Journal of Forecasting, Vol. 15, 49–61 (1996)

Forecast Combining with Neural Networks

R. GLEN DONALDSON
University of British Columbia, Canada

MARK KAMSTRA
Simon Fraser University, Canada

ABSTRACT

This paper investigates the use of Artificial Neural Networks (ANNs) to combine time series forecasts of stock market volatility from the USA, Canada, Japan and the UK. We demonstrate that combining with nonlinear ANNs generally produces forecasts which, on the basis of out-of-sample forecast encompassing tests and mean squared error comparisons, routinely dominate forecasts from traditional linear combining procedures. Superiority of the ANN arises because of its flexibility to account for potentially complex nonlinear relationships not easily captured by traditional linear models.

KEY WORDS forecast combing; artificial neural network; encompassing test

When combining n individual forecasts $f_1, \ldots, f_n$, the single combined forecast F is traditionally obtained by selecting β weights in the linear model $F = \beta_0 + \sum_{i=1}^{n} \beta_i f_i$, a popular example being the simple average across forecasts (i.e. $\beta_0 = 0$, $\beta_i = 1/n \forall i$).[1] However, a linear combination may not be optimal if the individual forecasts come from nonlinear models or if the true underlying conditional expectation is a nonlinear function of the information sets on which the individual forecasts are based. Consider, for example, the case of a dependent variable $y = \exp(\sum_{i=1}^{n} x_i) + \varepsilon$, where ε is an innovation and $x_1, \ldots, x_n$ are n explanatory variables known to us. If each of the $i = 1, \ldots, n$ individual forecasts are produced by $f_i = \alpha_i \exp(x_i)$, then any linear combination of the n individual forecasts will be inferior to the nonlinearly combined forecast $F = \prod_{i=1}^{n} f_i / \alpha_i$.

In this paper we investigate the incremental value of going from traditional linear forecast combining procedures to a particular class of nonlinear combining procedures based on Artificial Neural Networks (ANNs). Since ANNs have the ability to approximate arbitrarily well a large class of functions, they provide considerable flexibility to uncover hidden nonlinear relationships between a group of individual forecasts and realizations of the

[1] The forecast combining literature is much too vast to adequately cite here. For excellent reviews of the forecasting literature and discussions of traditional weight-selection techniques, however, see Clemen (1989), Granger (1989) and Min and Zellner (1993).

CCC 0277–6693/96/010049–13
© 1996 by John Wiley & Sons, Ltd.
Received October 1994
Accepted June 1995

variable being forecasted. Indeed, we develop in this paper 'optimally' combined ANN forecasts which generally outperform forecasts from a variety of traditional linear combining methods.

In the remainder of this paper we first present our ANN modelling procedure and describe the international stock market data we use to compare our nonlinear ANN to traditional linear combining methods. We then present model summary statistics and evaluate the combined forecasts on the basis of mean squared error, mean absolute error, and forecast encompassing tests to show that nonlinearly combined ANN forecasts perform at least as well as, and often better than, forecasts from a variety of traditional linear models. We conclude with a discussion of the practical significance of our results and a brief demonstration to show that the superiority of our ANN arises because of its flexibility to account for potentially complex nonlinear relationships not easily captured by traditional linear combining methods.

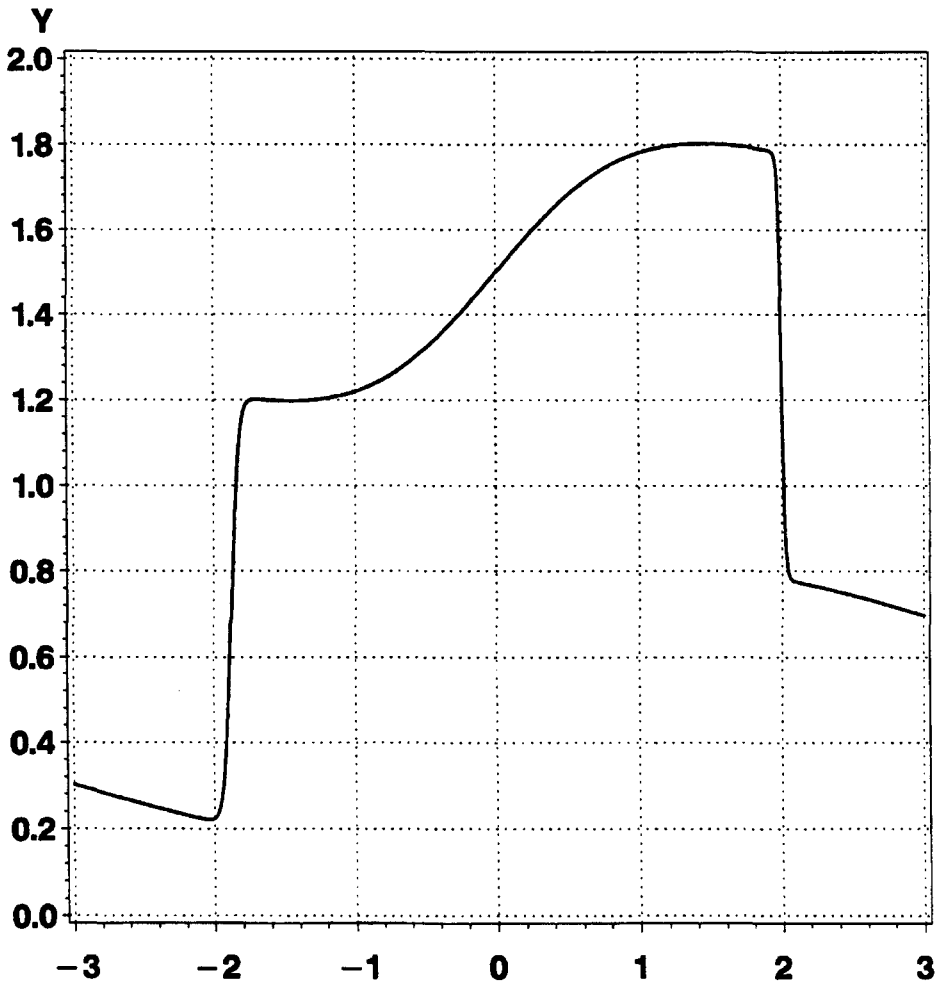

Figure 1. Example of a neural network. One linear and three nonlinear components.

