4.1 A SAMPLE SIZE SENSITIVITY TEST FOR MOS PRECIPITATION TYPE

Phillip E. Shafer *
Meteorological Development Laboratory
Office of Science and Technology
National Weather Service, NOAA
Silver Spring, MD

1. INTRODUCTION

Developing Model Output Statistics (MOS) (Glahn & Lowry 1972) guidance for rare events such as freezing precipitation often requires a long training sample in order to capture enough cases to obtain stable estimates of the regression parameters. Unfortunately, long samples of training data from a stable model often are not available, which can make forecasting rare events a challenge. Recently, about 29 years of reforecast data from the Global Ensemble Forecast System (GEFS) has become available (Hamill et al. 2013). The availability of this dataset provides an opportunity to assess the impact of sample size on the accuracy and skill of MOS guidance, particularly for rare events.

This paper describes a sample size sensitivity experiment for MOS precipitation type. Equations for the conditional probability of freezing, frozen, and liquid precipitation were developed from GEFS reforecasts with varying lengths of training data. A k-fold cross-validation was performed to assess the effect of sample length on the skill of the MOS guidance.

2. METHODOLOGY

2.1 *Observations*

MOS precipitation type guidance is developed from present weather observations at METAR sites (Allen & Erickson 2001a,b). Observations were examined for the period 1985-2013 which corresponds to the ~29-yr sample of GEFS reforecast data available for this test. In order to be used in the test, a site must report precipitation and must have had at least 50% of possible reports each year during the period. This limits the list of candidate stations to only those that are well-established and report present weather reliably, and also insures a consistent set of stations is available for any period one may wish to use for cross-validation. A set of 551 stations was retained, comprised of 506 CONUS stations, 26 Alaska stations, and 19 stations in Canada.

Present weather observations valid at 0000, 0600, 1200, and 1800 UTC were classified into one of three mutually exclusive categories: freezing, frozen, or liquid. A separate "null" category was used for cases when no precipitation of any type was reported or if the type of precipitation could not be determined. Thus, only precipitation cases of discernable type comprised the developmental sample. Table 1 lists the definitions of each MOS precipitation type category. As with previous MOS precipitation type developments, observations of sleet were classified as freezing and any mixture of liquid precipitation with snow was classified as liquid (Allen & Erickson 2001a,b; Shafer 2010; Shafer 2014). Freezing events are rare, comprising only 1.5% of precipitation cases over the CONUS and 0.5% of precipitation cases over Alaska.

2.2 *Predictors*

Predictor data for this sample size experiment was comprised of GEFS mean forecast fields archived on a 1-degree latitude / 1-degree longitude grid and interpolated to each of the 551 METAR stations. Several GEFS mean-based predictors were offered to the regression analysis. These include various thicknesses, temperature and wet bulb temperature at various levels, temperature advection, and a predictor based on the vertical profile of wet-bulb temperature. Geoclimatic information was incorporated based on logit transforms and gridded conditional relative frequencies of freezing, frozen, and liquid precipitation as predictors. As part of the most recent GFS MOS precipitation type development (Shafer 2010), logit 50% values were calculated at each station for several parameters that are good discriminators of precipitation type. These include 850-hPa temperature, 1000-850 hPa thickness, 1000-500 hPa thickness, and freezing level. The 50% values

*Corresponding author address:
Phillip E. Shafer, 1325 East-West Highway, Station 10434, Silver Spring, MD 20910-3283; e-mail: Phil.Shafer@noaa.gov

were subtracted from the GEFS mean forecast for each parameter to obtain a new "transformed" predictor that helps to capture localized effects that may not be well-resolved on the model scale. This allows data for all stations to be pooled together into one or more large regions for development, while still retaining station specificity in the equations. The ability to combine data into larger samples is critical when forecasting rare events such as freezing precipitation, since the number of cases in the sample is often limited. Since the equations were developed with GEFS mean-based predictors, any effects of the ensemble spread were not examined nor accounted for in the sample length tests presented here.

