTU real-time forecast system, which has
a total vertical resolution of 38 levels, are shown in blue, and the
same is shown for NOAA’s HRRR model, which has a total ver-
tical resolution of 51 levels, in green.
industries have also selected predictors over a given area around the station location (Perez et al. 2013; Herman and Schumacher 2016). The screening regression algorithm calculates the goodness of fit using the adjusted $R^2$ value between the observations and each potential predictor, and the best 20 predictors are selected. Each predictor is then a WRF Model variable located at a height and gridpoint location around the SODAR.

Once the optimal predictors are found, the MOS algorithm computes the multivariate regression coefficients. Using these coefficients, MOS forecasts are generated for future times and verified against observations. MOS was trained over all of the 24-h WRF forecast valid times, and also trained with 0000, 0600, 1200, and 1800 UTC initialization times separately for each of the horizontal and vertical resolutions and for each of the training periods all described in the WRF Model setup above. MOS was also developed for each of the five rotor layer heights measured by each of the two SODARs: 40, 60, 80, 100, and 120 m.

d. AnEn setup

The AnEn experiments performed in this study use optimal predictor weighting as described in Delle Monache et al. (2011) and Delle Monache et al. (2013). The AnEn experiments are performed on the 0000 UTC initialization time wind speed forecasts from the 2015 WRF dataset and the San Angelo SODAR observation location. The training period includes all forecast times within the 0000 UTC forecasts. This technique was also developed and verified at the five rotor layer heights: 40, 60, 80, 100, and 120 m.

The top four optimal predictors found using the screening regression during the MOS tests are used as four of the predictors for the AnEn method. The fifth predictor used is the WRF wind speed forecast at the SODAR heights for the initialization time of interest. These five predictors are ingested into Eq. (1) (Delle Monache et al. 2011) to find the best analogs:

$$F_t A_{t'} = \sum_{i=1}^{N_y} \sum_{j=-i}^{i} \left( F_{t,i+j} - A_{t,i+j} \right)^2.$$  

Equation (1) describes the method by which the distance between the current and past forecasts are calculated to determine the closest analog. Here, $F_t$ is the current deterministic WRF forecast for a location and time $t$, $A_{t'}$ is an analog at a past time $t'$ and at the same location and lead time as $F_t$, $N_y$ is the number of variables used as predictors, $w_i$ is the weights given to each predictor variable, $\sigma_i$ is the standard deviation of the predictor variables for a certain location over the training period of past forecasts, $i$ is half of the time window in which squared differences between analog and forecast values are calculated for a location, and $F_{t,t+i}$ and $A_{t,t+i}$ are the
values of the forecast and the analog for a variable in a
time window, respectively. The mean of this ensemble of
analogs forms a deterministic prediction that can be
compared to MOS forecasts and verified against obser-
vations. The AnEn predictor weights are found using
the static weighting strategy explained in Junk et al.
(2015b), as they found that the dynamic weighting ap-
proach did not improve the AnEn method over the static
weighting approach. Hourly data is used in this study,
so a half-time window of \( t = 1 \) is employed here as
in Delle Monache et al. (2011). Additionally, 10 analogs
are chosen for each forecast, again following Delle
Monache et al. (2011).

The AnEn method is first tested on the 2015 12-km,
38-level WRF dataset to quantify any skill added by
using the optimal weighting strategy described above for
all five predictors versus weighting the five predictors
equally. AnEn is run with the optimal predictor weights
and then again with equal weights applied to the pre-
dictors. AnEn with equally weighted predictors is found
to be not statistically significantly different in mean ab-
solute error (MAE) than AnEn with optimally weighted
predictors. The AnEn with equally weighted predictors is
used for the AnEn experiments throughout the rest of
the study because of its lower computational cost. The
mean of the AnEn members is used as a deterministic
forecast to compare with the deterministic MOS and
raw WRF forecasts.

