MEASURING IMPLIED VOLATILITY: IS AN AVERAGE BETTER? WHICH AVERAGE?

LOUIS H. EDERINGTON
WEI GUAN*

Options researchers have argued that by averaging together implied standard deviations, or ISDs, calculated from several options with the same expiry but different strikes, the noise in individual ISDs can be reduced, yielding a better measure of the market's volatility expectation. Various options researchers have suggested different weighting schemes for calculating these averages. In the forecasting literature, econometricians have made the same argument but suggested quite different weighting schemes. Ignoring both literatures, commercial vendors calculate ISD averages using their own weightings. We compare the averages proposed in both the options and econometrics literatures and the averages used by major commercial vendors for the S&P 500 futures options market. Although some averages forecast better than others, we find that the question of the best weighting scheme is of secondary importance. More

The authors acknowledge the helpful comments of Brian Barrett, workshop participants at the University of Oklahoma and Otago University, an anonymous referee, and especially the editor, Robert I. Webb. The final draft of this article was completed while Ederington was on sabbatical at Otago University. He thanks Otago for its support.
*Correspondence author, School of Management, Department of Accounting and Finance, Delaware State University, Dover, Delaware 19901-2277; e-mail: wguan@dsc.edu

Received November 2000; Accepted December 2001

- *Louis H. Ederington is an Oklahoma Bankers Professor of Finance at the Michael F. Price College of Business, University of Oklahoma, Norman, Oklahoma.*
- *Wei Guan is an Assistant Professor of Finance, School of Management, Department of Accounting and Finance, Delaware State University, Dover, Delaware.*

The Journal of Futures Markets, Vol. 22, No. 9, 811–837 (2002)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/fut.10034

important is the fact that the ISDs are upward biased measures of expected volatility. Fortunately, this bias is stable over time, so past bias patterns can be used to obtain unbiased volatility forecasts. Once this is done, most ISD averages forecast better than time series and naive models, and the differences between the averages produced by the various proposed weighting schemes are small. Jrl Fut Mark 22:811–837, 2002

INTRODUCTION

In most options markets, numerous options with the same expiry but different strike prices are traded. When plugged into the appropriate pricing model, the price of each yields an implied volatility of the underlying asset's return over the remaining life of the option, i.e., an implied standard deviation (ISD). Because they are forecasting volatility for the same underlying asset over the same period, if markets are perfect and the pricing model is correct, implied volatilities calculated from options with the same expiry but different strike prices should be identical, and so should the ISDs calculated from calls and puts. In reality, however, options with the same expiry but different strike prices often yield quite different ISD estimates, and the ISDs may also differ between puts and calls with the same expiry and strike. Consequently, it is unclear which ISD or combination of ISDs provides the best measure of the market's volatility expectation over the life of the options.

Although some ISD differences are persistent, as in the well-known volatility smile, others, specifically those due to market imperfections such as bid-ask spreads, nonsynchronous price observations, and discrete prices, are transitory. To the extent that market imperfections cause the implied volatility calculated from a single option's price on day t to differ from the market's true expectation, averaging together the ISDs calculated from the prices of numerous options with the same expiry should reduce the error.[1] Based on this reasoning, researchers have argued that averaging together the ISDs calculated from numerous options with the same expiry will yield a better measure of the market's volatility expectation over the life of the options than a single ISD. But what average? Should it be based on at-the-money options only or a

[1]Obviously this assumes the errors are not perfectly correlated. However, it is argued that the errors introduced by market imperfections should be relatively independent for different strikes, and could be negatively correlated for puts and calls with the same strike. For instance, as pointed out by Fleming, Ostdiek, and Whaley (1995), measurement errors due to nonsynchronous prices will tend to be of opposite signs for puts and calls.

wider set? Is a simple or weighted average better? If the latter, what weights are best? As reviewed below, starting with Latane and Rendleman (1976), several different weighting schemes have been proposed in the options literature. In addition, there is a large econometrics/forecasting literature on the question of combining forecasts in which numerous different weighting schemes have been proposed. Although this literature has focused on forecasts in general, not ISDs specifically, and the authors probably had in mind forecasts of such variables as GDP and inflation rather than market volatility, the principles are still applicable.

Although the proposed weighting schemes differ widely, both the options researchers and the econometricians invariably recommend basing the averages on a large set of forecasts. As Figlewski and Ulrich (1983) point out, the rationale is similar to that for diversification of portfolios. To the extent various forecasts are partially independent, combining them reduces the forecast error so the larger the set underlying the average, the smaller its error. However, ignoring the advice of both the econometricians and the options researchers, commercial vendors of implied volatility figures, such as Bloomberg and the CBOT, normally use only a few at-the-money options to calculate their implied volatility averages. Although most vendors employ an unweighted average, at least one uses a weighted average—but not one of the weighting schemes advocated in either the options or econometrics literatures.

In this article, we explore and compare various averaging schemes proposed by options researchers and econometricians, and also the schemes used by the major commercial vendors. First, we show that none of the averages—neither the weighted averages proposed in the academic literature nor the averages provided by commercial vendors—provide implied volatility estimates that forecast actual market volatility very well. Indeed, most fail to consistently beat a naive model that assumes volatility never changes. We further find that these models fail primarily because they do not adjust for the ISD biases implicit in the smile. Fortunately, because this bias is persistent, it is possible to obtain revised forecasts of future volatility that forecast considerably better than the unadjusted ISDs. When we correct the Black-Scholes implied volatilities for bias prior to constructing the averages, several of the corrected models significantly outperform the more naive models. Finally, we find that, at least in our market, i.e., options on S&P 500 futures, the weighting scheme matters relatively little after one corrects for bias because there is very little noise in individual ISD estimates to be averaged out. Although a few of the weighting schemes can be discarded as inferior, there is no significant forecasting difference between most, and it matters little whether one averages together many ISDs or just a few.

The article is organized as follows. In the next section, we review the literature on combining forecasts from both the econometrics and options literatures, and describe the combination procedures employed by major commercial vendors. In the Data and Measures section, we describe our data, define the models we compare, present our measures of forecast accuracy, and present initial results on forecast accuracy. In the Unbiased Forecasts section, we develop our bias adjustment and use adjusted models to forecast volatility. In the Weights Question section we explore the weights question. Then we conclude the article.

IMPLIED VOLATILITY AVERAGES IN THEORY AND PRACTICE

The Econometrics and Forecasting Literature

There is a fairly extensive econometrics literature on combining individual forecasts to obtain the best "consensus" forecast. Indeed, a 1989 survey by Clemen (1989) provided an annotated bibliography of over 200 articles on the topic.[2] Beginning with Bates and Granger (1969), most statistical models start with the objective of minimizing the mean squared error or error variance of the consensus forecast. Let $F_{j,t}$ be an unbiased forecast of Y_t and designate its forecast error as $E_{j,t} = F_{j,t} - Y_t$. Define as a weighted average of J different forecasts of Y_t:

$$\overline{F}_t = \sum_{j=1}^{J} w_{j,t} F_{j,t} \tag{1}$$

where w_j is the weight attached to forecast j. $\overline{F}_t$'s forecast error is $\overline{E}_t = \overline{F}_t - Y_t$. Suppressing the t subscript, the variance of $\overline{E}_t$ is:

$$\mathrm{Var}(\overline{E}) = \sum_{j=1}^{J} w_j^2 \sigma_j^2 + \sum_{j=1}^{J} w_j \sum_{i \neq j} w_i \sigma_{ij} \tag{2}$$

where σ_j^2 is the variance of $E_{j,t}$ (which is generally assumed constant over time), and σ_{ij} is the covariance of $E_{i,t}$ and $E_{j,t}$. As noted above, the most

[2]Many of these are psychological models dealing with how to combine judgmental and statistical forecasts, a question that we do not examine.

popular criterion for combining forecasts is to seek the weights, which minimize the $\mathrm{Var}(\overline{E})$ in Equation 2. If we take derivatives of $\mathrm{Var}(\overline{E})$ with respect to the w_j subject to the condition that the weights sum to 1 and set these equal to zero, we obtain a set of J first-order conditions:

$$2w_j\sigma_j^2 + \sum_{i \neq j} w_i \sigma_{ij} - \lambda = 0 \qquad \text{for all } j \tag{3}$$

where λ is the Lagrangian multiplier enforcing the condition that the weights sum to 1.0, and the other terms are as defined above.

