Evaluating Rank Histograms Using Decompositions of the Chi-Square Test Statistic

IAN T. JOLLIFFE

School of Engineering, Computing and Mathematics, University of Exeter, Exeter, United Kingdom

CRISTINA PRIMO

European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

(Manuscript received 12 April 2007, in final form 27 July 2007)

ABSTRACT

Rank histograms are often plotted to evaluate the forecasts produced by an ensemble forecasting system—an ideal rank histogram is "flat" or uniform. It has been noted previously that the obvious test of "flatness," the well-known χ^2 goodness-of-fit test, spreads its power thinly and hence is not good at detecting specific alternatives to flatness, such as bias or over- or underdispersion. Members of the Cramér–von Mises family of tests do much better in this respect. An alternative to using the Cramér–von Mises family is to decompose the χ^2 test statistic into components that correspond to specific alternatives. This approach is described in the present paper. It is arguably easier to use and more flexible than the Cramér–von Mises family of tests, and does at least as well as it in detecting alternatives corresponding to bias and over- or underdispersion.

1. Introduction

It is common to use rank histograms to evaluate the performance of ensemble forecasting systems (see Elmore 2005, hereinafter E05, and references cited therein). An ideal system produces a "flat" or uniform histogram, but because of sampling variation the histograms are almost never exactly flat. The question then arises: can observed deviations from "flatness" or uniformity be attributed to chance, or do they indicate deficiencies in the forecasts? An overall test of uniformity is provided by the well-known χ^2 goodness-of-fit test. This is a good test to use if *all* alternatives to flatness are of interest, because it has some power to detect any type of alternative. However, as noted by E05, if specific alternatives are of interest, such as bias or over- or underdispersion in the forecasts, then the χ^2 test is not very good at detecting them. Instead, it is advisable to use tests that concentrate their power on the specific alternatives of interest. E05 describes tests from the Cramér–von Mises family of tests that, unlike the χ^2 test, take into account the ordering of the bins and hence have improved power against certain hypotheses (see Choulakian et al. 1994 for more detail). Noceti et al. (2003) also examined the power of members of the Cramér–von Mises family of tests to detect various alternatives, although the context was different. They carried out tests on data from nonuniform distributions that were transformed to uniformity (see also Gneiting et al. 2007).

In the present paper, an alternative approach is suggested in which the χ^2 test statistic is decomposed into components that indicate whether the forecasts are biased, whether they are over- or underdispersed, and whether there are any other deviations from flatness once these two possibilities are accounted for. Other decompositions are also possible if different alternatives are of interest. The decompositions are arguably easier to use and more flexible than the Cramér–von Mises tests and are at least as good at detecting the alternatives of interest. Section 2 of the paper shows how decompositions can be constructed in a simple case, with detailed formulas for the decompositions deferred to the appendix. Section 3 applies the decomposition to the artificial examples given by E05 and to forecasts of 500-hPa geopotential heights. Finally, section 4 includes some concluding remarks and caveats.

Corresponding author address: Prof. Ian T. Jolliffe, 30 Woodvale Road, Gurnard, Cowes, Isle of Wight, PO31 8EG, United Kingdom.
E-mail: ian@sandloch.fsnet.co.uk

DOI: 10.1175/2007MWR2219.1

2. Decomposing the χ^2 test statistic

Consider a rank histogram for ensemble forecasts with $(k - 1)$ members, and hence k bins or classes in the rank histogram. Suppose that there are n $(k - 1)$-member ensemble forecasts from which the rank histogram is constructed. Then, if the underlying distribution is flat or uniform, there is the same probability p_0 of a verification observation falling in each class, and the expected number in each class is np_0. If n_i is the observed number in the ith class and $e_i = np_0$ is the corresponding expected number, then the test statistic for the usual χ^2 goodness-of-fit test is

$$T = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}. \tag{1}$$

Under the null hypothesis of a uniform underlying distribution, the statistic T has come from (approximately) a χ^2 distribution with $(k - 1)$ degrees of freedom.

