4.1 An Extension of the Reliability Diagram for 3-Category Probability Forecasts

Daniel S. Wilks
Cornell University, Ithaca NY
dsw5@cornell.edu

1. Introduction

Most often forecast quality is characterized using scalar (i.e., single-number) verification statistics such as mean squared error, or the ranked probability score. Although restricting attention to one or a small number of scalar statistics simplifies the verification process both computationally and conceptually, this practice inevitably masks some aspects of forecast performance even in the most restricted verification settings. The problem arises because the dimensionality (Murphy 1991) of the joint distribution of forecasts and observations (Murphy and Winkler 1987) is large, especially in settings involving probability forecasts. Distributions-oriented (Murphy 1997), or diagnostic (Murphy et al. 1989, Murphy and Winkler 1992), verification methods communicate the richness of verification data sets by respecting their dimensionality, but at the same time must overcome the problem of how best to portray the high-dimensional information.

Arguably the most effective approaches to communicating the full information content of a joint distribution of forecasts and observations have involved the use of well-designed graphics, although only a small number of these have yet been devised. By far the most commonly used diagnostic verification graphic is the reliability diagram (Murphy and Winkler 1977, Wilks 2011). Reliability diagrams show the calibration-refinement factorization of the joint distribution of probability forecasts for dichotomous ("yes/no") events, by plotting subsample event relative frequency as a function of forecast probability (the calibration), and including also a plot of the frequency-of-use of each of the possible scalar probability forecasts (the refinement).

Probability forecasts are most frequently issued for dichotomous events, so that the reliability diagram is appropriate and effective for graphically displaying the relevant joint distribution. However, probability forecasts may also be issued jointly for more than two predictand categories, notably including 3-category temperature or precipitation forecasts at lead times of a week and longer (e.g., Van den Dool 2007, Livezey and Timofeyeva 2008, Barnston et al. 2010). In this format the predictand categories are often defined in terms of the climatological terciles, so that the below-normal (cold, or dry), near-normal, and above-normal (warm, or wet) categories each have climatological occurrence probabilities of 1/3. Diagnostic verification of such forecasts has been approached by reducing the three-element probability forecast vectors to collections of dichotomous probabilities (e.g., Wilks 2000, Wilks and Godfrey 2002, Barnston et al. 2010) but that approach is not fully satisfying because it neglects relationships among probabilities assigned to the different events (Murphy and Hsu 1986).

This paper proposes an extension of the well-known reliability diagram, to verification of probability forecast vectors pertaining to three distinct outcome categories using a two-dimensional graphic called the calibration simplex, which represents the calibration-refinement factorization of the full joint distribution of these forecasts and their corresponding observations.

2. Structure of the Calibration Simplex

Diagnostic verification methods are those that communicate the joint frequency distribution of the forecasts and their corresponding observations (Murphy and Winkler 1987),

$$\Pr\{f_i, o_j\} = \Pr\{f_i \cap o_j\};\ i = 1,\cdots,I;\ j = 1,\cdots J \quad . \qquad (1)$$

Here the probabilities have been rounded to one of I distinct possible forecasts f_i, each of which pertains to J possible observations or outcomes o_j. In the case of forecasts for dichotomous outcomes, $J = 2$. For example, in that case there would be $I = 11$ probability forecasts if rounded to tenths, with the distinct forecasts ranging from $f_1 = 0.0$ through $f_{11} = 1.0$.

