Evaluating density forecasts: Model combination strategies versus the RBNZ

Chris McDonald Leif Anders Thorsrud*

June 16, 2010

Abstract

Forecasting the future path of the economy is essential for good monetary policy decisions. The recent financial crisis has highlighted the importance of having a good assessment of tail events. The central projection path is not enough. The whole range of outcomes should be forecasted, evaluated and accounted for when making monetary policy decisions. We compare forecasts published by the Reserve Bank of New Zealand to the performance of a suite of statistical models and the combination of these models. Densities used in this analysis have been constructed based on historic forecast performance. Therefore, they are implied density forecasts. Our results reveal that the model density forecasts are comparable in performance and sometimes better than the published forecasts across many different horizons and variables. We also find that the combination strategy performs better than relying on the best model in real time, that is the selection strategy.

*Address for correspondence: Leif Anders Thorsrud, Norges Bank, Postboks 1179, Sentrum, 0107 Oslo, Norway. Tel: +47 98837976. E-Mail: leif-anders.thorsrud@norges-bank.no. Chris McDonald, Reserve Bank of New Zealand, 2 The Terrace, Wellington, New Zealand. Tel: +64 4 471 3634. E-Mail: chris.mcdonald@rbnz.govt.nz. The views in this paper represent those of the authors and are not necessarily those of the Reserve Bank of New Zealand or Norges Bank.

1 Introduction

Economic analysis and forecasts are uncertain for many reasons. The state of the economy may be unclear and the available information set is prone to be revised. There is considerable uncertainty related to the economy's transitions mechanisms and also to the way in which the different economic variables interact with each other. To cope with these uncertainties, policy makers and economic agents lean upon a variety of information, economic theory, judgement and forecasts from econometric and statistical models when making decisions about the future. The recent financial crisis has however highlighted the importance of having not only good point forecasts, but also a good assessment of the likelihood of tail events. Evaluating the central projection path is not sufficient.

In this paper we assess the performance of the Reserve Bank of New Zealand's (RBNZ) forecasts against a model combination approach. The densities used in this analysis have been constructed based on historical forecast errors and assuming normality. They are implied density forecasts. We evaluate the calibration of both the model and published implied density forecasts, and compare different density weighting schemes against each other.[1]

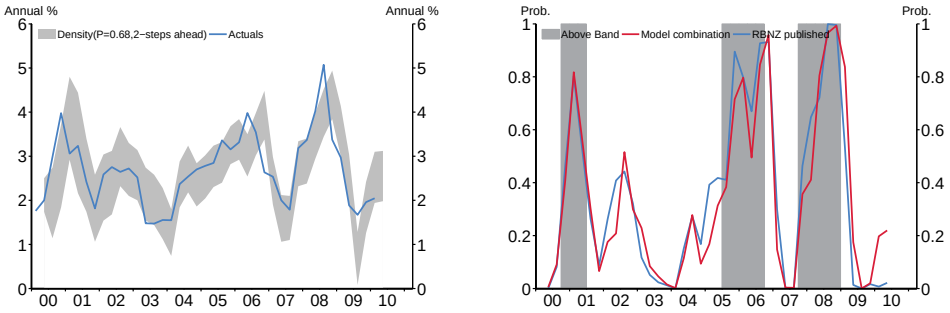
Typically inflation and GDP growth point forecasts have been evaluated in studies similar to this. In addition to focusing on densities, we also broaden the number of variables we evaluate, and assess the forecasting performance for four main macro variables in New Zealand: GDP, inflation, the 90-day interest rate and the exchange rate.

Our model combination approach has some key characteristics: We generate, evaluate, and combine density forecasts based on out-of-sample performance and model weights vary through the evaluation period. The uncertainty is thus time varying. For policy makers or forecasters this is important information since it affects the most likely outcome and the risk of other possible outcomes. As Garratt et al. (2003) writes: ´´In general,

[1]Currently, point forecasts from RBNZ's suite of statistical models are combined using a similar methodology to that described in this paper. These combined forecasts are presented as an alternative and robustness check to the central projection during the forecasting process at the RBNZ. A single combined forecast simplifies the outputs from the statistical models into one set of forecasts. This avoids the issue of focusing too much on any individual model. While this methodology is currently implemented on point forecasts, methods for doing density forecasts and combinations are being developed and discussed in this paper. Note that the published RBNZ forecasts apply judgement to forecasts from a DSGE model that includes an endogenous interest rate track. The statistical models we apply are run without judgement, and are thus purely statistical forecasts.