e. Experiment design

Each variation in configuration (experiment) is com-
pared to the use of the 2015 WRF training dataset that
has 12-km horizontal resolution and 38-level vertical
resolution that will be referred to as the base exper-
iment. To test the impact of horizontal resolution on
MOS and AnEn skill, the two techniques are trained
over one year (2015) for both the 12-km forecasts (base
experiment) and 4-km forecasts (horizontal hi-res ex-
periment). The vertical resolution is also varied for
MOS and AnEn using both 38 (base experiment) and
51 levels (vertical low-res experiment). There are fewer
model levels near the boundary layer (1 km and below)
in the 51-level model setup (9 levels) compared to
the 38-level model setup (13 levels) as can be seen in
Fig. 3. The 51 levels, which came from the HRRR op-
erational configuration, are concentrated more toward
the tropopause, whereas the 38 levels, from the TTU
real-time configuration, are focused more in the boundary
layer. Thus, here the 51-level configuration is considered
the low-resolution experiment. MOS was trained over
each initialization time separately (0000, 0600, 1200, and
1800 UTC) for the model resolution sensitivity tests and
the results are presented below in section 3. AnEn was
trained over the 0000 UTC initialization time for all
AnEn experiments and compared to the 0000 UTC

<table>
<thead>
<tr>
<th>Postprocessing technique</th>
<th>Aspect tested</th>
<th>WRF Model dataset used</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOS</td>
<td>No. of predictors</td>
<td>2015, 12 km</td>
</tr>
<tr>
<td></td>
<td>Training set length</td>
<td>6 month (2015), 12 km</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015, 12 km</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2014–15, 12 km</td>
</tr>
<tr>
<td>Vertical resolution</td>
<td>2015, 12 km, 51 levels</td>
<td>2015, 12 km, 38 levels</td>
</tr>
<tr>
<td>Horizontal resolution</td>
<td>2015, 12 km</td>
<td>2015, 4 km</td>
</tr>
<tr>
<td>AnEn</td>
<td>Optimal predictor weighting</td>
<td>2015, 12 km</td>
</tr>
<tr>
<td></td>
<td>Horizontal resolution</td>
<td>2015, 12 km</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2015, 4 km</td>
</tr>
<tr>
<td>Vertical resolution</td>
<td>2015, 12 km, 51 levels</td>
<td>2015, 12 km, 38 levels</td>
</tr>
<tr>
<td>Training set length</td>
<td>6 month (2015), 12 km</td>
<td>2015, 12 km</td>
</tr>
</tbody>
</table>

FIG. 5. (top) The San Angelo 0000 UTC MAE profile plot of
using a different number of potential predictors in the screening
regression and in the MOS model. The blue solid line represents
the 12-km MOS MAE over the larger screening area, the purple
solid line is 12-km MOS MAE over the original screening area,
and the solid black line is 12-km MOS MAE over using the
station location-only predictors. (bottom) As in the top panel,
but for Hereford. The “+” and “*” symbols indicate where
there is a statistical difference between the points with the same
 corresponding symbol.
MOS results. To provide an additional comparison to these experiments, a simple bias-removal technique was applied to the raw WRF forecasts at each model resolution tested. To apply this bias-removal technique, an overall model bias was calculated between the SODAR observations and the raw WRF Model over the training periods for the different model resolutions. Once this overall bias was calculated, it was subtracted from the raw WRF forecasts over the validation period to create a corrected forecast. This technique was included for completeness and visual comparison with the post-processing methods of interest (AnEn and MOS). For this reason, the results will focus mainly on the MOS and AnEn methods.

For the training dataset length sensitivity test, MOS and AnEn are trained over a 1-yr period, 2015 (base experiment), as well as over 6 months (short period experiment) of WRF reforecasts on the 12-km grid. The 6-month training period included February, April, June, August, October, and December 2015, sampling from all four seasons to allow for a fair comparison with the base experiment. In addition to the short period, MOS was also trained over a 2-yr period, 2014–15 (long period experiment).

To test how the number of predictors affects MOS skill, the numbers of MOS predictors are varied from 20 to 1. For this predictor test the base forecast dataset is used. An example of the predictors selected from the base screening regression area for the San Angelo SODAR using the base experiment training dataset 0000 UTC forecasts for 40-m wind speed as the predictand are shown in Fig. 4. The predictors and their locations are shown in the map in relation to the SODAR location.