It can be more informative to work with the joint distribution in terms of one of its factorizations (Murphy and Winkler 1987). The reliability diagram is based on the calibration-refinement factorization,

$$\Pr\{f_i, o_j\} = \Pr\{o_j \mid f_i\}\ \Pr\{f_i\};\ i = 1,\cdots,I;\ j = 1,\cdots J \quad . \qquad (2)$$

Thus, the full joint distribution can be decomposed into a collection of I conditional distributions $\Pr\{o_j|f_i\}$ of the observations given each of the possible forecasts f_i, called the calibration distributions; and a single I-element frequency distribution $\Pr\{f_i\}$ specifying the frequencies-of-use of the possible forecasts, called the refinement distribution. In the case of the reliability diagram, the calibration distributions are Bernoulli (i.e., binomial with $N = 1$) distributions, defined by the probability distribution function

$$\Pr\{o \mid f_i\} = p_i^{\,o}(1 - p_i)^{1-o} \quad , \ o = 0,\ 1 \ , \qquad (3)$$

where each p_i is estimated by its empirical (in the verification data set) conditional relative frequency $\hat{p}_i$ of the "yes" event occurring on occasions following the corresponding forecast f_i. Each of the I calibration distributions is fully characterized by its estimated Bernoulli probability $\hat{p}_i$, and collectively these define the vertical positions of the plotted points in the main portion of the reliability diagram. A histogram, or other quantitative representation of the refinement

distribution, completes the graphical portrayal of Equation (2).

Probability vectors $\mathbf{f}_i = [f_{B,i}, f_{N,i}, f_{A,i}]^T$ pertaining to the three mutually exclusive and collectively exhaustive outcomes "below-", "near-", and "above-normal", can be plotted in two dimensions because the three forecast probabilities in each vector must sum to 1. The geometrically appropriate graph in this case is the regular 2-simplex (Epstein and Murphy 1965, Murphy 1972), which takes the shape of an equilateral triangle. Each of the corners of the 2-simplex corresponds to forecast certainty (i.e., 100% probability) for one of the three outcomes being forecast. The point within the simplex at which a three-element probability vector $\mathbf{f}_i$ is plotted is located at distances proportional to the probabilities for each of the three outcomes, perpendicularly from the sides of the simplex opposite the respective corners. This plotting system generalizes the reliability diagram because the 1-simplex appropriate to 2-element probability forecasts for dichotomous events is the unit interval on the real line, which is the horizontal axis of the reliability diagram.

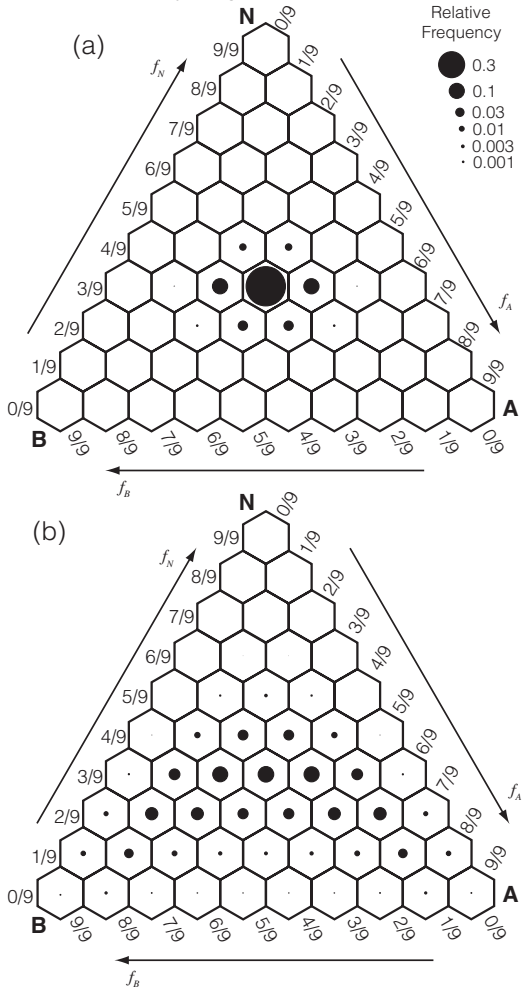

Figure 1. Calibration simplexes illustrating well calibrated forecasts exhibiting (a) lower and (b) higher sharpness.

