



Unanimity and Compromise among Probability Forecasters

Robert T. Clemen, Robert L. Winkler

Management Science, Volume 36, Issue 7 (Jul., 1990), 767-779.

Stable URL:

<http://links.jstor.org/sici?sici=0025-1909%28199007%2936%3C767%3AUACAPF%3E2.0.CO%3B2-7>

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

Management Science is published by INFORMS. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/informs.html>.

Management Science
©1990 INFORMS

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact jstor-info@umich.edu.

©2002 JSTOR

UNANIMITY AND COMPROMISE AMONG PROBABILITY FORECASTERS*

ROBERT T. CLEMEN AND ROBERT L. WINKLER

College of Business Administration, University of Oregon, Eugene, Oregon 97403
Fuqua School of Business, Duke University, Durham, North Carolina 27706

When two forecasters agree regarding the probability of an uncertain event, should a decision maker adopt that probability as his or her own? A decision maker who does so is said to act in accord with the unanimity principle. We examine a variety of Bayesian consensus models with respect to their conformance (or lack thereof) to the unanimity principle and a more general compromise principle. In an analysis of a large set of probability forecast data from meteorology, we show how well the various models, when fit to the data, reflect the empirical pattern of conformance to these principles.

(COMBINING PROBABILITIES; CONSENSUS; UNANIMITY; WEATHER FORECASTING)

1. Introduction

Imagine a situation in which two forecasters provide a decision maker (DM) with their probabilities for the occurrence of an uncertain event. If the forecasters' probabilities are equal, should DM take that probability as his posterior probability of the event or not? If the forecasters disagree, should DM's posterior probability be equal to one or the other of the probabilities, between them, or might it be reasonable for DM's posterior probability to be outside the range of the two probabilities?

If a decision maker adopts the commonly-stated probability of two forecasters, we will say that the decision maker abides by the "unanimity principle." This principle was one of the considerations that motivated Morris (1983) to develop an axiomatic approach to the combination of probabilities. Morris's paper in turn stimulated a debate involving the appropriateness of the unanimity principle as well as other issues (Clemen 1986; French 1986; Lindley 1986; Morris 1986; Schervish 1986; Winkler 1986). A special case of the unanimity principle, called the zero probability property (McConway 1981) or the zero preservation property (Genest and Zidek 1986), applies only when two or more forecasters all assign probability zero to an event. Many properties have been considered in the literature on combining probabilities; the unanimity principle and the zero preservation property are members of "a general class of axioms which would require the consensus distribution to embrace any aspect of the [forecasters'] personal opinions that are already the object of an (implicit) agreement between them" (Genest and Zidek 1986, p. 117).

Of course, the unanimity principle only makes sense when the forecasters agree on the probability of the event. In many situations, forecasters give different probabilities for the same event. A more general principle might be a compromise principle, by which DM's posterior probability should lie between the forecasters' probabilities; this subsumes the unanimity principle as a special case.

The combination of probabilities is a topic worth serious consideration. Many situations require a decision maker to deal with probability assessments from multiple sources. These situations include expert testimony in court cases, risk assessment and analysis, use of uncertainty in artificial intelligence applications, and others. Following Morris (1974), a Bayesian approach to the problem suggests that combining probabilities be

* Accepted by Irving H. LaValle; received January 13, 1989. This paper has been with the authors 1 month for 1 revision.

left up to the decision maker, whose responsibility it is to specify his or her own beliefs regarding the performance and inter-relatedness of the probability sources. To aid decision makers in this somewhat daunting task, a number of aggregation models have been proposed. While decision makers must carefully assess their individual situations in order to apply the models correctly, a natural issue that many decision makers may wish to consider is under what conditions a model conforms to the unanimity and compromise principles. If a decision maker decides to use a specific model, he or she should know, qualitatively at least, what the model implies.

In §2 of this paper we briefly discuss a number of probability-aggregation models that have been proposed. We restrict our attention to aggregation models that follow the Bayesian paradigm just described. (For discussions of models that do not follow the Bayesian approach, see Genest and Zidek 1986 and French 1985.) In general, we are able to identify three classes of models, and the classification has to do with how the idea of conditional independence is used in the development of the models. The three classes differ dramatically in the way they conform (or fail to conform) to the unanimity and compromise principles.

In many cases, a decision maker will have to combine probabilities using only subjective assessments of performance and dependence. However, in some instances sufficient data are available to examine empirically whether unanimity or compromise is reasonable. In §3 of the paper, we study a large sample of probability assessments from meteorology. In our analysis, we examine empirically the degree to which unanimity and compromise are followed, and we demonstrate how well (or how poorly) the various models described in §2 are able to mimic the empirical results.

Our aim is not to convince anyone that the unanimity and compromise principles are uniformly appropriate in all situations. Indeed, the unanimity and compromise principles might be viewed as “ad hoceries,” and we would argue that the proper normative use of expert information is via the Bayesian paradigm. Decision makers should think hard about their specific situations and choose models that are appropriate. In doing so, however, some decision makers may find it useful to think (at least initially) in terms of the unanimity and compromise principles. Our empirical analysis of meteorological assessments indicates the extent to which these principles are appropriate in this area, and in §4 we speculate on underlying reasons for this. We hope that our analysis and discussion will provide guidance for decision makers in assessing and understanding their situations and in applying available probability-aggregation models appropriately.

2. Models for Combining Probabilities

In this section we review a number of Bayesian probability-aggregation models that have been proposed. Our intent is to indicate the extent to which these models agree with or do not agree with the unanimity and compromise principles. In reviewing the literature, we have found that many authors have already examined their models in this regard; thus, our discussion here is brief. For more thorough discussions of the individual models, we refer the interested reader to the cited references.

