The Reliability and Skill of Air Force Weather's Ensemble Prediction Suites

THESIS

Derek A. Burns, 1st Lieutenant, USAF

AFIT-ENP-MS-16-M-059

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENP-MS-16-M-059

THE RELIABILITY AND SKILL OF AIR FORCE WEATHER'S ENSEMBLE PREDICTION SUITES

THESIS

Presented to the Faculty

Department of Engineering Physics

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Master of Science

Derek A. Burns, BS

1st Lieutenant, USAF

March 2016

AFIT-ENP-MS-16-M-059

THE RELIABILITY AND SKILL OF AIR FORCE WEATHER'S ENSEMBLE PREDICTION SUITES

THESIS

Derek A. Burns, BS
1st Lieutenant, USAF

Committee Membership:

Lt. Col. Kevin S. Bartlett, PhD
Chair

Evan L. Kuchera
Member

Lt. Col. Robert S. Wacker, PhD
Member

AFIT-ENP-MS-16-M-059

Abstract

Deterministic weather models are limited by the fact that they depict one of many plausible forecasts of the atmosphere. Weather models will always be prone to error, especially since sparse observations make it impossible to represent the true initial state of the atmosphere. Ensemble weather models that represent multiple plausible forecasts are the next progression of numerical weather prediction and need further operational testing. Ensembles provide estimates of the probability of certain weather forecast outcomes, which are especially valuable to decision makers who apply risk management to operational decisions. The Ensemble Prediction Suite (EPS) used at the 557th Weather Wing (557 WW) provides probability based forecasts for thousands of worldwide locations. These Point Ensemble Probability (PEP) bulletins are tailored specifically to the United States military and its criteria for operationally significant weather thresholds. During April to October 2013, a validation study by Clements was performed on the PEP bulletins from 557 WW's Global EPS, as well as the 20 km and 4 km resolution Mesoscale EPS across 10 geographically diverse locations. The study found that the PEP products over forecast lightning, while precipitation and wind forecasts improved with increased horizontal EPS resolution. Since then, significant changes have been made to how the EPSs generate products. This study assesses additional weather parameters and compares 557 WW global and mesoscale EPS at 17 Continental United States (CONUS) locations. The PEP bulletins will be compared to climatology, METARs, and Earth Networks Total Lightning Network (ENTLN) data to generate reliability diagrams and Brier Skill Scores (BSS). Results from April to October of 2015 show

that each EPS is underforecasting ceilings and visibility for most forecast hours at several locations. The underforecasting of ceilings is most severe at Vandenberg AFB, an area prone to frequent marine layer fog and stratus. The MEPS 4 km also shows significantly better lightning forecast skill compared to the other EPS grid scales. However, each EPS is susceptible to overforecasting lighting at night. Finally, in areas with complex terrain, wind forecasts are degraded with decreasing model resolution.

Acknowledgements

I would like to thank Lt Col Kevin S. Bartlett and Lt Col Robert S. Wacker for their expert advice, selfless commitment to my education, and for being outstanding professors and role models during my time here at AFIT. I would also like to thank Mr. Evan L. Kuchera and Mr. Jeff H. Zautner for their help and prompt responses to my needs and requests for this project. And of course, I thank my parents and all of my friends and family for their love, support, and the happiness they bring to my life.

Derek A. Burns

Table of Contents

Page

List of Figures

Figure Page

List of Tables

THE RELIABILITY AND SKILL OF AIR FORCE WEATHER'S ENSEMBLE PREDICTION SUITES

1. Introduction

1.1 Motivation

Communication of weather information in the present day still finds itself biased towards the paradigm of determinism. A deterministic weather model provides a single value forecast, answering questions such as how much will it rain or what the wind speeds will be for a specific time. Such information is useful and straightforward for forecasters and their customers because a yes or no answer to what will happen tomorrow is easy to communicate. However, model forecasts are always prone to error which generally grow with time, making long term forecasts less reliable and more uncertain.

Weather forecast uncertainty is sometimes misunderstood by military decision makers. Disregarding the potential uncertainty in a forecast can hinder mission optimization and success. For example, if a sensitive drone on a mission is vulnerable to winds over 35 kt, a deterministic forecast of 25 kt may not prompt any protective action. However, a probabilistic forecast of a 30 percent chance of winds over 35 kt might prompt action if operators are not willing to take substantial risk. Probabilistic forecasts are also known as stochastic forecasts. They objectively estimate the uncertainty by attempting to account for the inherent errors. Ensembles of weather models are the key to implementing stochastic forecasting methods. No forecast is complete without a quantification of its uncertainty (Ban,

2007), and uncertainties will always exist due to the chaotic nature of the atmosphere and limitations in model physics and resolution.

The DOD is beginning to migrate away from reliance on deterministic models in light of this fact. Ensemble products are available today, but over the next 5 years the Air Force plans to greatly improve the resolution of global and regional ensembles as they become the dominant source of information for weather forecasting. Until then, much work needs to be done to gain the trust of end users and weather forecasters in ensemble forecasting capabilities. Part of that work is assessing the skill and reliability of ensemble prediction, which is the main focus of this thesis.

1.2 Applications to Operational Risk Management (ORM)

Ensembles provide stochastic forecasts that enable decision makers to apply ORM. ORM maximizes gains or minimizes losses through assessing the risks and costs associated with making certain decisions. Sometimes we are willing to accept the risk of incurring substantial losses if it optimizes the mission costs and benefits over time (Eckel et al., 2008).

Eckel et. al. develop a typhoon evacuation scenario in which a decision maker chooses to evacuate or remain in place depending on the forecast. Each action has a cost which varies depending on the likelihood of typhoon impacts to the base. One decision-maker takes a deterministic approach and evacuates if the model exceeds a certain threshold. The second decision-maker takes a stochastic approach and decides on whether the probability of damaging weather exceeds a certain threshold. The author shows that the deterministic operator can occasionally make the better decision that saves the most resources. However over many simulations, the stochastic operator does best. With ORM, the stochastic operator avoided the most

costly outcomes more often. It is clear from the author's scenarios that stochastic models are valuable in providing an assessment of risk that is needed for sound ORM principles to optimize the mission. Although the benefits may not be realized immediately, probabilistic forecasting wins in the end.

1.3 Economic Value

The main sources of error in numerical weather prediction come from either external error (error growth due to model deficiencies) or internal error (the self-growth of error from the initial conditions) (Reynolds et al., 1994). External error can be reduced by improving the model representation of physics or by increasing resolution. Internal error can be reduced by using an ensemble forecast or by improving the analysis of the initial state of the atmosphere. It is not completely clear which of the two sources contributes the most error, but Leith theorized that error growth from external sources is linear, while error growth from internal sources is exponential (Leith, 1978). Studies by Wergen (1982), Wallace (1983), and Arpe et al. (1985), supported his theory. Under that assumption, Reynolds concluded that in the mid-latitudes, total forecast error can be reduced from reducing internal, rather than external, sources of error (Reynolds et al., 1994). An important implication here is that ensembles, which aim to reduce sources of internal error, should have the most impact on improving forecasts in the mid-latitudes.

Additionally, increasing model resolution on large scales requires exponential increases in computing power and resources. Ensembles also require additional resources to run multiple models, but the needs are not as significant as the computing power required to increase model resolution. In fact, ensemble models can be run at lower resolutions while maintaining as much, if not more, skill than their higher resolution deterministic counterparts (Tracton and Kalnay, 1993).

Deeper analyses on the economic value of ensembles are given by Richardson (2000) and Zhu (2002). So from both an economic and practical perspective, ensembles appear to have significant potential for improving the accuracy of weather forecasts over deterministic models, especially in the mid-latitudes.

1.4 Advantages Over Deterministic Forecasting

Ensembles present several utilities and advantages over conventional deterministic forecasting techniques. Ensembles forecast numerous outcomes that are all within the realm of possibility and important decisions may hinge on the probability that an outcome is realized. In some cases, an ensemble may detect the development of a new weather system long before the deterministic model can (Toth et al., 1997). Ensembles also display more consistency from day to day and improve the skill of medium to long range forecasts (Toth et al., 1997). Additionally, ensemble spread, which measures how much the ensemble member forecasts differ from each other, is a great indicator of confidence. If the ensemble spread is large, the forecaster is aware that a prediction may have a potential for significant error, while a low ensemble spread indicates a high level of certainty. Ensembles, therefore, support the major goals of numerical weather prediction (NWP), which are to improve and predict forecast skill in order to improve the overall utility of NWP products.

1.5 Research Topic and Objective

Given the value and need for a skillful EPS, the main objective of this research is to study the performance of output from the three 557 WW EPSs: the one-degree resolution Global Ensemble Prediction Suite (GEPS), and the 20 km and 4 km resolution Mesoscale Ensemble Prediction Suites (MEPS20 and MEPS4). A popular

557 WW EPS tool among Air Force weather forecasters the Point Ensemble Probability (PEP) bulletin. It is useful to DOD because it provides probabilistic information for point locations worldwide on operationally significant weather criteria. Probabilistic forecasts from PEPs will be assessed in this study based on how much the predicted probabilities match the actual frequency of occurrence. This validation will help 557 WW to make decisions on future implementations and provide forecasters in the field with a better metric for interpreting EPS data.

This work continues the work Clements (2014) who initiated validation of 557 WW PEP products in 2014. This study will expand upon his work by examining more forecast locations of interest and evaluating additional forecast variables.

1.6 Preview

Chapter 2 provides a general background on stochastic and ensemble forecasting with some detail on 557 WW EPS, followed by chapter 3 on the methodology used for conducting this research. Chapter 4 will discuss results, while chapter 5 will draw conclusions with recommendations for future research.

2. Background

2.1 Chaos and Uncertainty in the Atmosphere

In classical Newtonian physics natural phenomena are viewed deterministically. That is, a future state of a system is determinable given a set of physical laws and an initial state. This school of thought worked quite well for classical physicists who predicted the trajectories of planets in the solar system. It is still very applicable today, as many systems in nature are stable. A pendulum, for example, will always end up in the same stable state of rest no matter where it begins. If the atmosphere were such a system, forecasting and modeling would be easy. In reality, the atmosphere is chaotic and unstable with limited predictability. Lorentz pioneered studies in this field, assessing the non-periodicity of atmospheric flow and the theoretical limits of predictability (1963; 1969).

Lorentz made an important discovery when he was experimenting with a simple atmospheric model on a Royal-McBee LGP-30 computer capable of 60 multiplications per second Lorenz (1995). He ran the model and recorded the output, rounding it off only slightly. Lorentz ran the same model again later, using initial conditions from the rounded off data he recorded earlier. After some time, Lorenz found that with just a slight alteration of initial conditions, the model evolved into a state that had no resemblance to the original model run. Lorentz continued to study this issue for years. He concluded that, even if a model correctly simulates weather dynamics, and if the initial state is known almost exactly, there is always a finite limit of predictability in the atmosphere. A model will eventually make widely different forecasts if the initial conditions differ just slightly. Significant divergence can occur in just a few days on the synoptic scale, and in an even shorter time on the smaller mesoscales (Wilks, 2011).

Dynamical models are not perfect and our ability to accurately characterize the initial state of the the atmosphere is limited. Model dynamics are generally well-understood, but errors arise because models run at resolutions that are too large to explicitly resolve small-scale phenomena like precipitation and convection. Approximations are required to describe the development and effects of sub-grid processes from large scale information. Analyzing the state of the atmosphere is limited also because observations are sparse and prone to measurement errors. Data assimilation techniques are used to fill in the gaps, but those techniques are not perfect either. Errors also arise due to our imperfect description of soil type, vegetation properties, snow and ice cover, and sea surface temperatures.

Because of imperfect analysis techniques, our best guess at the initial state of the atmosphere is likely to be different from the true state. Therefore, errors are inevitable and thus the uncertainties of numerical weather prediction (NWP) are inevitable. Eady (1951) said that because of the unavoidable uncertainty, "forecasting is necessarily a branch of statistical physics in its widest sense: both our questions and answers must be expressed in terms of probabilities". This is the main motivation for taking a stochastic approach to forecasting.

2.2 The Stochastic Approach

Since there will always be degrees of uncertainty in the atmosphere, it is necessary to take a probabilistic approach in order to describe that uncertainty. Epstein (1969) made some of the first attempts to objectively quantify the uncertainty in initial state of the atmosphere and how it would evolve in the forecast. He understood that there could be many possible initial states around the best guess of the initial state. The collection of possible initial states could be thought of as a probability distribution function (PDF) in phase space (Wilks,

2011). The best guess of the initial state is analogous to the mean of some distribution where there exists an infinite number of possible solutions around the mean, with a likelihood that is proportional to the probability density. Epstein developed equations to estimate the PDF which captures all of the possible initial states and their likelihood. He proposed that model dynamics could evolve off of the distribution of initial states and lead to a distribution of possible future states.

Epstein referred to this type of forecasting as "stochastic dynamic prediction" (1969). Rather than run a model from one best guess analysis to arrive at a single future state, stochastic prediction involves a distribution of initial states that can lead to a number of possible future states. Although it is a valuable concept, keeping track of the evolution of the PDF from a system as complex as the atmosphere is impractical. In practice it is more feasible to sample a few initial states from the PDF and run an ensemble of models that are initialized from that sample. If the initial states are chosen wisely, an ensemble of forecasts can capture the majority of possible outcomes yielding an overall better forecast with more utility.

2.3 Ensemble Forecasting and Techniques

Ensemble forecasting begins with choosing a sample of initial states of the atmosphere. The initial states are intended to represent the range of possible errors in the best guess analysis. Each ensemble member is initialized a perturbed analyses. At first, the ensemble members begin from conditions that are very similar to each other. After running each ensemble for some time, the differences between the members may grow until each solution may be significantly different from the other. The ensemble of forecasts collectively represents a range of possible end states. It approximates how the probability distribution described by Epstein evolves in time.

Figure 2.1 provides an illustration of this concept with a two-dimensional phase space. The red shape on the left outlines the distribution of plausible initial states. Each red dot represents a sample of analyses from the distribution, with the assumed true state given as a black circle with a "T". The lines represent the evolution of each ensemble member through time. After 48 hours, the ensemble forecasts spread out from each other arriving at a distribution of forecasts. Each member is unique, yet all of the trajectories are plausible. It's impossible to know ahead of time which member will end up closest to the true state, but ensembles give insight as to where the true state will most likely be (Wilks, 2011).

The degree to which the ensembles spread apart tends to increase with time but ultimately depends on the stability of the dynamics in place. Some days the ensemble spread may be small, in which case there is less uncertainty in the forecast. Other times, when ensemble spread is large, there is more uncertainty. In this sense, an ensemble can predict how skillful a forecast may be. Whatever the spread, the mean of all the ensemble forecasts should perform better than any single control forecast based on the best analysis. Toth and Kalnay (1993) demonstrated this with a simple experiment on a 1991 version of the national meteorological center's global model. They compared the 5-day forecasts of a control model with an ensemble made of just two members. The ensemble members were perturbed from a control analysis. Over months of forecasting, the results showed that the ensemble outperformed the control model 80 % of the time in both hemispheres.

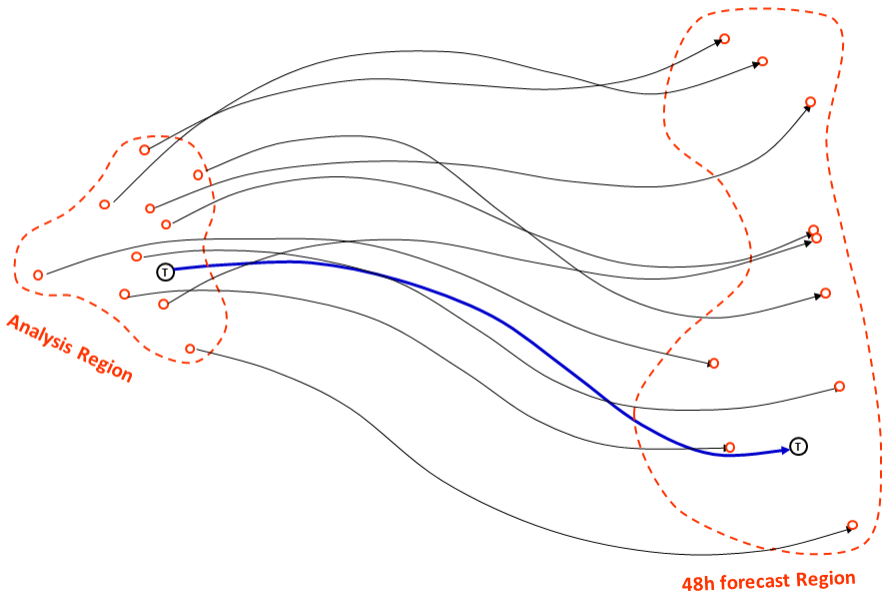


Figure 2.1. Two-dimensional phase space of an ensemble of forecasts. Figure used with permission from Evan Kuchera.

The question still remains as to how to actually perturb an analysis to sample plausible initial states of the atmosphere. First, it is necessary to know how an analysis is created. Generally, analyses are created with a combination of model and observational data. A model creates a forecast valid for the time of the analysis. This serves as a first guess. The first guess is a 6 hour forecast for most global models. Then, observations within a three hr window of the analysis time are used to correct the first guess through data assimilation (Kalnay, 2003; Warner, 2010). This serves as the best guess from which perturbations are generated.

After data assimilation, it is not known how far the analysis may deviate from the true state of the atmosphere. Generating perturbations essentially involves

guessing the errors believed to represent the expected variability of the atmosphere. The earliest techniques involved Monte Carlo methods that generated random perturbations that have amplitudes comparable to expected errors (Leith, 1974). This technique proved unsatisfactory because the random perturbations were unlikely to drive the atmosphere toward a new outcome and were more likely to develop gravity waves (Lacarra and Talagrand, 1988). Therefore the ensemble solutions would not diverge and represent the variability of forecasts.

Additional techniques have been developed to find perturbations that focus on the fast-growing unstable modes of the atmosphere. The techniques fall into three general categories: the breeding method (Toth and Kalnay, 1993), singular vectors (Kalnay, 2003), and ensemble Kalman filters (Houtekamer et al., 2005). The latter two are not applied in 557 WW ensembles and discussion of them is not needed here. However, the breeding method is simple and is used in part by GEPS. In this method, a control model is run alongside a model with a small, arbitrary perturbation. After six hours of integrating the models forward, the control forecast is subtracted from the perturbed forecast, creating a difference field. The difference between the two forecasts is scaled down, back to the size of an initial perturbation. That difference field is then added to the next analysis cycle as a new perturbation. This process is repeated several times at six hr intervals. By the end of the cycle, the new analysis contains the fastest growing modes. The slow growing modes are filtered out in the scaling process. Effectively, the method selects or "breeds" the fastest growing perturbations. Perturbations in today's ensembles are generated through modifications of the previously mentioned techniques or by hybrids of techniques.

Another technique is known as lagged average forecasting (Hoffman, 1983) and is applied to 557 WW MEPS. It is unique from other techniques because the initial

conditions are not directly perturbed. Instead, the ensemble is made of forecasts initialized from the current time as well as from previous times. Therefore, the ensemble is composed of forecasts of different age. For example, in 557 WW MEPS an ensemble member is initialized every two hours continuously. A MEPS forecast then contains forecasts from ensemble members that were initialized 2 hours ago, 4 hours ago, 6 hours ago and so on. The ensemble spread is produced by the variable forecast errors that develop from the newer and older members.

2.4 557 WW Ensembles

2.4.1 Global Ensemble Prediction Suite (GEPS).

GEPS is a global scale EPS. It consists of 62 total members that come from three established global ensembles: 21 members come from the Global Forecast System (GFS), 21 from the Global Environmental Multiscale (GEM) model (Cote et al., 1998), and 20 from the Navy Operational Global Atmospheric Prediction System (NOGAPS) (Hogan and Rosmond, 1991). GEPS is run twice per day at 00 UTC and 12 UTC with products available at 10 UTC and 22 UTC (Lisko, 2015). Output has 1 degree horizontal resolution at 6 hr intervals for up to 240 hr. Each ensemble system has particular resolutions, physics parameterizations, data assimilation, and ensemble perturbation techniques.

Global ensembles have lower resolution, but this allows the ensemble to contain more members that better represent the uncertainty in the atmosphere. They also have the ability to forecast longer lead times, sometimes on the order of a week. The disadvantage of lower resolution is that crude approximations sometimes have to be made to represent sub-grid processes like convection and boundary layer turbulence. The terrain is also heavily smoothed possibly leading to misrepresentations of phenomena like mountain waves and land/sea breezes.

2.4.2 Mesoscale Ensemble Prediction Suites (MEPS).

MEPS is a regional EPS capable of 20 km and 4k m horizontal resolutions. The 20 km resolution domain covers the northern hemisphere and a strip around the equator. The 4 km resolution domains cover areas including the Unites States, Europe, and the Middle East as well as relocatable domains. MEPS incorporates 12 versions of the Weather Research and Forecasting (WRF) model version 3.6.1 (Lisko, 2015). Each member uses different physics packages for parameterizing micro-physics, radiative transfer, planetary boundary layer effects, and convection. The 4 km MEPS, however, does not use cumulus convection schemes because the grid spacing is small enough to explicitly depict the general physics of convective systems (Done et al., 2004). Boundary conditions needed for each domain come from global forecast models that also vary depending on the member. Details on the model configurations can be found on the AFW-WEBS Wiki. The 20 km MEPS forecasts out to 132 hr at three hour intervals and the MEPS4 out to 72 hr at 1 hour intervals.

A significant change occurred to the MEPS on 27 July, 2015 when 557 WW introduced a rolling ensemble technique similar to lagged average forecasting. Before that, MEPS consisted of 10 members all run at the same time, providing one to two updates per day. In the rolling ensemble technique, each MEPS member is run individually at 2 hr intervals. Each run initializes from a new analysis based on the present observations through a type of data assimilation called Grid-point Statistical Interpolation (Kleist et al., 2009). The continuous assimilation of data reduces internal error contributed by analysis error and provides forecasters with more timely updates (Arpe et al., 1985).

The MEPS forecast is a combination of the previous 15 model forecasts valid for the same time. Each member is weighted equally. MEPS is ideal for short range

point forecasting due to its higher resolution. The lower resolution global models tend to blend the smaller scale features, but MEPS can resolve them and provide some indication of the uncertainty of those features.

2.4.3 Probability Generation.

The point probability products from MEPS and GEPS represent the expected probability of occurrence of various weather phenomena. The ensemble members do not predict probability directly, however. Ensemble forecasts are only sets of "deterministic realizations" and, therefore, are not "a prori" probabilistic forecasts (Jolliffe and Stephenson, 2012). To derive a probability from an ensemble, it is first necessary to make some assumption about the statistical behavior of the weather phenomena. The statistical behavior is described by a parametric probability distribution which can be used to calculate values of probability for weather phenomena.

A classic example of a parametric distribution is the normalized bell curve. Its shape is determined by parameters the mean and standard deviation of the data represented by the bell curve. These parameters enable the calculation of the probability that a data point falls within a certain range of the mean. The probabilities in PEPs are essentially calculated in this manner. The goal of this methodology is to represent the real data as much as possible with some distribution function. Although the distribution is abstract, having no physical connection to real-world phenomena, it approximates how weather parameters behave.

The normalized bell curve does not necessarily represent the behavior of the extreme weather phenomena relevant to PEPs, so other types of distributions are needed. PEPs give the probabilities that certain variables exceed a threshold, such as the probability of winds greater than 35 kt or visibility less than 3 sm. These

variables do not always follow an ideal Gaussian distribution. Instead, the "extreme value distribution" (Wilks, 2011) is used to calculate probability in MEPS and GEPS.

There is an important reason that extreme value distributions are applied in MEPS and GEPS. The theory of extreme value statistics states that a sample of maxima converges to an extreme value distribution as the number of sampled maxima increases (Coles et al., 2001). Coles called this the External Types Theorem, and it is analogous to the Central Limit Theorem (Wilks, 2011) which states that independent random variables eventually converge to a Gaussian distribution as the sample size increases. The External Types theorem is particularly applicable to PEPs because the thresholds are extrema of weather phenomena as opposed to a mean. PEPs use the Weibull distribution to calculate probabilities (Lisko, 2015).

A Weibull distribution is represented by the following equation (Wilks, 2011):

$$f(x) = \left(\frac{\alpha}{\beta}\right)\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1} exp\left[-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right], \quad x, \alpha, \beta > 0. \tag{1}$$

The parameters α and β are the shape and scale parameters, respectively. The variableγ is the shift parameter, and x is the variable output from the ensemble. The function outputs probabilities that are averaged among each ensemble in PEPs. For example with winds over land, the shift parameter is the sustained wind speed given by the ensemble member, α is 3 and β is sustained wind speed raised to the 0.75 power (Lisko, 2015). The shape of the distribution function depends on the type of variable to be forecast. Figure 2.2 shows Weibull distribution functions for various values of α. A summary of the algorithms used to calculate probabilities for each variable are listed in (Lisko, 2015).

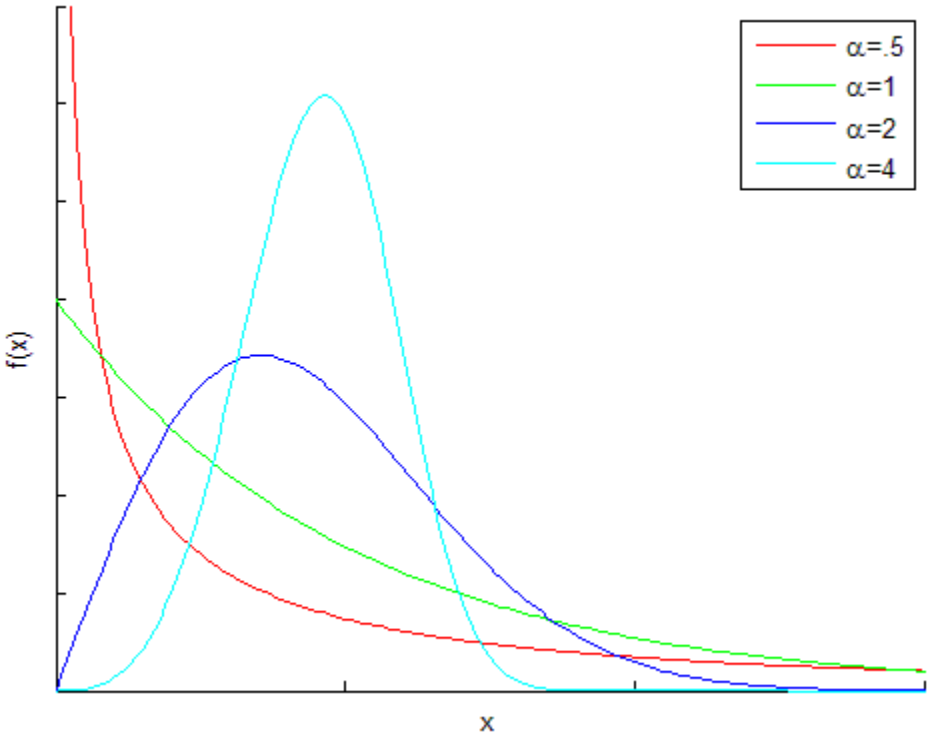


Figure 2.2. Example Weibull distribution functions for various values of the shape parameter α.

2.5 Previous Research

Validation of ensemble forecasts is not straightforward and has been the subject of much research. There are many options and metrics available to assess the quality of ensembles. For binary events (events that have only two outcomes), there is the Brier score (Brier, 1950). For categorical events (a forecast whose output is divided into multiple categories), there are the ranked probability score (Epstein, 1969) and the ignorance score (Roulston and Smith, 2002). For continuous outcomes, there is the continuous ranked probability score (Hersbach, 2000). The most applicable technique for this study is the Brier score and will be discussed further in the next chapter.

There are also graphical methods to depict the reliability of ensembles. One is the rank histogram proposed by (Hamill and Colucci, 1997). The rank histogram

can give some insight into whether the ensemble has a bias, is under or over dispersive, or is reliable in regards to a certain forecast variable, but it requires forecast information from each member. PEPs do not output data from each ensemble member, so the rank histogram is not applicable to this study. However, the reliability diagram is very applicable, and much information on the quality of ensemble forecasts can be gathered from it. A reliability diagram can depict the accuracy, resolution, skill and reliability of an ensemble (Hsu and Murphy, 1986). More on reliability diagrams and their application to this thesis will be given in next chapter.