Figure 1 illustrates the plotting of discretized forecast vectors onto the simplex, which has been rendered as a tessellation of hexagons. Each scalar forecast probability has been rounded to one of the K = 10 values 0/9, 1/9, . . ., 9/9, yielding $I = K(K+1)/2$ = 55 distinct possible vector forecasts $\mathbf{f}_i$, each of which is represented by one of the hexagons. The hexagons at the three vertices represent forecasts assigning all probability to the outcome labeled at that corner. Hexagons representing other forecast vectors are located at perpendicular distances from the respective opposite sides, that are indicated by the probability labels in the margins. For example, forecasts of equal probability for the middle category, f_N, are located along the same horizontal row of hexagons (perpendicularly upward from the horizontal bottom edge of the simplex), as indicated by the probability labels increasing upward along the left edge of the figure. Probability labels for the above-normal category, f_A, increase downward along the right edge of the simplex, and probability labels for the below-normal category, f_B, increase to the left along the bottom edge of the simplex. The result, for example, is that the large dot in the center of Figure 1a locates the climatological forecast vector $\mathbf{f}_{clim} = [1/3, 1/3, 1/3]^T$. Similarly, the forecasts in Figure 1a having the largest above-normal probability are $[2/9, 2/9, 5/9]^T$ and $[1/9, 3/9, 5/9]^T$, which are represented by the glyphs located furthest to the right in that diagram. Any two of the three forecast vector elements are sufficient to locate that vector's position on the simplex.

Figure 1 illustrates the graphical representation of refinement distributions Pr{$\mathbf{f}_i$} as glyph scatterplots (e.g., Wilks 2011), where the circle areas are proportional to the subsample sizes. Empty hexagons represent forecast vectors that were never used in the verification data sets under consideration.

Generalizing Equation 3 for the calibration distributions, in the case of three-element vector forecasts each of the I calibration distributions is multinomial, again with N = 1:

$$\Pr\{o \mid \mathbf{f}_i\} = p_{B,i}^{\,o_B}\, p_{A,i}^{\,o_A} (1 - p_{B,i} - p_{A,i})^{1-o_B-o_A},$$
$$o_B, o_A = 0, 1, \; o_B + o_A \le 1 . \qquad (4)$$

Thus each of the I calibration distributions are fully determined by any two of the three empirical (within the verification data set) conditional relative frequencies $\hat{p}_{B,i}$, $\hat{p}_{A,i}$, and $\hat{p}_{N,i} = 1 - \hat{p}_{B,i} - \hat{p}_{A,i}$ of the three events being forecast by $\mathbf{f}_i$, and therefore can be represented by a two-dimensional vector defined by, for example, $\hat{p}_{B,i}$ and $\hat{p}_{A,i}$ within the i^{th} hexagon of the simplex.

Figure 2 illustrates schematically how these conditional relative frequencies for the below- and above-normal outcomes are represented within each hexagon, using the corresponding miscalibration errors, or differences between the conditional average observations $\hat{p}_{B,i}$ and $\hat{p}_{A,i}$, and the respective forecast vector elements:

$$\begin{bmatrix} e_{B,i} \\ e_{A,i} \end{bmatrix} = \begin{bmatrix} \hat{p}_{B,i} - f_{B,i} \\ \hat{p}_{A,i} - f_{A,i} \end{bmatrix} . \quad (5)$$

These vector conditional miscalibration errors are plotted within their respective hexagons, using a coordinate system similar to that for the overall simplex, but with the origins at the centers of the hexagons. For a well-calibrated forecast subsample $\boldsymbol{f}_i$, both elements of Equation 5 will be zero, and the dot representing the corresponding subsample size will be plotted at the center of the i^{th} hexagon, as illustrated by the grey dot in Figure 2, and by all subsamples in both panels of Figure 1. In contrast, the solid dot in Figure 2 shows the plotting position when the above-normal category has been conditionally underforecast by 0.2 probability units, together with overforecasting of the below- and near-normal categories by 0.1 probability units each. In this case the location for the glyph representing the corresponding element of the refinement distribution will be displaced toward the "A" vertex of the simplex, consistent with the conditional outcome vector having a higher relative frequency than forecast for the above-normal category.

