Optimizing METAR Network Design for Verification of Cloud Ceiling Height and Visibility Forecasts

Eric Gilleland
National Center for Atmospheric Research
Research Applications Program
Boulder, CO 80307-3000
Email: ericg@ucar.edu

Keywords: cloud celing height, visibility, verification, coverage design

Abstract
Methods are given and explored for thinning METAR stations in order to make more meaningful verification analyses of cloud ceiling and visibility forecasts for use within the general aviation community. Verification of these forecasts is performed based on data from surface METAR stations, which for some areas are densely located and others only sparsely located. Forecasts, which are made over an entire grid, may be awarded or penalized multiple times for a correct or incorrect forecast if there are many METAR stations situated closely together. A coverage design technique in conjunction with a percent of agreement analysis is used to find an "optimal" network design in order to better score forecasts over densely located regions. Preliminary results for a network of 104 monitors in the New England area suggest that the removal of some stations is appropriate.

1 INTRODUCTION

Methods are explored here for optimizing the network design for verification of forecasts of cloud ceiling height and visibility for use by the general aviation (GA) community. Meteorological Aeronautical Report (METAR) data from surface stations are used for verification of these forecasts, but the placement of these stations may affect verification results. Specifically, in certain areas (usually urban areas and near airports) METAR stations are densely located, whereas in other areas (usually rural or mountainous areas) they tend to be sparsely located. For commercial aviation purposes such a network design may make sense. However, for GA use, where non-instrument rated pilots must avoid flying in any conditions of poor visibility regardless of whether the region is rural or not, this type of design may introduce bias not accounted for when computing the forecast bias. For instance, verification analyses in areas with densely located METAR stations may reward good forecasts–that cover several grid points in a region–too much for one good forecast. Converseley, poor forecasts might be penalized too much for one poor forecast. Subsequently, if there are stations located close to other stations, and these stations generally yield the same measurements; then it would make sense to use a thinned network that covers a region in such a way as to minimize any bias resulting from network design.

Section 2 gives details about the data. Section 3 presents the statistical methods used here. Section 4 are results, and Section 5 gives some discussion on the performance of these methods, and some thoughts on future and ongoing work.

2 METAR Cloud Ceiling Height and Visibility Data

Data used here are hourly data collected from METAR stations from January 1 to January 30, 2003. For this analysis, a subset of 104 stations in the New England area (Figure 1) is addressed. Gilleland (2004) and Fowler *et al* (2004) performed similar analyses on a subset of 48 stations in and around northern California. Some attention to this same subset is given here, but the primary focus of this paper is the New England subset.

