Bank of Japan Working Paper Series

Forecast Selection by Conditional Predictive Ability Tests: An Application to the Yen/Dollar Exchange Rate

Kei Kawakami*
kei@ucla.edu

No.08-E-1
January 2008

Bank of Japan
2-1-1 Nihonbashi Hongoku-cho, Chuo-ku, Tokyo 103-8660

* UCLA (formerly Monetary Affairs Department, Bank of Japan)

Papers in the Bank of Japan Working Paper Series are circulated in order to stimulate discussion and comments. Views expressed are those of authors and do not necessarily reflect those of the Bank.
If you have any comment or question on the working paper series, please contact each author.
When making a copy or reproduction of the content for commercial purposes, please contact the Public Relations Department (webmaster@info.boj.or.jp) at the Bank in advance to request permission. When making a copy or reproduction, the source, Bank of Japan Working Paper Series, should explicitly be credited.

FORECAST SELECTION BY CONDITIONAL PREDICTIVE ABILITY TESTS: AN APPLICATION TO THE YEN/DOLLAR EXCHANGE RATE♦

KEI KAWAKAMI*

【ABSTRACT】

In this paper, I propose a new method for forecast selection from a pool of many forecasts. My method has two features. The first is the use of the conditional predictive ability test proposed by Giacomini and White [2006]. Second, I construct a measure with two dimensions: "relative usefulness" and "signal predictability". The measure is designed to rank many forecasts in the order of ex-ante forecast accuracy. Therefore, the ranking can be useful not only for selection of a single forecast but also for forecast combinations. I apply the method to the monthly yen/dollar exchange rate. First, I evaluate the performance of base-line forecasting models including a forecast survey of Japanese companies. Second, I show empirically that my method of switching forecasting models reduces forecast errors compared with a single model.

♦ I would like to thank Raffaella Giacomini, Mototsugu Shintani, Tomoyoshi Yabu and the staff of the Bank of Japan for their helpful comments. The opinions expressed here, as well as any remaining errors, belong to the author and should not be ascribed to the Bank of Japan or the Monetary Affairs Department.

* UCLA, E-mail: kei@ucla.edu (formerly Monetary Affairs Department, Bank of Japan)

1. INTRODUCTION

Forecasters often face a problem of selecting forecasts. Although economic or econometric theories provide guidelines for relevant variables and specifications for forecasting, forecasters have discretion in forecast selection. I propose a new method for forecast selection from a pool of many forecasts and apply it to the monthly yen/dollar exchange rate. The Japan Center for International Finance (JCIF) survey and 28 model-based forecasts are used as primary forecasts, from which I select a single forecast (or multiple forecasts for forecast combinations) every period.

There is a vast body of literature on forecast selection. Diebold and Mariano [1995] propose a pairwise comparison method. West [1996] and Clark and McCracken [2001] generalize this framework to incorporate parameter uncertainty and comparison of nested models. Giacomoni and White [2006] further generalized it to allow for the use of conditional information. White [2000] and Hansen [2005] study methods of comparing more than two models for forecast accuracy. My main contribution is to construct ex-ante forecast selection criteria from many forecasts based on the conditional predictive ability test proposed by Giacomini and White. In their original paper, they propose a decision rule for selecting between two forecasting methods. My method makes it possible to use conditional information for forecast selection in the situation with more than two forecasts. In a similar vein, the forecast combination literature (see Timmermann [2006]) recommends ex-ante estimation of optimal weights for multiple forecasts, but my method is conceptually different. One merit of forecast combinations is that forecast errors cancel out (see Clements and Hendry [1998] and Kitamura and Koike [2002]). The motivation of forecast combinations is, therefore, analogous to that of basic portfolio theory, which suggests diversification (e.g. "Do not put all eggs in one basket," or more simply "Use many baskets"). The motivation for my forecast selection method is simple: "Pick a suitable basket every time you go". Although the underlying principle of my method is different, it can still be complementary to forecast combinations. Since my method ranks forecasts in the order of ex-ante forecast accuracy, it allows forecasters to combine forecasts based on the ranking.

With respect to the exchange rate forecasting, Meese and Rogoff [1983] find that

model-based forecasts are no better than a random-walk forecast in the short run out-of-sample forecasting. This finding has been corroborated by more recent research, such as work by Engel, Mark and West [2007]. My empirical results are consistent with the preceding literature in that beating a random walk by a single model is difficult, but also suggest that switching models based on conditional information might outperform a random walk forecast at least with a one-month forecast horizon.

The paper is constructed as follows. Section 2 introduces 29 primary forecasts including the JCIF survey. I overview the survey and explain other model-based forecasts. Their out-of-sample performances are evaluated in the period from May 2000 to March 2007. Section 3 explains my method of constructing forecast rankings using the conditional predictive ability test by Giacomini and White. In Section 4, I apply the method to a pool of primary forecasts. I examine the out-of-sample performances of the forecast based on my method and find that it reduces MSE (mean squared errors). Three robustness checks are added at the end of the section. Finally, Section 5 provides some conclusions.

2. PRIMARY FORECASTS

I introduce 29 primary forecasts from which I select one or multiple forecasts. Since the focus of the paper is to provide a rule of model selection, not to construct a single forecasting model, I do not pay much attention to variable choices, model specifications, data definitions, etc. If forecasters are not sure about any forecasting specifications, I suggest expanding a pool of primary forecasts by adding forecast sequences made by different specifications. First, I overview my 29 primary forecasts, and then evaluate their out-of-sample performances.

(1) The benchmark forecast

The simplest forecast is a forecast by the latest realized target value. I call this the RW (random walk) forecast. RW is used as a benchmark in this paper, because it is the easiest to construct, frequently used in practice and in the literature, and it has an

optimal property under certain conditions.[1]

(2) The JCIF survey

The survey covers the period from May 29, 1985 to April 26, 2007. It is conducted twice a month, once in the middle and again at the end of each month, the latter of which I use as a monthly forecast series.[2] I focus on the forecast horizon of one month; hence 264 time series forecasts are available. The respondents are categorized into four industries: (1) banks and brokers; (2) securities and trading companies; (3) export-oriented companies; and (4) life insurance and import-oriented companies. The number of respondents is time-varying. It was almost fixed at 44 but since 2001 the number has been decreasing with fluctuation and was 30 on April 26, 2007. I use the sample means of all the respondents as a forecast sequence and call it JC (Japanese companies) in the subsequent analysis. For further descriptions of the survey, see Ito [1990] and Hara and Kamada [1999, 2002].

(3) Model-based forecasts

I construct simple linear regression based forecasts. For all models, I use a rolling estimation scheme with an estimation window of size 120 months prior to the forecasting date. The first model is the autoregression (AR) forecast, where the lag length is chosen between 0 and 4 by the BIC criteria.[3] The other models include both the AR term and other variables. Lag lengths are chosen similarly, but separately for the AR term between 0 and 4 and for others between 1 and 4. Three kinds of variables are used: annual inflation rates; short-term interest rates; and trade statistics. For each variable, I construct a forecast with Japanese data only, with US data only, and with the difference of the two series. For example, there are three forecast sequences using inflation rates: a forecast with Japanese inflation (P1); a forecast with US inflation (P2); and a forecast with inflation differential (P3). Similarly, there are three forecast

[1] If a loss function is quadratic and a forecast target follows a random walk process, a forecast by the latest realization minimizes expected loss.

[2] The survey is usually conducted on the Tuesday two weeks before the final Tuesday and on the final Tuesday of each month. However, it skips the middle of August since 1989 and the end of December since 1991. Hence, strictly speaking, the forecast series I employ is not an end-of-month to end-of-month forecast.

[3] Lag 0 means forecasting by sample mean of the target.

sequences for interest rates (I1, I2, I3). Three forecast sequences are made from Japanese trade data: a forecast with Japanese exports (B1); a forecast with Japanese imports (B2); and a forecast with the Japanese trade balance (B3). Similarly, I construct three forecast sequences using trade data between Japan and the US (B4, B5, B6), and three more forecast sequences with US trade data (B7, B8, B9). Also, I construct regression models using two or three variables selected from above. For example, model PB1 includes the AR term, the inflation differential, and the Japanese trade balance as predictors. See Table 1 for the list of primary forecasts and Appendix for data sources. I do not claim that my list of forecasting models is complete; the use of only 29 forecast series is solely for presentational purpose and expanding the forecasting pool is not a problem for my method, as long as it is computationally manageable and data are available.

(4) Evaluation of primary forecasts

All primary forecasts span the period from May 1985 to April 2007. However, I evaluate them in the shorter period from May 2000 to March 2007.[4] For my forecast selection method, a large number of forecast results for each forecast series are required. That is, results of primary forecasts in the period from May 1985 to April 2000 are used as "inputs" for my forecast selection method from May 2000. Since I compare the performance of my method to that of each primary forecast, I set the same evaluation period for all forecast series. Now I investigate the performances of primary forecasts with special attention to RW and JC. Figure 1 shows the range of primary forecasts, the realized exchange rate, and JC. Forecast JC spans the upper and lower bound of the range most frequently among primary forecasts. It is worth noting that JC is neither always radical nor always conservative. For example, JC is the slowest to catch up with the actual rate in the yen appreciation episode in 2002, while it predicts the most radical yen appreciation in 2003. These changes in attitudes between being conservative and radical seem to be responsible for the larger swings of JC compared to other model-based forecasts.