ANN FORECAST COMBINING

An ANN is essentially a collection of nonlinear transfer functions which relate some output variable(s) of interest to some input variables, which may themselves be functions of even deeper explanatory variables.[2] One of the most commonly employed transfer functions in the ANN literature is the logistic $y = a + (1 + \exp[-(c + bx)])^{-1}$ in which the input x may itself be a subordinate function $g(z)$, where z is an element—or an even further function of other elements—of the forecasting information set. Fully linear functions, such as $y = a + bx$, can be used to augment the nonlinear functions if desired.

A simple network composed of one linear and three logistic functions is depicted in Figure 1. The network's input x is measured on the horizontal axis and output Y on the vertical axis, where $Y = \sum_{i=1}^{4} y_i$ with $y_1 = 1 - 0.1x$, $y_2 = -0.5 + (1 + \exp[-(75 + 40x)])^{-1}$, $y_3 = -0.5 + (1 + \exp[-(0 + 2x)])^{-1}$, and $y_4 = -1 + (1 + \exp[-(170 - 85x)])^{-1}$ being the four 'information nodes' of the network that filter information on the input x. Notice from the definitions of y_i that, when $x < -2$, $y_2 \approx -0.5$, $y_3 \approx -0.5$, and $y_4 \approx 0$, so Y's behaviour is determined largely by the slope of linear node y_1: i.e. $Y = \sum_{i=1}^{4} y_i \approx -0.1x$. However, as x rises past about -2, y_2 rapidly increases in value to essentially achieve its maximum of $y_2 = 0.5$ by the time x reaches roughly -1.6. The response of node y_2 therefore brings the network value of $Y = \sum y_i$ from 0.2 up to 1.2 as x rises from -2 to about -1.6, as seen in Figure 1. Then, as x continues its increase towards zero, the y_3 node begins to activate—although its response is less immediate given y_3's smaller gain (i.e. $2x$ instead of $40x$)—so that, by the time $x = 1.5$, $Y \approx 1.8$. Finally, as x rises past 2, the y_4 node deactivates to fall from 0 to -1 so that Y falls from 1.8 down to 0.8. The effect on Y of any further increases x are then determined largely by the linear node y_1, with slope -0.1, as seen in Figure 1.

The final shape produced by Figure 1's ANN is a complicated hump-pattern in which different segments have different slopes. By extension, one can see that, with many nodes activating and deactivating with different slopes and intercepts over various ranges of x, one could produce as an output response Y almost any desired function of the input x (or, with higher dimensionality, any desired function of a group of inputs $x_1, \ldots, x_n$). Indeed, as demonstrated by authors such as Hornik, *et al.* (1989, 1990), ANNs have the ability to approximate arbitrarily well a large class of functions.[3] ANNs are therefore ideally suited to the problem of forecast combining when the optimal combination of individual forecasts is potentially nonlinear.

In this paper we use ANNs to obtain a single consensus forecast F_t as the potentially nonlinear combination of two individual forecasts $f_{1,t}$ and $f_{2,t}$. To do this, let $\bar{A}$ and S_A denote the in-sample mean and in-sample standard deviation, respectively, of the variable being forecasted out of sample. The ANNs we investigate are then of the following form:[4]

$$F_t = \beta_0 + \sum_{j=1}^{k} \beta_j f_{j,t} + \sum_{i=1}^{p} \delta_i \Psi(z_i \gamma_i) \tag{1}$$

---

[2] Even a cursory discussion of the many applications to which ANNs have been put would consume several pages of text and is therefore well beyond the scope of this paper. For an excellent review of the econometric issues involved and some discussion of ANNs' many applications, however, see Kuan and White (1994) and Hertz *et al.* (1991).

[3] For further technical details see Hertz *et al.* (1991), Hornik (1991), Stinchcombe and White (1994) and White (1989, 1990).

[4] The model we employ is formally referred to as a single hidden-layer ANN (where the number of nodes will be selected optimally by reference to the data). In principle, an ANN can consist of many functional layers, with intermediate outputs used as inputs to ever higher layer of nodes until the ultimate output is finally produced. In practice, however, ANN researchers (e.g. White, 1990) have proven that, provided a sufficient number of nodes are placed on the first layer of the ANN, higher layers are not needed to establish a satisfactory connection between the initial raw inputs and the final output. For simplicity, we therefore entertain only one hidden layer in equations (1)–(3). The network whose output is graphed in Figure 1 is also a single hidden-layer ANN (with four nodes: y_1, y_2, y_3, y_4).

$$\Psi(z_t,\gamma_i) = (1 + \exp[-(\gamma_{0,i} + \gamma_{1,i}z_{1,t} + \gamma_{2,i}z_{2,t})])^{-1} \quad (2)$$

$$z_{j,t} = (f_{j,t} - \bar{A})/S_A \qquad j = 1, 2 \quad (3)$$

$$k \in \{0, 2\} \qquad p \in \{0, 1, 2, 3\}$$

Equation (3) gives the normalization of forecasts from each individual model required to prepare these forecasts for inputs into the logistic information node given in equation (2). Equation (1) states the manner in which the outputs from these nodes are to be weighted in producing the final combined forecast. Since $p \in \{0, 1, 2, 3\}$, up to three logistic nodes are permitted in the final network. The special case of a traditional linear model (i.e. $k = 2$, $p = 0$) is also permitted.

