THE UNIVERSITY OF
MELBOURNE

Department of Economics

Working Paper Series

Conditional Forecast Selection from Many Forecasts: An Application to the Yen/Dollar Exchange Rate

Kei Kawakami

JAN 2013

Research Paper Number 1167

ISSN: 0819 2642
ISBN: 978 0 7340 4518 8

Department of Economics
The University of Melbourne
Parkville VIC 3010
www.economics.unimelb.edu.au

Conditional Forecast Selection from Many Forecasts: An Application to the Yen/Dollar Exchange Rate

Kei Kawakami*

November 26, 2012

Abstract

This paper proposes a new method for forecast selection from a pool of many forecasts. The method uses conditional information as proposed by Giacomini and White (2006). It also extends their pairwise switching method to a situation with many forecasts. I apply the method to the monthly yen/dollar exchange rate and show empirically that my method of switching forecasting models reduces forecast errors compared with a single model.

JEL classification: C52; C53; F31; F37

Keywords: Conditional predictive ability; Exchange rate; Forecasting; Forecast combinations; Model selection

*Department of Economics, University of Melbourne, e-mail: keik@unimelb.edu.au

1 Introduction

Forecasters often face a problem of selecting forecasts. 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 forecast series are used as primary forecasting models, from which I select a single model (or multiple models for forecast combinations) at each forecasting date. The empirical results and Monte Carlo simulations show that my method of switching forecasting models can reduce forecasts errors compared with a single model.

The key feature of my forecast selection method is that it can respond to the new information much more quickly than forecast selection methods based on the past average performances. The following example illustrates the point. 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. Today, one tipster says "Horse A will win", while the other says "Horse B will win". If a naive gambler only knows that *on average* the first tipster has higher probability of identifying a winning horse, he might want to bet on the horse A. His decision is based on an *unconditional* test. On the other hand, an experienced gambler may evaluate the tipsters' predictions depending on their *conditions today.* For example, knowing that the first tipster gives more precise forecasts when he is sober (most of the time he is) than when he is drunk, and if he is drunk today, the experienced gambler will discount the first forecast. This is a *conditional* test. I show switching models by the timely use of conditional information can reduce forecast errors under structural breaks.

There is a vast body of literature on forecast selection.[1] My method is motivated by the conditional predictive ability test proposed by Giacomini and White (2006). In their paper, they propose a decision rule for selecting between two models. My method uses conditional

[1] Diebold and Mariano (1995) propose a pairwise comparison method, which is generalized by West (1996) and Clark and McCracken (2001) to incorporate parameter uncertainty and a comparison of nested models. Giacomini and White (2006) allows for the use of conditional information. White (2000), Hansen (2005) and Hansen, Lunde and Nason (2011) study methods of comparing more than two models.

information for forecast selection from more than two models. 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.[2] My empirical results are consistent with the preceding literature in that beating a random walk *by a single model* is difficult. However, I also find that switching models based on conditional information might outperform a random walk forecast.[3]

The paper is constructed as follows. Section 2 evaluates out-of-sample performances of 29 forecast series. Section 3 explains the new forecast selection method. Section 4 examines the out-of-sample performance of the forecast selection method. Robustness checks and Monte Carlo simulations are also presented. Section 5 concludes the paper.

2 Primary Forecasting Models

In this section, I introduce 29 primary forecasting models (henceforth *primary models*), from which one or multiple forecasts are selected (and combined) at each forecasting date. See **Table 1** for the list of primary models and the appendix 1 for data sources. A forecast by the latest realized target value shall be called a RW (random walk) forecast. This forecast series is easy to construct, frequently used both in practice and in the literature. Also, because I use a quadratic loss function to evaluate forecast performances, the random walk forecast minimizes the expected loss if a forecast target follows a random walk process. For these reasons, I use RW as a benchmark, upon which forecasters wish to improve. Forecast performances of other models as well as any forecast selection methods are measured by the decrease in the forecast loss relative to RW. A formal measure will be defined in the next subsection. I use the sample means of all the respondents in the JCIF survey as a forecast sequence and call it JC (Japanese companies). Other 27 models are simple linear-regression models. The appendix 2 gathers more details about the survey and regression-based models.

[2] This finding has been corroborated by more recent research. See Engel, Mark and West (2007).

[3] Altavilla and De Grauwe (2010) find that combining different forecasting procedures produces more accurate forecasts than a single model. While they combine a fixed set of forecasts based on average forecast performances, my method combines a time-varying set of forecasts using conditional information.

2.1 Evaluation of primary models

All primary models span the period of 264 months from May 1985 to April 2007. Forecast errors of primary models in the period from May 1985 to April 2000 are used as "inputs" for my forecast selection method from May 2000. Because I compare the performance of my forecast selection method to that of each primary model, I set the same evaluation period of 83 months from May 2000 to March 2007 for all primary models.[4]

I use the following forecast performance measure based on the mean squared errors (henceforth, MSE). Let $M = \{i \in I | i = 1, ..., 29\}$ be a set of primary models. Let $i = 1$ be the benchmark model RW. Other forecast series $i > 29$ can be constructed by selecting or combining models in M over time. Let MSE_i be MSE of a forecast series i, which can be a primary model as well as a forecast selection method. First, consider a hypothetical forecast selection method which can choose the primary model with the minimum absolute forecast error at each forecasting date. I call this *the ex post best selection* and denote the associated MSE by $\underline{MSE}$. Hence, $MSE_1 - \underline{MSE}$ is the maximum gains over the benchmark that can be expected from any forecast selection method which chooses one model in M at each time. To evaluate a forecast selection method, I use the performance measure f_i defined by

$$f_i \equiv \frac{MSE_1 - MSE_i}{MSE_1 - \underline{MSE}} * 100. \tag{1}$$

Thus, f_i measures the amount of loss reduction from the benchmark loss MSE_1, as a percentage to the maximum loss reduction by the ex post best selection. Note that $f_1 = 0$ and f_i is bounded above by 100% for any selection method i which selects one model at a time.[5]

Table 2 shows f_i for primary models and other forecast series. In this sample, $MSE_1 = 8.28$ and $\underline{MSE} = 2.56$. The model $I3$ achieves the smallest MSE of 7.40 and the largest $f_i = \frac{8.28-7.40}{8.28-2.56} * 100 = 15.36\%$ among the primary models. Because this is the best performance

[4] There are 84 forecasts including the last one made at April 2007. The last forecast can not be evaluated since I do not have a realization of the target for May 2007.

[5] f_i is not bounded below. For example, the ex post worst selection can have arbitrarily large negative f_i.

one could have achieved *by using a single model* in M, I call $I3$ *the ex post best model.* Comparing the performance of the ex post best model (15.36%) to that of the ex post best selection (100%), there seem to be large gains from switching models. In fact, my forecast selection method (shown as $GW1$ in **Table 2**) has a better performance than $I3$.

2.2 Why does switching models help?

How can switching between primary models reduce forecast errors? I check which primary models produced more accurate forecasts at each forecasting date, and how often so. A *primary ranking* at each forecasting date is defined by the order of absolute size of forecast errors. For example, if JC forecast made in May 2000 is closest to the realization of the June 2000 exchange rate in absolute terms among all primary models, JC's primary rank is 1st at this forecasting date. The primary ranking reveals interesting properties of the survey forecast JC and the benchmark model RW. **Figure 2(a)** shows the cumulative distributions of primary ranks for RW and JC. Note that the distribution for the ex post best selection would be degenerate at the 1st rank (a flat line at 100%). A 45 degree line ("+" markers in the figure) represents a uniform distribution over ranks. First, JC has the distribution with its left end above the 45 degree line and its remaining part below it. This means that JC forecasts move wildly between the top and the bottom of the primary ranking. In fact, JC is both "the most frequently 1st ranked" and "the most frequently worst ranked" forecast among all primary models. On the contrary, RW was never ranked either 1st or worst and its distribution is concentrated in the middle of the primary ranking. The contrast between these two forecasts indicates a trade-off between models: a model which can yield forecasts with pinpoint accuracy can also yield large forecast errors at other times (high and low primary ranks), while a model which surely avoids large forecast errors sacrifices pinpoint accuracy (middle primary ranks).