2.3 *Regression analysis*

Equations for the conditional probability of freezing, frozen, and liquid precipitation were developed for the 0000 UTC cycle for projections every 6 hours through 192 hours in advance. Equations were developed only for the cool season, defined as 1 September - 31 May. As with previous developments, multiple linear regression was used to derive the equations. This method, known as "Regression Estimation of Event Probabilities" (REEP), relates the binary predictands to a linear combination of predictor variables using a stepwise selection procedure (Miller 1964). The equations for each predictand (freezing, frozen, and liquid) were developed simultaneously; that is, the equations contained the same predictor variables but have different regression coefficients. The probability forecasts were normalized in a post-processing step to insure they are constrained to the 0 to 1 range and sum to 100%.

2.4 *Sample size experiments*

To evaluate the effect of sample size on the accuracy and skill of the GEFS MOS precipitation type guidance, equations were developed with varying lengths of training data as follows:

1) Test equations were developed with 1, 2, 3, 5, 10, and 15 years of daily forecasts, as well as sampling every third day from a 15-yr period (here one "year" is a 9-month cool season). The purpose of this latter test is to assess whether any benefit is realized from sampling over a longer period compared to using the most recent 5 years of data. One cool season was withheld as an independent sample.
2) Forecasts were generated for the withheld season using the equations produced from the sample length tests in (1).
3) A k-fold cross-validation was performed by repeating (1) and (2) over a period of 12 cool seasons, each time withholding a different season for testing. This procedure generated 12 seasons of independent forecasts for each sample length test.

For any given length of training data, the results can be influenced by how one chooses to develop the equations, be it by taking a single-station approach or by pooling stations into one or more regions. Pooling data into regions is often necessary when the event being forecast does not occur frequently enough at individual stations to obtain stable single-station equations, particularly for short training periods. To examine the sensitivity of the results to the development method used, the following additional tests were performed:

4) The test equations in (1) were developed by combining all stations into one large region – known as a "generalized operator equation (GOE)". This approach is expected to be most beneficial for the shorter sample length tests (i.e., 1 and 2 years), where the number of freezing and frozen cases in the sample is more limited.
5) The test equations in (1) were developed by grouping stations into 10 geographic regions over the CONUS and Alaska. The regions were chosen based on climatological and geographic similarity, and are the same as those used for MOS snowfall (see Cosgrove & Sfanos 2004).
6) Finally, for training periods of 5, 10, and 15 years, the test equations in (1) were developed with a single-station approach. For this test, the sample was restricted to stations that contained 5 or more cases of freezing precipitation and 10 or more cases of frozen precipitation.

3. RESULTS

3.1 *Sample length comparison*

The score most often used to assess the accuracy of multi-category probability forecasts is the Brier p-score (Brier 1950), which is essentially the mean squared error for the probability forecasts summed over each of the nominal binary events to which the probabilities relate (Wilks 2006). P-scores were calculated for the three-category

GEFS MOS guidance and for a reference climatology forecast. Here, climatology is simply the conditional relative frequency of freezing, frozen, and liquid precipitation for a 12-h period centered on the forecast valid times of 0000, 0600, 1200, and 1800 UTC. Scores were calculated for the aggregate of all 12 independent test seasons (see Section 2.4). Figure 1 shows the p-score percent improvement over climatology for the 3-category GEFS MOS precipitation type guidance. These results are for generalized operator (GO) equations. Curves are plotted for the 1, 2, 3, 5, 10, and 15-yr sample length tests. Only minimal differences in skill are seen for projections through 72 hours (Fig. 1a). For projections beyond 72 hours, the 1-yr sample length test performs the worst while sample lengths of 2 years or greater show similar skill. More sizeable differences in skill can be seen for projections beyond 144 hours (Fig. 1b). In the extended range projections, sample lengths of two years or longer have a clear advantage with the 15-yr test generally performing the best for most projections. Still, considerable clustering is present in the results, and for many projections it appears there is very little to be gained for training periods longer than 5 years.

3.2 *Generalized operator vs. regional*

A comparison of skill scores for GO and regional developments are shown in Figure 2 for the 5-yr and 15-yr sample tests. Overall, the regional development has better skill than GO, with differences of 3-5% seen for the early projections (Fig. 2a) and 2-3% in the extended range (Fig. 2b). If one wishes to make forecasts at stations then a regional development is usually the better choice as the equations will be more locally-tuned. However, for making gridded forecasts a GO approach alleviates the problem of boundary discontinuities between regions, and would produce forecasts that are more spatially consistent. Also plotted in Figure 2 are skill scores when sampling every third day from a 15-yr period (the equivalent of 5 years of training data). A slight advantage is seen in the extended range (Fig. 2b) for this test as compared to the 5-yr daily test, although the difference is very small.