Conditional independence is a typical assumption made by the Bayesian model builders who have generated the models we discuss here. However, it is possible to invoke conditional independence in a variety of ways. We have found it convenient to classify models according to the way conditional independence is used. The first possibility is to assume that the forecasters are conditionally independent given the event they are forecasting. A second possibility is to assume a parametric model with forecasters' data sets conditionally independent given the value of an unknown parameter (*not* given the event being forecast). A third possibility is to assume that the forecasters' data sets and hence their forecasts are not conditionally independent.

In the following pages, we will use the following conventions. The uncertain event of interest is random variable A which can take on values of 1 or 0. There are k sources (forecasters), and source i provides probability p_i for the occurrence of A . In some instances, source i 's probability is viewed as a random variable. We abuse notation slightly by using p_i to represent the random variable as well as its realization. The decision maker's prior probability is given by p_0 and posterior probability by $p^* = P(A|p_0, p_1, \dots, p_k)$.

2.1. Conditional Independence Given Event A or $\bar{A}$

With k sources, we begin with Bayes' theorem in odds form:

$$\frac{p^*}{1 - p^*} = \frac{p_0}{1 - p_0} \prod_{i=1}^k \frac{P(p_i|A, p_0, p_1, \dots, p_{i-1})}{P(p_i|\bar{A}, p_0, p_1, \dots, p_{i-1})}. \quad (1)$$

Assuming independence among the forecasters and DM implies that, conditional on either A or $\bar{A}$, the chance of forecaster i providing p_i does not depend on the values of $p_0, p_1, \dots, p_{i-1}$. Formally,

$$\frac{P(p_i|A, p_0, p_1, \dots, p_{i-1})}{P(p_i|\bar{A}, p_0, p_1, \dots, p_{i-1})} = \frac{P(p_i|A)}{P(p_i|\bar{A})} \quad (2)$$

for $i = 1, \dots, k$, from which (1) becomes

$$\frac{p^*}{1 - p^*} = \frac{p_0}{1 - p_0} \prod_{i=1}^k \frac{P(p_i|A)}{P(p_i|\bar{A})}. \quad (3)$$

We will call (3) the "Independence" model. Using this model with $k = 2$, $p_0 = 0.50$, $p_1 = 0.55$, $p_2 = 0.55$, and under suitable assumptions regarding the forecasters' performance, DM's posterior probability of rain is 0.60. If there were ten such independent forecasters, DM's posterior probability would be 0.90. These results are in clear disagreement with the unanimity principle.

In terms of the compromise principle, the Independence model is designed in such a way that if both p_1 and p_2 are greater (less) than 0.50, then p^* is greater (less) than the larger (smaller) of the two input probabilities. For more than two forecasters, it is necessary (but not sufficient) that at least one probability be greater and one less than 0.50 for p^* to be within the range of the forecasters' probabilities. Thus, the Independence model does not obey unanimity, nor does it obey the compromise principle uniformly.

Genest and Schervish (1985) (hereafter GS) use conditional independence in the same way to develop a similar aggregation rule in their Theorem 4.1. Their approach is ingenious, viewing the problem in terms of a DM willing to assess only certain aspects of the marginal distribution of source i 's probability p_i . Their combination formula, which we will call the "GS-I" model, is given by

$$p^* = \frac{p_0^{1-k} \prod_{i=1}^k \pi_i}{p_0^{1-k} \prod_{i=1}^k \pi_i + (1 - p_0)^{1-k} \prod_{i=1}^k (1 - \pi_i)}, \quad (4)$$

where $\pi_i = p_0 + \lambda_i(p_i - \mu_i)$, μ_i is DM's marginal expected value of p_i , and λ_i is interpreted as DM's subjective assessment of the coefficient of linear regression of A on p_i . The value π_i in (4) is GS's adjusted p_i ; that is, if DM consulted only source i and obtained probability p_i , (4) implies that $p^* = \pi_i$. Because (4) was developed under the assumption of independence of the p_i 's conditional on A or $\bar{A}$, each λ_i is assessed individually for each source i subject to constraints that ensure $0 \leq \pi_i \leq 1$.

In terms of unanimity and compromise, GS note that if all of the forecasters say exactly what DM expects them to say ($p_i = \mu_i$ for each i), DM's posterior probability is the same as his or her prior probability. On the other hand, if they all provide probabilities greater

(less) than expected, p^* will be greater (less) than the largest (smallest) of the π_i 's. However, p^* may still lie within the range of the unadjusted p_i 's.

2.2. Conditional Independence among Forecasters' Data Sets

An example of this approach is Morris's (1983) Bernoulli model for combining probabilities of events. It is also discussed in detail in Winkler (1986). The essence of the argument is that p_0 is considered to be the mean of a second-order probability distribution for p (or the mean of a distribution of the relative frequency p of occurrence of A if the situation could be repeated infinitely in a series of exchangeable trials). In particular, we might suppose that DM's distribution for p is a beta distribution with parameters r_0 and n_0 ,

$$f(p|r_0, n_0) \propto p^{r_0-1}(1-p)^{n_0-r_0-1}, \quad (5)$$

in which case $p_0 = P(A) = E(p|r_0, n_0) = r_0/n_0$. Likewise, we assume that DM views the i th forecaster's probability p_i as being based on information equivalent to a sample of n_i independent Bernoulli trials with probability p yielding r_i occurrences of A . If DM assumes that, conditional on p , the values $p_0, p_1, \dots, p_k$ are independent, then the appropriate combination of the probabilities is a convex combination of $p_0, p_1, \dots, p_k$ (Morris 1983):

$$p^* = \sum_{i=0}^k \beta_i p_i, \quad (6)$$

where $\beta_i = n_i / \sum n_j$. We will refer to (6) as the "Bernoulli" model. If DM is assumed to have very little information about p relative to the forecasters, then β_0 would be nearly zero. In the limit, as n_0 approaches zero, p^* becomes a convex combination of $p_1, \dots, p_k$, obeying the unanimity and compromise principles uniformly.

