

# The effect of nonstationarity on combined forecasts \*

Christopher M. Miller

*Jesse H. Jones Graduate School of Business, Rice University, Houston, TX 77251-1892, USA*

Robert T. Clemen

*College of Business Administration, University of Oregon, Eugene, OR 97403-1208, USA*

Robert L. Winkler

*Fuqua School of Business, Duke University, Durham, NC 27706, USA*

**Abstract:** Previous research on the combination of forecasts has, for the most part, implicitly assumed a stationary underlying process so that parameters could be estimated from historical data. While some models weight recent data more heavily in the estimation process in an attempt to provide more accurate parameter estimates in a nonstationary environment, no research to date has specifically examined the effects of nonstationarity on the performance of combining methods. This paper reports the results of a simulation study of the effects of nonstationarity (a shift in the process) on a range of forecast combination methods. Special attention is given to the relative performance of the methods in the time periods around the shift.

**Keywords:** Combining forecasts, Nonstationarity.

## 1. Introduction

Accurate forecasts are crucial to the success of any decision-making strategy. Research has shown that combining forecasts from different sources can improve the accuracy of forecasts [e.g., Newbold and Granger (1974), Makridakis and Winkler (1983)]. This improvement seems reasonable because multiple forecasts may provide different information about the event being forecast. Forecast combination is, in fact, used in practice; Dalrymple (1987) reports that 40% of firms frequently or usually combine sales forecasts in one form or another.

Over the past twenty years, a significant amount of research has been directed toward determining the best method of combining forecasts. Clemen (1989) presents an extensive review of the numerous combining methods that have been proposed and tested using both simulated and real-world data. Although this literature suggests that no single combining method is best in all situations, the simple methods, such as an average of the component forecasts, often perform as well as more statistically sophisticated methods [Figlewski and Urich (1983), Clemen and Winkler (1986), Holden and Peel (1988)].

One explanation for the relatively poor performance of the sophisticated normal model of Bates and Granger (1969), for example, has been that the combining weights can be very unstable [Cle-

\* This research was supported in part by the National Science Foundation under Grants IST-8600788, SES-9022616, and SES-9022573.

men and Winkler (1986), Kang (1986), Winkler and Clemen (in press)]. Using a variant of Akaike's (1974) information criterion, Schmittlein, Kim, and Morrison (1990) provide a way to decide whether to use the full normal model for combining forecasts or some simpler version thereof, thus mitigating the detrimental effects of the weight instability. However, it is important to note that estimation of the covariance matrix in the normal model using past data implicitly assumes that the process is stationary. Even the Schmittlein, Kim, and Morrison procedure relies on an assumption of stationarity. Under nonstationarity, though, past data on forecast accuracy may not reflect the current characteristics of the component forecasts, leading to poor estimates of the combining weights. Any nonstationarity is likely to reduce the accuracy of a combined forecast developed on the basis of the normal model.

Nonstationarity among component forecasts might occur for a variety of reasons. Consider the two following sales forecasting scenarios:

- (1) An unanticipated competitor aggressively enters the market. A sales-forecasting method that may have produced excellent forecasts prior to this event may now yield sales forecasts with different characteristics. In this case, the nonstationarity of the forecast's performance can be attributed to a structural shift in the marketplace.
- (2) The recent performance of a sales-forecasting technique may be deemed unacceptable, leading the forecaster to modify the technique or adopt a new one. Under this scenario, the nonstationarity of the forecast's performance is a result of a change in the forecasting method itself.

In both of these scenarios, the process that generates forecast errors (Forecast-Actual) has shifted. Past data on forecast errors may no longer reflect the current parameters of the process. As such, the past data may become less useful for estimating the parameters needed in more sophisticated forecast combination methods.

This study examines the effect of a structural shift in the process producing forecast errors on the performance of selected combining methods that have appeared in the literature. The next section presents the forecast combination meth-

ods that we consider. The following two sections discuss the simulation methodology and present the results, respectively. The methodology itself involves a single parameter change at a specific time and study of the resulting effects on combination techniques. Because few real-world situations involve single changes, a section entitled 'Extensions' reports tentative results of more realistic simulations involving gradual changes as well as multiple changes. The conclusion summarizes our results.

## 2. Methods for combining forecasts

The combining methods considered here can be categorized based on the amount and type of information utilized about previous forecast errors. The methods include some that disregard all or part of the available information regarding past forecast performance, some that implicitly ignore the possibility of structural shifts by equally weighting all past information, and others that weight the recent past more heavily.

The simplest method, the simple average, does not utilize any past information regarding the precision of the forecasts or the dependence among the forecasts. Let  $f_{it}$  be the  $i$ th forecast in time period  $t$ , and let  $n$  be the number of forecasts to be combined. The simple average is  $f_{ct} = \sum_{i=1}^n f_{it}/n$ . The simple average clearly ignores potentially useful information and assumes forecasts are exchangeable [Clemen and Winkler, (1986)].

The outperformance approach, developed by Bunn (1975, 1978), utilizes some of the information contained in previous forecast errors by calculating the proportion of times that each forecast has 'outperformed' the other forecasts in the sense of having the smallest error. Let  $p_{it}$  represent the proportion of times prior to time  $t$  that forecast  $i$  has the smallest error. The combined forecast  $f_{ct}$  is then calculated using  $p_{1t}, \dots, p_{nt}$  as the combining weights:  $f_{ct} = \sum_{i=1}^n p_{it} f_{it}$ . This approach and its extensions have performed well in empirical settings compared with other methods, especially when few data are available [Bunn (1979), Gupta and Wilton (1988)].

Another weighted average of forecasts bases combining weights on the relative precision of the component forecasts. The weight for forecast  $i$  at

time  $t$  is given by

$$w_{it} = \frac{\hat{\sigma}_{it}^{-2}}{\sum_{i=1}^n \hat{\sigma}_{it}^{-2}},$$

where  $\hat{\sigma}_{it}^2$  is the sample variance estimated from data for time periods 1 to  $t-1$  for the  $i$ th forecast. Combining based on the relative precision of forecasts has been found to perform well under real conditions [Newbold and Granger (1974), Winkler and Makridakis (1983)].

Yet another category of combining methods utilizes past information on the statistical interaction of the individual forecasts as well as their precision. One such model [Bates and Granger (1969), Newbold and Granger (1974), Winkler (1981)] treats the vector of forecast errors  $e_t = (f_{1t} - a_t, \dots, f_{nt} - a_t)'$ , where  $a_t$  is the actual value in period  $t$  and a prime denotes transposition, as normally distributed with zero mean vector and positive definite covariance matrix  $\Sigma$ . The combined forecast from the normal model is

$$f_{ct} = \frac{u' \Sigma^{-1} f_t}{u' \Sigma^{-1} u},$$

where  $f_t = (f_{1t}, \dots, f_{nt})'$ , and  $u$  is a conformable vector of ones. In practice,  $\Sigma$  is estimated on the basis of past data on forecast errors. In some cases this method has performed relatively well [e.g., Agnew (1985)], while in others its performance has been poor [e.g., Clemen and Winkler (1986)].