There are several examples of the previously stated techniques being used to evaluate EPSs. In a study by Wang et al. (2012), he used the techniques to compare the performance of a small-scale regional ensemble to a global ensemble with more members but lower resolution. The comparisons made by Wang et. al. are analogous to comparisons between MEPS and GEPS, with MEPS having fewer ensemble members but higher resolution. Wang concluded that a regional high-resolution ensemble with fewer members can provide more skill for near-surface weather variables, like sea level pressure and wind, but may have less skill when applied to upper-air variables (Wang et al., 2012).

In another study, Eckel and Mass (2005) evaluated two versions of the Short-Range Ensemble Forecasting (SREF) model with Brier skill scores and reliability diagrams. One version of SREF used a multi-model technique where the ensemble members are completely unique models. Eckel and Mass compared that version to an SREF model using a varied-model technique, where the ensemble is comprised of variations on a single model, such as different physics packages. MEPS, for example, is a model that applies the varied-model technique by using different versions of the WRF. Eckel used the skill and reliability metrics to show

that the multi-model SREF had the best overall performance.

Clements (2014) was the first to work on validation of PEPs from GEPS and MEPS. Clements used the Brier Skill Score with climatology as a reference and reliability diagrams. He chose 4 locations in the southeast US and 5 locations from military bases overseas. He discovered that lightning for all of the EPS lightning was substantially overforecasted leading to low reliability and skill scores for lightning within 20 nm and within 20 km (Clements, 2014). For winds, MEPS4 had the highest skill scores in 4 of the 5 locations that had sufficient sample sizes. MEPS4 was also the most reliable with precipitation, while GEPS was the least reliable. However, GEPS had more reliable precipitation forecasts with tropical systems.

3. Methodology

3.1 Time period and Location Selection

This study examines 17 U.S. Army and Air Force bases (AFB) across a variety of locations within the continental United States (CONUS). The locations' names and International Civil Aviation Organization (ICAO) identifiers are as follows in Table 3.1:

Location (ICAO)	
Air Force Academy (KAFF)	Offutt AFB (KOFF)
Scott AFB (KBLV)	Vandenberg AFB (KVBG)
Davis-Monthan AFB (KDMA)	McGuire AFB (KWRI)
Vance AFB (KEND)	Eglin AFB (KVPS)
Robert Gray Army Airfield (KGRK)	Destin Executive Airport (KDTS)
Holloman AFB (KHMN)	Hurlburt Field (KHRT)
Langley AFB (KLFI)	Duke Field (KEGI)
Little Rock AFB (KLRF)	Bob Sikes (KCEW)
Nellis AFB (KLSV)	

Table 3.1. List of locations from which PEP bulletins will be evaluated

Locations are illustrated in Figure 3.1 and were intended to sample a geographically diverse set of locations within CONUS for which DOD forecasters frequently forecast. More specifically, several locations were chosen in the Florida to study how the EPS performs with sea-breeze thunderstorm development. The weather parameters of interest include but are not limited to high winds, precipitation, lightning low, visibility, and low cloud ceilings. Due to storage limitations, ensemble output is not archived at the 557 WW regularly. Therefore, the time span of this study was limited to a single season spanning April to October 2015, covering the active summer weather season.

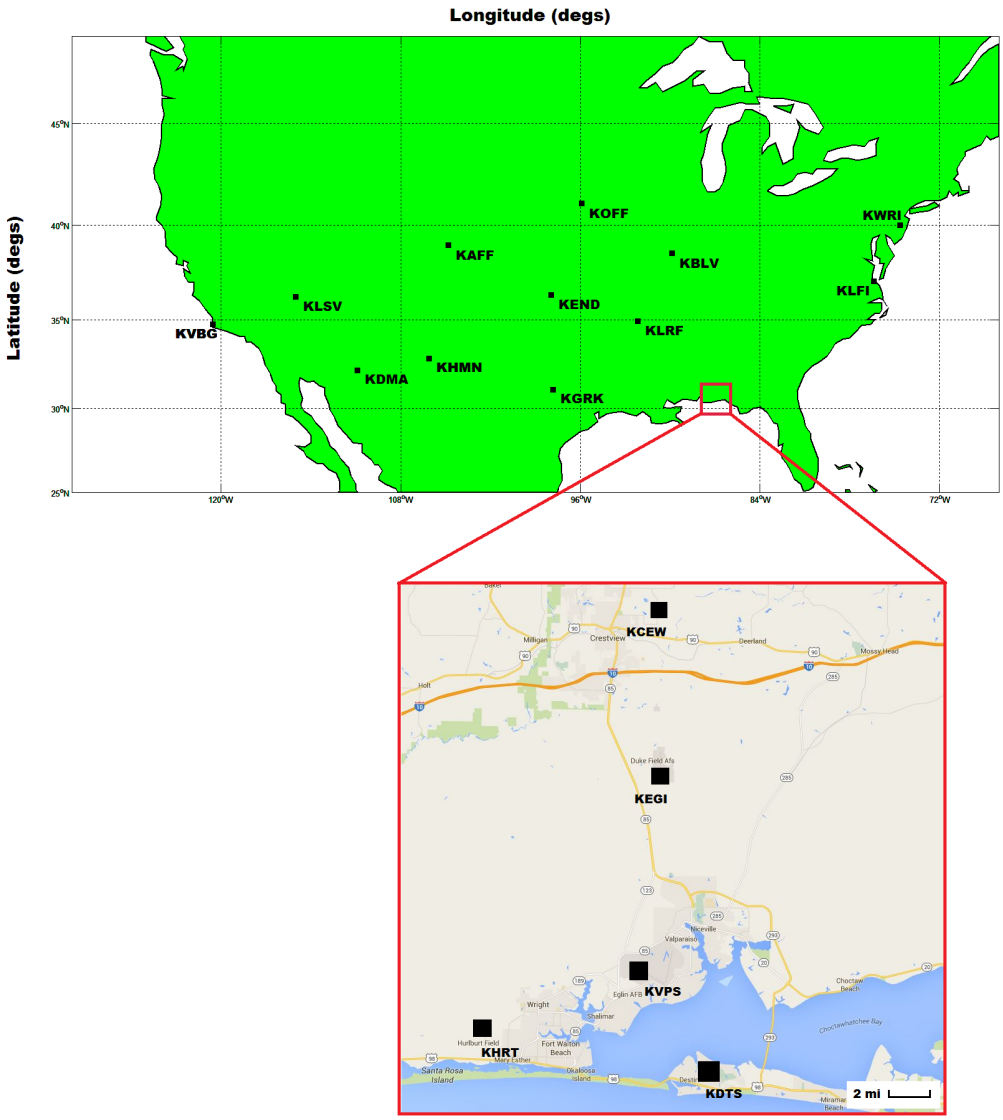


Figure 3.1. Map of selected locations labeled by ICAO.

3.2 Data Sources

3.2.1 Point Ensemble Probability Bulletins (PEPs).

PEP bulletins are HTML format files that display ensemble output in a simple format, shown in Figure 3.2. Each bulletin gives the name of the point location, the

type of the EPS, and the EPS run time in the upper left. The left column gives the criteria for operationally impactful weather. Each number to the right of the threshold gives the probability that the threshold will be exceeded. The probability within each column is valid from one minute after the previous forecast hour to the hour in the column heading. The color of the box (red, yellow, or green), highlights whether the probability reaches Air Force Weather's criteria for high, moderate, or low-risk potential. For GEPS that period is 6 hr, MEPS20 it is 3hr, and for MEPS4 it is 1 hr. PEP bulletins are disseminated daily for each EPS and location.

This research will validate the ensemble probabilities for the following GEPS forecast parameters: winds greater that 25, 35, and 50 kt, precipitation greater than 0.1 inches in 6 hr, precipitation greater than 2 inches in 12 hr, lightning within 20 km, visibility less than 5, 3, and 1 sm, and ceilings less than 3 kft, 1 kft and 500 ft. For MEPS, the same parameters validated, but there are a few differences in threshold values. For MEPS4, the lightning threshold is 20 nm as opposed to 20 km, and the 6 hr precipitation threshold is .05 in. The MEPS20 parameters are the same as GEPS. However, since the implementation of the rolling ensemble on July 27, 2015, the lightning threshold changed to 10 nm and the 6 hr precipitation threshold changed to .05 in. PEP bulletins were provided weekly by Evan Kuchera, the 557 WW 16th Weather Squadron Deputy Chief, Numerical Models Flight, Fine Scale Models and Ensembles Team Lead.

USAF ACADEMY AFLD — 38.9700 lat — -104.810 lon — 2003 meters elevation — THR 30

MODEL BOX INFO — 39.0089 lat — -104.756 lon — 2255 meters elevation

20 km MEPS 24/18Z — SAT 25 — SUN 26 — MON 27 — TUE 28 — WED 29

	SAT 25							SUN 26							MON 27							TUE 28							WED 29						
	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21
Surface winds																																								
Winds GT 25 KT	22	21	36	42	36	76	78	15	0	0	0	0	0	0	0	1	3	4	11	11	18	25	28	19	7	1	0	0	0	0	0	2	3	1	0	0	0	0	0	0	0	0
Winds GT 35 KT	0	0	0	1	0	10	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Winds GT 50 KT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Precipitation	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21
Precip GT 0.10 IN in 6 hr	13	1	0	0	0	0	0	7	34	17	0	0	0	0	11	21	62	60	66	71	63	60	53	45	34	30	14	1	0	0	3	11	14	14	1	0	0	0	0	0	0	0
Precip GT 2.0 IN in 12 hr	-99	-99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FZRA GT 0.01 IN in 3 hr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SN GT 0.1 IN in 6 hr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	20	40	58	90	99	90	16	0	0	0	0	0	13	10	0	0	0	0	0	0	0	0	0	0	0
SN GT 2.0 IN in 12 hr	-99	-99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	17	32	38	27	26	14	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SN GT 6.0 IN in 24 hr	-99	-99	-99	-99	-99	-99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	8	8	8	7	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Thunderstorms	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21
Lightning within 20 km	14	0	0	0	0	0	0	15	17	1	0	0	0	4	9	15	9	4	2	0	1	3	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0.75 IN Hail within 40 km	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ceiling and Visibility	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21
Visibility LT 5 miles	0	0	0	0	0	0	0	8	9	11	12	13	31	35	26	18	65	76	75	93	90	66	58	26	5	33	25	13	20	11	5	10	10	1	11	18	16	0	0	0	0	0
Visibility LT 3 miles	0	0	0	0	0	0	0	4	4	4	6	5	13	15	7	5	38	53	58	73	83	48	38	13	1	9	6	3	13	10	1	10	9	0	4	9	6	0	0	0	0	0
Visibility LT 1 mile	0	0	0	0	0	0	0	1	1	0	1	1	1	2	0	0	11	16	24	24	42	26	21	2	0	0	0	0	3	2	0	8	0	0	1	3	1	0	0	0	0	0
Ceiling LT 3000 feet	0	0	0	0	0	0	0	2	1	2	1	1	27	77	94	93	93	97	94	95	97	95	92	81	64	30	30	44	34	22	15	11	14	12	15	16	4	2	10	0	0	0
Ceiling LT 1000 feet	0	0	0	0	0	0	0	0	0	0	0	0	23	66	73	74	79	88	73	80	74	76	63	48	32	24	18	23	28	15	10	10	8	1	5	1	2	0	2	0	0	0
Ceiling LT 500 feet	0	0	0	0	0	0	0	0	0	0	0	0	6	14	17	16	19	22	17	16	14	14	10	7	4	4	3	5	5	4	3	3	2	0	1	0	1	0	0	0	0	0
Others	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21	0Z	03	06	09	12	15	18	21
Mean Temperature (F)	49	44	43	41	40	49	54	52	48	44	41	38	37	40	41	41	38	35	34	32	31	32	35	37	37	32	31	32	31	36	42	46	45	38	34	32	31	43	50	54	54	43
Mean Wind Speed (KT)	11	9	16	14	15	19	13	5	4	3	4	3	4	5	8	9	10	9	10	12	12	13	13	11	10	7	7	6	6	6	8	8	8	5	4	5	7	5	5	6	6	4
Prob GT 80% Cloud Cover	49	32	0	0	0	4	65	88	58	44	39	26	64	84	94	93	95	97	94	95	97	95	92	81	65	34	30	46	38	22	25	22	34	30	27	18	4	9	10	5	7	0
Pollution Trapping(1=LO,5=HI)	2	3	1	1	1	1	1	4	5	5	5	5	5	5	3	3	3	3	3	3	2	2	1	2	3	4	4	4	4	4	4	3	3	5	5	5	5	5	4	4	4	5
Station Adj ALSTG	2973		2981		2978		2980		2979		2981		2975		2977		2982		3000		3009		3025		3027		3034		3031		3032		3023		3028		3020		3022		3015		3018		3014		3018

Figure 3.2. Point Ensemble Probability Bulletin for KAFF from the 18z model run from MEPS20 on April 24, 2015. Each row represents a probability forecast for thresholds given on the left side of the chart. Each column represents the valid end time of the forecast, with the start time being one minute after the previous column's time. Note that the first forecast begins 6 hr after the model run time.

3.2.2 Weather Observations.

Ensemble output is compared with Aerodrome Routine Meteorological Reports (METARs) and Aerodrome Special Meteorological Reports (SPECIs). METARs are automated hourly weather observations formatted by the World Meteorological Office publication 306. SPECIs are disseminated when weather conditions change significantly in between METAR report times. As a result, it is possible for there to be multiple observations during a single forecast period. When that happens, the report used for validation is the one that represents the worst conditions (ex. highest winds, lowest visibility, lowest ceiling, etc.).

Additionally, data from the ENTLN will be used for lightning verification. The data includes the latitude, longitude, and time of every lightning strike within 20 nm of the site location from April to October 2015.

3.2.3 Climatology.

Climatology is based on a long history of observations and represents the frequency of occurrence of specified weather conditions. A forecast based purely on climatology requires no forecasting knowledge, allowing it to serve as baseline against which to measure PEP forecast skill. PEP forecasts will, therefore, be scored based on how much skill they have over climatology. A forecasting system has skill if it is correct more often than climatology. Climatology data for all 17 locations was provided by Mr. Jeff Zautner at the 14th Weather Squadron. The data spans from 2005 to 2014 and includes 1 hr, 2 hr, 3 hr, and 6 hr climatologies to maintain consistency with the forecast intervals of each EPS. For example, the percent frequency of winds greater than 35 kt from 00 UTC to 03 UTC in the month of April represents a 3 hr climatology for that period.

3.3 Implementation of Rolling Ensembles

An important change occurred to MEPS during this study. The change was effective on July 27, 2015, about halfway through the course of the study. Before the change, the MEPS initialized all members at the same time twice per day. Now, one member is initialized every 2 hr resulting in an ensemble of forecasts of different age. PEP data formats remained mostly the same for GEPS and the MEPS4, but there were significant changes made to MEPS20 data. MEPS20 forecasts changed from 3 hr forecasts to 2 hr forecasts. Additionally, the initial forecast times began 6 hr apart. Because of the differences, the datasets before and after the changes are incompatible with each other. MEPS20 data will, therefore, be split into two parts: one for the period from April to July 2015, and another starting in July after the changes were effective to October 2015. Additionally, the MEPS20 thresholds for 6 hour precipitation and lightning changed. Before, the MEPS20 threshold for precipitation was 0.01 inches in 6 hr and the range for lightning was 20 km. Now, the MEPS20 threshold for precipitation is 0.05 inches in 6 hr and the lightning range is 10 nm. These changes were accounted for in the analysis of data before and after the implementation of rolling ensembles on 27 July, 2015. MEPS4 and GEPS, on the other hand, maintained consistency and will blend over the entire period from April to October.

3.4 Validation

3.4.1 Software Utilization.

Following Clements (2014) MATLAB programs extracted PEP and METAR data for each ICAO and converted the data into text files to facilitate statistical analysis. For PEPs, each text file contained a table with columns for the month,

day, forecast hour, and each parameter category. Each row contained the forecast probabilities for each category with the corresponding date and time. Each PEP text file contained the data from one PEP file like the one shown in Figure 3.2. METARs, SPECIs, and lightning data sets were translated into monthly text files that contained columns for month, day, hour, and each parameter category. The rows for the observation text file contained a zero if the event was not observed and a one if the event was observed during the corresponding time. With all of the data translated into text files, a MATLAB program scanned the text files of a given ICAO and EPS for matching forecast periods and observation hours. Then, from the corresponding forecast probabilities and binary data, MATLAB calculated the scores and valuation metrics to be discussed in the following sections. For more details and illustration on this process, see Clements (2014). The process from this study is the same as in 2013 with the exception of section 3.3.4 in Clements's thesis regarding lightning verification. This study will use the national lightning detection network as a source of for lightning validation, rather than METARs, so the techniques used to modify probabilities as described in section 3.3.4 in Clement's thesis will not be necessary.

3.4.2 Observed Frequency vs. Probability: A Measure of Reliability.

A goal of probabilistic forecasting is for the forecast probability to match the actual frequency of occurrence. For instance, over all the times an EPS forecasts a 70% probability, the event should actually occur 70% of the time. The frequency of occurrence is the ratio of the number of occurrences to how many times the event was forecasted. A frequency of occurrence that exactly matches the ensemble probability is considered a perfectly reliable EPS forecast. This study calculates the frequency of occurrence from METAR and SPECI using the following equation:

$$P(y_i) = \frac{N_i}{n_i} \tag{2}$$

$P(y_i)$ is the observed frequency of a particular forecast value or range denoted as y_i. N_i is the number of actual occurrences of the event, and n is the number of forecasts (Wilks, 2011). In this study there are 11 possible forecast ranges for y_i. The first sub-sample is made of the 0 percent forecasts. The next sub-samples contain the 1 to 10 percent forecasts, followed by the 11 to 20 percent forecasts and so on to 100 percent. For example, if $y_i = 40 - 50\%$ then n would be the number of times that an EPS forecasted 40-50 percent, and N_i would be the number of times the event actually occurred when the EPS forecasted 40-50 percent. The observed frequencies are plotted on a reliability diagram, giving an objective assessment of the reliability of an EPS.

3.4.3 Brier Score.

The Brier score is a commonly used evaluation of error in probabilistic forecasts (Brier, 1950). It is applicable in dichotomous scenarios where an event either occurs or does not occur (Jolliffe and Stephenson, 2012). The Brier score simply averages the squared differences between the forecast probability and the corresponding binary outcomes. The Brier score is defined as:

$$Brier\ Score = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2 \tag{3}$$

f_i is the forecast probability for the ith forecast, and o_i is the ith outcome where $o_i = 1$ if the event occurs and $o_i = 0$ if it does not. In this study f_i is gathered from the PEP bulletin forecasts and o_i is decoded from the METARs, SPECIs, and lightning data. A perfect EPS would forecast 100% probability for every occurrence

and 0% for every non-occurrence, resulting in a Brier Score of 0. A perfectly incorrect EPS would do the opposite, resulting in a Brier score of 1. Therefore, the lower the Brier score is, the less error in the EPS.

Since the Brier Score is quadratic, it can be decomposed into the sum of three terms: reliability, resolution, and uncertainty. Murphy (1973) demonstrated how this decomposition is done, and how each of the three terms represents a unique measure of the EPS quality. The decomposed Brier score is represented by the following equation:

$$Brier\ Score = \underbrace{\frac{1}{n}\sum_{i=1}^{I} N_i \left(f_i - \bar{o}_i\right)^2}_{Reliability} - \underbrace{\frac{1}{n}\sum_{i=1}^{I} N_i \left(\bar{o}_i - \bar{o}\right)^2}_{Resolution} + \underbrace{\bar{o}\left(1 - \bar{o}\right)}_{Uncertainty} \tag{4}$$

Here $\bar{o}_i$ is observed frequency of the event when the forecast probability is f_i and $\bar{o}$ is the climatological frequency. The first term, reliability, summarizes any conditional biases in the forecasts (Wilks, 2011). Forecasts that are perfectly reliable will match the observed frequency, resulting in a reliability of zero. Reliability alone, however, is not sufficient for a useful EPS. For example, imagine a case where the EPS always forecasts the same probability as the climatological frequency. Over an extended period with invariable climatology, the ensemble forecasts would, in theory, match the observed frequency. The EPS may be entirely reliable but does not provide any valuable information over climatology (Toth et al., 2003).

In addition to reliability, a useful EPS also needs to be able to predict situations that lead to observed frequencies that may be higher or lower than climatology. Resolution measures how much the observed frequencies differ from climatology. It represents the ability to discriminate in advance between situations that lead to variable observed frequencies. Since the resolution term is subtracted in the Brier score equation, increasing resolution improves the score of the EPS.

The last term, uncertainty, is independent of the EPS forecast. It measures the climatological variation in the event occurrence (Ferro and Fricker, 2012). If an event is very rare or very common, then the uncertainty is low. However, if the event occurs 50 percent of the time, then we have the maximum amount of uncertainty. The Brier score, in theory, should improve with decreasing uncertainty.

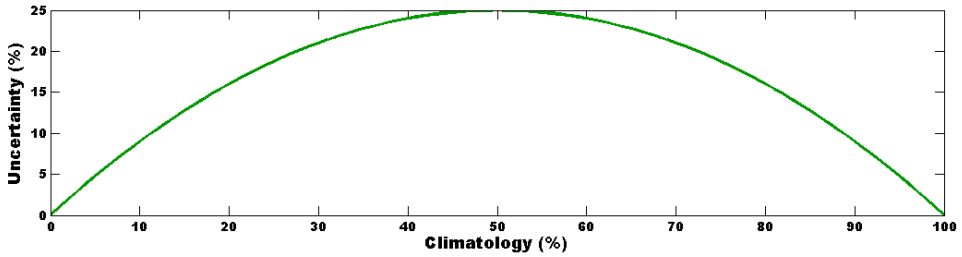


Figure 3.3. Graph of uncertainty for given values of climatology. Adapted from Clements (2014).

3.4.4 Brier Skill Score.

The Brier score by itself is not a complete measure of accuracy because there is no control forecast to compare with. The Brier Skill Score (BSS) makes that comparison possible. The BSS is defined by the following equation (Wilks, 2011):

$$BSS = \frac{BS - BS_{ref}}{0 - BS_{ref}} = 1 - \frac{BS}{BS_{ref}} \tag{5}$$

This score represents the level of improvement to the Brier score over a reference forecast strategy (Mason, 2004). The most commonly used approach for the reference forecast. In this case, BSS indicates the EPS's value over climatology. A perfect BSS is 1. A BSS of 0 indicates no skill over the reference forecast, and a negative BSS indicates less skill than the reference forecast.

Combining the BSS equation with the decomposition of the Brier score, the BSS

reduces to (Wilks, 2011):

$$BSS = \frac{Resolution - Reliability}{Uncertainty} \tag{6}$$

This equation shows how the BSS can be expressed in terms of reliability, resolution, and uncertainty. An implication here is that the BSS will be positive when the resolution of the EPS is greater than the reliability. Equation 6 provides some insight into the BSS in terms of reliability, resolution, and uncertainty. However, equation 6 is not used to calculate BSS since there are some inherent biases in the three terms that will be explained in the next section. Instead, the BSS is computed using equation 5.

3.4.5 The Reliability Diagram.

The reliability diagram offers a practical illustration of the quality of probabilistic forecasts. In addition to reliability, the diagrams can also reveal information on the skill and resolution of the EPS or whether the EPS has a bias (Hsu and Murphy, 1986). An example reliability diagram is given in Figure 3.4. The horizontal axis is the forecast probability and the vertical axis is the observed frequency. The forecast probabilities are separated into bins of width 10 percent except for the zero percent bin. The number of forecasts made in each bin is shown on the vertical axes of the box below the diagram. The green dots show the observed frequency for each bin. The diagonal dashed line is the zero (perfect) reliability line. Any green dots that lie on the zero reliability line indicate that the observed frequency matches the forecast probability. The horizontal dashed line is the climatological frequency of the event. The red shaded area forms the area of positive skill. It is bouded by the vertical line that intersects climatology and the zero reliability line, and a line bisecting the angle between climatology and zero

reliability. This line is where resolution equals reliability (Hsu and Murphy, 1986). Green points within the shaded area indicate areas of positive skill. It is possible to create a diagram for each forecast hour, so an enormous number of diagrams can be made. For instance, given the 240 hr forecast at 6 hr intervals for GEPS, we could make 40 diagrams per parameter, per location.

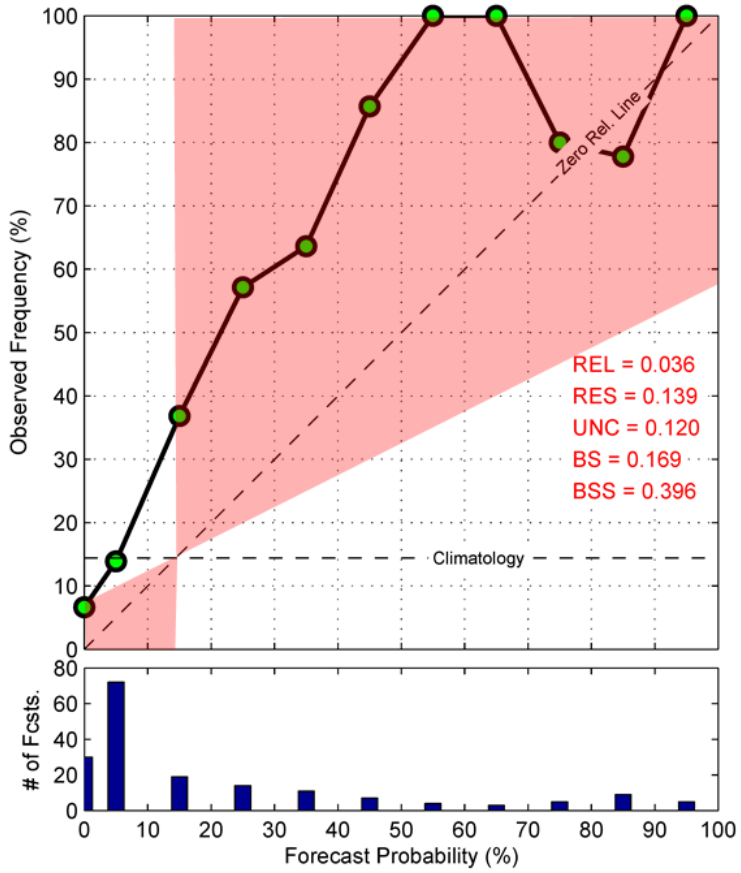


Figure 3.4. An example reliability diagram. The horizontal axis shows forecast probability. All forecast probabilities from the EPS are separated into bins with the number of forecasts in each bin shown in the box below the diagram. For example, the 0% percent bin is on the far left with about 30 forecasts, followed by the 1-10% bin with nearly 70 forecasts. On the vertical axis is the observed frequency, indicating how often the event occurred out of the times the EPS forecasted within a given forecast bin. The horizontal dashed line indicates how often the event occurs according to climatology. The 45 degree dashed line is the zero-reliability line that indicates the point where observed frequencies match the forecast probability. The red shaded area shows the area of positive skill. The BSS, Brier score and the three decomposition scores are shown in red on the right.

Consistently large deviations from the zero reliability line may indicate a forecasting bias. If the observed frequencies lie above the zero reliability line, the EPS may have an underforecasting bias which means that observed frequencies are consistently higher than the forecast. If the observed frequencies lie below the zero reliability line, then the EPS may have an overforecasting bias. When the forecasting bias is consistent for an EPS, it may be corrected by a re-calibration.

The resolution component of the Brier score for a forecast may be indicated by how far the green dots deviate from climatology. Resolution is large when the EPS can recognize and discern between events that occur more or less frequently than the climatological frequency (Toth et al., 2003). If the observed frequencies stay relatively close the climatology line, the EPS has poor resolution. Resolution increases as the EPS makes forecasts that are further from climatology which increases the BSS.