In §2.1 we described models in which the forecasts are independent at the level of providing information about Event A directly, whereas here the independence has to do with providing information about the parameter p . In the Bernoulli model, the likelihood ratios corresponding to those given in (2) for the Independence model are

$$\frac{P(p_i|A, p_0, p_1, \dots, p_{i-1})}{P(p_i|\bar{A}, p_0, p_1, \dots, p_{i-1})} = \frac{\int_0^1 P(p_i|p)f(p|A, p_0, p_1, \dots, p_{i-1})dp}{\int_0^1 P(p_i|p)f(p|\bar{A}, p_0, p_1, \dots, p_{i-1})dp}. \quad (7)$$

The conditional independence at the level of p in Model 2 does not translate into conditional independence at the level of predictions for A and $\bar{A}$.

2.3. Dependence among Forecasters' Probabilities

In this situation, dependence among expert probabilities is explicitly modeled. For example, Clemen (1987) provides a model similar to the Bernoulli model described above. As above, forecaster i 's probability is modeled as if it arises from observation of n_i independent Bernoulli trials. However, forecasters i and j may observe some observations in common, resulting in dependence among the stated p_i 's. DM aggregates this information into a probability distribution for p , the mean of which is p^* . In his article, Clemen demonstrates that this model does not obey the compromise principle uniformly.

Other models in this class include those in which the forecasters' log-odds of A are assumed to follow a multinormal distribution conditional on whether A or $\bar{A}$ occurs (French 1981; Lindley 1985; Clemen and Winkler 1987). Following Clemen and Winkler, define $q_i = \log [p_i/(1 - p_i)]$, and $q = (q_1, \dots, q_k)'$. The DM's posterior log-odds for A would be given by

$$q^* = \log \frac{P(A|q)}{P(\bar{A}|q)} = \log \frac{L(q|A)}{L(q|\bar{A})} + \log \frac{p_0}{1 - p_0}. \quad (8)$$

The likelihood functions $L(q|A)$ and $L(q|\bar{A})$ are modeled as being proportional to normal distributions for q with mean vectors M_1 and M_0 and covariance matrices Σ_1 and Σ_0 , respectively. The individual q_i 's might be thought of as arising from each forecaster's observation of independent trials (conditional on A or $\bar{A}$) from a normal process, and dependence among the elements of q (nonzero off-diagonal elements of Σ_1 and Σ_0) could arise through the observation of overlapping data by forecasters (Clemen 1987). A univariate version of this model also permits the adjustment or calibration of individual probabilities (Lindley 1982).

Given these specifications for the likelihood functions, DM's posterior log-odds are given by

$$q^* = \{\log(|\Sigma_0|/|\Sigma_1|) - q'(\Sigma_1^{-1} - \Sigma_0^{-1})q + 2q'(\Sigma_1^{-1}M_1 - \Sigma_0^{-1}M_0) - M_1'\Sigma_1^{-1}M_1 + M_0'\Sigma_0^{-1}M_0\}/2 + \log[p_0/(1 - p_0)]. \quad (9)$$

If the covariance matrices are equal ($\Sigma_0 = \Sigma_1 = \Sigma$), this expression reduces to

$$q^* = q'\Sigma^{-1}(M_1 - M_0) - (M_1 + M_0)'\Sigma^{-1}(M_1 - M_0)/2 + \log[p_0/(1 - p_0)]. \quad (10)$$

We will refer to equations (9) and (10) as the "Log-odds I" and "Log-odds II" models, respectively. Casual inspection of these equations reveals that neither one will necessarily obey either the unanimity or the compromise principles. Taking the simpler (10), q^* is clearly a linear combination of the q_i 's. However, the weights are given by $\Sigma^{-1}(M_1 - M_0)$, and do not generally sum to one. Furthermore, q^* also includes the constant term $-(M_1 + M_0)'\Sigma^{-1}(M_1 - M_0)/2 + \log[p_0/(1 - p_0)]$. Assuming that $p_0 = 0.50$, a rather odd set of sufficient conditions for unanimity and compromise to hold would be $M_0 = -M_1$, $\Sigma^{-1}M_1 > 0$, and $e'\Sigma^{-1}M_1 = 0.5$, where e' is a conformable vector of ones. These conditions, for which intuition fails us, are related to Lindley's (1982) conditions for probability calibration.

Genest and Schervish (1985) develop a general model for the combination of probabilities from multiple experts, of which the GS-I model discussed above is a special case. Theorem 3.2 in Genest and Schervish (1985) gives the following formula for combining $p_1, \dots, p_k$:

$$p^* = p_0 + \sum_{i=1}^k \lambda_i(p_i - \mu_i), \quad (11)$$

where μ_i is defined as above and the vector $\lambda = (\lambda_1, \dots, \lambda_k)$ is interpreted as the vector of coefficients of linear regression of A on $(p_1, \dots, p_k)$. (Actually, the published formula inadvertently left out p_0 from the expression due to a typographical error.) As in the GS-I model, the λ_i 's are subject to a number of constraints to ensure that $0 \leq p^* \leq 1$. We will refer to (11) as the GS-II model. Whether such a combination obeys unanimity and compromise as defined here clearly depends on the specific values of the p_i 's, μ_i 's, p_0 , and λ .

3. Aggregating Probabilities: Some Empirical Results

In order to study the unanimity principle empirically, we need a large sample of data, since the relevant sample sizes are the numbers of times different combinations of probability values occur. One area with extensive data on probability forecasts is weather forecasting. Since 1966 the National Weather Service (NWS) of the United States has formulated and issued probability of precipitation (PoP) forecasts. These forecasts indicate a probability of measurable precipitation at a particular location during a specified time period. Although precipitation can take on different forms, for convenience we will refer

to the occurrence and nonoccurrence of precipitation as “rain” ($A = 1$) and “no rain” ($A = 0$).