The final type of combining method considered in this paper modifies the normal model to give more recent data more weight in the estimation of  $\Sigma$  [Bates and Granger (1969), Newbold and Granger (1974)]. The motivation for doing so is to provide a mechanism by which the model can adapt rapidly to structural shifts by reducing the impact of data collected prior to the shift. With differential weighting, the  $(i, j)$ th element of  $\Sigma$  is estimated at time  $t$  by

$$\hat{\sigma}_{ijt}^2 = \frac{\sum_{s=1}^t g(s) e_{is} e_{js}}{\sum_{s=1}^t g(s)},$$

where  $g(s)$  is an increasing function evaluated at time period  $s$ . We examine two weighting functions, a linear weighting function where  $g(s) = s$  and a geometric weighting function where  $g(s) = b^s$ . The geometric constant  $b$  is set to 1.1 based on previous research results [Clemen and Winkler (1986)]. This approach is equivalent to a constrained weighted least squares (WLS) regression approach. Diebold and Pauly (1987) use unconstrained WLS to estimate combining weights in a regression framework.

For all of the combining methods, it is possible to estimate the average error by using past data and then to adjust the component forecasts before combining. We include debiased versions of all combination models except the simple average and outperformance. Thus, a total of ten different combining methods are considered:

- (1) The simple average (A).
- (2) Outperformance (o).
- (3),(4) Two variations of relative precision, one not accounting for bias (RP) and one accounting for bias (RPB).
- (5),(6) Two variations of the normal model, one not accounting for bias (N) and one accounting for bias (NB).
- (7),(8) Two variations of the normal model using a linear weighting function for past observations, one not accounting for bias (NL) and one accounting for bias (NLB).
- (9),(10) Two variations of the normal model using a geometric weighting function for past observations, one not accounting for bias (NG) and one accounting for bias (NGB).

### 3. Methodology

Our study is based on a simulation model for three forecasters with a multinormal random number generator. The generator obtains independent normal variates using the polar method [Knuth (1981)] and then transforms them into multivariate normal forecast errors using the lower triangular method of Scheuer and Stoller (1962). This process is repeated for each of 100 time periods to generate the required sequence of multivariate normal error vectors with the

Table 1  
Parameter values.

	Base values	
	Set 1	Set 2
Variance of 1	1.0	1.0
Variance of 2	0.9	0.7
Variance of 3	1.1	1.4
Correlation of 1 and 2	0.8	0.8
Correlation of 1 and 3	0.6	0.6
Correlation of 2 and 3	0.7	0.7
Bias of 1	0.0	0.0
Bias of 2	0.0	0.0
Bias of 3	0.0	0.0

Structural shift values <sup>a</sup>

Run number	Structural shift description	Base values	Variable	Initial	Change
1	No change	Set 1	–	–	–
2	No change	Set 2	–	–	–
3	Var. increase	Set 1	Variance of 1	1.0	1.7
4	Var. increase	Set 2	Variance of 1	1.0	1.8
5	Var. decrease	Set 1	Variance of 1	1.7	1.0
6	Var. decrease	Set 2	Variance of 1	1.8	1.0
7	Corr. increase	Set 1	Corr. 1 and 2	0.4	0.8
8	Corr. increase	Set 2	Corr. 1 and 2	0.4	0.8
9	Corr. decrease	Set 1	Corr. 1 and 2	0.8	0.4
10	Corr. decrease	Set 2	Corr. 1 and 2	0.8	0.4
11	Bias increase	Set 1	Bias of 1	0.0	1.0
12	Bias increase	Set 2	Bias of 1	0.0	1.0
13	Bias decrease	Set 1	Bias of 1	1.0	0.0
14	Bias decrease	Set 2	Bias of 1	1.0	0.0

<sup>a</sup> The ‘Variable’ column identifies which parameter is involved in the structural shift, ‘Initial’ refers to its value before the shift and ‘Change’ indicates its value after the shift. All other variables stay at the indicated base value before and after the shift.

specified biases, variances, and correlations. Each simulation ‘run’ comprises 3000 iterations. A fixed seed is used to begin each run.

A structural shift is simulated by changing one of the process parameters in time period 30. To examine the effect of different types of structural shifts, we examine increases and decreases in variances, correlations, and biases. Each type of change is studied in two separate simulation runs. These two different runs involve differing degrees of disparity among the forecast error variances. For convenience, we use the terms “parameter set 1” and “parameter set 2” to refer to the two sets of parameters in each case, with parameter set 1 having more homogeneous variances than parameter set 2. Table 1 displays the parameter values used in each simulation run. These param-

eter values reflect the experience of the authors in studying combined forecasts in a variety of real-world applications.

For each time period  $t$ , the data for time periods 1 to  $t-1$  are used to estimate the parameters needed to calculate each method’s combining weights. These weights are then used to find the combined forecast error for each method in time period  $t$ .

To examine each combining method’s bias and precision over time, we calculate the average combined forecast error  $\bar{e}_{ct}$  across iterations for each time period and the root mean squared error  $\text{RMSE}(e_{ct})$  of the combined forecasts for each time period. These measures allow us to capture changes over time for each combining method’s performance.

For another comparison of combining methods over time, in each time period we consider selected pairs of methods and calculate the proportion of iterations for which the first method outperforms the other. Although this does not capture information regarding the magnitudes of differences in errors, it provides insight into the comparative performance of the methods in each time period.

#### 4. Results and discussion

This section presents the results of the simulation runs described above. First, a baseline case with no shifts is discussed. Then the combining methods are investigated for the nonstationary runs. A comparison of the complex combining methods (N, NL, NG, NB, NLB, and NGB) emphasizes the impact of adjusting for bias and giving differential weight to past information in the normal model. Next, the performance of the simple methods (A, O, RP, and RPB) is discussed. Finally, the simple and complex methods are compared.

We present our results in the form of graphs showing the performance of the various combining methods on different criteria over time. The curves on all of the graphs have been smoothed using a spline-fitting algorithm; we are to some extent suppressing the sampling variability due to the simulation in favor of rendering the graphs more readable. The smoothing does not change the nature of the results. To save space we present here only a subset of the many graphs gener-

ated in this study. A complete set is available from the authors on request.

'No change.' In order to establish a base for examining the simulated structural changes, 'no change' simulations are run for two sets of parameter values. RMSE is graphed for each combining method as a function of time (*t*) for the "no change" case with each parameter set in Fig. 1. Because the bias values are all zero in both base cases, results for the bias-adjusted methods are not shown. The RMSEs for the complex combining methods (NG, NL, and N) are higher initially and decrease during the early time periods, while the simple combining methods (A, O, and RP) are relatively constant for all time periods. These differences between the simple and complex methods are due to small sample size effects on the estimation of the larger number of parameters used in the complex combining methods.

By the later time periods, each combining method reaches a fairly constant RMSE, and Fig. 1

provides an indication of the speed with which the RMSEs stabilize. Simulations using parameter set 1, with its highly similar variances, lead to the complex methods having higher RMSEs than the simple methods. For the unweighted normal model (N), this is a sample size effect, illustrating that very large samples are needed to estimate the parameters of N. In contrast, parameter set 2 simulations, with greater heterogeneity in variances, lead to the simple methods having considerably higher RMSEs than N and NL in later time periods. However, across the two sets of parameter values, the geometrically weighted normal model (NG) has a higher RMSE than the linearly weighted normal model (NL), which in turn has a slightly higher RMSE than the unweighted normal model (N). This result is consistent with the reduction in effective sample size with NG and the much smaller reduction with NL. All of the simple methods have similar RMSEs in the later time periods for parameter set 1; RP and O, which can