There is a result, dating back at least to Kendall and Stuart (1967), which states that T can be decomposed into $(k - 1)$ asymptotically independent components, each of which has an approximate χ^2 distribution with 1 degree of freedom. To achieve this decomposition, construct a $(k \times k)$ matrix $\mathbf{L}$, with elements l_{ri}, whose rows are orthonormal and whose final row is $(1/\sqrt{k})(1, 1, \ldots, 1)$. Because the rows of $\mathbf{L}$ are orthonormal, its elements should satisfy $\Sigma_{i=1}^{k} l_{ri}l_{si} = 1$ when $r = s$ and $\Sigma_{i=1}^{k} l_{ri}l_{si} = 0$ otherwise, for $r, s = 1, 2, \ldots, k$. Each of the rows of $\mathbf{L}$, except the last, is a so-called contrast, meaning that the sum of its elements is zero.

Next, define $x_i = [(n_i - e_i)/\sqrt{e_i}]$. Then $\Sigma_{i=1}^{k} x_i^2$ is the usual χ^2 statistic T. Finally, let $u_r = \Sigma_{i=1}^{k} l_{ri}x_i$, $r = 1, 2, \ldots, k$.

The last row of $\mathbf{L}$ is $(1/\sqrt{k})(1, 1, \ldots, 1)$ and $e_i = np_0$, $i = 1, 2, \ldots, k$, so $u_k = \Sigma_{i=1}^{k} l_{ki}x_i = (1/\sqrt{knp_0})\ \Sigma_{i=1}^{k} (n_i - e_i)$. But $\Sigma_{i=1}^{k} n_i = \Sigma_{i=1}^{k} e_i = n$, so $u_k = 0$.

Also, as is shown in a somewhat complex proof in Kendall and Stuart (1967), the other $(k - 1)$ u_r are asymptotically independent Gaussian random variables, each with mean zero and unit variance. Hence, asymptotically their squares are independent χ^2 random variables, each with 1 degree of freedom. Furthermore, $T = \Sigma_{i=1}^{(k-1)} u_i^2$ (Kendall and Stuart 1967).

A restricted form of decomposition of T has been used in the economics literature to detect deviations from uniformity in location, spread, skewness, and so on (see, e.g., Boero et al. 2004). With k bins, the elements of $\mathbf{L}$ are restricted to $(\pm 1/\sqrt{k})$.

To see how our result can be used to decompose T into components that are sensitive to specific alternative hypotheses, consider a very simple example with $n = 100$ and $k = 4$. More complicated examples will be considered in the next section. Here $e_i = 25$, $i = 1, 2, 3, 4$, and the last row of $\mathbf{L}$ is (0.5, 0.5, 0.5, 0.5). Specific vectors will now be given for the other rows of $\mathbf{L}$ in this example: general formulas for such vectors are provided in the appendix.

Suppose that a linear trend is of interest, corresponding to bias in the forecast. Define the first row of $\mathbf{L}$ to be a linear contrast $(1/\sqrt{20})(-3, -1, 1, 3)$. Then $u_1 = \Sigma_{i=1}^{k} l_{1i}x_i$ will be sensitive to trend. If a contrast between the observations in the two end categories and those in the middle categories is also of interest (corresponding to over- or underdispersion), define the second row of $\mathbf{L}$ to be $\frac{1}{2}(1, -1, -1, 1)$; then $u_2 = \Sigma_{i=1}^{k} l_{2i}x_i$ will be sensitive to over- or underdispersion. To complete the matrix $\mathbf{L}$, a third orthonormal contrast is needed. To achieve orthonormality with the other rows, it takes the form $(1/\sqrt{20})(1, -3, 3, -1)$. This contrast is included here simply to complete the matrix and show that T can be decomposed into $(k - 1)$ components, each with 1 degree of freedom. In practice, as will be seen in the examples, fewer than $(k - 1)$ components will usually be interpretable and of direct interest.