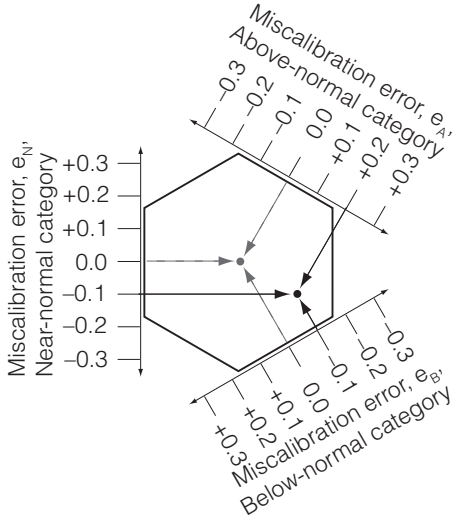

Figure 2. Plotting schematic showing locations of refinement distribution glyphs for calibrated forecasts (zero miscalibration errors, grey dashed arrows and point), and miscalibrated forecasts (nonzero calibration errors, black solid arrows and point).

3 Idealized Forecast Examples

Characteristic patterns of simplex glyph sizes and placements are diagnostic for different aspects of overall forecast performance, analogously to the situation for the reliability diagram (Wilks 2011, p. 335). These are illustrated in this section using simple idealized calibration and refinement distributions.

a. Sharpness: refinement distributions

Sharpness is a characteristic of the refinement distribution, independent of the relationship between the forecasts and observations. Forecast sharpness is commonly characterized using a measure of the dispersion of the refinement distribution, such as the standard deviation or variance (e.g., Wilks 2001). Forecasts that deviate frequently, including relatively large differences from, the climatological forecast are referred to as sharp. Forecasts exhibiting poor sharpness deviate rarely and relatively little from the climatological forecast.

Figure 1 shows two refinement distributions, with low (Figure 1a) and higher sharpness (Figure 1b), plotted as glyph histogram scatterplots on the simplex. The two distributions are defined quantitatively in Wilks (2013a). In both cases the resulting forecasts are shown as perfectly calibrated, so all of the circles representing subsample relative frequencies are plotted at the centers of their hexagons. For the low-sharpness forecasts in Figure 1a, it is clear that the most common forecast is $\boldsymbol{f}_{clim} = (1/3, 1/3, 1/3)^T$, which accounts for 68.3% of the forecasts. Furthermore, the individual forecast elements deviate no more than 2/9 from $\boldsymbol{f}_{clim}$, and only very rarely are those deviations larger than 1/9. In contrast, the sharper forecasts in Figure 1b often deviate quite strongly from $\boldsymbol{f}_{clim}$ (which accounts for only 10.7% of the forecasts), and take on values throughout the unit interval for the extreme-category probabilities f_B and f_A.

b. Conditional and unconditional biases: calibration distributions

Figure 3 shows calibration simplexes illustrating (a) overconfident forecasts, (b) underconfident forecasts, and (c) forecasts exhibiting an unconditional overforecasting bias for the above-normal category. Each of these cases is illustrated using the higher-sharpness refinement distribution shown in Figure 1b. Again, quantitative definitions of the three calibration distributions are provided in Wilks (2013a).

The overconfident forecasts shown in Figure 3a are conditionally biased, exhibiting overforecasting for probabilities above the climatological 1/3, and underforecasting for probabilities below 1/3. If such forecasts had been generated by a dynamical ensemble forecasting system, the overconfidence would be diagnostic for underdispersion (Wilks 2011, p. 372). Figure 3a shows that the signature for overconfidence in the calibration simplex is displacement of the relative frequency glyphs toward the center of the diagram, analogously to the calibration function in a conventional reliability or attributes diagrams being tilted from the ideal 45° diagonal toward the horizontal, climatological, "no resolution" (Murphy and Hsu 1986) line.