3.3 *Single-station tests*

A comparison of results for single-station (SS), regional, and GO developments is shown in Figure 3 for training periods of 5 years (Fig. 3a) and 15 years (Fig. 3b). Note that stations which did not contain sufficient cases of freezing or frozen precipitation to produce an equation are not included in the single-station verification. The scores for regional and GO developments do not vary significantly with training sample length, while the scores for SS equations are much more sensitive to sample size. For the 5-yr test, SS has a slight advantage through 36 hours then performance degrades quickly and becomes even worse than GO after about 108 hours. The rapid degradation in performance for the SS equations (Fig. 3a) is likely due to over-fitting, as a 5-yr training sample for one station is insufficient to capture enough freezing and frozen cases to produce a stable equation. However, for a training period of 15 years (Fig. 3b), performance for the SS equations is significantly improved, with skill scores superior to regional equations through around 138 hours and about the same through the extended range. If a single-station approach is desired, the results in Figure 3 suggest that a significant benefit can be gained from having a long sample of retrospective forecasts from a stable model.

3.4 *Reliability*

Aside from forecast accuracy, it is important that probability forecasts be well-calibrated and reliable. Forecasts are deemed reliable when the average forecast probability and observed relative frequency of the event are roughly the same in each probability bin. A comparison of reliability for the 5-yr and 15-yr sample tests is shown in Figure 4 for the 72-h projection. Reliability is generally good for all three categories, with perhaps a slight over-confidence in the lower probability bins and under-confidence in the higher bins. Since more freezing events can be captured in a longer training sample, reliability for that category is slightly better for the 15-yr test (Fig. 4b) than for the 5-yr test (Fig. 4a). Reliability for the other sample length tests (not shown) are very similar.

4. SUMMARY & CONCLUSIONS

Equations for the conditional probability of freezing, frozen, and liquid precipitation types were developed from GEFS reforecasts for training periods of 1, 2, 3, 5, 10, and 15 years and tested using a k-fold cross-validation procedure. The sensitivity of the results to the development method (i.e. regional, single-station) also was examined. For generalized operator and regional equations, the results from cross-validation suggest that a training period of no less than 2 years and no more than 5 years is sufficient to develop skillful and reliable MOS guidance for precipitation

type. Little difference was found for training periods of 5 years (daily) and sampling every third day from a 15-yr period. In general, a regional development was found to outperform generalized operator for training periods of 5 years or longer. If a single-station approach is desired, the results suggest that a significant benefit can be gained from having a long sample of retrospective forecasts from a stable model.

5. ACKNOWLEDGMENTS

The author wishes to thank Bob Glahn, Matthew Peroutka and Kathryn Gilbert for their helpful suggestions for the manuscript. Appreciation is also extended to Tom Hamill for his suggestions and for providing the archive of GEFS reforecasts used in this experiment.

6. REFERENCES

Allen, R. L., and M. C. Erickson, 2001a: AVN-based MOS precipitation type guidance for the United States. *NWS Technical Procedures Bulletin* No. 476, NOAA, U.S. Dept. of Commerce, 8 pp.

____, and ____, 2001b: MRF-based MOS precipitation type guidance for the United States. *NWS Technical Procedures Bulletin* No. 485, NOAA, U.S. Dept. of Commerce, 8 pp.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review*, **78**, 1-3.

Cosgrove, R. L., and B. Sfanos, 2004: Producing MOS snowfall amounts from cooperative observer reports. Preprints, 20th Conference on Weather, Analysis, and Forecasting, Seattle, WA, Amer. Meteor. Soc., **6.3**.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203-1211.

Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast data set. *Bull. Amer. Meteor. Soc.*, **94**, 1553-1565.

Miller, R. G., 1964: Regression estimation of event probabilities. Tech. Rept. No. 1, Contract CWB-10704, The Travelers Research Center, Inc., Hartford, Conn. 153 pp.

Shafer, P. E., 2010: Logit transforms in forecasting precipitation type. Preprints, *20th Conf. on Probability and Statistics in the Atmospheric Sciences*, Atlanta, GA, Amer. Meteor. Soc., **P222**.

