Combining Predictive Schemes in Short-Term Forecasting

K. FRAEDRICH* AND L. M. LESLIE

Bureau of Meteorology Research Centre, Melbourne, Victoria, 3001, Australia

(Manuscript received 22 September 1986, in final form 21 January 1987)

ABSTRACT

In this article, the theory is presented for a linear combination of two independent predictive techniques (either probabilistic or binary). It is shown that substantial gains might be expected for optimal weighting of the combination. The theory is general but also is applied to several special cases which may be useful for both short-term weather prediction and long-range forecasting. Using data from a recent operational evaluation of techniques for the short-term predicting of rainfall, a linear combination of two independent predictive techniques gives, in practice, improvement in skill compared with the techniques used individually. In the present case, a Markov chain and a numerical weather prediction (NWP) model were combined. The half-Brier score of the linear combination was 0.142 compared with individual scores of 0.164 for the Markov chain model and 0.258 for the NWP model. The combined Markov-NWP scheme may provide a possible simple alternative to the MOS approach for predictions up to 12 hours ahead.

1. Introduction

Optimal weighting of forecasts has received much attention in economics, management and statistics literature. In meteorology, also, it is known that consensus forecasts and the linear combination of predictions provide more accurate results than the individual forecasts which comprise the consenses. Although "this is the incontrovertible fact," and some combined schemes have been assessed (Perrone and Miller, 1985; Balzer, 1986) it still "does not appear to be widely recognized or accepted" (Thompson, 1977). In particular, short-term prediction (up to 12 hours) at single stations *and* long-range forecasting seem to be areas where a linear combination of the following two independent prediction schemes is promising: forecasts made by stochastic models, such as Markov chains for discrete or autoregressive models for continuous variables, can be combined with deterministic predictions, e.g., numerical weather predictions (NWP). This is especially true as the number of available predictive techniques has proliferated in recent years.

With the advantages of NWPs being obvious, Markov chains and other stochastic techniques have shown considerable short-term skill in forecasting various weather variables: the probabilities of 12-hourly rainfall predicted at midlatitude stations achieve half-Brier scores between 0.12 and 0.17, and furthermore, these predictions are only 20%–25% incorrect for wet and dry categories after deriving an optimal probability threshold value (Fraedrich and Müller, 1986). This level of skill for rainfall prediction is comparable to that of NWP models with (or without) applying MOS-techniques (Fraedrich and Leslie, 1986). Therefore, the purpose of this paper is to combine independent probabilistic and categorical NWP predictions to enhance forecasting skill.

In this article, Thompson's (1977) analysis, which combines two unbiased prediction schemes of continuous variables, is extended to probabilistic predictions, thereby constraining the variational principle to be applied. In some general format this has been described by Passi (1975) who, however, used unbiased estimates. In section 2, two independent probability forecasts are linearly combined to obtain the best estimate, and some basic examples are discussed briefly. In section 3 Markov and NWP predictions are combined to evaluate the forecast skill of the daytime probability of rainfall at a midlatitude station (Melbourne), which has direct impact on the quality of routine weather forecasts. The forecast skill is compared with that of the predictive schemes individually.

2. Combining two independent probability forecasts

Let ϕ denote the probability of an event, such as a wet or dry interval, forecast by some predictive scheme. This prediction may be either a probabilistic variable ($0 \leqslant \phi \leqslant 1$), or a binary variable ($\phi = 1$ or 0). The aim of the predictive scheme is to forecast the binary observation, δ, ($\delta = 1$ or 0) with a high degree of skill. One appropriate measure of success is the ensemble mean-square error, or half-Brier score. For example, the half-Brier score of a climatology prediction, ϕ_c,

* On leave from Institut für Meteorologie, Freie Universität Berlin, Berlin, Federal Republic of Germany.

where $\phi_c = \langle\delta\rangle$, equals the natural variance of the system

$$B_c = \langle(\delta - \phi_c)^2\rangle = \phi_c - \phi_c^2, \quad (1)$$

noting that $\langle\delta\rangle = \langle\delta^2\rangle = \phi_c$.