To implement estimation of equation (1) we follow the computationally simple approach of first choosing the γ_i with a uniform random number generator, so the γ_i lie between -1 and $+1$, and then estimating the δ and β parameters. Work by Stinchcombe and White (1994) yields a universal approximation result for ANNs with such an arbitrary choice of γ and, in practice, little difference is made in model performance with deviations from this convention. We therefore randomly choose 10 different sets of γ's which, together with $k \in \{0, 2\}$ linear and $p \in \{0, 1, 2, 3\}$ nonlinear nodes, produce 60 different ANN model specifications plus the purely linear model $k = 2$, $p = 0$. Since we do not know the true structure of the conditional expectation relationship, we employ a standard tenfold cross-validation selection procedure to choose the 'best' model specification from among the 61 possible specifications.[5] Information on particular specifications chosen for our study is provided below.

To gauge the performance of our nonlinear ANN we also estimate three popular fully linear combining models: the simple average of individual forecasts (AVE), the model in (1) with $p = 0$ and the βs estimated by Ordinary Least Squares (OLS), and the model in (1) with $p = 0$ and the βs estimated by the robust method of Mean Absolute Deviations (MAD).[6] Use of such traditional combining procedures is discussed in Clemen (1989) and Granger (1989) and the many references cited therein.

FORECAST DATA

Individual forecasts for use in our combining exercise are forecasts of the volatility in daily stock returns on the US Standard and Poor's 500 Stock Index (SP500), the Japanese Nikkei Stock Index (NIKKEI), Canada's Toronto Stock Exchange Composite Index (TSEC) and London's Financial Times Stock Exchange Index (FTSE)—for the period January 1969 to September 1987—as produced by two popular models of stock returns volatility: the MA variance model (MAV) and the GARCH(1,1) model (GAR). We employ these data for two

[5] Cross-validation calls for estimating the model on a subset of the in-sample data and then using the estimated model to forecast the remaining portion of the in-sample data. We then collect the 'out-of-sample' forecasts and repeat the process, leaving out a different subset of the in-sample data each time, until we have produced 'out-of-sample' forecasts for the entire in-sample data set. (A cross-validation estimation which has $1/N$ of the data omitted at a time is called an N-fold cross-validation.) The model specification which produces in-sample cross-validated forecasts with the lowest Mean Squared Error is then selected as 'best'. For a more complete discussion of this model selection procedure and its optimality properties, see Hjorth and Holmqvist (1981), Kavalieris (1989) and Stoica, *et al.* (1986).

[6] We choose the MAD estimator, as employed in Hallman and Kamstra (1989), over other robust possibilities, such as the Bayesian time-varying weight selection methods of Min and Zellner (1993), for two reasons. First, authors such as Min and Zellner (1993) have demonstrated that there is little benefit to using Bayesian techniques over classical methods at the particular forecasting horizon we investigate in our paper. Second, MAD is more easily implemented and thus more likely to be employed by practitioners.

reasons. First, while frequently combined quarterly or annual macroeconomic and accounting data offer a somewhat limited number of observations, the daily stock market data contain many thousands of observations, thereby providing the power necessary to discriminate between the various combining methods. Second, the MAV and GARCH stock volatility forecasting models, as specified below, are widely employed in the financial econometrics literature and have well understood properties that can aid us in interpreting test results and suggest useful specification checks of the individual and combined models.[7]

To obtain the MAV and GARCH individual forecasts, we follow the stock volatility literature and define R_t as the daily stock return with $R_t = \rho_0 + \rho_1 R_{t-1} + \varepsilon_t$, where ε_t is an error with zero mean and conditional variance $E(\varepsilon_t^2 \mid I_t) = \sigma_t^2$ (appropriately normalized, ε_t can be considered stationary). Our objective is to forecast σ_t^2: stock returns volatility. Let $\hat{\rho}_0$ and $\hat{\rho}_1$ be estimates of the parameters ρ_0 and ρ_1, and let $\hat{\varepsilon}_t = R_t - \hat{\rho}_0 - \hat{\rho}_1 R_{t-1}$. The MA variance model has the conditional variance forecast: $MAV_t = (1/n)\sum_{i=1}^{n} \hat{\varepsilon}_{t-i}^2$, with n chosen to minimize the Schwarz Criterion and the parameters ρ_0 and ρ_1 estimated with OLS. The GARCH(1,1) model has the conditional variance forecast: $GAR_t = \alpha_0 + \alpha_1 GAR_{t-1} + \alpha_2 \hat{\varepsilon}_{t-1}^2$, with parameters ρ_0, ρ_1, α_0, α_1 and α_2 estimated jointly with maximum likelihood methods. Explanations for why the MAV and GARCH models are widely employed for the purpose of forecasting stock market volatility can be found in Bollerslev *et al.* (1991) and Pagan and Schwert (1990) and the numerous references cited therein.