Figure 3b shows a calibration simplex for conditionally biased forecasts that are underconfident, in the sense that probabilities smaller than 1/3 are overforecast and probabilities larger than 1/3 are underforecast. Figure 3b shows that the signature for underconfidence in the calibration simplex is displacement of the relative frequency glyphs away from the climatological forecast at the center of the

diagram, and toward the corners of the simplex. Again this is analogous to the signature for underconfidence in the reliability diagram, in which the calibration function is tilted at an angle steeper than 45°, and away from the climatological horizontal "no resolution" line, and would be diagnostic for overdispersion of a dynamical ensemble.

Finally, Figure 3c shows a calibration simplex for unconditionally biased forecasts, exhibiting uniform overforecasting of the above-normal category equally at the expense of both below- and near-normal. Here the relative frequency glyphs are displaced upward and to the left, away from the "A" corner of the simplex. Equivalently, the below- and near-normal categories have been uniformly and equally underforecast, at the expense of the above-normal category, and the result is that the glyphs are displaced toward the simplex edge connecting the "B" and "N" corners.

4. Real Forecast Examples

This section illustrates use of the calibration simplex to understand performance of the Climate Prediction Center (CPC) "extended range" forecasts for average temperature and accumulated precipitation, at lead times of 6 to 10, and 8 to 14 days. These are subjective probability forecasts generated on weekdays, during 2001-2012 for the 6-10 day forecasts, and 2004-2012 for the 8-14 day forecasts, and interpolated to station locations from the graphical map products posted operationally at http://www.cpc.ncep.noaa.gov/.

Figure 4 shows the calibration simplexes for the temperature forecasts, which include n = 413,773 6-10 day forecasts (Figure 4a) and n = 313,523 8-14 day forecasts (Figure 4b). Glyphs are plotted only for forecasts having subsample sizes of 20 or more, and verifying categories have been determined relative to the 1971-2000 normals for forecasts made through April 2011, and using the 1981-2010 normals thereafter.

At each of the lead times, the overwhelming majority of forecasts include the climatological or near-climatological near-normal forecast f_N = 1/3, consistent with the conventional expectation that forecasts of the near-normal category will exhibit intrinsically weak skill (van den Dool and Toth 1991). Thus the sharpness for the near-normal probabilities is quite low. The most frequent forecast for each of the lead times is the climatological probability vector $\boldsymbol{f}_{clim}$, which was issued for 34.9% of the 6-10 day forecasts and 35.4% of the 8-14-day forecasts. These percentages correspond to n_{clim} = 144,211 and 111,010 for the climatological forecast $\boldsymbol{f}_{clim}$ (largest dots in the middles of the plots) in Figures 4a and 4b, respectively (compare glyph sizes in the legend). The corresponding error vectors $[e_B, e_N, e_A]^T$ are [–.085, .056, .029]^T (Figure 4a) and [–.088, .036, .052]^T (Figure 4b). Accordingly both of these glyphs have been displaced away from the "B" vertices, indicating too few below-normal verifications when the climatological temperature forecast was issued. The more extreme probability vectors were used correspondingly less frequently at the 8-14 day lead time, but overall the 6-10 day forecasts are only slightly sharper.

For both lead times, the temperature forecasts are only moderately well calibrated, with typical miscalibration errors in the range of 1/9 to 2/9. The