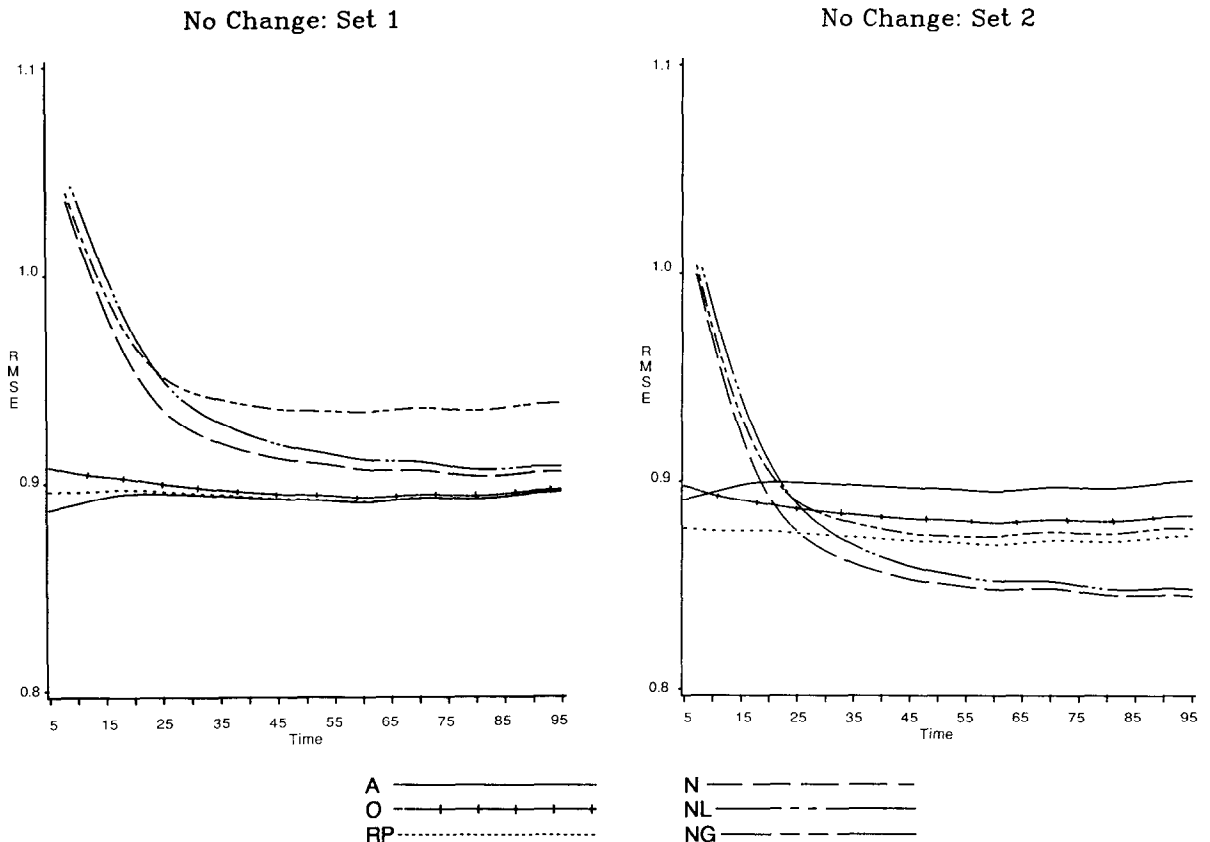


Fig. 1. RMSE for the 'no change' runs.

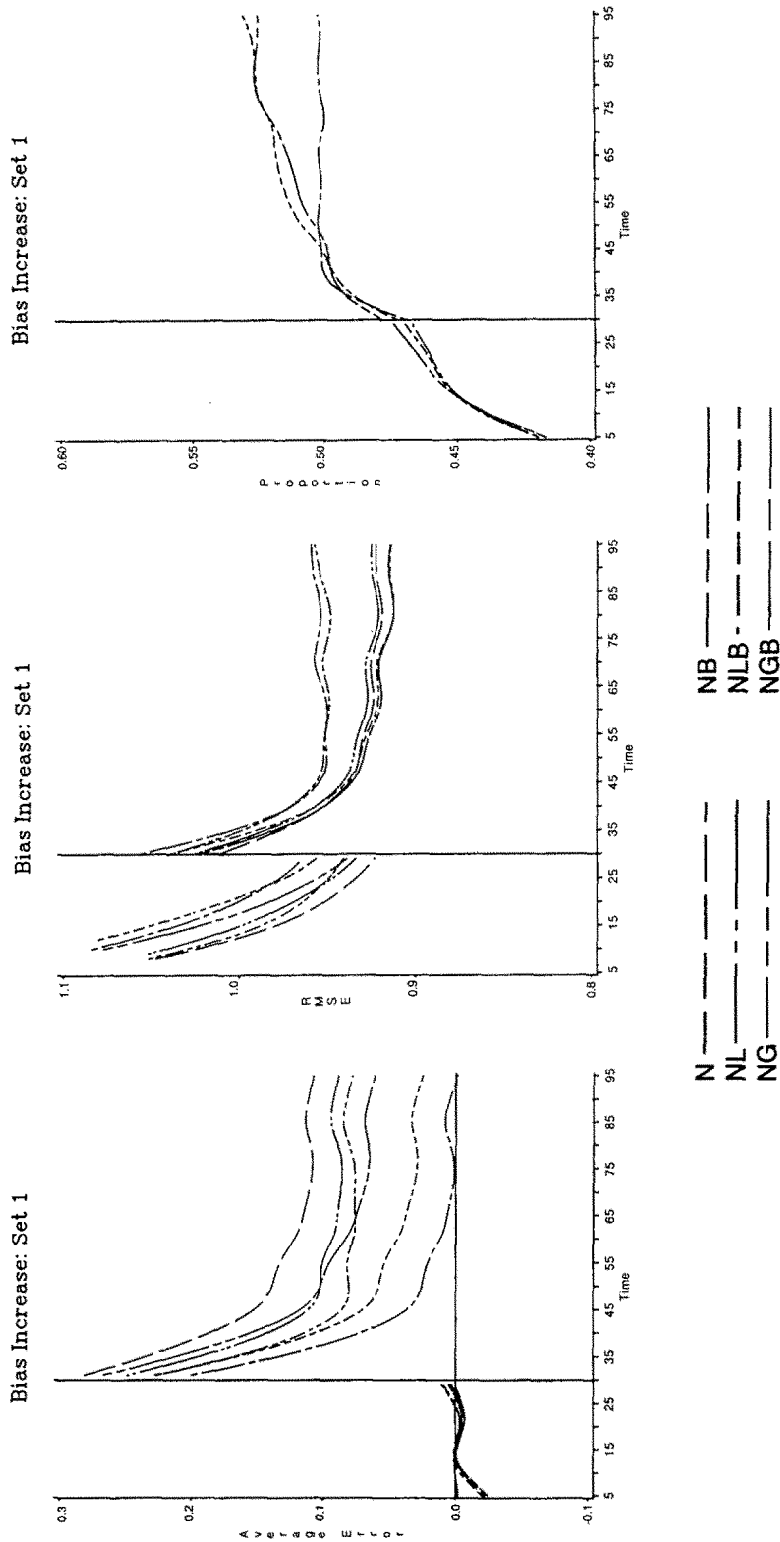


Fig. 2. Average error, RMSE, and proportion of iterations for which each bias-adjusted method beats the corresponding non-bias-adjusted method for run 11 (bias increase, set 1).