For any given forecast area and forecast period, two PoP forecasts are prepared. The forecast actually issued to the public is a local forecast made by a weather forecaster in the local NWS office. In addition, the NWS prepares forecasts based on a numerical-statistical model of the global atmospheric system. We call these forecasts “guidance forecasts” because they are supplied to the local forecasters for use in preparing the official local forecasts. Meteorologists have studied the relative performance of local and guidance forecasts; a review of this literature and a more complete review of the forecasting process are given in Murphy and Winkler (1984).

The data analyzed in this paper consist of local and guidance PoP forecasts for the NWS office in Boston. These data, covering the period from April 1972 through September 1983, were provided by the NWS Techniques Development Laboratory. Local and guidance forecasts are made twice each day, in the morning and evening. On each occasion forecasts are formulated for three consecutive 12-hour periods, or lead times: 12–24 hours, 24–36 hours, and 36–48 hours after the guidance forecast is made. In all, we analyzed a total of 12,885 pairs of local and guidance forecasts.

One concern with our analysis is that the overall probability of precipitation be roughly the same over time. Pooling data for which this is not the case can lead to misleading results; the pooled data may appear to conform to the unanimity and compromise principles even though the individual data sets do not. For our meteorological problem, the issue is whether the climatological probability of rain (overall relative frequency) is stable throughout the year. Meteorologists typically divide the year into cool (October–March) and warm (April–September) seasons. The climatological probabilities of rain for the warm and cool seasons were 0.2330 and 0.2145, respectively, with corresponding sample sizes of 2126 and 2169 forecast occasions. On the basis of these calculations we concluded that there was not a material difference in the two proportions and pooled the data from both warm and cool seasons for our analysis. The pooled climatological probability of rain was 0.2237.

For each (p_l, p_g) , where p_l represents the local PoP forecast and p_g represents the guidance PoP forecast, we looked at all occasions with forecast values (p_l, p_g) and found the relative frequency of occurrence of precipitation over those occasions. These relative frequencies, along with the number of occasions on which each is based, are given in Table 1. For example, the combination with $p_l = 0.30$ and $p_g = 0.40$ was observed 123 times, and precipitation occurred on 30.1 percent of these occasions.

Considering first the unanimity principle, we are interested in the cells in Table 1 for which $p_l = p_g = p$. These cells constitute 4181, or 32.5 percent, of the pairs in our data set. We can calculate the “average disagreement with unanimity” as the average absolute difference between p and the relative frequency ($\hat{p}$), weighted by the number of observations for each value of p . For our data, this average absolute difference is 0.028. Furthermore, it is interesting to note that in nine of the eleven cases we have $\hat{p} < p$, the only exceptions being when $p = 0.0$ or 0.8 .

With regard to the broader compromise principle, we must look at situations where $p_l \neq p_g$. Of such cells, 25 (marked with asterisks) have relative frequencies that are not between p_l and p_g , but 14 of these cases involve sample sizes of five or fewer observations. In total, these cells account for 752 observations, or 5.8 percent, of the pairs in our data set. Aside from those cells with five or fewer observations, all of the cells in disagreement with the compromise principle lie near the main diagonal and for values of $p_l \geq 0.30$ and $p_g \geq 0.50$. Of the eleven cells in question, $\hat{p} < \min(p_l, p_g)$ in eight cases, and $\hat{p} > \max(p_l, p_g)$ for the pairs $(0.7, 0.8)$, $(0.8, 0.7)$, and $(0.9, 0.8)$.

Overall, including all possible pairs (p_l, p_g) , we find 36 cells that do not conform to the compromise principle (which subsumes unanimity). These constitute 38.3 percent

TABLE 1
Relative Frequency of Precipitation and Sample Sizes

Local:	Guidance:														Total
	0.00	0.02	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00		
0.00	0.009	0.014	0.014	0.027	0.061	0.035	0.267	0.500	0.000		0.000			0.018	
0.10	1595	560	586	565	163	57	15	4	1					3547	
	0.026	0.068	0.063	0.063	0.103	0.161	0.143	0.167	0.333	*	0.000	*	0.000	0.073	
0.20	463	251	474	865	504	155	42	12	6		1	1	1	2778	
	0.027	0.052	0.080	0.125	0.129	0.235	0.268	0.271	0.226		0.571	*	1.000	0.164	
0.30	111	77	162	385	611	371	194	107	53		21	1	1	2103	
	0.135	0.294	0.136	0.169	0.233	0.274	0.301	*	0.275	*	0.290	*	0.000	0.244	
0.40	37	17	44	130	219	237	123	51	31		9	1		903	
	0.385	*	0.000	0.238	0.330	0.351	0.377	*	0.348	*	0.339	0.414	1.000	0.339	
0.50	13	4	21	61	109	148	167	92	62		29	3	2	716	
	*	0.600	*	0.000	0.361	0.350	0.355	0.410	0.457	*	0.395	0.618	0.667	0.417	
0.60	5	8	5	36	80	121	139	140	86		46	6	6	712	
	0.167	*	0.000	0.143	0.314	0.483	0.563	*	0.495	0.573	0.645	*	0.500	0.521	
0.70	6	1	7	23	51	58	103	107	124		62	10	3	587	
	0.000			0.333	0.471	0.613	0.500	0.589	0.658	*	0.833	0.700	1.000	0.654	
0.80	3			6	17	31	32	56	73		101	20	3	402	
	*	1.000	0.500	0.333	0.615	0.429	0.595	0.655	0.741	*	0.816	*	0.791	0.749	
0.90		1	2	3	13	21	42	58	81		125	67	10	537	
			*	1.000	*	1.000	0.833	0.750	0.750	*	0.933	0.893	0.963	0.848	
1.00					1	6	8	14	20		34	75	27	231	
	0.000			1.000	0.333	1.000	1.000	1.000	0.857	0.885	0.904	0.944	0.979	0.935	
	1			1	3	2	2	9	21	26	73	90	141	369	
Total:	0.019	0.039	0.050	0.085	0.164	0.279	0.376	0.435	0.520	0.691	0.813	0.836	0.938	0.224	
	2234	919	1302	2075	1771	1207	867	650	558	443	391	274	194	12885	