The latter property is even more pronounced for the mean forecast (a simple average of all primary models). **Figure 2(b)** shows the cumulative distribution of primary ranks for

RW, the mean forecast, and the ex post best model $I3$.[6] The mean forecast does not improve upon RW on the left side of the distribution, while there is a considerable improvement on the right side of the distribution. In other words, the mean forecast does not achieve the pin-point accuracy, but it avoids large losses by "diversifying" forecast errors over primary models. However, it fails to beat the performance of $I3$ as shown in **Table 2**.

These empirical findings are the basis of my forecast selection method. As long as I use a single model or simply average models, I cannot escape the trade-off shown above. However, it might be possible to switch models to overcome the trade-off and reduce forecast losses.

3 Conditional Forecast Selection

This section explains how to construct a selection criterion using conditional information. This involves three steps: (i) forecast loss differences, (ii) check statistical reliability of the forecasts, and (iii) construct a ranking of primary models using both (i) and (ii). I explain each step in separate subsections.

3.1 Forecasting loss differences

The loss[7] difference of primary model i at time t is defined by

$$A_{i,t} \equiv e^2_{1,t|t-1} - e^2_{i,t|t-1}, \tag{2}$$

where $e_{i,t|t-1}$ is an error of the forecast made by primary model i at time $t-1$ for the target value at time t. Note that $e_{i,t|t-1}$ becomes available to a forecaster at time t. By construction, $A_{1,t} = 0$ for all t. The positive value of (2) suggests that primary model i should have been used instead of the benchmark model at time $t-1$. The negative value suggests the opposite. Also, $A_{i,t} > A_{j,t}$ suggests the better performance of primary model i

[6] Here, primary ranks were recalculated with the mean forecast. Hence, the bottom rank is 30th.

[7] I use a quadratic loss function but it can be replaced by other loss functions.

over j concerning forecasts made at time $t-1$.

At time t, a forecaster wishes to know one-period-ahead loss differences $\{A_{i,t+1}\}_{i=2}^{29}$ to select the best model. I employ two types of conditional information to forecast $A_{i,t+1}$. The first variable is the realized loss differences up to time t. This is expected to capture systematic mistakes that some models may typically, if not always, make. If such a systematic pattern exists, past loss differences can predict future loss differences. The second variable is the mean deviation of the forecast target, where the mean is that of the most recent 12 months' data available at time t. This time-varying mean is meant to capture structural changes in exchange rate formation in a market. Given these two kinds of conditional information, I propose the following regression-based forecast of $A_{i,t+1}$:

$$A_{i,t+1} = \alpha_i + \sum_{s=1}^{p_i} \beta_{i,s} A_{i,t-s+1} + \sum_{s=1}^{q_i} \gamma_{i,s} B_{t-s+1} + \varepsilon_{i,t+1}, \tag{3}$$

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

The lag lengths p_i and q_i for the two signals are chosen by the BIC criteria between 0 and 2. First, I estimate the parameters in (3) by expanding the data window (i.e., by all the data up to the forecasting date t), and then use the estimated parameters and the latest signals to forecast $A_{i,t+1}$. Let $\widehat{A}_{i,t+1}$ denote the forecast of $A_{i,t+1}$. Note that if $p_i = q_i = 0$, $\widehat{A}_{i,t+1}$ will be the average of past loss differences. Therefore, my method makes use of conditional information only if the BIC criterion chooses $p_i > 0$ or $q_i > 0$. In Section 4, I provide a robustness check with respect to the choice of p_i and q_i.

3.2 Statistical reliability

Now I have $\left\{\widehat{A}_{i,t+1}\right\}_{i=2}^{29}$, forecasts of $\{A_{i,t+1}\}_{i=2}^{m}$, but I do not know which one to trust. In practice, it is possible to construct large $\widehat{A}_{i,t+1}$ by arbitrary choice of signals so that primary model i looks good. Therefore, it is necessary to take into account the statistical reliability of $\widehat{A}_{i,t+1}$ estimated with a specific choice of conditional information. For this purpose, I use *one* *minus the p-value* of the conditional predictive ability test by Giacomini and White (2006). The authors prove the following result under mild conditions.

$$N\left(\frac{1}{N}\sum_{t}h_{i,t-1}A_{i,t}\right)'V^{-1}\left(\frac{1}{N}\sum_{t}h_{i,t-1}A_{i,t}\right)\xrightarrow{d}\chi^2_{\dim(h_i)}, \tag{4}$$

where N is the number of out-of-sample forecasts, $h_{i,t}$ is a vector of signals used to forecast $A_{i,t+1}$, V is a consistent estimate[8] of $Var[h_{i,t-1}A_{i,t}]$, and $\dim(h_i)$ is the number of signals.[9] Note that $\dim(h_i) = 1 + p_i + q_i$ depends on the BIC result for regression (3). Intuitively, this test statistic (4) ("GW statistic" henceforth) detects the correlation between the conditional information and the one-month-ahead loss difference. If the correlation is larger, I can use the chosen conditional information to predict $A_{i,t+1}$ with more confidence. Therefore, I discount $\widehat{A}_{i,t+1}$ by multiplying one minus the p-value of the test and denote the discount factor by $P_{i,t} \in [0,1]$. If $P_{i,t}$ has a large value, it is a good sign for $\widehat{A}_{i,t+1}$. If $P_{i,t}$ is very small, I do not want to give much credit for $\widehat{A}_{i,t+1}$.[10]

3.3 Ranking measure

Given forecasts of the loss difference by (3) and their statistical reliability by (4) for all $i \in M$, consider the following ranking measure:

$$K_{i,t}(\eta) \equiv sign\left(\widehat{A}_{i,t+1}\right)\left|\widehat{A}_{i,t+1}\right|^{\eta} P_{i,t}^{1-\eta}, \tag{5}$$

where the sign operator $sign(\cdot)$ returns the sign (plus or minus) of the argument and $\eta \in [0,1]$ is a parameter which controls the relative importance of $\widehat{A}_{i,t+1}$ to $P_{i,t}$. When $\eta = 0.5$, (5) is equivalent to the product $\widehat{A}_{i,t+1}P_{i,t}$ in terms of the ranking they produce. When $\eta = 0$,

[8] For longer forecast horizons, V is replaced by the HAC (heteroskedasticity and autocorrelation consistent) estimator.

[9] Technical assumptions behind the convergence result (4) is discussed in the appendix 3.

[10] In the example given in the introduction, tipsters, their conditions (sober/drunk) today, and discounting correspond to models (i), conditional information (h_i), and conditional predictive ability test (P_i).

only the p-value of the GW test and the sign of $\widehat{A}_{i,t+1}$ matter.[11] When $\eta = 1$, only $\widehat{A}_{i,t+1}$ matters for the ranking. I propose the ranking with $\eta = 0.5$ as a main case, but I also study the effect of the weight η on the forecast performance. The ranking defined by the decreasing order of (5) is called the *GW-ranking.* By combining the top x primary models for $x \in \{1, .., 29\}$ at each forecasting date, I construct 29 forecast series $GW1$, $GW2$, $GW3$, . . . , $GW29$. I also compare two weighting schemes to combine selected models. The first is the simple mean (referred to as using *mean weights*). The second scheme uses the inverse of the GW-ranks as weights (referred to as using *rank weights*). Relative to the mean weights, the rank weights give more weights for forecasts that are higher in the GW-ranking.[12] The next section evaluates the performances of $GW1$, . . . , $GW29$.

4 Results

As shown in the bottom of **Table 2**, the MSE of $GW1$ forecast series is 7.04 and its f_i is 19.6%. Thus, $GW1$ achieves a smaller MSE than the ex post best model $I3$. **Figure 2** shows $GW1$ with the range of forecasts made by primary models and realized exchange rates. In **panel b**, the yen appreciation started in July 2003. Before then, RW was repeatedly selected in $GW1$. Immediately after RW incurred a large forecast loss, $GW1$ switched to JC, which was the most radical forecast at that time. Similar cases can be found in the figure. This indicates that the GW-ranking can change flexibly to reduce forecast errors under structural breaks. **Figure 1(c)** shows the cumulative distribution of primary ranks of $GW1$ contrasted with those of RW and the mean forecast. Unlike the mean forecast, it improves upon RW on the left side of the distribution.