If the variances and covariances are known, the variance minimizing weights can be obtained by solving the set of J equations in Equation 3 for the w_j. In reality, of course, these variances and covariances are unknown; so much of the forecasting literature deals with how best to proceed in light of this uncertainty. If a past history exists, the various variances and covariances in Equation 3 may be estimated from this data, and these estimates can be used to estimate the weights w_j. In practice, however, this "unconstrained" approach has not been very successful. Negative weights are commonly obtained, and the resulting weighted average forecast usually does not forecast very well (Clemen, 1989; Figlewski & Ulrich, 1983; MacDonald & Marsh, 1994; Newbold & Granger, 1974; Winkler, 1981; and Winkler & Makridakis, 1983). Apparently, the covariances are either difficult to estimate or unstable, so that past sample covariances are a poor guide to future population covariances. Consequently, a number of researchers have proposed imposing constraints on the form of the variance-covariance matrix.

One popular constraint is to assume that the various forecasts are independent. Because different forecasts are normally based on similar information sets, it is generally recognized that this assumption is unrealistic in most cases, but the resulting forecasts often forecast reasonably well. Assuming the J forecasts are independent, i.e., assuming all the covariances σ_{ij} are zero, the variance minimizing weights from the J equations in 3 reduce to:

$$w_j = \frac{1}{\sigma_j^2 \sum_{i=1}^{J} \frac{1}{\sigma_i^2}} \tag{4}$$

Weighted average forecasts based on these weights have generally outperformed those based on unconstrained solutions to the set of equations in 3 (Clemen, 1989; Figlewski, 1983; Newbold & Granger, 1974; Winkler, 1981; Winkler & Makridakis, 1983). Note that in 4, the more accurate forecasts receive higher weights. More specifically, the weights are inversely proportional to the variances of their error terms σ_j^2. Also, if the J forecasts are both independent, i.e., $\sigma_{ij} = 0$, and equally accurate, i.e., $\sigma_j^2 = \sigma^2$ for all j, then the optimal weights reduce to $w_j = \lambda/2\sigma^2 = 1/J$, that is, a simple average of the J forecasts.

As described in Figlewski and Ulrich (1983) and Palm and Zellner (1992), an alternative is to assume that the covariance between forecasts i and j is not zero but has a particular form. Specifically, it is often assumed that j's forecast error can be decomposed into two independent components: an expectational error, η_t, which is common to all forecasts of Y_t, and an error specific to forecast j, $\varepsilon_{j,t}$:[3]

$$E_{j,t} = F_{j,t} - Y_t = \varepsilon_{j,t} + \eta_t \tag{5}$$

If $\varepsilon_{j,t}$ and $\varepsilon_{i,t}$ are assumed independent, $\text{Covar}(E_{j,t}, E_{i,t}) = \text{Var}(\eta_t)$. Assuming further that the variance of η is time invariant and, therefore, suppressing the time subscript, the variance of the forecast errors is:

$$\text{Var}(\overline{E}) = \text{Var}(\eta) + \sum_{j=1}^{J} w_j^2 \sigma_j^{*2} \tag{6}$$

where σ_j^{*2} is the variance of $\varepsilon_{j,t}$. Setting the first derivatives of 6 with respect to the w_j equal to zero and solving the resulting set of equations yields the variance minimizing set of weights:

$$w_j = \frac{1}{\sigma_j^{*2} \sum_{i=1}^{J} \frac{1}{\sigma_i^{*2}}} \tag{7}$$

Note that Equation 7 is identical to Equation 4 except that we replace the variance of the forecast errors, E_j, with the variance of the forecast specific error, ε_j.[4]

The problem in implementing Equation 7 is obtaining reliable estimates of σ_j^{*2}, which requires separating out the common error η_t. One possibility is to calculate the forecast errors $E_{j,t}$ for each of J forecasts each past period t and then regress these T^*J forecast errors on a series of dummy variables for each of the T periods D_t, where $D_t = 1$ for

[3]Sometimes it is assumed (e.g., Figlewski & Ulrich, 1983) that the error is proportional to η. For example, it is assumed the error term is: $E_{j,t} = F_{j,t} - Y_t = \varepsilon_{j,t} + \lambda_j \eta_t$, where the parameter λ_j varies from forecast to forecast.

[4]This is sometimes referred to as the single index model.

observations in period t and zero otherwise. In other words, the regression takes the form:

$$E_{j,t} = \sum_{t=1}^{T} \hat{\eta}_t D_t + \hat{\varepsilon}_{j,t} \tag{8}$$

The coefficients of the dummy variables provide estimates of the common expectational errors η_t in Equation 5 and the residuals provide measures of the forecast specific errors, $\varepsilon_{j,t}$. The sample variances of the residuals, $\hat{\varepsilon}_{j,t}$, provide estimates of the σ_j^{*2}. Substituting these sample variances into Equation 7 provides the weights.

As summarized in Granger (1989), Clemen (1989), and Palm and Zellner (1992), several general results have emerged from empirical comparisons of some of these alternatives to date. First, average or consensus forecasts usually out-forecast the individual forecasts on which they are based and often are substantially better. Second, averages obtained by solving the set of J equations in 3 utilizing unconstrained sample covariances generally perform worse than those based on Equation 4 using estimated variances only. Third, simple averages in which all forecasts are weighted equally often perform better than weighted averages, such as those based on Equation 4, in which those forecasts with the best track record receive higher weights.

The Options Literature

Averaging prescriptions in the options literature differ in two important respects from those in the econometrics literature. First, while econometricians have focused on the general question of how to combine individual forecasts, $F_{j,t}$, to obtain the best consensus forecast, options researchers have focused on the more specific question of how best to combine individual implied volatilities, $\text{ISD}_{j,t}$. Second, although the former generally base their suggested weights on the track record of the individual forecasts, the latter base their weights on the properties of the options at time t.

The first proposal due to Latane and Rendleman (1976) was to calculate a weighted average of the ISDs from a set of options J using Equation 1 with weights equal to the option's relative vega, i.e.,

$$w_{j,t} = \left(\frac{\partial C_{j,t}}{\partial \text{ISD}_{j,t}}\right) \Big/ \sum_{i=1}^{J} \left(\frac{\partial C_{i,t}}{\partial \text{ISD}_{i,t}}\right)$$

where $C_{j,t}$ is the price of option j on day t and $\text{ISD}_{j,t}$ is the implied standard deviation calculated from $C_{j,t}$ using the Black-Scholes formula.[5] Because vega measures the sensitivity of option j's price, $C_{j,t}$, to changes in $\text{ISD}_{j,t}$, the same noise in option prices $C_{j,t}$ due to market imperfections should create more noise in the implied volatilities of small vega options than in large vega options. Hence, similar to the weights obtained from equation 4, weighting option j's ISD by its relative vega should give more weight to the lower variance options. Because vega is maximized when the strike price is just slightly above the underlying asset price, this scheme puts more weight on ISDs calculated from at-the-money options than options based on deep in-the-money or far out-of-the money options.

Chiras and Manaster (1978) propose a variation on this in which the weights are set equal to the option's price elasticity with respect to $\text{ISD}_{j,t}$, specifically $w_{j,t} = [(\partial C_{j,t}/\partial \text{ISD}_{j,t})(\text{ISD}_{j,t}/C_{j,t})] \div [\sum_{i=t}^{J}(\partial C_{i,t}/\partial \text{ISD}_{i,t})(\text{ISD}_{i,t}/C_{i,t})]$. Due to the addition of the term $(\text{ISD}_{j,t}/C_{j,t})$, this scheme puts more weight on out-of-the-money options (with low prices $C_{j,t}$) than the vega weights model.