Figure 1
Historical densities and the probability of inflation above band.

(a) *Historical densities* **(b)** *Probability of inflation above band*



where the loss function underlying the decision problem is non-quadratic and/or one or more of the constraints facing the decision maker are non-linear, the solution to the decision problem invariably involves a comparison of the probability of an event (...) to the cost-benefit ratio of taking a decision."[2]

Figure 1 illustrates the usefulness of the density evaluation approach in this respect. Figure 1a displays the actual annualised inflation rate from 2000 to 2010. The shaded area is the two quarters ahead 68 percent confidence interval forecasts given at each point in time. In figure 1b, the shaded area is the ex-post defined periods when inflation was above the target band. The blue and red lines are the two quarter ahead probability forecasts of such an event for the published RBNZ and the combined model forecasts respectively. Note how the width of the density forecast in figure 1a changes during the evaluation period.

Our choice to use a model combination approach is motivated by at least three factors. Figlewski and Urich (1983), Kang (1986), Diebold and Pauly (1987), Makridakis (1989), Hendry and Clements (2002) and Aiolfi and Timmermann (2006) all note that combining forecasts from models with different degrees of adaptability to structural breaks will outperform forecasts from individual models. Individual forecasting models may be subject to mis-specification bias of unknown form, a point stressed by Clemen (1989), Makridakis (1989), Diebold and Lopez (1995) and Stock

[2]It can of course be debated whether or not the loss function underlying the decision problem in central banks is non-quadratic. May studies suggest it is, see Bjørnland et al. (forthcoming) for some references and further discussion.

and Watson (2004), giving a second argument for combining forecasts. A third argument in favour of combining forecasts is advocated by Timmermann (2006), who notes that the underlying forecasts may be based on different loss functions. If for example any of these loss functions makes the underlying forecasts biased, forecast combinations that apply a more symmetric loss function can purify the forecast from this bias.[3]

Further, knowledge about the forecasting performance is important for policy makers since the statistical models can be considered as separate advisors when monetary policy decisions are made. The density combination approach naturally facilitates this.

Our work resembles work by Romer and Romer (2008) who analysed the usefulness of the Federal Open Market Committee (FOMC) forecasts against the staff forecasts using US data, and Groen et al. (2009) who do a similar exercise evaluating the Bank of England inflation and GDP growth forecasts against a suite of statistical models and a simple combination strategy. Also Adolfson et al. (2007) and Bjørnland et al. (2009) relates to this literature, evaluating the Sveriges Riksbank's and the Norges Bank point forecasts respectively.[4] However, in contrast to these earlier studies, we are interested in the whole distribution of future possible outturns. Again, we believe that assessing the mean or median projections is not enough.

The literature on density combinations is relatively new and unexplored, at least in an economic context. Genest and Zidek (1986) summarise the literature on combinations of densities up to the mid 80s. Clements (2004), Elder et al. (2005), Hall and Mitchell (2007), Eklund and Karlsson (2007), Kascha and Ravazzolo (2010) provide more recent examples of empirical density evaluations. This paper uses the same methodology as outlined in Bjørnland et al. (forthcoming) and Gerdrup et al. (2009), who assessed the relative performance of different density combination and ensemble strategies to more naive approaches.

Our results show that the suite of statistical models is able to generate density forecasts comparable in performance and calibration to densities

[3]There are of course also numerous arguments against using forecast combinations. Diebold and Pauly (1990) and Yang (2004) highlight that estimation errors can seriously contaminate the combination weights, and might therefore be a serious problem for many combination techniques. Palm and Zellner (1992) is only one of many who argue that structural breaks can make it difficult to estimate combination weights that perform well. Lastly, as Timmermann (2006) notes, when the full set of predictor variables used to construct different forecasts is observed by the forecast user, the use of combination strategies instead of attempting to identify a single model can be challenged.

[4]Interestingly enough, many of these studies confirm empirically the theoretical advantages of a model combination approach compared to a model selection strategy.

based on the published RBNZ forecasts. Using the log score as our main scoring criteria, we find that the GDP growth forecasts from the suite of models seem to perform relatively well compared to the published forecasts, while the RBNZ's published inflation forecasts outperform the statistical forecasts on all horizons evaluated. For the 90-day interest rate and exchange rate forecasts the picture is less clear. Further, the combination strategy applied in this paper performs markedly better than the model selection strategy; for all variables, and nearly all horizons.