Next, in order to assess which primary forecasts are relatively closer to a realization at each forecasting date, I introduce a primary ranking defined as the order

[4] I can not evaluate April 2007 forecasts since I do not have realization for May 2007.

of absolute size of forecast error. For example, if JC in May 2000 is primary rank 1st, it is closest to the realization of the June 2000 exchange rate in absolute terms among primary forecasts. The primary rankings are calculated for 83 months, for each of which there exist 1st, 2nd, …, 29th forecasts. Table 2 lists the frequency of primary ranks and it shows a sharp contrast between RW and JC. Looking at the frequency of the 1st rank, JC is most frequently ranked together with B2 and IB3. Looking at the frequency of the 29th rank, JC is again most frequently ranked. Thus, JC is both "the most frequently 1st ranked" and "the most frequently worst ranked" forecast. On the contrary, RW shows up at right top, right bottom, and left middle in the table. This means that RW is rarely (actually never) ranked either 1st or worst, and it is most frequently ranked in the middle ranks. The contrast between these two forecasts in terms of primary rankings appears to show a certain trade-off when using a single forecasting model. Namely, a forecasting model which sometimes yields forecasts with pinpoint accuracy yields large forecast errors at other times, while a forecasting model which completely avoids large forecast errors sacrifices pinpoint accuracy. No single forecasting model can have both features. Figures 2 and 3 confirm this trade-off from a more general perspective. Figure 2 is histograms of primary rankings for "the most frequently 1st-ranked forecasts" (JC, B2, IB3). It shows an M shape, which means that these forecasts move wildly between the top and bottom of primary rankings. Figure 3 is the same histograms for "the least frequently worst-ranked forecasts" (RW, AR, B4).[5] They show a bell shape with short tails, which means that these forecasts are concentrated in the middle of primary rankings. Table 3 shows MSEs for all primary forecasts. MSEs for both types of models shown by bold characters are not especially small.[6]

These empirical findings are the basis of my forecast selection method. As long as

[5] Since there are many forecasts with frequency zero of the worst primary rank, I picked models with the smallest frequency of the bottom three primary ranks among them (RW, AR, B4, B6, B8) and further narrowed models with the smallest frequency of bottom ten primary ranks.

[6] Readers who see Table 3 might think that I should simply use I3, the model with the smallest MSE. However, since MSE is unconditional evaluation of the model, it is not necessarily a good criterion from conditional perspective. See Timmermann [2006], Fujiwara and Koga [2002] for support on this point, and also Section 4 where I investigate MSE-based ranking.

I stick to a single model, I cannot escape the trade-off shown above. However, given the large amount of data showing the performance of each forecasting model, it might be possible to switch models when appropriate. If that is possible, I can overcome the trade-off caused by sticking to a single model and can expect smaller forecast losses. The next section explains my selection method in detail.

3. A FORECAST SELECTION METHOD BY GIACOMINI-WHITE STATISTICS

In this section, I explain how to construct ex-ante forecast rankings for primary forecasts in the order of forecast loss size. This involves three steps: (1) forecast relative usefulness; (2) check signal predictability; and (3) construct a ranking measure using both (1) and (2). I explain each step in separate subsections.

(1) Relative usefulness

I start by defining "relative usefulness" of forecast *i*. This is given by the loss differential between a benchmark forecast RW and forecast *i*:

$$A_{i,t} = \left(e_{rw,t|t-1}\right)^2 - \left(e_{i,t|t-1}\right)^2, \quad (1)$$

where i = 1, …, 28 denotes primary forecasts except RW, and $e_{i,t|t-1}$ is an error of forecast *i* made at time *t*-1, which is available to a forecaster at time *t*. I assume that the loss function is quadratic but it can be replaced by other loss functions depending on the goal of forecast. Thus defined, the positive value of relative usefulness suggests that I should have used forecast *i* instead of RW at time *t*-1. On the contrary, the negative value suggests the opposite. Also, $A_{i,t} > A_{j,t}$ suggests the relative usefulness of forecast *i* over *j*.

Now, I need to know one-period ahead relative usefulness, but not the past realization. I employ two signals to forecast $A_{i,t+1}$. The first is the past realized relative usefulness. The second is mean deviation of forecast target, where the mean is that of the most recent 12 months' data in the estimation window.[7] I expect that the first signal will capture systematic mistakes that some forecasting models typically, if not always, make. In that case, knowing past relative usefulness will help predict future

[7] Since the mean is time varying, the second signal is revised for every forecasting date.

relative usefulness. The second signal tries to capture structural changes in exchange rate formation in a market. Typically for asset markets, "market sentiment" or "story" drives the markets in the short run. Stories might focus on the US trade deficit sometimes and the interest rate differential at other times. If changes of ruling stories affect the data generating process of the target, the signal defined by mean deviation can be helpful for identifying them. Given these two signals, I conduct the following regression-based forecast of $A_{i,t+1}$:

$$A_{i,t+1} = \alpha + \sum_{s=1}^{p} \gamma_s A_{i,t-s+1} + \sum_{s=1}^{q} \beta_s B_{t-s+1} + \varepsilon_{t+1}, \quad (2)$$

where B_t denotes a signal defined by the mean deviation of the forecast target.

The lag lengths for two signals are chosen separately by the BIC criteria between 0 and 2. First, I estimate the parameters in equation 2 by expanding the data window (i.e., by all the data up to the forecasting date), and then use the estimated parameters and the latest signals to forecast $A_{i,t+1}$. I shall denote the forecast of $A_{i,t+1}$ by $\hat{A}_{i,t+1}$.

Given $\hat{A}_{i,t+1}$, my concern is twofold. First, "How big are $\hat{A}_{i,t+1}$?" Second, "Are $\hat{A}_{i,t+1}$ (forecasts of $A_{i,t+1}$ by given signals) trustworthy?" The first point is straightforward, since I want to select a forecast with the largest $A_{i,t+1}$ (without "hat"). Since the second point is somewhat complex, I discuss it in the next subsection.

(2) Signal predictability

Now I have $\hat{A}_{i,t+1}$, which are forecasts of $A_{i,t+1}$, but I do not know in general whether they are reliable, because it is not difficult to construct large $\hat{A}_{i,t+1}$ in practice. For example, I can expand my forecasting pool and find arbitrary signals to make $\hat{A}_{i,t+1}$ as large as possible (this is an example of forecasters' discretion).[8] Therefore, it is necessary to take into account the reliability of forecasts of $A_{i,t+1}$ by signals. I call this "signal predictability" and formally define it below. Note however that there is nothing new in this concept, since it is just (one minus) the p-value of the conditional predictive ability test by Giacomini and White. These authors prove that the following relationship holds under mild conditions:

[8] This concern is better described by White [2000]. There, forecasters' taking advantage of discretion is called "data snooping".

$$GW_{i,t} \equiv N\left(N^{-1}\sum_{t} h_{t-1}A_{i,t} \right)' V^{-1}\left(N^{-1}\sum_{t} h_{t-1}A_{i,t} \right) \xrightarrow{d} \chi^2_{\dim(h)}, \quad (3)$$

where N is the sample size, h is a vector of signals used to forecast $A_{i,t+1}$, V is a consistent estimate[9] of $Var\left(h_{t-1}A_{i,t}\right)$, and *dim(h)* is the number of signals. In my application, *dim(h)* depends on the BIC result for equation 2 and is given by 1+*p*+*q*. Equation 3 holds under the null hypothesis $H_0 : E\left[h_{t-1}A_{i,t}\right] = 0$.[10] Intuitively, this test statistic ("GW statistic" henceforth) detects the correlation between current signals and one-month-ahead relative usefulness. If the correlation is strong, the GW statistic becomes bigger and I can use the signals to predict $A_{i,t+1}$ with confidence. Therefore, I define signal predictability as one minus the p-value of the test and denote it by $P_{i,t}$. If I observe a large value of $P_{i,t}$, it is a good sign for $\hat{A}_{i,t+1}$. On the contrary, if $P_{i,t}$ is very small, I do not want to give much credit for $\hat{A}_{i,t+1}$.[11]

(3) A ranking measure with two dimensions

After obtaining a forecast of relative usefulness by equation 2 and signal predictability by equation 3 for all *i*, I am ready to rank primary forecasts based on $\left(\hat{A}_{i,t+1}, P_{i,t}\right)_{i=1\ldots28}$. I use $\hat{A}_{i,t+1} \times P_{i,t}$ as a ranking measure for primary forecasts, since in this way it appropriately reflects my twofold motivation. Since I do not provide a rigorous theory justifying this measure, I conduct a robustness check in the next section. I call the ranking defined by the decreasing order of $\hat{A}_{i,t+1} \times P_{i,t}$ GW ranking. I

[9] This test can be extended to longer forecast horizons. In that case, *V* is replaced by HAC (heteroskedasticity and autocorrelation consistent) estimator.

[10] The null (hence the alternative) hypothesis depends on signals and a benchmark model. Acceptance of the null hypothesis does not necessarily imply that forecast *i* is useless, but implies that signals *h* are not reliable to forecast relative usefulness defined with *i* and RW.

[11] Since the point made here is central to my method, I give an illustrative example. Suppose you are a gambler and want to know which racehorse will win. There are two tipsters by the racetrack and you know what they have said in the past. One says "Horse G will win by more than three lengths". The other says "Horse W will probably win by a narrow margin". A naive gambler might want to believe the first tipster, since it seems more certain to bet on a horse with a potential lead of three lengths rather than just a nose difference. But since you are an experienced gambler, you will discount their forecasts depending on their track records. Discounting may depend on other information. Knowing that the first tipster gives precise forecasts only when he is sober, if he is drunk today, you will heavily discount his forecast. In this example, tipsters, discounting, and their track records (and sober/drunk) correspond to *i*, *P*, *h*.

construct a GW ranking every forecasting date and pick up the 1st-ranked primary forecast. Note that GW ranking can drastically change in one period, because the sign of $\hat{A}_{i,t+1}$ matters. I call the thus-constructed forecast sequence GW1. In the next section, I investigate the nature of GW ranking and evaluate the performance of GW1.

4. RESULTS

The previous section demonstrated how to construct GW ranking. In this section, I report the result of its application to monthly exchange rate forecasts.

(1) The nature of GW ranking

Table 4 shows the frequency of GW rank 1st. GW1 picked out 8 forecasts out of 28 forecasts (RW is excluded by construction) during a forecasting period from May 2000 to April 2007. First, it is noteworthy that GW rankings put JC in the 1st rank three times. Since JC would never be selected if selection were based on average (hence unconditional) performance like MSE, this is a surprising result made possible by the conditional nature of GW ranking. Figures 4 and 5 show another aspect of GW ranking. Comparing histograms of GW ranks for P3, I3, and B3 with those for PIB1, PIB2, and PIB3, the latter large models are ranked relatively lower in GW rankings. It is intuitively reasonable to discount forecasts by large models more heavily, because they have more parameters to be estimated.[12]

Next, I check how the forecast behaves as a result of GW selection. Figures 6 to 8 show GW1 with the range of primary forecasts, realized exchange rate, and JC. In Figure 6, note the yen appreciation episode in 2002. Before appreciation started, GW1 picked IB1. The last time IB1 was used, it incurred a large forecast loss due to sudden appreciation. However, GW1 swiftly switched to I2, which followed the actual rate relatively well among primary forecasts. In Figure 7, there is another sudden yen appreciation episode in 2003. Before then, B3 was extensively used by GW1 forecast. Immediately after B3 incurred a large forecast error, GW1 switched to JC, which was

[12] This should be especially the case here, since the estimation window size is the same for all models.

the most radical forecast at that time. These are two major episodes of radical forecast switching, and more cases can be found in these figures. This shows that GW ranking can change flexibly and might contribute to controlling forecast errors under structural breaks.

