The Accuracy of Short-Term Forecast Combinations

Eleonora Granziera, Corinne Luu and Pierre St-Amant, Canadian Economic Analysis

- This article examines whether, and under what circumstances, combining forecasts of real GDP from different models can improve forecast accuracy. It also considers which model-combination methods provide the best performance.
- In line with the previous literature, we find that combining forecasts from different models generally improves forecast accuracy when compared with various benchmarks.
- Unlike several previous studies, we find that assigning equal weights to each model is not always the best weighting scheme. Unequal weighting based on the past forecast performance of models tends to improve accuracy when forecasts across models are substantially different.

In conducting monetary policy, central banks need to regularly assess the current and the future state of the economy. To do this, they combine expert judgment with the results of several models, since no single model can provide the most-accurate results in all circumstances and at all forecasting horizons. For example, while some models do well at forecasting the current period, others do well at forecasting one or two quarters ahead. In addition, with the flow of new data, structural changes in the economy and the introduction of new modelling techniques, the relative usefulness of individual models tends to change over time. Economists at the Bank of Canada therefore regularly update the set of models they use in their current analysis and short-term forecasting.

Uncertainty about the appropriateness of individual models has led researchers to propose using combinations of forecasts from different models, i.e., a diversification strategy, since this strategy may produce forecasts that are less vulnerable to structural breaks and may mitigate the risk that decisions are based on the results of poorly performing models. Indeed, researchers have often found that forecasts generated by combinations of models are more accurate and more robust than those of individual models (Stock and Watson 2004).

This article presents the key findings of a recent project that assessed the potential for various combinations of models to improve the accuracy and robustness of forecasts. The project focused on models for Canadian real gross domestic product (GDP) that the Bank of Canada has used to

backcast (predict the previous quarter before data for that quarter are released by Statistics Canada), nowcast (predict the current period) and forecast over short horizons (typically one or two quarters ahead).[1] We first briefly describe the models and explain how these models were estimated and their forecasts produced. We then explain how forecasts were combined and present the results from those combinations.

Models: Descriptions and Forecasts

To assess the benefits of combining forecasts, this article focuses on a group of simple models, as well as more-complex forecasting tools that the Bank has used to predict quarterly growth in Canadian real GDP, measured at market prices, from Statistics Canada's National Income and Expenditure Accounts. Some of the models in our sample are built to forecast quarterly growth in real GDP over the very short term, while others are designed to produce more accurate forecasts at longer horizons, up to four quarters following the latest release of real GDP (**Table 1**). The individual predictions from these tools are combined to forecast quarter-over-quarter annualized real GDP growth over the short term (i.e., the two quarters following the release by Statistics Canada of the latest quarterly data on real GDP), as well as over slightly longer horizons (i.e., the third or fourth quarter after the latest available data).

To assess and combine the forecasts from the various models, we need to generate predictions from these models in a manner similar to how they would have been generated in practice when forecasting real GDP. The

Table 1: Models of real GDP used in the forecast combinations

Name	Type of model	Forecast horizon[a]	Variables used
State-Space Nowcasting (SSN) Model	Factor model	1–2 quarters	Weekly financial data and monthly data (including total hours worked, monthly real GDP and housing starts)
Bayesian Vector Autoregression (BVAR) Model	Bayesian vector autoregression model	1–4 quarters	Key Canadian and U.S. macroeconomic variables (including U.S. real GDP growth, core inflation and interest rates)
Regional Aggregate Model (RAM)	Univariate models for each region in Canada, aggregated to the national level	1–2 quarters	Provincial-level indicators (e.g., provincial economic accounts, manufacturing sales, employment and retail sales)
Supply-Side Bridge Equation (SSBE) Model	Linear univariate model	1–2 quarters	Wholesale trade, housing starts, interest rates, U.S. retail sales and U.S. personal consumption
Investment-Saving (IS) Curve Models (two models)	Linear univariate models	1–4 quarters	Global output, interest rates, exchange rates and commodity prices. One model includes consumer confidence.
Yield Curve Model (YCM)	Linear univariate model	1–4 quarters	Lagged yield curve (difference between the overnight rate and the rate for 10-year Government of Canada bonds)
Canadian Composite Leading Indicator (CLI)[b]	Linear univariate model	1–2 quarters	Based on the CLI of real activity, composed of indicators of real GDP (e.g., housing index, money supply (M1), TSE 300 stock price index and U.S. Conference Board leading indicator)
Hours Model (HM)	Linear univariate model	1–4 quarters	Based on growth in total hours worked
Autoregressive Distributed Lag (ADL) Model	Linear univariate model	1–4 quarters	Financial variables, composite leading indicator, business credit, employment and U.S. real GDP growth
Narrow Money Model (MM)	Linear univariate model	1–2 quarters	Money supply (M1+)