____, 2014: Experimental MOS precipitation type guidance from the ECMWF model. Preprints, *22nd Conf. on Probability and Statistics in the Atmospheric Sciences*, Atlanta, GA, Amer. Meteor. Soc., **6.4**.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. International Geophysics Series, Vol. 91, Academic Press, 627 pp.

Table 1. Definitions of MOS precipitation type categories.

Freezing	Frozen	Liquid
Freezing rain (FZRA)	Snow (SN)	Drizzle (DZ)
Freezing drizzle (FZDZ)	Snow showers (SHSN)	Rain/drizzle (RADZ)
Ice pellets (PL)	Snow grains (SG)	Rain (RA)
Any precipitation in combination with any of the above.		Rain shower (SHRA)
		Thunderstorm (TSRA)
		Any mixture of liquid precipitation with snow.

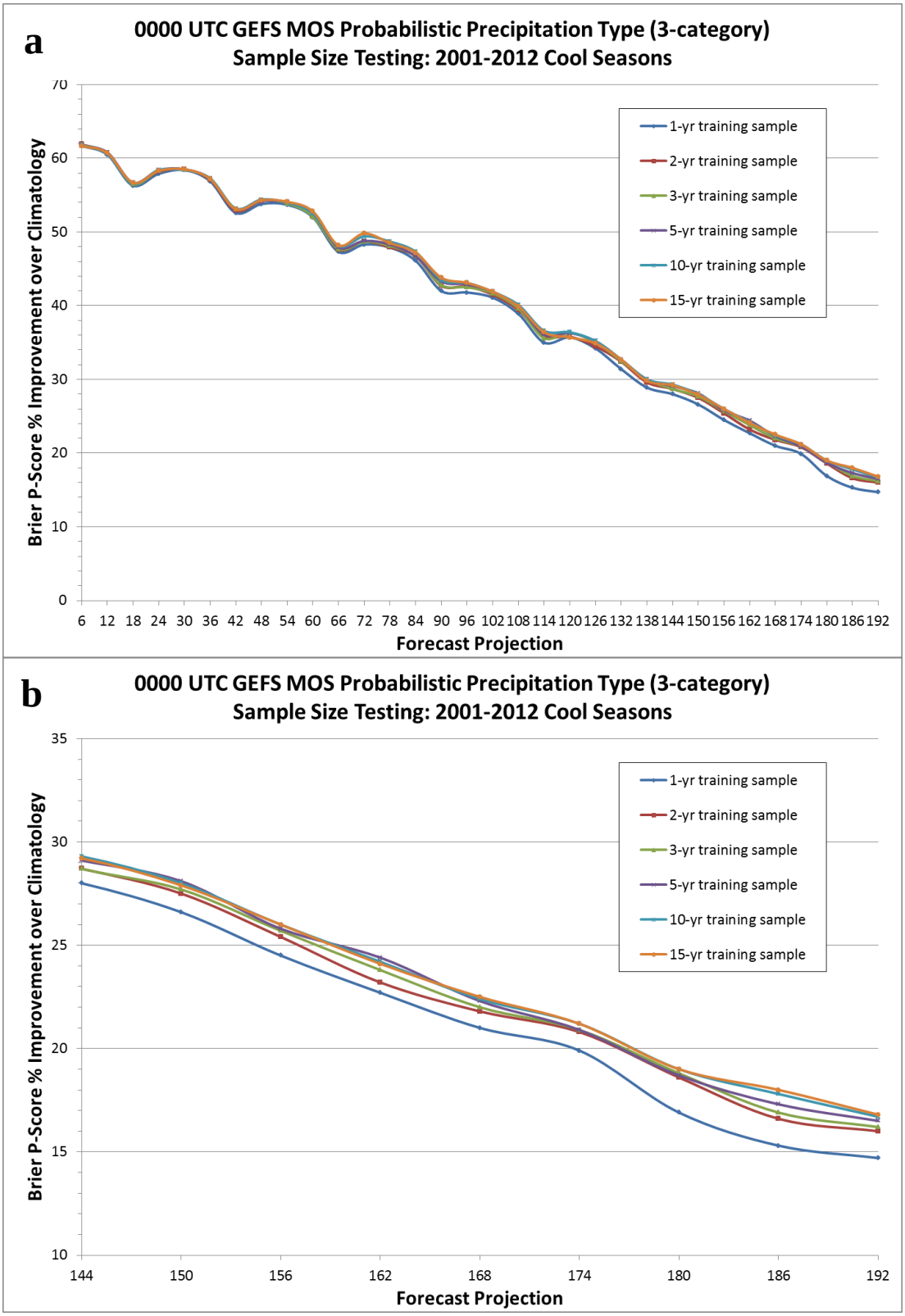
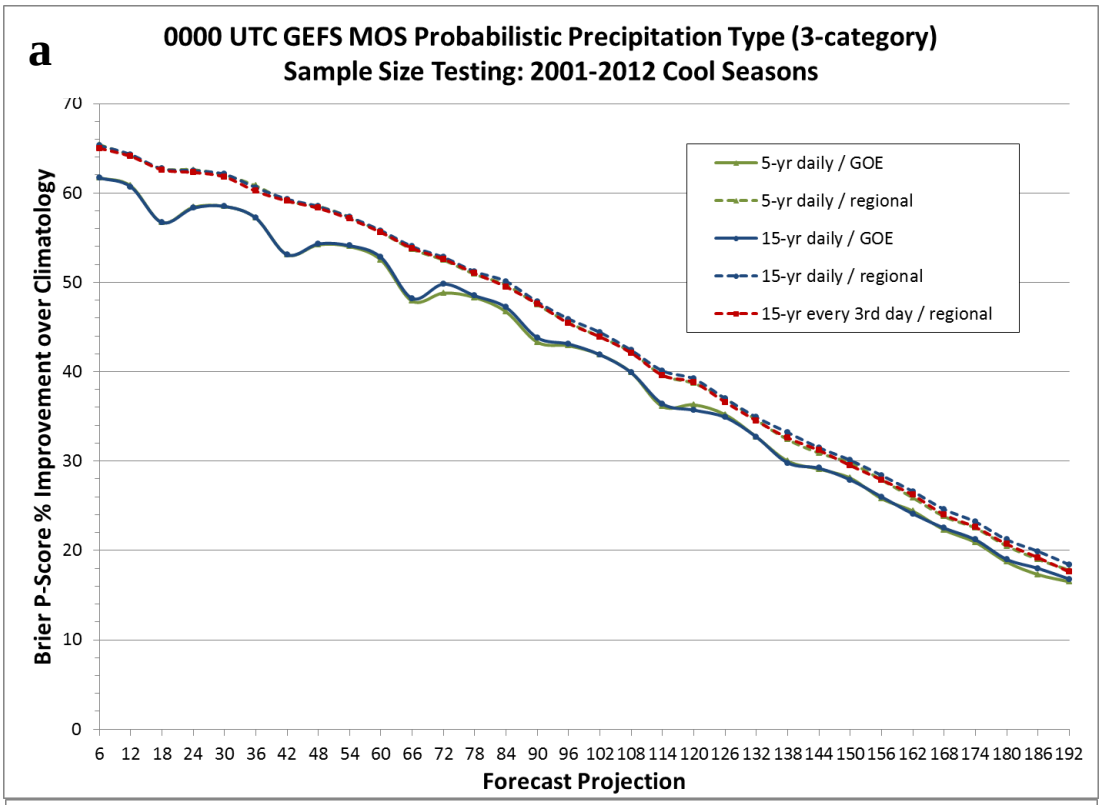


Figure 1. P-score percent improvement over climatology for 3-category GEFS MOS precipitation type guidance developed with varying lengths of training data. Results for all projections (a) and extended-range projections (b) are shown.