3.4.6 Limitations and Sources of Uncertainty.

The data and the methods used in this thesis have possible sources of error and uncertainty. For each EPS, the location of the nearest model grid box center does not exactly match the location of the observing site. For most parameters this is not an issue because the forecast is the same for the entire grid box that covers both points. However, the difference in locations may cause some misrepresentations of the lightning forecasts that involve multiple grid boxes. The range ring (10nm, 20 km, or 20nm depending on the EPS) that surrounds the grid center may not exactly match the range rings that surround the actual location. Lightning occurrences were based on the distance of the lightning strike from the actual location as opposed to the location of the model grid center. Therefore, it is possible for lightning to be coded as "occurred" even though the strike happened outside the

range ring surrounding model grid box center. In most cases these kinds of errors are likely to be very infrequent and would only effect the higher resolution MEPS4.

Additionally, the BSS shown in Eq. 6 does not necessarily match the BSS in Eq. 5. The differences are introduced by biases inherent in the reliability and resolution terms as shown in Eq. 4. Stephenson (2008) explained that the bias results from binning the forecast probabilities. Eq. 4 does not account for "within-bin variations" that cause variance and covariance between forecasts and observations. According to Brocker, the true reliability tends to be higher than what is given in Eq. 4, and the true resolution tends to be lower regardless of the sample size (2012). Therefore, the reliability and resolution curves shown in the BSS plots are not to be considered true measures. Instead, the terms are graphed to show trends, to show the relative magnitude of each term, and to compare with the previous study by Clements (2014).

Some caution also needs to be taken in the interpretation of the BSS from Eq. 5 as well. The BSS may be zero or negative in some cases, but this does not necessarily indicate that the forecast has no value compared to climatology. A forecast may actually contain some useful information even if the BSS is zero or negative (Mason, 2004). Mason recommended that the BSS alone should not be used as a measure of forecast skill over climatology. If the EPS has some resolution (ability to forecast events with more/less frequency than climatology), then all that may be needed is some calibration to make the BSS positive. In this study, plots of reliability, resolution, and uncertainty offer the needed additional information.

4. Results

4.1 Skill and Reliability Overview

The format of reliability diagrams and skill plots/tables given in this chapter remain consistent with those given in the previous 2014 study by Brad Clements. Skill plots for each parameter contain two charts. The top chart shows the BSS versus forecast hour and the bottom chart shows the value of the reliability, resolution, and uncertainty terms versus forecast hour. The utility of the BSS plots and reliability depends on how frequently the event occurs. If the event almost never happens, then BSS scores and reliability may be erratic providing no valuable information (Clements, 2014). For example, Figure 4.1 shows the behavior of the BSS for the GEPS forecast of ceilings less than 500ft at KLSV, which almost never occurs at KLSV. As in the previous study, the results for the parameters precipitation greater than 2.0in 12 hr and winds greater than 50 kts are left out because they do not occur frequently enough.

There are too many charts to potentially present in this chapter, therefore tables are presented to summarize the data for each parameter. Each table contains the mean positive BSS and the fraction of the forecast that had positive BSS. The mean positive BSS is calculated by taking the average of the BSS scores which were positive. This number is used because sometimes the BSS for some forecast hours fell to abnormally large negative values, likely due to limited data points used in BSS calculation. In the interest of preventing these outliers from skewing results, the mean is taken over only positive BSS. The percentage of the forecast with positive skill supplements the mean positive BSS. Percent positive skill is calculated by dividing the number of forecast hours that had positive skill by the total number of forecast hours in the forecast period. It indicates how much of the total forecast

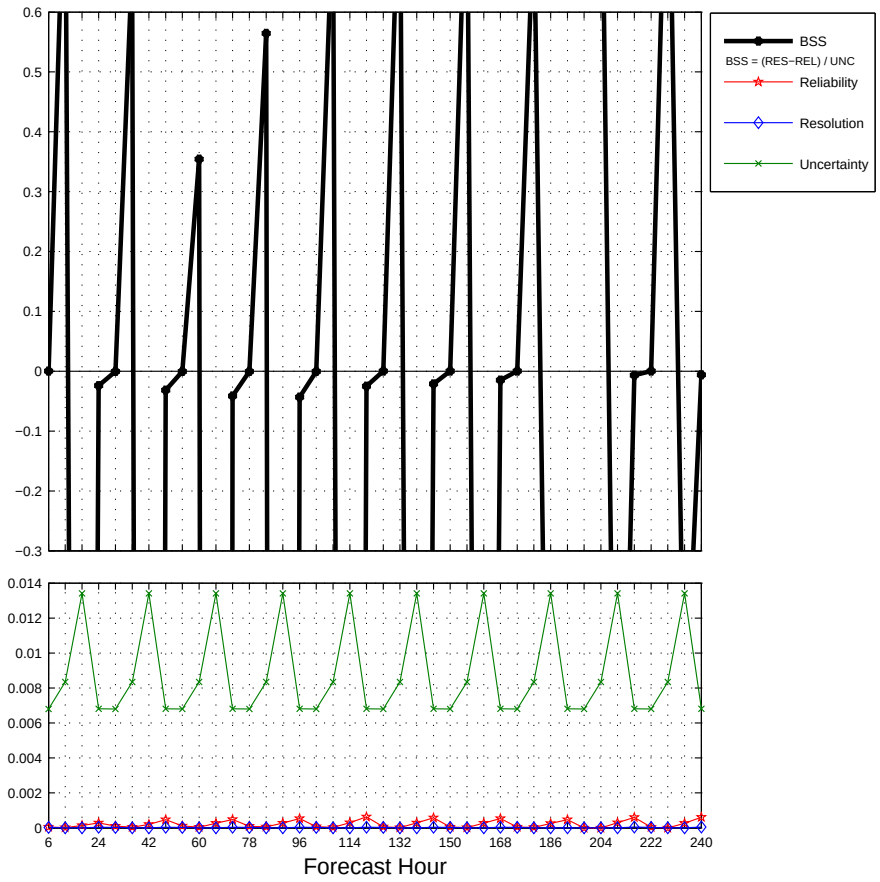


Figure 4.1. KLSV GEPS BSS for ceilings less than 500 ft from Apr-Oct 2015. The BSS trends are very volatile due to a rarity of forecasts or observations.

had positive BSS. Both numbers together summarize performance over the entirety of the forecast. The second reason for reporting these numbers is to maintain continuity with Clements (2014).

An initial overview of the data reveals some consistent trends to be explained further in later sections. Ceilings of each category tended to be underforecasteded by each EPS. The degree of underforecasting depended on the location and time of day. Generally, the skill of ceilings forecasts from MEPS improved considerably over GEPS. The same generalizations were true for visibility, however EPS skill did not have a significant dependence on the horizontal resolution EPS. Lower categories of visibility performed worse and suffered from the most underforecasting bias.

Lightning skill and reliability fell during the evening and improved during the day in a similar fashion to what was shown in Clements' 2014 study. Lightning skill generally increased with increasing resolution of the EPS, but skill sometimes jumped to large negative or relatively small values. Wind forecasts also improved with increasing resolution, with GEPS being prone to miss events especially in areas with terrain. For 6 hr precipitation, skill generally remained positive and improvements to average skill primarily depended on location rather than the type of EPS.

4.2 Ceilings

As demonstrated in Table 4.1, the MEPS 4 km and MEPS 20 km were able to maintain positive skill for the majority, if not all, of the total forecast period for ceilings less than 3kft. The average positive skill scores are given for the first 72 hr of forecasts from each EPS for comparison. The 72 hr average positive skill showed that skill scores tended to improve with increasing model resolution. There is one exception at KOFF, where GEPS matched MEPS 20 km skill. At KLRF, GEPS skill exceeded the MEPS 20 km skill for the first 72 hr. This result is likely because, for KLRF AFB, the MEPS 20 km model run began at a different time from GEPS causing the forecasts to cover different times of the day. GEPS positive skill percent dropped significantly from MEPS, with the highest positive skill percentage being just 58.4% at KEND. None of the EPSs were able to produce any positive skill at KVBG. KVBG is an exceptional case from the other locations because it is prone to frequent marine fog and stratus events that are difficult to forecast and will be discussed in detail later.

One thing is clear from the reliability diagrams of ceilings for all locations: each EPS has a tendency to underforecast the probability of ceilings less the 3kft, 1kft

and 500ft thresholds. The first example is shown in Figure 4.2 of the 8hr MEPS4 forecast of ceilings less than 3kft at KGRK. The observed frequencies for each forecast bin are clearly above the zero reliability line except for the 71-80% bin. There were 67 forecasts between 1-10%, the highest amount of all bins. The observed frequency for the 1-10% was 31.3%, a difference of over 21% from the forecast probability. The deficit increases by over 45% for the next 11-20% forecast bin, with the observed frequency at 67%. For the subsequent bins, the underforecasting continues as the observed frequency increases to 100%.

The underforecasting is also evident from the observed frequency in the 0% bin. In this bin, it would be most accurate for the observed frequency be zero. However, it is 8.7% which means that some of the MEPS4 8hr forecasts missed the event completely. There are 46 forecasts in the 0% bin. Therefore, MEPS4 missed 4 events.

This 8hr forecast is not the only example of underforecasting 3kft ceilings at Robert Gray. By averaging the observed frequencies for every forecast (1-72h) we find that the average observed frequencies for the 0% to 91-100% bins are respectively; 10.9%, 21.8%, 56.5%, 78.0%, 89.9% 93.0%, 96.7%, 97.8% 99.1% and 97.7%. These averages clearly show that MEPS4 underforecasts ceilings less than 3kft at KGRK.

Despite the underforecasting, the MEPS4 forecast of 3kft ceilings at Robert Gray is actually not bad compared to climatology. The BSS in Figure 4.2 is positive at .384. The forecast may look unskillful from the reliability diagram, but it has some skill over climatology. Climatology is about 6% so the uncertainty is low which contributes to higher skill values. Also, if we draw the area of positive skill on Figure 4.2 as shown in Figure 3.4, we see all of the bins greater than the 10-20% percent bin lie in the area of positive skill. Each bin contributes to the total BSS in

Site	EPS	Avg Positive Skill (0h-72h)	Avg Positive Skill (Total)	% Positive Skill (0h-72h)
KAFF	GEPS	.106	.128	50.0
	MEPS20	.101	.109	86.9
	MEPS4	.263	.263	100
KBLV	GEPS	.152	.228	50.0
	MEPS20	.198	.241	95.6
	MEPS4	.386	.386	100
KEND	GEPS	.139	.218	58.4
	MEPS20	.361	.415	100
	MEPS4	.366	.366	100
KGRK	GEPS	.165	.227	50.0
	MEPS20	.172	.186	100
	MEPS4	.403	.403	100
KLFI	GEPS	.0271	.0271	25.0
	MEPS20	.263	.296	100
	MEPS4	.420	.420	100
KLRF	GEPS	.167	.223	50.0
	MEPS20	.171	.193	78.3
	MEPS4	.354	.353	100
KOFF	GEPS	.118	.167	50.0
	MEPS20	.185	.168	100
	MEPS4	.290	.290	100
KVBG	GEPS	0	0	0
	MEPS20	0	0	0
	MEPS4	0	0	0
KWRI	GEPS	.0767	.101	50.0
	MEPS20	.175	.206	85.1
	MEPS4	.393	.393	100
KDMA	GEPS	.0530	.0858	100
	MEPS20	.121	.182	43.5
	MEPS4	.0847	.0847	52.2
KHMN	GEPS	.0491	.0891	56.5
	MEPS20	.172	.221	42.5
	MEPS4	.139	.139	95.5
KLSV	GEPS	.0361	.0526	80.6
	MEPS20	-	-	-
	MEPS4	.113	.113	80.5

Table 4.1. Ceilings less than 3 kft BSS average for both the first 72 hr and the whole forecast period, with 72 hr percent positive skill.

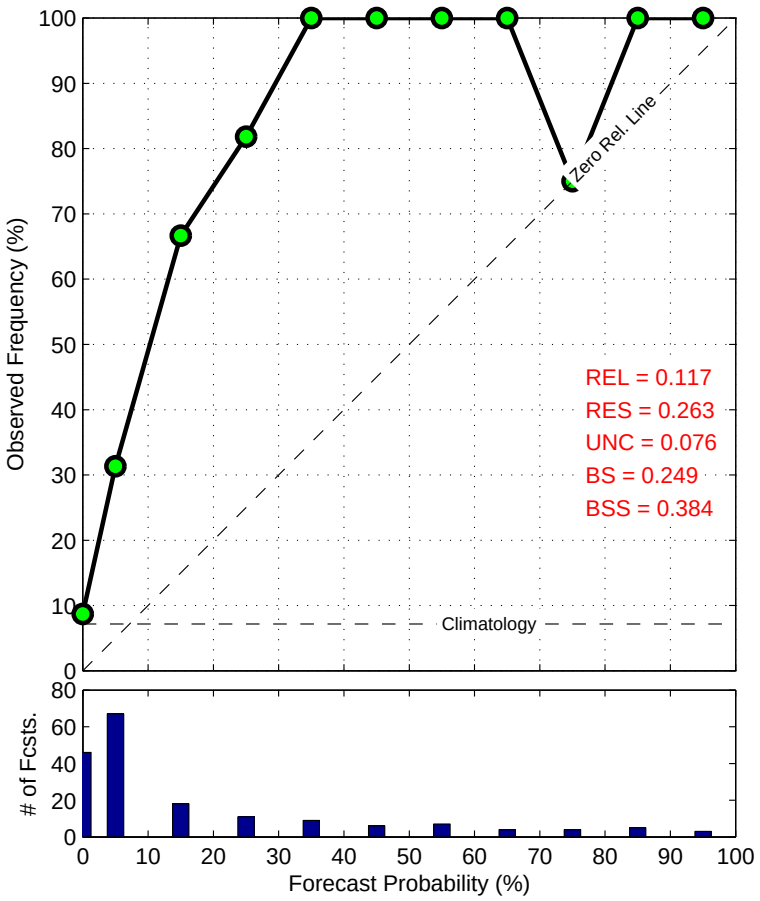


Figure 4.2. KGRK MEPS4 8 hr reliability diagram for ceilings less than 3 kft indicating underforecasting.

an amount proportional to the number of forecasts. The 20% and higher bins have a smaller number of forecasts; however when added together they account for about 50% of the total number of forecasts. Therefore, the 20-100% bins contribute significantly to the total skill.

Another example shown in Figure 4.3, is a reliability diagram for the MEPS20 33h forecast of 3kft ceilings at KBLV from April-Jul. This additional example shows that MEPS20 also has an underforecasting bias 33 hr into the forecast. However, in this example, the underforecasting is not quite as severe. For the first 5 forecast bins, the observed frequencies exceed the forecast probabilities by no more than 25%. The 0% probability bin indicates that MEPS20 missed 4 events, and the last

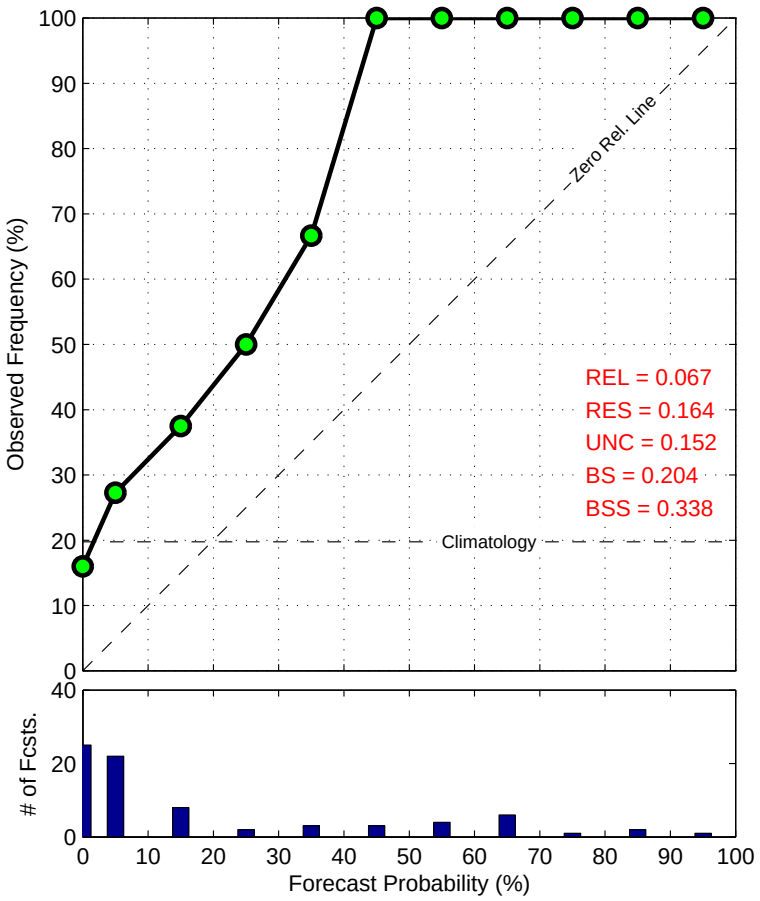


Figure 4.3. KBLV MEPS20 33 hr reliability diagram for ceilings less than 3 kft

six bins all have an observed frequency of 100%. Averaging the observed frequencies over the entire 132 hr forecast reveals an overall underforecasting bias again. The average observed frequencies for KBLV over the entire periods yields the following for bins 0% to 91-100% respectively; 14.2%, 30.8%, 51.5%, 60.6%, 72.2%, 76.5%, 85.0%, 79.4%, 79.3%, 92.7%, 94.1%. The most significant underforecasting occurred in the lower 4 bins that were also the most populated bins for most forecast hours.

Not every reliability diagram indicated underforecasting of ceilings. There are several examples where the reliability improved. One instance is shown in Figure 4.4 of the GEPS 18hr forecast of 3kft ceilings at KWRI from April to October. For the first 5 bins (0%-40%), the observed frequencies were very close to the zero reliability line. The last two bins had 100% observed frequencies, yet only accounted for 7% of

the total number of forecasts from the GEPS 18hr forecast. The reliability for Figure 4.4 is better compared to Figure 4.2 and 4.3 at .019. The better reliability score leads to a better Brier score; however the BSS is lower than the previous examples. The lower BSS is attributed to the fact that the 6-hr climatology is much higher at KWRI, resulting in more uncertainty that also decreases resolution.

The GEPS 18hr forecast for ceilings is just one example and does not reflect the entire GEPS forecast for KWRI. Again, if we average the observed frequencies for the much longer 240 hr forecast we get the following results for the 0% to 51-70% bins (No forecasts greater than 60% occurred) respectively: 1.6%, 20.6%, 45.5%, 63.8%, 75.1%, 81.5%, 93.0% and 100%. So for GEPS at KWRI, there was still an overall underforecasting trend as well.

Although the underforecasting trend seems consistent for each EPS, the degree to which the to which the EPS underforecasted fluctuated depending on the location. For some places, each EPS showed some consistent diurnal trends of forecast skill and reliability. The first example is illustrated in Figure 4.5 of MEPS4 forecasts of ceilings less than 1kft at KGRK. Each diagram was taken from a different time of the day. Figure 4.5(a) is from the 22 hr forecast, which would be 10Z or 0600 local time. 4.5(b) is the 28 hr forecast or noon local time. The underforecasting is clearly less significant in the 22 hr forecast at 0700 local time when the event is climatologically more frequent. The reliability for the 22 hr forecast was better and uncertainty was higher. The decline in uncertainty along with a reduction in reliability and skill in the 28 hr forecast indicate that MEPS4 is worse at forecasting 1kft ceilings when the event becomes climatologically less frequent.

The diurnal trend in skill and reliability further illustrates itself in Figure 4.6 of MEPS4 ceilings less than 3kft at Scott AFB. The top plot shows the BSS versus

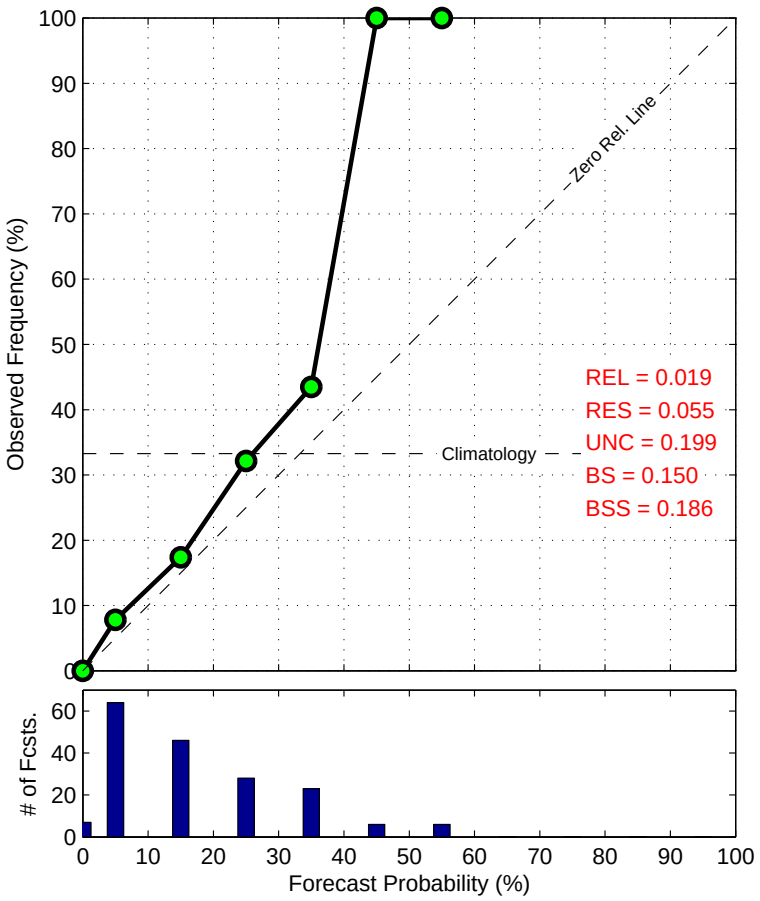


Figure 4.4. KWRI GEPS 18 hr reliability diagram for ceilings less than 3 kft

forecast hour, and the bottom plots show resolution, reliability, and uncertainty as calculated in Eq. 4 versus forecast hour. A diurnal trend is evident from the plot of BSS. The BSS peaks at the 22 and 46 hr forecasts, and bottoms near the 15, 37, and 61 hr forecasts. In local time, the peaks occur around 0300 and troughs at 2100 local time. Climatology shows that ceilings less than 3kft occur more frequently near 0300 and the least frequently during 2100 local time, so it is likely that trends in BSS and climatology are related. The trend is also evident within the lower plot of reliability resolution and uncertainty. As expected, the Murphy reliability value goes down as the skill increases and the opposite when skill decreases.

Diurnal trends did not exist at all locations. At KLFI, the MEPS4 BSS for ceilings less than 3kft remained relatively steady as shown in Figure 4.7. Instead,

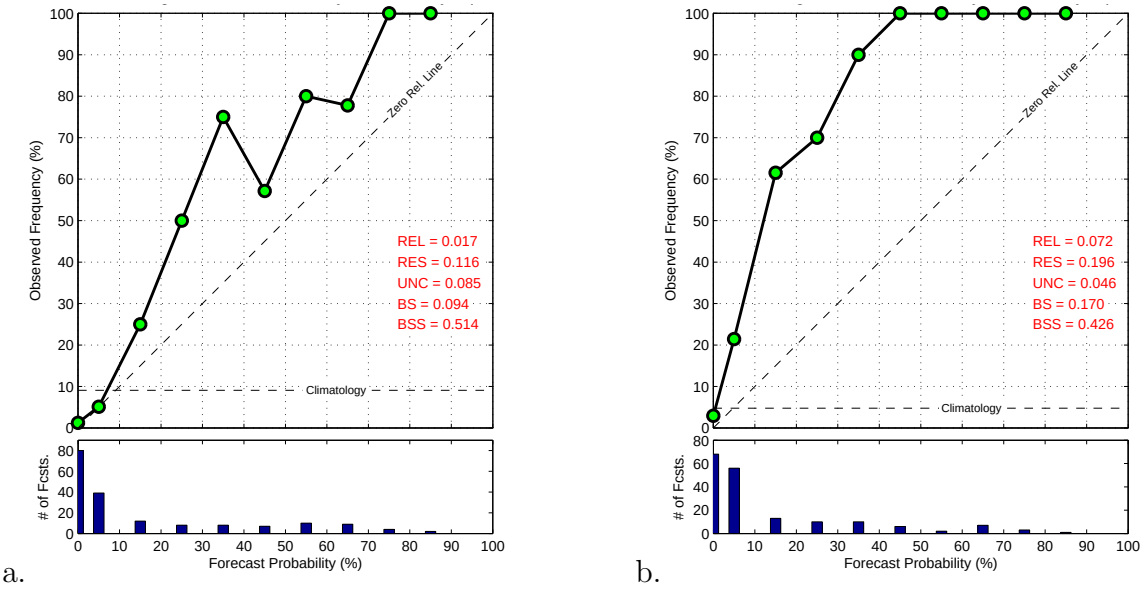


Figure 4.5. KGRK MEPS4 reliability diagram comparison for ceilings less than 1kft: a: 22 hr forecast b: 28 hr forecast

the plot indicated a gradual increase in skill through the whole forecast period with no clear evidence of skill or reliability following a diurnal trend. Interestingly the lowest BSSs appear near the beginning of the forecast and the highest BSSs appear near the end forecast. Typically, skill decreases as the length of the forecast increases while the decline in skill is due to the growth of internal and external error. It is possible that for MEPS4, the error growth is not significant over a 72 hr period.

The day to night fluctuation in skill and reliability may be caused by issues modeling low ceilings associated with convection. Or perhaps the EPS is not accurately modeling the diurnal variations in the boundary layer that can effect the ceiling height. It is also possible that skill increased because of a larger number of forecasts and observed occurrences, where as forecast periods with different amounts of forecasts may reflect differently on EPS skill.

Figure 4.8 shows a direct comparison of skill between MEPS4, MEPS20, and GEPS forecasts of 3kft ceilings at KOFF. MEPS4 and 20 km maintain positive skill while GEPS fluctuates diurnally between positive and negative skill values. Skill

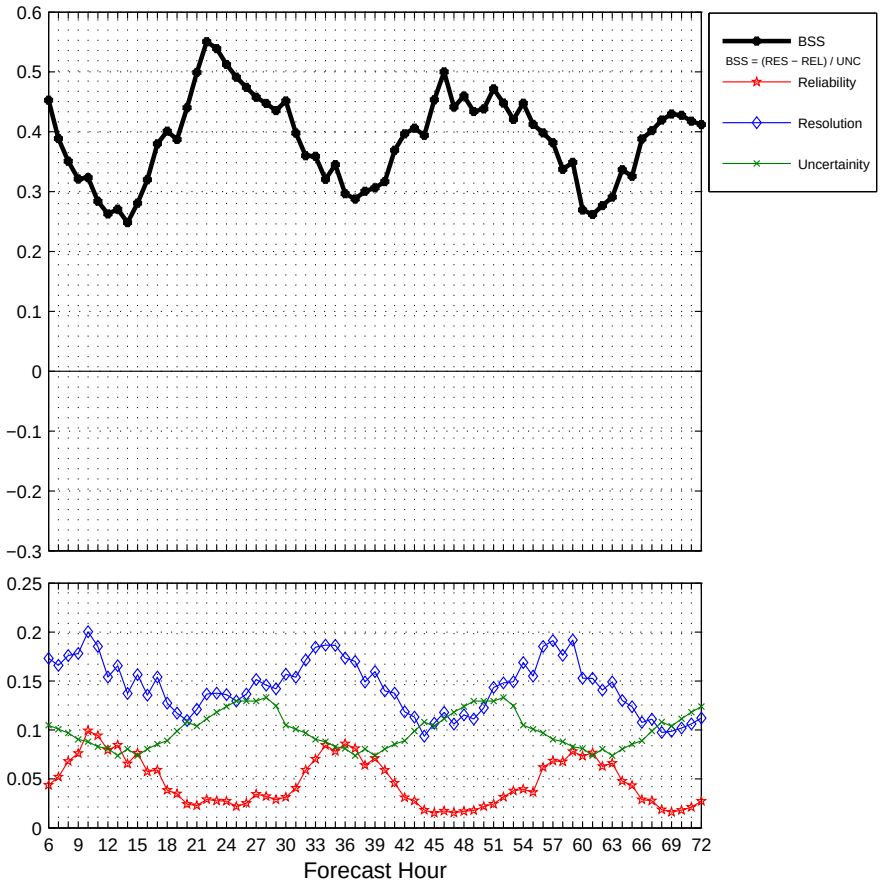


Figure 4.6. KBLV MEPS4 BSS for ceilings less than 3 kft from Apr-Oct 2015

clearly improves with increasing model resolution. Mean skill for MEPS remains consistent while GEPS mean skill decreases over time. This phenomena is likely to due to the nature of error growth in the global ensembles. Recall that the differences between ensemble members in GEPS are from perturbations to the initial conditions. This process introduces internal error which grows exponentially according to (Reynolds et al., 1994). Since GEPS forecasts longer into the future, we see the exponential error growth having negative impact on skill during forecast hours with longer lead times.