The GARCH and MAV forecasts we use in our study are one-step-ahead out-of-sample forecasts beginning the first day of trading in January 1980. Daily data from the first day of 1969 to the last day of 1979 are used to estimate the MAV and GARCH model parameters and then produce the one-step-ahead out-of-sample MAV and GARCH forecasts for the first trading day of 1980. We next update our data set—while keeping sample size constant—by adding the first trading day of 1980 and dropping the first observation from 1969. We then re-estimate the MAV and GARCH models and produce one-step-ahead out-of-sample forecasts for the second day in 1980. This recursive updating and one-step-ahead out-of-sample forecasting procedure is repeated until one-step-ahead out-of-sample MAV and GARCH forecasts of daily returns volatility are produced for each trading day from 1 January 1980 to 30 September 1987. These constitute the individual out-of-sample forecasts used in the combining exercise. For our purposes, the most important feature of these forecasts is that the MAV and GARCH models used to produce them employ partially non-overlapping information sets; MAV uses as least twelve times as many lags of the ε^2 error as GARCH, but does not use any lagged conditional variance estimates. Thus, there may be an advantage to using a combined forecast as opposed to either of the individual forecasts.[8]

The next step in our procedure is to divide the MAV and GARCH out-of-sample forecasts into two subsamples: 1 January 1980 to 21 June 1983, and 22 June 1983 to 30 September 1987. Data on MAV and GARCH forecasts from 1 January 1980 to 21 June 1983 are then used, with the cross-validation procedure described in the previous section, to select the following optimal

---

[7] Note that our data extend only to September 1987 in order to avoid including the infamous stock market crash of October 1987. Out-of-sample model comparisons, such as ours, make sense only in the context of stable conditional data generating processes. The Crash of '87 resulted in such large and abrupt changes in stock volatility that it is not perfectly clear that the assumption of stability is valid over this period.

[8] Note that, in practice, we would combine forecasts only if we did not have access to forecasters' information sets. However, we obviously do have the information sets used to produce both of our individual forecasts. The application to conditional variance forecasts in our paper should therefore be viewed primarily as an exercise to compare the combining methods. Use of ANNs to forecast stock volatility using all available information is the subject of related work in Donaldson and Kamstra (1994).

specifications for our ANN(k, p) models outlined in equations (1)–(3): SP500 = (0, 1); NIKKEI = (0, 2); TSEC = (1, 1); FTSE = (0, 1). Notice that three of the four data series do not require linear terms (i.e. $k = 0$) and that, in all cases, our cross-validation procedure selects only one or two nonlinear nodes. The implications of these particular model specifications for forecast combining are discussed below.

Given our model specifications, we next use the 1 January 1980 to 21 June 1983 data to estimate β weights for AVE, OLS and MAD—and β, γ weights for ANN—to obtain one-step-ahead out-of-sample combined forecasts for 22 June 1983. Having done this, we update our information set by one day to obtain new weights and new one-step-ahead out-of-sample combined forecasts for 23 June 1983. This updating/combining/forecasting procedure is recursively repeated until we have obtained combined one-step-ahead out-of-sample forecasts for each of our four combining models—AVE, OLS, MAD and ANN—on each of the four stock indices we study—SP500, NIKKEI, TSEC and FTSE—for the period 22 June 1983 to September 30, 1987. These are the out-of-sample forecasts used to evaluate the performance of the combining models.

COMPARING FORECASTS

In this section we evaluate the out-of-sample forecasting ability of our various combining models. When considering our evidence, it is particularly important to note that all of our tests focus on out-of-sample comparisons of the combining models, as obtained with the procedure described above. Since we look *exclusively* at out-of-sample comparisons of the combining methods, we do *not* automatically favour the method with the most flexibility to fit the data in sample; i.e. the ANN method. While ANN would clearly be expected to dominate in sample, since it nests the fully linear specification as a special case, there is in fact no guarantee that ANN will dominate out of sample. Indeed, it is possible that ANN could overfit the data in sample and thus produce out-of-sample ANN forecasts that are inferior to forecasts from the simpler linear combining models. By focussing on out-of-sample tests we are therefore better able to assess the practical significance of going to a nonlinear ANN specification in environments in which forecasters are trying to predict unknown future events.

Summary statistics

We begin our assessment of the various combining methods by comparing the models' abilities to reproduce broad features of the data. To this end, we present in Table I some summary statistics on the various individual and combined volatility forecasts, and on the implied standardized returns from our various individual and combining models, for the out-of-sample period 22 June 1983 to 30 September 1987. The first two columns in Table I list the index name and forecast method, respectively.

Columns three and four of Table I report the mean and standard deviation of actual stock returns volatility for each index, as well as the mean and standard deviation of the stock volatility forecasts produced by each individual and combined model. We would expect the mean of the various volatility forecasts to be similar to the actual mean for each series and the standard deviation of the forecasts to be smaller than the standard deviation of the actual data. This is true for every combining method except MAD. Forecasts of volatility produced by the MAD combining method are typically one half to one third the

Table I. Out-of-sample statistics for stock returns and volatility from 22 June 1983 to 30 September 1987