Beckers (1981) and Whaley (1982) propose estimating a consensus ISD forecast at time t by finding the single ISD that minimizes

$$\sum_{j=1}^{J} v_{j,t}[C_{j,t} - BS_{j,t}(\text{ISD})]^2$$

where BS refers to the Black-Scholes price calculated from that ISD. In Whaley's model $v_{j,t} = 1/J$, while in Becker's it is the relative vega. Although all the other models that we have discussed form weighted averages of the ISD's from J options using Equation 1, Beckers and Whaley's solves for a single ISD based on a weighted average of the squared pricing errors. However, according to Beckers, this is approximately equivalent to a traditional weighted average ISD model in which the weights are the squared vegas.[6] Because the weights are approximately proportional to the vegas raised to a high power, this scheme puts even more weight on close-to-the-money options than the vega weighting scheme of Latane and Rendleman.

[5]Actually, Latane and Rendleman used a weighted geometric average not an arithmetic average as in Equation 1. Consequently, as shown by Chiras and Manaster (1978), their weights did not sum to 1, and the resulting ISD average was biased toward zero. However, because the basic idea of vega weights was theirs, we, like some others, credit the version of Equation 1 with vega weights to them.
[6]In fact, we find that in our data set this ISD measure is most highly correlated ($r = 0.99998$), with a weighted average ISD model where the individual ISDs are weighted by their relative vegas raised to the fourth power.

Several studies have compared subsets of these weighting schemes empirically with differing results. Beckers (1981) found that, in the market for individual equity call options, the ISD for the single call with the highest vega tended to forecast slightly better than Becker's own scheme and better than a weighted average with vega weights. However, his data period was quite short: October 1975 through January 1976. Roughly consistent with Beckers, Feinstein (1989) found that in the S&P 500 options market, the single nearest but out-of-the-money call forecast marginally better than other models considered. In contrast, Gemmill's (1986) best performing ISD was the single ISD calculated from a deep-in-the-money call. Turvey (1990) found that in two commodity options markets, soybeans and cattle, a weighting scheme with vega weights forecast better than either an equally weighted model, elasticity weights, and a single at-the-money put. Note that in three out of four studies, a single ISD forecast better than the averages that they considered—certainly not a good sign for the averaging models. Despite this finding, none reject the averaging idea—perhaps due to its intuitive appeal. Moreover, they do not explore the reason for this result or explain why one scheme forecasts marginally better than another.

Commercial Vendors

As noted above, while academic researchers propose obtaining consensus ISDs with little noise by averaging together ISDs from a large number of options j, commercial providers normally use just a few ISDs from at-the-money options. For instance, Bloomberg's call and put ISDs are equally weighted averages of the two nearest-the-money calls and two nearest-the-money puts, respectively. Knight Ridder's implied volatilities are an equally weighted average of four individual option ISDs, the two nearest-the-money calls and the two nearest-the-money puts. The CBOE's Market Volatility Index, VIX, uses a weighted average of the same four ISDs, but weights them so that the mean strike equals the underlying asset price.[7] Although the procedures due to Latane and Rendleman (1976), Beckers (1981), and Whaley (1982) put more weight on the ISDs from options that are near-the-money, the Bloomberg, Knight-Ridder, and CBOE models focus on these exclusively. Although the commercial providers do not provide an explicit reason for this choice, one possible

[7]The VIX index also averages together these same four implied volatilities from two different times-to-maturity to maintain a constant average term-to-maturity.

rationale is the belief that ISDs of away-from-the-money options are biased upward. Nonetheless, restricting the set to at-the-money options would appear to ignore a lot of information and forego much of the noise reduction benefits of averaging.

DATA AND MEASURES OF FORECAST ABILITY

Data

Our data consist of Wednesday closing prices of S&P 500 futures and options on S&P 500 futures traded on the Chicago Mercantile Exchange from January 6, 1988 through April 29, 1998.[8] Our data set begins in January 1988 for three reasons. One, trading was light in the first years (1983–1987) of the market so it may not have been as efficient. Two, until serial options were introduced in August 1987, only options maturing in March, June, September, and December were traded. After that date, we have a continuous monthly series. Three, several studies report a sizable, permanent shift in the smile pattern (from a smile to a sneer) after the 1987 market crash.

To calculate the ISDs, we utilize weekly Wednesday settlement price observations on the nearest to maturity options with at least 11 trading days to expiration. We switch from the nearby option to the second nearby on the 10th day before expiration of the former because (1) the time value of very short-lived options relative to bid-ask spreads is small making the ISDs less stable, and (2) to determine how well implied volatilities forecast actual realized volatility we need a sufficient sample of days over which to calculate realized volatility. Data are collected for only one maturity each day.

Each Wednesday we observe both calls and puts at a variety of strike prices. In the S&P 500 options market, these strikes are in the increments of five points, for example, 825, 830, 835, etc. Because trading is light in some far-from-the-money contracts, we restrict our sample to the first eight out-of-the-money and first eight in-the-money contracts

[8]Using options on futures rather than on the underlying index itself has a couple of advantages: (1) both the options and futures markets close at the same time, alleviating the problem of nonsynchronous price quotes, and (2) the option price (and therefore implied volatility) does not depend on assumed dividends. Of course, the S&P 500 index and its nearby futures prices are very highly correlated. As in many other studies, Wednesday was chosen because few holidays fall on Wednesday (in which case we use the Tuesday close).

relative to the underlying S&P 500 futures price. In summary, each Wednesday we collect data on *up to* 16 calls and *up to* 16 puts. However, because not all 32 options trade each day, we do not have 32 observations on all days.

Using Black's (1976) model for options on futures, day t settlement prices for both the option and S&P 500 futures, and 3-month T-bill rates, we solve for the implied standard deviation, $\mathrm{ISD}_{j,t}$, on each of the (32 or less) options j observed on day t. With the exception of the Beckers-Whaley measure, these ISDs are then combined in various weighted or unweighted averages using Equation 1 to obtain a volatility forecast over the remaining life of the option, i.e.,

$$\overline{F}_t = \sum_{j=1}^{J} w_{j,t}\,\mathrm{ISD}_{j,t}.$$

Testing the ability of these models to forecast actual volatility requires a measure of actual realized (or ex post) volatility over this period, RLZ_t, which is calculated as the annualized standard deviation of returns over the period from day $t + 1$ through the option expiration date:

$$\mathrm{RLZ}_t = \sqrt{252 \times \left[\frac{1}{N - t - 1}\sum_{s=t+1}^{N}(R_s - \overline{R})^2\right]} \tag{9}$$

where N is the expiration date of the options observed on day t, $R_s = \ln(P_s/P_{s-1})$, where P_s is the closing futures price on day s, and $\overline{R}$ is the average of the R_t. As noted above, $11 \leq N - t \leq 35$. Because all options j observed on day t expire on the same day, RLZ only has a t subscript.