Using probability integral transforms (PITs) we evaluate the width and bias of the combined density forecasts and the implied published forecast. We find that both forecasts have tended to have a negative bias and, if anything, the densities under-estimate the amount of uncertainty for most variables and horizons. For GDP and the 90-day interest rate, the PITs suggest the combination densities were slightly better estimates of uncertainty, though for inflation and the exchange rate the differences are less clear.

The rest of this paper is organized as follows: In section 2 we describe the individual models, how we derive the individual model weights, and finally how we produce the combined densities. Section 3 outlines the real-time out-of-sample forecasting experiment and our evaluation criteria, while in section 4 we present the results. Section 5 concludes.

2 Model combination

The model combination approach provides the modeller with many possibilities for choosing weights and combination methods. Below we describe how we derive the individual model weights using scoring rules, and also describe how we combine the individual model densities. Finally, the models themselves will be outlined. For details and a more thorough description of possible scoring rules, combination strategies and derivations, see for example Hall and Mitchell (2007) and Timmermann (2006). As already mentioned, our approach follows Bjørnland et al. (forthcoming) closely.

2.1 Deriving the weights

In this application we apply three types of weights: equal weights, logarithmic score (log score) weights and weights based on the continuous ranked probability score (CRPS). These weighting methods are relevant for density forecasts and sufficiently different to give interesting results.

Equal weighting is simply $1/N$, where N is the number of models. These weights are constant, that is, they do not change throughout the evaluation period. The two other weighting schemes are both recursively updated, and thus time varying.

2.1.1 Recursive log score weights

The log score is the logarithm of the probability density function evaluated at the outturn of the forecast. As discussed in Hoeting et al. (1999), the log score is a combined measure of bias and calibration. The preferred densities will thus have probability mass centred on the correct location. Following Hall and Mitchell (2007) we define the log score weights as:

$$w_{i,\tau,h} = \frac{exp[\sum_{\underline{\tau}}^{\tau-h} ln\ g(y_{\tau,h}|I_{i,\tau})]}{\sum_{i=1}^{N} exp[\sum_{\underline{\tau}}^{\tau-h} ln\ g(y_{\tau,h}|I_{i,\tau})]}, \quad \tau = \underline{\tau},...,\overline{\tau} \tag{1}$$

where N is the number of models in total, $\underline{\tau}$ and $\overline{\tau}$ the period over which the weights are derived, and $I_{i,\tau}$ is the information set used by model i to produce the density forecast $g(y_{\tau,h}|I_{i,\tau})$ for variable y. Two things are important to note about this expression. The weights are derived based on out-of-sample performance, and the weights are horizon specific.

Note that maximising the log score is the same as minimising the Kullback-leibler distance between the models and the true but unknown density. Mitchell and Wallis (2008) show the difference in log scores between an "ideal" density and a forecast density, that is the Kullback-Leibler information criterion (KLIC), can be interpreted as a mean error in a similar manner to the use of the mean error or bias in point forecast evaluation.

A perhaps not so satisfactory property of the the logarithmic score is that it involves a harsh penalty for low probability events and therefore is highly sensitive to extreme cases. Other studies have noted similar concerns and considered the use of trimmed means when computing the logarithmic score, for example Gneiting and Raftery (2007). In our application, where the sample size already restricts the analysis, we instead test another scoring rule; the continuous ranked probability score (CRPS).

2.1.2 Recursive CRPS weights

Bjørnland et al. (forthcoming) describes the CRPS as an error measure: if forecasters could correctly anticipate all future events, all the probability mass would be centred on the soon-to-be realised outcome, and the corresponding cumulative density function would be a step function. The

CRPS can be conceptualized as a measure of deviation from this step function. Following Gneiting and Raftery (2007), we define the so called negative orientation of the CRPS as:

$$CRPS_{i,\tau,h} = E_F|Y_{\tau,h|I_{i,\tau}} - y_{\tau,h}| - \frac{1}{2}E_F|Y_{\tau,h|I_{i,\tau}} - Y'_{\tau,h|I_{i,\tau}}|, \tag{2}$$

where Y and $Y^{'}$ are independent copies of the forecast with distribution function F, E_F is the expectation of this distribution, y is the realised value, and i, τ, I and h are defined above.