Table 3 lists primary models in the order of the average GW-ranks for 84 months from

[11] This corresponds to the following procedure: (i) for models with a positive prediction $\left\{i \in M | \widehat{A}_{i,t+1} \geq 0\right\}$, attach higher ranks to model i with larger $P_{i,t}$, (ii) for models with a negative prediction $\left\{i \in M | \widehat{A}_{i,t+1} < 0\right\}$, attach lower ranks to model i with larger $P_{i,t}$.

[12] If $I1$ and $P1$ are the top 2 models by the GW-ranking, then $GW2$ (rank weights) forecast is the weighted sum of the two forecasts with the weights given by (1, 1/2) normalized to sum up to one.

May 2000 to April 2007. The second column shows the average of the BIC lag selection results for the regression (3). The remaining columns show the numbers of times where each model is ranked 1st, 2nd, .., and 5th in the GW-ranking. The benchmark RW has the best average rank 4.4, while JC has the worst average rank 25.6. That RW is ranked high by my forecast selection method is consistent with the literature on the exchange rate forecast. On average, it is difficult to beat the random walk by a single model. What is more surprising is that JC was ranked 1st three times, even though it has the worst average rank and the largest MSE.[13] The forecast sequence $GW1$ picked out 8 models during the forecasting period. There are only 6 primary models for which average p_i is positive. Out of these 6 models, 5 models were used in $GW1$ for 52 periods.[14] Recall that $(p_i, q_i) = (0, 0)$ means that the *unconditional* predictive ability test was used for such models. Interestingly, none of these models was used in $GW1$. **Figure 3** shows $(P_{i,t}, \widehat{A}_{i,t+1}, A_{i,t+1})$ for the primary models with $(p_i, q_i) = (0, 0)$. **Panel a** shows that some of these models are heavily discounted. More importantly, **Panel b** shows that the values of $\widehat{A}_{i,t+1}$ look completely different from the actual $A_{i,t+1}$ in **Panel c**. For these models, the estimate $\widehat{A}_{i,t+1}$ is based on the average over the expanding sample period and it fails to respond to the current information. **Figure 4** shows $(P_{i,t}, \widehat{A}_{i,t+1}, A_{i,t+1})$ for the primary models used in $GW1$. **Panel a** shows that discount factors $P_{i,t}$ are above 70% and stable over time. **Panel b** shows large fluctuations of $\widehat{A}_{i,t+1}$ reflecting even larger fluctuations of $A_{i,t+1}$ in **Panel c**. This implies frequent model switching in $GW1$.

In sum, the forecast selection based on the GW-ranking shows two features. First, through the BIC lag selection, both conditional and unconditional tests are used in the forecast selection. However, when the unconditional test was applied, models were heavily discounted and not ranked high in the GW-ranking. Second, when the conditional test was applied, $\widehat{A}_{i,t+1}$, rather than $P_{i,t}$, mainly drives the forecast selection. In fact, the next

[13] This can be due to its "directional" accuracy. I checked how often the actual loss difference $A_{i,t+1}$ had the same sign with the prediction $\widehat{A}_{i,t+1}$ for each primary model. It turns out that JC has the highest frequency: 68.7% of the time, predicted loss differences were in the right directions.

[14] For the remaining 32 periods, RW and two other models ($B3$ and $B5$, both of which have average $(p_i, q_i) = (0, 1)$) were used in $GW1$.

subsection shows that the performance difference between $\eta = 0.5$ and $\eta = 1$ is small.

4.1 Robustness checks

This subsection presents robustness checks for the proposed forecast selection method from three perspectives. First, I study the performance of forecast combinations based on the GW-ranking, and compare it with forecast combinations based on the MSE-based ranking. Second, I investigate how changing the value of η from 0.5 affects the result. Finally, I conduct a robustness check with respect to the lag selection method of (p_i, q_i) in (3).

For each forecasting date t, denote by $mse_{i,t}$ the real-time MSE of $i \in M$, and rank the primary models in the increasing order of $mse_{i,t}$ (*the MSE-ranking*). By combining the top x primary models based on this ranking for $x \in \{1, .., 29\}$, I construct forecast series $MSE1$, $MSE2$, $MSE3, \ldots, MSE29$ (both mean and rank weights are considered as before). **Figure 5** shows the performances of $GW1 - 29$ and $MSE1 - 29$. First, the performance of forecast combinations based on the MSE-ranking is worse than forecast combinations based on the GW-ranking. In particular, MSE-based forecasts do not come close to the ex post best model ($I3$, $f_i = 15.4\%$) and the improvement over the mean forecast is small.[15] Second, for $GW1 - 29$, the use of rank weights resulted in the better performance than mean weights. More surprisingly, the forecast performance roughly *decreases* in the size of combinations. This contrasts the benefit of my forecast selection method with that of forecast combinations. We usually expect that forecast errors from different models cancel out – the benefit of diversification. However, if conditional information tells us which model is better, pooling a bad model worsens the forecast performance and this may outweigh the benefit of diversification. The MSE-ranking does not contain much information about which model is better *at the moment*. Hence the benefit of diversification is not severely compromised. In the GW-ranking, models are ranked using more information, and adding worse models can worsen forecast performance. **Figure 5** is consistent with this argument.[16]

[15] Note that $GW29$ (mean weight), $MSE29$ (mean weight) and the mean forecast are equivalent.

[16] The literature on forecast combinations (see Timmermann (2006) and Clements and Hendry (1998))

Figure 6(a) shows the change in f_i when η is changed from 0.5. The impact of changing η between 0.5 and 1 is small, while that of changing η between 0 and 0.5 can be large. **Figure 6(b,c)** show the same picture for different choice of lag lengths for the regression (3). The forecast performance is more sensitive to η when the lag length is fixed. **Table 4** shows f_i for different lag selections. The BIC with the maximum lag length 2 is the benchmark case. In the row of "Lag 1", the performance with the BIC with the maximum lag length 1 is shown in the left panel, while the performance with the fixed lag length 1 is shown in the right panel. Shaded cells indicate the better performances than the ex post best model $I3$. Except when the maximum lag length is 1, the performance is not so sensitive to the choice of the maximum lag length for the BIC, because the average BIC lag length does not change much. On the other hand, the performance seems more sensitive to the choice of the fixed lag length. All in all, except the very small value of η, the result is not too sensitive to the value of η, and using the BIC criteria and $\eta = 0.5$ seems to be a sensible choice.

4.2 Monte Carlo evidence

This subsection presents Monte Carlo experiments to support my forecast selection method. I consider the following data generating process.

$$\begin{aligned} &A_{1,t+1} \equiv 0, \text{ and for } i \geq 2, \\ &A_{i,t+1} = \rho_i A_{i,t} + 4(1-\rho_i)\mu_i(S_t - 0.5) + \varepsilon_{i,t+1}, \quad \varepsilon_{i,t+1} \sim \text{i.i.d. } N(0,\sigma^2), \end{aligned} \tag{6}$$

where $S_t = 1$ or 0 with equal probabilities for each period.[17] Primary models differ in two respects of the loss difference: (i) persistence $\rho_i \in \left\{\underline{\rho}, 0, \overline{\rho}\right\}$ and (ii) jump size $\mu_i \in \left\{0, \frac{\overline{\mu}}{2}, \overline{\mu}\right\}$ in response to the state S_t. The noise term $\varepsilon_{i,t+1}$ is i.i.d. both in time-series and in cross-section.

recommends real-time estimation of optimal weights for multiple forecasts. An effort is targeted to find weights so that forecast errors from combined models cancel out as much as possible. In my method, an effort is directed to find relevant information in order to choose the currently best model.

[17] This is a modified version of the simulation design in section 5.2 in Giacomini and White (2006). Since the forecast *loss* is directly generated, forecast combinations cannot be studied. A Monte Carlo experiment of forecast combinations requires a careful design of cross-sectional relationship among forecast *errors*, which is left for the future work.