Index name	Forecast method	Volatility Forecasts Mean $\times 10^{-5}$	St. dev $\times 10^{-5}$	Standard deviation	Standardized returns Skewness	Excess kurtosis	ARCH p-value
SP500	Actual	7.02	13.51	1.00	−0.12	1.75	0.062
	MAV	6.92	3.36	1.07	0.03	1.37	0.336
	GAR	6.94	2.38	1.02	−0.03	1.28	0.474
	AVE	6.93	2.81	1.04	−0.00	1.31	0.420
	OLS	6.82	1.71	1.02	−0.12	1.61	0.482
	MAD	2.50	0.82	1.73	−0.12	1.75	0.056
	ANN	6.83	1.42	1.02	−0.11	1.77	0.377
NIKKEI	Actual	5.52	12.08	1.00	−0.32	2.89	0.000
	MAV	5.48	4.83	1.11	−0.62	2.95	0.009
	GAR	5.12	4.46	1.07	−0.61	2.28	0.709
	AVE	5.30	4.37	1.06	−0.61	2.47	0.309
	OLS	4.99	3.78	1.05	−0.51	1.82	0.209
	MAD	1.86	1.52	1.73	−0.53	1.76	0.160
	ANN	4.97	3.90	1.05	−0.49	1.72	0.068
FTSE	Actual	8.80	12.47	1.00	−0.24	0.10	0.000
	MAV	8.81	6.02	1.20	0.06	1.86	0.048
	GAR	10.14	3.38	0.94	−0.26	−0.02	0.786
	AVE	9.48	4.52	1.01	−0.24	0.18	0.180
	OLS	9.30	3.14	0.98	−0.28	−0.04	0.656
	MAD	4.23	1.32	1.45	−0.28	−0.03	0.735
	ANN	9.64	3.45	0.97	−0.29	0.04	0.461
TSEC	Actual	3.55	6.51	1.00	0.06	1.37	0.001
	MAV	3.55	2.53	1.17	0.01	4.52	0.778
	GAR	6.87	3.66	0.77	0.24	2.47	0.002
	AVE	5.21	2.53	0.85	0.04	1.52	0.841
	OLS	4.92	2.03	0.88	−0.02	1.99	0.942
	MAD	1.60	0.70	1.52	0.01	1.51	0.995
	ANN	4.74	1.87	0.90	0.01	2.04	0.966

magnitude of the other models' forecasts, on average, with a correspondingly smaller standard deviation.[9]

Columns five to eight of Table I report statistics on the standardized residual returns from each index; i.e. $\hat{\varepsilon}_t/\sqrt{F_t}$. When divided by its forecasted standard deviation $\sqrt{F_t}$, the returns residuals $\hat{\varepsilon}_t$ should have a standard deviation of 1, as seen down column five in the 'Actual' rows of Table I (where actual return residuals are divided by their actual standard deviations). It is therefore interesting to note from column five that, for the AVE, OLS and ANN combining methods, standard deviations of the forecasted standardized returns are generally within 15% of their desired value of one. Conversely, MAD delivers standardized returns with standard deviations much greater than unity, consistent with the downward bias of the MAD conditional variance forecasts observed in column four. In terms of higher moments, it is well known (e.g.

[9]Mechanically this occurs because, while the estimated coefficients on the MAV and GARCH individual forecasts for all combining models except MAD sum close to unity, the estimated MAD coefficients on the MAV and GARCH individual forecasts sum well below unity without a compensatingly large β_0 estimate. This may occur because the MAD estimation method relies on symmetry of the forecast error about its mean and, as Table I reveals, this is unlikely to be satisfied when modelling conditional variances of stock returns. Conversely, weight selection by methods such as OLS is remarkably robust to non-normal distributions, if not to outliers.

Bollerslev *et al.* 1991) that no volatility model yet developed can fully account for all the skewness and excess kurtosis in stock market data and, as seen in columns six and seven, our models are no exception.

Column eight of Table I reports the p-value from an LM test of the null hypothesis that the standardized residuals do not display autoregressive conditional heteroscedasticity at 24 lags. As expected, the actual data generally display strong evidence of ARCH (i.e. p-values near zero). Stock volatility-forecasting models are designed to remove ARCH from returns without changing the lower moments of the standardized returns distribution. A reliable method for combining volatility forecasts should therefore also have this property. It is thus interesting to note, from column eight of Table I, that all our combining models remove ARCH even when the individual forecasts being combined do not. For example, GARCH does not remove ARCH in the TSEC at the 1% level of significance ($p = 0.002$) but all the combining methods do. This occurs because the MAV and GARCH forecasts are somewhat independent so that, during time periods when one model fails to capture ARCH, the other model may not fail. Thus, assuming that each forecast has some relevant information that is not contained in the other forecast, the combined forecast is less likely to fail to capture ARCH effects than either of the individual MAV or GARCH models alone. The non-overlapping nature of information in MAV and GARCH is substantiated with the encompassing tests below.

Mean and absolute forecast errors

Table II reports the root mean squared forecast error (RMSE) and root mean absolute forecast error (RMAE) for each of the individual models and each of the combining methods for the out-of-sample period 22 June 1983 to 30 September 1987. In terms of RMSE, the OLS and ANN combined forecasts share common performance characteristics with both models generally performing as well as, or better than, the other models. Only the performance of the MAD combining method is noticeably worse than that of its competitors in terms of RMSE. Conversely, as we would expect, the Mean Absolute Deviations combined forecast has consistently lower mean absolute errors than the other models. The ANN and OLS forecasts again perform about the same, with both having lower RMAE than the simple AVE combined forecast.