a. Beyond the latest release of data on real GDP growth

b. Since the project was conducted, Statistics Canada has discontinued publication of the CLI; this model has therefore been modified to incorporate a different, albeit similar, measure of activity.

1 While the Bank still uses some of these models, others have been dropped and new ones have been added.

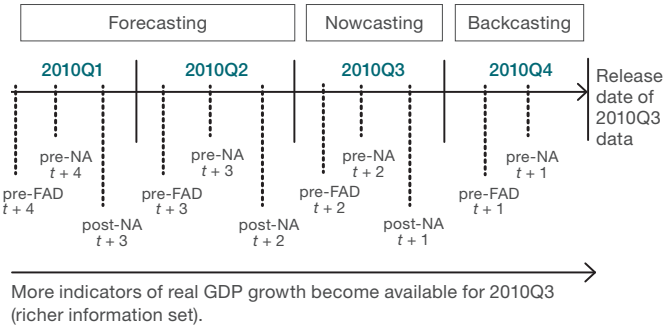
Box 1

Timeline for Real GDP Forecasts

To illustrate the timeline in our analysis, **Figure 1-A** shows the dates when forecasts were made for real GDP growth in the third quarter of 2010. The first forecast (pre-FAD $t + 4$) took place in January 2010. Owing to publication lags for the National Accounts (NA), at this time, quarterly data for real GDP were available only up until 2009Q3; consequently, this prediction for 2010Q3 is considered a four-step-ahead ($t + 4$) forecast. This forecast coincided with a major briefing provided by Bank staff to senior management before the January fixed announcement date (FAD) for a monetary policy decision. The next forecast was produced immediately before the release of real GDP data for 2009Q4 (pre-NA $t + 4$). This second forecast would therefore be at the end of February, which would still be considered a four-step-ahead forecast. Even though no new information regarding quarterly real GDP was released between the January and February forecasts, new weekly and monthly data would have become available, and therefore the forecasts for 2010Q3 may change in some models.

After the release of real GDP data for 2009Q4, the estimation period for the models was extended to include the new data and a new forecast for 2010Q3 (post-NA $t + 3$) was made, which would be a three-step-ahead forecast ($t + 3$). This process was continued until immediately before the actual release of the 2010Q3 real GDP data in November 2010. Eleven forecasts, including those preceding the first FAD, were made for each quarter. The forecast horizons ranged from four steps ahead to one step ahead.

Figure 1-A: Timeline of forecasting the real GDP growth rate in the third quarter of 2010[a]



pre-FAD: forecasts produced about one week before the first fixed announcement date in each quarter

pre-NA: forecasts produced immediately before the release of real GDP (also referred to as the National Accounts)

post-NA: forecasts produced immediately after a real GDP release

a. This is the timeline of the production of forecasts for 2010Q3 in this project. However, the Bank would have made forecasts for 2010Q3 much earlier based on other modelling methods, since this quarter would have been part of the longer-term projection using the Terms-of-Trade Economic Model (ToTEM) (see Coletti and Kozicki (this issue) for information on the projection process).