We compute the CRPS weights using the weighting scheme:

$$w_{i,\tau,h} = \frac{\frac{1}{CRPS_{i,\tau,h}}}{\sum_{i=1}^{N} \frac{1}{CRPS_{i,\tau,h}}} \tag{3}$$

2.2 Combining densities

We use the linear opinion pool to combine the individual densities:

$$p(y_{\tau,h}) = \sum_{i=1}^{N} w_{i,\tau,h}\, g(y_{\tau,h}|I_{i,\tau}), \quad \tau = \underline{\tau}, ..., \overline{\tau} \tag{4}$$

where τ, h, y, N, i and $g(y_{\tau,h}|I_{i,\tau})$ are defined above. The combined density is thus simply a linear combination of the individual densities, where the density combination may be uni-model, skewed and non-normal. Other alternative combination methods do exist, for example the logarithmic opinion pool. However, from a theoretical perspective, no scheme is obviously superior to the other.[5] Our combination strategy is the same as that used in Bjørnland et al. (forthcoming), and is very standard in the literature.

2.3 The individual models

As described in Bloor (2009), at the Reserve Bank of New Zealand, the forecasts underlying policy decisions are formed as part of a rigourous forecasting process. Two main classes of models are used when making these forecasts: Statistical models that exploit statistical patterns in the data, and

[5] Bjørnland et al. (forthcoming) describes the differences between the two methods, and also find evidence that indicates that the so called logarithmic opinion pool might yield better results than the linear opinion pool.

more structural models that draw on economic theory when making predictions for the future. The output from the two model classes provides a basis for incorporating judgement into the final published forecasts.

When combining models in this application, we solely use the output from the statistical models, which can be categorized into six different types: Autoregressive (AR) models, Bayesian vector autoregressive (BVAR) models, Factor models, Indicator models, Factor augmented vector autoregressive (FAVAR) models and Term structure models. This suite of models resembles the suite of models used in the forecasting process at other central banks, for example at the Bank of England (United Kingdom), The Riksbank (Sweden) and at Norges Bank (Norway), see Bjørnland et al. (2009) for an overview. In this application, we only use the output from the statistical models and combine these forecasts into separate, combined forecasts. Our combination strategy is naive in the sense that we do not incorporate any judgement into the forecasting process.[6]

Each of the six different model types may consist of one or more individual models of that type, with either different dynamic specifications or data. Thus, even though our model suite may seem rather limited, with "only" six model types compared to the model suite usually applied in model combination applications, see for example Mitchell et al. (2008), the number of individual models applied is actually much larger. The combination strategy in this paper is therefore actually a two step procedure. The individual models inside each of the six different groups of models are first combined using different in-sample criteria.[7] The combined forecasts from each group are then combined into a single forecast using out of sample evaluation criteria as discussed in section 2.1.

For a full description of the different model groups used at the Reserve Bank of New Zealand as well as the forecasting process, see Bloor (2009). Bloor and Matheson (2009) give a detailed description of the BVAR model, Matheson (2005) documents the factor model as well as the indicator models, Matheson (2007) outlines the FAVAR model, while Krippner and Thorsrud (2009) documents the term structure model.

[6]Developing statistical and econometric models to describe and forecast the behaviour of the economy is however subject to many important decisions that can have a material impact on the output – e.g. forecasts – of the models. Examples of such decisions are the choice of the data set, the choice of the estimations techniques and the dynamic specification of the models.

[7]The BVAR and FAVAR models are exceptions. The BVAR model consists of two models, and both are used in the final combination step.

3 The experiment

The models are evaluated on a horizon and variable specific basis. That is, we first estimate all the models using information up to 1999Q4 and then forecast one to eight quarters ahead. One quarter of information is added to the information set for all the models before the models are re-estimated and another vintage of out-of-sample forecasts are made. This procedure is repeated until we have 37 out-of-sample forecast vintages. The real-time evaluation period runs from 2003Q1 to 2010Q1.

The model weights are derived recursively using the available information set at each point in time. This means that we lose one observation at the beginning of the evaluation period for the one step ahead forecast, two observations for the two step ahead forecast etc. The evaluation sample used to derive the weights grow as more and more vintages become available. This makes the weights vary through time.[8]

Neither the individual models or the RBNZ produce density forecasts directly. As such, all of the individual densities used in this analysis have been constructed based on historical forecast errors and assuming normality. The forecast errors from which the densities are constructed are recursively updated as we move through the evaluation period, following the same structure as described above. The fact that the densities are implied density forecasts is an unsatisfactory feature of our analysis since it does not allow for skewed and possibly multi modal distributions. However, since our method for constructing density forecasts for the individual models is the same for all models, our procedure makes it easier to disentangle the effect of using different weighting criteria. Further, since we are using the linear opinion pool to combine the models, the combined density forecast may very well be both multi-modal and skewed.