The Averages

Using all of the (up to 32) individual $\mathrm{ISD}_{j,t}$'s available on each day t we compute volatility forecast averages, $\overline{F}_t$, using each of the major models used in practice or proposed in the literature. These are listed and described in Table I. Two, ISD4 and VIX, are measures calculated and reported by major commercial vendors. Three are measures proposed in the econometrics/forecasting literature. EC1 utilizes weights as defined in Equation 4, i.e., weights based on the error variances only, that is, without the covariances. EC2 utilizes weights obtained by applying the two-stage procedure outlined in Equations 7 and 8, which assumes that

TABLE I
The Volatility Forecast Models

Name	*Description*
ISD4	An equally weighted average of the ISDs of the two closest-to-the-money calls and two closest-to-the-money puts. This is the measure used by Knight Ridder and close to Bloomberg's measures, which report separate averages for the two closest-to-the-money calls and two closest-to-the-money puts.
VIX	A weighted average of the ISDs of the two closest-to-the-money calls and two closest-to-the-money puts where the weights are chosen so that the average strike price equals the underlying futures price. This is essentially the Market Volatility Index distributed by the CBOE except that the time to maturity of VIX matches the horizon of RLZ instead of being adjusted to a constant 1-month horizon.
EC1	A weighted average of the ISDs of all traded options where each ISD's weight is proportional to the reciprocal of the variance of the option's forecast error (Equation 4).
EC2	A weighted average of the ISDs of all traded options where each ISD's weight is proportional to the reciprocal of the variance of the option's forecast error (Equation 7) after removing the error common to all 32 ISDs (Equation 8).
EC3	A weighted average of the ISDs of all traded options (up to 32) where the weights are chosen to minimize the variance of the forecast error (Equation 2) subject to the conditions that all weights are non-negative and that they sum to 1.
LR	A weighted average of all traded options where each ISD_j is weighted by the option's relative vega as proposed by Latané and Rendleman (1976).
CM	A weighted average of all traded options where each ISD is weighted by the option's elasticity with respect to σ as proposed by Chiras and Manaster (1978).
BW	The single ISD that minimizes the sum of the squared deviations of the observed prices of all traded options from their Black–Scholes values as proposed by Whaley (1982) and similar to that proposed by Beckers (1981). Unlike the other measures, BW is not an average of the individual ISDs (Equation 1).
ISD32	An equally weighted average of the ISDs calculated from all traded options with the same expiry.
HIS40	The standard deviation of returns over the last 40 trading days as defined by Canina and Figlewski (1993).
CONSTANT	0.13 (the approximate average of realized volatility over the 1988–1998 period).

the covariance can be represented by a common error. EC3 uses the weights obtained by minimizing the forecast variance in Equation 2 subject to the constraints that the weights sum to one *and* that all are non-negative. We also calculated weights by solving the set of 32 simultaneous equations shown in Equation 3 without imposing the non-negativity constraint. However, when this resulted in several negative weights and several that were greater than 1, we did not proceed further with this average.[9] The forecasts LR, CM, and BW are those proposed by Latane and Rendleman (1976), Chiras and Manaster (1978), and Beckers (1981) and Whaley (1982), respectively.[10] Finally, ISD32 is a simple average of all ISDs (up to 32) observed on day t.

As benchmarks, we also calculate two relatively naive non-ISD based forecasts. HIS40 is the standard deviation of returns over the last 40 trading days. Canina and Figlewski (1993) find that this simple time-series alternative out-forecasts ISD-based measures, although this conclusion has been challenged by Christensen and Prabhala (1998), among others. CONSTANT is a constant value of .13, which is the approximate mean of RLZ_t over the 1988–1998 period.

The fact that all 32 options do not trade each Wednesday necessitates an adjustment to EC1, EC2, EC3, and ISD32 to ensure that the weights sum to 1 each period. If one or more ISDs are not observed in period t, we divide the remaining weights by their sum. For instance, suppose that a particular procedure assigns a weight, w_i, of 0.05 to a single ISD, which is not observed in period t. In this case, $w_{j,t} = w_j/0.95$ for the remaining j ISDs that are observed.

Measures of Forecast Accuracy

We employ three measures of the ability of the various averages to forecast actual volatility: the root-mean-squared forecast error (RMSE), the mean absolute forecast error (MAE), and the mean absolute

[9]Consequently, the model forecast negative variances in some periods, and variances could not be calculated for some periods when all 32 options were not observed. With EC1, EC2, EC3, and ISD32, if all 32 options are not observed on day t we adjust the weights of the remaining options so that they sum to 1.0. For instance, suppose that an option that normally receives a weight of 0.05 does not trade on day t. Then the weights of the remaining 31 are adjusted upward by dividing each by 0.95 or (1.0 − 0.05). If we apply this procedure to a measure without the non-negativity constraints and the weight for an unobserved option exceeds 1.0, then this adjustment cannot be applied. Suppose, for example, that the normal weight of the unobserved option is 3.0, meaning, of course, that the remaining weights sum to −2.0. There is no logical adjustment to the remaining weights, which both preserves their original ranking and ensures they sum to 1.

[10]For the latter, we utilize Whaley's weights.

percentage forecast error (MAPE), where the forecast error is defined as $RLZ_t - \overline{F}_t$.[11] Obviously, RMSE puts more emphasis on large errors than MAE and MAPE.

Initial Results

In Table II, we report RMSE, MAE, and MAPE for all 11 models over the 1988–1998 period.[12] We also rank the models from best (1) to worse (11) by each criterion. The findings reported in Table II are on an ex post basis in that the sample variances and covariances used to obtain the weights for EC1, EC2, and EC3 are calculated over the same 1988–1998 period used to measure forecast accuracy.[13] Although this allows easier analysis of the models, it gives the three econometrics measures an advantage versus the models based solely on information available at time *t*. Later we shall present purely ex ante results.

All three forecast accuracy criteria tell basically the same story. First, the most accurate forecasting procedure by all three measures is EC3, the measure obtained by minimizing the variance of the forecast error (Equation 2) subject only to the non-negativity constraint and the condition that the weights sum to 1. This is not surprising, because the objective on which EC3 is based is to minimize the RMSE and the results in Table II are ex post. Its ex ante forecasting ability could be quite different.[14] Two, among the remaining ISD-based models, the two best are the two commercial models based on the four closest-to-the money options, ISD4 and VIX, with the latter forecasting slightly better. Three, with the exception of EC3, none of the ISD models forecast volatility very well. At the bottom of Table II we report RMSE, MAE,

[11]Specifically,

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\text{RLZ}_t - \overline{F}_t)^2}$$

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|\text{RLZ}_t - \overline{F}_t|$$

and,

$$\text{MAPE} = \frac{1}{T}\sum_{t=1}^{T}\frac{|\text{RLZ}_t - \overline{F}_t|}{\text{RLZ}_t}$$

[12]One of the four options used to calculate ISD4 and ISDVIX was not traded on 11 of the 533 Wednesdays in the 1988–1998 period, so these are excluded from the sample, leaving a final sample of 522 trading days.

[13]Likewise, the value for CONSTANT is based on mean volatility over the entire period.

[14]Also, as described below, EC3 is an average of only four of the options.

TABLE II
Ex-Post Forecast Accuracy of Unadjusted Volatility Forecasting Models

	RMSE		*MAE*		*MAPE*	
Model	*Level*	*Rank*	*Level*	*Rank*	*Level*	*Rank*
ISD4	0.05468	$3^{\dagger\dagger}$	0.04117	5	0.3656	5
VIX	0.05450	$2^{\dagger\dagger}$	0.04100	4	0.3635	3
EC1	0.05962	7	0.04759	8	0.4414	9
EC2	0.05955	6	0.04715	7	0.4351	7
EC3	0.04966	$1^{**\dagger\dagger}$	0.03618	$1^{**\dagger\dagger}$	0.3108	1^{**}
LR	0.06025	9	0.04763	9	0.4401	8
CM	0.06451	11	0.05295	11	0.4986	11
BW	0.05603	$4^{\dagger\dagger}$	0.04286	6	0.3851	6
ISD32	0.06294	10	0.05112	10	0.4791	10
CONSTANT	0.05603	5	0.04057	2	0.3637	4
HIS40	0.05996	8	0.04083	3	0.3264	2

Note. We report three measures of forecast accuracy: the root-mean-squared forecast error (RMSE), the mean absolute forecast error (MAE), and the mean absolute percentage forecast error (MAPE) for each of the 11 volatility forecasting models described in Table I. The models are also ranked from the most accurate forecast (1) to the least (11). All models except BW, HIS40, and CONSTANT are weighted averages of the implied standard deviations calculated for puts and calls with the same expiry but different strike prices. The forecast error is measured as the difference between actual and forecast volatility over the life of the options. Forecast accuracy measures are calculated over the period 1/6/88 through 4/29/98. Forecast accuracy measures for the EC models are ex-post because weights are based on variances and covariances for the 1988–1998 period. An * (†) on a model's RMSE, MAE, or MAPE rank indicates a model whose mean squared error, mean absolute error, or mean absolute percentage error respectively is significantly less than that for CONSTANT (HIS40) at the 0.05 level. ** or †† denote significant differences at the 0.01 level.

and MAPE for the two naive forecasts: CONSTANT and HIS40. Excepting EC3, CONSTANT has the lowest MAE and HIS40 the lowest MAPE among the 11 models examined, and their RMSE is not much higher than that of ISD4 and ISDVIX. Based on these results it would appear (as argued by Canina & Figlewski, 1993) that implied volatility contains little useful information regarding future volatility beyond the information from past volatilities.