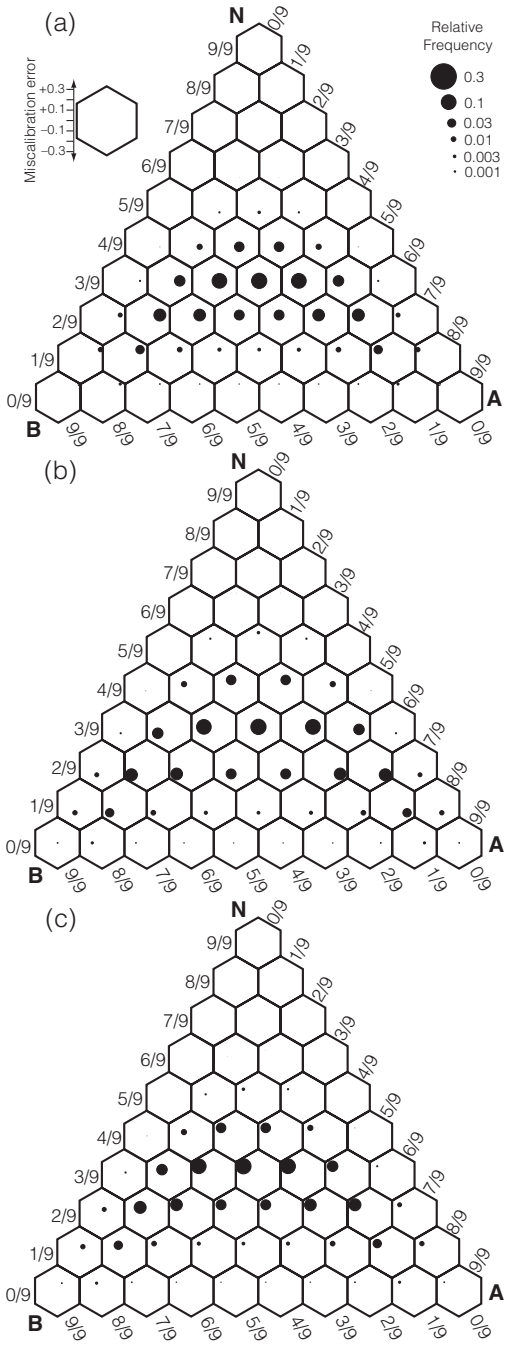

Figure 3. Calibration simplexes illustrating (a) overconfident forecasts, (b) underconfident forecasts, and (c) unconditionally biased forecasts. In each case the refinement distribution is the same as in Figure 1b.

above-normal outcome is underforecast for the larger (≥4/9) above-normal and near-normal probabilities, as these subsample size glyphs are displaced toward the "A" corner of the simplex. The glyph representing the very large subsample of $\boldsymbol{f}_{clim}$ forecasts is displaced toward the edge connecting the "N" and "A" vertices, indicating a smaller fraction of below-normal outcomes than the climatologically expected 1/3. Both of these results are consistent with the baseline 30-year normals lagging the quasi-linear warming trend that has been evident in U.S. temperature data since the mid-1970s (e.g., Livezey et al. 2007, Wilks 2013b, Wilks and Livezey 2013) together with the mean forecasts not fully tracking the warming, as has also been observed for seasonal tercile forecasts (Wilks 2000, Wilks and Godfrey 2002, Barnston et al. 2010). On the other hand, displacement of glyphs away from the simplex center for below-normal probabilities of 4/9 and larger indicates an overall underconfidence (consistent with the pattern of glyph dispersion away from the center of Figure 3b), suggesting that these forecasts could be somewhat sharper without degrading their skill.

Calibration simplexes for the precipitation forecasts are shown in Figure 5, which include n = 406,856 6-10 day forecasts (Figure 5a) and n =306,594 8-14 day forecasts (Figure 5b). The precipitation forecasts exhibit notably less sharpness than their temperature counterparts in Figure 4, with 44.9% of the 6-10 day forecasts and 44.7% of the 8-14 day forecasts using the climatological probabilities $\boldsymbol{f}_{clim}$. As was the case for the temperature forecasts, f_N = 1/3 is overwhelmingly the most common near-normal forecast. The most extreme probabilities have been used somewhat less frequently in the 8-14 day precipitation forecasts, but overall these are only slightly less sharp that the 6-10 day forecasts. At both lead times there is a clear tendency for the subsample-size glyphs to be displaced toward the "B" vertex, indicating underforecasting of the below-normal category, and corresponding overforecasting in roughly equal proportions of the near-normal and above-normal outcomes. Figure 5 also indicates strong overconfidence in the larger (≥4/9) probabilities for the near-normal category, for both lead times.