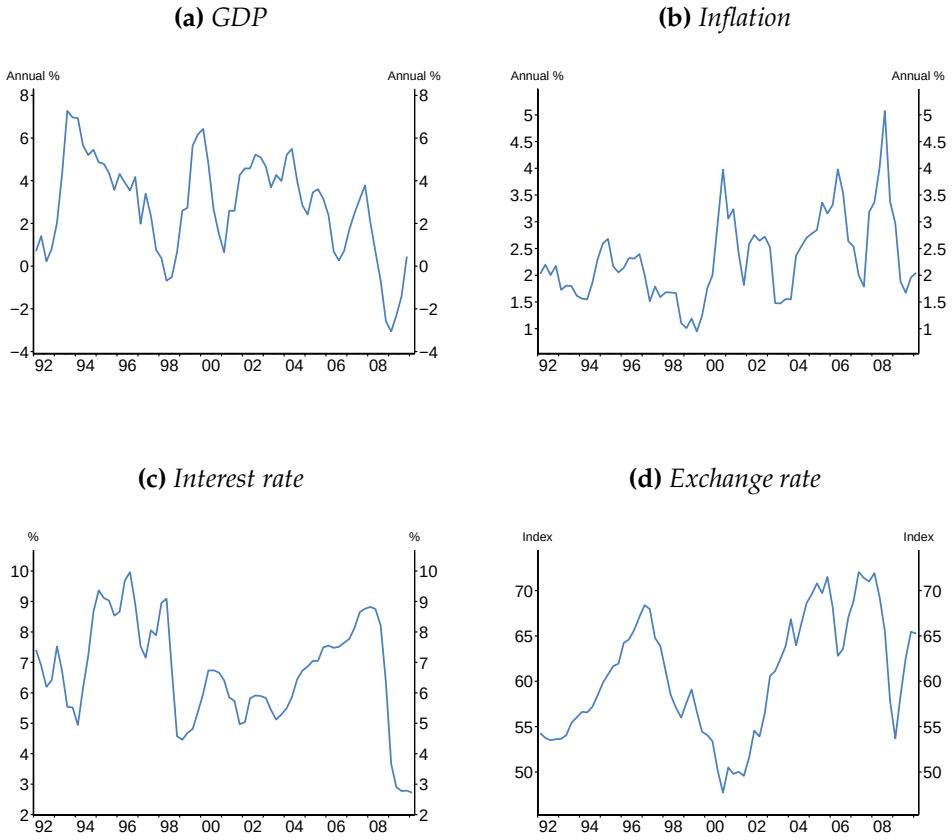
3.1 Data

We forecast and evaluate four variables: GDP, headline inflation, the 90-day interest rate, and the exchange rate[9], see figure 2.

[8] We always use the latest real time vintage to update the weights. Following the real time literature, other vintages or combination of vintages could have been used. We have not explored these possibilities in this analysis.

[9] However, not all models give forecasts for all variables. For example, the term structure model only forecasts economic growth. This model will thus only be included in the model suite when we evaluate the combination strategy for GDP. For some of the vintages, some of the models have not produced real time forecasts. We have replaced these missing observations with forecasts from the BVAR model.

Figure 2
NZ Data: 1992Q1 - 2010Q1



The gross domestic product (GDP) measure we use is total production GDP, seasonally adjusted and in real terms. We get this series from the System of National Accounts. We use the headline consumer price index as our measure of CPI inflation. For both GDP and CPI, we evaluate the forecasts of the annual percent change at each quarterly horizon. These series are released quarterly by Statistics New Zealand. Our measure of the exchange rate is its trade-weighted average value relative to trading partner currencies. For both the exchange rate and the 90-day bank bill interest rate, we take an average over the quarter and evaluate the level forecasts.

All the models are estimated on real time data vintages. Where required, real time forecasts were produced using data from the Reserve Bank of New Zealand's real time database. Real time uncertainty is of course foremost related to real variables published within the National

Accounts. The RMSE of the ultimate net revisions to quarterly real time GDP is for example 0.85 for the vintages spanning the time period 2000Q1 to 2009Q1. That is large relative to other OECD countries.[10]

3.2 Evaluation criteria

In this analysis we are interested in investigating the forecasting performance of model combination forecasts versus the RBNZ published forecasts. Since our focus is on density forecasting, we have chosen to use the log score as our main scoring criteria.[11] The log score is easy and fast to compute, and has some nice interpretations since it can viewed as a density forecast error, as described in section 2.[12]

To help us judge whether the densities are biased in a particular direction, and whether the width of the densities has been roughly correct on average we use PITs. The PITs summarise what we call the calibration of the densities, and are the ex-ante inverse predictive cumulative distribution evaluated at the ex-post actual observations.