The various MOS and AnEn sensitivity experiments performed in this study are summarized in Table 1.

**f. Verification**

Verification against observations is done to measure the relative skill of different MOS and AnEn configurations and determine the added value to raw WRF forecasts. The primary verification metric used here is...
MAE calculated over the 2016 verification period for all the experiments.

3. Results and discussion

a. MOS and AnEn sensitivity to different configurations

1) Area over which predictors are chosen

The screening regression to select predictors to use within MOS and AnEn selects from a pool of potential predictors from an area that can vary in size. Here we compare MOS results using predictors from the original base screening area (equivalent to about $6 \times 6$ grid points centered around the SODAR), using predictors from a screening area that is 3 times larger, and using predictors at the SODAR station location only (interpolated from surrounding model grid points). This comparison is performed using all forecast hours within the 0000 UTC initialization WRF runs, and MOS is trained for both San Angelo and Hereford for all five SODAR heights.

Figure 5 shows the MAE profiles of each of the three screening areas within MOS for San Angelo and Hereford. It is found that using the predictors from the SODAR station location, which is typically how MOS is configured operationally and in previous studies, has the largest MAE overall but shows an improvement over raw WRF forecasts of 36%, averaged over all levels at San Angelo (not shown). This method also yields an average improvement over raw WRF forecasts of 11% at Hereford (not shown). The MOS larger screening area has slightly less error compared to the original screening area, but they are very similar for San Angelo and Hereford locations. The original screening area MOS shows average improvement over raw WRF forecasts of 37% at San Angelo and 14% at Hereford while the larger screening area shows an average improvement of 40% at San Angelo and 14% at Hereford. These three screening areas are not significantly different from each other except for at the Hereford location at 40 and 60 m where the difference of MAE between the MOS using the station-only predictors and the
larger screening area shows statistically significant differences at a 95% confidence level. Here, the larger screening area shows significant improvement in the MOS forecasts over the MOS using station-only predictors. These results reveal that there could be a justification in using a smaller screening area to select predictors for MOS because, overall, there are not statistically significant differences in the improvement of the MOS forecasts and using a smaller area can aid in decreasing computational cost. Though the MAE differences are not statistically significant, it is also seen that the larger the screening area the smaller the MOS forecast MAE. For the purposes of this study, the smaller original base screening area was used due to it being less computationally intensive while still providing improved MAE compared to using station selected predictors.

2) NUMBER OF MOS PREDICTORS

The relationship between the number of predictors and MOS forecast skill is evaluated by MAE for each initialization time, both SODARs, and each of the five heights. Figure 6 shows results for both SODARs and each initialization for 100-m wind speeds (results are extremely similar at the other four levels). In all cases, a single MOS predictor provides significant improvements to the raw WRF forecasts. MAE values also show that for both SODARs and for each initialization time, little additional skill is gained beyond about six MOS predictors.

3) HORIZONTAL RESOLUTION

MOS results involving different horizontal resolutions are shown here for both SODARs, and for all four forecast initializations, at all five heights located within the rotor layer (40, 60, 80, 100, and 120 m). Figures 7 and 8 show profiles of MAE for the horizontal hi-res and base (4 and 12 km) MOS forecasts as well as the corresponding raw WRF forecasts for the San Angelo and Hereford SODARs, respectively. At the San Angelo SODAR and all four initializations, the raw WRF 4-km forecast MAE values are smaller than those on the 12-km grid. It is expected that the higher resolution 4-km forecasts would have larger errors compared to the
base 12-km forecasts. Using verification metrics, like MAE which was used in this study, tend to favor smoother (more spatially averaged) lower resolution forecasts leaving higher resolution forecasts to be doubly penalized for containing higher spatial variation (Gallus 2002). At Hereford, the MAEs between the 4- and 12-km WRF are closer to each other and do not show statistically significant differences unlike at San Angelo. The two locations show different relationships in error between the two horizontal resolutions neither of which is exactly what we would expect to see (higher resolution WRF with larger MAEs). This could potentially be due to topographic and land surface differences at the two locations. San Angelo has much more terrain variation compared to Hereford. These terrain impacts would be much better resolved at the higher resolution. However, further investigation that is beyond the scope of this study would be required to fully understand this behavior.