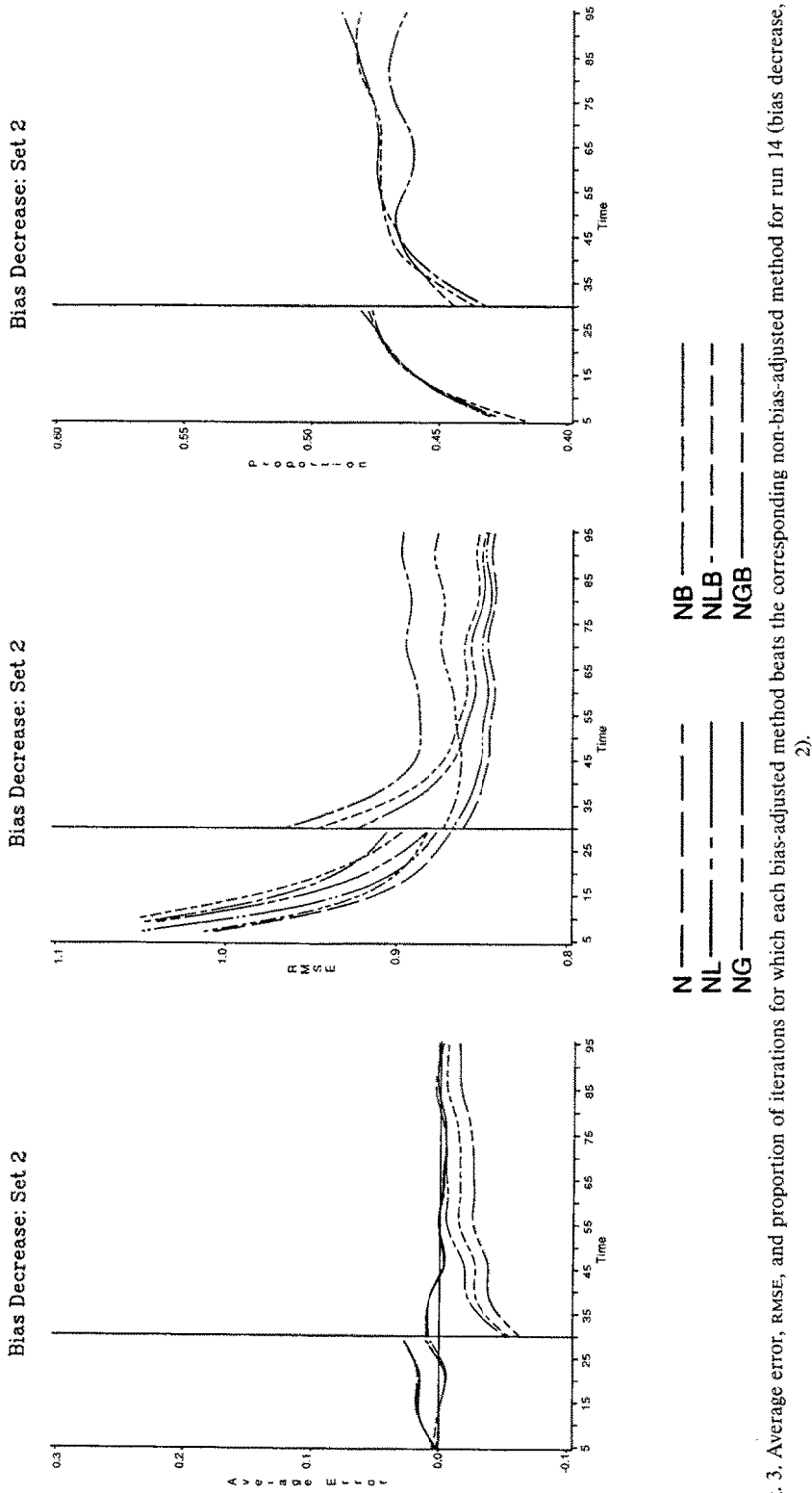


Fig. 3. Average error, RMSE, and proportion of iterations for which each bias-adjusted method beats the corresponding non-bias-adjusted method for run 14 (bias decrease, set 2).

react somewhat to the differences in variances, perform better than A for parameter set 2.

*A comparison of the complex methods.* We call the normal model (N) and its variants (NL, NG, NB, NLB, NGB) the complex methods because they require estimation of considerably more parameters than the other methods. The six complex methods can be differentiated easily along two lines: whether there is an adjustment for bias, and whether past observations on forecast errors are weighted equally or differentially. Our discussion of the performance of these methods will be organized accordingly.

To study the impact of adjusting for bias, we compare the forecast errors from each bias-adjusted method with those from the corresponding non-adjusted forecasts. Furthermore, since the only simulation runs with any non-zero bias are the bias-shift runs, we focus on those cases. Results involving a bias increase and a bias decrease with parameter set 1 are shown in Figs. 2 and 3.

Each figure includes three graphs, which show the average error, the RMSE, and the proportion of times the bias-adjusted method beats the corresponding non-bias-adjusted method.

In Fig. 2, the bias increase causes an immediate jump in average error for all combining methods, with the non-adjusted methods jumping only slightly more than the adjusted methods. The average errors of the bias-adjusted methods adapt more rapidly to the bias increase, but only NGB reaches an average error of zero before the end of the simulation. In RMSE, the bias-adjusted methods do worse than their non-adjusted counterparts immediately after the shift. Over time, the adjusted forecasts improve more rapidly and wind up with lower RMSEs than the non-adjusted forecasts, but the improvements are very slight. As for the pairwise comparisons presented in Fig. 2, the adjusted forecasts beat the non-adjusted forecasts in less than half of the iterations right after the shift; in later periods, NB and NLB beat N