4 Results

4.1 Log score performance

Table 1 summarises the log scores of the three different combination strategies described in section 2, the published density forecasts, and a selection strategy.[13] The selection strategy is constructed by ex-ante choosing the best model up to that point in time and using this model to forecast into the future. Note that the selection strategy is also done in real time and is horizon specific. Different models can thus enter into this strategy throughout

[10] See Sleeman (2008) for a full description of how the real time database for New Zealand has been constructed and also for documentation on the relatively big revisions that New Zealand GDP measures undertake compared to other OECD countries.

[11] Other scoring rules do exist, and Gneiting and Raftery (2007) give a nice overview.

[12] However, unlike point forecast evaluation, where the root mean squared forecast error is an often used criteria, a high log score is better than a low log score.

[13] We have also done the same forecasting experiment evaluating only point forecasts and using so called MSE weights, see for example Timmermann (2006). Our results show that the model combination approach using MSE weights performs more or less as good as the published forecasts from the Reserve Bank of New Zealand. Further results can be provided from the authors on request.

the evaluation sample and for different horizons at each point in time.[14]

Looking at table 2a, we see that the best combination strategy for GDP is using the log score weights. Only at the longest forecasting horizons are equal weights and CRPS weights better or equally good. The combination approach also performs better than the published forecasts at all horizons according to the log score criteria. Compared to the selection strategy, the combination strategies are better at almost all forecasting horizons.

Table 2b displays the results for the inflation evaluation. At almost all horizons the published forecasts get a higher log score than any of the combination strategies. However, the different combination strategies still perform just as well or better than the selection strategy at almost all horizons.

The results for the 90-day interest rate forecasts, see table 2c, gives a more dispersed picture than for the previous inflation evaluation. At horizons 1-4 quarters ahead the published forecasts generally gets a higher log score than the combination strategies, while equal weighting performs best at the longer forecasting horizons. The selection strategy only has a better log score than any of the combination strategies at the very first horizons.

Finally, table 2d displays the log score evaluation for the exchange rate forecasts. Different combination strategies outperform the RBNZ forecasts on nearly all horizons. Both equal weights and CRPS weights do as well or better than the RBNZ forecasts. As seen for the other variables we have evaluated, the selection strategy generally gets a lower log score than the combination strategies.

Summarising the results, three main points stand out: The model combination strategy performs better than the selection strategy for most variables at almost all forecasting horizons, and the combination strategy also performs on average just as well as the published forecasts.[15] Further, no combination strategy seems to be dominant. For some variables log score weights are best, for other variables equal weighting or weights derived using the CRPS are better.[16]

[14]Comparing the model combination strategy with the ex-post best individual model is not a reasonable comparison since this strategy uses information that would not have been available in real time.

[15]For GDP and the exchange rate, the combination strategies generally got a better log score while the published forecasts were better for inflation and the interest rate.

[16]See the appendix A for individual model scores.

Table 1

Average log scores. All forecasting horizons.

(a) *GDP: model combinations and published*

	1	2	3	4	5	6	7	8
equal	-1.22	-1.42	-1.63	-1.78	-1.96	-2.06	**-2.14**	-2.07
logScore	**-1.14**	**-1.41**	**-1.61**	**-1.73**	**-1.95**	-2.07	-2.15	**-2.04**
crps	-1.20	-1.42	-1.63	-1.77	-1.96	**-2.05**	-2.15	-2.08
RBNZ	-1.21	-1.44	-1.63	-1.88	-2.07	-2.19	-2.25	-2.17
Selection strategy								
bestLogScore	**-1.11**	-1.43	-1.66	**-1.72**	-2.06	-2.10	-2.24	-2.06

(b) *Inflation: model combinations and published*

	1	2	3	4	5	6	7	8
equal	-0.07	-0.78	-1.07	-1.13	-1.27	**-1.15**	-1.30	-1.31
logScore	0.15	-0.70	-1.08	-1.18	-1.43	-1.26	-1.32	-1.31
crps	-0.01	-0.75	-1.05	-1.11	-1.28	-1.16	-1.29	-1.31
RBNZ	**0.21**	**-0.62**	**-1.00**	**-1.07**	**-1.22**	-1.17	**-1.11**	**-1.18**
Selection strategy								
bestLogScore	0.12	**-0.65**	-1.08	-1.18	-1.54	-1.30	-1.35	-1.31