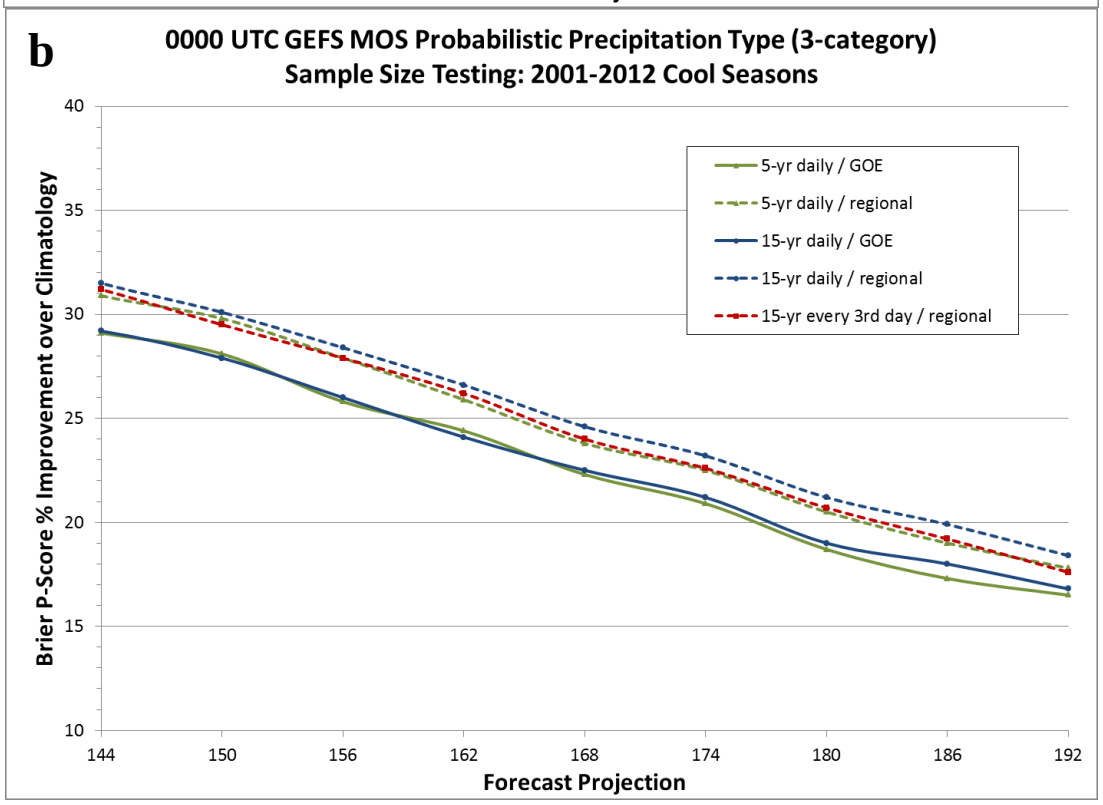
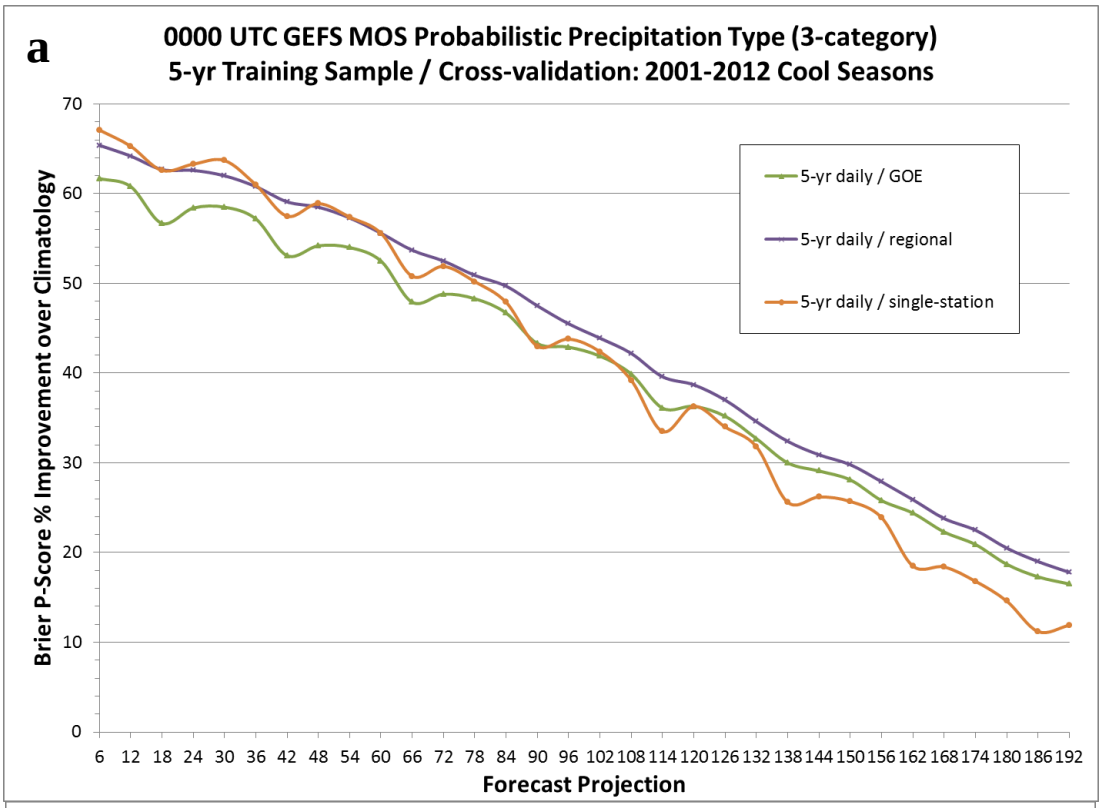