(2) The performance of GW1 forecast

Since GW1 selects one of the primary forecasts every time, I can directly compare its primary rank to those of the primary forecasts presented in Section 2. Figures 9 and 10 are the same as Figures 2 and 3 except that a histogram of primary ranks for GW1 is added. Compared to "the most frequently 1st-ranked forecasts" (JC, B2, IB3) in Figure 9, GW1 decreases the frequency of being in bottom 10 and increases that of being in top 10. As a result, the histogram recovers its bell shape, with its peak on the left. Compared to "the least frequently worst-ranked forecasts" (RW, AR, B4) in Figure 10, although GW1 slightly increases the frequency of being in bottom 3, it increases the frequency of being in top 10. As a result, the histogram has thicker tails and is shifted further to the left. Finally, I check the MSE of GW1 forecast. Table 5 is the same as Table 3 except that GW1 is added and bold characters show the primary forecasts used in GW1. It shows that GW1 achieves a smaller MSE than any other primary forecasts.

(3) Robustness checks

In this subsection, I investigate my forecast selection method in further detail from three perspectives. First, I compare my method with two other forecasting methods: mean forecast and forecast selection by MSE-based ranking. Second, I introduce a generalized ranking measure which includes $\hat{A}_{i,t+1} \times P_{i,t}$ as a special case. Finally, I change the data availability timing to check the performance of my method in a situation closer to a real-time setting.

First, I compare my method with two alternative forecasting methods. The first is the mean forecast, which is the simple mean of all primary forecasts. In order to compare my method to the mean forecast method, I introduce GW-based forecast combinations. I combine two primary forecasts which are ranked 1st and 2nd by GW ranking and call it GW2. Similarly, by combining the top X primary forecasts and calling it GWX, I obtain GW2, GW3, …, GW28. Also, I use two ways to combine

forecasts. The first is the simple mean. The second is to use the inverse of the GW rank as weight. For example, if AR, JC, P1 are the top 3 in this order by GW ranking at some date, then GW3 (rank weight) forecast at this date is the weighted sum of these three forecasts with the weights given by (1, 1/2, 1/3). Figure 11 shows how MSE changes as I increase the number of combined primary forecasts based on GW ranking. The MSE of the mean forecast is shown by a single marker on the right and it is much larger than GW1. In typical forecast combinations, MSE tends to decrease as the number of combined forecasts increases. However, forecast combination based on GW ranking shows roughly the opposite. Figure 12 shows histograms of primary rankings[13] for forecast combinations. The more primary forecasts are combined, the more concentrated a histogram is around the center, hence a forecast becomes more conservative. The second forecasting method I consider is forecast selection by MSE-based ranking. At each forecasting date, I calculate MSE for each primary forecast and rank them in increasing order of MSE. The ranking changes much more slowly over time than the GW ranking because MSE measures the average property of the past. Figures 13 and 14 compare MSE-based forecast combination with GW-based forecast combination. The MSE-based forecast combination results in a higher MSE than GW-based forecast combination regardless of the number of combined forecasts.

For the second robustness check, I empirically investigate my ranking construction method based on $\left(\hat{A}_{i,t+1}, P_{i,t}\right)_i$. Consider the following generalized ranking measure:

$$K_{i,t} = sign\left(\hat{A}_{i,t+1}\right) \left|\hat{A}_{i,t+1}\right|^{\alpha} P_{i,t}^{1-\alpha}, \tag{4}$$

where the sign operator returns the sign (plus or minus) of the argument and $\alpha \in [0,1]$. Note that this measure yields the equivalent ranking to $\hat{A}_{i,t+1} \times P_{i,t}$ when $\alpha = 0.5$. Also, the measure includes two extreme cases where $\alpha = 0$ or $\alpha = 1$. Figure 15 illustrates these three cases. The left figure is the case where only the p-value of the GW test and the sign of $\hat{A}_{i,t+1}$ matter. The figure in the middle corresponds to the measure proposed in Section 3. The right figure is the case where only $\hat{A}_{i,t+1}$ matters

[13] Here, primary ranking is calculated among 30 forecasts (primary forecasts + one of the four forecasts in Figure 12).

for the ranking. By changing α between 0 and 1, I can check the effect of the relative weight of $\hat{A}_{i,t+1}$ and $P_{i,t}$ in the ranking measure on the forecast performance. Figure 16 shows how MSE changes as I change the value of α for GW1 and GW3 (rank weight). Though MSE becomes less sensitive to α for GW3, both GW1 and GW3 attain smaller MSE in the middle ($\alpha = 0.5$) than on both sides ($\alpha = 0$ or $\alpha = 1$). This implies that when constructing a ranking of primary forecasts, caring about both relative usefulness and signal predictability, not just one of them, is likely to increase forecast accuracy.

Finally, I change the data availability timing and see if the main results still hold. In the analysis so far, I simply took the one-month difference between the predictor variables and the target variable based on statistic dates. However, some data are not available with this timing due to the lag in publishing data. Since it is reasonable to suppose that exchange rates and interest rates are available with the timing employed above, I make an adjustment only for prices and trade data. I conduct the same out-of-sample exercise using a one-month-older data set for prices and trade data. Figures 17 and 18 show how MSE changes by this adjustment in data timing. Naturally, the performance becomes worse. Especially, the GW ranking incurs uncertainty for the top 2 ranks and GW1 does not yield the smallest MSE in this case. However, GW3 (rank weight) still beats every primary forecast.[14] Therefore, in the case of real-time application of my method, it is recommended that forecasters watch not only the top primary forecast but also other top-ranked forecasts in the GW ranking.

5. CONCLUSION

This paper introduces a new method for selecting forecasts from many forecasts. Using GW statistics, I constructed a two-dimensional measure which enables me to rank forecasts in the order of ex-ante forecast accuracy. I apply the method to the monthly yen/dollar exchange rate and show empirically that it is useful for increasing forecast accuracy.

[14] MSE of GW3 is 7.39 with this data timing. Among primary forecasts, I3 still gives the minimum MSE shown in Table 3.

Much work remains to confirm the empirical results presented here. Most importantly, theoretical work needs to be done to investigate why my method works. I provided only a heuristic argument and believe it would be fruitful to further formalize the idea. Second, as another robustness check, it is straightforward to extend the empirical work to longer forecast horizons. Finally, since the method is based on the asymptotic property of GW statistics, care needs to be taken for its performance with a finite sample in any application.

Table 1: List of primary forecasts

Forecast number	Code	Name of variable (process)
Benchmark	RW	The latest realization of target
1	JC	JCIF survey
2	AR	Auto regression
3	P1	Japanese inflation
4	P2	US inflation
5	P3	Inflation differential (P1 - P2)
6	I1	Japanese interest rate
7	I2	US interest rate
8	I3	Interest rate differential (I1 - I2)
9	B1	Japanese export
10	B2	Japanese import
11	B3	Japanese trade balance (B1 - B2)
12	B4	Export (from Japan to US)
13	B5	Import (from US to Japan)
14	B6	Trade balance between Japan and US (B4 - B5)
15	B7	US export
16	B8	US import
17	B9	US trade balance (B7 - B8)
18	BB	B3, B9
19	PB1	P3, B3
20	PB2	P3, B6
21	PB3	P3, B9
22	PI	P3, I3
23	IB1	I3, B3
24	IB2	I3, B6
25	IB3	I3, B9
26	PIB1	P3, I3, B3
27	PIB2	P3, I3, B6
28	PIB3	P3, I3, B9

Note: Forecasts 3 to 28 also include AR terms.

Table 2: Frequency of primary ranks

Primary ranks	Most frequent	Second	Third	…	Least frequent
1st	JC/B2/IB3	---	---		RW/AR/etc.
Top 3	IB3	B2	I2/PB3		AR
Top 10	IB1	B3	I3		RW
11th to 19th	RW	AR	P1		PIB2
Bottom 10	JC	PIB2	PIB3		RW
Bottom 3	JC	PIB2	PIB3		P1
29th	JC	B2	PIB3		RW/AR/etc.

Note: For the row "1st", I counted the number of times each model was ranked 1st in primary ranks and listed models with high and the lowest frequency. For the row "Top 3", I counted the number of times each model was ranked in top 3 (1st, 2nd, and 3rd) in primary ranks. Similarly for the other rows.

Table 3: MSE ranking for primary forecasts

MSE rank	Code	MSE
1	I3	7.40
2	B3	7.79
3	I2	7.85
4	IB1	7.90
5	B5	8.03
6	PI	8.04
7	PIB1	8.08
8	**B4**	8.15
9	**AR**	8.15
10	B7	8.17
11	**RW**	8.28
12	P1	8.33
13	I1	8.47
14	IB2	8.47
15	BB	8.49
16	PB1	8.64
17	P2	8.66
18	B6	8.68
19	B8	8.69
20	**B2**	8.75
21	B9	8.88
22	P3	8.89
23	B1	8.93
24	**IB3**	9.31
25	PIB2	9.57
26	PB2	9.64
27	PIB3	9.91
28	PB3	10.3
29	**JC**	10.7

Note: Forecasts in bold are the ones mentioned at the end of Section 2 (or in Figures 2 and 3).