On the basis of absolute errors alone, we would select MAD as the 'best' combining method while, on the basis of squared errors alone, ANN or OLS would be preferred. However, these comparisons do not account for the issues raised in the previous section in which we documented the failure of MAD to capture key features of the data being forecasted. Furthermore, neither comparison by RMSE nor by RMAE provides us with an indication of

Table II. Root mean squared error and root mean absolute error for out-of-sample from 22 June 1983 to September 1987

Index →	SP500		NIKKEI		FTSE		TSEC	
Forecast Method ↓	RMSE $\times 10^{-4}$	RMAE $\times 10^{-3}$	RMSE $\times 10^{-4}$	RMAE $\times 10^{-3}$	RMSE $\times 10^{-4}$	RMAE $\times 10^{-3}$	RMSE $\times 10^{-4}$	RMAE $\times 10^{-3}$
MAV	1.36	8.71	1.17	7.69	1.30	9.30	0.66	6.21
GAR	1.35	8.66	1.16	7.48	1.24	9.40	0.77	7.66
AVE	1.35	8.68	1.16	7.53	1.25	9.30	0.68	6.83
OLS	1.35	8.62	1.16	7.46	1.23	9.22	0.67	6.73
MAD	1.42	7.94	1.23	7.02	1.31	8.64	0.67	5.63
ANN	1.35	8.62	1.17	7.46	1.24	9.30	0.67	6.65

whether any one model's performance is significantly better than that of other models, in a formal statistical sense. We therefore investigate in the next section an additional means of comparison between forecasting models that allows for tests of significantly better performance: comparison by forecast encompassing.

Encompassing tests

The test for encompassing-in-forecast we employ was introduced by Chong and Hendry (1986) and applied to the out-of-sample comparison of combined forecasts by Hallman and Kamstra (1989). This test formalizes the intuition that model j should be preferred to model k if model j can explain what model k cannot explain, without model k being able to explain what model j cannot explain. As such, the test provides a useful method for ranking out-of-sample forecasts.

The formal test for encompassing in forecast is based on a set of OLS regressions of the forecast error from one model on the forecast from the other model. Thus, letting E_j be model j's forecast error and F_k be model k's forecast, the tests for encompassing involve testing for significance of the θ_1 parameter in the regression

$$E_{jt} = \theta_0 + \theta_1 F_{kt} + \eta_t \tag{4}$$

Given forecasts from two models, j and k, we test the null hypothesis that neither model encompasses the other. We first regress the forecast error from model j on the forecast from model k, as in equation (4). Call this 'regression jk' and the resulting estimate of the θ_1 coefficient $\hat{\theta}_1(jk)$. We then regress the forecast error from model k on the forecast from model j. Call this 'regression kj' and the resulting θ_1 estimate $\hat{\theta}_1(kj)$. If $\hat{\theta}_1(jk)$ is *not* significant at some predetermined level, but $\hat{\theta}_1(kj)$ *is* significant, then we reject the null hypothesis that neither model encompasses the other in favour of the alternative hypothesis that model j encompasses model k. Conversely, if $\hat{\theta}_1(kj)$ is *not* significant, but $\hat{\theta}_1(jk)$ is significant, then we say that model k encompasses model j. However, if both $\hat{\theta}_1(kj)$ and $\hat{\theta}_1(jk)$ are significant, or if both $\hat{\theta}_1(kj)$ and $\hat{\theta}_1(jk)$ are not significant, then we fail to reject the null hypothesis that neither model encompass the other. Multicollinearity can lead to both estimated coefficients being insignificant, while sufficiently non-overlapping information sets can lead to both estimated coefficients being significant.

Columns three to eight of Table III contain p-values associated with the heteroscedasticity-robust t-statistics on θ_1 for all possible j–k comparisons. P-values less than 0.010 reveal that the forecast from the model listed along the top of the table explains, with 1% significance, the forecast error from the model listed down the left side of the table and thus that the model listed down the side cannot encompass the model listed along the top, at the 1% level. For example, the p-value of 0.942 in the ANN row and AVE column of Table III in the TSEC reveals that the AVE combining method's forecast of volatility on the Toronto Stock Exchange Composite Index does *not* explain the ANN combining method's forecast error at the 1% significance level. Conversely, the p-value of 0.000 in the ANN column and AVE row for the TSEC reveals that the ANN forecast can indeed explain part of the AVE forecast error at the 1% level. Thus, from these two p-values, one would conclude that ANN encompasses AVE at the 1% level in the TSEC.

Pairwise comparisons for encompassing, as conducted in the preceding example, reveal that ANN is the only model whose forecast is *not* encompassed by at least one other forecast in at least one index at the 1% level of significance. Every other model is encompassed at least once. For example, ANN encompasses AVE in the SP500, FTSE and TSEC. ANN also encompasses OLS in the TSEC and encompasses MAD in the NIKKEI. Thus, we conclude that ANN is significantly preferred to other forecast combining techniques on the basis of encompassing tests.

Table III. Tests for out-of-sample forecast encompassing p-values on θ_1 from: $E_{it} = \theta_0 + \theta_1 F_{kt} + \eta_t$

Index name ↓	Forecast error E_{jt} from ↓	Forecast F_{kt} from ↓ MAV	GAR	AVE	OLS	MAD	ANN
SP500	MAV	—	0.000	0.000	0.003	0.194	0.000
	GAR	0.011	—	0.011	0.008	0.038	0.000
	AVE	0.000	0.001	—	0.005	0.092	0.000
	OLS	0.970	0.799	0.930	—	0.006	0.032
	MAD	0.032	0.029	0.028	0.375	—	0.600
	ANN	0.794	0.847	0.944	0.559	0.210	—
NIKKEI	MAV	—	0.635	0.059	0.245	0.401	0.116
	GAR	0.725	—	0.268	0.141	0.195	0.119
	AVE	0.092	0.248	—	0.182	0.278	0.113
	OLS	0.934	0.750	0.903	—	0.593	0.409
	MAD	0.000	0.001	0.000	0.001	—	0.003
	ANN	0.825	0.707	0.753	0.445	0.541	—
FTSE	MAV	—	0.000	0.000	0.000	0.000	0.000
	GAR	0.115	—	0.059	0.091	0.050	0.141
	AVE	0.000	0.000	—	0.000	0.000	0.000
	OLS	0.106	0.152	0.109	—	0.093	0.104
	MAD	0.173	0.066	0.115	0.127	—	0.173
	ANN	0.019	0.108	0.031	0.041	0.044	—
TSEC	MAV	—	0.785	0.157	0.016	0.000	0.000
	GAR	0.581	—	0.000	0.000	0.022	0.017
	AVE	0.002	0.005	—	0.000	0.001	0.000
	OLS	0.249	0.539	0.407	—	0.004	0.000
	MAD	0.203	0.218	0.170	0.445	—	0.527
	ANN	0.239	0.712	0.924	0.014	0.032	—