a. General case

Now consider two independent prediction schemes of probabilistic variables ϕ_1 and ϕ_2. For the linear combination ϕ_*

$$\phi_* = a\phi_1 + b\phi_2 \quad (2)$$

to be a probabilistic variable requires that the weights a, b satisfy

$$a + b = 1. \quad (3)$$

The related ensemble mean-square error (or half-Brier score) is

$$\begin{aligned} B_* &= \langle(\delta - \phi_*)^2\rangle \\ &= \langle\delta^2\rangle - 2\langle\delta\phi_2\rangle + \langle\phi_2^2\rangle + 2a(\langle\delta\phi_2\rangle - \langle\phi_2^2\rangle \\ &\quad - \langle\delta\phi_1\rangle + \langle\phi_1\phi_2\rangle) + a^2(\langle\phi_1^2\rangle + \langle\phi_2^2\rangle - 2\langle\phi_1\phi_2\rangle) \quad (4) \end{aligned}$$

where $a = 1 - b$ has been introduced; $\langle\phi_1\phi_2\rangle$ is the covariance between the predictions ϕ_1 and ϕ_2, $\langle\delta\phi_1\rangle$ and $\langle\delta\phi_2\rangle$ are the covariances between either ϕ_1 and the observation δ or ϕ_2 and δ; $\langle\phi_1^2\rangle$ and $\langle\phi_2^2\rangle$ are measures of the prediction variance; and $\langle\delta^2\rangle$ is the observed variability. Note that all these second moments are not taken as deviations from their respective means.

As the half-Brier score (4) can be partitioned into two additive terms denoting the (square of the) mean error of the forecasts (bias or reliability) and its variance, it appears to be the most appropriate measure of skill. Therefore, we shall choose a such that the half-Brier score (and thus bias and error variance) is minimized by the combination forecast. Regarding the half-Brier score as a function of a, we deduce from (4) that

$$\begin{aligned} \frac{dB_*}{da} &= \langle\delta\phi_2\rangle - \langle\delta\phi_1\rangle - \langle\phi_2^2\rangle + \langle\phi_1\phi_2\rangle \\ &\quad + a(\langle\phi_1^2\rangle + \langle\phi_2^2\rangle - 2\langle\phi_1\phi_2\rangle) = 0. \quad (5) \end{aligned}$$

Solving for the weight $a = 1 - b$ in terms of the second moments,

$$a = \frac{\langle\delta\phi_1\rangle - \langle\delta\phi_2\rangle + \langle\phi_2^2\rangle - \langle\phi_1\phi_2\rangle}{\langle\phi_1^2\rangle + \langle\phi_2^2\rangle - 2\langle\phi_1\phi_2\rangle}. \quad (6)$$

Note that this weight a is not constrained to lie between 0 and 1, but in general could also be either negative or greater than 1. Furthermore, it appears that a is simply the slope of the regression of $\delta - \phi_2$ on $(\phi_1 - \phi_2)$, which arises due to the constraint that the weights (3) sum to unity. With the optimum weight we can determine the mean-square error B_* of the best estimate, ϕ_*, by introducing (6) into (4). After some algebra the result is

$$\begin{aligned} B_* &= \langle(\delta - \phi_2)^2\rangle \\ &\quad - \frac{(\langle\delta\phi_1\rangle - \langle\delta\phi_2\rangle + \langle\phi_2^2\rangle - \langle\phi_1\phi_2\rangle)^2}{\langle\phi_1^2\rangle + \langle\phi_2^2\rangle - 2\langle\phi_1\phi_2\rangle}, \quad (7) \end{aligned}$$

where $\langle(\delta - \phi_2)^2\rangle = B_2$ is the half-Brier score for the model ϕ_2.