UNBIASED FORECASTS

The ISD Bias

The reason that most of the ISD models in Table II forecast worse than the two naive models HIS40 and CONSTANT is because $ISD_{j,t}$ is a biased estimator of RLZ_t for all j. In Table III we report time series means of the $ISD_{j,t}$ over the 1988–1998 period for all 32 options using the following nomenclature to identify calls and puts and their strike prices. The first letter, "C" or "P," indicates a **c**all or a **p**ut, the second,

TABLE III
The ISD Bias by Strike Price

Strike Price	*Mean ISD*	*% of obs > RLZ*	α_0	α_1	*Obs*
Calls					
CI8	0.2199**	94.01**	0.0832**	0.2009**	267
CI7	0.2107**	93.33**	0.0558**	0.3400**	315
CI6	0.2028**	93.28**	0.0329*	0.4712**	357
CI5	0.1904**	91.71**	0.0236	0.5470**	410
CI4	0.1817**	90.44**	0.0249	0.5638**	460
CI3	0.1727**	86.32**	0.0232	0.6108**	497
CI2	0.1637**	83.43**	0.0219	0.6564**	525
CI1	0.1562**	78.53**	0.0239*	0.6789**	531
CO1	0.1491**	74.06**	0.0282*	0.6801**	532
CO2	0.1437**	69.74**	0.0287*	0.7022**	532
CO3	0.1398**	67.99**	0.0282*	0.7244**	531
CO4	0.1380**	66.29**	0.0236*	0.7641**	528
CO5	0.1376**	64.91**	0.0220*	0.7743**	493
CO6	0.1409**	67.15**	0.0207	0.7718*	414
CO7	0.1465**	66.25**	0.0274	0.7266*	320
CO8	0.1517**	66.67**	0.0300	0.7248	231
Puts					
PO8	0.2269**	95.67**	0.0398*	0.3925**	508
PO7	0.2188**	95.07**	0.0356*	0.4286**	527
PO6	0.2074**	94.15**	0.0311	0.4746**	530
PO5	0.1964**	92.67**	0.0307	0.5038**	532
PO4	0.1857**	90.24**	0.0267	0.5560**	533
PO3	0.1752**	87.05**	0.0263	0.5914**	533
PO2	0.1653**	83.46**	0.0269	0.6237**	532
PO1	0.1560**	79.06**	0.0295*	0.6390**	530
PI1	0.1487**	74.29**	0.0305*	0.6618**	525
PI2	0.1430**	70.40**	0.0290**	0.6883**	500
PI3	0.1405**	67.36**	0.0275*	0.7226**	432
PI4	0.1389**	64.72**	0.0262*	0.7497**	343
PI5	0.1441**	67.19**	0.0207	0.7614**	256
PI6	0.1535**	70.41**	0.0167	0.7856*	196
PI7	0.1623**	72.99**	0.0280	0.6672**	137
PI8	0.1771**	74.56**	0.0617*	0.4833**	114

Note. We report time-series means of implied standard deviations, $ISD_{j,t}$ by strike price. We also report estimations of the regression: $RLZ_t = \alpha_{0j} + \alpha_{1j} ISD_{j,t} + u_{j,t}$ where RLZ_t is the realized volatility from day t through the expiration of the options and $ISD_{j,t}$ is the implied standard deviation for option j on day t. Weekly observations from 1/1988–4/1998 on S&P 500 futures options are utilized. In the "Strike Price" column, the first letter, C or P, stands for a **C**all or a **P**ut; the second letter, I or O, refers to **I**n-the-money or **O**ut-of-the-money; and the last digit indicates the position of the option's strike relative to the futures price where 1 indicates that the option is the nearest-to-the-money. * and ** on the mean ISDs designate means that are significantly greater than that for RLZ at the 0.05 and 0.01 levels respectively. * and ** on the regression parameters designate parameters that are significantly different from zero for α_0 or 1.0 for α_1 at the 0.05 and 0.01 levels, respectively.

"I" or "O", indicates whether the option is **i**n or **o**ut of the money, and the last digit, "1" through "8," reports the strike price position relative to the underlying futures price where "1" is the closest to the money and "8" is the furthest in- or out-of-the-money. For example, CI3 indicates an in-the-money call option whose strike price is the third (or 10 to 15 points) below the futures price.

As reported in Table III, the differences in the time series means of $ISD_{j,t}$ by strike price are sizable. The mean ISDs at the top of the smile are over 50% larger than those at the bottom. The null that the means of $ISD_{j,t}$ are the same for all j is easily rejected at the 0.0001 level. Moreover, all exceed the mean of RLZ_t (0.130). The null that the mean of $ISD_{j,t}$ is no greater than that of RLZ_t is rejected at the 0.0001 level for all j except CO8 and PI4. In Table III we also report the percentage of times $ISD_{j,t}$ exceeds RLZ_t. For all 32 options, this percentage exceeds 50%.

From Table III, it is also clear that the reason EC3, ISD4, and VIX forecast better than the other procedures is because the options on which they are based are less biased. EC3 puts all its weight on CO3, CO4, PI2, PI5, and PI7, which are among the least biased of the options in Table III. VIX and ISD4 put all their weight on the four at-the-money options CO1, CI1, PO1, and PI1, which tend to have less bias than most farther-from-the-money options. Nonetheless, the unbiasedness hypothesis is rejected for these at-the-money options as well.

Obtaining Unbiased ISDs and Averages

The econometric analysis reviewed earlier assumed that the individual forecasts on which the averages are based are unbiased. Consequently, the econometricians suggesting these averaging procedures are unanimous in advising that the individual forecasts be adjusted to remove any bias before combining into any average. If the bias is persistent over time, then this can be done based on the historical bias. Fortunately, numerous studies (e.g., Christensen & Prabhala, 1998; Fleming, 1998; Jorion, 1995) have found a fairly consistent linear relation between realized and implied volatility:

$$RLZ_t = \alpha_{0j} + \alpha_{1j}\,ISD_{j,t} + u_{j,t} \tag{10}$$

If an ISD is an unbiased estimator of RLZ_t, $\alpha_{0,j} = 0$ and $\alpha_{1,j} = 1$, but most empirical studies find $\alpha_{0,j} > 0$ and $\alpha_{1,j} < 1$. In Table III we report results of the estimation of Equation 10 for each of the 32 options j over the 1988–1998 period. For all 32 options, $\alpha_{0j} > 0$ and $\alpha_{1j} < 1$. Moreover, the null hypothesis that the ISD is unbiased is rejected for all j.

The regressions in Table III also provide a means of removing the bias. If the bias is persistent and takes the form in Equation 10, then unbiased ISD measures $\mathrm{ISD}^*_{j,t}$ may be derived as:

$$\mathrm{ISD}^*_{j,t} = \hat{\alpha}_{0j} + \hat{\alpha}_{1j}\,\mathrm{ISD}_{j,t} \tag{11}$$

where $\hat{\alpha}_{0j}$ and $\hat{\alpha}_{1j}$ are the parameter estimates from Equation 10. Fleming (1998) finds that $\mathrm{ISD}^*_{j,t}$'s forecast error is well behaved, i.e., orthogonal to determinants of conditional volatility, implying that this linear model does provide an effective forecast.