Now suppose that the observed values are (15, 22, 28, 35), which have a clear trend or bias. Then $T = 8.72$, $u_1^2 = 8.712$, $u_2^2 = 0$, and $u_3^2 = 0.008$, with the last three numbers adding to T. Under the null hypothesis, T has approximately a χ^2 distribution with 3 degrees of freedom and u_1, u_2, and u_3 have (approximately) independent χ^2 distributions, each with 1 degree of freedom. Here, $T = 8.72$ corresponds to a p value of 0.033, while $u_1^2 = 8.712$ corresponds to a p value of 0.003. The two other components u_2 and u_3 are clearly not significant. Thus, the overall test suggests a deviation from uniformity, but the decomposition pins down the form of that deviation as trend or bias.

Suppose now that the observed values are (31, 18, 19, 32). In this case the deviation from flatness indicates underdispersion of the ensemble. Here $T = 6.80$, $u_1^2 = 0.032$, $u_2^2 = 6.76$, and $u_3^2 = 0.008$; $T = 6.80$ corresponds to a p value of 0.079, while $u_2^2 = 6.76$ corresponds to a p value of 0.009. The two other components u_1 and u_3 are clearly not significant. The overall test statistic is not large enough to reach the 5% significance level for χ^2 with 3 degrees of freedom, but u_2^2 gives strong evidence of a deviation from uniformity in the sense of underdispersion.

3. Examples

In this section we illustrate the decomposition approach on Elmore's artificial examples and on a real example.