2011Q2 vintage of National Accounts data[2] is used for all estimations (albeit with a sample that lengthens over time), while the initial observation used to estimate each model varies depending on the model. At each point in time, predictions are produced for up to four quarters beyond the latest release of quarterly real GDP (**Box 1**). The models start to produce forecasts as early as 10.5 months before the actual release of real GDP data for the quarter of interest. A total of 11 forecasts are made: a week before the first fixed announcement date (FAD) in each quarter, as well as immediately before and immediately after the release of National Accounts (NA) data.[3] The forecasts are based on the information available up to that period (initial estimates use data up to 1999) for quarterly real GDP and for the monthly or weekly variables included in the models. When data for a new quarter are available, the models are re-estimated and the forecast cycle is repeated with forecasts produced up to 2011. The models are then evaluated by comparing these forecasts with the actual quarterly growth in real GDP (using the 2011Q2 vintage of data). Since data tend to be revised over time, this exercise should be considered a proxy for forecast accuracy in real time, and the results of this analysis should be interpreted with caution.

2 A vintage is the latest estimate for a given series at a particular time.

3 Some models produce only very short-term forecasts, i.e., the first or second quarter following the latest release of real GDP data (Table 1); they would therefore produce fewer than 11 forecasts for a given quarter.

Forecast Combinations

An initial objective of the forecast combination exercise is to assess whether and under what circumstances competing forecasts may be combined to produce a pooled forecast that performs better than the individual benchmark models. A second objective is to determine whether the relative success of combination methods changes with the forecast horizon.

To evaluate the accuracy of the forecasts (of each model as well as the combined forecasts), we use the root-mean-square prediction error (RMSPE), which measures the discrepancy between the model forecasts and the actual realizations. A lower RMSPE indicates a better performance, or, equivalently, more accurate model forecasts.

To combine forecasts, it is necessary to assign a weight to each model, and the success of the combination may depend on how these weights are assigned. There are several combination schemes with different degrees of sophistication, from simple averaging to complex methods where weights change over time.[4] This article considers the methodologies used most often in the literature, which can be distinguished according to the importance they assign to the past forecast performance of the models. **Box 2** provides a technical description of these combination schemes.

The **simple-average (SA) scheme**, which weighs forecasts equally, has the major advantage of not requiring the use of statistical methods to estimate the weights, since they are determined simply by the number of models.

Box 2

Forecast Combination Schemes

The combined h-step-ahead forecast, y^C_{t+h}, is constructed as a weighted average of the N single-model forecasts:

$$y^C_{t+h} = w^1 y^1_{t+h} + w^2 y^2_{t+h} + \ldots + w^N y^N_{t+h},$$

where $y^i_{t+h}\ i = 1, \ldots N$ is the forecast from model i.

The weights are computed using forecasts up to time t and differ according to the combination scheme:

(i) Simple average: $w^i = 1/N$

(ii) Inverse RMSPE: $w^i = \frac{RMSPE_{i,t+h}{}^{-1}}{\sum_{j=1}^{N} RMSPE_{j,t+h}{}^{-1}}$, where $RMSPE_{i,t+h} = \sqrt{\frac{1}{t}\sum_{\tau=1}^{t}\left(y_{\tau+h} - y^i_{\tau+h}\right)^2}$

(iii) Inverse rank: $w^i = \frac{RANK_{i,t+h}{}^{-1}}{\sum_{j=1}^{N} RANK_{j,t+h}{}^{-1}}$, where $RANK_{i,t+h} = 1$ if model i has the lowest RMSPE up to time t; $RANK_{i,t+h} = 2$ if model i has the second-lowest RMSPE; etc.

(iv) Least squares: weights are the ordinary least-squares coefficients estimated from

$$y_{t+h} = w^1 y^1_{t+h} + w^2 y^2_{t+h} + \ldots + w^N y^N_{t+h} + \varepsilon_{t+h}.$$

4 See Timmermann (2006) for a comprehensive review.

Several studies document the success of the SA scheme relative to more sophisticated weighting schemes, at least when forecasting beyond the current quarter (e.g., Stock and Watson 1999, 2004). The imprecision of statistical methods when estimating weights with short samples of data is cited as the reason for this phenomenon. Theoretically, the SA scheme can be shown to be the optimal combination scheme if the models included in the combination have the same predictive accuracy (as measured by the RMSPE) and the correlations between forecasts from any two models are identical (Smith and Wallis 2009).[5] Intuitively, under these conditions, taking into account the past performance of models and the correlation of forecasts across models does not compensate for the imprecision introduced by estimating the weights.