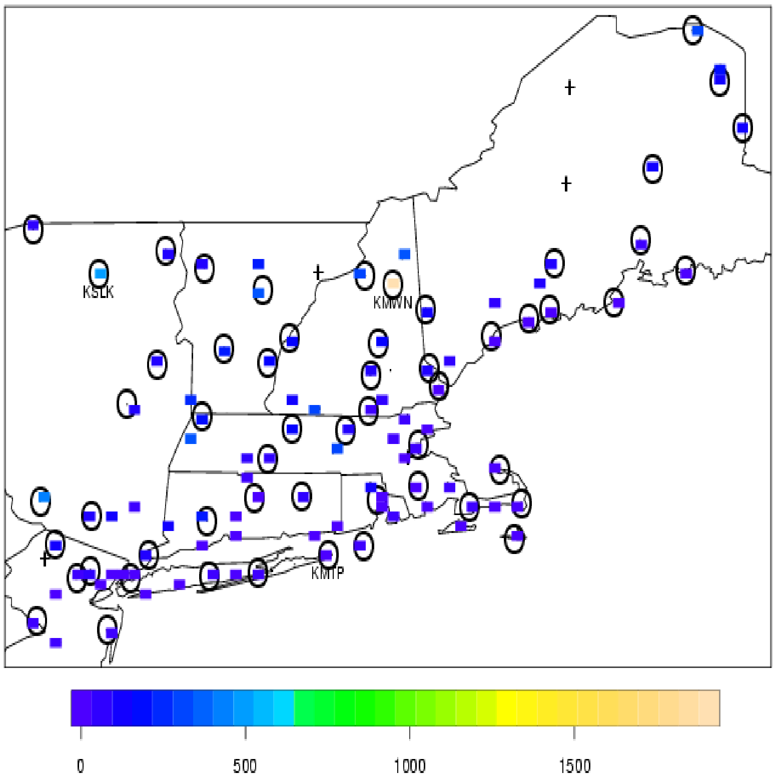


Figure 1: ***New England subset of METAR station locations used for exploring network design issues for verification of cloud ceiling height and visibility. Circled are the best 56 of 104 stations found from the analysis. Elevations are shown by shading and the "+" symbol indicates stations with no visibility data in January 2003.***

There are 719 hourly time points yielding a total of $74,880$ data points, with $3,311$ missing data points for cloud ceiling height and $7,747$ for visibility–5 stations do not have any visibility data recorded for this time period(Figure 1). Although theoretically continuous, the data are measured rather discretely (Gilleland (2004)) suggesting the use of categories.

The most natural categories are the flight rules (Table 1), or an aggregate of two of them: poor visibility (LIFR and IFR) and clear skies (MVFR and VFR).

Table 1: Designations for Low Instrument Flight Rules (LIFR), Instrument Flight Rules (IFR) Marginal Visual Flight Rules (MVFR), and Visual Flight Rules (VFR).

	Flight rules	Cloud ceiling height	Visibility
Poor Visibility	LIFR	< 500 feet	< 1 mile
	IFR	< 1000 feet	< 3 miles
Clear Skies	MVFR	< 3000 feet	< 5 miles
	VFR	> 3000 feet	> 5 miles

3 STATISTICAL ANALYSES

For network design it is of interest to know whether information on a variable of interest from a group of sites is sufficient for an area or not. Often, spatial sampling design is used for sampling points from a grid of locations and being able to rearrange the design locations (see, for example, Angulo (2003) or Müller (2003)). Here, however, points are not from a grid nor is it possible to move stations; it is desired only to thin an existing network. Additionally, because of the discrete nature of these data, standard geostatistic techniques are not appropriate. Categorical or indicator kriging is more appropriate (Gilleland (2004)); however, another alternative, used here, is to look at how the percent agreement (PA) among neighboring stations changes as the network is thinned (see Fowler *et al* (2004)). This latter approach is the most natural when considering the application to forecast verification because of its analogy to probability of detection (POD) or false alarm rate (FAR).

The correlation function of Gilleland (2004) (pair-wise probabilities of two stations, say $\mathbf{x}$ and $\mathbf{x}'$, separated by a distance vector, h, (here $h = |\mathbf{x} - \mathbf{x}'|$ and $|\mathbf{x} - \mathbf{x}'|$ is the great circle distance between $\mathbf{x}$ and $\mathbf{x}'$), being in the same category through time) is used here. Where feasible, a correlation function, $\rho(h)$, is fit, and $1 - \rho(h)$ is used as a measure of dissimilarity in a coverage design (see Nychka and Saltzman (1998) or Johnson (1990)). Here, the function *cover.design* from the R package fields (Nychka *et al* (2003)) is used for calculating the coverage designs.

That is, for a given set of candidate points, C, denote the set of n design points as D, where $D \subset C$; then an overall average criterion is an L_q average of cover points in the design region.

$$\left(\sum_{\mathbf{x}' \in C} \left(\sum_{\mathbf{x} \in D} d(\mathbf{x}, \mathbf{x}')^p\right)^{q/p}\right)^{1/q} \tag{1}$$

where $p < 0$, $q > 0$ are parameters and $d(\mathbf{x}, \mathbf{x}')$ is a distance metric, or in the present case the dissimilarity metric $d(\mathbf{x}, \mathbf{x}') = 1 - \rho(h)$.