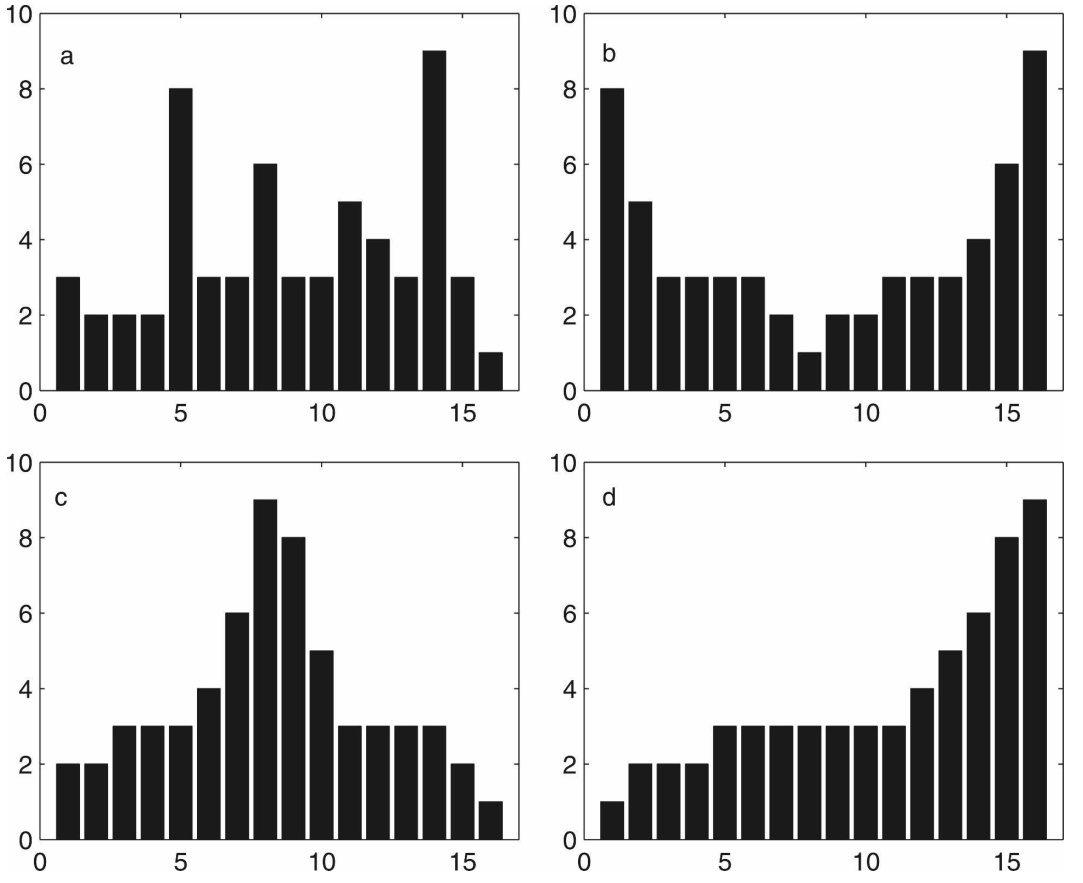

FIG. 1. The same rank histograms as E05's Fig. 1: (a) 60 observations generated randomly from a flat (uniform) distribution over 16 bins. The rest of the panels shows the same 16 numbers rearranged in ways that are suggestive of (b) underdispersion, (c) overdispersion, and (d) bias.

a. Elmore's artificial examples

E05 illustrated the lack of power of the usual χ^2 goodness-of fit test for specific deviations from uniformity, using two artificial examples. Figure 1 gives the same rank histograms as E05's Fig. 1. Figure 1a shows 60 observations generated randomly from a flat (uniform) distribution over 16 bins, and Figs. 1b–d show the same 16 bin frequencies rearranged in ways that are suggestive of underdispersion, overdispersion, and bias. The value of T—19.467, with corresponding p value 0.193—is identical for all four graphs. The conclusion is that there is no evidence against the null hypothesis of uniformity, a conclusion that seems reasonable for the data in Fig. 1a, but not for Figs. 1b–d.

E05 introduces alternative test statistics from the Cramér–von Mises family, and shows how they can detect the nonflatness that is apparent in Figs. 1b–d. They can do this because they are based on cumulative frequencies in rank histogram, rather than individual frequencies and hence, unlike the χ^2 test, they take into account the ordering of the bins. However, they are not designed to be powerful against specific alternatives to flatness.

Alternatively, the χ^2 statistic T can be decomposed as described in the previous section. Here we have 16 classes, and to look for bias and under- or overdispersion we use the 16-element vectors $\mathbf{l}_1 = (1/\sqrt{1360})(-15, -13, \ldots, -3, -1, 1, 3, \ldots, 13, 15)$, $\mathbf{l}_2 = (1/\sqrt{112})(7, -1, -1, \ldots, -1, -1, 7)$, respectively. General forms for these and other contrasts that may be of interest are given in the appendix. To complete the matrix $\mathbf{L}$, we need 13 more orthogonal rows, but if the 2 rows already defined represent the only deviations from uniformity that are of interest, then the simple decomposition $T = u_1^2 + u_2^2 + \Sigma_{i=3}^{(k-1)} u_i^2$ can be used. Under the null hypothesis of uniformity, the three terms in the decomposition asymptotically have independent χ^2 distributions, with 1, 1, and $(k - 3)$ degrees of freedom, respectively. To calculate the third term we do not need all the individual rows of $\mathbf{L}$ and corresponding u_i, $i = 3, \ldots, k - 1$. We simply subtract the

TABLE 1. Decomposition statistics and p values for Elmore's first artificial example.

		Linear	Ends	Resid_1	V-shape	Resid_2
Figure 1a	Statistic	0.512	1.876	17.088	1.146	17.808
	(p value)	0.474	0.172	0.195	0.284	0.165
Figure 1b	Statistic	0.314	13.752	5.401	14.679	4.474
	(p value)	0.575	0.000 21	0.965	0.000 13	0.985
Figure 1c	Statistic	0.113	3.086	16.268	14.679	4.474
	(p value)	0.737	0.079	0.235	0.000 13	0.985
Figure 1d	Statistic	14.937	0.952	3.578	2.314	2.216
	(p value)	0.000 11	0.321	0.995	0.128	0.9996

first two terms from T. The third term represents all deviations from uniformity that are independent of those represented by the first two terms.