Note that the covariance between model and observation can be deduced from the mean-square error (half-Brier score) B_i of the model ϕ_i ($i = 1$ or 2)

$$\begin{aligned} B_i &= \langle(\delta - \phi_i)^2\rangle = \langle\delta^2\rangle - 2\langle\delta\phi_i\rangle + \langle\phi_i^2\rangle \\ \langle\delta\phi_i\rangle &= \frac{1}{2}(\phi_c + \langle\phi_i^2\rangle) - B_i/2. \quad (8) \end{aligned}$$

For the observations, $\delta = 1$ or 0, the mean and second moment are equal; that is,

$$\langle\delta\rangle = \langle\delta^2\rangle = \phi_c \quad (9)$$

where ϕ_c is climatology. For practical purposes ϕ_c is estimated by the relative frequency of the event $\delta = 1$ (for wet episodes). The same holds for a model prediction $\phi_i = 1$ or 0, which is associated with the model climate

$$\langle\phi_i\rangle = \langle\phi_i^2\rangle = \phi_{c_i}.$$

For a probabilistic prediction $0 \leqslant \phi_j \leqslant 1$, the model climate yields $\langle\phi_j\rangle = \phi_{cj}$ but $\langle\phi_j^2\rangle \neq \phi_{cj}$.

In the following, some simplifications are introduced to show the skill improving by the combination of two independent probability predictions.

b. Special cases

1) BINARY (YES/NO) PREDICTION AND CLIMATOLOGY COMBINED

Let the binary prediction scheme be the variable $\phi_1 = 1$ or 0; for example, the NWP model or persistence forecasting can be a wet or dry time interval. Then the following statistical relations are easily deduced:

$$\begin{aligned} \langle\phi_1\rangle &= \langle\phi_1^2\rangle = \phi_{c_1} \\ \langle\delta\phi_1\rangle &= \frac{1}{2}(\phi_c + \phi_{c_1} - B_1). \quad (10) \end{aligned}$$

The model climate is the relative frequency of events with $\phi_1 = 1$. The mean-square prediction error of ϕ_1 is determined by

$$B_1 = \langle(\delta - \phi_1)^2\rangle,$$

which is the relative frequency of incorrect ϕ_1-predictions of the events or realizations $\delta = 1$ or 0.

The second prediction scheme is the observed climatology, $\phi_2 = \phi_c$. The associated statistical relationships yield

$$\left.\begin{aligned}\langle\phi_2\rangle=\phi_c,\quad\langle\phi_2{}^2\rangle=\phi_c{}^2\\ \langle\delta\phi_2\rangle=\langle\delta\phi_c\rangle=\phi_c\langle\delta\rangle=\phi_c{}^2\\ B_2=B_c=\langle(\delta-\phi_c)^2\rangle=\phi_c-\phi_c{}^2\end{aligned}\right\}. \quad (11)$$

The optimum linear combination $\phi_* = a\phi_1 + (1 - a)\phi_2$ of both forecasts ϕ_1 and ϕ_2 is found by minimizing the mean-square error B_* associated with the weight a, by substituting (10) and (11) into (6) and (7):

$$a=\frac{\langle\delta\phi_1\rangle-\phi_{c_1}\phi_c}{\phi_{c_1}+\phi_c{}^2-2\phi_{c_1}\phi_c},$$

$$B_*=B_c-\frac{(\langle\delta\phi_1\rangle-\phi_{c_1}\phi_c)^2}{\phi_{c_1}+\phi_c{}^2-2\phi_{c_1}\phi_c}. \quad (12)$$

If the prediction scheme ϕ_1 is assumed to be unbiased, i.e., with model variance $B_{c_1} = \langle(\phi_1 - \phi_{c_1})^2\rangle$ equal to the natural variance $\langle(\delta - \phi_c)^2\rangle = B_c$ (e.g., persistence), then

$$\phi_c=\phi_{c_1}\quad\text{and}\quad B_c=B_{c_1}=\phi_c-\phi_c{}^2.$$

Now the weight a and the combined Brier scores B_* yield, after some algebra,

$$a=1-\frac{B_1}{2B_c},\quad B_*=B_1-\frac{B_1{}^2}{4B_c}. \quad (14)$$

Furthermore, if the mean-square error or the relative frequency of incorrect ϕ_1-predictions B_1, coincides with the natural variance, B_c, then combining ϕ_1 with climatology or ϕ_c-forecasts, yields $a = 0.5$ and $B_* = 3B_c/4$ which is a 25% improvement in skill. This optimal combination is essentially a recalibration of categorical forecasts so that they are reliable probability forecasts. Such a conversion has been recently advocated for scoring categorical forecasts (Murphy, 1986).