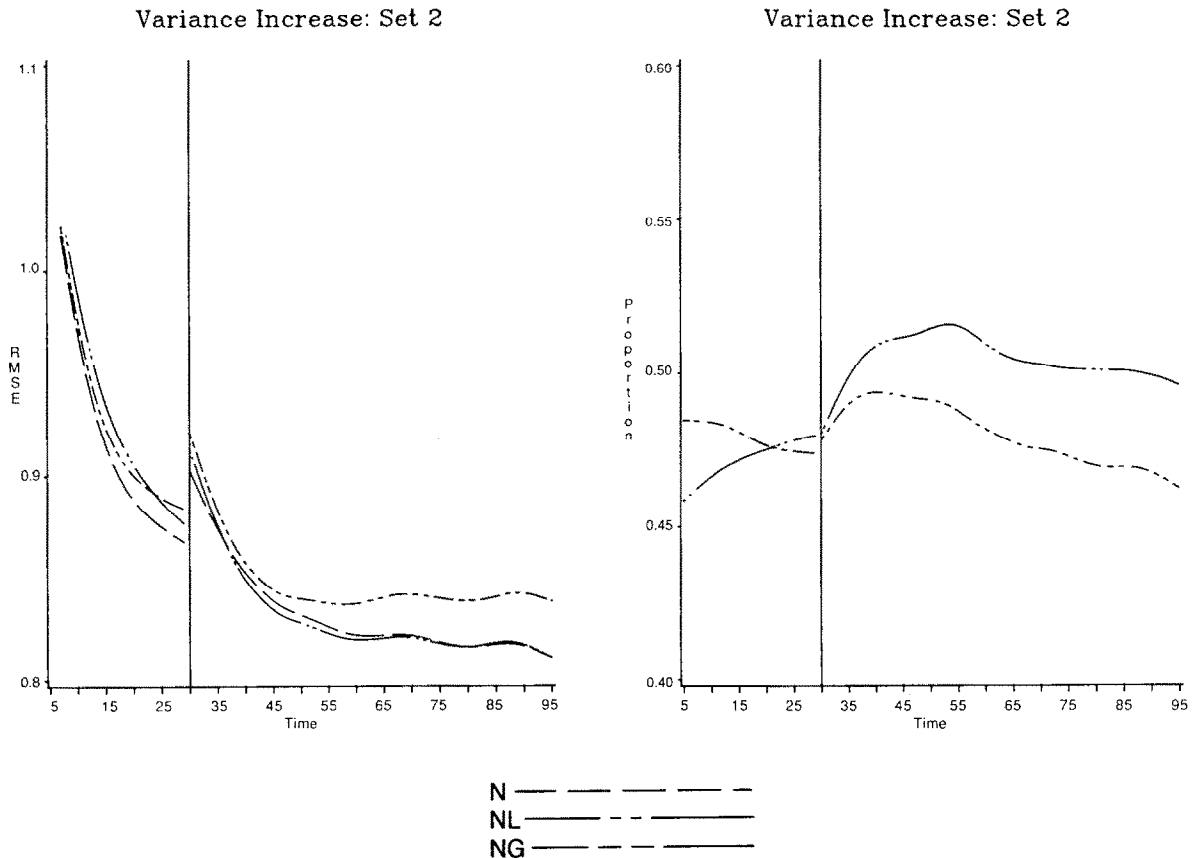


Fig. 4. RMSE and proportion of iterations for which NL and NG beat N for run 4 (variance increase, set 2).

and NL, respectively, on slightly more than half of the iterations.

The overall picture that emerges from Fig. 2 is that any advantage from adjusting for bias is not evidenced soon after the bias shift. The bias-adjustment methods apparently take some time to incorporate and adjust adequately for the shift. Moreover, even in the long run the gains are limited, as can be seen from the RMSE curves in Fig. 2. This is just one case, of course, but it is by far the case most favorable to the bias-adjusted methods. For the bias increase with parameter set 2 (not shown), the average errors are small for all methods and the non-bias-adjusted methods have slightly smaller RMSEs than their bias-adjusted counterparts for all time periods. And for the bias decrease with parameter set 2 illustrated in Fig. 3, the non-adjusted forecasts are substantially better than the adjusted forecasts right after the shift and maintain an advantage through all time periods. As might be expected, the bias-adjusted methods do worse uniformly for the runs without shifts in bias. In summary, although there might be cases (such as the bias increase with parameter set 1) where adjusting for bias can yield some gains, these gains may not be substantial and may only show up in the long run. In many other situations, the bias-adjusted methods appear to be at a disadvantage.

Next, we compare N, NL, and NG to investigate the impact of weighting the past observations when using the normal model. From Fig. 1, we see that in the no-change runs, N consistently has lower RMSE than NL, which in turn is lower than NG. For the nonstationary runs, a consistent pattern emerges, and this pattern is illustrated in Fig. 4 for the variance increase with parameter set 2. Immediately before and after the shift, N has lower RMSE than NL which in turn is lower than NG. By 5–10 periods after the shift, NL has caught and slightly passed N in Fig. 4, and they are quite close for the remaining time periods. NG, on the other hand, stays close (but still behind) for about 15 periods and then levels off while N and NL continue to improve in RMSE.

Although NL is always a bit worse than N right after a shift, the linear scheme for weighting past observations apparently allows it to adapt quickly and catch up to N. The geometric weighting in NG also demonstrates adaptation, but this adaptation is not sufficiently great or long-lived to recom-

mend it over N or NL near the time of a shift, which is exactly when NG might be expected to have an advantage because it so heavily weights recent observations. The great reductions in effective sample size for estimation with NG render it inefficient in a stable environment. In summary, NG is clearly dominated by N in our study. As for N versus NL, N is consistently better under stationarity but NL shows some advantages in terms of more rapid adaptation to shifts.

*A comparison of the simple methods.* The simple methods include the simple average (A), which requires no parameter estimation and ignores past data; outperformance (O), which just considers the proportion of times each method has been best in the past; and relative precision (RP), which estimates the error variances from past data but ignores correlations. RP is the least simple of the methods we label as “simple”, but it utilizes a simple weighting scheme with only  $n$  parameters. The bias-adjusted version of relative precision (RPB) performs poorly relative to RP, as in the case of bias-adjusted versions of the complex methods, and we will ignore RPB in the discussion in order to save space.

From Fig. 1, the simple methods are all very similar in performance with parameter set 1. With parameter set 2, however, RP is consistently better than O, which in turn is better than A. Apparently RP is best able to take advantage of the greater heterogeneity in variances in parameter set 2, while at the other extreme A ignores the heterogeneity entirely. In the more homogeneous case, RP and O yield weights that are close to equal and therefore perform about the same as A.

In the simulation runs with shifts, the relative performances of A, O, and RP before and after the shifts is consistent with the results for the no change scenarios in the sense that the results seem to be dictated by the degree of heterogeneity in the variances. For example, consider the RMSE curves given in Fig. 5 for the variance decrease with parameter set 2. In this situation the variances are heterogeneous before the shift, and we see that RP has the lowest RMSE, followed by O and A. After the shift, the variance decrease leads to lower RMSEs for all three methods, but they are still in the same order because heterogeneity is still present. In contrast, the variance decrease with parameter set 1 (also shown in Fig. 5) exhibits heterogeneous variances before the