The first three columns of Table 1, labeled Linear, Ends, and Resid_1, give the values of the three terms in the decomposition, together with corresponding p values, for Figs. 1b–d, respectively. It is seen that, although the overall test statistic T provides no evidence of departures from flatness in these instances, its components do. Specifically, the test for trend or bias has a p value of 0.000 11 for Fig. 1d. This p value is comparable to that quoted by E05 for the Cramér–von Mises and Anderson–Darling tests. Also, in Fig. 1b the test for under- or overdispersion has a p value of 0.000 21, an order of magnitude smaller than for the tests used by E05. Of course, the p values are calculated using approximate null distributions, but it is clear that decomposing χ^2 provides a viable, powerful alternative to the Cramér–von Mises family.

The p value for the Ends test in Fig. 1c is not particularly small. This is understandable, as the test is designed to detect deviations from flatness in the end classes, while the deviations here are spread over all classes. However, the decomposition approach is sufficiently flexible to cope with this. Suppose that we expect deviations from flatness to be as in Figs. 1b and 1c, rather than concentrated at the ends. For convenience, such deviations will be referred to as V-shaped even though one of them is an inverted V. In the same way that a single contrast can be used to detect positive or negative linear trend, and a single contrast will detect both under- and overdispersion, a single contrast can also detect both V and inverted-V shapes. A positive value of the corresponding u_i will correspond to one shape, and a negative value to the inverted version. In the present example, the appropriate contrast is $(1/\sqrt{336})(7, 5\ \ 3, \ldots, -5, -7, -7, -5, \ldots, 3, 5, 7)$—a general formula is given in the appendix. This is orthonormal to the linear contrast found earlier, but not to the Ends contrast. Thus Linear and Ends, or Linear and V-shape, can be used together in a decomposition of T into independent χ^2 terms, but Ends and V-shape cannot.

The first, fourth and fifth columns of Table 1, labeled Linear, V-Shape and Resid_2, give values, and p values, for three terms in an alternative decomposition to that involving Ends. It is seen that the p value for the component corresponding to V-shape in Figs. 1b and 1c is very small, and is an order of magnitude smaller than those quoted by E05 for the Watson test.

E05 has another similar example, with the same number of classes but with 540 rather than 60 observations distributed among the classes. Results for the Cramér–von Mises family of tests are similar in E05's larger example to those in the smaller example. The same is true for our χ^2 decomposition approach.

In identifying whether a specific form of trend or bias is present, an alternative approach to decomposition of T or the Cramér–von Mises family of tests is to construct a null hypothesis corresponding to the specific form of the trend or bias. This has two disadvantages: first, it relies on being able to exactly specify the form of bias or trend of interest, and second, it is never possible to accept a null hypothesis. The most that can be done is to fail to reject it.

b. *Northern Hemisphere 500-hPa geopotential height forecasts*

The data examined here are daily 24-h forecasts of NH 500-hPa heights for the 2006/07 winter season. They were created from the National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS). There were 14 ensemble members and hence 15 bins. The total number of forecasts is 290 304, comprising 84 days of forecasts for 3456 grid points. The rank histogram, with percentages rather than absolute numbers on the vertical axis, is plotted in Fig. 2. The value of T, the overall χ^2 statistic, is over 33 000. Not surprisingly, with such a large sample size, the deviations from uniformity achieve extreme levels of significance, as do all components in the decompositions described earlier. This may seem a