2) Two binary prediction schemes combined

Let the prediction schemes ϕ_1 and ϕ_2 attain the values 1 or 0; then for (6) and (7) we have

$$\begin{aligned}&\langle\delta\rangle=\langle\delta^2\rangle=\phi_c;\quad\langle\phi_1\rangle=\langle\phi_1{}^2\rangle=\phi_{c_1},\\ &\langle\phi_2\rangle=\langle\phi_2{}^2\rangle=\phi_{c_2},\quad\langle\delta\phi_1\rangle=\frac{1}{2}(\phi_c+\phi_{c_1}-B_1),\\ &\langle\delta\phi_2\rangle=\frac{1}{2}(\phi_c+\phi_{c_2}-B_2).\end{aligned} \quad (15)$$

Note that the Brier scores B_1 and B_2 are now equivalent to the relative number of incorrect predictions of the two predictive schemes ϕ_1 and ϕ_2. Assuming that both prediction models are unbiased with the same skill $B_1 = B_2 = B$, then

$$\begin{aligned}&\phi_c=\phi_{c_1}=\phi_{c_2}\\ &\langle\delta\phi_1\rangle=\langle\delta\phi_2\rangle=\phi_c-B/2.\end{aligned} \quad (16)$$

Now one obtains for the combination of the two independent forecasts, $\phi_* = a\phi_1 + (1 - a)\phi_2$ the following weights and the Brier score after substituting (15) and (16) into (6) and (7):

$$a=0.5,\quad B_*=B-\frac{1}{2}(\phi_c-\langle\phi_1\phi_2\rangle), \quad (17)$$

where $\langle\phi_1\phi_2\rangle$ is the second mixed moment between the two prediction schemes. Perfect correlation between both schemes $\phi_1 = \phi_2$ or $\langle\phi_1\phi_2\rangle = \phi_{c_1}$ does not lead to skill improvement ($B_* = B$), which is not surprising. If there is no correlation between the schemes, then

$$\langle\phi_1'\phi_2'\rangle=\langle\phi_1\phi_2\rangle-\langle\phi_1\rangle\langle\phi_2\rangle=0,$$

or

$$\langle\phi_1\phi_2\rangle=\phi_{c_1}\cdot\phi_{c_2}=\phi_c{}^2. \quad (18)$$

Clearly there is some skill improvement as we then have after substituting (18) into (17), $B_* = B - \frac{1}{2}B_c$. If the models' skills equal the natural variance (or climatology predictions), B_c, then the combination forecast skill improves to $B_* = 0.5\ B_c$, which is a 50% improvement in skill.

3. The prediction schemes

There are two basic prediction schemes discussed in this article: a Markov chain model and an operational limited area NWP model; a MOS scheme which utilizes the NWP model forecasts will be used for further comparison. These prediction schemes will be subsequently briefly discussed.

The Markov chain model is an extension of that devised by Fraedrich and Müller (1983). A second-order Markov chain is approximated by a first-order Markov chain in which the transition probabilities are taken to be linear functions of the previous state (Miller and Leslie, 1985). Monthly changes are included by fitting different models each month and allowance is made for incorporating the effects of other surface data. Four states based upon cloud cover and rainfall are defined, and the probability of precipitation from current state j, at current time t, for h hours ahead is

$$P(m,j,t,h)=a(m,j,t,h)+\sum_{\kappa=1}^{3}b_\kappa(j,h)X_\kappa \quad (19)$$

where m is the month, and the covariates X_κ ($\kappa = 1, 2, 3$) are the station pressure, the dewpoint depression, and the east–west wind component. Notice that while different intercepts, a, are used for each month, common slopes, b, are fitted for all months. The model has been fitted by ordinary least-squares to the data. For the Melbourne weather station, other covariates such as the pressure change from the previous state were not found to be significant.