The process (6) implies that $E[A_{i,t+1}] = 0$ for all i, but for $i \geq 2$,

$$E[A_{i,t+1}|A_{i,t}, S_t] = \begin{cases} \rho_i A_{i,t} + 2(1-\rho_i)\mu_i \text{ if } S_t = 1 \\ \rho_i A_{i,t} - 2(1-\rho_i)\mu_i \text{ if } S_t = 0 \end{cases} .$$

Therefore, unconditionally all primary models have the equal predictive ability, while conditionally their performances can be different. This makes the conditional information $(\{A_{i,t}\}_{i \in M}, S_t)$ potentially useful for the forecast selection. A benchmark model $i = 1$ is implicit, and the other models $i \in \{2, ..., 10\}$ are characterized by 9 combinations of (ρ_i, μ_i). I consider the following four cases of parameter values.

	Case 1	Case 2	Case 3	Case 4
$(\underline{\rho}, \overline{\rho}, \overline{\mu}, \sigma)$	$(-0.2, 0.2, 1, 20)$	$(-0.2, 0.2, 1, \mathbf{10})$	$(-0.2, 0.2, \mathbf{2}, 20)$	$(-\mathbf{0.3}, \mathbf{0.3}, 1, 20)$

The parameters for the case 1 were chosen such that the unconditional variance $Var[A_{i,t+1}]$ matches that in the data used in this paper. The parameters for the other three cases are chosen to capture better environments for the conditional forecast selection because of small noise (case 2), large jump (case 3), and high persistence (case 4). For each case, a time series sample is generated by (6). The size of estimation window, the forecasting period, the evaluation period are as described in section 3. This is repeated for 3,000 times.[18]

Table 5 shows the average (for 3,000 simulations) performance measure f_i of $GW1$ with $\eta \in \left\{0, \frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}, 1\right\}$ and different lag selection methods for each of 4 cases. Shaded cells imply that the average performance of $GW1$ is better than that of the ex post best model. Except when η is small, $GW1$ has a better average performance than the ex post best model.[19] Finally, the effect of η on the average performance is minor except when $\eta = 0$. By deviating from $\eta = 0.5$, there can be both losses and gains depending on the lag selection

[18] MSE-based forecast selection is not useful in this experiment, because all primary models are designed to perform equally *on average*. In fact, the average f_i of $MSE1$ is nearly zero for all four cases.

[19] I also check the frequency (% out of 3,000 simulations) that $GW1$ outperforms the ex post best model. For the benchmark method (the BIC lag 2 and $\eta = 0.5$), $GW1$ is better than the ex post best forecast for about 60% for case 1-3. With higher persistence (case 4), the frequency goes up to above 90%.

method, but losses seem greater than gains. This supports the use of $\eta = 0.5$ and the BIC lag selection as a practical benchmark.

5 Conclusion

This paper proposed a new method of selecting (and combining) forecasts from many models. Motivated by the work of Giacomini and White (2006), I constructed a time-varying measure which uses conditional information to rank many models according to their relative forecast accuracy. I applied the method to the monthly yen/dollar exchange rate and showed empirically that it improved the forecast performance compared with a single model.

Much work remains to confirm the empirical results presented here. First, as another robustness check, it is straightforward to extend the empirical work to longer forecast horizons. Second, since the method is based on the asymptotic property of the conditional predictive ability test, care needs to be taken for its performance with a finite sample. Third, a mean forecast could be included in a set of primary models and used as a benchmark instead of the random walk model. Such an exercise will shed more light on the relationship between forecast selection and forecast combinations. Finally, recent research shows that there is significant heterogeneity between forecasters, which follows a systematic pattern (e.g., Beine et al., 2007; Ruelke et al., 2010). This additional information may be useful to improve the forecast performance. These are left for the future work.

Acknowledgments I would like to thank the editor, anonymous referees, Raffaella Giacomini, Kouichirou Kamada, Marc Luppino, Mototsugu Shintani, Tomoyoshi Yabu and the staff of the Bank of Japan for their helpful comments. The financial support from the faculty of business and economics at the university of Melbourne is gratefully acknowledged. Any remaining errors are mine.

Appendix 1. Data sources

The following table includes the definitions and sources of the data. The sample is monthly and spans the period from January 1973 through April 2007 unless otherwise stated in the table. 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 sources

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 1972 - April 2007)	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.

Data sources (continued)

Data	Definitions and Sources
Short-term interest rates	Japan: Uncollateralized 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 1985 - April 2007)	Exchange rate forecast. Forecast horizon = one month. Source: Market Data Survey, Japan Center for International Finance.

Appendix 2. Primary forecasting models

The JCIF 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. I focus on the forecast horizon of one month; hence 264 time series forecasts are available. 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 used in this paper is not an end-of-month to end-of-month forecast. 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. For further descriptions of the survey, see Ito (1990) and Hara and Kamada (1999).

For regression-based models, I use a rolling estimation scheme with an estimation window of size 120 months prior to the forecasting date. Each model has the autoregression term and/or other variables (prices, interest rates, and trade related data). Lag lengths are chosen by the BIC criterion for each model between 0 and 4 for the autoregression term and between 1 and 4 for the other terms.[20] Three kinds of variables are used: (i) annual inflation rates, (ii) short-term interest rates, and (iii) 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 series 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 series for interest rates ($I1$, $I2$, $I3$). Three forecast series 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 series using trade data between Japan and the US ($B4$, $B5$, $B6$), and three more forecast series with US trade data ($B7$, $B8$, $B9$). Also, I construct models using two or three variables selected from above. For example, the model $PB1$ includes the AR term, the inflation differential, and the Japanese trade balance as predictors.

Appendix 3. Conditional predictive ability test

Convergence (4) holds under the null hypothesis

$$H_{0,h_i} : E[h_{i,t-1}A_{i,t}] = 0 \ \forall t.$$

[20] Lag zero means forecasting by the sample mean of the target.

The null and the alternative hypotheses depend on the choice of conditional information and a benchmark model. More precisely, the test is $H_{0,h_i} : \forall t, E[h_{i,t-1}A_{i,t}] = 0$ vs. $H_{1,h_i} : \exists t, E[h_{i,t-1}A_{i,t}] \neq 0$. These are exhaustive under stationarity, but not necessarily so under heterogeneity. Under heterogeneity, the test may have no power against important alternatives. I assume stationarity in this paper. Acceptance of the null hypothesis does not necessarily imply that model i is useless, but implies that conditional information $\{h_{i,t}\}$ is not reliable to forecast the loss difference between model i and RW.

References

[1] Altavilla, C. and P. De Grauwe (2010): “Forecasting and Combining Competing Models of Exchange Rate Determination,” *Applied Economics*, 42, 3455-3480.

[2] Beine M., A. Benassy-Quere, and R. MacDonald (2007): “The Impact of Central Bank Intervention on Exchange-rate Forecast Heterogeneity,” *Journal of The Japanese and International Economies*, 21, 38-63.

[3] Clark, T. E., and M. W. McCracken (2001): “Tests of Equal Forecast Accuracy and Encompassing for Nested Models,” *Journal of Econometrics*, 105, 85-110.

[4] Clements, M. P., and D. F. Hendry (1998): “Forecasting Economic Time Series,” Cambridge University Press.

[5] Diebold, F. X., and R. S. Mariano (1995): “Comparing Predictive Accuracy,” *Journal of Business and Economic Statistics*, 13, 253-265.

[6] Engel, C. M., N. C. Mark, and K. D. West (2007): “Exchange Rate Models Are Not As Bad As You Think,” NBER Working Paper Series, No. 13318.

[7] Giacomini, R., and H. White (2006): “Tests of Conditional Predictive Ability,” *Econometrica*, 74, 6, 1545-1578.

[8] Hansen, P. R. (2005): “A Test for Superior Predictive Ability,” *Journal of Business and Economic Statistics*, 23, 4, 365-380.

[9] Hansen, P. R., A. Lunde, and J. M. Nason (2011): “The Model Confidence Set,” *Econometrica*, 79, 2, 453-497.

[10] Hara, N., and K. Kamada (1999): “Yen/Dollar Exchange Rate Expectations in the 1980-90’s,” Bank of Japan Research and Statistics Department Working Paper Series, No. 99-1.

[11] Ito, T. (1990): "Foreign Exchange Rate Expectation: Micro Survey Data," *American Economic Review*, 80, 3, 434-449.