Other combination schemes weigh the models based on their past performance. In the **inverse-RMSPE (I-RMSPE) scheme**, larger weights are assigned to models that have better forecast accuracy over the forecast sample. Similarly, the **inverse-rank (I-Rank) scheme** assigns to each model a weight inversely proportional to its rank, which is based on the performance of the model over the forecast sample. These schemes do not require the estimation of the correlation between single-model forecasts and do well in practice: for example, a recent Bank of Norway study finds that the I-RMSPE methodology is superior to the simple average (Bjørnland et al. 2012).

Finally, the **least-squares weighting** scheme takes into account not only the past performance of the models but also the correlation of forecasts from different models. The weights assigned in this combination tend to be larger for more accurate forecasts that are less correlated with other forecasts. Although, theoretically, the estimated weights obtained through this method are optimal (Timmermann 2006), they may be biased, especially when the sample size is small.[6] We consider three variants of this scheme: (i) weights are unconstrained (LS); (ii) combined weights sum to one but can take negative values (LSnp); and (iii) combined weights sum to one and must be positive (LSp).

The RMSPE of each weighted combination of forecasts is assessed against that of several benchmarks. The first benchmark is a simple autoregressive (AR) model of the quarterly real GDP growth rate that forecasts future real GDP growth based only on its past values. Although this model is the most commonly used benchmark model in the literature, it is not likely to be very successful at short forecasting horizons, since it does not use higher-frequency information from monthly indicators of economic activity. The second benchmark is an AR model that forecasts quarterly real GDP at market prices based on the monthly series for real GDP at basic prices.[7] The third benchmark is a forecasting strategy in which, at each period, a researcher would select the model that has been most accurate up to that period (best ex ante) and use it to forecast the next quarter.

5 The correlations between forecasts are identical if, for example, the forecasts from model A and model B have a correlation of 0.7 and the forecasts from model A and model C as well as forecasts from model B and model C also have a correlation of 0.7.

6 When the observations available to the researcher are few, the estimates might not be precise, i.e., they might be biased.

7 Basic prices exclude taxes and subsidies on products. The two measures of real GDP are highly correlated on a quarterly basis (at 0.99 since 2007). The benchmark model forecasts the growth rate of real GDP at basic prices on a monthly basis using its own lags. It then aggregates the monthly forecasts to quarterly forecasts. The quarterly growth rate of real GDP at basic prices is taken as a forecast of the quarterly growth rate of real GDP at market prices.

Results

Table 2 shows the RMSPE for forecast horizon $h = 1, \ldots, 4$, for each benchmark model and for the most accurate combination of models. The weights assigned to the models are initially computed on the sample from the 1999Q4–2005Q1 period. The benchmarks and combinations are then evaluated, based on their RMSPEs, over the remaining quarters, from 2005Q2 to 2011Q2.

For each forecasting model and combination, the RMSPE increases with the forecast horizon. Conversely, forecasts become more accurate as the release date approaches, since more information is available.

Combined forecasts are substantially more accurate than the AR benchmark model for real GDP growth at any forecast horizon considered, and the relative performance of the best combination improves as the release date approaches.

The comparison of the best combination and the AR model based on monthly real GDP at basic prices is limited to backcasting and nowcasting, where the relative performance of the quarterly AR model was least successful. Overall, the best combination proves more accurate than the AR monthly model, with relative gains rising as the forecast horizon increases. When backcasting immediately before the release of the quarterly National Accounts, however, the simple benchmark based on monthly GDP data has a slightly lower RMSPE than the combination. This is because two out of the three monthly figures of real GDP at basic prices are available and aggregating them to the quarterly frequency provides a very accurate indicator of real GDP at market prices (see Binette and Chang in this issue).

A comparison of the RMSPE of the best ex ante model with the RMSPE from the best combination indicates that relying on a single forecasting model will typically decrease accuracy: the best model is difficult to identify in advance, and choosing only the model that has performed best up to the time of the forecast produces systematically worse results than combining models.