Table 4: Frequency of GW rank 1st

Frequency of GW rank 1st	Code
23	AR
17	I2
14	B5
13	B3
8	I1
3	JC/BB/IB1
0	Others

Table 5: MSE ranking for primary forecasts and GW1

MSE rank	Code	MSE
1	GW1	7.04
2	I3	7.40
3	**B3**	7.79
4	**I2**	7.85
5	**IB1**	7.90
6	**B5**	8.03
7	PI	8.04
8	PIB1	8.08
9	B4	8.15
10	**AR**	8.15
11	B7	8.17
12	RW	8.28
13	P1	8.33
14	**I1**	8.47
15	IB2	8.47
16	**BB**	8.49
17	PB1	8.64
18	P2	8.66
19	B6	8.68
20	B8	8.69
21	B2	8.75
22	B9	8.88
23	P3	8.89
24	B1	8.93
25	IB3	9.31
26	PIB2	9.57
27	PB2	9.64
28	PIB3	9.91
29	PB3	10.3
30	**JC**	10.7

Note: Forecasts in bold are the ones used in GW1.

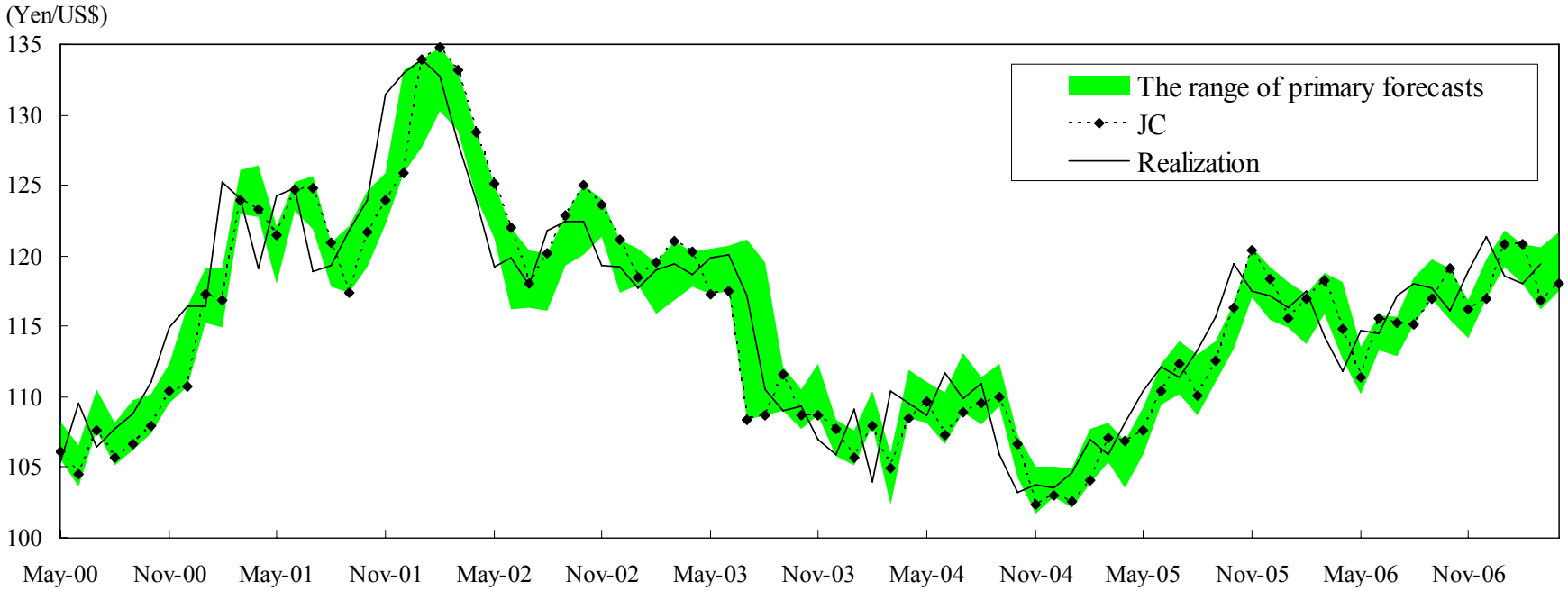

Figure 1: Primary forecasts

Note: Dates of forecasting timing are shown on the x-axis. Namely, forecast values and realization above "May 2000" on the x-axis are those of June 2000.

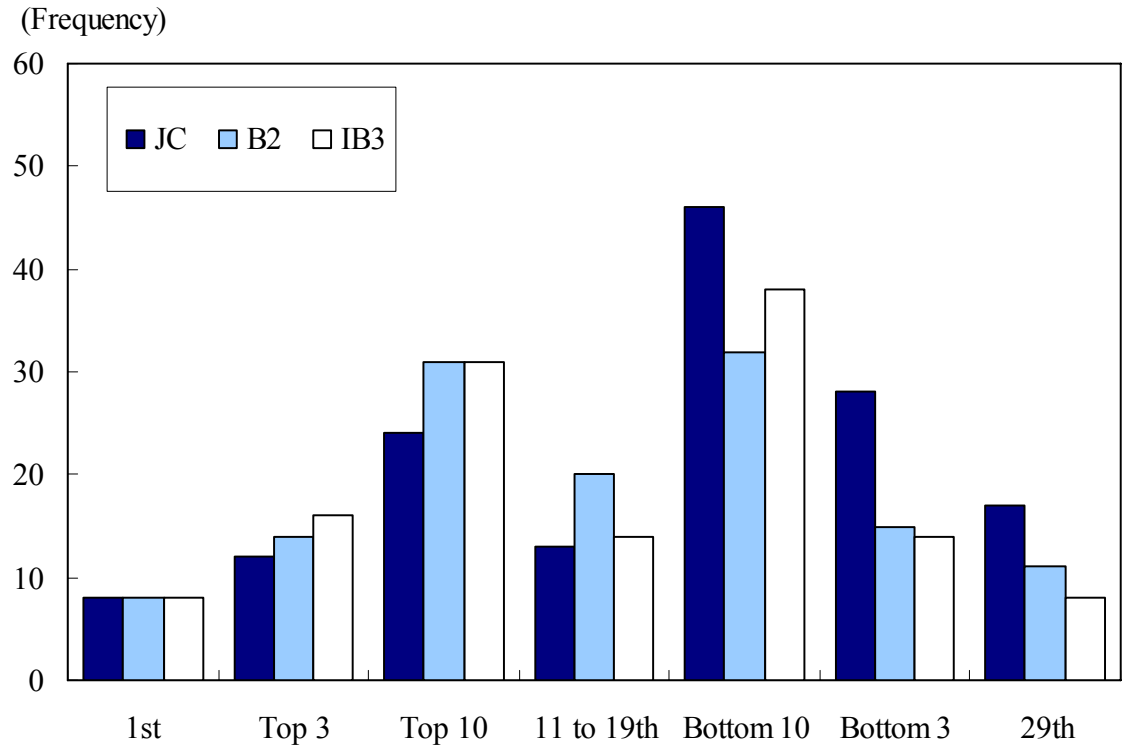

Figure 2: Histograms of primary ranks for the most frequently 1st ranked models

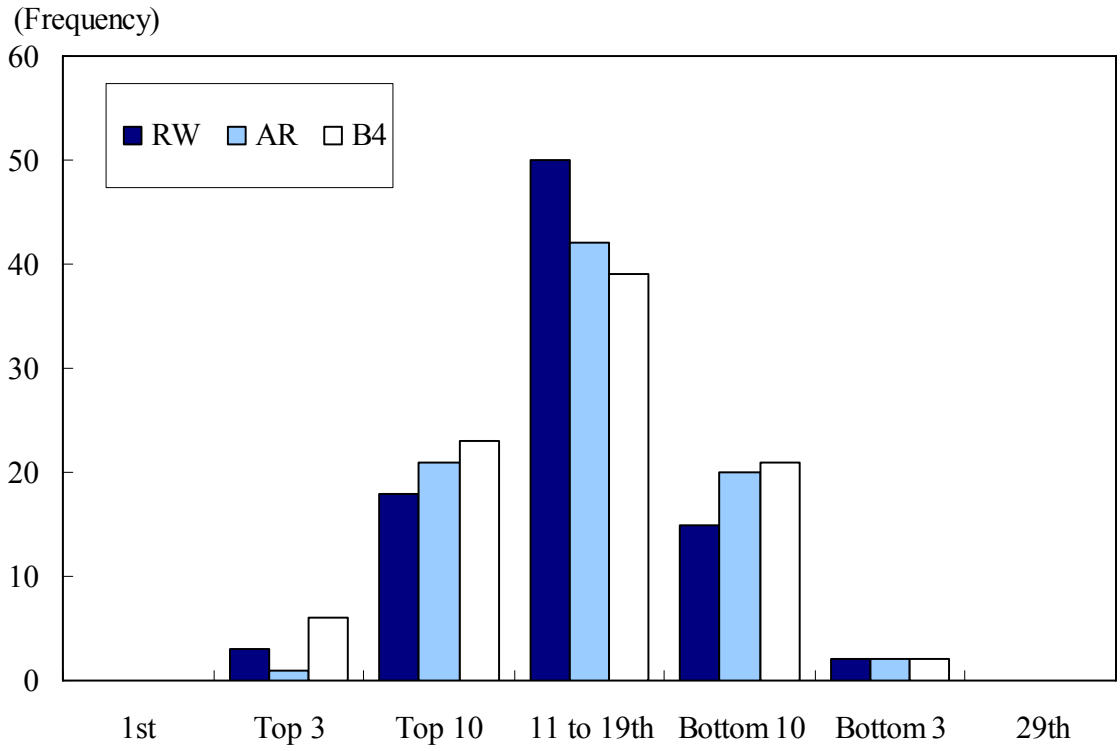

Figure 3: Histograms of primary ranks for the least frequently worst ranked models