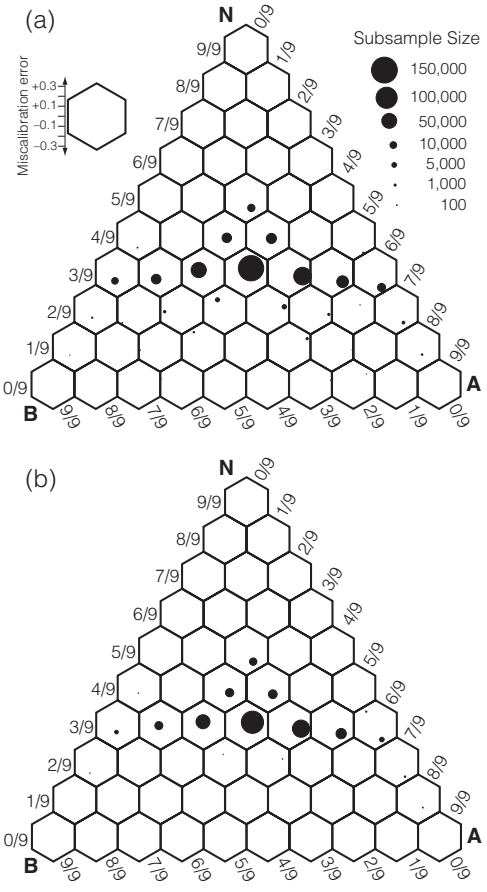

Figure 4. Calibration simplexes for (a) 6-10 day, and (b) 8-14 day temperature forecasts.

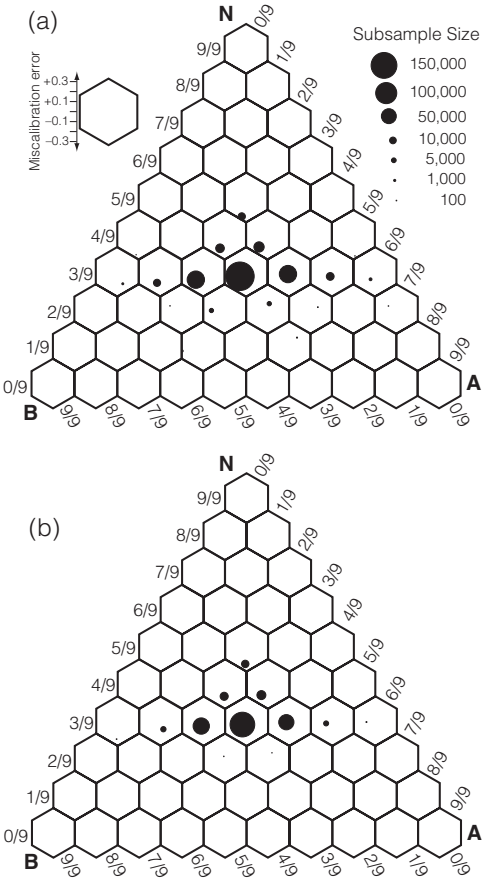

Figure 5. As Figure 4, for precipitation forecasts.

5. Summary and Conclusions

Gaining a full appreciation of the performance of a set of forecasts requires investigation of the joint distribution of the forecasts and corresponding observations (Murphy and Winkler 1987), and a graphical approach to this exposition will often be the most immediately informative. The well-known reliability diagram for probability forecasts of dichotomous events is the most commonly encountered such graphical device. This paper has defined and illustrated a natural extension of the reliability diagram to probability forecasts for three disjoint events, called the calibration simplex. It displays the refinement distribution for such forecasts using a glyph scatterplot histogram of the $K(K+1)/2$ possible vector forecasts that result when each probability element has been rounded to one of K discrete values. Simultaneously it shows the two empirical outcome relative frequencies conditional on each forecast vector that are necessary to fully characterize the corresponding refinement distributions, using displacements of the glyphs from central plotting locations.