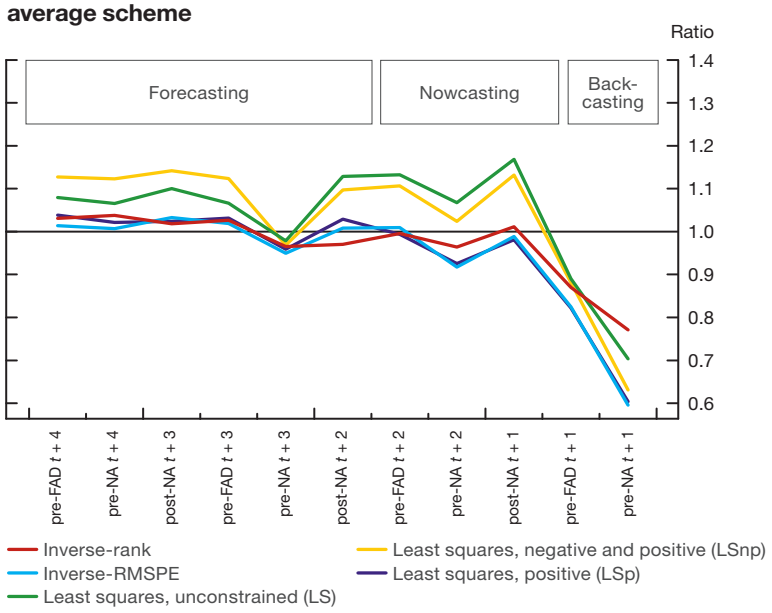
◂ *Choosing only the model that has performed best up to the time of the forecast produces systematically worse results than combining models*

Table 2: Root-mean-square prediction errors (RMSPEs) from benchmark models and combinations, 2005Q2–2011Q2

Horizon (steps ahead)	Timing of forecast	Benchmark			Best combination[a]	
		Autoregressive	Autoregressive monthly	Best ex ante	RMSPE	Scheme
$t + 4$	pre-FAD	3.92		3.14	2.89	SA
$t + 4$	pre-NA	3.92		3.15	2.91	SA
$t + 3$	post-NA	3.92		2.96	2.75	SA
$t + 3$	pre-FAD	3.60		2.89	2.62	SA
$t + 3$	pre-NA	3.60		2.52	2.46	I-RMSPE
$t + 2$	post-NA	3.60		2.38	2.30	I-RANK
$t + 2$	pre-FAD	2.97	4.73	2.31	2.15	I-RANK, LSp
$t + 2$	pre-NA	2.97	3.08	2.12	1.78	I-RMSPE
$t + 1$	post-NA	2.97	1.87	2.19	1.71	LSp
$t + 1$	pre-FAD	2.90	1.42	1.29	1.27	I-RSMPE, LSp
$t + 1$	pre-NA	2.90	0.68	0.73	0.78	I-RMSPE

pre-FAD: forecasts produced about one week before the first fixed announcement date in a quarter
pre-NA: forecasts produced immediately before the release of real GDP (also referred to as the National Accounts)
post-NA: forecasts produced immediately after a real GDP release
a. These two columns show the results from the combination with lowest RMSPE for each horizon.

Chart 1: Relative RMSPEs of performance-based schemes over the simple-average scheme



pre-FAD: forecasts produced about one week before the first fixed announcement date in a quarter
pre-NA: forecasts produced immediately before the release of real GDP (also referred to as the National Accounts)
post-NA: forecasts produced immediately after a real GDP release
Source: Bank of Canada calculations

The performance of different combination schemes is now compared. Because of the documented success of the simple average, the performance-based combination methods are evaluated against the easier-to-implement SA scheme. **Chart 1** shows, for each forecast horizon, the RMSPE of each performance-based combination method relative to the RSMPE of the simple average. A number above (below) one suggests that a performance-based combination method is less (more) accurate than the SA scheme.

To begin, we focus on longer forecast horizons. In line with the previous literature, we find that the SA scheme generally performs well at longer horizons; however, the improvements in accuracy over the other combination schemes are not uniform. The increased accuracy in forecasts using a simple average compared with two of the least-squares variants is substantial (for example, the ratio of the LSnp RMSPE to the SA RMSPE is 1.14), while it is more modest with respect to other methods. This supports the finding that pooling techniques that take into account the correlation of the forecasts (i.e., LS) might be affected by small-sample estimation bias. As **Chart 1** shows, this bias is reduced by imposing the constraint that weights are positive and sum to one (LSp), or by using combination techniques that do not estimate the correlation across forecasts (I-RMSPE or I-Rank). These weighting schemes reduce the uncertainty around the estimates of the weights and can therefore improve the performance of the combination.