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

[13] Miyao, R. (2005): "Use of the Money Stock in the Conduct of Japan's Monetary Policy: Re-Examining the Time-Series Evidence," *Japanese Economic Review*, 56, 2, 165-187.

[14] Ruelke, J. C., M. R. Frenkel, and G. Stadtmann (2010): "Expectations on the Yen/Dollar Exchange Rate – Evidence from the Wall Street Journal Forecast Poll," *Journal of The Japanese and International Economies*, 24, 355-368.

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

[16] West, K. (1996): "Asymptotic Inference about Predictive Ability," *Econometrica*, 64, 5, 1067-1084.

[17] White, H. (2000): "A Reality Check for Data Snooping," *Econometrica*, 68, 5, 1097-1126.

Tables

Table 1: List of primary models

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

Note: Models 4 to 29 may also include AR terms depending on the lag selection result.

Table 2: Performance of primary models and other forecast series

(a) Primary models

Rank	Code	MSE	f_i (%)
1	I3	7.40	15.36
2	B3	7.79	8.65
3	I2	7.85	7.55
4	IB1	7.90	6.58
5	B5	8.03	4.36
6	PI	8.04	4.14
7	PIB1	8.08	3.52
8	B4	8.15	2.35
9	AR	8.15	2.30
10	B7	8.17	1.86
11	RW	8.28 (MSE_1)	0
12	P1	8.34	-1.04
13	I1	8.468	-3.29
14	IB2	8.472	-3.36
15	BB	8.49	-3.77
16	PB1	8.64	-6.34
17	P2	8.66	-6.66
18	B6	8.68	-6.94
19	B8	8.69	-7.17
20	B2	8.75	-8.25
21	B9	8.88	-10.5
22	P3	8.89	-10.6
23	B1	8.93	-11.3
24	IB3	9.31	-18.0
25	PIB2	9.57	-22.6
26	PB2	9.64	-23.7
27	PIB3	9.91	-28.4
28	PB3	10.3	-34.7
29	JC	10.7	-41.9

(b) Other forecast series

---	Mean forecast	7.84	7.70
---	GW1	7.04	19.58
---	Ex post best selection	2.56 ($\underline{MSE}$)	100

Table 3: GW ranking

	Average	Average	Number of top 5				
Code	rank	(p_i,q_i)	1st	2nd	3rd	4th	5th
RW	4.4	---	10	15	10	15	12
AR	6.1	2,1	21	16	10	1	7
BB	6.2	0,0	0	3	12	10	15
I3	7.1	0,0	0	0	5	11	13
B4	7.7	2,1	0	11	9	16	9
B3	8.2	0,1	8	10	5	5	5
I2	9.0	1,0	17	2	3	4	0
PI	10.2	0,0	0	0	0	0	0
I1	10.6	2,1	8	7	8	6	0
B5	11.2	0,1	14	9	9	1	7
PB1	13.0	0,1	0	1	3	3	0
B1	14.6	0,1	0	0	0	0	3
IB1	15.46	2,1	3	5	3	1	3
B2	15.55	0,1	0	0	1	4	0
B6	15.7	0,0	0	0	0	0	0
P2	16.08	0,0	0	0	0	0	1
IB2	16.11	0,1	0	0	0	3	0
B9	16.3	0,0	0	0	0	0	3
B8	16.6	0,1	0	0	2	4	0
IB3	17.3	0,1	0	0	0	0	0
PIB1	19.1	0,1	0	3	2	0	2
B7	19.3	0,1	0	0	0	0	2
PB2	20.5	0,1	0	1	2	0	0
P3	21.3	0,0	0	0	0	0	0
PIB3	22.0	0,1	0	0	0	0	0
P1	22.4	0,1	0	0	0	0	0
PB3	23.4	0,1	0	0	0	0	0
PIB2	23.9	0,1	0	0	0	0	0
JC	25.6	1,0	3	1	0	0	1

Notes: (i) Models that appear in bold letters are used in GW1. (ii) The second column reports average GW ranks for 84 periods. (iii) The third column reports the average lag selection results.

Table 4: Lag selection

	BIC				Fix			
		f_i (%)				f_i (%)		
Lag	(p,q)	GW1	GW3	GW29	(p,q)	GW1	GW3	GW29
1	0.14,0.72	7.45	15.36	13.65	1,1	7.57	11.22	11.36
2	0.24,0.68	19.58	18.75	15.33	2,2	11.32	16.70	14.27
3	0.22,0.70	17.23	18.88	15.49	3,3	14.95	20.64	16.33
4	0.33,0.64	16.15	19.74	15.13	4,4	22.59	21.57	16.98

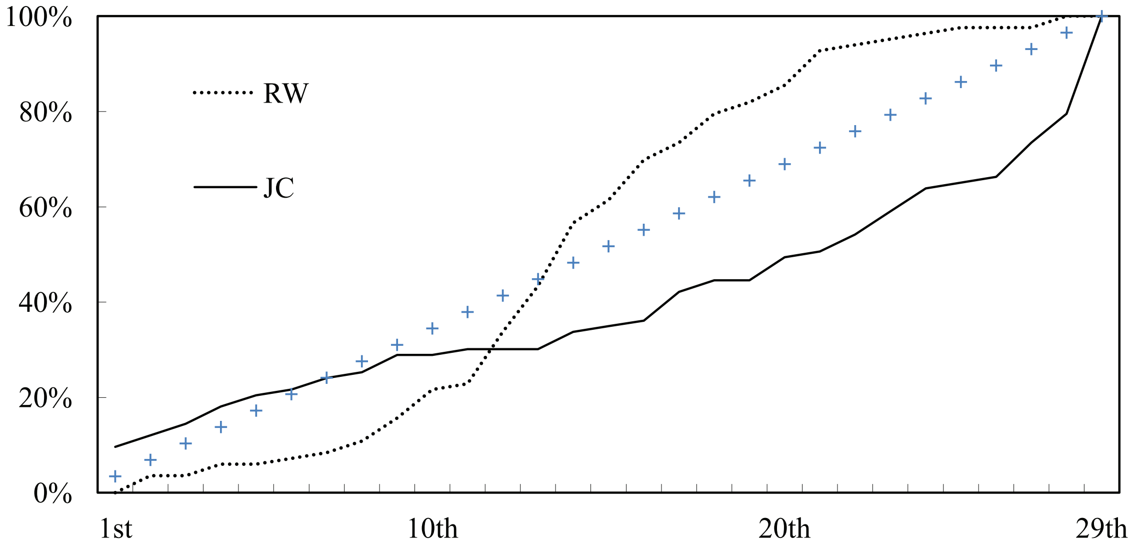
Notes: (i) Lag length (p,q) is an average number for 84 periods and 28 primary models other than RW. (ii) The shaded cells indicate a better performance than the ex post best model. (iii) The benchmark forecast selection method uses BIC, Lag 2, GW1 ($f_i = 19.58$).