CONCLUSIONS

The preponderance of the statistical evidence presented above suggests that ANN combined forecasts generally outperform forecasts from a variety of traditional combining methods in the international stock market data we employ. The practical significance of this result is evident from the out-of-sample nature of the tests employed. Although our ANN nests traditional linear models as special cases, and would therefore be expected to dominate these models in sample, there was no *a priori* guarantee that ANNs would dominate out of sample, especially if the ANNs overfit the in-sample data. The fact that our ANNs did out-perform traditional models in the out-of-sample tests therefore reveals that flexible ANNs may be preferred to more restrictive traditional models in environments in which forecasters are actually trying to predict unknown future events.

The reason for the ANNs' relative success is seen in Figure 2, which plots the ANN and OLS combining functions for the TSEC data. The curved surface of crosses in Figure 2 plots the functional relationship—estimated on the in-sample data 1 January 1980 to 21 June 1983—between the MAV and GARCH individual forecasts and the ANN combined forecast.

Similarly, the flat surface of circles reveals the way in which the MAV and GARCH individual forecasts are combined to produce the OLS forecast. The pillars rising up from the floor of Figure 2, to pierce the forecasting surfaces, represent the actual out-of-sample data points—22 June 1983 to 30 September 1987—which the OLS and ANN models combine (the height of the pillars is slightly raised so that they are easily visible through the surfaces). The surfaces in Figure 2 therefore indicate the difference in response we can expect from the ANN and OLS

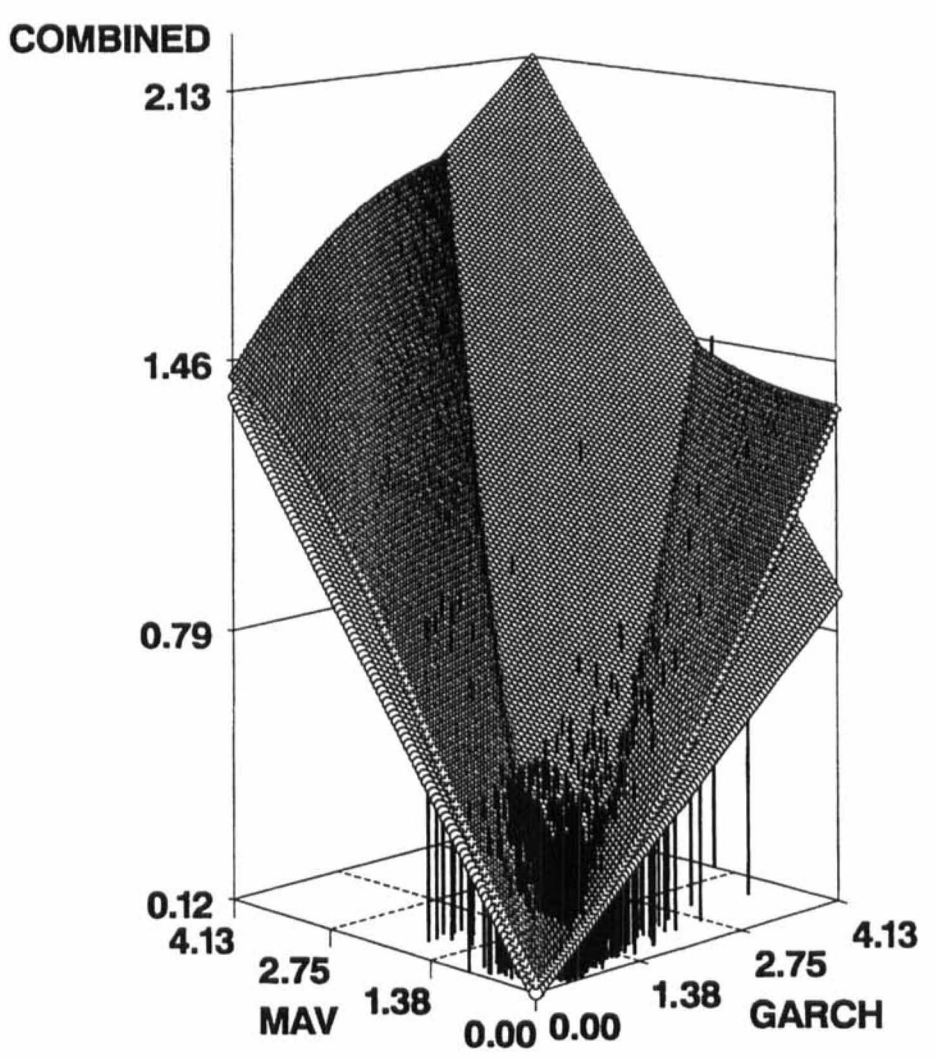

Figure 2. Forecasts for TSEC volatility. Values scaled up by 10,000.

models for various pairs of GARCH-MAV forecast inputs, while the pillars provide some indication as to whether the differences in response are relevant for the out-of-sample period we investigate. We expect to find statistically important differences between the ANN and OLS combined forecasts when the surfaces are dissimilar in the range where the forecasts (i.e. pillars) occur.