In terms of the added value of MOS, both SODAR locations reveal that MOS MAE values for both the horizontal hi-res and base experiments are significantly smaller than their respective raw WRF forecast MAE profiles. Further, both horizontal resolutions’ MOS MAE profiles shown in Figs. 7 and 8 are nearly the same and are not statistically different at a reasonable confidence interval. Interestingly, this indicates that applying MOS to the 12 km is able to achieve nearly identical skill as the horizontal hi-res MOS configuration, even though the raw WRF 12-km forecast has significantly larger error. At least with respect to average point forecasts of wind speed in the lowest 120 m, this indicates substantial computational cost can be saved through running lower resolution configurations and applying MOS. These results do not appear to be sensitive to the initialization time of the WRF forecast as they do not vary much at different initialization times. Therefore, the rest of these results will be presented using the 0000 UTC initialization time. The 12-km base MOS shows an average improvement of 37% over raw WRF forecasts at San Angelo, and an average improvement of 14% over raw WRF forecasts at Hereford. By comparison, the horizontal hi-res MOS shows an average improvement of 25% over raw WRF forecasts at San Angelo, and an average improvement of 11% over raw WRF forecasts at Hereford. Though the horizontal hi-res experiment shows less improvement overall, since the 4-km WRF has less error the resulting MOS forecasts from the two different resolutions end up having similar MAE values.

Figure 9 shows the MAE profiles at San Angelo of the horizontal hi-res and base experiment (4 and 12 km) AnEn mean, MOS, raw WRF forecasts, and a simple bias removal technique is also shown for comparison that was applied to the two raw WRF forecasts and the different resolutions. The horizontal hi-res AnEn mean MAE is the lowest of all the methods shown, though only slightly better than the base AnEn mean. The horizontal hi-res AnEn mean consistently has lower MAE than the horizontal hi-res MOS. This difference is greater than the difference between the 12-km AnEn mean and the 12-km MOS, which could imply that the higher horizontal resolution of the raw NWP model may play a greater role in the skill of AnEn than for MOS. The base and horizontal hi-res AnEn means show average improvements of 37% and 27% over the raw WRF forecasts, respectively. These values of improvement compared to the average percent improvement over WRF from the base and horizontal hi-res MOS are similar (37% and 25% respectively). While the horizontal hi-res AnEn mean is significantly better than the raw WRF forecasts and has the lowest error of the postprocessing methods tested, it is not statistically significantly better than the either resolution MOS, or the base 12-km AnEn mean. Ultimately, this indicates that either the MOS or AnEn method at coarser horizontal resolution is able to provide similarly good results than at higher resolution, thus saving on computational and storage requirements.

4) VERTICAL RESOLUTION

Figures 10 and 11 show MAE values at the five rotor layer levels for the 38- and 51-level MOS forecasts as
well as the corresponding raw WRF forecasts for the San Angelo and Hereford SODARs, respectively, for all four initializations. Both figures show that the raw WRF 51-level forecast (vertical low-res) MAEs are larger than those using 38 levels. This difference is statistically significant at all heights and initialization times at both SODAR sites.

For both SODARs it is also seen in Figs. 10 and 11 that the MOS MAE profiles for both vertical low-res and base experiments are significantly smaller than their respective raw WRF forecast MAE profiles. Therefore, similar to the horizontal resolution test, MOS is significantly better than the raw WRF forecasts for both locations, all heights within the rotor layer, and for all initialization times. Also, like the horizontal resolution test, these results are not sensitive to the initialization time and so only the 0000 UTC initialization time results will be provided. The base 38-level MOS shows an average improvement of 37% over raw WRF forecasts at San Angelo and an average of 14% over raw WRF forecasts at Hereford. The vertical low-res MOS shows an average improvement of 49% over raw WRF forecasts at San Angelo and an average of 25% over raw WRF forecasts at Hereford.