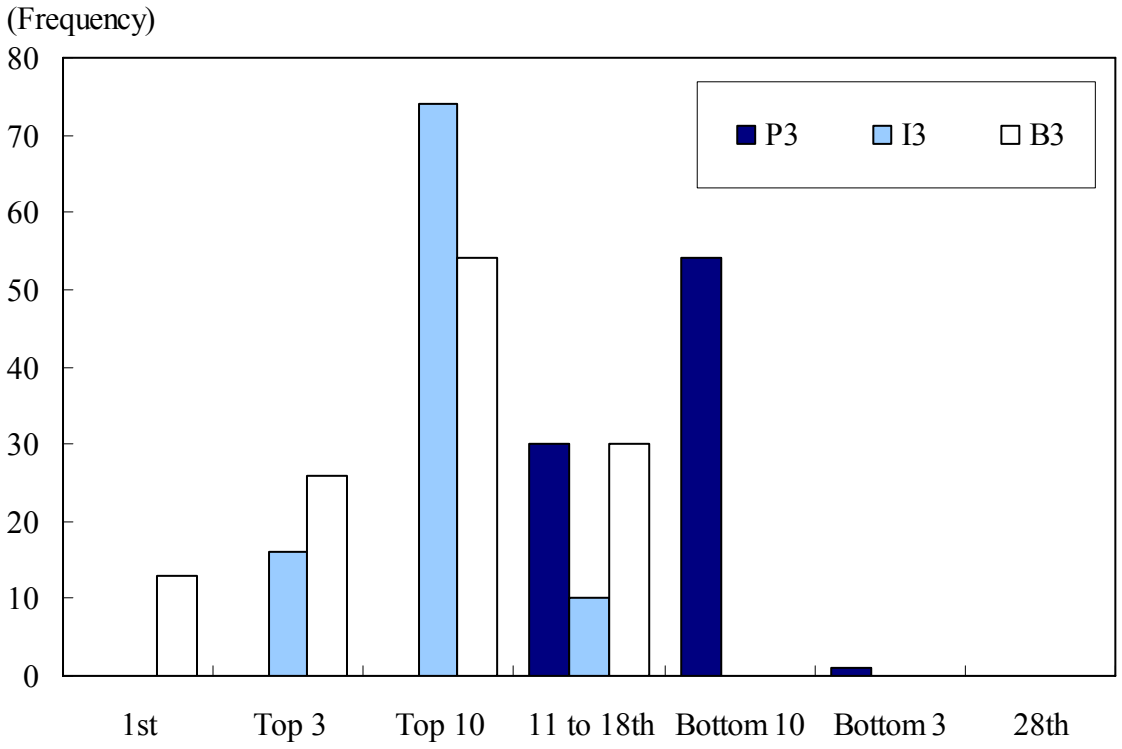

Figure 4: Histograms of GW ranks for small models

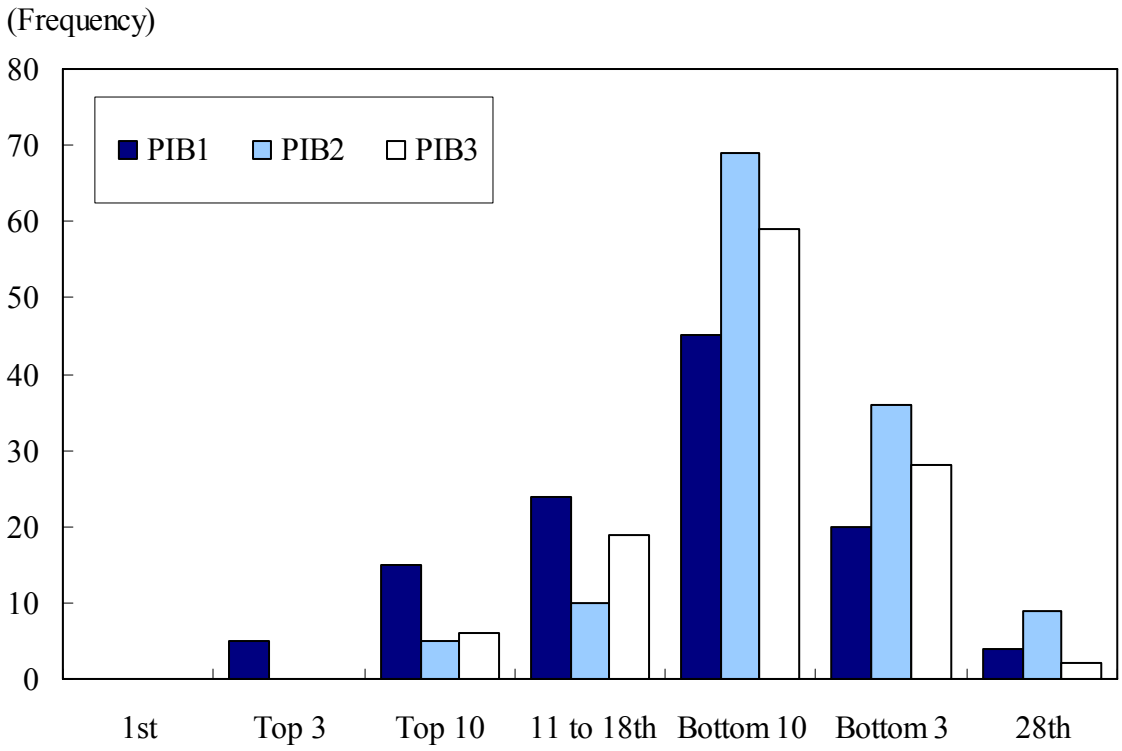

Figure 5: Histograms of GW ranks for large models

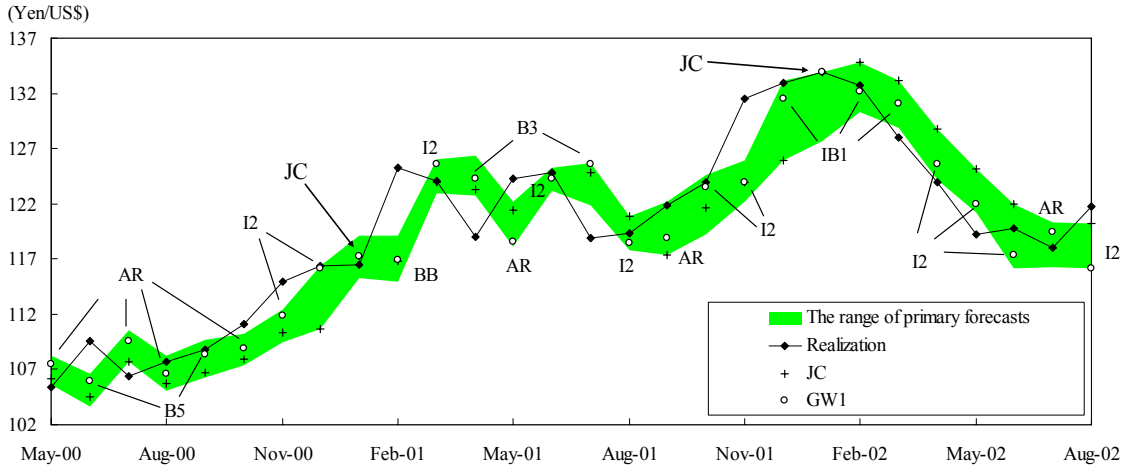

Figure 6: GW1 forecast (May 2000 – Aug 2002)

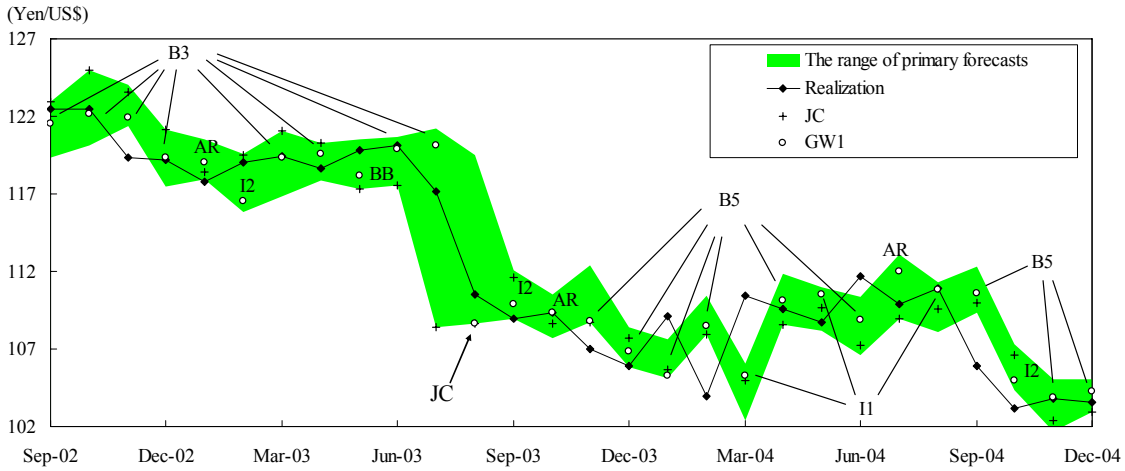

Figure 7: GW1 forecast (Sep 2002 – Dec 2004)

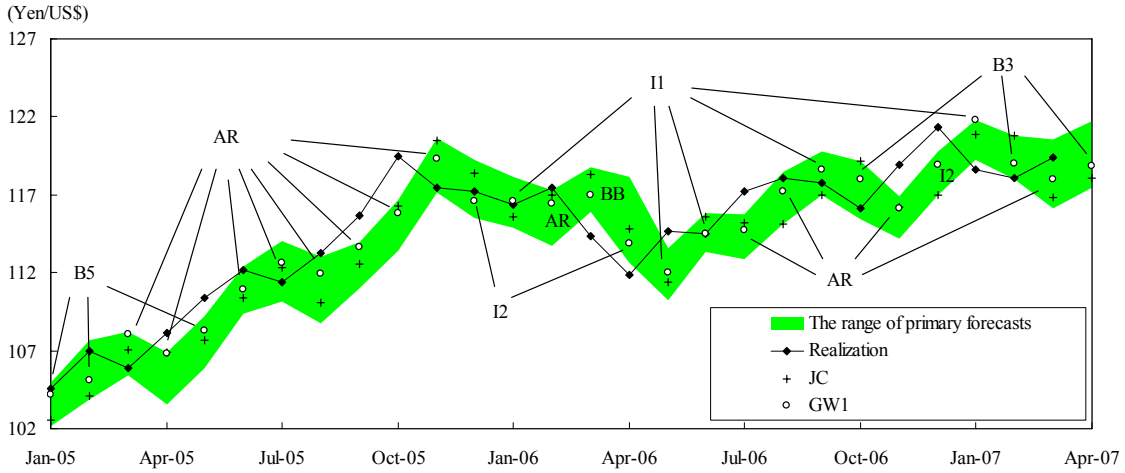

Figure 8: GW1 forecast (Jan 2005 – Apr 2007)