Figure 2. P-score percent improvement over climatology for 3-category GEFS MOS precipitation type guidance. Results for generalized operator and regional equations are shown for all projections (a) and extended-range projections (b).



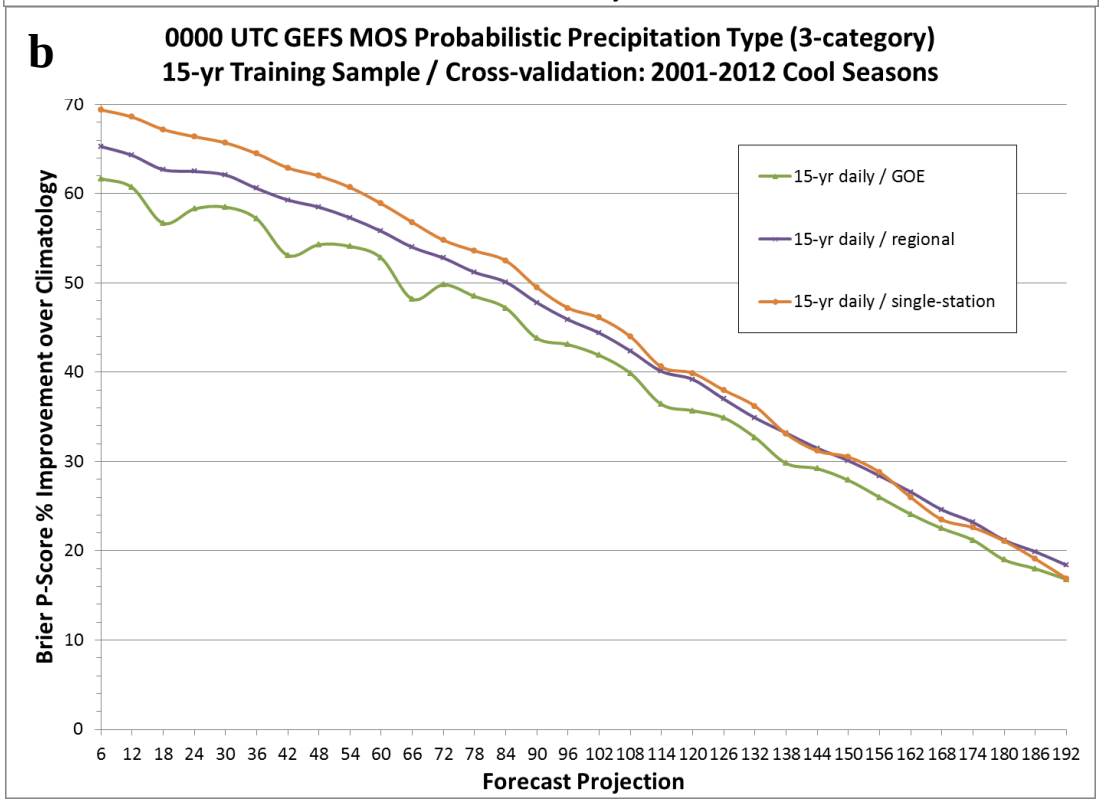


Figure 3. P-score percent improvement over climatology for 3-category GEFS MOS precipitation type guidance. Results for generalized operator, regional, and single-station equations are shown for sample sizes of (a) 5 years and (b) 15 years.