The lower thresholds of 1kft and 500ft ceilings do not occur as often, so reliability diagrams do not clearly illustrate forecast biases but trends in reliability and skill for the lower thresholds were similar to the 3kft threshold patterns.

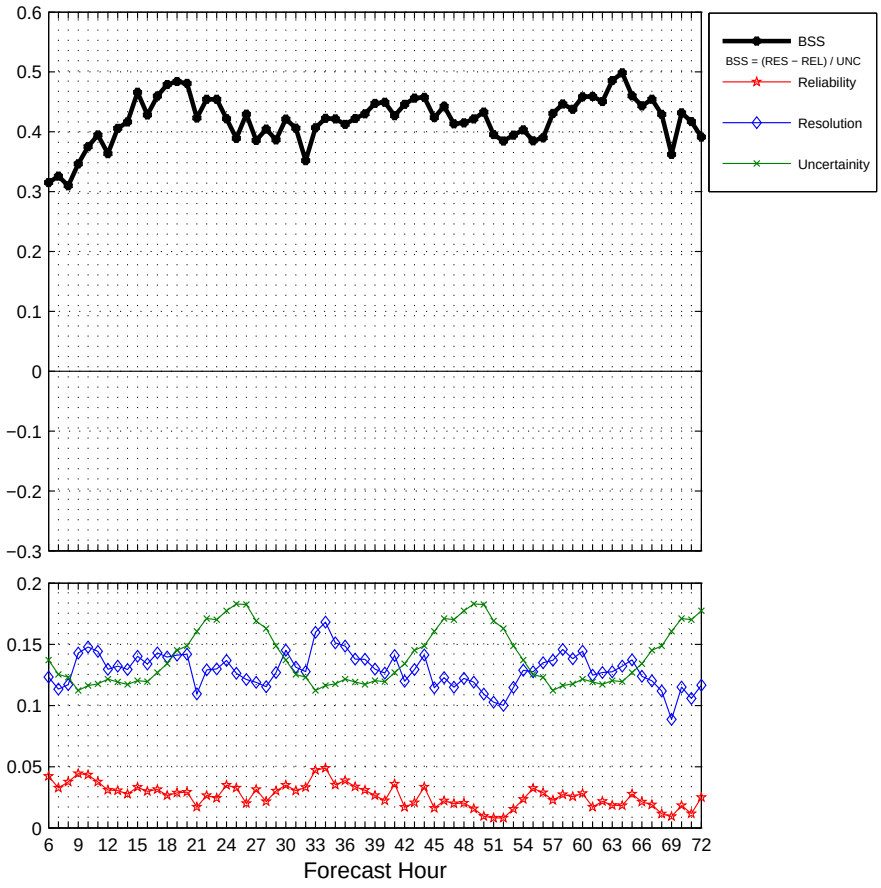


Figure 4.7. KLFI MEPS4 BSS for ceilings less than 3 kft from Apr-Oct 2015

However, the average BSS of the 500ft and 1kft ceiling categories were much lower than 3 kft BSS. What was clear from the reliability diagrams for 500ft and 1 kft ceilings was that most of the forecast probabilities were in the 0-10 bin, with a small number of forecasts in the higher probability bins. Also the observed frequencies were typically higher for the majority if the forecast period. Therefore, the underforecasting bias affected the lower threshold ceiling categories as well.

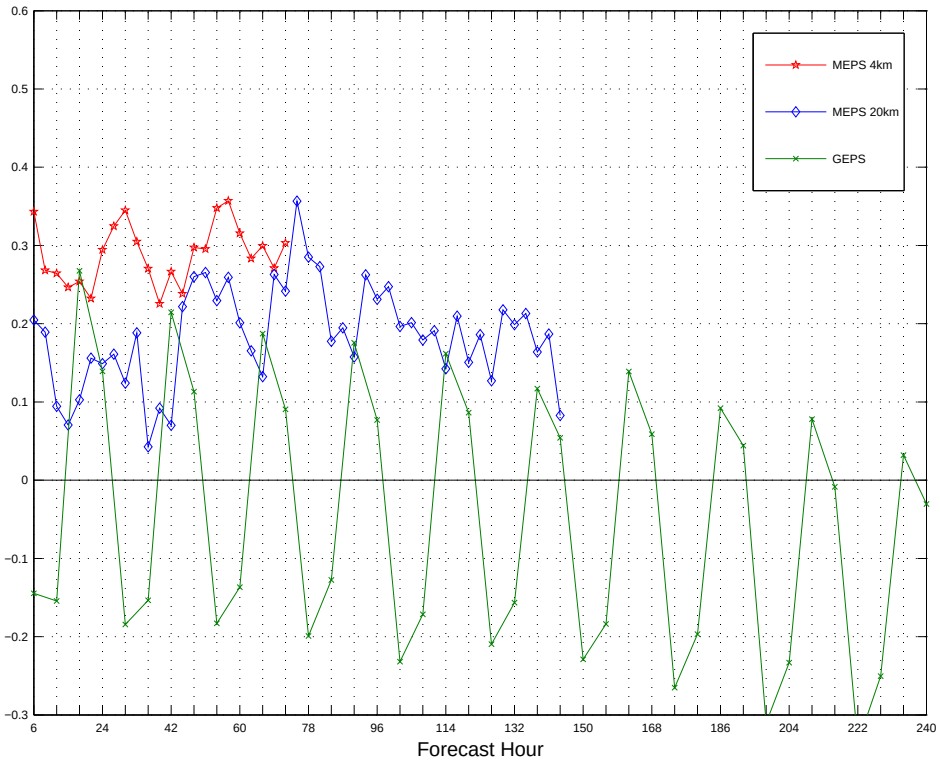


Figure 4.8. KOFF BSS model comparison for ceilings less than 3 kft

4.2.1 Vandenberg AFB.

The EPS forecast trends for ceilings at KVBG were unique from all other cases. Fog and low stratus conditions at KVBG often result from the onshore movement of condensed water from the Pacific Ocean. The development and duration of low clouds at KVBG is highly dependent on the height of the marine layer inversion, moisture content, microphysical processes, and pressure gradients between coastal waters and land. These small scale processes and the challenges involved with modeling the boundary layer make model forecasts more unreliable, especially for probability forecasts produced by the ceiling probability algorithms used in 557 WW's EPS. The result is that nearly all of the ceiling forecasts at KVBG had poor skill and reliability. Reliability diagrams show that underforecasting was severe, and the ability to discern events of varying observed frequencies was poor.

The first example is shown in Figure 4.9 of the MEPS4 8 hr forecast of less than 3kft ceilings at KVBG. The observed frequencies for each bin are upwards of 70%, indicating a significant amount of underforecasting especially in the lower probability bins. The most heavily weighted bins between 0-30% lie outside the area of positive skill. The most glaring issue shown in the diagram is the number of missed events. The 0% bin contains 60 forecasts with an observed frequency of 70%. Therefore, MEPS4 missed 42 events for this hour. Average observed frequencies show underforecasting for the 72 hr forecast as a whole. The averages are as follows for bins 0% to 100% respectively at the following: 48.2%, 72.3%, 85.8%, 87.2%, 90.1% 90.8%, 90.2%, 93.7%, 94.7%, 89.9%, and 97.9%. Clearly, 3kft ceilings at KVBG occurred much more frequently than what was forecasted by MEPS4.

Figure 4.10 shows similar issues with the MEPS20 and GEPS. Again, there is evidence of low reliability and underforecasting. Most observed frequencies show differences from climatology, which improves the resolution term of the BSS. However, reliability and skill remain small. For 3kft ceilings, GEPS and MEPS20 achieved no positive skill over the whole forecast period

The other ceiling categories of less than 1kft and 500ft performed similarly to 3kft ceilings forecasts at KVBG. Skill remained negative for the entire forecast, while the underforecasting was significant.

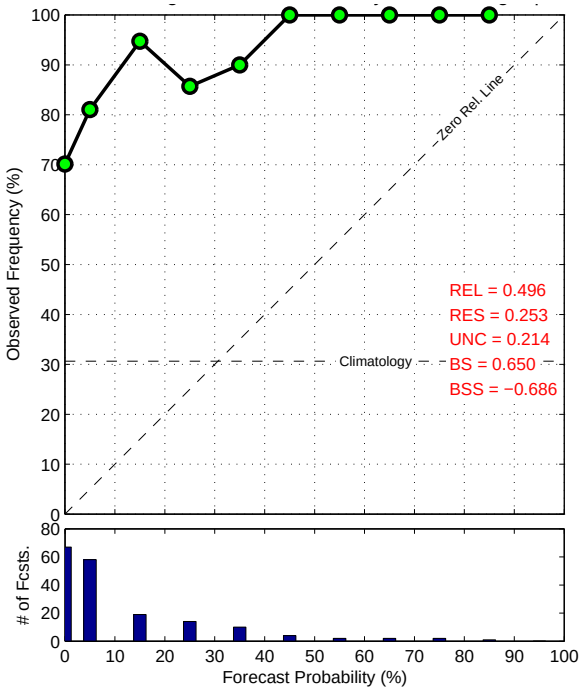
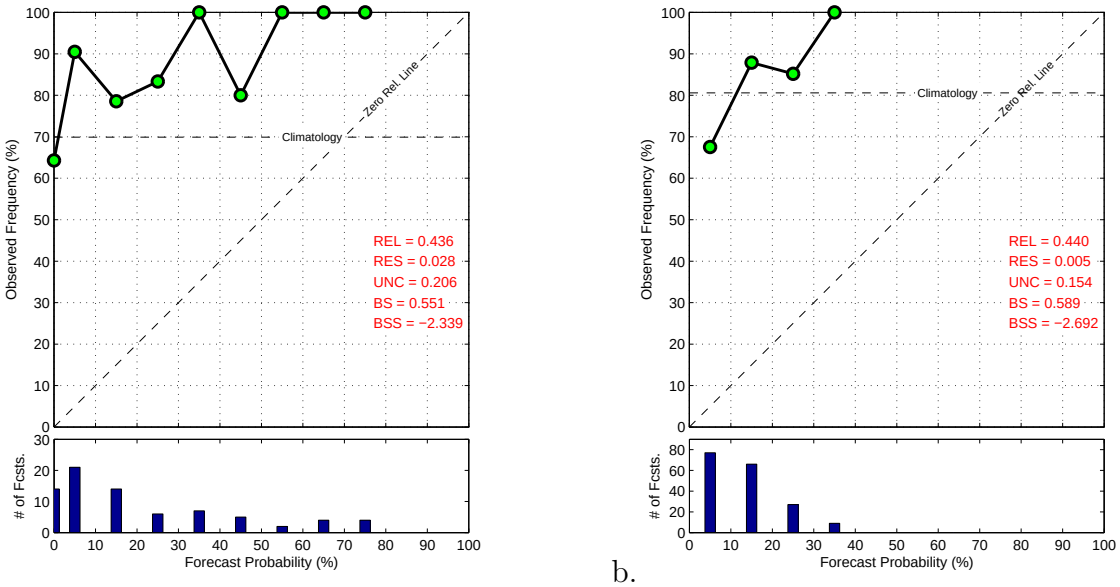


Figure 4.9. KVBG MEPS4 8 hr reliability diagram ceilings less than 3 kft indicating severe underforecasting.



a. b.

Figure 4.10. KVBG MEPS20 (a) and GEPS (b) reliability diagrams for ceilings less than 3kft.

4.3 Visibility

Table 4.2 shows a summary of BSS for the EPS forecasts of visibility less than 5 sm. Generally, skill for visibility shows less of a dependence on the model horizontal resolution and sometimes had no effect at all. The locations where skill had an insignificant dependence on model resolution were KOFF, KGRK, KWRI, KEND and KVBG where MEPS4 maintained positive skill for the majority of the forecast period at each location. GEPS struggled to achieve positive skill at KLFI and KBLV while MEPS4 showed significant improvements at those sites. At other locations like KGRK and KLRF, MEPS20 and GEPS performed similarly. For all of the EPS, the positive skill averages over the total forecast period for each site did not exceed .22. The worst skill scores came from KVBG, KHMN, KDMA, and KLSV. At KVBG, it is likely the case that the decreased skill is again associated with marine layer fog. For the drier desert locations like KDMA KLSV and KHMN, visibility restrictions were so infrequent that most BSS trends were unreliable or did not show significant improvements over climatology.

Usually, a higher resolution EPS has increased skill over a low-resolution EPS(Wang et al., 2012). Brad Clements (2014) also showed that MEPS4 most often outperformed MEPS20 and GEPS because of its higher resolution. However, depending on the location, the skill of visibility forecasts from GEPS and MEPS did not show significant differences from each other. Figure 4.11 shows one example at KGRK of forecasts of visibility less than 5 sm. In the initial 72 hr forecast period, MEPS4, MEPS20, and GEPS BSS peak at approximately the same values. MEPS20 and GEPS skill then trough to levels lower than MEPS4. In this case if the higher resolution EPS outperformed the lower resolution EPS, the increase in skill values tended to be small and occurred over short periods of time.

Visibility reliability diagrams illustrated a mix of results as some locations had

Site	EPS	Avg Positive Skill (0h-72h)	Avg Positive Skill (Total)	Skillful % of Forecast
KAFF	GEPS	.172	.164	62.5
	MEPS20	.220	.210	76.6
	MEPS4	.167	.167	100
KBLV	GEPS	.042	.043	50
	MEPS20	.092	.096	48.9
	MEPS4	.115	.115	97
KEND	GEPS	.136	.093	62.5
	MEPS20	.180	.156	87.2
	MEPS4	.163	.163	98.5
KGRK	GEPS	.140	.125	65
	MEPS20	.150	.170	63.8
	MEPS4	.137	.137	100
KLFI	GEPS	.055	.032	22.5
	MEPS20	.108	.086	46.8
	MEPS4	.111	.111	92.5
KLRF	GEPS	.060	.067	70
	MEPS20	.070	.067	61.7
	MEPS4	.104	.104	91.04
KOFF	GEPS	.088	.063	32.5
	MEPS20	.032	.044	59.6
	MEPS4	.073	.073	89.6
KVBG	GEPS	.090	.058	25
	MEPS20	0	.00890	2.1
	MEPS4	.051	.0796	61.2
KWRI	GEPS	.092	.061	37.5
	MEPS20	.084	.065	61.7
	MEPS4	.119	.119	94
KDMA	GEPS	.025	.018	37.5
	MEPS20	-	-	-
	MEPS4	.010	.010	34.3
KHMN	GEPS	.065	.047	57.5
	MEPS20	-	-	-
	MEPS4	.034	.034	67.2
KLSV	GEPS	.015	.015	52.5
	MEPS20	-	-	-
	MEPS4	-	-	-

Table 4.2. Visibility less than 5 sm BSS summary showing positive skill duration, skillful percentage of forecast, and average positive skill.

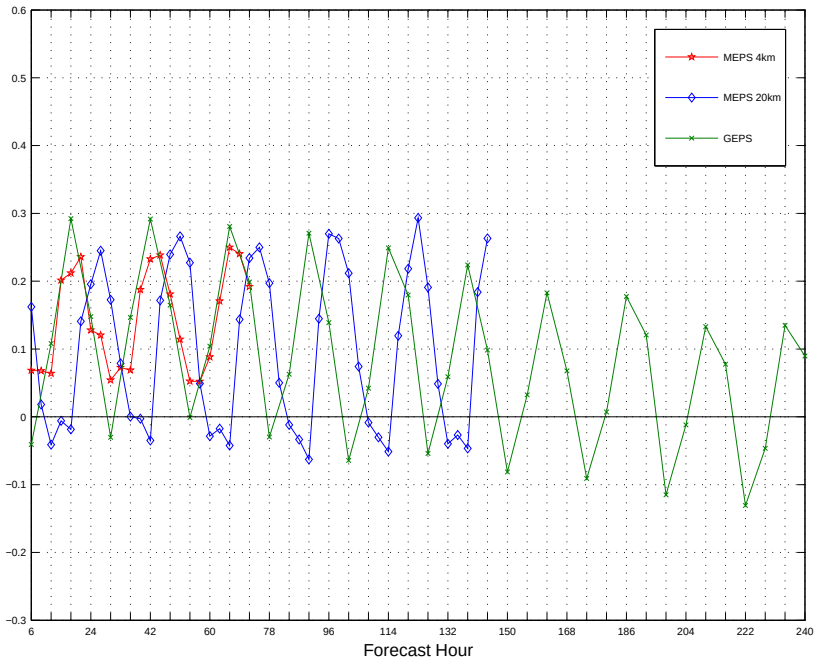


Figure 4.11. KGRK BSS model comparison for visibility less than 5 sm.

reliability diagrams with observed frequencies that closely matched forecast probability, yet had just as many examples of underforecasting depending on the time of day. At KBLV each EPS underforecasted at nearly all forecast hours. For the lower visibility thresholds of 3 sm and 1 sm, cases of underforecasting were more frequent and generally had less skill than forecasts of 5 sm visibility. This possibly occurred because significant skill over climatology is harder to achieve if the event occurs very infrequently. For example, if climatology is near zero and the EPS forecasts mostly zero percent, the EPS may be reliable but does not demonstrate much improvement over climatology.

An example of good reliability is shown in Figure 4.12 of the GEPS 18 hr and 12 hr forecast at KEND. In Figure 4.12(a), the most heavily weighted bins have observed frequencies that closely match the forecast probability. The other less populated bins with higher forecast probabilities show some slight underforecasting.

Figure 4.12(b) is from the same location and model but is for the 12 hr forecast.

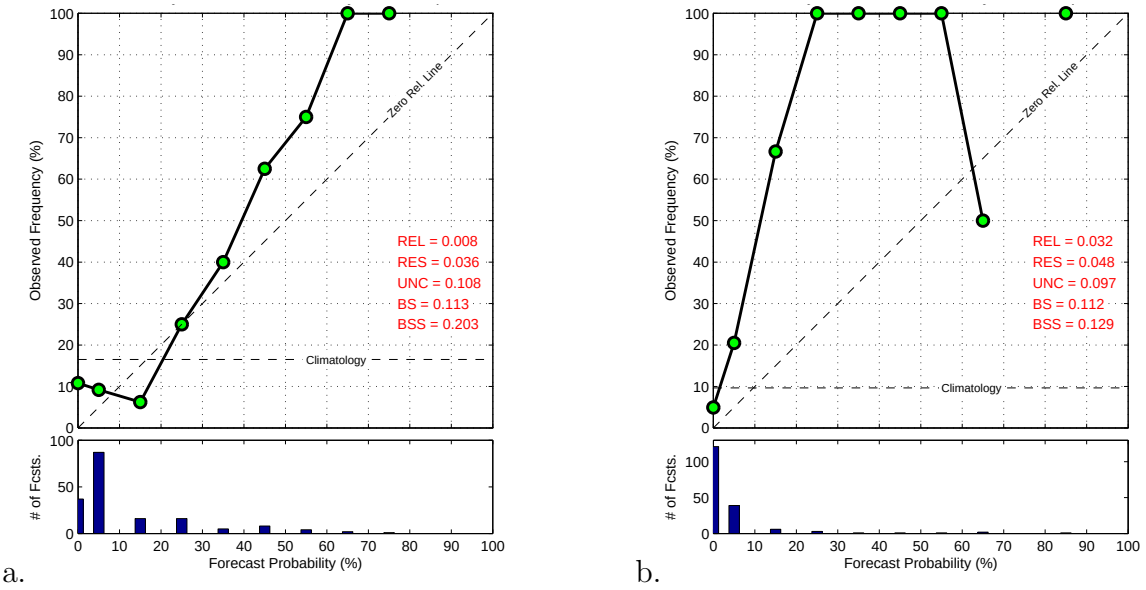


Figure 4.12. KEND GEPS reliability diagram comparison of visibility less than 5 sm. a: 18 hr forecast b: 12 hr forecast.

In the 12 hr forecast, the degree of underforecasting is more significant than the 18 hr forecast. The observed frequency is 20% for the 1-10% bin and 67% for the 11-20% percent bin. Plots of reliability, resolution and uncertainty for GEPS visibility forecasts reveal that reliability is best during times when the uncertainty is highest, and it is worst when uncertainty is low. Therefore, the reliability is best at times when visibility less than 5 sm occurs most frequently according to climatology. Therefore, the reliability is best from 6 pm to 6 am each day at KEND when visibility tends to drop due to overnight fog.

MEPS4 had several examples of good reliability but, there were also as many examples of underforecasting. Averaging the observed frequencies over the entire forecast most often indicated that the EPS overall underforecasted visibility. Figure 4.13 shows a good reliability example from MEPS4 at KGRK. Each observed frequency is relatively close to the forecast probability for their respective bins. The 51-60% had a 100% percent observed frequency. Since only one forecast occurred for this bin however so it contributes less to the overall reliability and skill. However, if

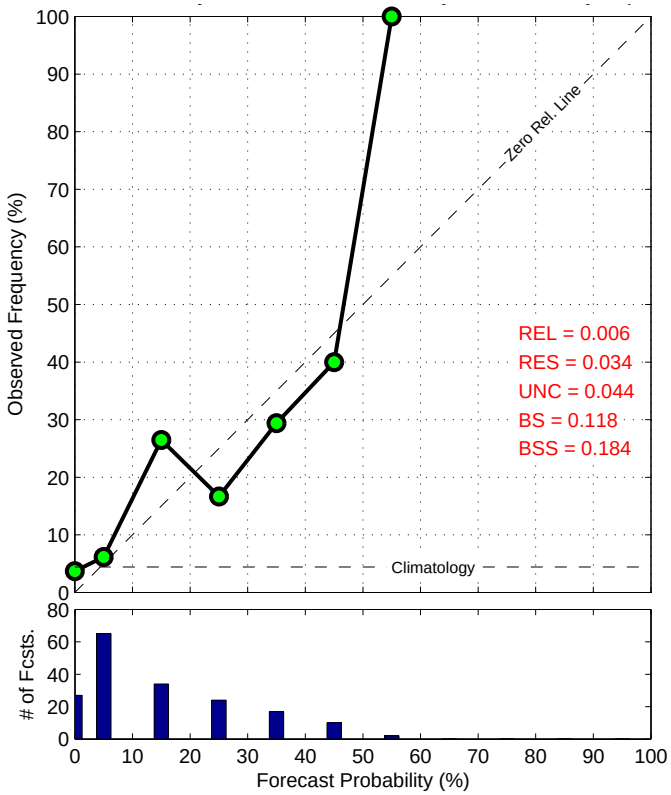


Figure 4.13. KGRK MEPS4 22 hr reliability diagram for visibility less than 5 sm.

we average the observed frequencies at KGRK over every forecast hour, we get the following for bins 0% through 71-80% bins respectively: 6.5%, 24.5%, 46.8%, 60.1%, 69.7%, 69.07%, 100% and 100%. These averages reveal an overall underforecasting trend. For other locations, it was usually the case that each EPS had some degree of underforecasting forecasting visibility, but the severity of underforecasting depended upon the location.

Like ceilings, the BSS for visibility forecasts followed diurnal patterns for several locations including KAFF as shown in Figure 4.14. The figure depicts the BSS falling to near zero values for a short period at around the same time each day. In local time, the period of decreased skill is during the afternoon hours. In the summer time, the afternoon also happens to be when the Air Force Academy receives the largest frequency of thunderstorms according to climatology. Therefore,

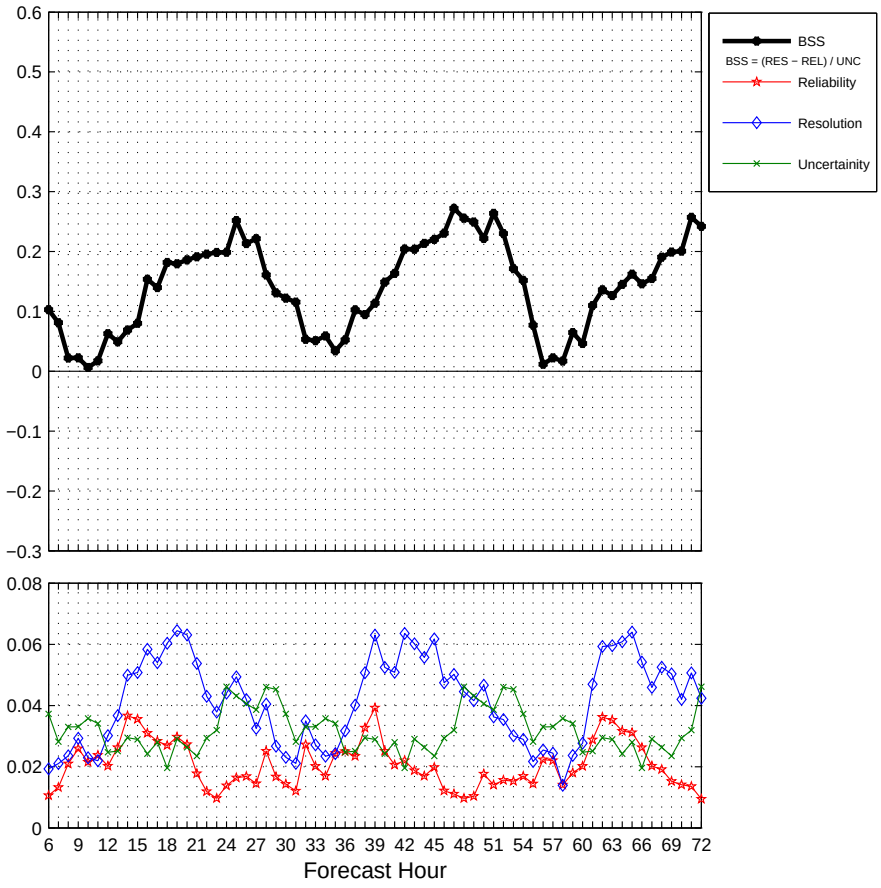
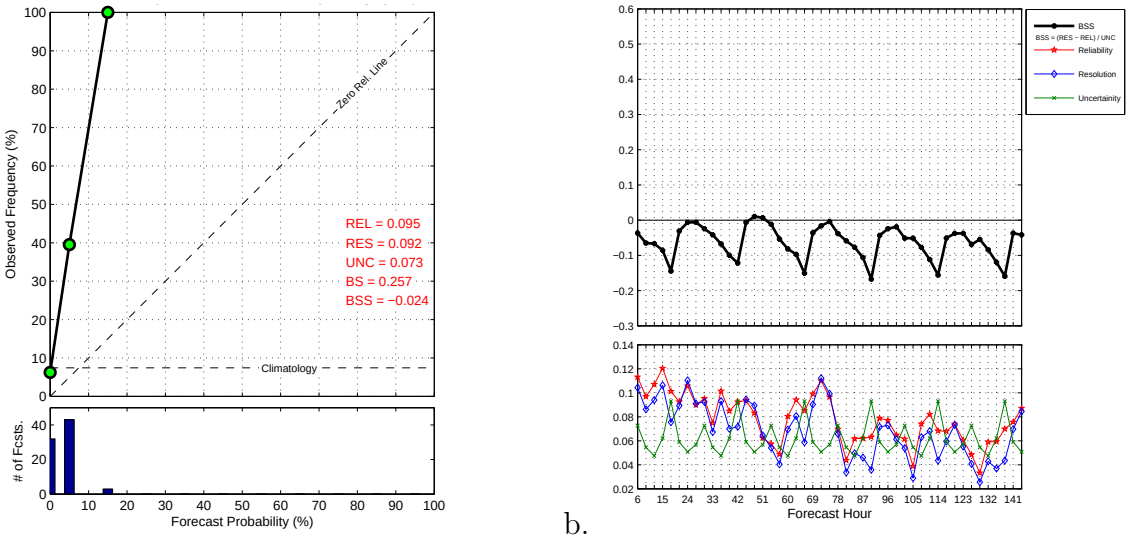


Figure 4.14. KAFF MEPS4 BSS for visibility less than 5 sm from Apr-Oct 2015

it is possible that the EPS has less skill with forecasting decreased visibility caused by heavy rain in thunderstorms.