Using Equation 11 and the parameter estimates from Table III, we derive bias corrected ISDs, $\mathrm{ISD}^*_{j,t}$ for each j, and then use these volatilities, in place of $\mathrm{ISD}_{j,t}$, in Equation 1 to obtain bias-corrected averages. For ISD4, VIX, ISD32, LR, and CM the weights used to calculate the bias corrected averages are unchanged from those used to calculate the raw averages reported in Table II. For EC1, EC2, and EC3, new weights are obtained based on the variances and covariances of the ISD^*_j. Note that such a bias correction is not possible for the BW measure because it is not an ISD average.

Ex Post Results

Ex post forecast accuracy measures for the bias corrected averages are reported in Table IV. For comparison, we also repeat forecast accuracy

TABLE IV
Ex Post Forecast Accuracy of Bias Corrected Volatility Forecasting Models

	RMSE		*MAE*		*MAPE*	
Model	*Level*	*Rank*	*Level*	*Rank*	*Level*	*Rank*
ISD4	0.04729	3*†	0.03151	3*†	0.2612	3*†
VIX	0.04725	2*†	0.03146	2*†	0.2605	2*†
EC1	0.04778	5*†	0.03192	5*†	0.2669	6*†
EC2	0.04783	6*†	0.03198	6*†	0.2668	5*†
EC3	0.04715	1*†	0.03106	1*†	0.2535	1*†
LR	0.04770	4*†	0.03179	4*†	0.2643	4*†
CM	0.04804	8*†	0.03223	8*†	0.2712	8*†
ISD32	0.04788	7*†	0.03204	7*†	0.2685	7*†
CONSTANT	0.05603	9	0.04057	9	0.3637	10
HIS40	0.05996	10	0.04083	10	0.3264	9

Note. We report three measures of forecast accuracy: the root-mean-squared forecast error (RMSE), the mean absolute forecast error (MAE), and the mean absolute percentage forecast error (MAPE) for the eight implied standard deviation, ISD, averages described in Table I after the underlying ISDs are adjusted for their historical bias using the regression results in Table III. We also report results for a simple time series model, HIS40, and a naive CONSTANT of 0.13. The models are also ranked from the most accurate forecast (1) to the least (10). The forecast error is measured as the difference between actual and forecast volatility over the life of the options. Forecast accuracy measures are calculated over the period 1/6/88 through 4/29/98. All are ex-post because the ISD bias adjustment is based on the 1988–1998 period. An * (†) on a model's RMSE, MAE, or MAPE rank indicates a model whose mean squared error, mean absolute error, or mean absolute percentage error respectively is significantly less than that for CONSTANT (HIS40) at the 0.01 level.

measures for the two naive models CONSTANT and HIS40. Removing the bias considerably improves the forecasting ability of all eight ISD averages. For the eight ISD averages as a group, correcting the ISDs for their usual bias reduces the RMSE an average of 17.7%, the MAE 29.4%, and the MAPE 35.4%. For instance, the mean absolute percentage forecast error for the simple average ISD32 is cut almost in half from 47.9% in Table II to 26.9% in Table IV. The improvement in EC3's mean absolute percentage error is more moderate but still substantial, dropping from 31.1 to 25.4%.

Again, all three measures of forecasting ability tell basically the same story, and several results stand out from Table IV. First, after correcting for bias, all eight implied volatility measures forecast actual volatility significantly better than the two naive measures HIS40 and CONSTANT. Although CONSTANT's MAPE is 36.4%, the worst MAPEs for an ISD average is only 27.1% (for the Chiras-Manaster average). For all eight implied volatility measures, the null hypothesis that its RMSE is no lower that of the naive measure CONSTANT is rejected at the 0.01 level. The same result holds for MAE and MAPE, and when we compare the ISD averages with HIS40.

Second, as in Table II, the most accurate average by all three measures is EC3. Third, after EC3, the two most accurate ISD averages are the two commercial measures, ISD4 and VIX, based on the four at-the-money options. The worst among the bias-corrected averages is the Chiras-Manaster average that puts more weight on out-of-the-money options but even it forecasts better than the two naive models (and the Beckers-Whaley measure, which does not lend itself to a bias correction).

Fourth, despite the consistency of the relative rankings, after applying the bias correction, forecasting ability differs relatively little among the eight averages. RMSE varies from a low of 0.04715 for EC3 to a high of 0.04804 for CM—a difference of less than 2%. MAE varies from 0.03106 for EC3 to 0.03223 for CM—a difference of about 3.8%. Clearly, correcting the underlying ISD for bias before averaging is more important than choosing the best weighting scheme. We explore the reasons for this result in the section below.

Ex Ante Results

The forecast accuracy measures for all the averages in Table IV are ex post in that Equation 10, which was estimated over the 1988–1998 period, was then used to correct for bias over the same period. Likewise, in both Tables II and IV, the weights used to calculate EC1, EC2, and EC3 were based on statistics for the same period used to evaluate the models. Obviously, this assumes insight that market participants do not, in fact, have prior to 4/30/1998. Consequently, the question arises whether these results and rankings would be different if the adjusted ISDs and averages were based solely on information available at the time the ISDs were observed. Consequently, we next compute ex ante measures of forecast accuracy reporting the results in Table V.

To obtain ex ante forecasts, we first use data over the 4-year period 1/6/88–12/31/91 to obtain initial estimates of the regression parameters in Equation 10. These are used to calculate bias corrected ISD measures

TABLE V

Measures of Ex Ante Forecast Accuracy of Volatility Forecasting Models

	RMSE		*MAE*		*MAPE*	
Model	*Level*	*Rank*	*Level*	*Rank*	*Level*	*Rank*
Panel A: Bias Corrected Models						
ISD4	0.04192	2**††	0.03171	5**†	0.3156	7**
VIX	0.04188	1**††	0.03168	3**†	0.3152	6**
EC1	0.04266	4**†	0.03160	2**†	0.3084	2**
EC2	0.04281	6**†	0.03185	7**†	0.3122	5**
EC3	0.04199	3**††	0.03146	1**††	0.3048	1**
LR	0.04267	5**†	0.03170	4**†	0.3099	3**
CM	0.04321	8**†	0.03219	8**†	0.3167	8**
ISD32	0.04289	7**†	0.03179	6**†	0.3104	4**
Panel B: Models Without a Bias Correction						
ISD4	0.04514	11**	0.03596	12**	0.3713	11**
VIX	0.04495	9**	0.03582	11**	0.3692	10**
EC1	0.04924	15	0.04065	15	0.4410	15
EC2	0.04836	14*	0.03964	14	0.4249	14*
EC3	0.04496	10**	0.03580	10**	0.3792	12**
LR	0.04998	16	0.04145	17	0.4467	17
CM	0.05324	19	0.04484	19	0.4939	19
BW	0.04637	12**	0.03738	13*	0.3910	13**
ISD32	0.05227	18	0.04391	18	0.4808	18
Panel C: Naive Models						
CONSTANT	0.05189	17	0.04144	16	0.4423	16
HIS40	0.04757	13	0.03497	9	0.3191	9

Note. We report three measures of forecast accuracy: the root-mean-squared forecast error (RMSE), the mean absolute forecast error (MAE), and the mean absolute percentage forecast error (MAPE) for various volatility prediction models. Results for models based on ISDs that have been adjusted to an unbiased basis based on past bias are reported in panel A, while results for unadjusted models are reported in panel B. In panel C, results for a simple time series model, HIS40, and a naive CONSTANT of 0.13 are reported. The models are also ranked from the most accurate forecast (1) to the least (19). Forecast accuracy measures are calculated over the period 1/1992 through 4/1998. All are strictly ex-ante with the models updated every 6 months based on new data. An * (†) on a model's RMSE, MAE, or MAPE rank indicates a model whose mean squared error, mean absolute error, or mean absolute percentage error respectively is significantly less than that for CONSTANT (HIS40) at the 0.05 level. ** or †† denote significance at the 0.01 level.