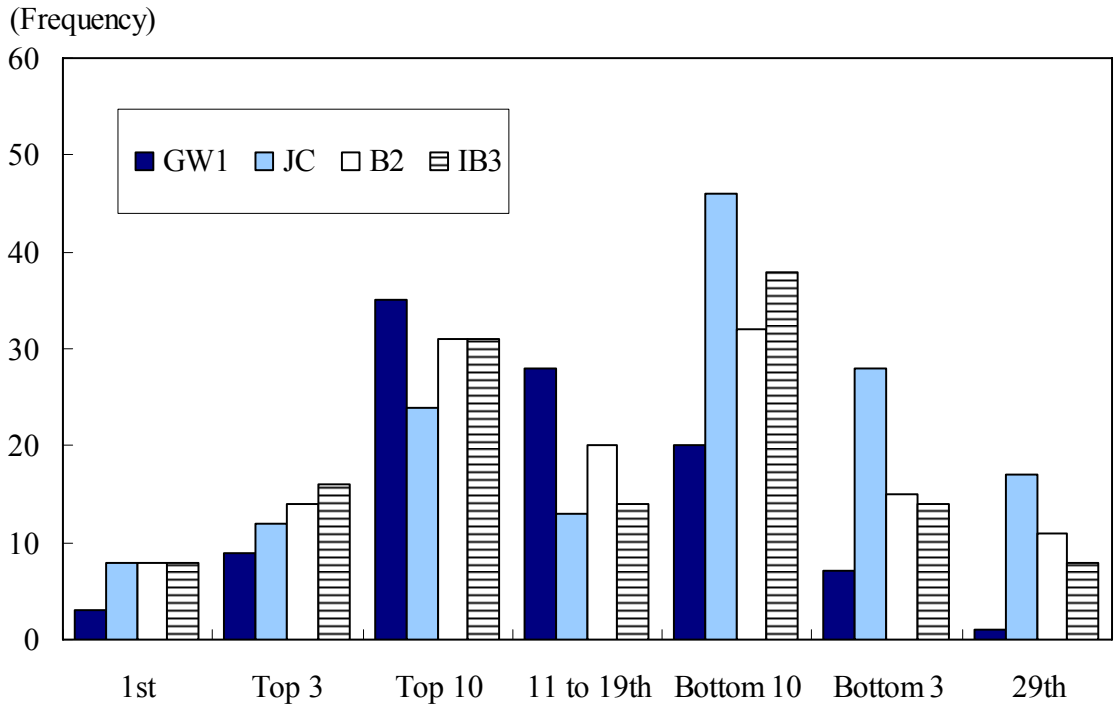

Figure 9: Histograms of primary ranks for GW1 with the most frequently 1st ranked models

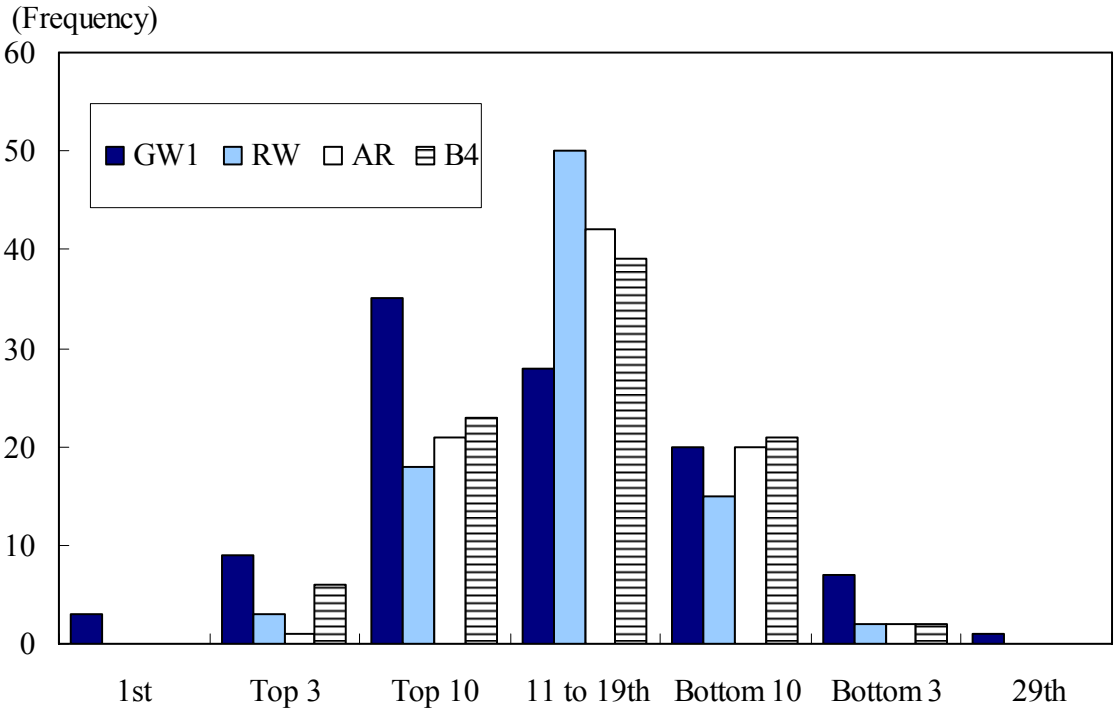

Figure 10: Histograms of primary ranks for GW1 with the least frequently worst ranked models

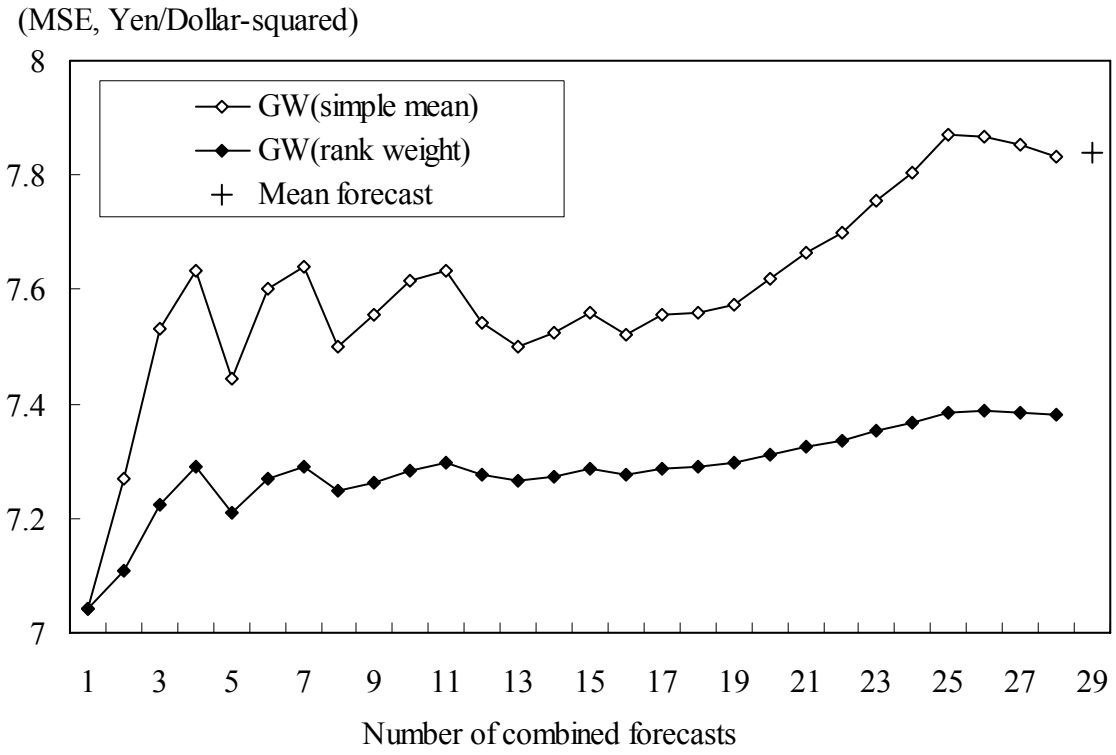

Figure 11: MSE for GW-based forecast combinations with mean forecast

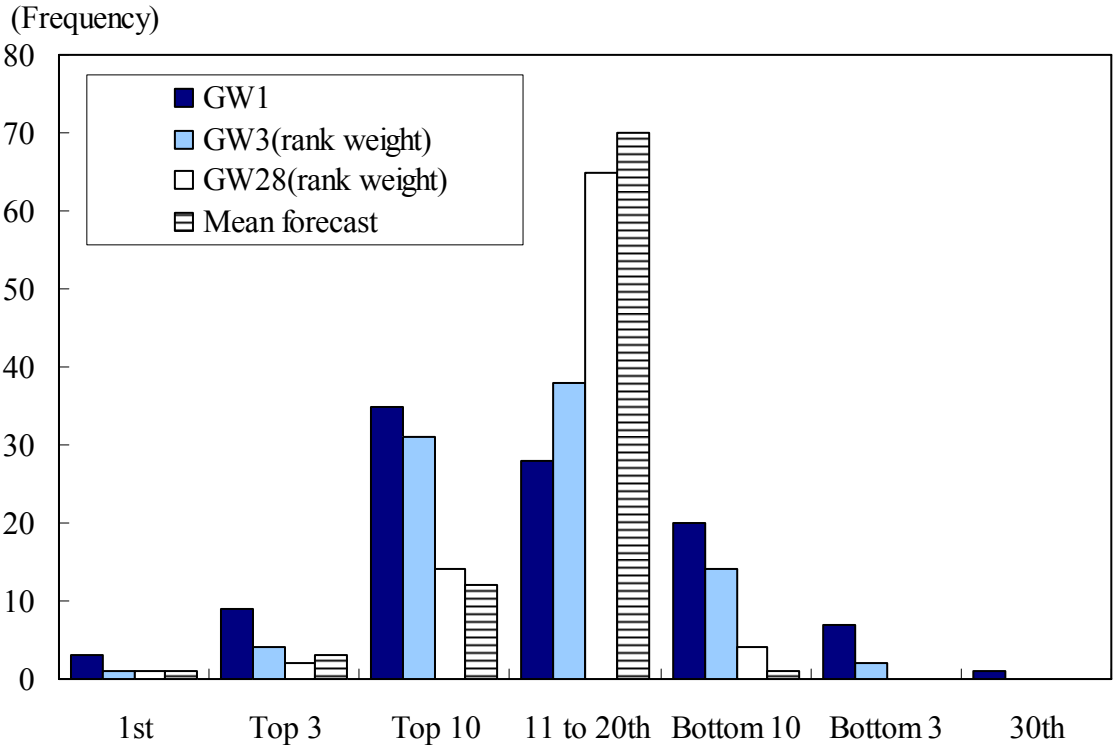

Figure 12: Histograms of primary ranks for GW-based forecast combinations with mean forecast