(c) *Interest rate: model combinations and published*

	1	2	3	4	5	6	7	8
equal	0.96	-0.75	-1.35	-1.63	-1.84	**-1.92**	**-2.10**	**-2.12**
logScore	1.41	-0.59	-1.56	-1.70	-1.99	-2.07	-2.29	-2.24
crps	1.16	-0.66	**-1.31**	-1.62	-1.85	-1.95	-2.13	-2.16
RBNZ	**1.47**	**-0.50**	-1.43	**-1.58**	**-1.83**	-2.00	-2.17	-2.25
Selection strategy								
bestLogScore	**1.44**	**-0.54**	-1.53	-1.64	-1.89	-2.02	-2.43	-2.24

(d) *Exchange rate: model combinations and published*

	1	2	3	4	5	6	7	8
equal	-0.95	-2.51	-2.96	-3.17	**-3.29**	**-3.31**	**-3.36**	**-3.35**
logScore	-0.93	-2.51	-3.00	-3.30	-3.37	-3.34	-3.42	**-3.35**
crps	**-0.83**	-2.51	-2.96	**-3.16**	-3.30	**-3.31**	-3.37	**-3.35**
RBNZ	-0.86	**-2.45**	**-2.94**	-3.25	-3.45	-3.52	-3.58	-3.63
Selection strategy								
bestLogScore	-0.95	**-2.50**	-3.07	-3.41	-3.45	-3.41	-3.54	-3.37

Notes: The columns displays the forecasting horizon, and the rows the weighting strategy. RBNZ refers to the Reserve Bank of New Zealand's published forecasts, while the bestLogScore row is the selection strategy. The log scores are averages over the whole real time sample. A high log score is better than a low log score. The best log score among the combinations strategies and the RBNZ forecasts is marked with bold. bestLogScores are compared to the different combination strategies (and not with RBNZ). If the selection strategy is better than the combination strategies it is marked with bold.

As can clearly be seen in table 1 though, the differences in log scores are usually very small, and too strong a conclusion can and should not be drawn from these results. In addition, our evaluation sample is rather short and includes a dramatic turning point in the economy, due in part to the financial crisis (see figure 2). These facts are of course important for the log score evaluation, especially since the log score weights themselves are so sensitive to outliers. Still, the model combination strategy performs very similarly to the published forecasts, which we think is very encouraging.

4.2 Probability integral transforms

In this section we compare the probability integral transforms (PITs) for the combined density forecast and the implied published density forecast from the Reserve Bank of New Zealand.

The published implied GDP density forecasts, see figure 3b, seem to underestimate the uncertainty. Too many observations end up in the tail of the distributions. Especially evident is the tendency to overestimate the growth on longer horizons, where the outturns have been in the lower end of the density too often. On average across all the forecasting horizons the combined density forecasts (figure 3a) look better calibrated than the published forecasts. However, the combination approach also has a negative bias on the longest forecasting horizons.

Figures 3c and 3d display the PITs for the inflation forecasts. Compared to the PITs evaluation for GDP, the difference between the published and combined inflation forecasts is less striking. There is a tendency for both the published and the combined forecasts to underestimated the inflation pressure, though this is more evident for the combination forecast.

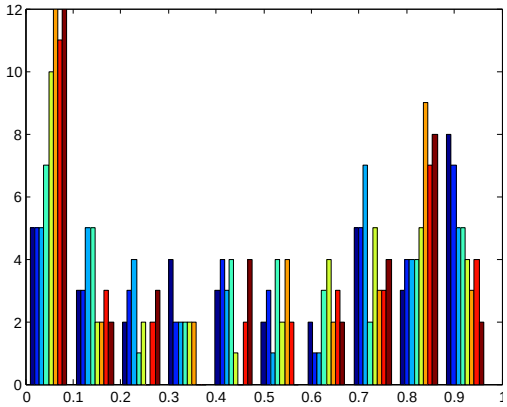
Over the evaluation sample, the RBNZ has overestimated the future path of the 90-day interest rate. Figure 4a shows how the long-run forecasts were too often in the lower tail of the distribution. Further, a marginal U-shape suggests this density was also too narrow. For the combined density forecasts the PITs are more uniform, though a slight upward slope suggests a negative bias.

Figure 4c reveals that the published exchange rate density forecasts significantly underestimate the true uncertainty. Too many observations end up in both the upper and lower end of the forecast densities. The combined density forecasts on the other hand, see figure 4d, have a clear negative bias on nearly all forecasting horizons.