of the pairs in our data set. Aside from the off-diagonal cells with sample sizes less than five, the overall pattern is one in which $\hat{p}$ most often is less than both p_l and p_g , with a few exceptions, mostly for fairly high values of p_l and p_g .

Our next step is to fit the combination models discussed in §2 to these data. That is, suppose that DM was interested in aggregating p_l and p_g via one of the models to arrive at p^* . In the absence of additional information, DM would want p^* to approximate $\hat{p}$ reasonably well for the possible combinations of p_l and p_g . Our data set is large enough to estimate the required parameters for each model and to calculate p^* for all (p_l, p_g) pairs. Also, we can compare the models in terms of the goodness of the approximation of $\hat{p}$ with p^* .

Before we proceed, however, it is important to consider the characteristics of our data and the appropriateness of the various models. Given that the guidance forecast is available and used in the formulation of the official local forecast, it does not make sense to assume that the two forecasts are conditionally independent given the weather. Indeed, we might expect them to be highly dependent, and our primary candidates for models appear to be those that permit conditional dependence among the forecasts. On the surface it would appear that the Independence and GS-I models are inappropriate. Also, since the data sets used by the two forecasters overlap, the Bernoulli model appears inappropriate. However, a substantial and growing literature in the forecasting area has documented the phenomenon that simple forecast combination methods perform well under a wide variety of circumstances. In particular, it has been shown (e.g., Granger and Newbold 1977, Winkler and Makridakis 1983, Clemen and Winkler 1986) that combination models that ignore forecast dependence often perform better in practice than those that attempt to exploit the dependence. We speculate that the same may be true in the case of combining probabilities, and so we believe that it is worthwhile to fit the simpler models in order to compare their performance with that of the more complicated models.

Table 2 presents summary information regarding the models that we fit to the PoP forecast data. In all cases, the local and guidance forecasters played the part of forecasters 1 and 2, respectively, in the models. In reporting the results we have retained the subscripts l and g for convenience. The prior probability p_0 was taken to be equal to the climatological probability of rain (0.2237).

TABLE 2
Model Estimation and Performance Statistics. Estimates Are So Designated with a “^”

Model	Estimation	MSE	MAD
Independence	$p^* = \frac{P(p_l A)P(p_g A)p_0}{P(p_l A)P(p_g A)p_0 + P(p_l \bar{A})P(p_g \bar{A})(1 - p_0)}$ (See Table 3)	0.01222	0.07112
Bernoulli	$p^* = 0.601p_l + 0.399p_g$	0.00272	0.03461
GS-I	$p^* \text{ given by equation (4) with:}$ $\hat{\mu}_l = 0.252 \quad \hat{\lambda}_l = 0.600$ $\hat{\mu}_g = 0.239 \quad \hat{\lambda}_g = 0.575$	0.00331	0.03427
GS-II	$p^* = 0.558p_l + 0.346p_g$ ($\hat{\mu}_l$ and $\hat{\mu}_g$ as in GS-I)	0.00268	0.03243
Log-odds I	$q^* = 0.067 + 0.525q_l + 0.360q_g$ $- 0.116q_l^2 - 0.120q_g^2 + 0.219q_lq_g$	0.00425	0.04108
Log-odds II	$q^* = -0.190 + 0.476q_l + 0.422q_g$	0.00265	0.02872

TABLE 3
TABLE 3a Likelihood Function: Relative Frequency of Joint Forecast Given Rain (A)

Guidance:	0.00	0.02	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	$P(p_i A)$
Local: 0.00	0.0048	0.0028	0.0028	0.0052	0.0035	0.0007	0.0014	0.0007	0.0000	0.0000	0.0000	0.0000	0.0000	0.0218
0.10	0.0042	0.0059	0.0104	0.0188	0.0180	0.0087	0.0021	0.0007	0.0007	0.0000	0.0007	0.0000	0.0000	0.0701
0.20	0.0010	0.0014	0.0045	0.0167	0.0274	0.0302	0.0180	0.0101	0.0042	0.0042	0.0017	0.0003	0.0000	0.1197
0.30	0.0017	0.0017	0.0021	0.0076	0.0177	0.0226	0.0128	0.0049	0.0031	0.0014	0.0007	0.0000	0.0000	0.0763
0.40	0.0017	0.0000	0.0017	0.0045	0.0125	0.0180	0.0219	0.0111	0.0073	0.0042	0.0007	0.0000	0.0007	0.0843
0.50	0.0010	0.0003	0.0000	0.0045	0.0097	0.0149	0.0198	0.0222	0.0118	0.0090	0.0073	0.0014	0.0010	0.1030
0.60	0.0003	0.0000	0.0003	0.0028	0.0056	0.0097	0.0201	0.0184	0.0246	0.0139	0.0080	0.0017	0.0007	0.1062
0.70	0.0000	0.0000	0.0000	0.0007	0.0028	0.0066	0.0056	0.0115	0.0167	0.0243	0.0173	0.0049	0.0010	0.0913
0.80	0.0000	0.0003	0.0003	0.0003	0.0028	0.0031	0.0087	0.0132	0.0208	0.0323	0.0364	0.0184	0.0028	0.1395
0.90	0.0000	0.0000	0.0003	0.0000	0.0003	0.0017	0.0021	0.0024	0.0052	0.0090	0.0146	0.0232	0.0090	0.0680
1.00	0.0000	0.0000	0.0000	0.0003	0.0003	0.0007	0.0007	0.0031	0.0062	0.0080	0.0229	0.0295	0.0479	0.1197
$P(p_g A)$	0.0149	0.0125	0.0226	0.0614	0.1006	0.1169	0.1131	0.0982	0.1006	0.1062	0.1103	0.0795	0.0631	