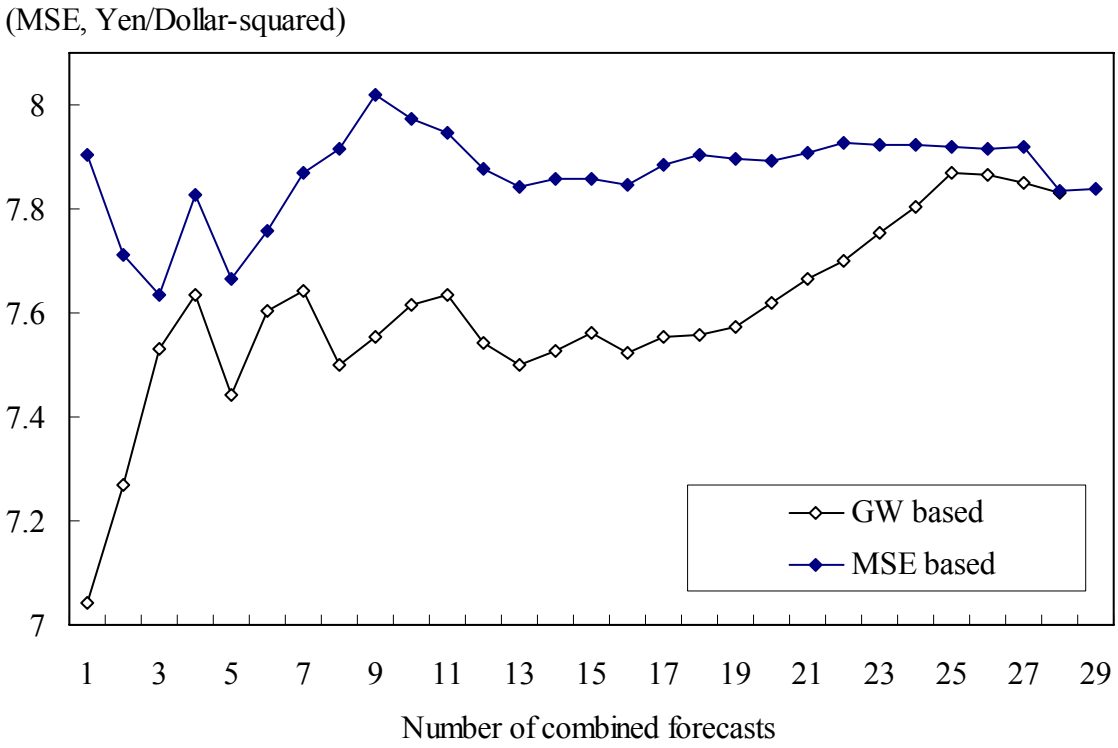

Figure 13: MSE of GW-based forecast combinations and MSE-based combinations (simple mean)

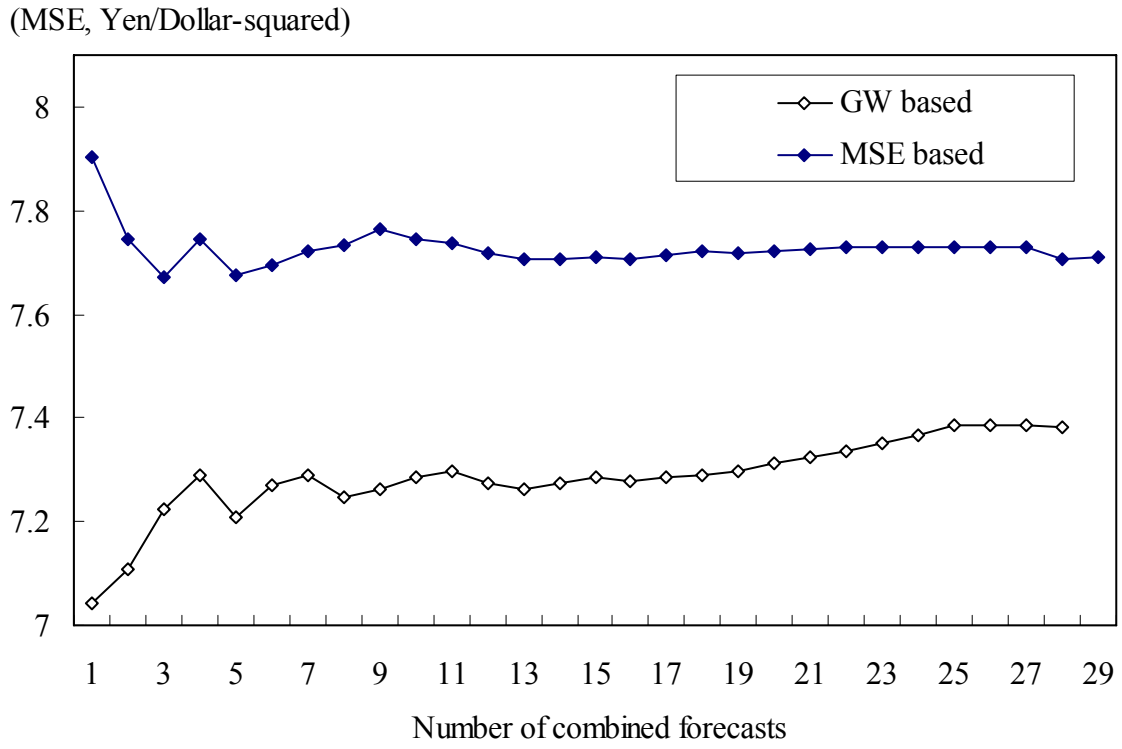

Figure 14: MSE of GW-based forecast combinations and MSE-based combinations (rank weight)

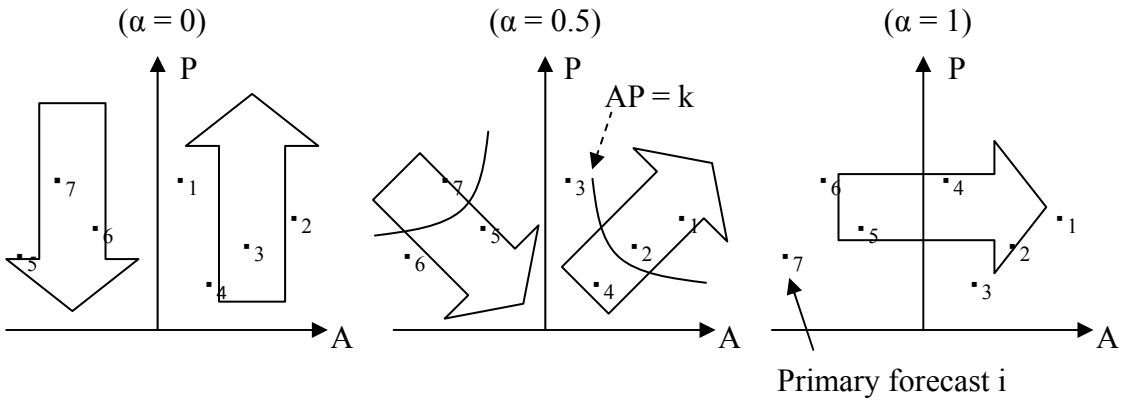

Figure 15: Three extreme cases of a generalized ranking measure

Note: The block arrows show ranking directions for each case.