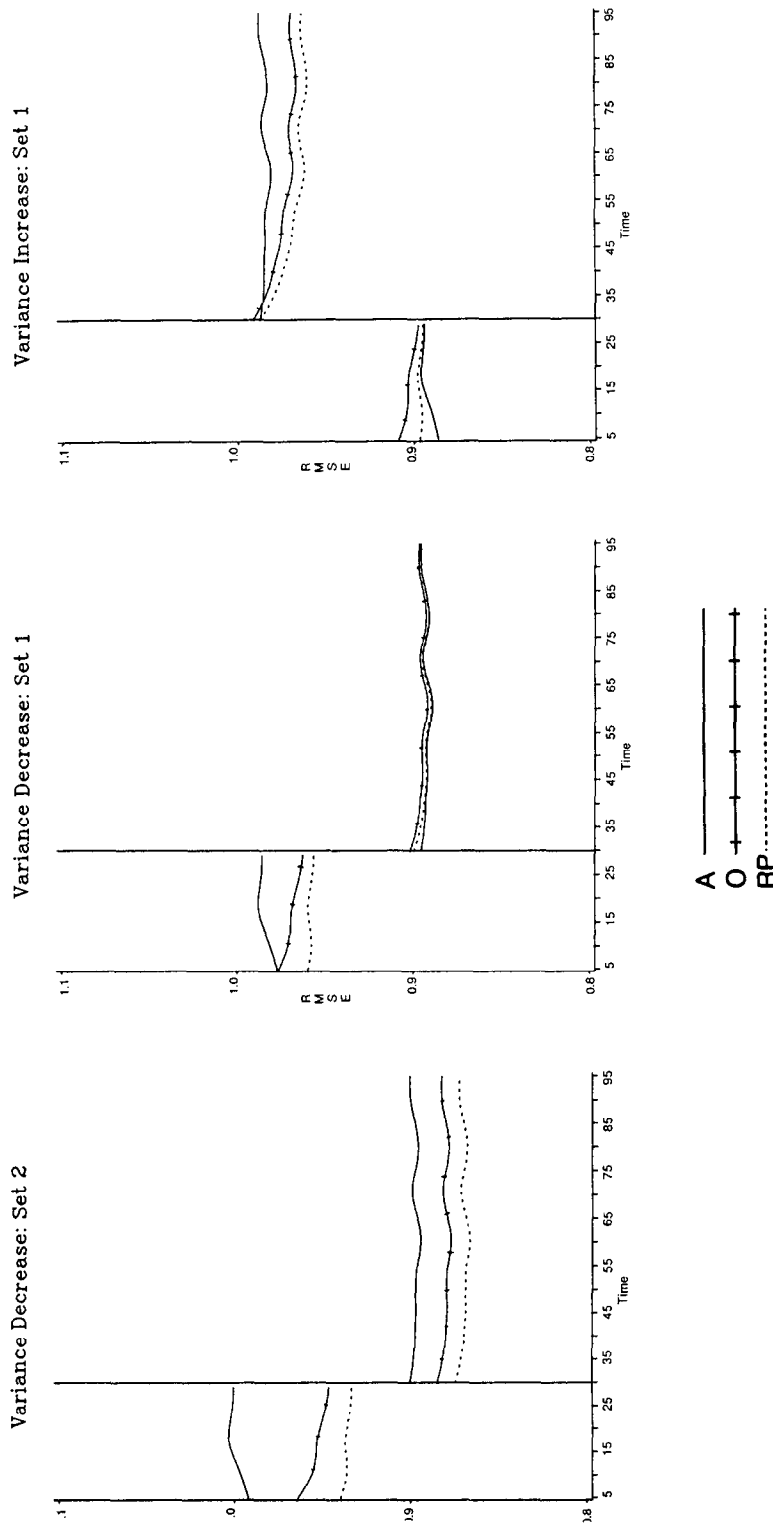


Fig. 5. RMSE for simple methods for selected runs.

shift and homogeneous variances after the shift. As a result, the RMSEs of A, O, and RP are barely distinguishable after the shift. In the final graph in Fig. 5, the variance increase with parameter set 1 is considered. Here the variances are homogeneous before the shift, and the three simple methods are therefore similar in RMSE (A leads at the beginning because of the small samples for the estimation in RP and O). After the shift, the variances are heterogeneous. The three methods

are still close immediately after the shift (with higher RMSEs because of the variance increase), but RP and O adapt to the heterogeneity and improve their RMSEs over time while A remains at the same level.

Among the simple methods, then, RP is better than the others in some situations and about the same in other situations. Thus, it would seem to be the method of choice within this group. In a situation where homogeneous variances are antic-

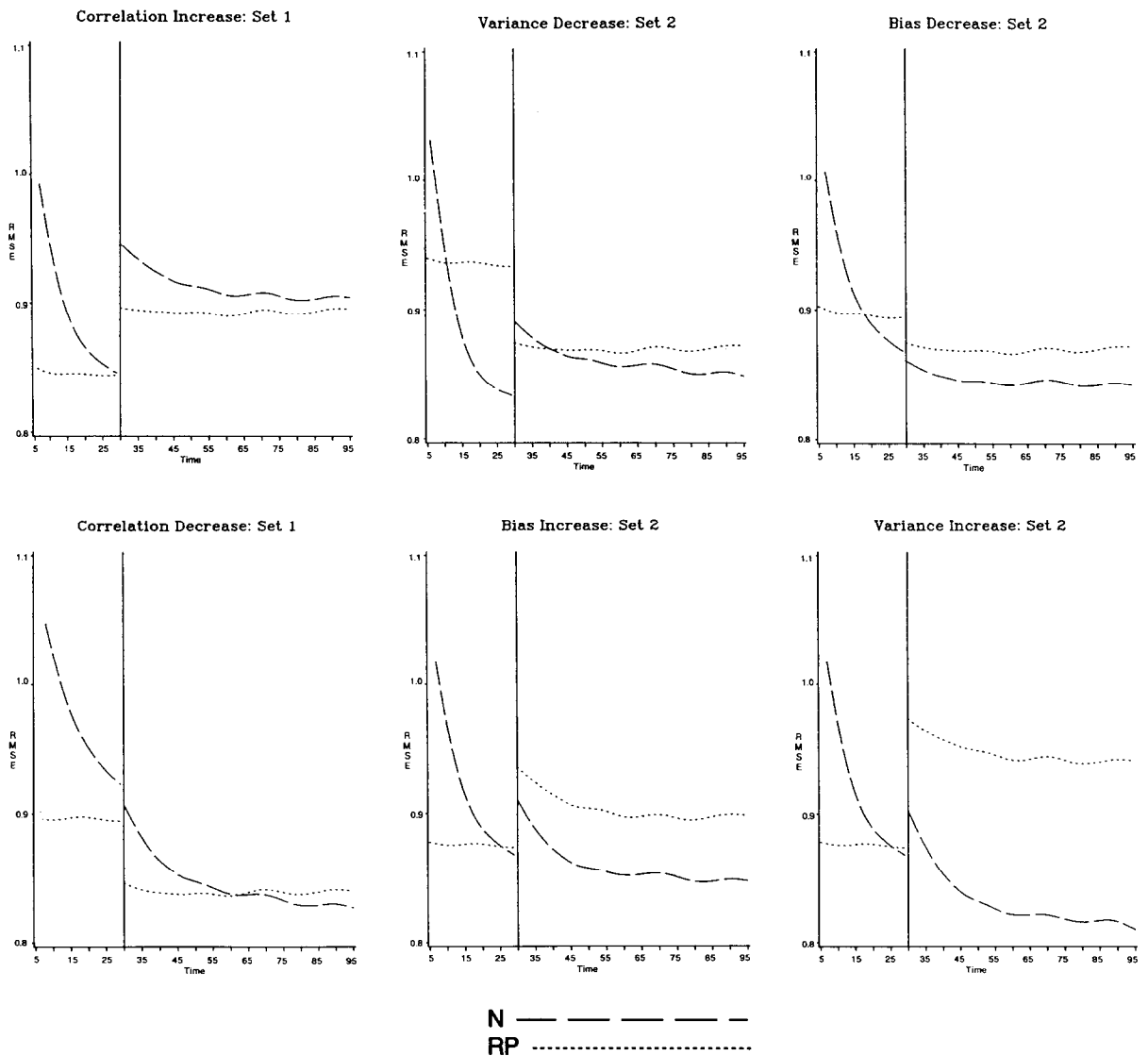


Fig. 6. RMSE for N and RP for selected runs.

ipated, of course, the simple average does well and has the advantage of being the simplest combining method to use.

*A comparison of simple and complex methods.* As mentioned at the beginning of the paper, one of the issues in the combination of forecasts is the relatively strong performance of the simple methods in various empirical studies, and one speculation has been that the poor performance of the complex methods has been due to the presence of nonstationarity in the forecasting process. In this section, we compare the performances of simple and complex methods in our simulation study. Based on the discussion of the results so far,  $N$  and  $N_L$  were preferred among the complex methods and  $RP$  is the method of choice from among the simple ones. In our comparison of simple and complex methods in this section, we focus on comparisons of  $N$  and  $RP$  without losing the essence of the comparison between simple and complex methods. Moreover, comparing only two methods renders the graphs considerably more understandable.

First, consider the relative performances of  $N$  and  $RP$  in the no-change case as shown in Fig. 1. With parameter set 1,  $N$  gradually improves in RMSE through time until it is performing almost as well as  $RP$ , which has roughly a constant RMSE through time. On the other hand, with parameter set 2  $N$  improves enough so that by period 30 it is performing slightly better than  $RP$ . As mentioned above, these results follow from the fact that parameter set 1 includes a highly homogeneous set of variances, whereas the variances are much more heterogeneous with parameter set 2. In the heterogeneous case,  $N$  can improve substantially on  $RP$  by including information about the correlations, but some time is required for  $N$  to “learn” about the correlations. In the homogeneous case, both methods approach roughly equal weights, but  $N$  takes longer to do so.