Criterion (1) is minimized over several space-filling designs of a given size to obtain a "coverage design" from among the class of space-filling designs. It is possible to fix points in the design so that they cannot be swapped out. Note that this method gives a subset of a predetermined size, n, that is "best"; but does not determine the "best" size of the network. It is usual to investigate the effects on prediction error or cross validation on nested designs. For this analysis, it is more appropriate to look at the effects of PA (as described in Fowler *et al* (2004)) on nested designs because of the semi-discrete nature of these data, as well as the application.

4 RESULTS

Gilleland (2004) and Fowler *et al.* (2004) found that the California subset showed little spatial correlation for cloud ceiling height or visibility, indicating it should not be thinned–particularly for cloud ceiling height verification. However, a reasonable subset of 20 stations was obtained for visibility data based on fixing some stations for elevation and some from a coverage design obtained from a smaller subset around the San Francisco Bay. A verification analysis of the entire subset and the thinned network shows some improvement in the forecast verification without any change in bias. POD and FAR both improved slightly (by 2%) when using fewer stations (*possibly* resulting from removing many stations in the San Francisco Bay area; a region where making correct cloud ceiling and visibility forecasts is likely very difficult). Poor forecasts in this area are penalized less using the 20 stations than with all 48 stations. Verification statistics are still quite close, and the bias is unchanged; indicating that nothing was likely lost by using 20 stations instead of 48. These results suggest that the California subset may be nonstationary, and thinning the network may still be reasonable if such nonstationarity is taken into account.

The analysis of Gilleland (2004) is based on all four flight rules categories, whereas that of Fowler *et al.* (2004) is based on two categories (Table 1). Using the two categories with the approach of Gilleland (2004) instead of all four did not show any improvement.

Using all four flight rules categories for New England is not very promising for network thinning. Empirical correlations for both cloud ceiling height and visibility are relatively scattered with some low values at relatively short distances (Figure 2)–particularly for visibility–indicating that network thinning may be inappropriate here.

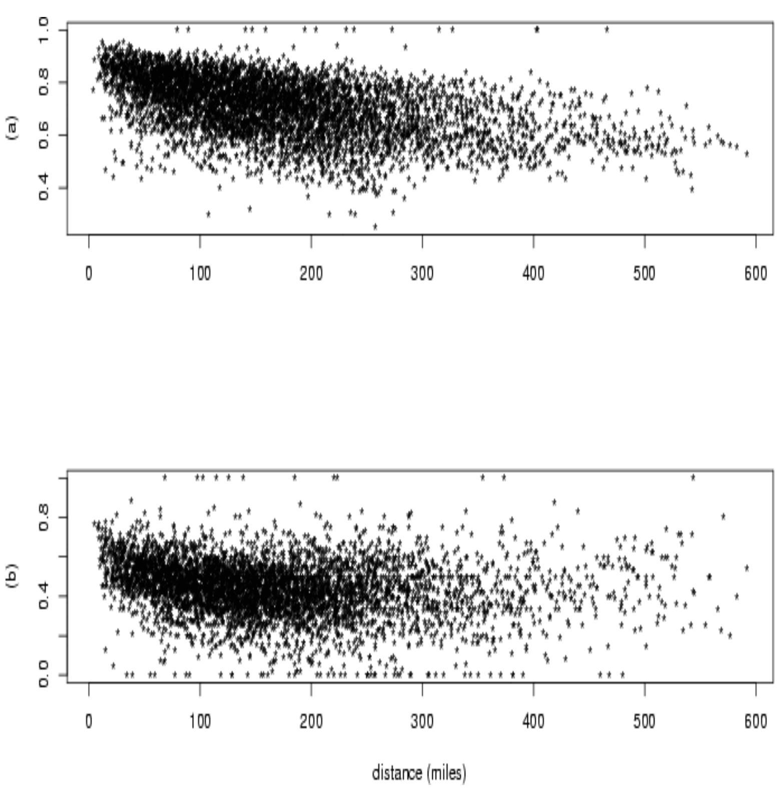


Figure 2: ***Distance vs. empirical correlations (pair-wise probabilities of two stations being in the same class) for (a) cloud ceiling height and (b) visibility data using all four flight rules categories.***