Much like the results at different horizontal resolutions, the MOS-corrected forecasts at both SODARs using both vertical resolutions shown in Figs. 10 and 11 are very similar to each other and are not statistically significantly different. These results reiterate that larger errors at different resolutions are corrected by MOS to a larger degree, resulting in postprocessed forecast skill that is nearly the same regardless of the resolution of the NWP forecast (at least for the resolutions tested in this study). Forecasts at finer grid spacing require substantially more computational resources. Due to these results, in which similar skill can be obtained through MOS at lower resolution, substantial computing cost can be avoided. In turn, such computational cost
savings may allow the enhancement of other modeling aspects (e.g., larger ensemble size or more detailed physics parameterizations) that could further improve forecast skill.

Figure 12 shows the MAE profiles at San Angelo of the base and vertical low-res AnEn mean, MOS, WRF, and the bias removal method. Figure 12 does not show a large difference between MAEs for either of the two AnEn mean vertical resolutions or between the vertical low-res AnEn and MOS. Unsurprisingly, these small differences between these postprocessing tests are not statistically significant. The base AnEn mean shows an average improvement of 37% over the raw WRF, and the vertical low-res AnEn mean improves upon the raw WRF by an average of 49%. Both of these improvements are identical to that of the MOS improvements at the different vertical resolutions and are statistically significant improvements by AnEn compared to raw WRF.

Therefore, either of the methods with either resolution will show improvement over the raw WRF forecasts significantly and in a similar way. Thus, a coarser resolution model, both in the horizontal and vertical, is sufficient for either AnEn or MOS to improve raw WRF forecasts for average wind speeds in the lowest 120 m of the atmosphere.

5) TRAINING SET LENGTH

Different MOS training periods are tested to understand the dependence of MOS skill on training set length. Results can be valuable to guide MOS applications since training datasets can be large and difficult to manage. Figure 13 shows MAE profiles for the different training periods as well as those for the raw WRF forecasts for the San Angelo and Hereford SODAR, respectively, for the 0000 UTC initializations. The short period (6 month) MOS shows average improvement of 36% over raw WRF forecasts at San Angelo and 12% average

**Fig. 11.** Hereford vertical resolution MAE profile panel plot. The blue solid line represents the 38-level vertical hi-res MOS, and the green solid line is the 51-level vertical low-res MOS MAE profile. The blue dashed line is the 38-level WRF, and the green dashed line is the 51-level vertical low-res WRF MAE profile. Significant differences between lines are marked by matching symbols at the top of the plots.
improvement over raw WRF forecasts at Hereford. The base 1-yr MOS shows average improvement of 37% over raw WRF forecasts at San Angelo and 14% average improvement over raw WRF forecasts at Hereford. The long period (2 year) MOS shows average improvement of 37% over raw WRF forecasts at San Angelo and 15% average improvement over raw WRF forecasts at Hereford. Figure 14 shows the additional AnEn MAE profiles using the base and short training periods at San Angelo. It is revealed from this that both length periods of AnEn have an average improvement over raw WRF of 37%. This is similar improvement compared to the long, base, and short period trained MOS.

All statistical postprocessing configurations in this training period experiment significantly reduced the error of the raw WRF forecast while each configuration is statistically similar to each other. Skill did generally improve somewhat with longer training periods, due to the computational cost of producing and storing longer training periods, using only 6 months or 1 year of training data is sufficient here.

Even though the different training periods tested were found to result in similar reduction in error among the postprocessing methods, in the future, it would be useful to compare AnEn and MOS with training periods that are much longer (much greater than two years) to see if there are statistically significantly better forecasts. AnEn skill in particular benefits from as long a training period as possible, in order to draw from a larger sample of past observations from which to draw the analog ensembles, thus yielding more skillful AnEn means. Forecasting for extreme events (i.e., the tails of the distribution of observations) with AnEn would benefit the most from a multiyear training period that captured more of those extremes.

b. Detailed MAE comparison of MOS and AnEn

To more thoroughly compare the MOS and AnEn methods to each other, their MAEs were stratified by month, wind speed, and wind direction. Their frequency distributions are also compared with the observed wind speeds. This allows for a better investigation into the characteristics of these methods’ errors and help to determine which technique may perform better under certain conditions.