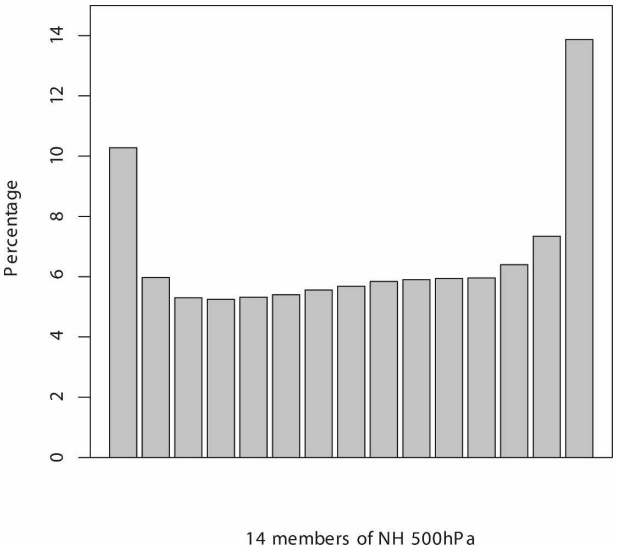

FIG. 2. Rank histogram of 14 ensemble members of daily 24-h forecasts of NH 500-hPa heights for the 2006/07 winter season. They were created from the NCEP GEFS system. The vertical axis represents percentages rather than absolute numbers.

trivial example, but it is not. As noted by Hamill (2001), to use the χ^2 approximation to the distribution of T, the observations should be independent. Clearly they are not here—there is both spatial and temporal dependence between the observations. Indeed, it has been suggested (Z. Toth 2007, personal communication) that there are approximately 25 degrees of freedom in the NH, and that the number of time points should be divided by 5. Thus the 290 304 observations are roughly equivalent to 420 independent observations.

Suppose now that the rank histogram resulted from 420 observations. Then $T = 48.36$ with a p value of 0.000 01. This is still very small, but large enough to make it interesting to investigate its decomposition. We do not pretend that reducing the number of observations from 290 304 to 420, but retaining the same percentages, addresses the problem of testing the flatness of the original dependent dataset in an optimal way—the rank histogram is smoother than it should be if it were truly based on 420 independent observations. Other ways of dealing with dependence in the original data are possible, as is discussed in section 4, but for the purpose of illustration the same shape of histogram as for the full dataset, but with reduced sample size, is assumed here.

Table 2 gives information for this example in similar format to that of Table 1. The Linear, V-shape, and Ends components are all highly significant. Visually, the deviation from flatness due to bias or trend is not particularly obvious, but the small p value for the Linear component shows it to be statistically significant.

TABLE 2. Decomposition statistics and p values for 500-hPa height forecasts.

	Linear	Ends	Resid_1	V-shape	Resid_2
Statistic	7.397	17.495	23.465	6.426	34.534
p value	0.0065	0.000 029	0.024	0.011	0.0006

The p value for Ends is three orders of magnitude smaller than that for V-shape. This confirms the visual impression in Fig. 2 that any deviation from uniformity due to under- or overdispersion is concentrated mainly in the end categories. Another aspect of this result is that the p value for the residual deviations after removing the Linear and V-shape components is much smaller than that for residuals after removal of the Linear and Ends components. This indicates that the extent of unexplained deviations from uniformity is much greater for the former decomposition.

The two examples in this section have similar numbers of bins, but the technique is easy to use for any number of bins. We have applied it to examples with the number of bins ranging from 10 to 51.

4. Discussion

This paper has shown that the lack of power of the overall χ^2 test statistic T against specific alternatives can be overcome by decomposing T into components that home in on alternatives of special interest. The approach provides an alternative to the use of test statistics from the Cramér–von Mises family, advocated by E05. The power of the new approach is competitive with, or better than, that of the Cramér–von Mises statistics in the examples examined. Furthermore, it has some advantages in terms of ease of use and flexibility.

First, the technique needs only χ^2 distributions to assess significance and compute p values, whereas the Cramér–von Mises tests need special tables or formulas to derive critical values or p values. In terms of flexibility, two possible decompositions have been considered here, Linear + V-shape + Residual and Linear + Ends + Residual. The appendix gives simple formulas for the rows of the matrix $\mathbf{L}$ needed to implement these. However, if different alternatives are of interest, then components in these decompositions can be replaced, or extra components added. The appendix demonstrates this for U-shaped alternatives.