On the other hand, empirical correlations using only two categories shows that thinning the network is, in fact, appropriate (Figures 3 and 4). There are some seemingly strange gaps in

Figure 4 (a) where it would seem that some areas are less correlated with nearby stations than the norm for this subregion. However, these are easily explained by three monitoring stations that are relatively uncorrelated from nearby stations–specifically, KMWN, KMTP and KSLK (Figure 1). KMTP has zero correlation with all the other stations, which is not surprising considering there is a sharp difference in elevation between this station and all other stations in the region. The lower spatial correlation between these three stations and all other stations respectively suggests fixing these three stations along with at least one nearby station in the network designs. Figure 4 (b) shows the empirical correlations leaving these three stations out. Clearly, the spatial correlation among the remaining 96 stations is quite high suggesting that network thinning is appropriate for visibility in this region provided certain stations are not removed. The empirical correlations suggest fitting a mixture of exponentials correlation function.

$$\rho(h) = \alpha \exp(-h/\theta_1) + (1 - \alpha) \exp(-h/\theta_2) \tag{2}$$

The fitted function (2) for cloud ceiling height data (Figure 3) has parameter estimates (std. dev. in parentheses): $\hat{\alpha} \approx 0.30$ (0.003), $\hat{\theta}_1 \approx 15.04$ miles (0.93 miles) and $\hat{\theta}_2 \approx 1262.11$ miles (36.57 miles). For visibility data (Figure 4 (b)): $\hat{\alpha} \approx 0.099$ (0.002), $\hat{\theta}_1 \approx 40.33$ miles (1.74 miles) and $\hat{\theta}_2 \approx 7672.46$ miles (443.70 miles). Correlation model (2) allows the spatial field to be interpreted as the sum of two independent spatial processes with possibly different correlation scales without changing the smoothness of ρ at zero, but the shape will be modified for short distances in a similar fashion to the Matérn family (Gilleland and Nychka (2004)).

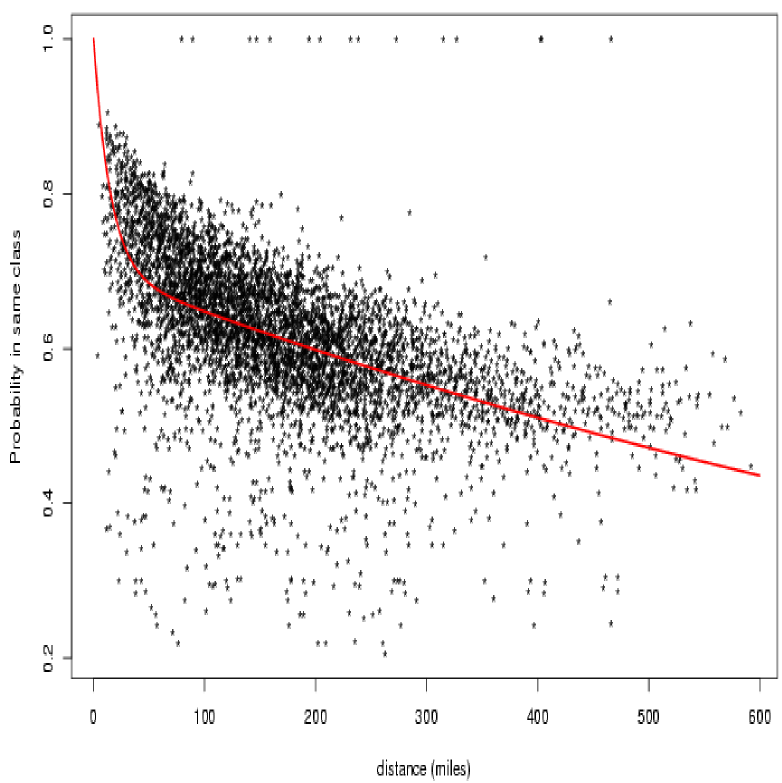


Figure 3: ***Distance vs. empirical correlations (pair-wise probabilities of two stations being in the same class) using only two classes: poor visibility (LIFR and IFR) and clear skies (MVFR and VFR): for cloud ceiling height data.***