Table 5: Average performance (f_i, %) of GW1 in 3,000 Monte Carlo experiments

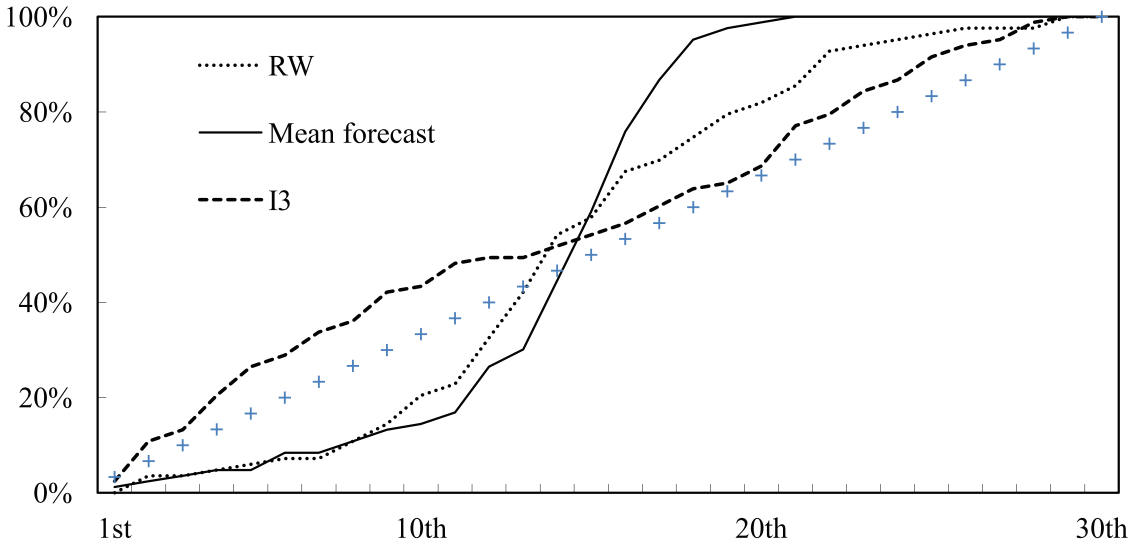
Case 1: (Ex post best model = 11.43)																
	BIC								FIX							
Lag	(p,q)	η = 0	1/6	2/6	3/6	4/6	5/6	1	(p,q)	η = 0	1/6	2/6	3/6	4/6	5/6	1
1	0.50,0.03	9.39	13.46	13.47	13.50	13.45	13.43	13.42	1,1	9.49	13.98	14.23	14.26	14.34	14.35	14.33
2	0.51,0.03	9.28	13.31	13.32	13.33	13.31	13.29	13.31	2,2	8.73	12.49	12.91	13.04	13.02	13.05	13.00
3	0.51,0.03	9.28	13.29	13.31	13.31	13.29	13.28	13.26	3,3	7.98	11.36	11.71	11.84	11.86	11.77	11.69
Case 2: Small noise (Ex post best model = 11.35)																
	BIC								FIX							
Lag	(p,q)	η = 0	1/6	2/6	3/6	4/6	5/6	1	(p,q)	η = 0	1/6	2/6	3/6	4/6	5/6	1
1	0.50,0.04	9.42	13.58	13.59	13.65	13.64	13.63	13.62	1,1	9.35	14.07	14.31	14.45	14.53	14.52	14.53
2	0.51,0.05	9.36	13.41	13.45	13.45	13.46	13.45	13.44	2,2	8.74	12.65	12.91	12.97	12.98	12.97	12.82
3	0.51,0.05	9.28	13.37	13.40	13.38	13.39	13.40	13.39	3,3	7.93	11.41	11.74	11.83	11.85	11.77	11.63
Case 3: Large jump (Ex post best model = 11.50)																
	BIC								FIX							
Lag	(p,q)	η = 0	1/6	2/6	3/6	4/6	5/6	1	(p,q)	η = 0	1/6	2/6	3/6	4/6	5/6	1
1	0.50,0.04	9.31	13.44	13.43	13.44	13.45	13.43	13.42	1,1	9.48	14.01	14.29	14.40	14.36	14.41	14.41
2	0.52,0.04	9.21	13.28	13.31	13.32	13.33	13.31	13.30	2,2	8.78	12.56	12.91	13.04	12.98	12.98	12.91
3	0.66,0.03	9.19	13.16	13.25	13.25	13.28	13.26	13.25	3,3	8.03	11.46	11.90	12.00	12.02	11.94	11.84
Case 4: High persistence (Ex post best model = 11.88)																
	BIC								FIX							
Lag	(p,q)	η = 0	1/6	2/6	3/6	4/6	5/6	1	(p,q)	η = 0	1/6	2/6	3/6	4/6	5/6	1
1	0.66,0.03	15.83	24.446	24.449	24.44	24.43	24.41	24.40	1,1	15.69	24.173	24.172	24.13	24.06	24.03	23.98
2	0.68,0.03	15.81	24.36	24.35	24.33	24.32	24.31	24.29	2,2	15.12	22.98	23.00	22.96	22.88	22.76	22.60
3	0.68,0.03	15.78	24.36	24.35	24.34	24.31	24.31	24.31	3,3	14.44	21.92	21.98	21.89	21.71	21.56	21.35

Notes: (i) Lag length (p,q) is an average number for 9 primary models. (ii) The benchmark forecast selection method uses BIC with Lag 2 and η = 0.5.