The forecasts for lower thresholds were more biased than the 5 sm visibility category. For the lower limits of 3 sm and 1 sm visibility, each EPS tended to forecast small probabilities, while the observed frequencies were often much larger. An example is shown in Figure 4.15(a) of the MEPS20 30 hr reliability diagram of visibility less than three sm at KLFI. The figure clearly shows an underforecasting bias. Most of the forecasts lie in the 0-10 percent bin where the observed frequency was 40%. The 20-30 percent bin had three forecasts with a 100% observed frequency. In 4.14(b), skill values for the forecasts of less than 3 sm visibility remained small if not negative, due to the more significant underforecasting bias.

BSS values fluctuated between 0 and -.2 for the full forecast.



a. b.

Figure 4.15. KLFI MEPS20 visibility less than 3 sm reliability diagram (a) & BSS plot (b). Significant underforecasting resulted in mostly negative BSS values.

Some sites, however, achieved positive skill with the lower thresholds, but results overall depended on the location. Figure 4.16 shows a comparison of the skill of forecasts for each category of visibility at KWRI. Skill of the 5 sm visibility forecasts remained almost 100% positive, 3 sm visibility forecasts fell below zero more often, and 1sm visibility forecast skill stayed mostly below zero.

In summary, visibility forecasts often showed positive BSS and acceptable reliability, however for the lower thresholds of visibility less than 3 and 1 sm, skill and reliability were less. An underforecasting bias was present overall that varied with location and time of day. The bias was more significant at the lower 3 sm and 1 sm visibility forecasts.

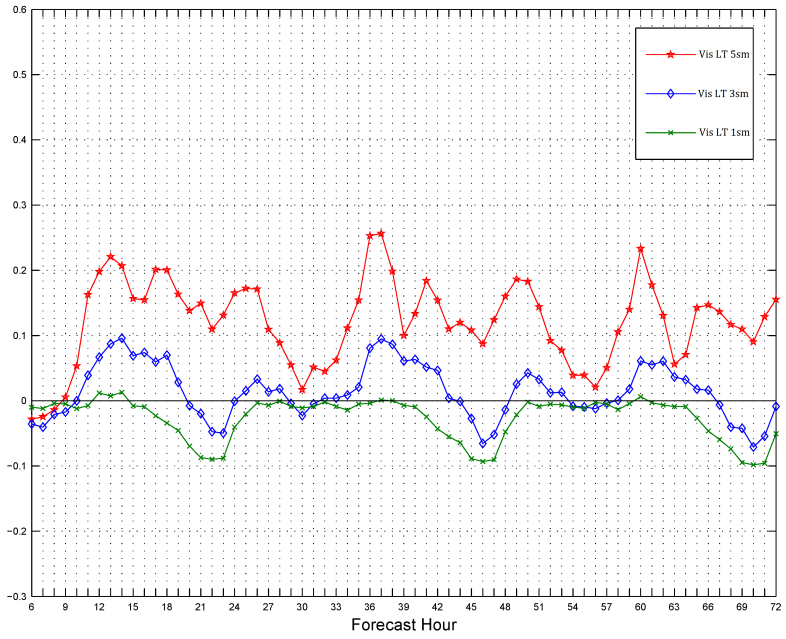


Figure 4.16. KWRI MEPS4 BSS comparison of forecasts of visibility for each threshold from Apr-Oct 2015.

4.4 Precipitation

In Brad Clements 2014 study, he found that precipitation forecasts improved with increasing model resolution. He distinguished between cases of precipitation caused by synoptic forcing and by thunderstorms that develop through daytime heating and small-scale lifting mechanisms. In small scale convective systems, reliability and skill increased sharply with model resolution. However, in the case of synoptically forced precipitation, model resolution did not have as much of an effect. This study produced results that mostly agree with these conclusions.

There are several locations in this study that receive precipitation in the summer from convective systems induced by small-scale mechanisms. Those include 5 locations from the Florida panhandle and 3 that receive terrain induced convection like KAFF, KLSV and KHMN during the summer monsoon. MEPS20 and MEPS4 had a higher BSS than GEPS at 6 of these 8 locations. At the other ten sites, GEPS sometimes had better skill scores like at KLRF and KLFI but was still

outperformed most of the time.

Brad Clements showed that at Kunsan Air Base, Korea, each EPS produced similar scores and concluded that each model is equally capable of forecasting precipitation associated with synoptic weather systems. In this 2015 study, GEPS performed as well as MEPS in some instances for likely the same reasons. Besides Florida and the locations within the Rocky Mountains, each location lies in mid-latitudes over flat terrain. Therefore, each of those locations typically experiences synoptically forced thunderstorms and precipitation. Figure 4.17 supports the idea showing that BSS averages were smaller for locations that mostly receive small scale precipitation events.

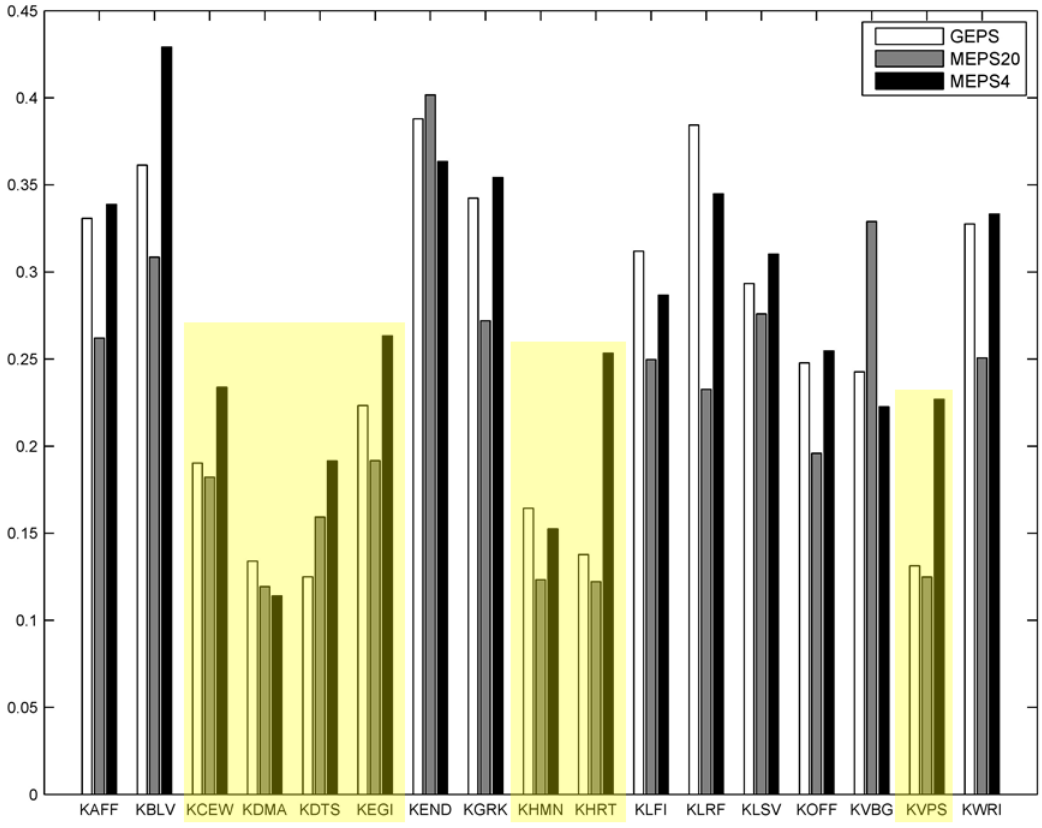


Figure 4.17. Mean BSS model comparison for 6 hr precipitation forecasts at all locations. Averages are taken over the first 72 hr of forecasting from GEPS, MEPS20, and MEPS4 at each location. Highlighted areas show locations that had the lowest BSS averages.

GEPS is capable of forecasting out to 240 hr or 10 days. For some weather models, forecasts are deemed uncertain past about 5 days. GEPS demonstrated that, for precipitation forecasts, it was able to maintain good skill over climatology even out to the end of the ten-day forecast. At Langley, GEPS achieved a mean skill of .18 for the whole 240 hr and .31 for the first 72 hr. As expected, the skill of the model drops near the end of the forecasts. The drop occurs because over time internal errors grow exponentially in each ensemble member. Therefore by the end of the forecast, model solutions tend to diverge significantly, creating more uncertainty and an ensemble mean that may drift farther from the actual state of the atmosphere. The GEPS KLFI precipitation forecast is a clear example of that. Shown in Figure 4.18, the skill drops to lower levels near the end of the forecast, as expected. It is interesting to note that the reliability did not degrade with time. Reliability stayed steady while resolution began to fall by the 96 hr point. The degradation of resolution and the consistency of reliability indicates that over time, GEPS continues to output reliable forecast probabilities but the forecasts deviate less from climatology.

The effect of falling resolution on GEPS is shown in Figure 4.19. Figure 4.19(a) is a reliability diagram from the 48 hr forecast, and 4.18(b) is from the 240 hr forecast. Each example shows good reliability. The observed frequencies from each model closely match the zero reliability line. However, one notable difference between the diagrams is how the forecasts are more spread out in Figure 4.19(a). The 240 hr chart shows most of the forecasts in the 1-30% percent bin while the 48 hr diagram contains forecasts in each bin from 0 to 80%. The difference is the effect of decreased resolution as given in the brier score decomposition. Forecasts with a high resolution score distinguish between cases that occur more or less than climatology. An EPS with a low resolution score offers less value over climate

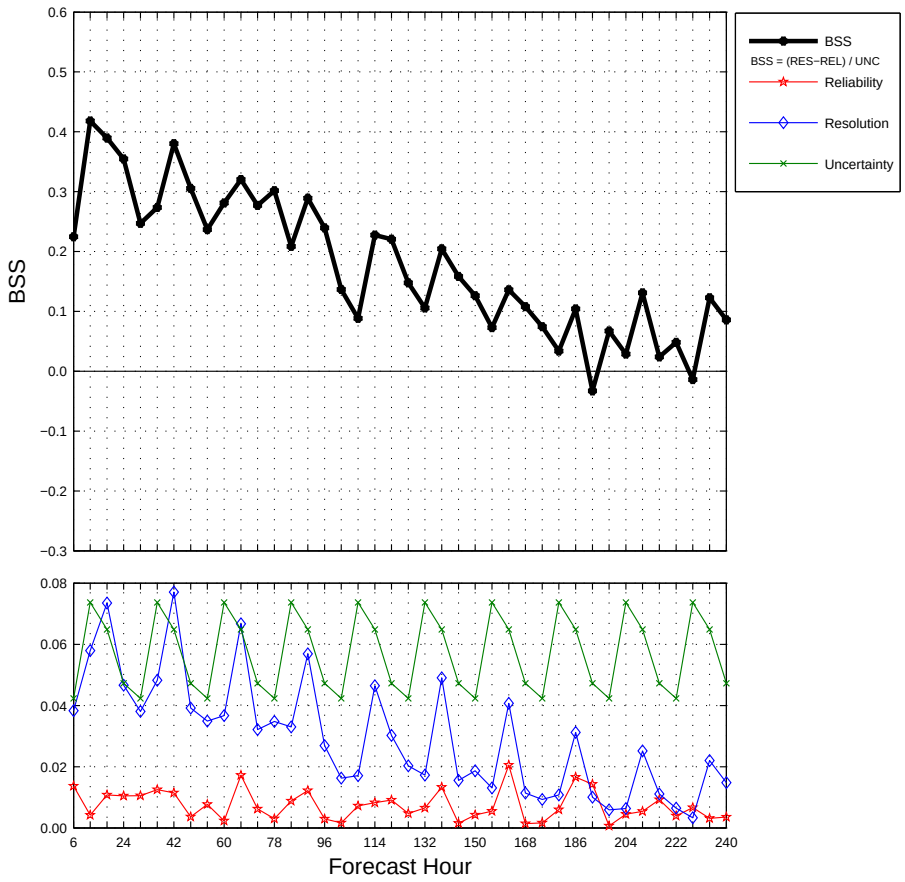


Figure 4.18. KLFI GEPS BSS diagram for 6 hr precipitation from Apr-Oct 2015.

because of how close the forecasts are to climatology. For GEPS, skill typically degraded at around the 96 hr point due to the decreased resolution score of the forecasts.

GEPS proved to be reliable at forecasting precipitation, and in several instances, the reliability of GEPS probabilities were better than MEPS20 and MEPS4. Table 4.3 shows nine instances where GEPS had better reliability than at least one Mesoscale EPS. While it is desired for probability forecasts to be reliable, it is not the only measure of skill. It is also important to have resolution (not to be confused with horizontal resolution), and the table shows that at each site, MEPS had better resolution than GEPS. The higher resolution values indicate that the EPS is forecasting a wider range of probabilities that differ from climatology. So although

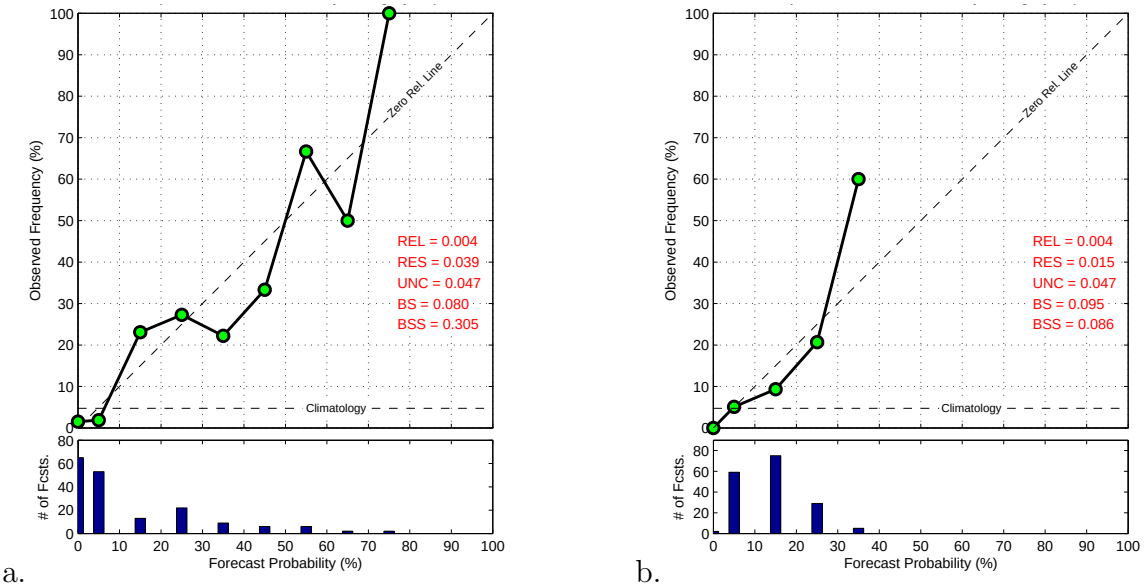


Figure 4.19. KLFI GEPS reliability diagrams for 6 hr precipitation: a: 48 hr forecast. b: 240 hr forecast.

the forecasts from MEPS may not be as reliable for some locations in the table, they still are providing a more detailed forecast. Its is possible that MEPS, the more resolved model, may just need some calibration to make forecasts closer to zero reliability. As shown in the Table 4.3, MEPS proved that the increased resolution over GEPS made it more skillful over climatology.

Site	EPS	Avg Reliability	Avg Resolution	Avg. Positive Skill
KAFF	GEPS	.0116	.0395	.214
	MEPS20	.0363	.0976	.266
	MEPS4	.0132	.0478	.339
KVPS	GEPS	.0184	.0190	.106
	MEPS20	.0356	.0441	.130
	MEPS4	.0146	.393	.227
KCEW	GEPS	.0503	.0164	.107
	MEPS20	.0298	.0517	.177
	MEPS4	.0194	.0422	.234
KHRT	GEPS	.0168	.0164	.102
	MEPS20	.0278	.0272	.115
	MEPS4	.0161	.0414	.253
KEGI	GEPS	.0119	.0240	.130
	MEPS20	.0346	.0597	.189
	MEPS4	.023	.0522	.263
KLRF	GEPS	.00742	.0356	.262
	MEPS20	.0227	.0558	.233
	MEPS4	.0143	.0518	.345
KDMA	GEPS	.00410	.00947	.0826
	MEPS20	.00755	.00579	.133
	MEPS4	.00670	.0132	.114
KHMN	GEPS	.00414	.0093	.0995
	MEPS20	.0148	.0184	.119
	MEPS4	.00917	.0158	.153
KLSV	GEPS	.00284	.0049	.132
	MEPS20	.0021	.0005	.276
	MEPS4	.00481	.0088	.310

Table 4.3. Average Reliability, Resolution, and Skill for 6 hr precipitation over the entire forecast period of each EPS.

4.5 Lightning

In Clements (2014) study, MEPS and GEPS were shown to have a significant overforecasting bias during the overnight hours when thunderstorms are less frequent. Upon review of lightning forecasts for new locations, there are several examples of the same overforecasting trend. Also in the previous study, MEPS20 proved in some cases to have a higher average positive skill than MEPS4, but

MEPS4 had far more hours of positive skill proving that it was overall the best performer. In the current study, MEPS4 consistently had higher positive skill averages over the first 72 hr and the number of hours of positive skill typically matched MEPS20. However, there were several cases where GEPS had the most hours of positive skill during the first 72 hr.

Figure 4.20 shows that in 14 out of 17 cases that MEPS4 had the highest positive score average over the first 72 hr of forecasting. Outliers exist at KLSV and KVBG due to a very limited number of cases of lightning. Another point to note about the figure is that at KDMA, KAFF, and KHMN, GEPS performance was especially weak while MEPS4 outscored it by a relatively large margin. The vast gap in scores at these locations is likely due to their proximity to mountains that is not depicted accurately by the reduced horizontal resolution of GEPS. Similarly, for the Florida coast locations, GEPS had lower BSS where small-scale sea breeze convergences affect the development of thunderstorms and were better depicted by higher resolution MEPS.

Figure 4.21 illustrates model performance from a different perspective, revealing the effect of the length of the forecast interval on score. The figure shows the fraction of hours of the first 72 that had positive skill. In 8 out of 17 cases, GEPS maintains a longer period of positive skill over the first 72 hr despite having lower average scores. In many of those instances, GEPS had 100 percent positive skill. Improved EPS accuracy is not necessarily the cause. Rather, it is likely due to the length of the forecast period of each model. GEPS forecasts at 6 hr intervals, while MEPS4 forecasts at 1 hr intervals.

Due to the intermittent nature of lightning and the difficulty in predicting onset, a 1 hr forecast is harder to verify than a 6 hr forecast. The shorter forecast causes more instances of false alarms and overforecasting that result in more variability in

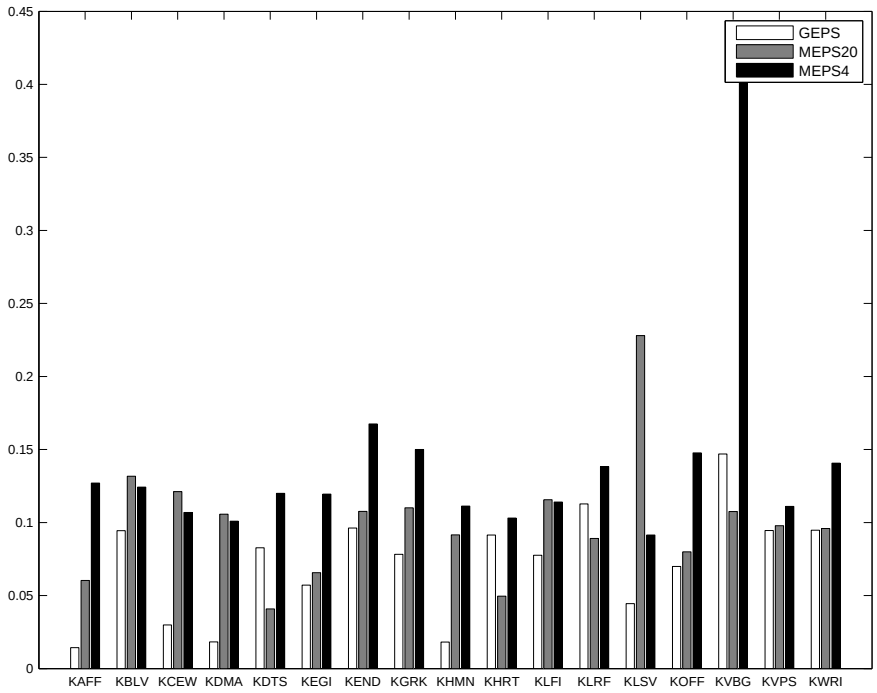


Figure 4.20. Mean positive BSS model comparison for lightning forecasts at each location/EPS over the first 72 hours.

BSS as shown in Figure 4.22. Also, most of the instances where GEPS had a higher percentage of positive skill were at locations that are mostly influenced by synoptic weather patterns. As shown from the precipitation forecasts, GEPS lightning performance improves in locations where synoptic forcing dominates thunderstorm generation. So despite lower numbers of positive skill hours, MEPS4 still shows higher BSS in agreement with results in Clements (2014). In any case, the 1 hr forecasts are much more valuable from a forecaster's perspective for forecasting thunderstorms than a 6 hr forecast. Therefore, GEPS forecasts have less value regardless of their apparent positive skill.

During certain times of the day and depending on location, each EPS demonstrated a tendency to over forecast thunderstorms. Typically overforecasting was most significant during the evening hours once daytime heating was over, and most thunderstorm activity broke down. In other instances, the ensembles would

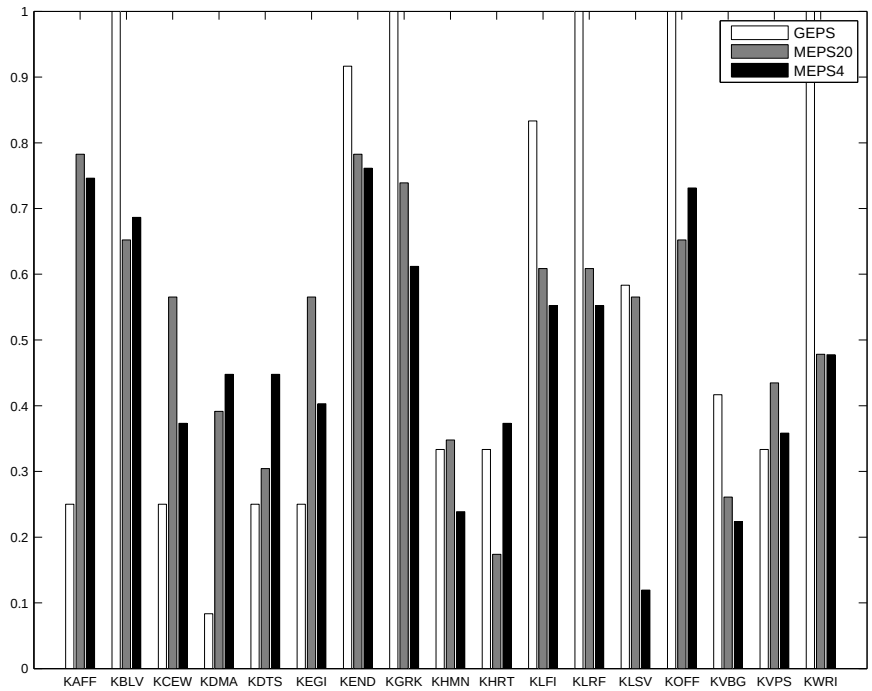


Figure 4.21. Percentage of hours with positive BSS model comparison for lightning forecasts at each location/EPS over the first 72 hours

overforecast lightning just before peak hours of thunderstorm activity. One example is illustrated by reliability diagrams in Figure 4.23 of the MEPS4 lightning forecast at KDMA. Figure 4.23(a) shows the 7 hr forecast for 1000-1100 local time. Figure 4.23(b) is the nine hr forecast for 1200-1300 local time. It is clear that the 7 hr forecast severely overforecasted lightning as most of the forecasts were a false alarm. There were a total of 57 forecasts ranging between 1% and 80%. Of those, there were just three occurrences of lightning within 20 nm.

The forecast for 2 hr later, however, was much more reliable as shown in diagram Figure 4.23(b). Reliability, resolution, and BSS all improve for the nine hr forecast, with the observed frequencies close to each forecast probability. This trend between high and low quality forecasts repeated daily at regular intervals over the 72 hr period. MEPS4 started to overforecast lightning at 2200 local time at KDMA and continued to do so until 1100 local time. The cycle of overforecasting bias alligns

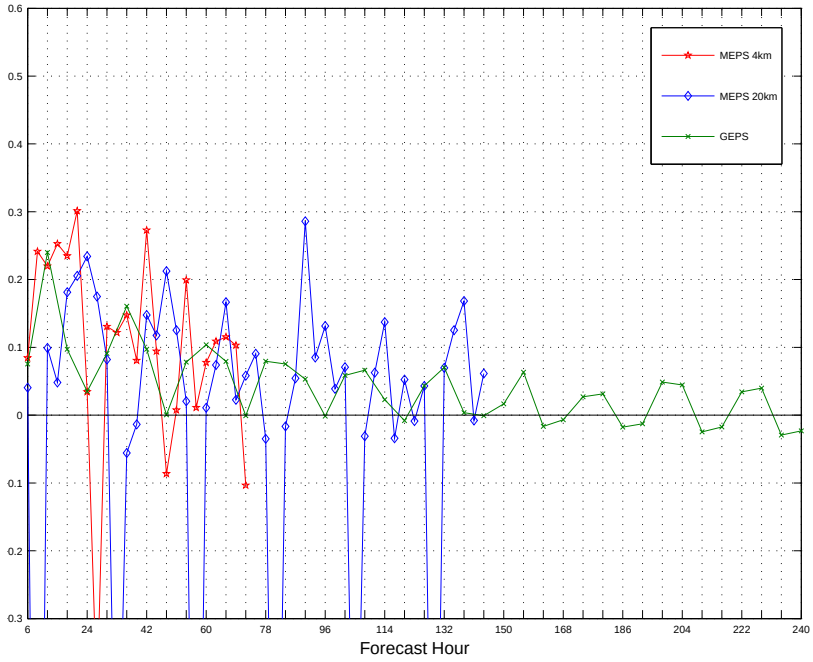


Figure 4.22. KEND BSS model comparison for lightning forecasts. GEPS BSS remains relatively steady while MEPS BSS shows wide variability and sharp drops at regular intervals but, overall higher skill than GEPS.

with the daily cycle of thunderstorms activity which follows diurnal heating patters. As a result of the overforecasting bias, MEPS had 44% positive skill. 44% was still much better than GEPS, which had just 8.3% positive skill in the first 72 hr.