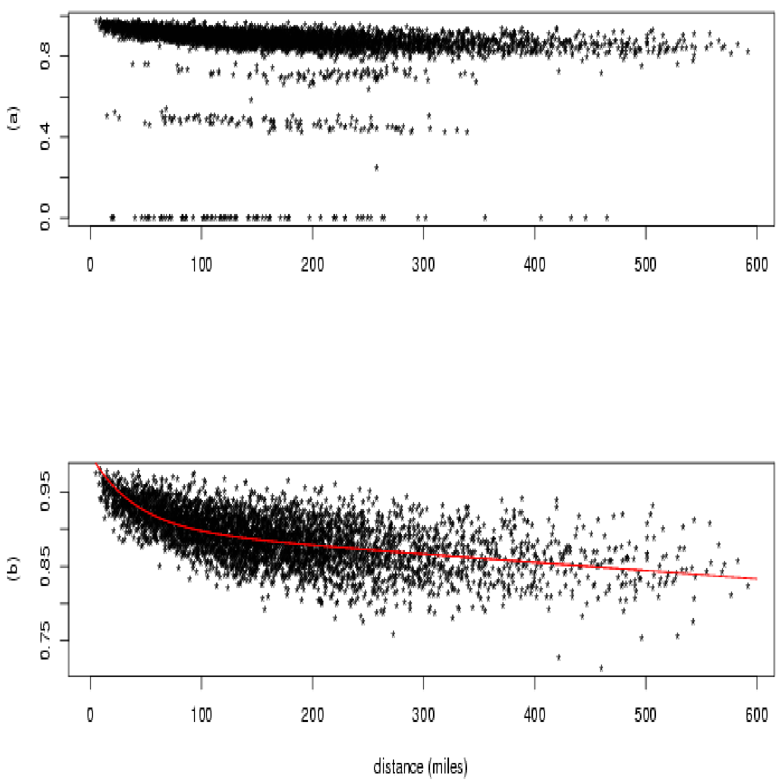
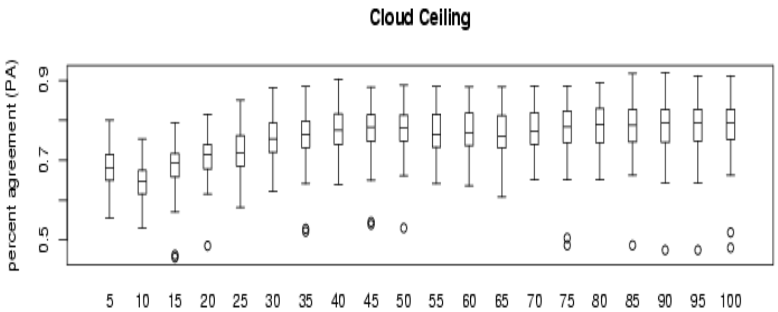


Figure 4: ***Distance vs. empirical correlations (pair-wise probabilities of two stations being in the same class) for visibility using only two classes: poor visibility (LIFR and IFR) and clear skies (MVFR and VFR): for visibility data with (a) all stations and (b) three stations removed.***

For cloud ceiling height using the New England subset, it appears that the design with 55 stations is reasonable because the distribution of PA does not change considerably beyond this number of stations, and is fairly high; similarly for visibility (Figure 5). It should be noted that the designs for cloud ceiling and visibility were found independently, and so are not the same. It is desirable to use the same design for each if possible. The correlation structure for cloud ceiling height is indicative that this will not be a problem; and, in fact, the mean PA for cloud ceiling height using the best design from the visibility data analysis is about 76%. The range is a bit wider (about 48% to 92%), but the interquartile range is high (about 71% to 81%); suggesting that the best design (of size 56) for the visibility data also works very well for cloud ceiling height.