$ISD^*_{j,t}$ for all 32 j for each Wednesday from 1/2/92 to 6/30/92. We also use the variances and/or covariances of the error terms from these regressions to obtain the weights used to calculate EC1, EC2, and EC3. Using the $ISD^*_{j,t}$ and equation 1, we calculate ISD averages for each model for each Wednesday from 1/2/1992 to 6/30/1992. On 6/30/92, we reestimate the regressions using data from 1/6/1988 through 6/30/1992 and use these to calculate the forecasts each Wednesday from 7/1/1992 to 12/31/1992. We also update the weights for EC1, EC2, and EC3. We proceed in this manner updating the parameter estimates every 6 months through 4/29/98.

The resulting ex ante measures of forecast accuracy, RMSE, MAE, and MAPE, are reported in Table V for the 1/1992–4/1998 period. Results for the bias corrected models are reported in Panel A, results for the models without a bias correction in Panel B, and results for the two naive models: CONSTANT, and HIS40 in Panel C.

As with the ex post results in Tables II and IV, the importance of correcting the ISDs for bias stands out in Table V. All the bias corrected models forecast significantly better than their uncorrected counterparts. Indeed, by all three measures, the worst-performing bias-corrected model forecasts better than the best-performing uncorrected model. On average, adjusting the ISDs based on their past bias reduces MAPE 26%, MAE 20%, and RMSE 12%. These results make clear that the bias patterns are relatively stable over time so that observed bias patterns in the recent past can be used to effectively adjust future forecasts. Supposedly, one reason that the commercial purveyors use at-the-money ISDs only is because they are viewed as less biased. It is clear from comparing the results in Panels A and B that restricting the sample to at-the-money ISDs is not sufficient. The corrected versions of ISD4 and VIX forecast considerably better than their unadjusted counterparts.

According to all three accuracy measures, all the bias-corrected models in Panel A forecast better than CONSTANT and HIS40. Except for MAPE differences between the various averages and HIS40, all differences are significant at at least the 0.05 level. Clearly, the market bases its volatility forecasts on more than recent past volatility, and this information has predictive value. Thus, our results collaborate those of Christensen and Prabhala (1998) and others who find that implied volatilities forecast better than time series models.

Compared with the ex post rankings in Table IV, the ex ante rankings of the bias corrected models in Panel A of Table V are not as consistent across the three measures. The only consistency is that the CM model is the least accurate of the eight. EC3 has the lowest MAE and MAPE of the eight, and its RMSE ranks third. Based on the results in Tables II, IV, and V, if we were forced to choose a "winner," it would be probably be EC3, but this measure has its disadvantages. Specifically, it normally assigns positive weights to only four or five options as we discuss further below.

However, what is more clear is that there is no real winner among the eight models after they are adjusted for bias. As in Table IV, the differences in RMSE, MAE, and MAPE among the eight models in Panel A of Table V is slight. The lowest RMSE is VIX's at 0.04188, while the highest is CM's at 0.04321—a difference of only 3.1%. MAE varies from a low of 0.03146 for EC3 to a high of 0.03219 for CM—a difference of only 2.3%. All the attention paid to the weighting question in the literature seems to be misplaced. It really doesn't matter much. What is important is that the ISDs be adjusted for bias based on their past history.

THE WEIGHTS QUESTION

In both the ex post results in Table IV and the ex ante results in Table V, we observe that after correcting for bias, all of the ISD averages forecast roughly equally well. By most measures EC3 forecasts a little better than the other averages, and by all measures CM forecasts a little worse but the differences are small. The RMSE difference between the eight averages is less than 3%, and the MAE difference is less than 4%.

This leads to the question of whether the various averages forecast roughly equally well on average because their volatility forecasts don't differ much each period or whether they do differ but one average forecasts better in one period and another average in another period. In Table VI we report correlations between the various averages over the 1/1988–4/1998 period. Because the lowest correlation coefficient is

TABLE VI
Correlations Between ISD Averages

	ISD4	*VIX*	*EC1*	*EC2*	*EC3*	*LR*	*CM*	*ISD32*
ISD4	1.0000							
VIX	0.9997	1.0000						
EC1	0.9899	0.9896	1.0000					
EC2	0.9932	0.9926	0.9983	1.0000				
EC3	0.9892	0.9895	0.9857	0.9824	1.0000			
LR	0.9942	0.9938	0.9982	0.9995	0.9848	1.0000		
CM	0.9820	0.9815	0.9980	0.9947	0.9789	0.9939	1.0000	
ISD32	0.9884	0.9879	0.9998	0.9985	0.9829	0.9982	0.9981	1.0000

Note. Correlation coefficients are reported for eight bias corrected ISD averages on a weekly basis for the period 1/1988–4/1998.

0.9789 (between EC3 and CM), it is clear that the various averages forecast roughly equally well because they are produce approximately the same volatility forecast each period.

Certainly the various averages are not coming up with similar forecasts because they weight the various ISDs similarly. In Table VII we report the average weights each procedure placed on the various ISDs

TABLE VII
Option Weightings

Strike Price	*ISD4*	*VIX*	*EC1*	*EC2*	*EC3*	*LR*	*CM*	*ISD32*
			Calls					
CI8			0.0250	0.0051		0.0175	0.0034	0.0313
CI7			0.0253	0.0091		0.0199	0.0036	0.0313
CI6			0.0293	0.0168		0.0230	0.0041	0.0313
CI5			0.0292	0.0289		0.0280	0.0049	0.0313
CI4			0.0302	0.0575		0.0343	0.0060	0.0313
CI3			0.0304	0.0731		0.0420	0.0079	0.0313
CI2			0.0308	0.0562		0.0498	0.0106	0.0313
CI1	0.2500	0.2418	0.0316	0.0438		0.0555	0.0150	0.0313
CO1	0.2500	0.2582	0.0321	0.0402		0.0558	0.0215	0.0313
CO2			0.0323	0.0296		0.0492	0.0310	0.0313
CO3			0.0324	0.0243	0.0282	0.0381	0.0439	0.0313
CO4			0.0359	0.0196	0.0028	0.0272	0.0594	0.0313
CO5			0.0344	0.0201		0.0191	0.0735	0.0313
CO6			0.0349	0.0200		0.0145	0.0831	0.0313
CO7			0.0372	0.0209		0.0125	0.0891	0.0313
CO8			0.0323	0.0192		0.0117	0.0931	0.0313
			Puts					
PO8			0.0269	0.0096		0.0167	0.0776	0.0313
PO7			0.0272	0.0125		0.0199	0.0711	0.0313
PO6			0.0282	0.0191		0.0239	0.0628	0.0313
PO5			0.0289	0.0304		0.0289	0.0547	0.0313
PO4			0.0294	0.0511		0.0352	0.0462	0.0313
PO3			0.0301	0.0791		0.0424	0.0377	0.0313
PO2			0.0306	0.0814		0.0499	0.0293	0.0313
PO1	0.2500	0.2418	0.0315	0.0671		0.0553	0.0215	0.0313
PI1	0.2500	0.2582	0.0319	0.0466		0.0556	0.0149	0.0313
PI2			0.0385	0.0331	0.6582	0.0488	0.0099	0.0313
PI3			0.0337	0.0237		0.0373	0.0066	0.0313
PI4			0.0355	0.0198		0.0266	0.0047	0.0313
PI5			0.0391	0.0176		0.0197	0.0040	0.0313
PI6			0.0304	0.0111		0.0155	0.0033	0.0313
PI7			0.0343	0.0080	0.3108	0.0131	0.0026	0.0313
PI8			0.0207	0.0056		0.0129	0.0028	0.0313

Note. The weight each option's ISD receives in eight weighted average models (defined in Table I) is reported. These are weights for the ex-post models reported in Table IV. For VIX, LR, and CM average weekly weights over the 1/1988–4/1998 period are reported; for the other models, the weights are constant. In the "Strike Price" column, the first letter, C or P, stands for a **C**all or a **P**ut; the second letter, I or O, refers to **I**n-the-money or **O**ut-of-the-money; and the last digit indicates the position of an option relative to the futures price where 1 indicates that the option is the nearest-to-the-money.

over the 1/1988–4/1998 period, that is the weights underlying the results in Table IV. As shown there, the weights differ dramatically. As reported there, ISD4, VIX, and EC3 are all based on only four ISDs, but the four options underlying EC3 are completely different from the four at-the-money options on which ISD4 and VIX are based. Yet the correlation between EC3 and ISD4 is 0.9892, while that between EC3 and VIX is 0.9895. EC1, EC2, and LR tend to put more weight on ISDs with close-to-the-money strikes, while CM puts more weight on deep out-of-the money ISDs; yet the correlation between CM's forecast and any of the other three is 0.99 or greater. Clearly, the conclusion has to be that the weights don't matter very much.