Of all of the locations in this study, KAFF had the highest number of lightning forecasts and occurrences. For MEPS4, there were a total of 3719 forecasts greater than 0% with 1035 verified events. With a larger amount of data any biases become more conclusive. When a forecast probability bin has very few forecasts there is a greater margin of error. MEPS4 proved to be quite reliable during times of peak heating at KAFF. The times of peak heating are shown in Figure 4.24 by the diurnal variability of uncertainty. When the uncertainty is higher, the climatology is closer to 50% thus, lightning tends to be more frequent during those hours. The BSS diagram shows that MEPS4 maintained positive BSS for most of the 72 hr forecast including times of peak heating. The BSS tended to fall or go negative

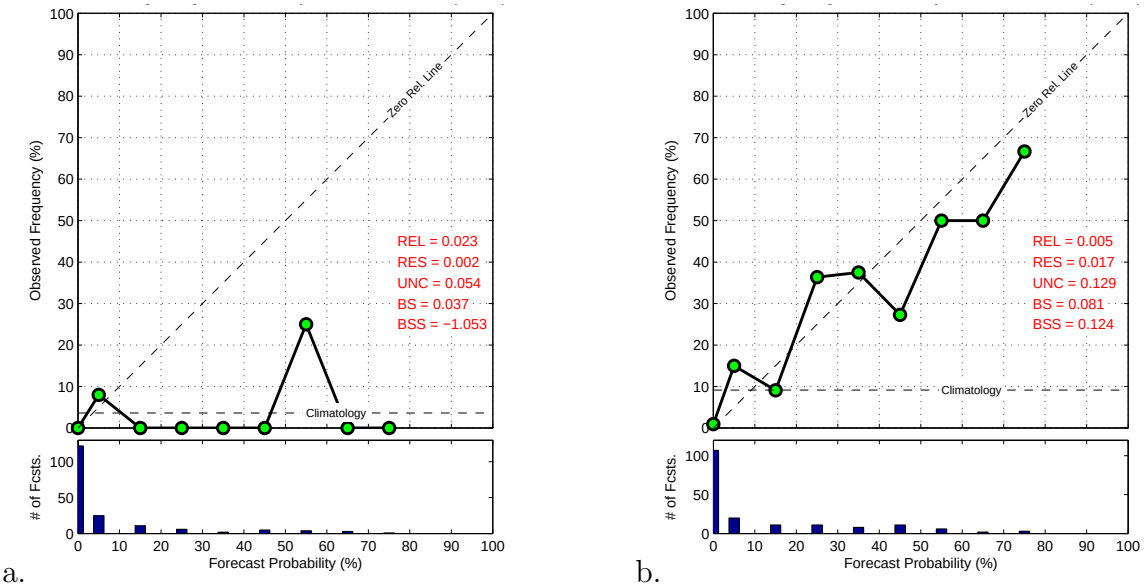


Figure 4.23. KDMA MEPS4 reliability diagrams for lightning within 20 nm. a: 7 hr forecast b: 9 hr forecast.

during the hours leading up to peak heating or just after.

The example reliability diagram in Figure 4.25 shows good reliability for the MEPS4 34 hr forecast of lightning at KAFF. In this figure, there are a total of 117 forecasts above 0%, with the majority of forecast probabilities closely matching the observed frequency. Over the whole 72 hr forecast, the average observed frequencies for bins 0% through 91-100% respectively were: 1.8%, 12.1%, 20.6%, 25.0%, 33.6%, 42.1%, 50.9%, 65.64%, 57.60%, 67.26%, and 0%. Except for the last three bins, the total 72 hr forecast was overall very reliable. The last three bins only accounted for 3% of the total number of forecasts, so there is more margin for error in those bins.

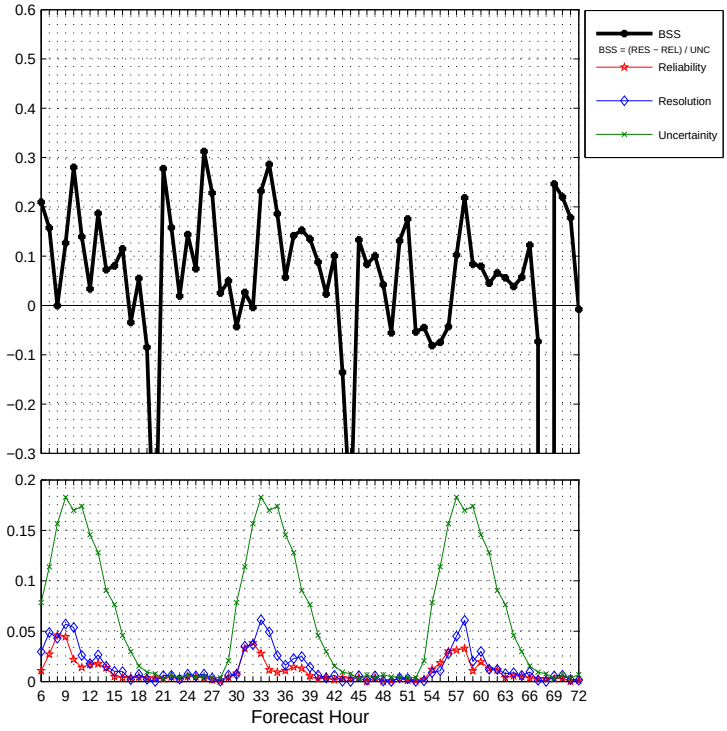


Figure 4.24. KAFF MEPS4 BSS for lightning within 20 nm.

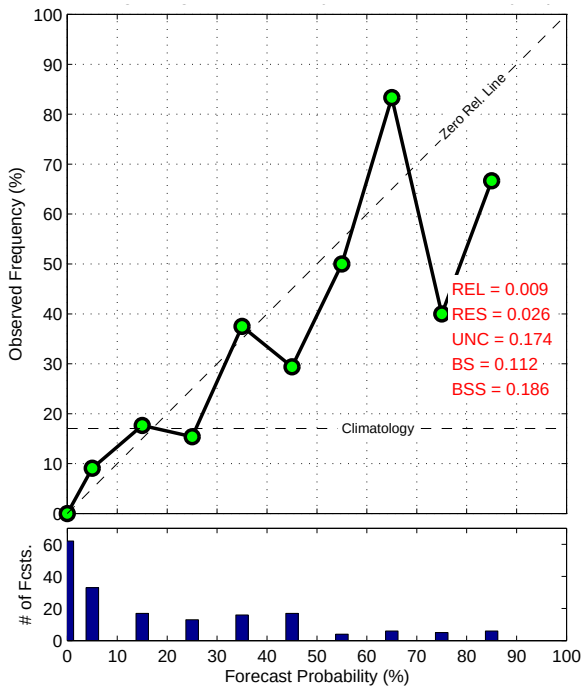


Figure 4.25. KAFF MEPS4 35 hr reliability diagram for lightning within 20nm.

4.5.1 Sea Breeze Thunderstorm Forecasts in Florida.

This study included 5 locations from the Florida panhandle to examine EPS performance on small-scale, localized sea breeze induced thunderstorms. Those locations are Duke Field (KEGI), Eglin AFB (KVPS), Fort Walton Beach (KDTS), Hurlburt Field (KHRT), and Bob Sikes Airfield (KCEW). KVPS, KDTS, and KHRT are within 3 mi of the gulf coast while KEGI and KCEW are further inland.

Since GEPS provides 6 hr forecasts, it is not as useful to predict the onset of thunderstorm activity that varies considerably from hour to hour. Therefore, this section examines MEPS20 and MEPS4 exclusively, providing comparisons between the two. Recall that MEPS20 parameterizes convection, while MEPS4 does not, providing two unique approaches to lightning forecasting. Also recall that MEPS4 forecasts lightning for a 20nm radius of the site location while MEPS20 forecasts for lightning within a 20 km radius. The MEPS20 data in this section includes forecasts from May-July while MEPS4 data spans from May-Oct. The 00z and the 06z model run were chosen for MEPS4 and MEPS20 respectively since they had the most available PEP bulletins.

MEPS4 produced the highest BSS score averages at 4 out of the 5 Florida locations. The highest skill values achieved from each model also typically came from MEPS4. However, it attained the most hours of positive skill at just 2 of the 5 locations. Figure 4.26 shows direct comparisons of MEPS4 and MEPS20 over the first 72 hr of forecasting. Although MEPS4 was able to achieve higher skill overall, it had some issues of overforecasting during certain hours causing the MEPS4 BSS to drop significantly more than MEPS20.

Figure 4.27 compares the BSS of the MEPS4 00 UTC model run and the MEPS20 6z model run for lightning forecasts at KCEW beginning at 12 UTC. Similar to KDMA and KAFF, MEPS4 skill drops significantly at regular intervals

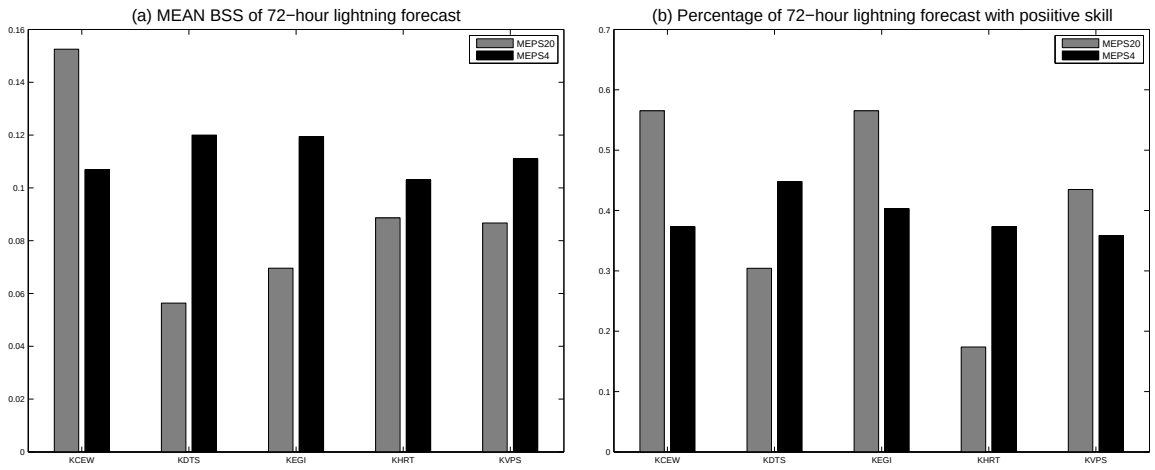


Figure 4.26. Mean BSS (a) and percent positive BSS (b) of lightning forecasts at Florida locations.

while MEPS20 is able to maintain better BSS. Generally, the drop begins around 00 UTC (1900 local) and remains low or falls to large negative values until the next morning. For the rest of the time, MEPS4 BSS generally exceeds the MEPS20 BSS. This relationship between MEPS20 and MEPS4 was not only true for Bob Sikes but for all locations in Florida. The drop in skill for MEPS4 coincides with a time when the model had significant overforecasting bias.

The negative BSS values from MEPS4 and MEPS20 are attributed to an overforecasting bias in most instances. Among afternoon thunderstorm occurances the model had the most reliability when the climatological frequency was relatively high. Reliability fell, and overforecasting was more common during times when the climatological frequency was lower. The times of overforecasting coincides with overnight or the early morning hours when convective available potential energy is low. One example is shown by reliability diagrams in Figure 4.28 of MEPS4 lightning forecasts at KHRT. Figure 4.28(a) coincides with 0600 local time. Clearly there is an overforecasting bias, with even the 91-100% bin having a zero percent observed frequency. Figure 4.28(b) shows the 46 hr forecast valid at 2200 local time. There are far more forecasts during this period, and they are much more reliable.

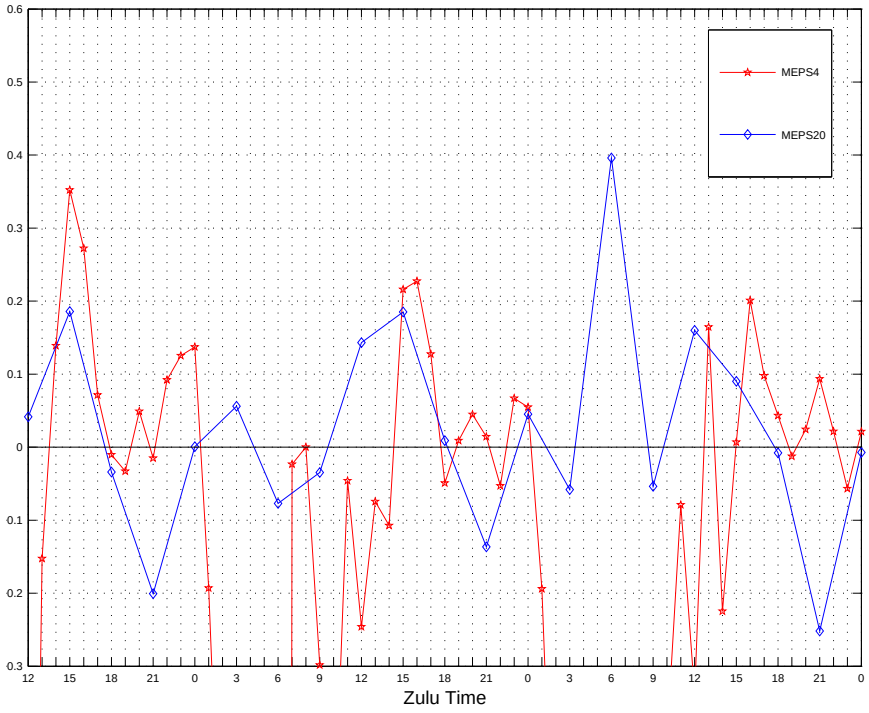


Figure 4.27. KCEW MEPS model comparison of lightning within 20 nm (MEPS4) and 10 nm (MEPS20).

4.28(a) highlights one of the the worst performing forecasts and 4.28(b) highlights one of the best. For other hours, reliability and skill was somewhere in between the two.

Figure 4.29 shows the reliability and skill trends for KHRT. It is evident that the BSS is higher during times of increasing uncertainty when thunderstorms are most common. Although skill is overall better during these times, BSS drops when the uncertainty peaks at forecast hours 18, 44, and 68. Overall, skill remained positive but low during the day and dropped to negative values overnight. Patterns are not as clear in the resolution and reliability values, but it is evident that the drops in BSS coincided with decreases in resolution. The BSS chart in Figure 4.29 along with the diagrams from Figure 4.28 show the diurnal variability of skill and reliability of lightning forecasts at KHRT. Similar trends in reliability resolution and BSS also apply to the other 4 Florida locations.

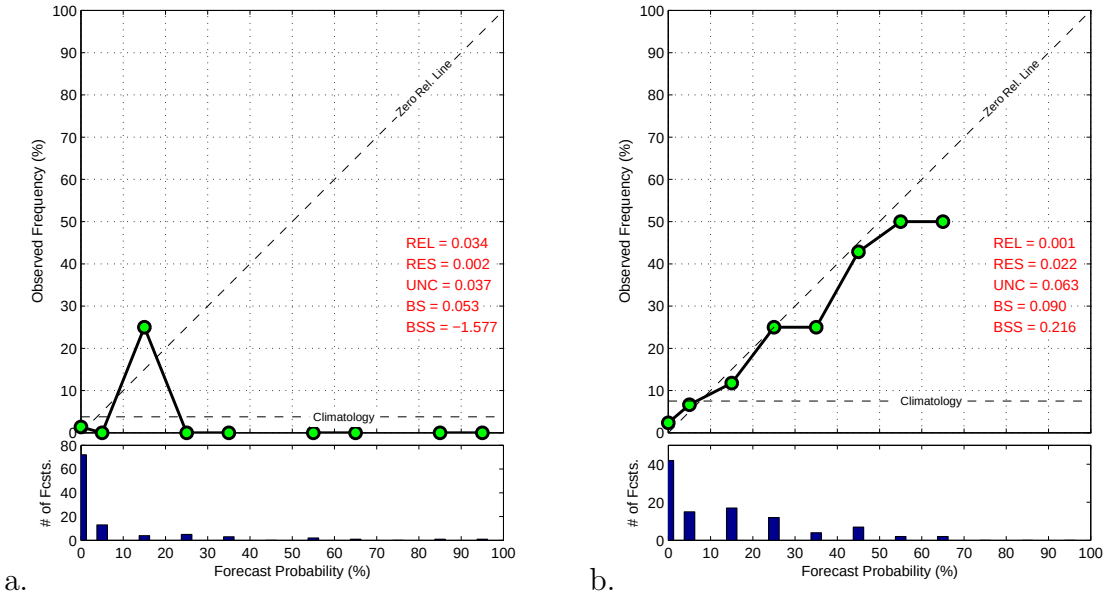


Figure 4.28. MEPS4 KHRT reliability diagram comparison for lightning within 20 nm. a: 11 hr forecast. b: 46 hr forecast.

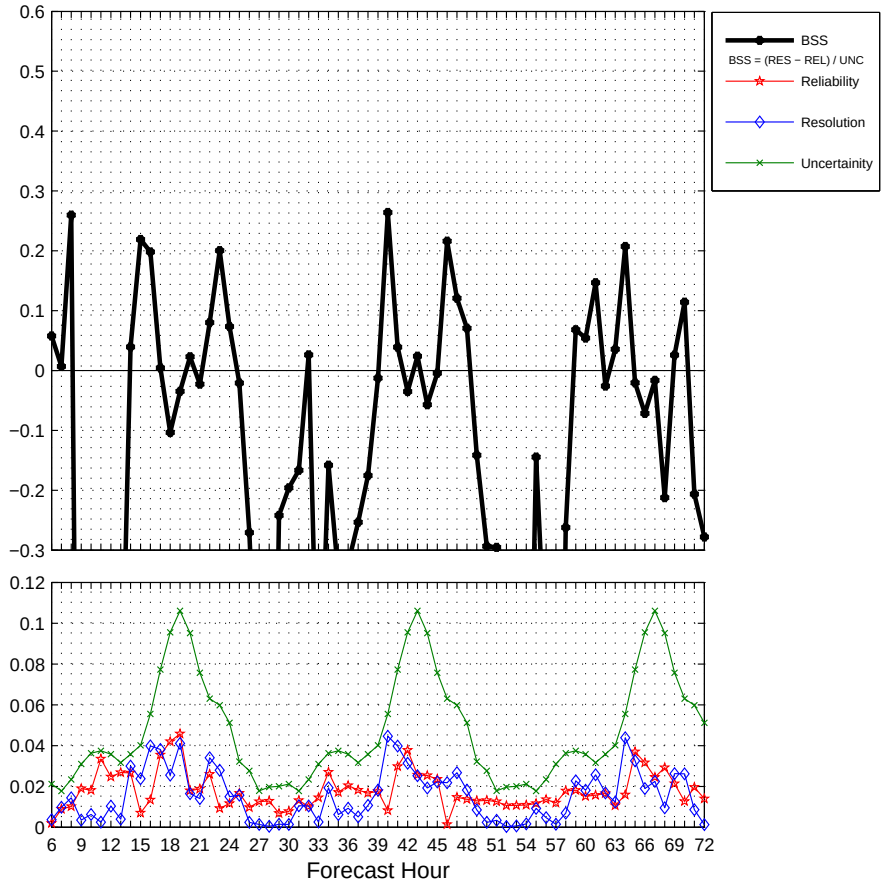


Figure 4.29. KHRT MEPS4 BSS for lightning within 20nm forecasts from Apr-Oct.

4.6 Winds

Stronger wind events are infrequent during the summer season. If they do occur in the summer, the events are typically associated with strong thunderstorm activity or diurnal mountain/valley breezes. High wind events such as 35 kt or 50 kt winds are rare, especially in the summer, and may occur just a few times or not at all depending on the location. The limited data means that BSS and reliability diagrams for 35 and 50 kt winds were discarded due to unreliable and irrational trends. There was an acceptable amount of data however for 25 kt winds at most locations. These results follow the conclusion from Clements (2014) that BSS and reliability improve with increasing model resolution. There were several exceptions, however, where MEPS20 BSS increased over MEPS4 BSS, but MEPS4 had better reliability in each case.

Reliability diagrams of 25 kt wind forecasts at KDMA show that MEPS4 improved over MEPS20 and had far more reliability than GEPS. Shown in Figure 4.30(b), the GEPS 25 knot wind 12 hr forecast missed several events at KDMA. There are 118 forecasts in the 0% bin, yet the observed frequency is 23.7%. Therefore, GEPS missed 29 events of 25 kt winds at KDMA over the 6 month period. An underforecasting bias is also apparent, which affected most of the KDMA wind forecasts as well as all sites within vicinity of complex terrain including KAFF, KHMN and KLSV. The corresponding MEPS20 forecast (not shown in Figure 4.30) had less bias, increased reliability at .012, and similar BSS at .063 while missing 2 events. The MEPS4 11 hr forecast shown in 4.30(a) improved reliability and BSS to .007, and .25 respectively while missing zero events.

Clements (2014) concluded that a lower resolution EPS does not as accurately depict terrain and elevation resulting in more unreliable wind forecasts. The same conclusion also applies in this study to additional locations at KAFF, KHMN,

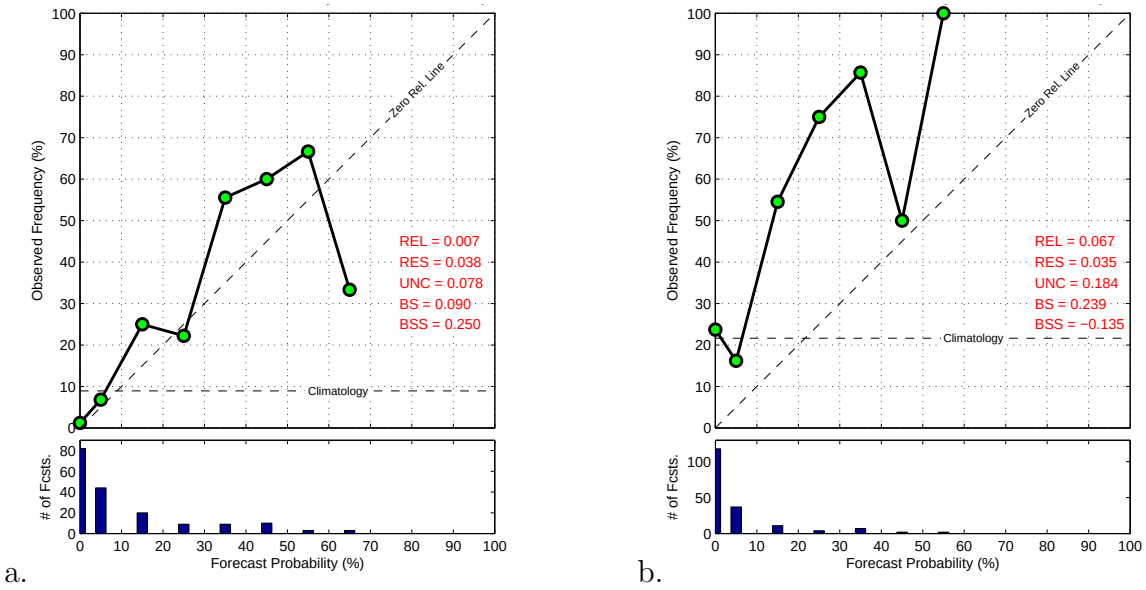


Figure 4.30. KDMA MEPS4 reliability diagram comparison for winds greater than 25 kt. a: 11 hr forecast b: 12 hr forecast.

KLSV, and KDMA, which frequently observe winds influenced by terrain effects. GEPS has a significant underforecasting bias for such locations. Also, both MEPS20 and GEPS are more prone to completely missing events. MEPS4 improves upon both with higher resolution, skill, reliability, and fewer missed events when terrain is a factor

BSS and reliability for winds repeated a diurnal pattern like several other parameters discussed previously. In particular, in Figure 4.31 the BSS, reliability, and resolution were best around noon local time at KDMA. BSS improved as the number of wind events increased, then at night around 1900 local BSS fell to negative values. During the night, MEPS4 suffered from an overforecasting bias when skill dropped to negative values. This bias particularly affected KAFF, KVBG, and KDMA but was present to some degree at all locations in the evening hours. It is likely that MEPS4 is failing to model the setup of a boundary layer inversion which helps to block high winds aloft from mixing to the surface. Also, since it was evident from section 4.5 that MEPS4 tended to over forecast lightning,

it is possible that MEPS4 is predicting higher winds associated with thunderstorm activity resulting in overforecasting of wind speeds.

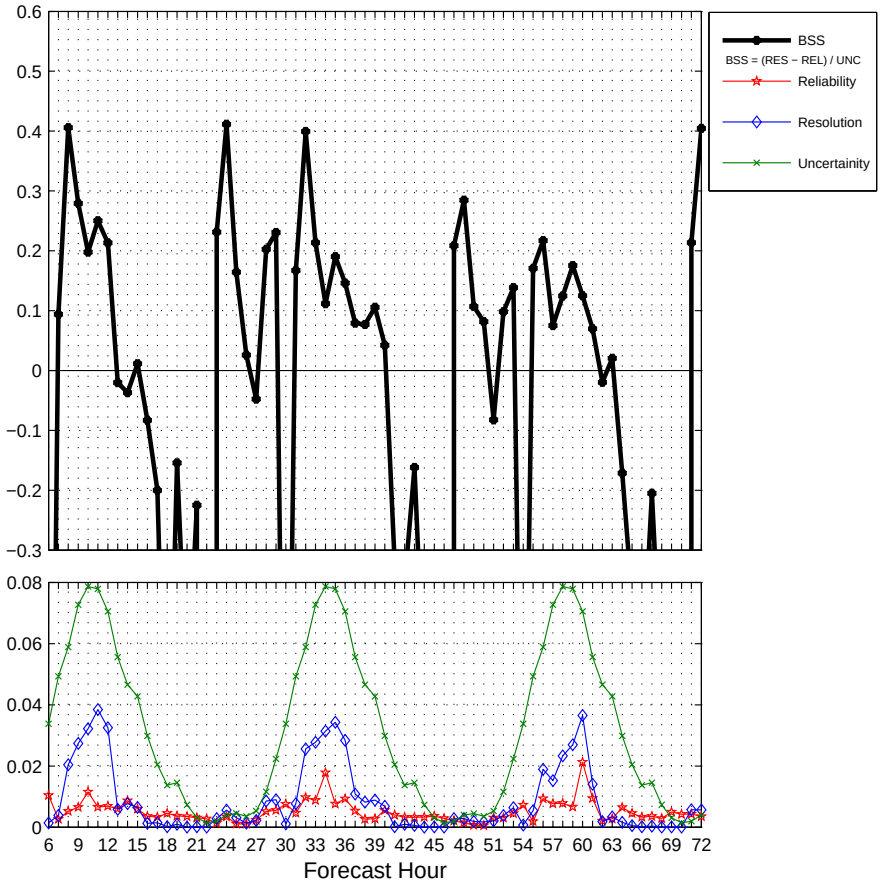


Figure 4.31. KDMA MEPS4 BSS for winds greater than 25 kt from Apr-Oct 2015.

For locations surrounded by flat terrain like those in central CONUS like KBLV and KEND, winds are not as difficult to accurately predict. Each EPS proved to have much more reliable and skillful forecasts for such locations. MEPS4, for example, had excellent reliability for 25 kt wind forecasts at KEND. Figure 4.32 shows the 8 hr MEPS4 forecast of 25 kt winds at KEND. The figure shows good resolution with a broad range of forecast probabilities from 0 to 100%. The forecasts closely match the zero-reliability line. Both factors result in a relatively high BSS of .545. Over the entire forecast period, the forecast probabilities average to the

following for bins 0 to 91-100% respectively proving MEPS4 had good reliability at KEND: 0.3%, 3.4%, 11.7%, 21.3%, 31.8%, 50.3%, 65.7%, 87.1%, 88.46%, 100% and 100%.