The NWP model is the Australian Bureau of Meteorology's new operational limited-area model (Leslie et al., 1985) which has a horizontal resolution of 150 km and has 12 levels in the vertical. The model produces 36 hour forecasts twice-daily at 0000 UTC and 1200 UTC. Prediction of precipitation consists of two adjustment procedures applied successively. The first step is a simulation of cumulus convection based on a modification of the Kuo scheme. The second step is a large-scale saturation adjustment applied to all model gridpoints when the mixing ratio exceeds 95% of the saturation mixing ratio.

The model output statistics (MOS) scheme used by the Australian Bureau of Meteorology was developed for the seven major Australian cities, including Melbourne (see Tapp et al., 1986). Probability of precipitation is one of a number of predictands in the regression scheme which uses the operational NWP model output as predictors.

4. Results

The practical advantages of combining independent probabilistic and/or binary predictions are demonstrated in this section, using some of the data from a recently completed operational trial in which the skill of various rainfall prediction schemes was compared. The trial was carried out on a daily basis for the winter season, June–August 1986, at the Melbourne weather station. The forecast period each day was the 12 hour interval 0600 to 1800 local time. Full details and results of the trial will be published shortly by Fraedrich and Leslie (1986). Among the individual schemes evaluated in the following are the Markov chain with no lead time, and NWP and MOS models with both 21 h lead time. The performance of these schemes is given in the upper half of column 3 of Table 1, for which the half-Brier scores for July 1986 are shown and compared with climatology. Not unexpectedly, the Markov chain yields the best skill, with a score of 0.164, while the NWP model has a score (0.258) comparable with climatology (0.259); MOS obtains a half-Brier score of 0.184.

For the purpose of this article it was decided to combine the Markov chain and NWP models, as they are independent schemes, and to compare them with the uncombined forecasts. If we denote the Markov probabilistic predictions by ϕ_1 ($0 \leqslant \phi_1 \leqslant 1$) and the NWP model binary predictions by ϕ_2 (=1 or 0), then an optimal combination

$$\phi_* = a\phi_1 + (1-a)\phi_2, \tag{20}$$

was obtained from four years (1982–85) of July Markov and NWP forecasts with half-Brier scores of $B_1 = 0.154$ and $B_2 = 0.234$ (see column 2 of Table 1). The optimal value of the weight a for the combined scheme (20) was calculated from Eq. (6) and found to be 0.7. When scheme (20) was used in this optimal hindcast form, a half-Brier score of $B_* = 0.136$ was obtained. Turning from the hindcast mode to a true forecast mode, the value of $a = 0.7$ was retained and the combined scheme (20) was applied to the July 1986 data. This leads to a half-Brier score of $B_* = 0.142$. Clearly, then, the advantages of the combined schemes implied by the theory in section 2a have been verified in practice. As can be seen in Table 1, column 3, the combined half-Brier score for scheme (20) is about 15% lower than that of the Markov scheme alone, and about 45% lower than that of the NWP model alone.

In a broad sense, the MOS scheme can be seen as combining NWP and climatology and as a rough test of this idea, climatology and the NWP model were combined. The theory of section 2b implies an optimal weight of 0.5, and yielded a half-Brier score of 0.196 for July 1986 (see Table 1, column 3). This was only 7% worse than the score for MOS (0.184), thereby adding support to our suggestion. Furthermore, interpreting the different Brier scores as error variances of independent forecast samples taken from the same normal distribution, one cannot reject the null hypothesis of similar or equal error variances of both MOS and climate–NWP (on a 5% significance level). Note that this also holds for the MOS and Markov–NWP comparison.

As the MOS scheme available to the Australian Bureau of Meteorology has been developed only for 23 UTC (0900 local time) data, the predictions made by the MOS scheme for the period 0600 to 1800 local time involve a lead time of 21 hours, which presumably imposes a limitation on its performance. Moreover, for predictions beyond 12 hours, the Markov chain model discussed here would be of doubtful validity and the MOS scheme would be preferred.