Fig. 6 shows a selection of six RMSE graphs for several different kinds of shifts. The three graphs on top represent shifts toward less heterogeneous forecasters, while the three on the bottom represent shifts toward more heterogeneity. We select the particular cases that are displayed in Fig. 6 because they illustrate the range of results obtained in comparing  $N$  and  $RP$ .

Taking the cases of decreasing heterogeneity, consider first the correlation increase with pa-

parameter set 1. The results show that the shift results in a greater increase in RMSE for  $N$  than for  $RP$ , but that  $N$  gradually improves over time until its RMSE is close to  $RP$ 's approximately constant RMSE. In the other two decreasing-heterogeneity cases, the same individual patterns are apparent, but the difference between  $N$  and  $RP$  right after the shift is less.  $N$  continues to improve over time so that by period 100 it is noticeably better than  $RP$ .

The three increasing-heterogeneity graphs show somewhat more variability than the other three. In the correlation decrease for parameter set 1,  $N$  has a higher RMSE than  $RP$  right after the shift but catches up and passes  $RP$  in the later time periods. For the other two graphs, after the shift  $N$  has the lower RMSE, and over time the difference between  $N$  and  $RP$  increases as  $N$  adapts to the changes.  $RP$  also adapts in these two graphs but at a much slower rate than  $N$ .

The essential message of these graphs is that in cases of decreasing heterogeneity,  $N$  may be

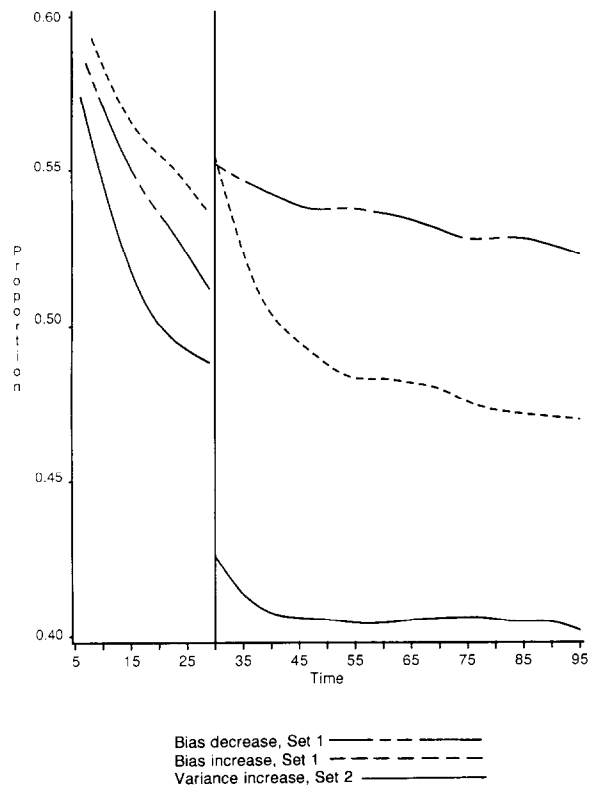


Fig. 7. Proportion of iterations for which  $RP$  beats  $N$  for selected runs.

slightly better or worse than  $RP$ . However, for increasing heterogeneity  $N$  can be much better than  $RP$ , especially after sufficient time has passed for  $N$  to adapt to the change. It is also instructive to consider the performance of the two methods right after the shift. The fact that  $N$  requires some time to adapt to the shift means that either (1)  $N$  begins relatively close to  $RP$  after the shift and then improves relative to  $RP$ , or (2)  $N$  begins above  $RP$  and gradually adapts until the two are approximately equal. The overall effect is that  $RP$  tends to look better relative to  $N$  immediately after the shift than it does in the later periods.

Fig. 7 shows pairwise comparisons between  $N$  and  $RP$ . The curves show the proportion of iterations for which  $RP$  has a smaller error than  $N$  for each of three different structural shifts indicated. In general, these graphs yield similar conclusions to those drawn from the RMSE graphs. For example, the variance increase with parameter set 2 shows that  $RP$  performs poorly relative to  $N$  after the shift. In the bias increase for set 1, also an increasing-heterogeneity situation,  $RP$  performs well immediately after the shift, but  $N$  adapts and overtakes  $RP$ , bringing the curve below 0.50. The third case, a bias decrease with parameter set 1, is a case of decreasing heterogeneity. In this case  $RP$  is better than  $N$  for all time periods. As with the RMSE graphs, in all three cases  $RP$  compares

more favorably with  $N$  immediately after the shift than it does later.

## 5. Extensions

Our methodology described above has allowed us to look at the effects of nonstationarity under “laboratory” conditions: one parameter change at a specific time. However, the real world is rarely so well behaved. To understand better the extent to which our results may generalize to more complex – more realistic – types of nonstationarity, we have run a number of simulations that incorporate simultaneous changes of multiple parameters, gradual changes, and changes of multiple parameters at many different points in time. We have performed a large number of these more complicated simulations, and Table 2 describes five representative runs.

Before presenting the general results obtained from this preliminary study, we must emphasize that the results are tentative. Nonstationarity may occur in many forms, and we do not pretend to have performed an exhaustive study via a careful experimental design. We only hope to give a flavor of the ways in which our previous results may extend to more complicated situations.

As with the previous simulations, the simple

Table 2  
Five representative complicated simulations.

Description	Results (RMSE)
(1) Var(1) gradually increases over 10 time periods. Homogeneity decreases.	$A$ and $RP$ are better before change begins but are worse after changes. For 5–10 periods after the shift, all methods are about the same.
(2) Var(1) gradually decreases over 10 periods. Homogeneity increases.	$RP$ performs best at the beginning. All methods about the same at the time of the shift, but $A$ and $RP$ are much better after.
(3) One-time reduction in variance accompanied by an increase in bias and increased correlation. Homogeneity increases.	RMSE jumps up for all methods at $t = 30$ . Debaised methods perform much better after the shift.
(4) A forecaster substantially improves performance through a one-time reduction in bias and variance. Correlation with other forecasters increases. Homogeneity increases.	$N$ and $NL$ perform best before the shift, but all methods are similar afterward. $A$ and $R$ improve slightly on other methods after shift.
(5) Multiple changes in multiple time periods. In this simulation, one forecaster experiences a relatively gradual change in variance, another has occasional swings, and the third has some large changes. Biases also change periodically.	Large positive spikes occur at times of shifts (especially changes in bias). At the spikes, $A$ tends to jump the least, followed by $RP$ . Overall good performance of $A$ and $RP$ , especially through periods of change.

average and relative precision methods beat the more complex methods in early time periods in the more complicated simulations. In many of the complicated simulations, relative precision performs nearly as well as, or better than, the simple average. Tentatively, it appears that relative precision tends to lose out to the simple average under conditions when changes can be characterized as homogeneity-increasing and relatively dramatic (rapid or large changes).

Within the set of complex methods, adjusting for bias yields gains only in situations of relatively large increases in bias accompanied by a relatively long period of stability. Geometric weighting of past observations in the estimation of the covariance matrix is dominated by linear weighting in virtually all of the more complicated simulations. Linear weighting leads to improvements over the normal model in some situations, but the improvement tends to be meaningful for only a limited number of stable time periods after a change.