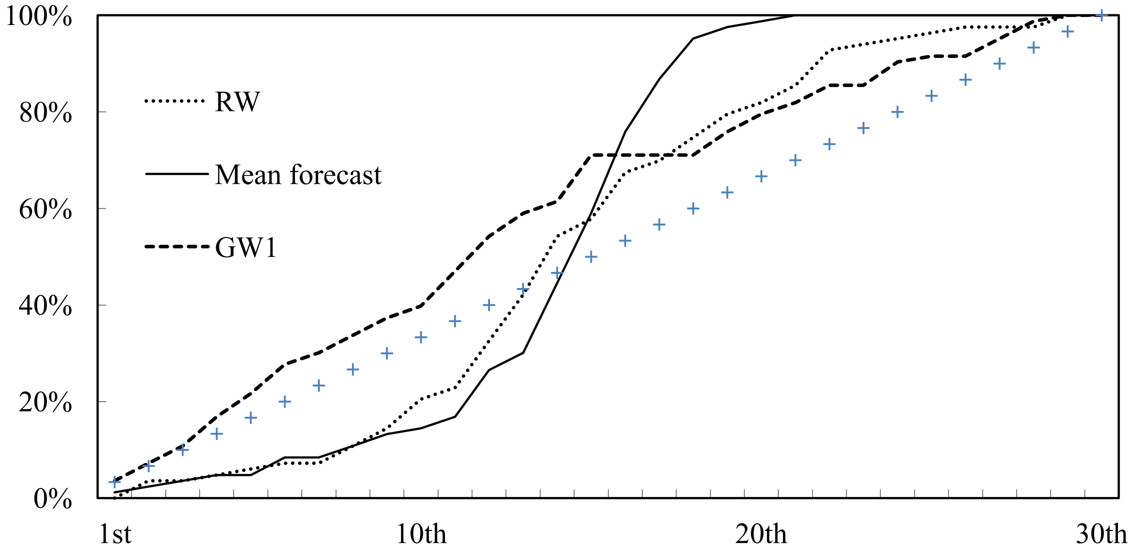
Figure1



(a) RW, JC, I3



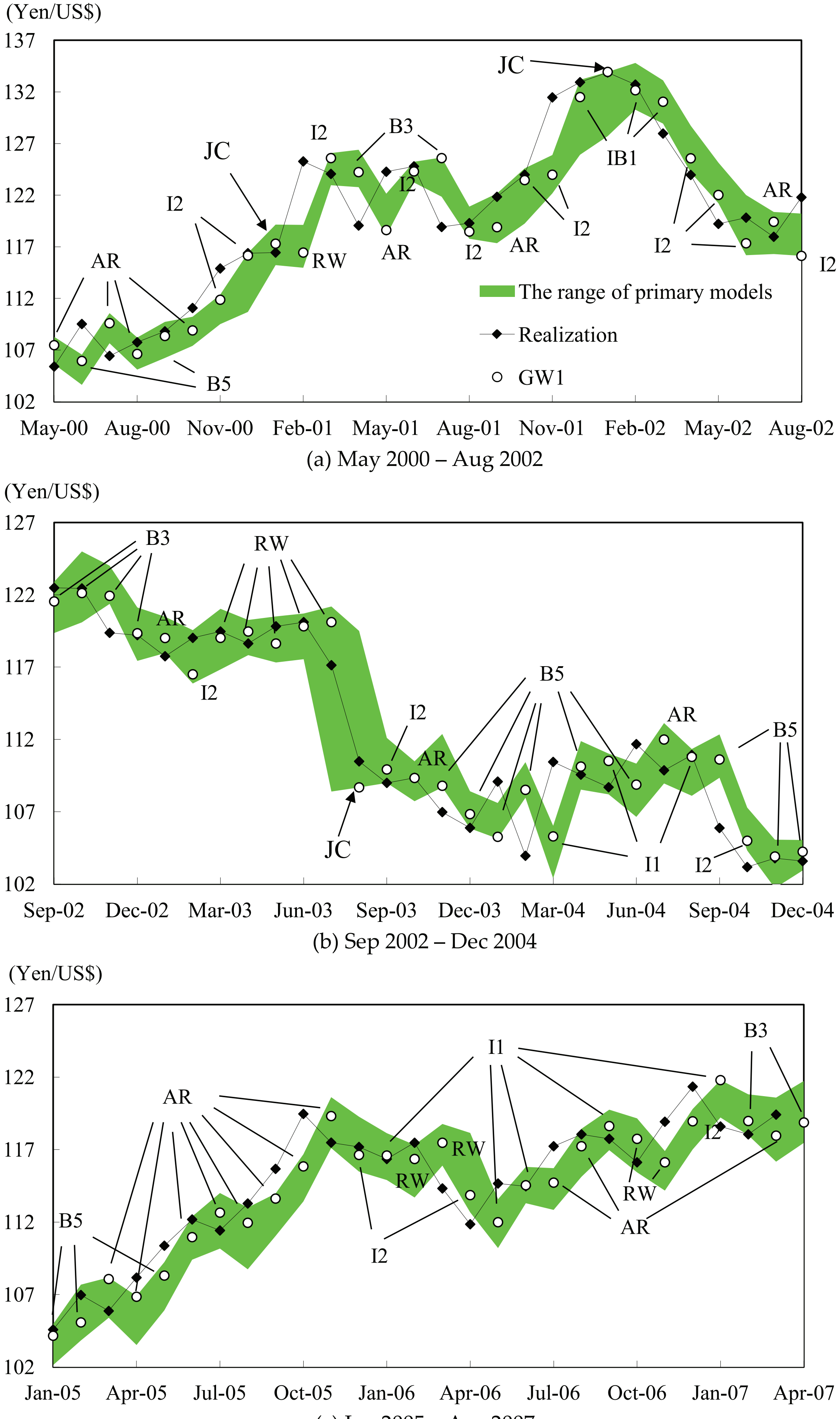
(b) RW, Mean forecast, I3



(c) RW, Mean forecast, GW1

Figure 1: Cumulative distributions of primary ranks

Figure2



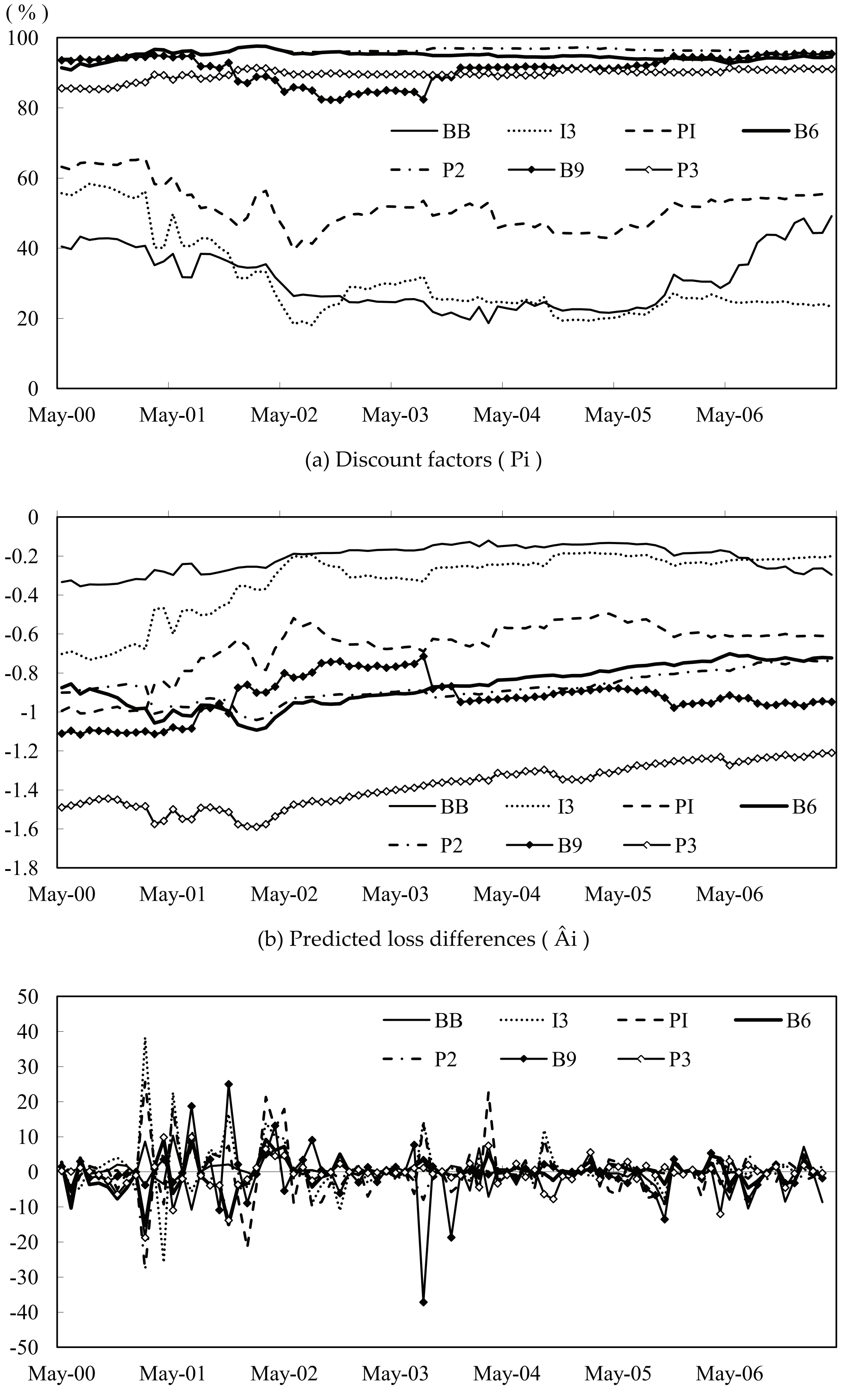
(a) May 2000 – Aug 2002

(b) Sep 2002 – Dec 2004

(c) Jan 2005 – Apr 2007

Figure 2: GW1 forecast series

Figure3



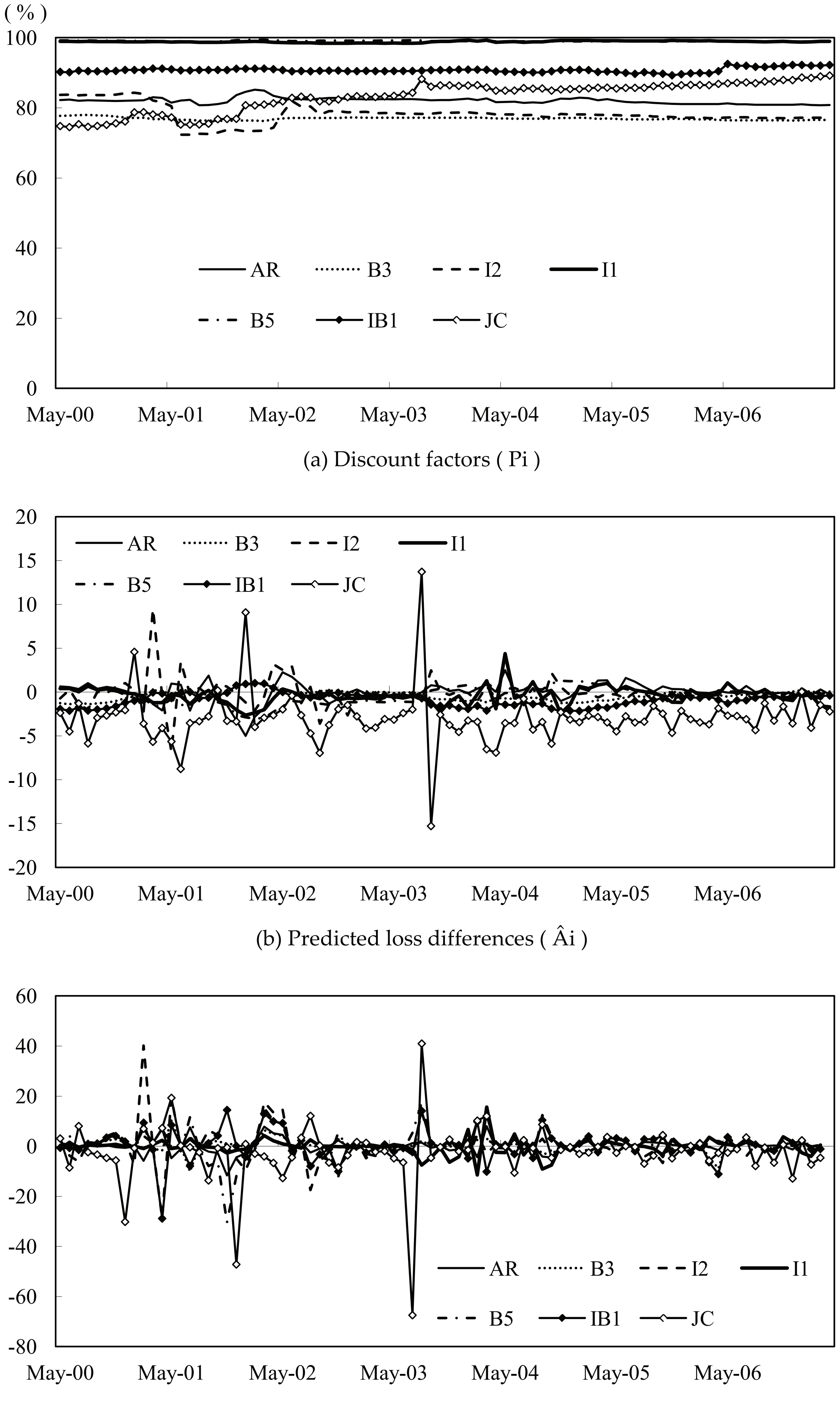
(a) Discount factors (Pi)