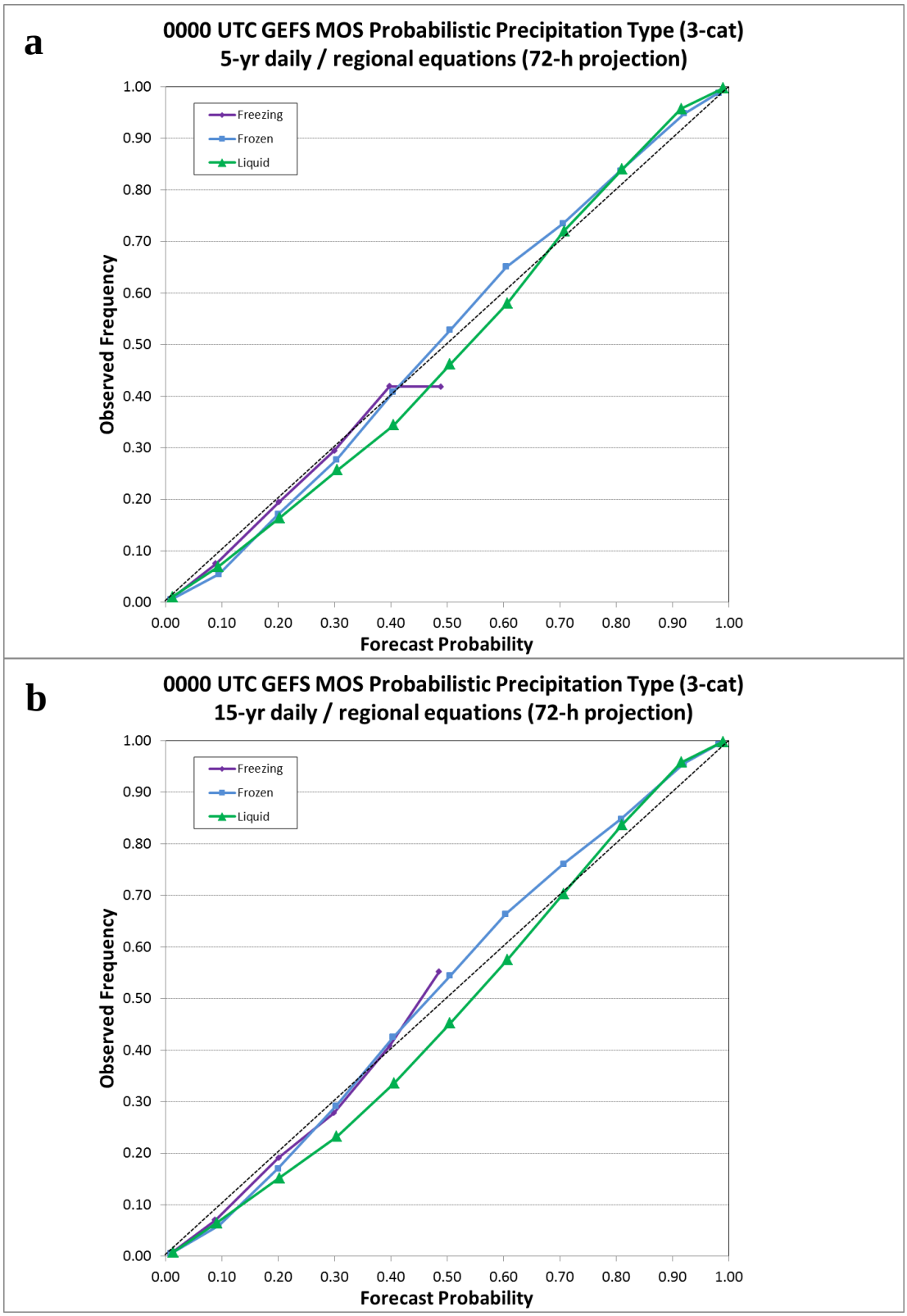


Figure 4. Reliability diagrams for 3-category GEFS MOS precipitation type guidance. Results for the 72-h projection are shown for (a) the 5-yr daily and (b) 15-yr daily tests.