TABLE 3b Likelihood Function: Relative Frequency of Joint Forecast Given No Rain ($\bar{A}$)

Guidance:	0.00	0.02	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	$P(p_i \bar{A})$
Local: 0.00	0.1581	0.0552	0.0578	0.0550	0.0153	0.0055	0.0011	0.0002	0.0001	0.0000	0.0001	0.0000	0.0000	0.3483
0.10	0.0451	0.0234	0.0444	0.0811	0.0452	0.0130	0.0036	0.0010	0.0004	0.0001	0.0001	0.0001	0.0001	0.2575
0.20	0.0108	0.0073	0.0149	0.0337	0.0532	0.0284	0.0142	0.0078	0.0041	0.0009	0.0004	0.0000	0.0001	0.1757
0.30	0.0032	0.0012	0.0038	0.0108	0.0168	0.0172	0.0086	0.0037	0.0022	0.0005	0.0002	0.0001	0.0000	0.0683
0.40	0.0008	0.0004	0.0016	0.0048	0.0073	0.0096	0.0104	0.0060	0.0041	0.0017	0.0003	0.0003	0.0000	0.0473
0.50	0.0002	0.0007	0.0005	0.0023	0.0052	0.0078	0.0082	0.0076	0.0052	0.0020	0.0013	0.0002	0.0003	0.0415
0.60	0.0005	0.0001	0.0006	0.0015	0.0035	0.0030	0.0045	0.0054	0.0053	0.0022	0.0009	0.0005	0.0001	0.0281
0.70	0.0003	0.0000	0.0000	0.0004	0.0009	0.0012	0.0016	0.0023	0.0025	0.0031	0.0010	0.0006	0.0000	0.0139
0.80	0.0000	0.0000	0.0001	0.0002	0.0005	0.0012	0.0017	0.0020	0.0021	0.0021	0.0020	0.0014	0.0002	0.0135
0.90	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0005	0.0008	0.0003	0.0008	0.0001	0.0035
1.00	0.0001	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000	0.0003	0.0003	0.0007	0.0005	0.0003	0.0024
$P(p_g \bar{A})$	0.2190	0.0883	0.1237	0.1897	0.1481	0.0870	0.0541	0.0367	0.0268	0.0137	0.0073	0.0045	0.0012	

TABLE 4
*Estimates of Mean Vectors and Covariance Matrices
of Forecast Log-Odds for the Log-Odds I and
Log-Odds II Models. The First Element
Represents the Local Forecast.*

$\hat{M}'_1 = (0.39, 0.09)$	$\hat{M}'_0 = (-2.67, -2.54)$
$\hat{\Sigma}_1 = \begin{bmatrix} 3.59 & 2.48 \\ 2.48 & 3.03 \end{bmatrix}$	$\hat{\Sigma}_0 = \begin{bmatrix} 4.53 & 2.57 \\ 2.57 & 3.45 \end{bmatrix}$
$n_1 = 2882$	$n_0 = 10,003$
Pooled estimate of covariance matrix for Log-odds II model:	
$\hat{\Sigma} = \begin{bmatrix} 4.18 & 2.55 \\ 2.55 & 3.35 \end{bmatrix}$	

For the Bernoulli model, β_l and β_g were estimated using least squares, with these coefficients constrained to sum to one. For the GS-I and GS-II models, the μ_i 's were estimated with the sample means ($\hat{\mu}_i$), and the λ_i 's were estimated using least squares, subject to the appropriate constraints.

Because the Independence, Log-odds I, and Log-odds II models require the likelihood functions, we have disaggregated the data and displayed it in Tables 3a and 3b. These tables show the relative frequency of forecast pairs (p_l, p_g) given rain (A) and no rain ($\bar{A}$), respectively. The marginal entries in these tables give the relative frequencies $P(p_l|A)$, $P(p_l|\bar{A})$, $P(p_g|A)$, and $P(p_g|\bar{A})$ which are used in the Independence model. The Log-odds models require the fitting of normal distributions to the two joint likelihood functions after a transformation to log-odds. For these models, forecast values of 0 and 1 were replaced with values of 0.005 and 0.98, respectively. Table 4 shows estimates of mean vectors M_1 and M_0 and of covariance matrices Σ_1 and Σ_0 for the Log-odds I model and the pooled estimate Σ for the Log-odds II model. These estimates were used in equations (9) and (10).

Table 2 also shows some summary performance measures. MSE (mean squared error) was calculated in the usual way. For each forecast pair (p_l, p_g) we found the squared difference between the empirical relative frequency ($\hat{p}$) as shown in Table 1 and the model's predicted relative frequency (p^*). MSE is then the weighted average of these differences, the weights being the proportion of observations in each cell. MAD measures mean absolute deviations; it is calculated the same as MSE but with absolute rather than squared differences.

On the basis of Table 2 we can conclude that the Independence model fits the data poorly compared to the other models. On the other hand, the GS-I model performs reasonably well, with error statistics that are comparable to the other models. Considering the remaining four models, the Bernoulli model, GS-II, and Log-odds II are virtually equivalent in terms of MSE, but Log-odds II has a slightly lower MAD.