Use of the calibration simplex has been illustrated using the CPC 6-10 and 8-14 day subjective temperature and precipitation forecasts. The temperature forecasts exhibit an unconditional bias consistent with the average forecasts lagging the ongoing climate warming, as has been observed also for seasonal forecasts, but these graphs also indicate that greater sharpness for the below- and above-normal elements of the forecast vectors could be employed without degrading overall accuracy or skill. The near-normal element of the precipitation forecast vectors were seen to be strongly overconfident; with overall underforecasting of the below-normal category, and corresponding overforecasting in roughly equal proportions of the near-normal and above-normal outcomes.

Acknowledgements

I thank Scott Handel for supplying the CPC forecasts and corresponding verifications. This research was supported by the National Science Foundation under grant AGS-1112200.

References

Barnston, A.G., S. Li, S.J. Mason, D.G. DeWitt, L. Goddard, and X. Gong, 2010: Verification of the first 11 years of IRI's seasonal climate forecasts. *Journal of Applied Meteorology and Climatology*, **49**, 493-520.

Epstein, E.S., and A.H. Murphy, 1965: A note on the attributes of probabilistic predictions and the probability score. *Journal of Applied Meteorology*, **4**, 297-299.

Livezey, R.E., and M.M. Timofeyeva, 2008: The first decade of long-lead U.S. seasonal forecasts – insights from a skill analysis. *Bulletin of the American Meteorological Society*, **89**, 843-854.

Livezey, R.E., K.Y. Vinnikov, M.M. Timofeyeva, R. Tinker, and H.M. van den Dool, 2007: Estimation and extrapolation of climate normals and climatic trends. *Journal of Applied Meteorology and Climatology*, **46**, 1759-1776.

Murphy, A.H., 1972: Scalar and vector partitions of the probability score: Part II. *N*-state situation. *Journal of Applied Meteorology*, **11**, 1183-1192.

Murphy, A.H., 1991: Forecast verification: its complexity and dimensionality. *Monthly Weather Review*, **119**, 1590-1601.

Murphy, A.H., 1997: Forecast verification. In: R.W. Katz and A.H. Murphy, eds., *Economic Value of Weather and Climate Forecasts*. Cambridge University Press, Cambridge. 19-74.

Murphy, A.H., B.G. Brown, and Y.-S. Chen, 1989: Diagnostic verification of temperature forecasts. *Weather and Forecasting*, **4**, 485-501.

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

Murphy, A.H., and R.L. Winkler, 1977: Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics*, **26**, 41-47.

Murphy, A.H. and R.L. Winkler, 1987: A general framework for forecast verification. *Monthly Weather Review*, **115**, 1330-1338.

Murphy, A.H., and R.L. Winkler, 1992: Diagnostic verification of probability forecasts. *International Journal of Forecasting*, **7**, 435-455.

Van den Dool, H., 2007: *Empirical Methods in Short-Term Climate Prediction*. Oxford University Press, Oxford, 215 pp.

Van den Dool, H., and Z. Toth, 1991: Why do forecasts for "near normal" often fail? *Weather and Forecasting*, **6**, 76-85.

Wilks, D.S., 2000: Diagnostic verification of the Climate Prediction Center long-lead outlooks, 1995-98. *Journal of Climate*, **13**, 2389-2403.

Wilks, D.S., 2001: A skill score based on economic value for probability forecasts. *Meteorological Applications*, **8**, 209-219.

Wilks, D.S. 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd Edition. Academic Press, Amsterdam, 676 pp.

Wilks, D.S., 2013a: The calibration simplex: a generalization of the reliability diagram for 3-category probability forecasts. *Weather and Forecasting* **28**, 1210-1218.

Wilks, D.S., 2013b: Projecting "normals" in a nonstationary climate. *Journal of Applied Meteorology and Climatology*, **52**, 289-302.

Wilks, D.S., and C.M. Godfrey, 2002: Diagnostic verification of the IRI Net Assessment forecasts, 1997-2000. *Journal of Climate*, **15**, 1369-1377.

Wilks, D.S. and R.E. Livezey, 2013: Performance of alternative "normals" for tracking climate changes, using homogenized and non-homogenized seasonal U.S. surface temperatures. *Journal of Applied Meteorology and Climatology*, 52, 1677-1687.