Simple methods tend to outperform the more complex methods during and immediately after structural changes when the resulting homogeneity of forecasts is high. Additionally, as the number and magnitude of shifts increase, the simple methods tend to outperform the complex methods because the complex methods have less time to adjust. However, these generalizations comparing simple and complex methods across the complicated simulations should be considered especially tentative because the precise relative performance of the methods depends upon the nature, type, and sequence of structural changes.

## 6. Summary and conclusions

Our results provide some insight into the relative performance of a variety of forecast combination methods over time, with particular interest in the impact of different types of structural shifts in the underlying process generating forecast errors. Specific results vary from run to run, but the important conclusions can be summarized in a few key points.

- As might be expected, the simple methods do considerably better than the complex methods in the early periods, when the amount of evidence available to estimate parameters is very

limited. The complex methods improve rapidly, and in the absence of nonstationarity (the no change runs), their ability eventually to beat the simple methods depends on the nature of the process. The complex methods are able to take advantage of heterogeneity in the variances, but when such heterogeneity is not present, the complex methods do not outpace the simple methods even in the later periods of the simulation.

- Although there might be cases where adjusting for bias can yield some gains, these gains may not be substantial and may only show up in the long run. The bias-adjusted methods apparently take some time to incorporate and adjust adequately for shifts, and in many situations (e.g., small biases, reductions in bias), the bias-adjusted methods appear to be at a disadvantage.

- Linear weighting of past observations in the estimation of the covariance matrix in the normal model allows for rapid adaptation to changes and leads to improvements in some situations (but not others) over the standard normal model. Geometric weighting of past observations, on the other hand, seems to be dominated by linear weighting or no weighting even near the time of a shift, which is exactly when it might be expected to have an advantage, if ever.

- Among the simple methods, relative performance before or after structural shifts in the process depends on the degree of heterogeneity in the variances. Relative precision adapts to heterogeneous variances better than outperformance and, of course, than the simple average, which ignores any information about the process. When variances are relatively homogeneous, the simple methods yield almost identical results.

- In a comparison of simple and complex methods, the results appear to depend on the nature of the process. When the forecasters display a high degree of homogeneity, then simple methods perform well. As homogeneity decreases, the complex methods have an advantage in the long run, but because they require additional adaptation time, the advantage may be absent or considerably diminished immediately after a shift.

Although tentative in nature, the results of the more complicated simulations tend to support the results of the previous simulations in this study. In some ways these results are consistent with past empirical studies. For example, adjustments

for bias and nonlinear weighting of past observations have not been particularly successful in practice. On the other hand, the consistently good performance of the simple average in a variety of studies has not been duplicated here. In real-world applications of combining methods, the forecasting procedures or expert forecasters whose forecasts are being combined are typically chosen because they are expected to be “good”, and within this set we might expect similar error variances. In this case, the simple average should perform well, as it does in our simulations when the forecasters have similar error variances.

## References

- Agnew, C.E., 1985, “Bayesian consensus forecasts of macroeconomic variables”, *Journal of Forecasting*, 4, 363–376.
- Akaike, H., 1974, “New look at the statistical model identification”, *IEEE Transactions on Automatic Control*, 19, 716–723.
- Bates, J.M. and C.W.J. Granger, 1969, “The combination of forecasts”, *Operational Research Quarterly*, 20, 451–468.
- Bunn, D.W., 1975, “A Bayesian approach to the linear combination of forecasts”, *Operational Research Quarterly*, 26, 325–329.
- Bunn, D.W., 1978, *The Synthesis of Forecasting Models in Decision Analysis* (Birkhauser, Basel).
- Clemen, R.T., 1989, “Combining forecasts: A review and annotated bibliography”, *International Journal of Forecasting*, 5, 559–583.
- Clemen, R.T. and R.L. Winkler, 1986, “Combining economic forecasts”, *Journal of Business and Economic Statistics*, 4, 39–46.
- Dalrymple, D.J., 1987, “Sales forecasting practices”, *International Journal of Forecasting*, 3, 379–391.
- Diebold, F.X. and P. Pauly, 1987, “Structural change and the combination of forecasts”, *Journal of Forecasting*, 6, 21–40.
- Figlewski, S. and T. Urich, 1983, “Optimal aggregation of money supply forecasts: Accuracy, profitability and market efficiency”, *Journal of Finance*, 28, 249–258.
- Gupta, S. and P.C. Wilton, 1988, “Combination of economic forecasts: An odds-matrix approach”, *Journal of Business and Economic Statistics*, 6, 373–379.
- Holden, K. and D.A. Peel, 1988, “The accuracy of forecasts of the UK economy”, *Prévision et Analyse Économique*, 7(3), 35–52.
- Kang, H., 1986, “Unstable weights in the combination of forecasts”, *Management Science*, 32, 683–695.
- Knuth, D.E., 1981, *The Art of Computer Programming: Volume 2 / Semi-numerical Algorithms*, 2nd ed. (Addison-Wesley, Reading, MA).
- Makridakis, S. and R.L. Winkler, 1983, “Averages of forecasts: Some empirical results”, *Management Science*, 987–996.
- Newbold, P. and C.W.J. Granger, 1974, “Experience with forecasting univariate time series and the combination of forecasts”, *Journal of the Royal Statistical Society A*, 137, 131–164.
- Scheuer, E. and D.S. Stoller, 1962, “On the generation of normal random vectors”, *Technometrics*, 4, 278–281.
- Schmittlein, D.C., J. Kim and D.G. Morrison, 1990, “Combining forecasts: Operational adjustments to theoretically optimal rules”, *Management Science*, 36, 1044–1056.
- Winkler, R.L., 1981, “Combining probability distributions from dependent information sources”, *Management Science*, 27, 479–488.
- Winkler, R.L. and R.T. Clemen, in press, “Sensitivity of weights in combining forecasts”, *Operations Research*.
- Winkler, R.L. and S. Makridakis, 1983, “The combination of forecasts”, *Journal of the Royal Statistical Society A*, 146, 150–157.

**Biographies:** Christopher M. MILLER is Assistant Professor of Administrative Science at the Jesse H. Jones Graduate School of Administration at Rice University. He received his B.A. in Economics and M.S. in Applied Economics from the University of California at Santa Cruz, and his Ph.D. in Marketing at the University of Oregon. His current research interests include combining forecasts, managerial decision making, and interdependent demand models. He is a member of TIMS and AMA.

Robert T. CLEMEN holds a Ph.D. in Business from Indiana University and is Associate Professor of Decision Sciences and Robert C. Braddock Research Scholar at the University of Oregon. He has also held a visiting position at the Fuqua School at Duke University. His research interests include decision analysis, decision theory, and forecasting, especially the use and aggregation of expert information. His articles have appeared in a wide variety of scholarly journals, including *Management Science*, *Operations Research*, *Journal of Forecasting*, and *International Journal of Forecasting*.

Robert L. WINKLER is James B. Duke Professor in the Fuqua School of Business and the Institute of Statistics and Decision Sciences at Duke University, Durham, NC 27706, where he is also serving as Senior Associate Dean for Faculty and Research in the Fuqua School. He received a B.S. from the University of Illinois and a Ph.D. from the University of Chicago, was at Indiana University prior to moving to Duke, and has held visiting positions at the University of Washington, the International Institute for Applied Systems Analysis, Stanford University, the National Center for Atmospheric Research, and INSEAD. His primary research interests include Bayesian inference, decision analysis, probability forecasting, combining forecasts, and risk assessment.