1) MAE BY MONTH

Figure 15 displays the MAE of the raw WRF Model, MOS, AnEn, and bias removal forecasts by month for each of the five rotor layer heights. The raw WRF Model
has higher error in the winter months than the others except for April. There also doesn’t appear to be much sensitivity to rotor layer height with these results. The MOS and AnEn methods show large improvements over the raw model in every month except March where the MOS MAE is larger on average compared to the raw WRF Model and the AnEn forecasts’ MAEs. This may be due to the raw WRF Model having low MAE overall in March and therefore the MOS correction being applied to the already good WRF forecasts is causing additional error. Aside from March, The MOS and AnEn methods are similar to each other in terms of their monthly MAEs and not one method looks to handle different seasons better than the other.

2) MAE BY WIND SPEED

Figure 16 shows the MAE by wind speed of the raw WRF Model, MOS, AnEn, and bias removal forecasts for the five rotor layer heights. The results are shown only for wind speed bins with greater than 10 records. This comparison was made to determine which method performs best at different wind speeds particularly near typical cut-in and cut-out thresholds which we assumed to be 3 and 25 m s⁻¹, respectively. However, there were not enough records at the higher wind speeds to be able to consider cut-out wind speeds. MOS and AnEn MAEs are very similar at every wind speed. At higher wind speeds (near ~9–10 m s⁻¹), the raw WRF MAE is lower and the statistical postprocessing methods’ MAE are higher. The raw model does better so the methods may be over correcting the forecasts by applying a larger correction than what would be needed. Also, there is a lower record count at higher wind speed bins that could be skewing the data here.

3) MAE BY WIND DIRECTION

MAEs by wind direction for raw WRF, MOS, AnEn, and the bias removal method are shown in Fig. 17 to determine how successful these methods are compared to each other at different wind directions. The AnEn and MOS MAEs are similar at every wind direction and not one method is significantly better than the other. These results are not sensitive to the different rotor layer wind speeds.

4) FREQUENCY DISTRIBUTION COMPARISON

The relative frequency distributions for the MOS and AnEn forecasted wind speeds for each of the five rotor layer heights are compared to the observed wind speed distribution in Fig. 18. This comparison was done to determine if there could be any differences in terms of variability in the two postprocessing methods’ deterministic forecasts. As is shown in Fig. 18, there appears to be many similarities between the MOS and AnEn forecast distributions. They both have clear infrequency at the tails of the distributions with higher frequency at the medium wind speeds compared to the observed distributions. Unsurprisingly, a limitation that is shared by both types of postprocessed forecasts are that they are too smooth compared to the observations and some of the true variability is lost. In the future, it would be interesting to investigate object-based MOS and AnEn using metrics like timing. This could be especially interesting to the wind industry in how these object-based forecasts could lead to possible improvement in ramp forecasts.

4. Summary and conclusions

Two statistical postprocessing techniques, MOS and AnEn, were applied to raw WRF rotor layer wind speed forecasts to determine the skill that may be added to the deterministic model forecasts after each postprocessing method was applied. In addition to testing the methods’ forecast skill compared to the raw model and compared to each other, different aspects of each method were tested to determine which were most important to improving the WRF Model forecasts. Both MOS and AnEn significantly improved the raw deterministic WRF Model forecasts for each configuration tested and these methods were statistically similar to each other in general. Either of these methods with any configuration
could be used to improve the WRF Model forecasts of rotor-layer wind speed forecasts.

The first aspect that was evaluated in this study was the screening regression area used to select potential predictors. When using a larger area of potential predictors, there was no statistical significance among the three different screening areas, and they all significantly improved over WRF forecasts. Therefore, using a larger area, there are more potential predictors to choose from and results can be made marginally better, but a smaller area will still significantly improve the WRF forecasts and is computationally less expensive.

The optimal number of predictors for the MOS equations was also tested by training MOS with 1–20 predictors. The MOS model skill for each model with a different number of predictors showed that the error decreased after more predictors were added, but the error generally became nearly constant after around six predictors. Therefore, using more than about six predictors is unnecessary in this case and does not substantially enhance the MOS forecast.