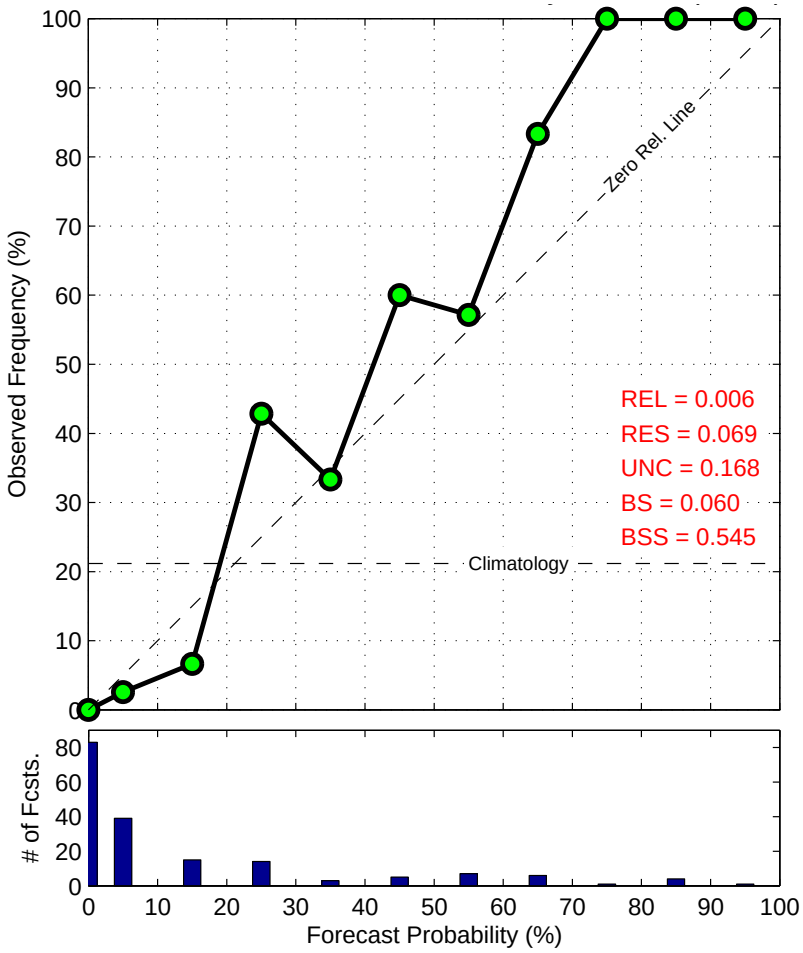


Figure 4.32. KEND MEPS4 8 hr reliability diagram for winds greater than 25 kt

At KEND, the BSS remained positive for most forecast hours. The scores in Figure 4.33 show that MEPS4 outperformed climatology significantly for the majority of the forecast. Most forecast hours achieved a score between .3 and .5, indicating that MEPS4 had good reliability and resolution despite relatively higher values of uncertainty. Some forecast hours scored negative, however, revealing biases in MEPS4 during certain times. The forecast hours that had negative skill were hours 20, 44, 45, and 68-71. These forecast hours coincide with times between 0300

and 0600 local time when winds are typically light. During these times, MEPS4 had a significant overforecasting bias. Each site displayed varying degrees of bias during particular times of the day. For some like KEND, the bias appeared just over a small window of time. For others, the bias lasted longer, creating longer windows of negative skill values.

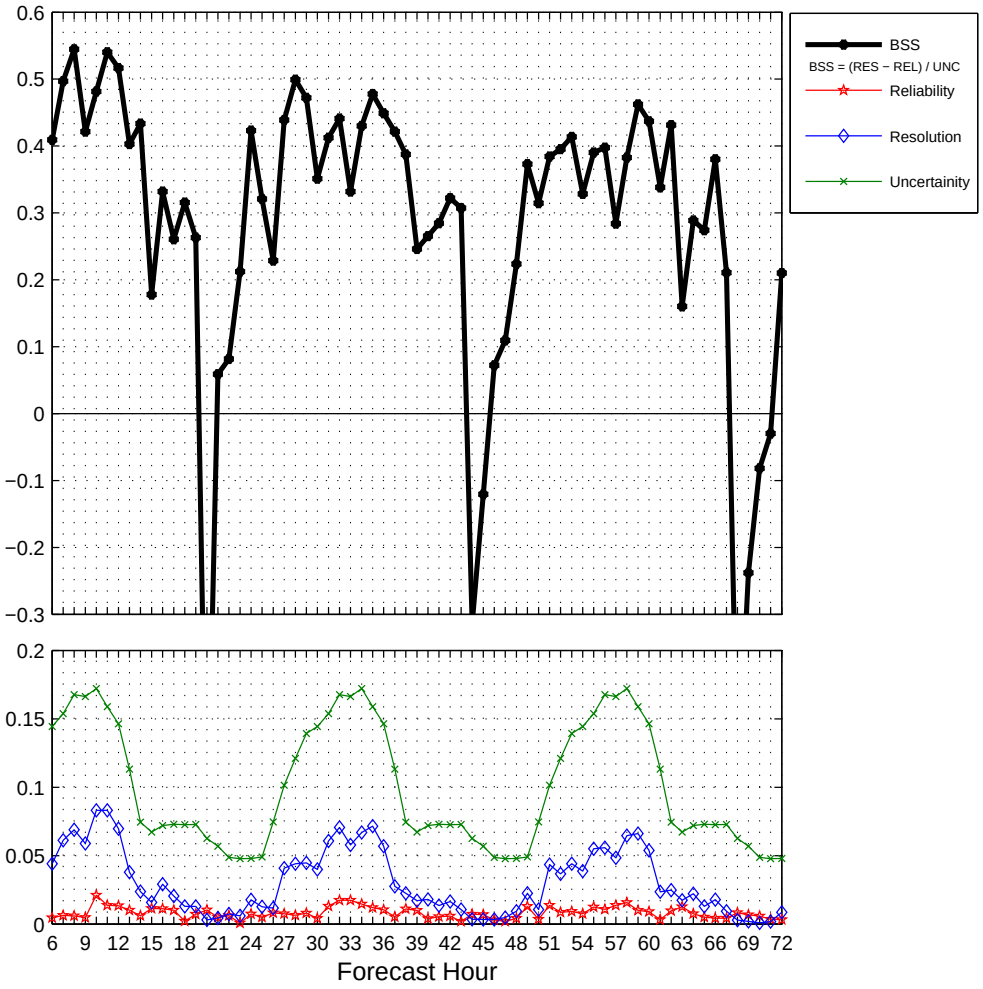


Figure 4.33. KEND MEPS4 BSS for winds greater than 25 kt from Apr-Oct 2015.

A glaring example of the overforecasting of winds in shown in Figure 4.34. The forecast is from MEPS4 for the 7 hr forecast of winds greater than 25 kt at KVBG. In this example, the event never occurred, even though the EPS predicts much higher than zero probabilities.The majority of bins lie outside the area of skill,

resulting in a large negative BSS. Typically, forecasts with significant bias occurred during overnight hours when winds are relatively weak, and gusts are rare.

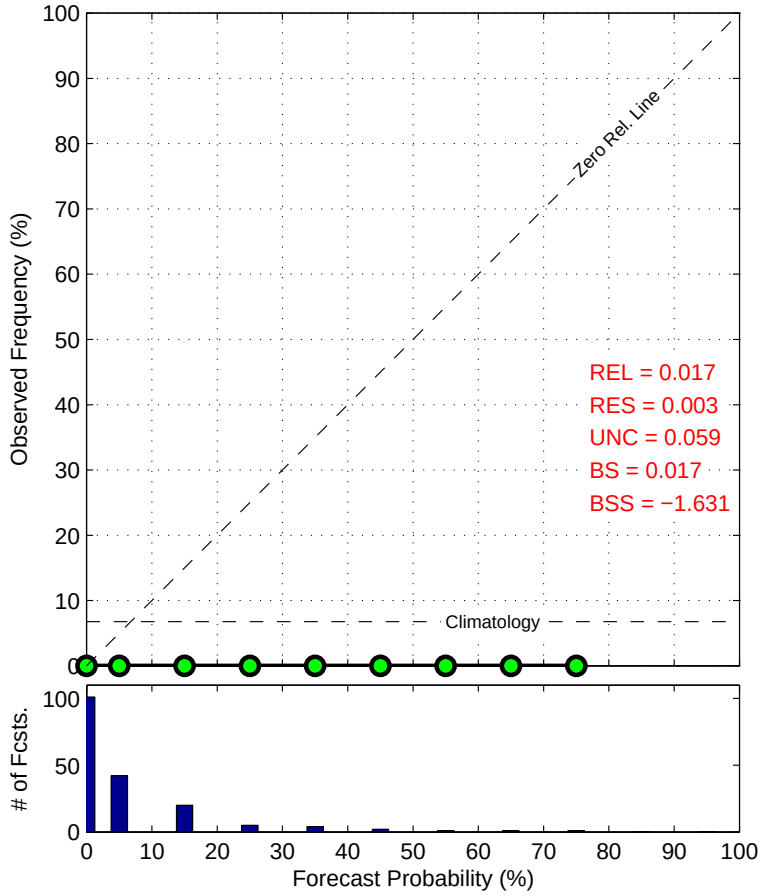


Figure 4.34. KVBG MEPS4 7 hr reliability diagram for winds greater than 25 kt.

Table 4.4 contains a summary of the average skill and percent positive skill over the first 72 hr for winds greater than 25 kt. The table also includes average resolution and reliability over the entire forecast period of the EPS. Some data is left out of the table due to that fact the majority of forecast hours for that particular location and EPS did not produce a large enough sample of forecasts to provide any reliable data. The table reveals some interesting statistics that contradict the conclusion that increasing horizontal resolution results in a more positive BSS. MEPS4, with the highest resolution, did not always produce the

greatest BSS. There were 5 locations in which MEPS20 or GEPS outperformed MEPS4 BSS: KAFF, KDMA, KGRK, KHMN, KVBG. Therefore, the theory that higher horizontal resolution leads to better BSS is not absolute.

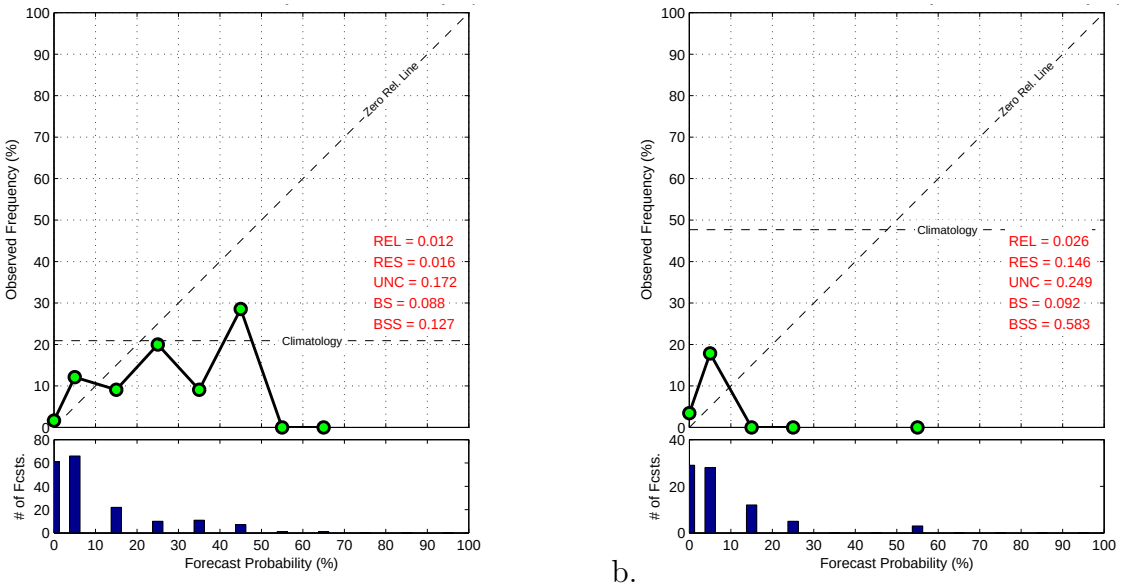
With several factors affecting BSS, it is possible to determine which may have led the result of MEPS4 not having the highest BSS. At the 5 locations with this result, Table 4.4 shows that MEPS4 had far better reliability than the other models. Where MEPS4 did worse was with the resolution component of the BSS. Figure 4.35 is an example of how MEPS20, a model with more coarse horizontal resolution, could have had a better BSS than MEPS4 even with less reliability. Figure 4.35(a) is the MEPS4 25 kt wind forecast at KAFF, and Figure 4.35(b) is the MEPS20 forecast for approximately the same local time. Figure 4.35(a) has a BSS of .127 and Figure 4.35(b) has a skill of .583, although it is not obviously clear why the MEPS20 has such a significant increase in BSS.

The difference in BSS can be explained by the distinct difference in the climatological frequency between the two forecasts. For MEPS20 climatology is almost 50% while it is just 20% for MEPS4. One reason for the difference in climatology is the different forecast interval length from each EPS. Another reason is the fact that the climatology for the MEPS20 forecast averages the first three months of the study as opposed to a six-month period for MEPS4. The effect of different climatologies is evident from the differences in the resolution between the two forecasts. For MEPS4, the resolution is .016 and for MEPS20, it is .146. Resolution measures how far the observed frequencies deviate from climatology. Higher resolution results in more skill since it indicates the model is forecasting situations of 25 kt winds that occur either more or less frequently than climatology. Figure 4.35(a) shows most of the observed frequencies are close to climatology causing low resolution. However, for MEPS20, the observed frequencies were much

Site	EPS	Avg Reliability	Avg Resolution	Average Skill	Percent Positive Skill
KAFF	GEPS	-	-	-	-
	MEPS20	.0140	.0390	.424	74
	MEPS4	.00650	.00998	.211	75
KBLV	GEPS	.00302	.00280	.086	58
	MEPS20	-	-	-	-
	MEPS4	.00182	.00201	.139	51
KLRF	GEPS	-	-	-	-
	MEPS20	-	-	-	-
	MEPS4	.00221	.00106	.151	15
KDMA	GEPS	-	-	-	-
	MEPS20	.0182	.0119	.204	39
	MEPS4	.00516	.00876	.174	58
KEND	GEPS	.0384	.0308	.093	58
	MEPS20	.0173	.0418	.315	61
	MEPS4	.00817	.00309	.345	89
KGRK	GEPS	.0105	.00629	.143	25
	MEPS20	.0173	.0123	.217	78
	MEPS4	.00448	.00629	.175	57
KHMN	GEPS	.0506	.0120	.118	25
	MEPS20	.0190	.0275	.160	65
	MEPS4	.00805	.0124	.126	51
KLFI	GEPS	.00888	.0171	.292	100
	MEPS20	.00282	.00479	.356	82
	MEPS4	.00756	.0270	.559	100
KLSV	GEPS	-	-	-	-
	MEPS20	-	-	-	-
	MEPS4	.00864	.00159	.157	60
KOFF	GEPS	.00914	.0135	.287	100
	MEPS20	.0124	.0163	.340	52
	MEPS4	.00694	.0135	.287	48
KVBG	GEPS	.00188	.00841	.389	100
	MEPS20	.0131	.0183	.397	48
	MEPS4	.00741	.00786	.288	40
KWRI	GEPS	.00473	.00583	.182	67
	MEPS20	.00402	.00406	.261	61
	MEPS4	.00560	.00997	.320	73

Table 4.4. Table of the average BSS and 72 hour percent positive skill values for winds greater than 25 kt. The table also shows average resolution and reliability for the whole forecast period for each EPS.

less than the 50% climatology leading to more resolution. The higher resolution score of MEPS20 gives the forecast a higher BSS, despite the fact that the MEPS20 forecast seems to be more affected by an overforecasting bias. The Brier score for MEPS4 is closer to zero however which is better, reflecting EPS performance in a different way. In this example, and for many others among the cases where MEPS4 did not have the best BSS, it was often a result of a different climatology. Other performance metrics like reliability and Brier score indicated that increasing resolution improved the performance of 25 kt wind forecasts.



a. b.

Figure 4.35. KAFF MEPS reliability diagram comparison for winds greater than 25 kt. a: MEPS4 35 hr forecast valid from 22-23z. b: MEPS20 42 hr forecast valid from 20-23Z.

4.7 Effect of MEPS Modifications Implemented in July 2015

In late July, MEPS4 and MEPS20 underwent significant changes stated previously in Section 3.3. Because changes to the length of the forecast period in MEPS20, it was feasible to separate the April through October data set into 2 data sets beginning and ending on the day PEP bulletins changed format. Having two data sets allows for a comparison between the two versions of MEPS20. The old

version contains 15 WRF members run simultaneously with a forecast starting at 6Z. The new version is an ensemble of 15 WRF versions initialized consecutively at two hr intervals up to the most current member that begins at 12Z. The following will demonstrate effects on skill and reliability between the two versions.

The data displayed a variety of results, but in the end, supported the conclusion that MEPS20A (MEPS20 after July 17, 2015) BSS and reliability increased over MEPS20B (MEPS20 before July 17, 2015) for the majority of parameters and locations. One case is shown in Figure 4.36. Where MEPS20A significantly increased the BSS of ceiling forecasts at KGRK, with sizable increases at around 12Z. Besides KGRK, there were nine other locations out of 17 where the average BSS of 3kft ceiling forecasts improved over MEPS20B.

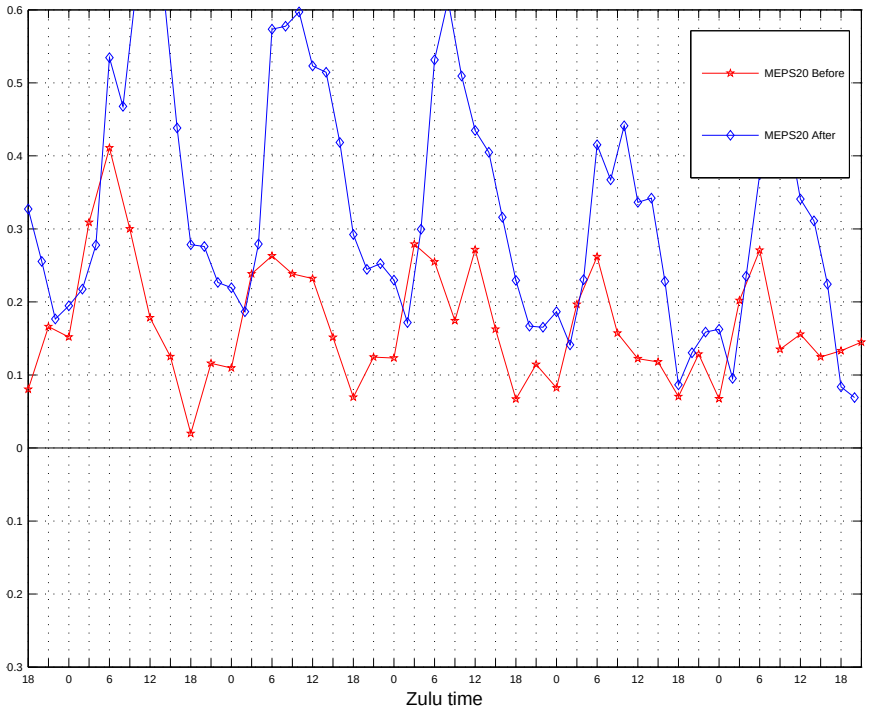


Figure 4.36. KGRK BSS comparison between MEPS20 before and after model update for ceilings less than 3kft.

In other instances, the BSS of MEPS20A was dramatically smaller than

MEPS20B. For example, at KAFF, the average positive BSS for visibility less than 5mi forecasts from MEPS20B was .24, which dropped profoundly to .03 for MEPS20A. However, at the same time the Brier score (the measure of error in probabilistic forecasts) and reliability for MEPS20A were better. The question arises of what is causing the evaluation metrics to contradict each other. The answer lies in the climatology used in the BSS.

A problem with using BSS to compare models from two different time periods is that the climatology for each forecast hour is often not the same. Different climatologies take away the control aspect of using climate as a reference forecast strategy in the BSS. The BSS differences shown in Figure 4.36 are possibly a result of variable climatology instead of from the accuracy of the model itself.

To clarify the cause of changes to BSS, Figure 4.37 takes a closer look at specific forecasts of 3kft ceilings at KGRK from both versions at a time where BSS changed significantly. Each diagram is valid at 12z. 4.37(a) is from MEPS20B, showing substantial underforecasting bias as demonstrated in Section 4.2. Figure 4.37(b) is from MEPS20A with less bias. Climatology is at about 25 percent for MEPS20A as opposed to nearly 43 percent for MEPS20B. The higher climatology of MEPS20B has the following effect: the observed frequencies of MEPS20B deviate further from climatology, resulting in a greater value for resolution. MEPS20A, on the other hand, has a better Brier score and reliability value. Both Brier score and reliability are not based on the value of climatology, therefore, the different climatology percentage has no direct influence. Since climatology in this case does not represent a control factor, reliability and Brier score are the less biased metrics than resolution and BSS as indicators of forecast accuracy. Therefore, the improved Brier score and reliability indicate that MEPS20A indeed improved the accuracy of 3kft ceiling forecasts at KGRK without influence from different climatology forecasts.

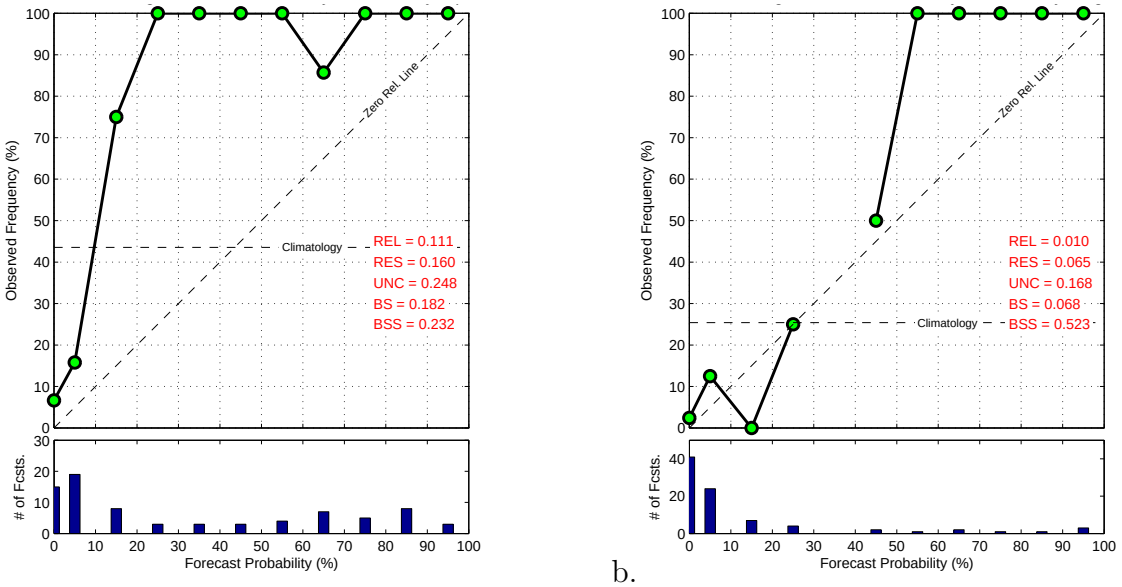


Figure 4.37. KGRK reliability diagrams for ceilings less than 3kft: a: MEPS20B 54 hr forecast. b: MEPS20A 48 hr forecast. Both forecasts are valid for 12Z local time.

The issues caused by variable climatology means that BSS is not by itself an indicator of forecast accuracy and quality in comparisons between MEPS20B and MEPS20A. Brier score is better for comparisons, and shown in Figure 4.38 is the Brier score for forecasts of visibility less than 5 sm at KVBG. Recalling that a lower Brier score equates to more accuracy, we see that MEPS20A improves between 6-9Z each day.

Table 4.5 shows performance metrics of additional visibility categories at KVBG and KAFF. The table suggests that at KVBG MEPS20A improved nearly all of the performance metrics for each category of visibility. Interesting results came from KAFF, with MEPS20A showing much lower BSS, but better Brier score and reliability than MEPS20B. The lower BSS is likely due to the fact that MEPS20A had poor resolution at KAFF bringing the BSS down with it. At KAFF and several other locations, it took more than BSS to show forecast quality.

For other parameters, it was clear that the changes made to MEPS20 overall improved Brier score and thus the accuracy of MEPS20 forecasts for most locations.

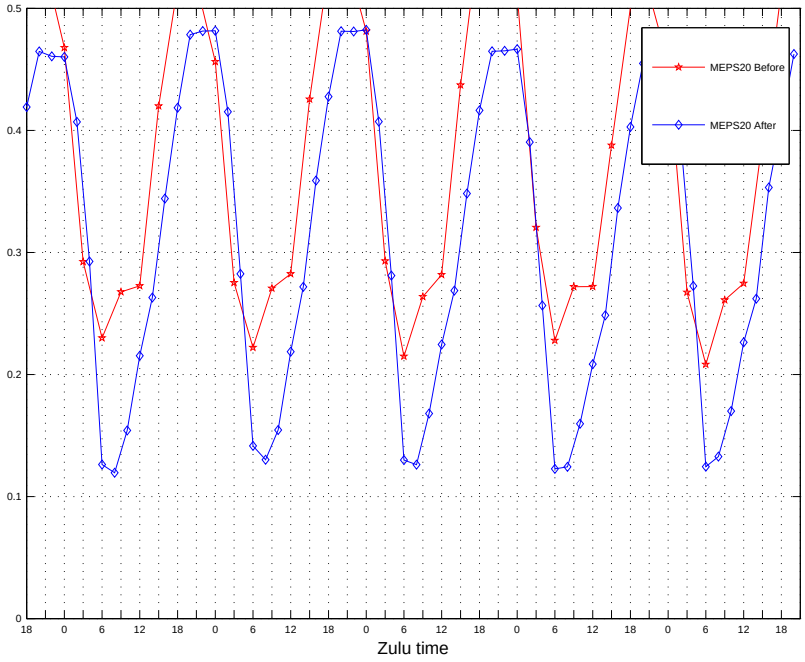


Figure 4.38. Comparison of MEPS20A and MEPS20B Brier score for visibility less than 5 sm forecasts at KVBG.

KVBG		5mi visibility	3mi visibility	1mi visibility
Average positive BSS	MEPS20 A.	.183	.211	.209
	MEPS20 B.	.00891	0.0	.0860
Average Brier score	MEPS20 A.	.041	.305	.230
	MEPS20 B.	.170	.340	.250
Average Reliability	MEPS20 A.	.112	.103	.0657
	MEPS20 B.	.147	.102	.0716
Average Resolution	MEPS20 A.	.0721	.0610	.0410
	MEPS20 B.	.0696	.0476	.0399
KAFF		5mi visibility	3mi visibility	1mi visibility
Average positive BSS	MEPS20 A.	.0298	.00692	-
	MEPS20 B.	.218	.0680	-
Average Brier score	MEPS20 A.	.0650	.0420	-
	MEPS20 B.	.180	.12	-
Average Reliability	MEPS20 A.	.00884	.0030	-
	MEPS20 B.	.0482	.0230	-
Average Resolution	MEPS20 A.	.00512	.00145	-
	MEPS20 B.	.085	.0281	-

Table 4.5. Comparison of KVBG forecasts of visibility from MEPS20 before and after model update. Averages are over the first 130 hr of forecasting.

Figure 4.39 summarizes the results with bar graphs of the mean Brier score for 3kft ceilings, 6 hr precipitation, lightning, and 25 kt winds. Each chart indicates than at most locations MEPS20A improved Brier score of forecasts over MEPS20B. MEPS20A scored better at 12 out of 17 sites for 3kft ceilings, 14 out of 17 for 6 hr precipitation, 16 out of 17 for lightning, and 10 out of 17 for 25 kt wind forecasts.

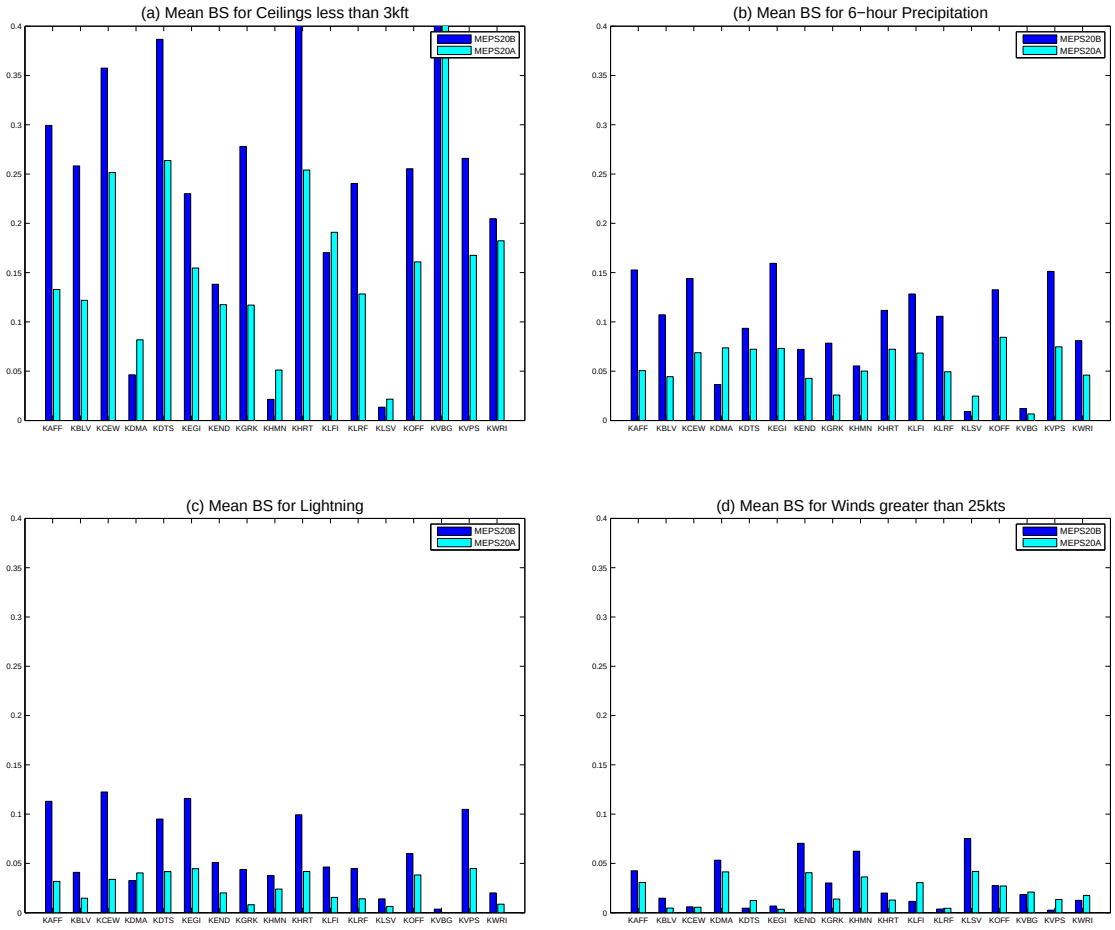


Figure 4.39. Mean Brier score comparison for (a) ceilings, (b) 6 hr precipitation, (c) lightning within 20 km (MEPS20B)/10 nm (MEPS20A), and (d) winds from MEPS20 before and after the model update.