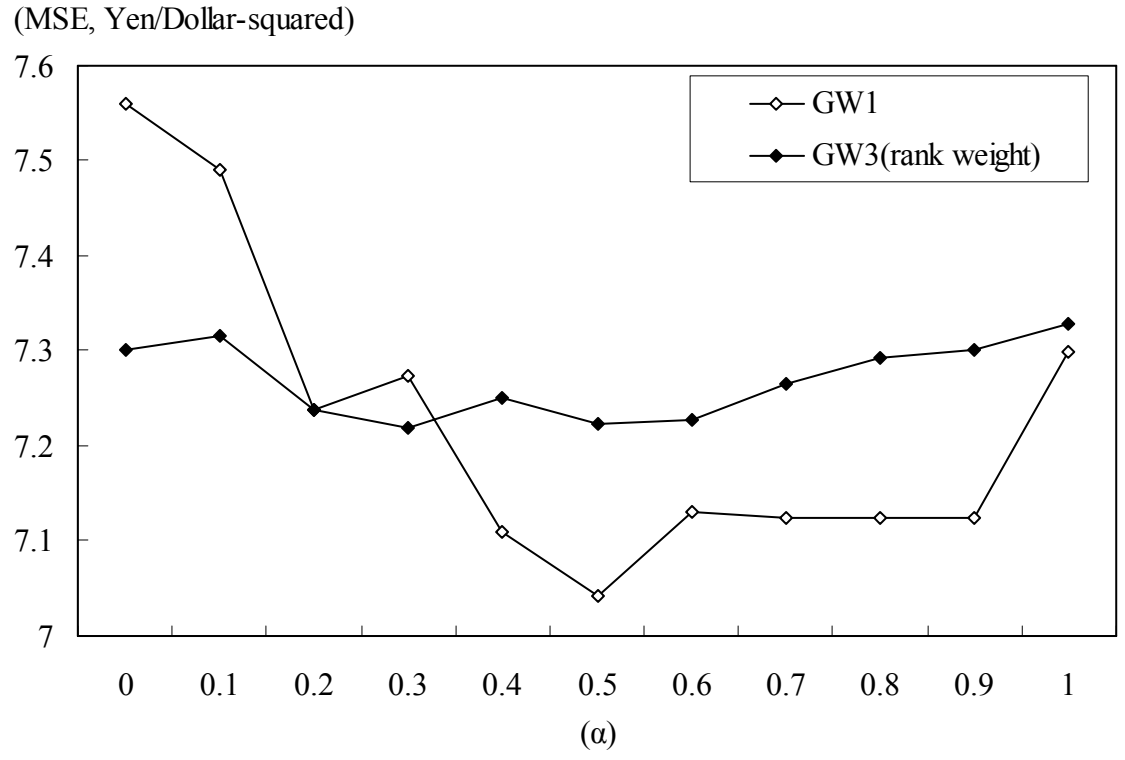

Figure 16: MSE for different α of a generalized ranking measure

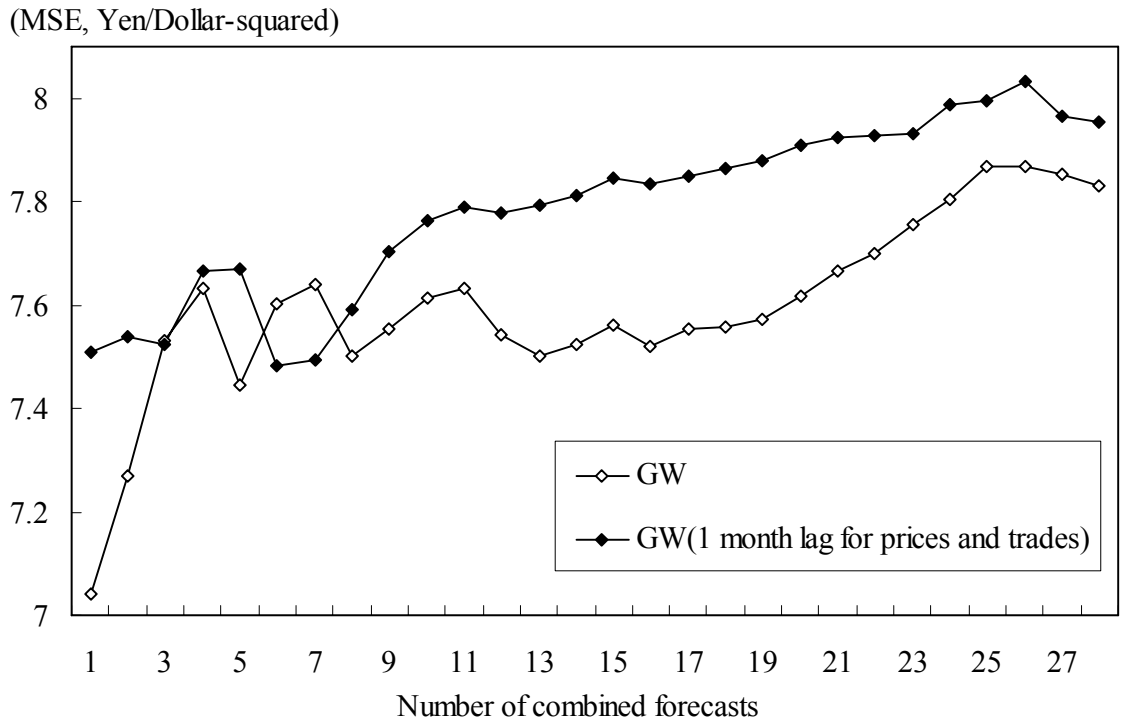

Figure 17: MSE for forecast combinations with older data timing (simple mean)

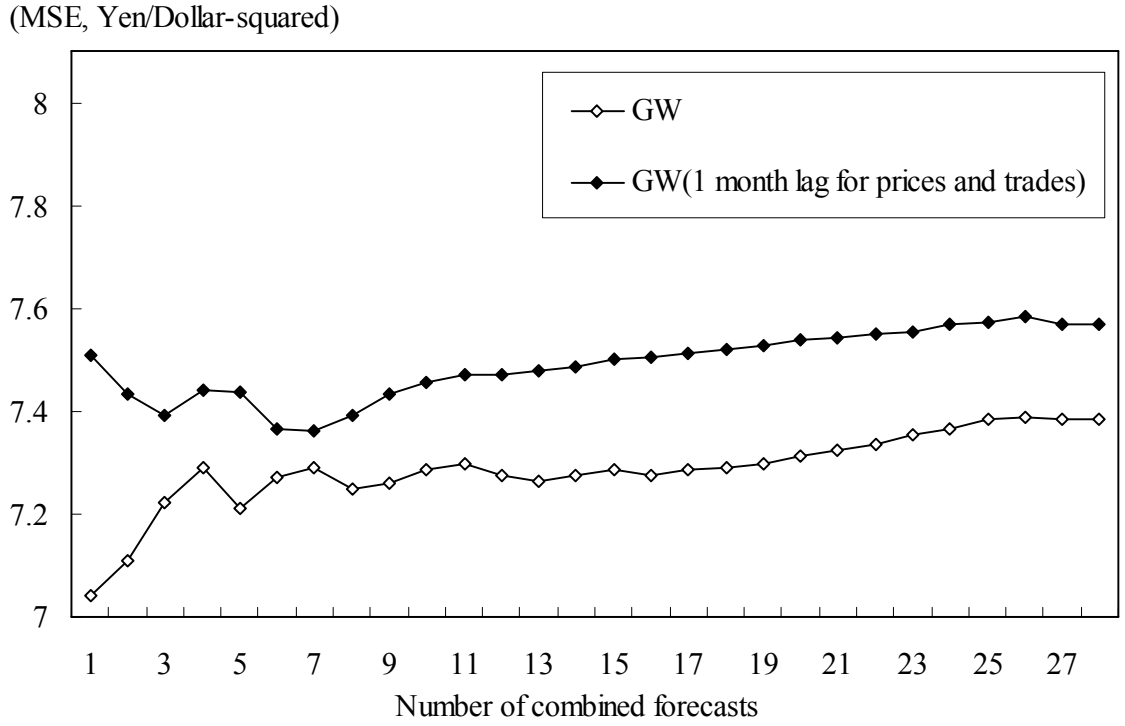

Figure 18: MSE for forecast combinations with older data timing (rank weight)

APPENDIX. DATA SOURCES

The following list includes the definitions and sources of the data used in this paper. The sample is monthly and spans the period from January 1973 through April 2007 except prices and the JCIF survey. Only trade data are seasonally adjusted. For US-Japan trade data, seasonal adjustment by Census X12 is used for the original series from January 1970 to July 2007.

Data	Definitions and Sources
Exchange rate	Yen/dollar spot rate. Interbank rate at Tokyo market. End of month. Source: *Financial and Economic Statistics Monthly*, Bank of Japan.
Prices (January 72 - April 07)	Japan: Consumer Price Index. General, excluding "fresh food." Year 2005 = 100. Source: *Consumer Price Index*, Ministry of Internal Affairs and Communications
	US: Consumer Price Index. All items less food and energy. Year 1982-84 = 100. Source: *Consumer Price Index*, Department of Labor Bureau of Labor Statistics
Short-term interest rates	Japan: Uncollateralized overnight call rate (or collateralized overnight call rate). Source: *Financial and Economic Statistics Monthly*, Bank of Japan. Note: Uncollateralized rate since July 1985. Prior to this, collateralized rates are used, adding the mean spread between uncollateralized and collateralized rates, as in Miyao (2005).
	US: Federal funds rate. Source: *Federal funds effective rate*, Board of Governors of the Federal Reserve System
Exports/Imports	Japan, Japan-US: Exports, customs. Imports, customs. Source: *Trade Statistics*, Ministry of Finance
	US: Exports, F.O.B. Imports, C.I.F. Source: *International Financial Statistics (IFS)*, IMF
Survey (May 85 – April 07)	Exchange rate forecast. Forecast horizon = one month. Source: *Market Data Survey*, Japan Center for International Finance

REFERENCES

Clark, T. E., and M. W. McCracken, "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics*, Vol. 105, 2001, pp. 85-110.

Clements, M. P., and D. F. Hendry, "Forecasting Economic Time Series," Cambridge University Press, 1998.

Diebold, F. X., and R. S. Mariano, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, Vol. 13, 1995, pp. 253-265.

Engel, C. M., N. C. Mark, and K. D. West, "Exchange Rate Models Are Not As Bad As You Think," *NBER Working Paper Series*, No. 13318, 2007.

Fujiwara, I., and M. Koga, "A Statistical Forecasting Method for Inflation Forecasting," *Bank of Japan Research and Statistics Department Working Paper Series*, No. 02-5, 2002.

Giacomini, R., and H. White, "Tests of Conditional Predictive Ability," *Econometrica*, Vol. 74, No. 6, 2006, pp. 1545-1578.

Hansen, P. R., "A Test for Superior Predictive Ability," *Journal of Business and Economic Statistics*, Vol. 23, No. 4, 2005, pp. 365-380.

Hara, N., and K. Kamada, "Kawasesoubayosoku to Syouhisyabukka: kawaseyosoku saabei wo mochiita jissyou bunseki," *Bank of Japan Research and Statistics Department Working Paper Series*, No. 02-7, 2002 (in Japanese).

----------, "Yen/Dollar Exchange Rate Expectations in the 1980-90's," *Bank of Japan Research and Statistics Department Working Paper Series*, No. 99-1, 1999.

Ito, T., "Foreign Exchange Rate Expectation: Micro Survey Data," *American Economic Review*, Vol. 80, No. 3, 1990, pp. 434-449.

Kitamura, T., and R. Koike, "The Effectiveness of Forecasting Methods Using Multiple Information Variables," *Institute for Monetary and Economic Studies Discussion Paper Series*, No. 2002-E-20, 2002.

Meese, R. A., and K. Rogoff, "The Out-of-Sample Failure of Empirical Exchange Rate Models: Sampling Error or Misspecification," *Journal of International Economics*, Vol. 14, 1983, pp. 3-24.

Miyao, R., "Use of the Money Stock in the Conduct of Japan's Monetary Policy: Re-Examining the Time-Series Evidence," *Japanese Economic Review*, Vol. 56, No. 2, 2005, pp. 165-187.

Timmermann, A., "Forecast Combinations," Chapter 4 in *Handbook of Economic Forecasting*, ed. by G. Elliot, C. W. J. Granger, and A. Timmermann. Amsterdam: North-Holland, 2006.

West, K., "Asymptotic Inference about Predictive Ability," *Econometrica*, Vol. 64, No. 5, 1996, pp. 1067-1084.

White, H., "A Reality Check for Data Snooping," *Econometrica*, Vol. 68, No. 5, 2000, pp. 1097-1126.