Cramér–von Mises tests are not designed to be powerful against specific alternatives to flatness. They, too, can be decomposed, and elements in the decomposition can, in some cases, be identified with certain alternatives, but the decomposition is more rigid and the prac-

tical application more complicated than decomposing T (see Choulakian et al. 1994 and references therein).

In addition to the advantages of the approach, there are, of course, drawbacks. All the testing and the calculations of p values rely on χ^2 approximations to distributions of test statistics. We have no formal results to tell us when the approximations can be used. However, the conditions for the χ^2 approximation to the distribution of T to be good are well known, and we would be surprised if the conditions required for the components in the decomposition are much more stringent.

Another point to remember is that the decomposition requires orthogonal rows in $\mathbf{L}$. Thus, if both the V-shape and Ends components are included in a decomposition for which the $\mathbf{l}$ vectors are not orthogonal, then the two terms are not independent and the residual term will not have a χ^2 distribution.

The flexibility of the decomposition also has a potential pitfall. Ideally, the types of decomposition of interest should be decided *before* looking at the rank histogram. Choosing alternatives after seeing the histogram is an instance of using the same data to formulate and test the same hypothesis, which will usually overstate the significance of the test statistic.

It should also be noted that although significant values for the Linear component have been taken as corresponding to bias in the forecasts, and the Ends or V-shape components to under- or overdispersion, this may not be the case. As noted by Hamill (2001), U-shaped rank histograms can occur for a number of reasons other than underdispersion. The same is true for V-shaped histograms, and similarly a Linear rank histogram need not imply bias. Conversely, a flat histogram does not necessarily indicate reliability of the ensemble forecasts Hamill (2001). Of course, the problems noted in this paragraph relate to interpretation of rank histograms in general, and not to the choice of how to detect whether or not they are flat.

One aspect of a dataset that may cause the problems noted by Hamill (2001) is that it may be inhomogeneous. Provided that the dataset is large, this can be overcome by dividing the data into more homogeneous subsets. Decomposition of T can then be done separately in each subset, and insights can be gained about the behavior of an ensemble forecasting system if the deviations from uniformity (or lack of them) change for different time periods or different geographical regions.

Finally, as with other tests for uniformity, there is an assumption of independence of the observations that make up the rank histogram. This was noted in one of the examples above. One possible way to avoid dependence is to systematically sample at much coarser spatial and temporal separations. Hamill (2001) demonstrates how to investigate whether nonindependence is present, and how to determine the temporal and spatial spacing needed to avoid it. Another possibility is to treat the values at different spatial locations as different variables and use a minimum spanning tree (MST) approach to reduce the many variables to a one-dimensional problem in which a rank histogram is constructed based on lengths of MSTs when the verification observation replaces each ensemble member in turn (see Wilks 2004; Smith and Hansen 2004). Slightly different considerations are needed for rank histograms based on MSTs (Wilks 2004), but the decomposition of T would still be applicable. If both spatial and temporal correlations are present, the spatial aspect can be handled by the MST approach, while adjustments for temporal correlation can be made to the critical values of T (Wilks 2004).

Various caveats concerning the use of the decomposition technique have been noted in this section, but despite these we are convinced that it is a valuable additional tool in assessing deviations from uniformity in a rank histogram.

Acknowledgments. We are grateful to Kimberley Elmore, Jim Hansen, and an anonymous reviewer for comments and suggestions that led to improvements in the paper. Yuejian Zhu and Zoltan Toth provided the 500-hPa height data, and suggested appropriate degrees of freedom for the data.

APPENDIX

Formulas for the Decomposition of the χ^2 Test Statistic

This section provides formulas for the Linear, V-shape, and Ends contrasts that have been used in the earlier examples. U-shaped contrasts are also given.