The Brier score is an evaluation of EPS accuracy, but some caveats need to be explained due to the fact that forecasts from each model cover different time periods. The issue is that if there is a large difference in the number of weather

events between each period, then the model that makes fewer forecasts for fewer events will likely have a better Brier score. The reason is that the model with fewer events will forecast 0% more often which are 100% accurate if the event does not occur. Therefore, the Brier score in such a scenario may only reflect that one model had fewer events to forecast.

This factor had an influence on the Brier score in some cases, for example with lightning at KAFF. From April through July at KAFF there were 509 MEPS20 verifications of lightning within 20nm and just 193 from August through October. The effect was that MEPS20A put out more 0% forecasts that were most often correct, contributing to a better Brier score.

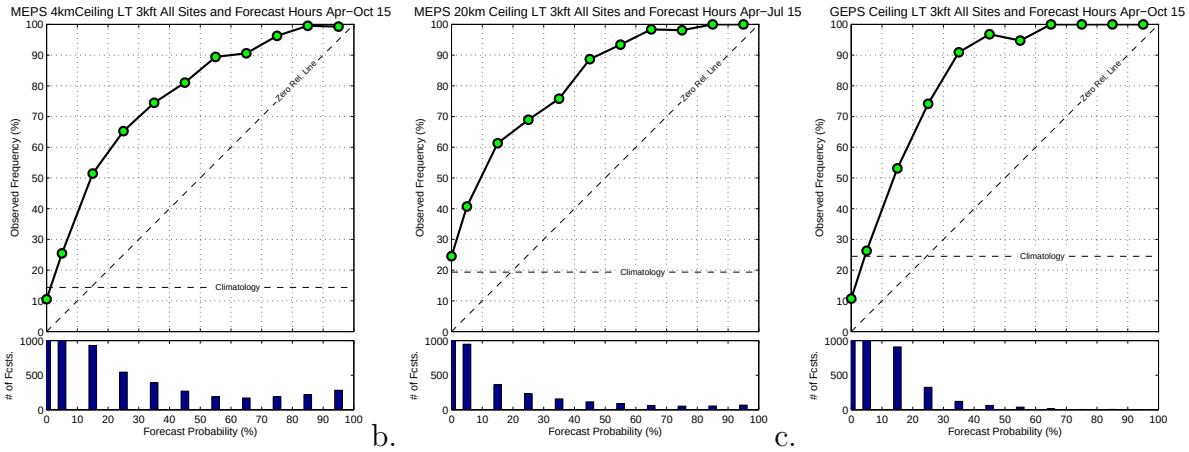
Although the result that MEPS20A improved Brier score over MEPS20B may not prove an increase in performance at this point, it can at least be said with confidence that model performance did not degrade after the change. One of the primary goals of the MEPS modifications is to provide more current updates to forecasters. Therefore, it is valuable to know that the changes resulted in, if not an improvement, at least consistent performance from MEPS20B.

4.8 Summary Reliability Diagrams for All Forecast Hours and Locations

So far this work has shown reliability diagrams for individual forecasts at various locations. They have been useful for showing trends in reliability within the forecast period and how it depends on location. It is also valuable to compile the data for all locations and forecasts hours to make a single reliability diagram that summarizes the performance of an EPS on a single parameter. This section presents reliability diagrams that evaluate observed frequencies for a forecast bin that contains the forecasts from all forecast hours and locations within the given bin interval. When this is done, the bar showing the number of forecasts in the 0% bin reached a very

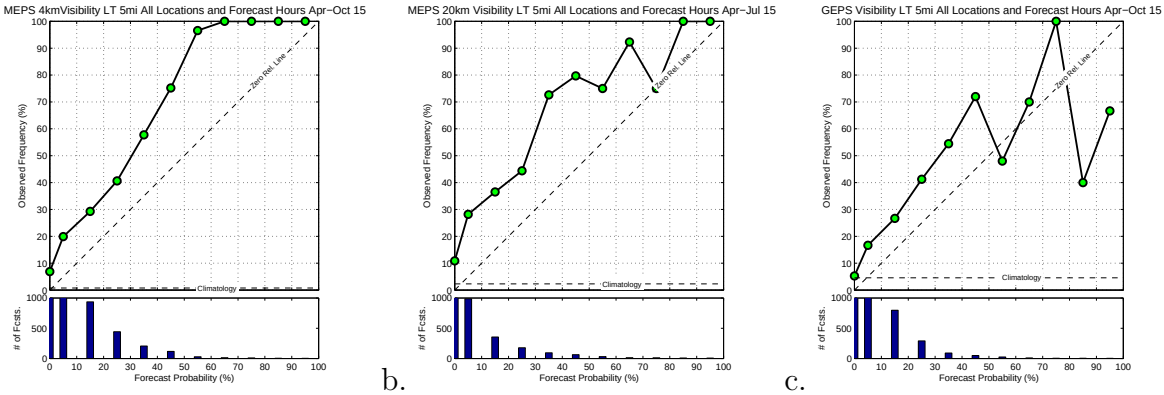
large amount of forecasts (typically around 8000) and overshadowed the number of forecasts in the bins greater than 0%. To make the number of forecasts in bins greater than 0% easier to see, the limit of the bar graph was set to 1000. In each figure there are three diagrams, MEPS4 (a) is on the left, MEPS20 (b) is in the middle, and GEPS (c) is on the right.

Overall, trends in observed frequencies shown in the diagrams in Figures 4.40 through 4.45 are much smoother than the diagrams of individual forecasts. Underforecasting bias is clearly evident in Figure 4.40 of the forecasts of ceilings less than 3kft. Similar trends were observed for the ceilings less than 1kft and 500ft categories. There is marginal underforecasting bias shown in the less than 5 sm visibility forecasts in Figure 4.41. Visibility less than 3 sm forecasts in Figure 4.42 were quite different, however, with MEPS4 being the most reliable and MEPS20 and GEPS having an apparent overforecasting bias in the higher probability bins. The number of forecasts in the higher probability bins in Figure 4.43 were still small even after compiling data from all locations and forecast hours. The precipitation summary reliability diagrams in Figure 4.44 showed that overall each EPS performed reliably. In Figure 4.44(a) MEPS4 observed frequencies for lightning closely match the forecast probabilities for the majority of bins, while it is clear there is some underforecasting bias in Figures 4.44(b &c) with MEPS20 and GEPS respectivley. In Figure 4.45, MEPS appears to have an overforecasting bias with winds greater than 25 kts while GEPS significantly underforecasts winds greater than 25 kts.



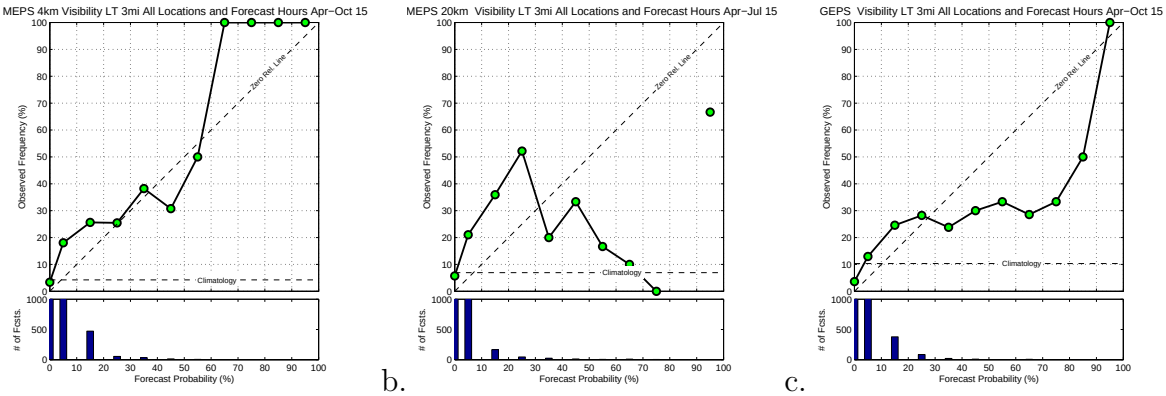
a. b. c.

Figure 4.40. Ceilings less than 3kft reliability diagram compiling data from all forecast hours and locations for each EPS.



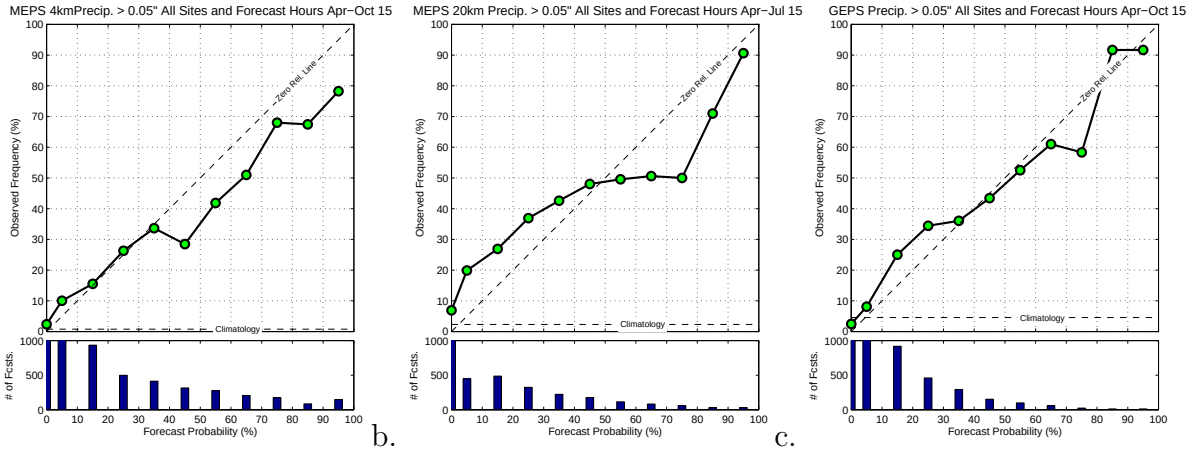
a. b. c.

Figure 4.41. Visibility less than 5 sm reliability diagram compiling data from all forecast hours and locations for each EPS.



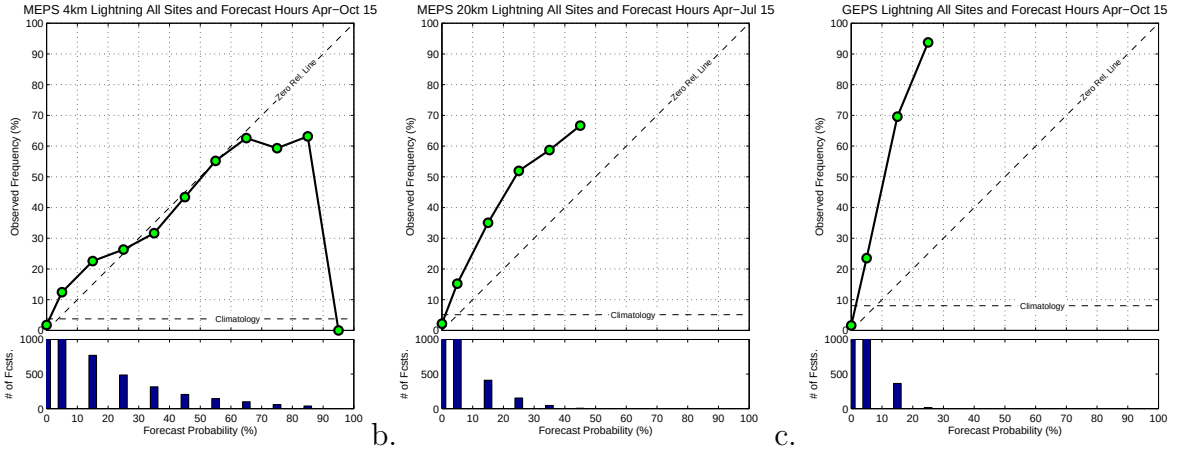
a. b. c.

Figure 4.42. Visibility less than 3 sm reliability diagram compiling data from all forecast hours and locations for each EPS.



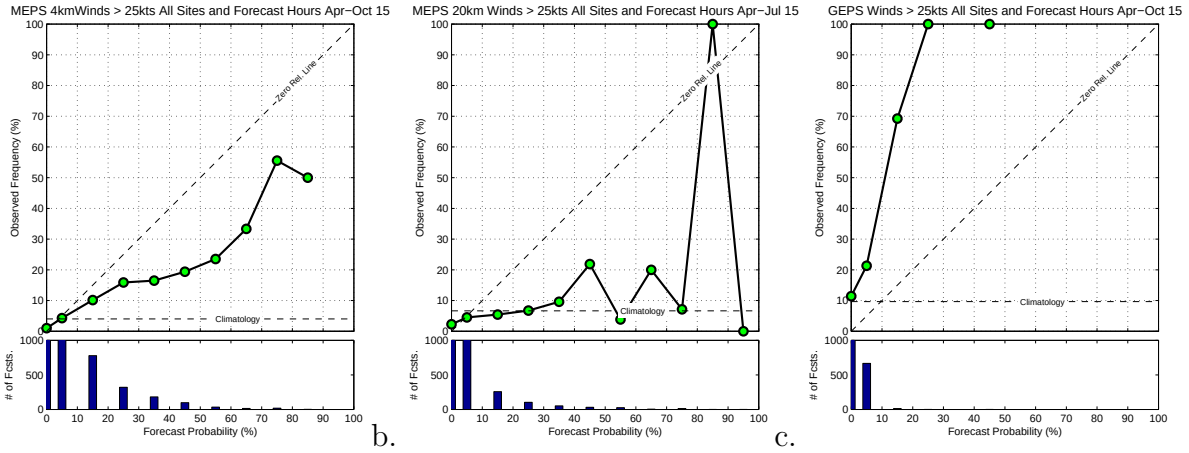
a. b. c.

Figure 4.43. Precipitation reliability diagram compiling data from all forecast hours and locations for each EPS.



a. b. c.

Figure 4.44. Lightning reliability diagram compiling data from all forecast hours and locations for each EPS.



a. b. c.

Figure 4.45. Winds greater than 25 kts reliability diagram compiling data from all forecast hours and locations for each EPS.

5. Conclusions and Future Work

5.1 Conclusions

With just one season of data, a variety of results from many locations, and numerous factors affecting the skill and reliability of probabilistic forecasts, it is difficult to make absolute conclusions about 557 WW's EPS. However, the trends evident from reliability diagrams and Brier scores from the current study support several conclusions drawn by Clements (2014). Additional this study provides evidence of new biases and tendencies of 557 WW's EPS. For the spring to early fall season (April-October) of 2015, the results support the following conclusions:

Ceiling and visibility thresholds are PEP parameters that have yet to be analyzed with BSS and reliability diagrams. The results indicate that, for ceilings, each EPS was, for the most part, capable of providing convincing value over climatology. However, for each EPS and at all locations, an underforecasting bias was significant and affected the majority of the forecast period. The bias was even more prevalent for the lower ceiling thresholds of 1kft and 500ft. Each EPS had poor performance with ceilings at Vandenberg, indicating that probability algorithms have large issues with forecasting low ceilings associated with marine layer fog and stratus. EPS skill improved with finer horizontal resolution. Underforecasting bias also affected visibility but proved to be more dependent on location. Bias also increased for the lower thresholds of 3 and 1 sm visibility. In fact, the forecast probabilities for the 3 and 1 sm often did not exceed 30-40% with most forecasts in the 1-10% bin, but the observed frequencies were often much higher. Decreasing EPS horizontal resolution had some impact on the average positive skill, but the largest impact was on the percentage of the forecast that had positive skill. MEPS4 demonstrated large improvements to the skillful percent of the forecast.

For precipitation, Clements again found that EPS performance was also improved with increasing resolution. The BSS in 2013 tended to be lower at locations where precipitation is mostly caused by small-scale factors, and higher when precipitation is induced synoptically. The same tendencies were shown in the current study. When terrain or small-scale processes like sea-breezes were involved in the development convective precipitation, the BSS was lower compared with locations largely influenced by synoptic systems. This supports the theory that the cumulus and precipitation parameterization schemes from MEPS20 and GEPS are sufficient for depicting precipitation for large-scale events like frontal systems, but lacking with respect to small scale convective systems in comparison with MEPS4. GEPS, capable of forecasting out to 240 hours, had BSS that began to degrade, typically, after 72 hr due to a decrease in the resolution component as the forecasts tended more towards climatology.

For lightning forecasts, Clements (2014) showed higher resolution improved BSS, but each model still had a tendency to overforecast lightning during times when climatological lightning frequency is low. The current study supported the conclusion of increasing skill with increasing resolution, and examples of overforecasting bias also existed. The significance of the bias depended on location. Sometimes overforecasting bias was most significant in the morning, just before the peak expectancy of lightning, or in the evening, after factors supporting thunderstorm development break down, and sometimes present during both periods. At the 5 Florida locations, for example, MEPS4 overforecasted lightning during both morning and evening hours. MEPS20 tended to alleviate biases shown from MEPS4, but MEPS4 on average produced higher Brier skill averages. It is important to note that limited amounts of data may have made biases look more significant when shown in reliability diagrams for specific forecast hours. Evidence

from the summary reliability diagrams that compile data from all forecast hours and locations show that MEPS4 was overall reliable. So it is possible that biases evident from specific forecast hours in MEPS4 may have been reduced with more data. The different length of forecasts periods from each EPS may have impacted results as well. Lightning by nature is sporadic, so over a one hour period the possibility of a false alarm is more likely versus a six hour period that lasts much longer. This may explain why the BSS for MEPS4 sometimes dropped significantly at times when lightning was infrequent.

Wind results also match Clements' results at the additional locations. In the current study, 35 kt wind events were too rare to produce any meaningful results. However, for 25 kt winds, BSS and reliability diagrams indicate that model performance improved with increasing horizontal resolution. As model resolution decreased, skill decreased and missed wind events occurred more frequently. GEPS had a tendency to underforecast at locations surrounded by complex terrain. An overforecasting bias was present in MEPS at some sites that repeated a diurnal pattern and was most significant overnight. The bias was clearly shown from the reliability diagrams that summarized all locations and forecast hours, with MEPS20 having the most significant overforecasting bias.

Since MEPS20 data had to be split into two data sets from April-July and July-October due to the implementation of changes described in Section 3.3, the current study did a comparison of reliability and skill before and after the changes. Results indicated that the Brier score improved overall, and the Brier skill improved sometimes after the change. However, the seasonal variability between the two periods makes the results less conclusive. The results do seem to indicate that skill and reliability did not degrade in any significant way after the changes to MEPS.

5.2 Future Work

The main limiting factor in this study was due to a lack of data. Often, several bins from the reliability diagrams had 10 or fewer forecasts. A small number of probability forecasts in a bin increases the margin of error and makes and the indication of possible biases in the EPS to be less conclusive. Jolliffe and Stephenson (2012) showed a way to quantify the margin for error on a reliability diagram based on the number of forecasts within a bin, which may be useful to depict on a future study with reliability diagrams. It would also be valuable to apply a similar analysis to BSS diagrams through depicting some sort of confidence interval to show the margin of error on BSS when sample sizes are low. In any case it would be better to have a study spanning several seasons with no model changes. This would be much more conclusive in regard to biases and levels of skill in the EPS. An accessible database of PEP bulletins for a reasonable amount of locations would be a good start to enabling the use of larger datasets.

So far, studies on EPS reliability and skill have been just for the summer season. There would be much value from analyzing the data from a winter season. Non-convective wind events occur much more frequently in the winter, especially downslope wind events. Snow forecasts can also be analyzed. There also may be sufficient amounts of data to test the BSS for the 35 kt wind category for the winter season.

Another valuable analysis would be to examine the reliability and skill of more specific weather regimes. In this thesis the results have summarized weather over an entire season, but it would be useful to only select data when certain kinds of weather occurred to observe the model performance on that particular type of weather event. Clements (2014) did this with tropical cyclones forecasts over Kadena AB, Japan by only evaluating data during the presence of a tropical cyclone.

This type of evaluations could be expanded upon for other types of weather.

If model data from each EPS can be separated and analyzed individually, it may be possible to create a rank histogram. The rank histograms would give some insight into how the ensemble members are performing relative to each other. It would indicate whether there is too little or too much spread between ensemble forecasts. A study on impacts of standard deviation between ensemble members would also be valuable. Higher standard deviation indicates ensemble forecasts are spread farther apart. It would be interesting to see if that may have some correlation with forecast accuracy.

As Mason (2004) discussed, negative BSS values can hide valuable information content on forecast quality. It may be beneficial to use a random guessing strategy instead of climatology as a reference strategy for the calculation of the BSS. The strategy would be especially applicable in comparing models over different seasons, ensuring that a different climatology would not skew results.

Bibliography

Arpe, B. K., A. Hollingsworth, M. Tracton, A. Lorenc, S. Uppala, and P. Kållberg, 1985: The response of numerical weather prediction systems to fgge level iib data. part ii: Forecast verifications and implications for predictability. *Quarterly Journal of the Royal Meteorological Society*, **111 (467)**, 67–101.

Ban, R., 2007: Completing the forecast: Characterizing and communicating uncertainty for better decisions using weather and climate forecasts (2006). *Sixth Communications Workshop*.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Monthly weather review*, **78 (1)**, 1–3.

Brocker, J., 2012: Estimating reliability and resolution of probability forecasts through decomposition of the empirical score. *Climate dynamics*, **39 (3-4)**, 655–667.

Clements, W. B., 2014: Validation of the air force weather agency ensemble prediction systems. M.S. thesis, AFIT/ENP/14-M-04. School of Engineering and Management, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, (ADA598471).

Coles, S., J. Bawa, L. Trenner, and P. Dorazio, 2001: *An introduction to statistical modeling of extreme values*, Vol. 208. Springer.

Cote, J., S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth, 1998: The operational cmc-mrb global environmental multiscale (gem) model. part i: Design considerations and formulation. *Monthly Weather Review*, **126 (6)**, 1373–1395.

Done, J., C. A. Davis, and M. Weisman, 2004: The next generation of nwp: Explicit forecasts of convection using the weather research and forecasting (wrf) model. *Atmospheric Science Letters*, **5 (6)**, 110–117.

Eady, E., 1951: The quantitative theory of cyclone development. *Compendium of meteorology*, 464–469.

Eckel, F. A., J. G. Cunningham, and D. E. Hetke, 2008: Weather and the calculated risk: exploiting forecast uncertainty for operational risk management. *Air & Space Power Journal*, **22 (1)**, 71–83.

Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Weather and Forecasting*, **20 (3)**, 328–350.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus A*, **21 (6)**.

Ferro, C. A., and T. Fricker, 2012: A bias-corrected decomposition of the brier score. *Quarterly Journal of the Royal Meteorological Society*, **138 (668)**, 1954–1960.

Hamill, T. M., and S. J. Colucci, 1997: Verification of eta-rsm short-range ensemble forecasts. *Monthly Weather Review*, **125 (6)**, 1312–1327.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15 (5)**, 559–570.

Hoffman, R. E. K., 1983: Lagged average forecasting, an alternative to monte carlo forecasting. *Tellus*, **35A**, 100–118.

Hogan, T. F., and T. E. Rosmond, 1991: The description of the navy operational global atmospheric prediction system's spectral forecast model. *Monthly Weather Review*, **119 (8)**, 1786–1815.

Houtekamer, P. L., H. L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen, 2005: Atmospheric data assimilation with an ensemble kalman filter: Results with real observations. *Monthly weather review*, **133 (3)**, 604–620.

Hsu, W., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, **2 (3)**, 285–293.

Jolliffe, I. T., and D. B. Stephenson, 2012: *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons.

Kalnay, E., 2003: *Atmospheric modeling, data assimilation, and predictability*. Cambridge university press.

Kleist, D. T., D. F. Parrish, J. C. Derber, R. Treadon, W.-S. Wu, and S. Lord, 2009: Introduction of the gsi into the ncep global data assimilation system. *Weather and Forecasting*, **24 (6)**, 1691–1705.

Lacarra, J.-F., and O. Talagrand, 1988: Short-range evolution of small perturbations in a barotropic model. *Tellus A*, **40 (2)**.

Leith, C., 1974: Theoretical skill of monte carlo forecasts. *Monthly Weather Review*, **102 (6)**, 409–418.

Leith, C., 1978: Objective methods for weather prediction. *Annual Review of Fluid Mechanics*, **10 (1)**, 107–128.

Lisko, S., 2015: Afw-webs wiki: Operational afwa ensemble information. Retrieved on 29 September 2015.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *Journal of the atmospheric sciences*, **20 (2)**, 130–141.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus A*, **21 (3)**.

Lorenz, E. N., 1995: *The essence of chaos*. University of Washington Press.

Mason, S. J., 2004: On using climatology as a reference strategy in the brier and ranked probability skill scores. *Monthly Weather Review*, **132 (7)**, 1891–1895.

Murphy, A. H., 1973: A new vector partition of the probability score. *Journal of Applied Meteorology*, **12 (4)**, 595–600.

Reynolds, C. A., P. J. Webster, and E. Kalnay, 1994: Random error growth in nmc's global forecasts. *Monthly weather review*, **122 (6)**, 1281–1305.

Richardson, D., 2000: Skill and relative economic value of the ecmwf ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **126 (563)**, 649–668.

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, **130 (6)**, 1653–1660.

Stephenson, D. B., C. A. Coelho, and I. T. Jolliffe, 2008: Two extra components in the brier score decomposition. *Weather and Forecasting*, **23 (4)**, 752–757.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at nmc: The generation of perturbations. *Bulletin of the American Meteorological Society*, **74 (12)**, 2317–2330.

Toth, Z., E. Kalnay, S. M. Tracton, R. Wobus, and J. Irwin, 1997: A synoptic evaluation of the ncep ensemble. *Weather and forecasting*, **12 (1)**, 140–153.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. Wiley, 137–163 pp.

Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the national meteorological center: Practical aspects. *Weather and Forecasting*, **8 (3)**, 379–398.

Wallace, J. M., S. Tibaldi, and A. J. Simmons, 1983: Reduction of systematic forecast errors in the ecmwf model through the introduction of an envelope orography. *Quarterly Journal of the Royal Meteorological Society*, **109 (462)**, 683–717.

Wang, Y., S. Tascu, F. Weidle, and K. Schmeisser, 2012: Evaluation of the added value of regional ensemble forecasts on global ensemble forecasts. *Weather and Forecasting*, **27 (4)**, 972–987.

Warner, T. T., 2010: *Numerical weather and climate prediction*. Cambridge University Press.

Wergen, W., 1982: Forced motion in the tropics. *ECMWF Workshop on the Current Problems in Data Assimilation*.

Wilks, D. S., 2011: *Statistical methods in the atmospheric sciences*, Vol. 100. Academic press.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bulletin of the American Meteorological Society*, **83 (1)**, 73–83.