Figure 1 provides graphic views of the patterns of disagreement with compromise for the Independence (Figure 1a), Log-odds I and Log-odds II (Figure 1b), and GS-I and GS-II (Figure 1c) models. (The Bernoulli model is not included since it conforms to unanimity and compromise uniformly.) Furthermore, the empirical pattern of disagreement [$\hat{p} < \min(p_l, p_g)$ or $\hat{p} > \max(p_l, p_g)$] is shown via shadings of the cells. (No comparisons are made for cells with less than ten empirical observations.) Thus, one can obtain a qualitative impression of the extent to which the pattern of conformance of the models to the principles mimics the empirically observed pattern.

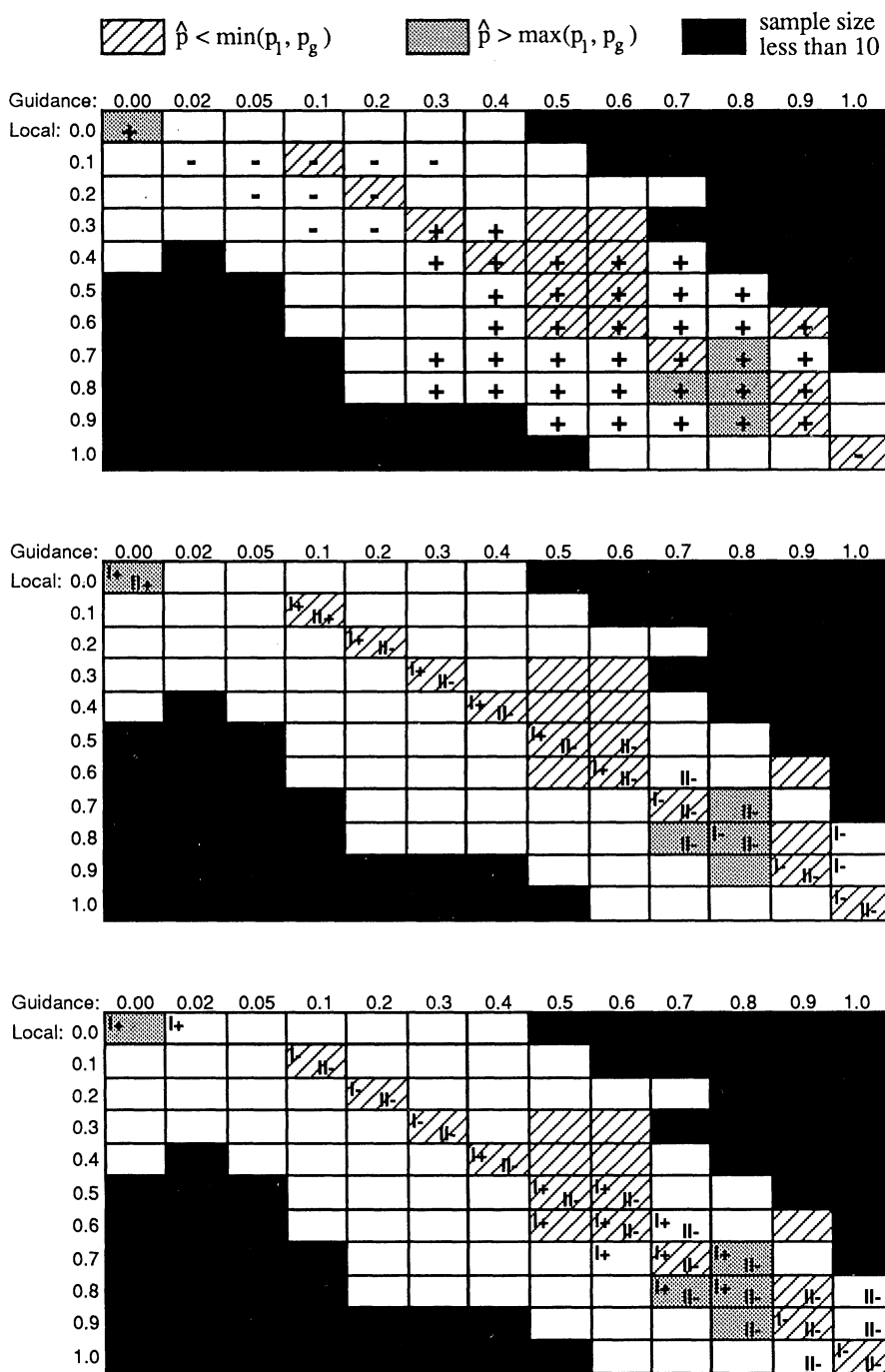


FIGURE 1. Patterns of conformance with unanimity and compromise. A “+” indicates a cell in which $p^* > \max(p_l, p_g)$, and a “-” indicates a cell in which $p^* < \min(p_l, p_g)$. The shading shows the empirical pattern of conformance according to the legend.

FIGURE 1a. Independence Model.

FIGURE 1b. Log-Odds I and Log-Odds II Models.

FIGURE 1c. GS-I and GS-II Models.

The patterns displayed in Figure 1a show clearly why the Independence model is unacceptable for this set of data. For a large portion of the table, the model disagrees with unanimity and compromise. The model pattern does not come close to reflecting the empirical pattern.

Turning to Figure 1b, we find that the Log-odds II model provides a reasonable reflection of the empirical pattern. Most of its disagreement lies in cells on or near the diagonal, generally having $p^* < \min(p_l, p_g)$. On the other hand, Log-odds I is less satisfactory. While its disagreement with unanimity and compromise lies mostly in cells where $p_l = p_g$, the model has $p^* > \max(p_l, p_g)$ for many of the cells in which the empirical result was the reverse.