Performance-based weights deliver more accurate combinations than equal weights when nowcasting or backcasting

Performance-based weights deliver more accurate combinations than equal weights when nowcasting or backcasting.[8] The improvement is particularly substantial when backcasting, with the relative RMSPE plunging to 0.6 (**Chart 1**). Consistent with the previous discussion, there are significant gains from unequal weighting, because the forecast accuracy of individual models varies greatly at

8 At these shorter horizons, the improvements in accuracy obtained by allowing for performance-based weights more than compensate for the uncertainty introduced by computing the weights in the I-RMSPE or I-Rank schemes.

this horizon (with their RMSPEs ranging from 0.73 to about 2.79) and, among the models, the forecasts tend to be very different (correlation across individual models can be as low as 0.2 and as high as 0.87 for the pre-NA $t + 1$ horizon). In contrast, the optimal weights are close to equal weights for longer horizons, since the model forecasts tend to converge to the mean of real GDP growth and forecasts across models have similar correlation.[9]

Overall, the performance-based schemes that do not take into account the correlations between forecast errors, in particular, the I-RMSPE, are the most robust combination schemes across forecast horizons, since they achieve increased forecast accuracy at shorter horizons and their performance is comparable with the simple average at longer horizons.

The performance-based schemes that do not take into account the correlations between forecast errors are the most robust across forecast horizons

Conclusion

Combining forecasts from several models is more accurate than relying on a single model at all horizons. At longer horizons (three to four quarters ahead), the simple-average scheme outperforms, or does as well as, more sophisticated weighting schemes that take into account the past performance of the models. These results are in line with previous studies.

However, in contrast to much of the existing literature, combined forecasts using performance-based weights significantly increase accuracy at shorter horizons. This result occurs because the models we consider produce very different forecasts at these horizons. Some models are more accurate than others and therefore receive more weight.

Our results support the Bank's approach of using a wide range of models in a flexible manner rather than relying solely on a single model. Although the set of models the Bank uses changes over time, the finding that there are gains from combining forecasts is likely to be an enduring result.

There are certain caveats associated with our work, however. First, because some of the required real-time data were not available, our assessment of model combinations does not take into account the implications of data revisions for forecast accuracy. That is, when simulating the models, we use the 2011Q2 vintage of data rather than the data that were available at the time forecasts were made. Second, the sample we use for estimating models and assessing forecasts is small. The accumulation of progressively longer time series will help to address this limitation in future work.

9 For example, for the pre-FAD $t + 4$ horizon, the difference between the lowest and highest RMSPE is 0.20 and the correlation across model forecasts ranges from 0.87 to 0.99.

Literature Cited

Binette, A. and J. Chang. 2013. "CSI: A Model for Tracking Short-Term Growth in Canadian Real GDP." *Bank of Canada Review* (Summer): 3–12.

Bjørnland, H. C., K. Gerdrup, A. S. Jore, C. Smith and L. A. Thorsrud. 2012. "Does Forecast Combination Improve Norges Bank Inflation Forecasts?" *Oxford Bulletin of Economics and Statistics* 74 (2): 163–79.

Coletti, D. and S. Kozicki. 2013. "Introduction: Tools for Current Analysis at the Bank of Canada." *Bank of Canada Review* (Summer): 1–2.

Smith, J. and K. F. Wallis. 2009. "A Simple Explanation of the Forecast Combination Puzzle." *Oxford Bulletin of Economics and Statistics* 71 (3): 331–55.

Stock, J. H. and M. W. Watson. 1999. "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series." In *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive W. J. Granger*, 1–44. Edited by R. F. Engle and H. White. Oxford: Oxford University Press.

———. 2004. "Combination Forecasts of Output Growth in a Seven-Country Data Set." *Journal of Forecasting* 23: 405–30.

Timmermann, A. 2006. "Forecast Combinations." *Handbook of Economic Forecasting*, 135–96. Edited by G. Elliott, C. W. J. Granger and A. Timmermann. Amsterdam: North Holland.