The nonlinear nature of the ANN is easily seen in Figure 2 as the ANN surface lies above the OLS when both MAV and GARCH are small and when either the MAV or GARCH are very large. However, the ANN surface scoops below the OLS surface in the area where MAV and GARCH yield similar forecasts. As seen by the cluster of pillars in the centre front of Figure 2, many of the data points occur in the area where the OLS and ANN forecasting functions intersect. Thus, we would expect the ANN and OLS combined forecasts to produce forecasts of return variance and standardized returns which are quite similar *on average* and, as revealed by the statistics in Tables I and II, this is indeed the case. However, from Figure 2 we also see that numerous data points occur away from the curve intersections (e.g. in the centre and edges of the figure). Since the ANN and OLS combining functions are quite different in this region, we would also expect to see some important differences between the ANN and OLS forecasts, and we do. Indeed, the differing functional treatment of these out-of-sample data points accounts for the ANN's ability to encompass the OLS forecast in Table III for TSEC at the 0.1% level of significance.

Results similar to those in Figure 2 can be demonstrated for the other stock market indices and traditional combining methods we study. In sum, combining with ANNs produces out-of-sample forecasts that are in general at least as accurate (and often considerably more accurate) than forecasts produced by a variety of traditional techniques. We have shown that the superiority of our ANN combined forecasts arises because of the ANN's flexibility to account for potentially complex nonlinear relationships not easily captured by traditional linear combining methods.

ACKNOWLEDGEMENTS

For helpful comments and suggestions we thank an anonymous referee and the editor of this journal, as well as Burton Hollifield, Lisa Kramer, and participants at the 1994 Conference on Neural Networks in the Capital Markets in Pasadena, California. Financial support from the Social Sciences and Humanities Research Council of Canada is gratefully acknowledged. The usual disclaimer applies.

REFERENCES

Bollerslev, T., Chou, R. T. and Kroner, K. F., 'ARCH modelling in finance: a review of the theory and empirical results,' *Journal of Econometrics*, **54** (1991), 1–30.

Chong, Y. Y. and Hendry, D. F., 'Econometric evaluation of linear macroeconomic models', *Review of Economic Studies*, **53** (1986), 671–90.

Clemen, R. T., 'Combining forecasts: a review and annotated bibliography', *International Journal of Forecasting*, **5** (1989), 559–83.

Donaldson, R. G. and Kamstra, M., 'An international comparison of alternative stock volatility models', Simon Fraser University Department of Economics Discussion Paper No. 94–06, 1994.

Granger, C. W. J., 'Invited review: combining forecasts—twenty years later', *Journal of Forecasting*, **8** (1989), 167–173.

Hallman, J. and Kamstra, M., 'Combining algorithms based on robust estimation techniques and cointegrating restrictions', *Journal of Forecasting*, **8** (1989), 189–98.
Hertz, J., Krogh, A. and Palmer, R. G., *Introduction to the Theory of Neural Computing*, Redwood City, CA: Addison-Wesley, 1991.
Hjorth, U. and Holmqvist, L., 'On model selection based on validation with applications to pressure and temperature prognosis', *Applied Statistics*, **30** (1981), 264–74.
Hornik, K., 'Approximation capabilities of multilayer feedforward networks', *Neural Network*, **4** (1991), 251–7.
Hornik, K., Stinchcombe, M. and White, H., 'Multi-layer feedforward networks are universal approximators', *Neural Network*, **2** (1989), 359–66.
Hornik, K., Stinchcombe, M. and White, H., 'Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks', *Neural Network*, **3** (1990), 551–60.
Kavalieris, L., 'The estimation of the order of an autoregression using recursive residuals and crossvalidation', *Journal of Times Series Analysis*, **10** (1989), 271–8.
Kuan, C., and White, H., 'Artificial neural networks: an econometric perspective', *Econometric Reviews*, **13** (1994), 1–91.
Min, C. and Zellner, A., 'Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates', *Journal of Econometrics*, **56** (1993), 89–118.
Pagan, A. and Schwert, W., 'Alternative models for conditional stock volatility', *Journal of Econometrics*, **45** (1990), 267–90.
Stoica, P., Eykhoff, P., Janssen, P. and Soderstrom, T., 'Model selection by cross-validation', *International Journal of Control*, **43** (1986), 1841–78.
Stinchcombe, M. and White, H., 'Using feedforward networks to distinguish multivariate populations', *Proceedings of the International Joint Conference on Neural Networks*, 1994.
White, H., 'Some asymptotic results for learning in single hidden layer feedforward network models', *Journal of the American Statistical Association*, **84** (1989), 1003–13.
White, H., 'Connectionist nonparametric regression multilayer feedforward networks can learn arbitrary mappings', *Neural Network*, **3** (1990), 535–49.

Authors' biographies:
Glen Donaldson is the Finning Ltd Associate Professor of Finance in the Faculty of Commerce and Business Administration at the University of British Columbia. His research areas include applied asset pricing and financial econometrics.
Mark Kamstra is an Assistant Professor of Economics at Simon Fraser University. His research areas include econometrics, neural networks, and applied finance.

Authors' addresses:
Glen Donaldson, Faculty of Commerce and Business Administration, University of British Columbia, Vancouver BC, Canada, V6T 1Z2.
Mark Kamstra, Department of Economics, Simon Fraser University, Burnaby BC, Canada, V5A 1S6.