In Figure 1c we can examine the patterns of disagreement of the GS-I and GS-II models. The GS-I model shows, not surprisingly, an effect similar to that of the Independence model, with $p^* > \max(p_l, p_g)$ for values of p_l and p_g between 0.4 and 0.8. For the most part, this is contrary to the empirical pattern. The GS-II model fits reasonably well, disagreeing with compromise [$p^* < \min(p_l, p_g)$] primarily for cells with equal p_l and p_g and for some of the cells with large p_l and p_g .

4. Summary and Conclusion

We have studied a variety of approaches for combining probabilities. The focus of our inquiry has been how these models relate to principles of unanimity and compromise. Those models that provide for the most general patterns of dependence among sources are the most complex in terms of their conformance to the principles. We also examined a large set of probability of precipitation forecasts in terms of the empirical conformance and disagreement with the principles. Finally, we fit the models to the data in an effort to understand which models could mimic most accurately the empirical pattern of combined probabilities. Because of the interaction of the two forecasts, it makes sense a priori to use a model that explicitly accounts for dependence between the forecasts. Our analysis bears out this intuition. Two models that permit dependence among the forecasts, Log-odds II and GS-II, were best able to reflect the patterns in the data, both in terms of MAD and MSE, as well as in terms of our qualitative evaluations in Figure 1.

As stated in the introduction, we do not claim that the unanimity and compromise principles are universally appropriate, nor do they have normative appeal. Instead, our goal is to provide insight into the probability-aggregation models that are available so that they may be applied more appropriately. On one hand, we can imagine situations in which two separate sources provide probability estimates which, when taken together, result in a very high (or very low) probability of A . Schervish (1986) provides such an example. On the other hand, it is easy to imagine, in the case of the weather forecasters, why this might not be the case. Both the local and guidance forecasts have developed over the years to be informative and calibrated forecasting systems. A reviewer of an early draft of this paper suggested that evolutionary pressures might lead p_l and p_g to be similar, and that, if both of these are good forecasts, then it would be reasonable to suspect that a convex combination of them would also be good.

This line of argument, we believe, holds a good deal of truth for competing real-world forecasts, and might be considered to be an informal argument in favor of the unanimity and compromise principles as a kind of "default" basis for combining forecasts in general. For example, when combining econometric forecasts, Clemen and Winkler (1986) concluded that combinations of the forecasts in which the weights were constrained to be positive and sum to one performed better in terms of MAD and MSE than did combinations without this constraint. For our weather forecasters studied here, the Bernoulli model performed reasonably well relative to the alternative models. In many situations,

with weather forecasting providing only one example, the unanimity and compromise principles might make sense in terms of providing a first-cut approximation.¹

¹ We gratefully acknowledge the helpful comments by Christian Genest, Dennis Lindley, and Mark Schervish on an earlier version of this paper, and we thank Gary Carter of the NWS Techniques Development Laboratory for providing the PoP forecasts analyzed in §3. This research was supported in part by the National Science Foundation under Grant IST 8600788.

References

- CLEMEN, R. T., "Calibration and the Aggregation of Probabilities," *Management Sci.*, 32 (1986), 312–314.
- , "Combining Overlapping Information," *Management Sci.*, 33 (1987), 373–380.
- AND R. L. WINKLER, "Combining Economic Forecasts," *J. Business and Economic Statist.*, 4 (1986), 39–46.
- AND ———, "Calibrating and Combining Precipitation Probability Forecasts," in *Probability and Bayesian Statistics*, R. Viertl (Ed.), Plenum, New York; 1987, 97–110.
- FRENCH, S., "Consensus of Opinion," *European J. Oper. Res.*, 7 (1981), 332–340.
- , "Group Consensus Probability Distributions," in Bernardo, J. M. et al. (Eds.), *Bayesian Statistics 2*, North-Holland, Amsterdam; 1985, 183–197.
- , "Calibration and the Expert Problem," *Management Sci.*, 32 (1986), 315–321.
- GENEST, C. AND M. J. SCHERVISH, "Modeling Expert Judgments for Bayesian Updating," *Ann. Statist.*, 13 (1985), 1198–1212.
- AND J. V. ZIDEK, "Combining Probability Distributions: A Critique and an Annotated Bibliography," *Statistical Sci.*, 1 (1986), 114–135.
- GRANGER, C. W. J. AND P. NEWBOLD, *Forecasting Economic Time Series*, Academic Press, New York; 1977.
- LINDLEY, D. V., "The Improvement of Probability Judgments," *J. Roy. Statist. Soc. Ser. A*, 145 (1982), 117–126.
- , "Reconciliation of Discrete Probability Distributions," in Bernardo, J. M. et al. (Eds.), *Bayesian Statistics 2*, North-Holland, Amsterdam; 1985, 375–387.
- , "Another Look at an Axiomatic Approach to Expert Resolution," *Management Sci.*, 32 (1986), 303–306.
- MCCONWAY, K. J., "Marginalization and Linear Opinion Pools," *J. Amer. Statist. Assoc.*, 76 (1981), 410–414.
- MORRIS, P. A., "Decision Analysis Expert Use," *Management Sci.*, 20 (1974), 1233–1241.
- , "An Axiomatic Approach to Expert Resolution," *Management Sci.*, 29 (1983), 24–32.
- , "Observations on Expert Aggregation," *Management Sci.*, 32 (1986), 321–328.
- MURPHY, A. H. AND R. L. WINKLER, "Probability Forecasting in Meteorology," *J. Amer. Statist. Assoc.*, 79 (1984), 489–500.
- SCHERVISH, M. J., "Comments on Some Axioms for Combining Judgments," *Management Sci.*, 32 (1986), 306–312.
- WINKLER, R. L., "Expert Resolution," *Management Sci.*, 32 (1986), 298–303.
- AND S. MAKRIDAKIS, "The Combination of Forecasts," *J. Roy. Statist. Soc. Ser. A*, 146 (1983), 150–157.