The reason that the weights don't matter much is that there is very little noise in the individual ISDs to be averaged out. This becomes clear if we review the results of the estimation of Equation 8, which was used to obtain EC2. This regression was of the form:

$$E_{j,t} = \sum_{t=1}^{T} \hat{\eta}_t D_t + \hat{\varepsilon}_{j,t}$$

where $E_{j,t}$ is the forecast error, $\mathrm{RLZ}_t - \mathrm{ISD}_{j,t}$, for ISD$j$ in week t, where j varies from 1 to 32 representing each of the 32 different ISDs listed in Tables III and VII, and where t represents the week. This is regressed on a series of 532 zero-one dummy variables, D_t, for each of the weeks from January 1988 through April 1998. The coefficient of D_t, $\hat{\eta}_t$, measures the forecast error common to all of the 32 (or less) ISD_js. The error terms $\hat{\varepsilon}_{j,t}$ then capture the differences between the 32 various ISDs on day t. Because it is common to all 32 ISDs η_t cannot be averaged out. It will be common to all our averaging models, for example, ISD4, ISD32, EC3, etc. On the other hand, the $\varepsilon_{j,t}$ represent the ISD specific errors that should be reduced by averaging. The most efficient averaging procedure would place the most weights on the ISD_js, which tend to have the smallest $\varepsilon_{j,t}$s and the lowest correlation with the other $\varepsilon_{j,t}$s.

The adjusted r-square for this regression is 0.9712, meaning that for our data period over 97% of the squared forecast errors were common to all 32 ISDs and only about 3% was ISD specific. It is therefore clear why RMSE, MAE, and MAPE differ little between the various averaging procedures in Tables IV and V. On average, 97% of the squared forecast error is common to all ISDs on day t so cannot be averaged out no matter which weights are used. This also explains why averages based on just a few options, like ISD4, VIX, and EC3, forecast as well or better than the averages based on all 32 options. Because only 3% of the squared forecast error is ISD specific, it is clear that there is just very little noise to be averaged out. In summary, the weights are largely irrelevant.

CONCLUSIONS

In the options literature, considerable attention has been paid to the question of how implied volatilities from options with the same maturity but different strikes should be combined to obtain a single best forecast of future volatility. Several different weighting procedures have been proposed, and subsets of these have been compared in various horse race-type studies. However, no one has analyzed why one procedure forecasts better than another or when. At the same time but independently, there has evolved a voluminous literature in econometrics on how to combine various forecasts into a single best forecast, and this literature suggests quite different weighting schemes. Finally, commercial vendors such as Bloomberg and the CBOT ignore both literatures and use their own weighting schemes.

We find that, at least in our data set, the question of the best weighting scheme is of minor importance *if* one is working with unbiased forecasts. When applied to unadjusted ISDs, a constrained mean squared error minimization procedure as suggested by the econometrics literature yields the best forecast followed by procedures used by the commercial vendors. However, this is because they place more weight on the least biased ISDs. None of the procedures forecast significantly better than a naive model, which assumes future volatility will be the same as past volatility.

The reason none of the averaging procedures forecast very well is because *all* the ISDs on which they are based, both those in, out, or near the money, are biased predictors of actual volatility. All tend to over predict. Fortunately, it turns out that this bias is pretty stable through time so that estimates of the bias from past periods can be used to adjust the forecasts for future periods with good results. When we correct the underlying ISDs for this bias, all of the averaging procedures forecast significantly better than the naive and time series models.

We further find that once the underlying ISDs are adjusted for bias, the averaging scheme matters very little. This is because there is simply very little ISD specific "noise" to be averaged out, i.e., that each of the ISD yields virtually the same forecast. In other words, most of the differences between the ISDs from different strike prices at time t are permanent, not transitory. When they are adjusted for this systematic difference, the ISDs are quite similar.

We would add a large caveat to this conclusion that the averaging scheme doesn't matter. Our market, that for options on S&P 500 futures, is one of the most actively traded and most liquid in the world. Consequently, bid-ask spreads are reportedly low. The options and the underlying futures trade side by side on the same floor, and both markets close at exactly the same time so nonsynchronous trading is minimized. Because these are options on futures and not the underlying S&P 500 index, arbitrage between the options and the underlying futures is easy. For all of these reasons, individual ISDs in this market would be expected to contain less noise than options from less liquid markets. Consequently, the averaging procedure could matter much more in less liquid markets, such as individual equity options, but we will leave this question to subsequent researchers.

BIBLIOGRAPHY

Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. Operational Research Quarterly, 20, 451–468.

Beckers, S. (1981). Standard deviations implied in option prices as predictors of future stock price variability. Journal of Banking and Finance, 5, 363–381.

Black, F. (1976). The pricing of commodity contracts. Journal of Financial Economics, 3, 167–179.

Canina, L., & Figlewski, S. (1993). The informational content of implied volatility. Review of Financial Studies, 6, 659–681.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. International Journal of Forecasting, 5, 559–583.

Chiras, D., & Manaster, S. (1978). The information content of option prices and a test of market efficiency. Journal of Financial Economics, 10, 213–234.

Christensen, B. J., & Prabhala, N. R. (1998). The relation between implied and realized volatility. Journal of Financial Economics, 50, 125–150.

Feinstein, S. (1989). A theoretical and empirical investigation of the Black-Scholes implied volatility. Unpublished doctoral dissertation, Yale University.

Figlewski, S. (1983). Optimal price forecasting using survey data. Review of Economics and Statistics, 65, 13–21.

Figlewski, S., & Ulrich, T. (1983). Optimal aggregation of money supply forecasts: Accuracy, profitability, and market efficiency. Journal of Finance, 38, 695–710.

Fleming, J., Ostdiek, B., & Whaley, R. (1995). Predicting stock market volatility: A new measure. Journal of Futures Markets, 15, 265–302.

Fleming, J. (1998). The quality of market volatility forecasts implied by S&P 100 index options. Journal of Empirical Finance, 5, 317–345.

Gemmill, G. (1986). The forecasting performance of stock options on the London traded options market. Journal of Business Finance and Accounting, 13, 535–546.

Granger, C. W. (1989). Combining forecasts—Twenty years later. Journal of Forecasting, 8, 167–173.

Jorion, P. (1995). Predicting volatility in the foreign exchange market. Journal of Finance, 50, 507–528.

Latané, H., & Rendleman, R. (1976). Standard deviation of stock price ratios implied by option premia. Journal of Finance, 31, 369–382.

MacDonald, R., & Marsh, I. W. (1994). Combining exchange rate forecasts: What is the optimal consensus measure. Journal of Forecasting, 13, 313–332.

Newbold, P., & Granger, C. W. (1974). Experience with forecasting univariate time series and the combination of forecasts. Journal of the Royal Statistical Society, Series A, 137, 131–149.

Palm, F. C., & Zellner, A. (1992). To combine or not to combine? Issues of combining forecasts. Journal of Forecasting, 11, 687–701.

Turvey, C. G. (1990). Alternative estimates of weighted implied volatilities from soybean and live cattle options. Journal of Futures Markets, 10, 353–356.

Whaley, R. E. (1982). Valuation of American call options on dividend-paying stocks: Empirical tests. Journal of Financial Economics, 10, 29–58.

Winkler, R. L. (1981). Combining probability distribution from dependent information sources. Management Science, 27, 479–488.

Winkler, R. L., & Makridakis, S. (1983). The combination of forecasts. Journal of the Royal Statistical Society, Series A, 146, 150–157.