The next tested aspect was the resolution of the WRF Model (both horizontal and vertical) to determine if having a higher resolution model influences the skill of the MOS and AnEn forecasts. This was done by training MOS and AnEn on two different horizontal resolutions (4 and 12 km) and on two vertical resolutions (38 and 51 levels). There was no statistically significant difference in error between either of the two horizontal resolution-trained forecasts and using either of the horizontal resolutions significantly improved the WRF Model forecasts. These results are similar for training

![Figure 15](image Link)
MOS and AnEn using two different vertical resolutions as well. The MOS- and AnEn-corrected error was statistically similar for both vertical low and hi-res WRF forecasts despite a difference in the number of vertical model levels near the surface. Because both horizontal resolution and vertical resolution MOS and AnEn models have significant MAE improvement over raw WRF, the coarser resolution model can be used for average raw WRF forecast improvement by MOS or AnEn thus facilitating wiser use of limited computational resources.

MOS was trained over three different periods, 6 months, 1 year, and 2 years, to determine if a shorter training period affects skill of the MOS forecasts. AnEn was also trained over a 1 year and 6-month period to compare with MOS. The 6-month trained MOS and AnEn had slightly higher MAE than the 1- and 2-yr training periods, but this difference was not statistically significant. The 6-month MOS, like the 1- and 2-yr MOS, significantly reduced the error of the raw WRF Model. Like MOS, the longer the training period for AnEn, the better the skill of the forecast, though the difference in skill with training length was not statistically significant, nor was the difference between the AnEn forecasts and the 6-month MOS forecast MAE. The 6-month AnEn forecast did significantly improve WRF forecasts, however. It can be computationally expensive to produce more than two years of training data and based on the San Angelo results, it may not be necessary to train MOS or AnEn over more than one year of data, or even just 6 months of data that samples the full annual cycle.

Overall, both MOS and AnEn methods were similar and sufficient in significantly reducing the errors in the
raw deterministic WRF Model forecasts of wind speed in the rotor layer in west Texas. Even when these errors are stratified by month (season), wind speed, or wind direction, both MOS and AnEn show similar MAEs and improvement over the raw WRF Model forecasts. There is an exception, however, at high wind speed forecasts where the raw WRF gives better predictions. Both postprocessing methods also produce much smoother forecasts when their distributions are compared to the true observed wind speeds. They are able to reduce the overall systematic bias but much of the variability is lost. With these postprocessing methods, a coarser resolution model configuration and at least a 6-month training period was found to be sufficient to improve the WRF Model forecast skill. A few benefits of AnEn is that it is less computationally expensive because it uses only the ten closest 24-h forecasts (analogs) in the training period, and therefore, fewer points compared to MOS, which uses all data in the training period for its calculations. AnEn also uses the corresponding observations found at the times of the analogs to serve as ensemble members instead of applying a correction to the mode forecasts as is done with MOS. Therefore, AnEn forecasted wind speeds are able to mimic the variability of the observations, whereas MOS forecasts tend to follow the smoothness of the raw model. Much of the variability in the AnEn may be lost though by taking the mean for the deterministic forecast. However,
both methods are easily able to be integrated into any forecast model and will lead to significantly improved low-level wind speed forecasts for any given observation location with sufficient observations despite the methods’ specific configurations as they relate to number of predictors, forecast model resolution, and training period length over samples of the full annual cycle.

**Acknowledgments.** This work was funded by the National Science Foundation Grant 1450177. The National Center for Atmospheric Research is sponsored by the National Science Foundation. I would like to acknowledge the West Texas Mesonet (WTM) for providing the SODAR data that were used in this study. Also, this could not have been accomplished without the Texas Tech High Performance Computing Center (TTU HPCC) and NCAR for providing facilities and computing for this work. This research could not have been possible without the analog ensemble code provided by Luca Delle Monache (NCAR) and Will Cheng (NCAR) as well as valuable support and conversations regarding the code.

**REFERENCES**