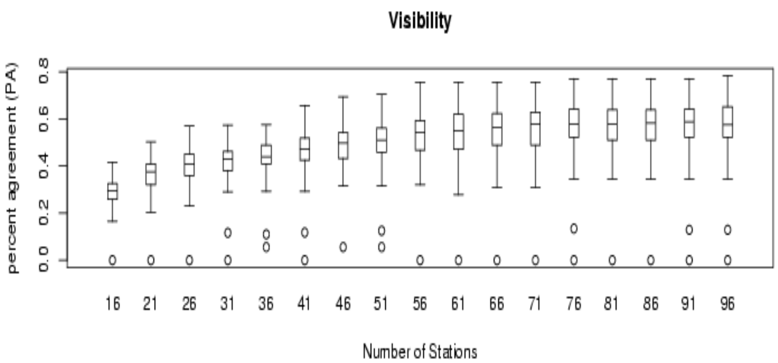


Figure 5: ***PA for predicting the event of poor visibility for cloud ceiling height*** *(*$n = 5, 10, 15, \ldots, 100$*)* ***and visibility*** *(*$n = 16, 21, 26, \ldots, 96$*)* ***for the entire New England subset from nested designs. Only the ten nearest neighbors used for prediction with design sizes greater than or equal to ten.***

Figure 1 shows the best coverage design for 56 stations obtained for the visibility data–and analyzed for the cloud ceiling data. Some good features of this design are: all stations near those with missing visibility data for the entire month are retained, the design is spread out across the entire region and nearly half of the stations are removed.

5 DISCUSSION

Compared with the California subset, New England seems to have conditions that are more amenable to network thinning. Elevation seems to factor into the analysis for only two stations: KMWN and KSLK. The other station with unique characteristics (KMTP) could possibly be explained by instrumentation or operational differences. Further investigation of this station would be required to determine if this is the case.

Future and ongoing work is primarily focused on testing these methods on more data. More winter months for 2003 in addition to data from more winter seasons is necessary to fully justify these network designs for the winter. Summer data is also needed, and will be looked at separately from the winter data in case of any seasonal effects. Stratification of the data by day and night also needs to be analyzed to account for any daily cycles.

ACKNOWLEDGEMENTS

This research is in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA. NCAR is sponsored by the National Science Foundation.

REFERENCES

Angulo, J.M., Ruiz-Medina, M.D., Alonso, F.J., Bueso, M.C., July 2003: Generalized approaches to spatial sampling design, *The ISI International Conference on Environmental Statistics and Health: Conference Proceedings*, 11-19.

Fowler, Tressa L., Braid, Jamie T. and Holmes Anne, 2004: Selecting METARs to verify ceiling and visibility forecasts, Preprints: 17th Conference on Probability and Statistics in the Atmospheric Sciences, Seattle, WA, 11-15 Jan, 2.4, American Meteorological Society (Boston).

Gilleland, Eric, 2004: Improving forecast verification through network design, Preprints: 17th Conference on Probability and Statistics in the Atmospheric Sciences, Seattle, WA, 11-15 Jan, 2.5, American Meteorological Society (Boston).

Gilleland, Eric and Nychka, Doug, 2004: Statistical models for monitoring and regulating ground-level ozone, *Special Issue of Environmetrics* (submitted).

Johnson, M.E., Moore, L.M., and Ylvisaker, D., 1990: Minimax and maximin distance designs, *Journal of Statistical Planning and Inference*, 26, 131-148.

Müller, W.G., July 2003: Spatial design methods in case of correlated observations: a comparison, *The ISI International Conference on Environmental Statistics and Health: Conference Proceedings*, 21-32.

Nychka, Douglas and Saltzman, Nancy, 1998: Design of Air Quality Monitoring Networks, *Lecture Notes in Statistics: Case Studies in Environmental Statistics, Springer*, 175 Fifth Avenue, New York, NY 10010.

Nychka, D., Bailey, B., Ellner, S., Haaland, P., O'Connell, M., Hardy, S., Baik, J., Meiring, W., Royle, J.A., Fuentes, M., Hoar, T., Tebaldi, C. and Gilleland, E., 2003: fields: a collection of programs based in R/S for curve and function fitting with an emphasis on spatial data, *http://www.cgd.ucar.edu/stats/Software/Fields/*.