(b) Predicted loss differences (Âi)

(c) Realized loss differences (Ai)

Figure 3: (Pi, Ai, Âi) for models with (pi,qi) = (0,0)

Figure4



(a) Discount factors (Pi)

(b) Predicted loss differences (Âi)

(c) Realized loss differences (Ai)

Figure 4: (Pi, Ai, Âi) for models used in GW1

Figure5

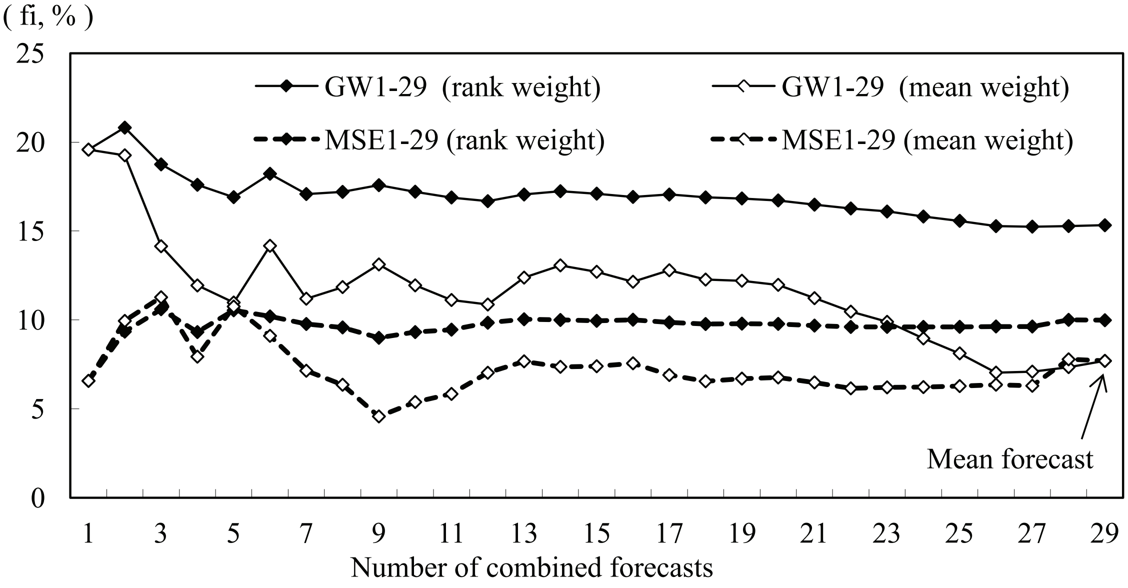
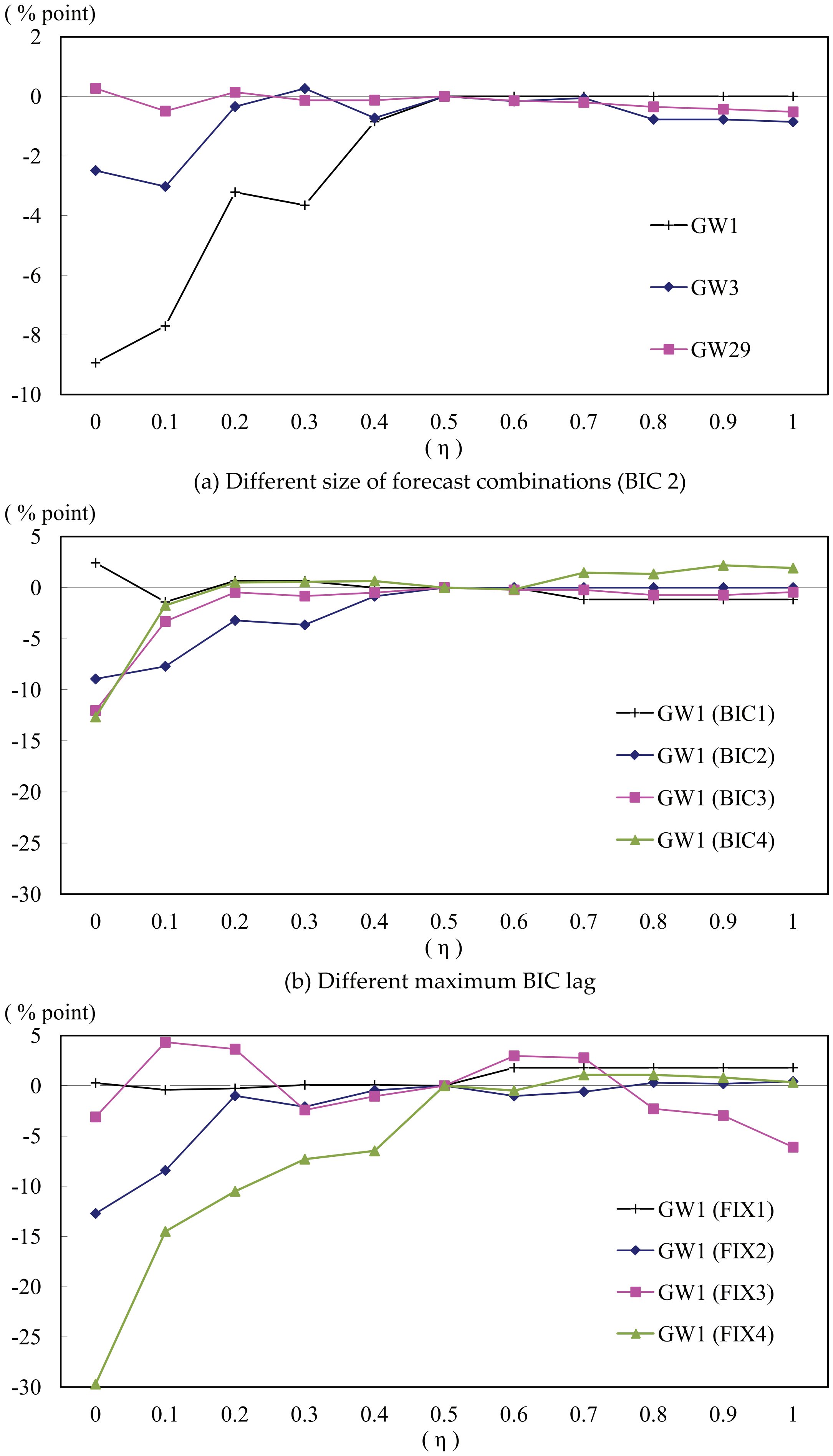


Figure 5: Performances of forecast combinations

Figure6



(a) Different size of forecast combinations (BIC 2)

(b) Different maximum BIC lag

(c) Different fixed lag

Figure 6: Performance difference relative to η = 0.5