Given that the Markov chain is simple and immediate to use, and requires minimal computing resources, it would appear to provide an effective alternative to the MOS method which has gained such widespread acceptance. Furthermore, only weighting coefficients have to be newly evaluated for the combined Markov–NWP scheme, if a NWP-model is substantially changed; the local Markov chains remain unchanged.

TABLE 1. Half-Brier scores for July 1982–85 and July 1986. Asterisks denote hindcasts.

Model	1982–85	1986
Individual predictions		
Markov	0.154	0.164
NWP	0.234	0.258
MOS	—	0.184
Climatology	0.243*	0.259
Combined predictions		
Markov and NWP	0.136*	0.142
Climatology and NWP	0.177*	0.196

5. Conclusions

In both theory and practice, two independent prediction schemes for short-term (up to 12 hours) prediction can be linearly combined in an optimal manner to obtain a skill level that exceeds the skill of each prediction scheme independently.

The optimal weighting coefficient was calculated for a linear combination of a Markov chain model and the Australian Bureau of Meterology's limited area NWP model, using four years (1982–85) of July daily 12-hour precipitation predictions. When the combined Markov–NWP model was run in a hindcast mode with the optimal weighting, an improved half-Brier skill score of 0.136 was obtained, compared with individual scores of 0.154 and 0.234 for the uncombined Markov and NWP models, respectively. In a forecast made for July 1986, using the weighting calculated from the 1982 to 1985 data, the margin of improvement was retained, with a half-Brier score of the combined scheme of 0.142 compared with individual Markov and NWP scores of 0.164 and 0.258. The performance of the combined Markov–NWP predictions suggests that they might be an alternative to the MOS approach for short-term forecasting, at least up to 12 hours ahead.

Acknowledgments. Thanks are due to D. Pike for computational assistance, R. Tapp for discussions, referees for their helpful comments and Ms. F. Nikou for typing.

REFERENCES

Balzer, K., 1986: Eine selbstlernende, optimale Mensch–Maschine–Kombination von operativen Immissionsprognosen. *Z. Meteor.*, **36,** 127–133.

Fraedrich, K., and K. Müller, 1983: On single station forecasting: Sunshine and rainfall Markov chains. *Beitr. Phys. Atmos.*, **56,** 108–134.

——, and K. Müller, 1986: On single station forecasting: Probability of precipitation in Berlin. *Beitr. Phys. Atmos.*, **59,** 427–434.

——, and L. M. Leslie, 1987: Evaluation of techniques for the operational, single station, short-term forcasting of rainfall at a midlatitude station (Melbourne). *Mon. Wea. Rev.*, **115,** 1645–1654.

Leslie, L. M., G. A. Mills, L. W. Logan, D. J. Gauntlett, G. A. Kelly, M. J. Manton, J. L. McGregor and J. M. Sardie, 1985: A high-resolution primitive equations NWP model for operations and research. *Aust. Meteor. Mag.*, **33,** 11–36.

Miller, A. J., and L. M. Leslie, 1985: Short-term single-station probability of precipitation forecasting using linear and logistic models. *Beitr. Phys. Atmos.*, **58,** 517–527.

Murphy, A. H., 1986: Comparative evaluation of categorical and probabilistic forecasts: Two alternatives to the traditional approach. *Mon. Wea. Rev.*, **114,** 245–249.

Passi, R. M., 1975: Statistical estimation of meteorological parameters with correlated observations. *Mon. Wea. Rev.*, **103,** 521–527.

Perrone, T. J., and R. G. Miller, 1985: Generalized exponential Markov and model output statistics: A comparative verification. *Mon. Wea. Rev.*, **113,** 1524–1541.

Tapp, R. G., F. Woodcock and G. A. Mills, 1986: The application of model output statistics to precipitation prediction in Australia. *Mon. Wea. Rev.*, **114,** 50–61.

Thompson, P. D., 1977: How to improve accuracy by combining independent forecasts. *Mon. Wea. Rev.*, **105,** 228–229.