Suppose that a rank histogram has k bins. The vector $\mathbf{l}_{\mathrm{LIN}}$ defining the linear contrast is proportional to $[a, a + b, a + 2b, \ldots, a + (k - 1)b]$, and the vector $\mathbf{l}_{\mathrm{ENDS}}$ is proportional to $(a, -b, -b, \ldots, -b, -b, a)$, with different constants a, b, depending on k, for the two cases. The vector $\mathbf{l}_{\mathrm{V}}$ defining the V-shaped contrast is different for k even and k odd. For k even ($=2h$), the vector is proportional to $[a, a - b, \ldots, a - (h - 1)b, a - (h - 1)b, \ldots, a - b, a]$, and for k odd ($=2h + 1$), the vector is proportional to $[a, a - b, \ldots, a - (h - 1)b, a - hb, a - (h - 1)b, \ldots, a - b, a]$, again for different a, b.

Table A1 gives formulas for a and b in terms of k or h. To implement the decomposition technique, the vectors implied by Table A1 must be standardized to have

TABLE A1. Values of a and b in the formulas for vectors defining Linear, Ends, and V-shape contrasts for k bins.

		a	b
Linear	$k = 2h + 1$ (odd)	$-h$	1
	$k = 2h$ (even)	$-(2h - 1)$	2
Ends	$k = 2h + 1$ (odd)	$(2h - 1)$	2
	$k = 2h$ (even)	$(h - 1)$	1
V-shape	$k = 2h + 1$ (odd)	h^2	$2h + 1$
	$k = 2h$ (even)	$(h - 1)$	2

length 1, by dividing all elements by the square root of the sum of squares of the elements.

These contrasts are not the only possible ones. For example, a U-shaped, or quadratic, alternative may be of more interest than the Ends or V-shaped alternative. For this contrast, as with the V-shaped contrast, different formulas are relevant for k even and k odd. For k odd ($=2h + 1$), the relevant vector $\mathbf{l}_U$ is proportional to $[h^2 - b, (h - 1)^2 - b, \ldots, 1 - b, -b, 1 - b, \ldots, (h - 1)^2 - b, h^2 - b]$, where $b = h(h + 1)/3$. For k even ($=2h$), the relevant vector is $[(2h - 1)^2 - b, (2h - 3)^2 - b, \ldots, 9 - b, 1 - b, 1 - b, 9 - b, \ldots, (2h - 3)^2 - b, (2h - 1)^2 - b]$, where b is now $(4h^2 - 1)/3$.

Yet other contrasts are possible and limited only by the imagination of the user, but in practice those discussed above are likely to correspond to the most relevant alternatives to flatness.

REFERENCES

Boero, G., J. Smith, and K. Wallis, 2004: The sensitivity of chi-squared goodness-of-fit tests to the partitioning of data. *Econ. Rev.,* **23,** 341–370.

Choulakian, V., R. A. Lockhart, and M. A. Stephens, 1994: Cramér–von Mises statistics for discrete distributions. *Can. J. Stat.,* **22,** 125–137.

Elmore, K. L., 2005: Alternatives to the chi-square test for evaluating rank histograms from ensemble forecasts. *Wea. Forecasting,* **20,** 789–795.

Gneiting, T., F. Balabdaoui, and A. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.,* **69B,** 243–268.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.,* **129,** 550–560.

Kendall, M. G., and A. Stuart, 1967: *The Advanced Theory of Statistics.* 2nd ed. Vol. 2, *Inference and Relationship,* Charles Griffin, 690 pp.

Noceti, P., J. Smith, and S. Hodges, 2003: An evaluation of tests of distributional forecasts. *J. Forecasting,* **22,** 447–455.

Smith, L., and J. Hansen, 2004: Extending the limits of ensemble forecast verification with the minimum spanning tree. *Mon. Wea. Rev.,* **132,** 1522–1528.

Wilks, D., 2004: The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. *Mon. Wea. Rev.,* **132,** 1329–1340.