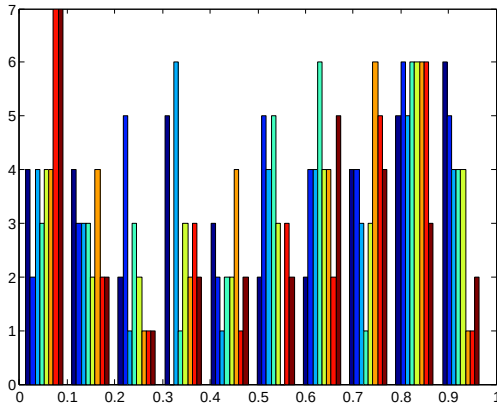
Figure 3

Probability integral transforms. Forecasting horizons one to eight. Each bar colour relates to one horizon. A well specified density should have a uniform distribution.

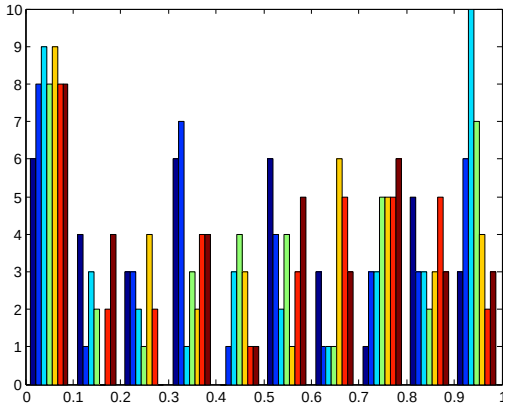
(a) *GDP: RBNZ*



(b) *GDP: Model combination*



(c) *Inflation: RBNZ*



(d) *Inflation: Model combination*

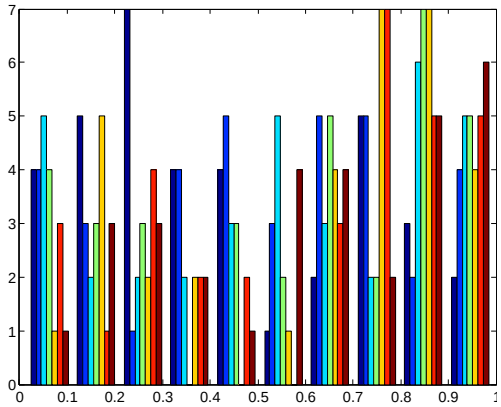
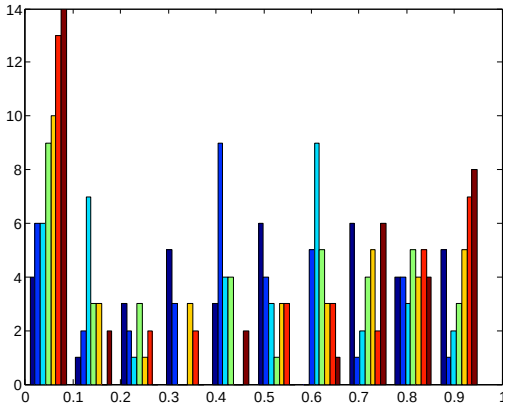


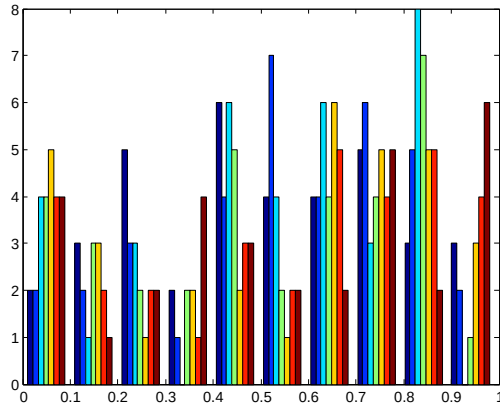
Figure 4

Probability integral transforms. Forecasting horizons one to eight. Each bar colour relates to one horizon. A well specified density should have a uniform distribution.

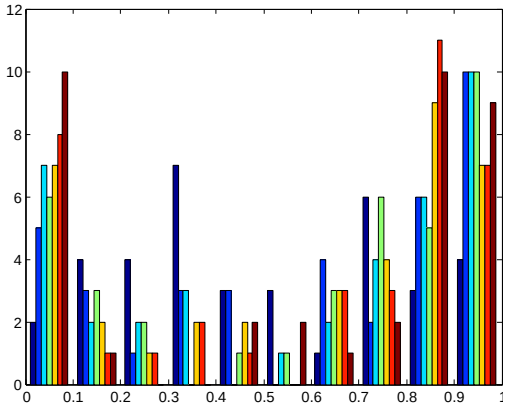
(a) *Interest rate: RBNZ*



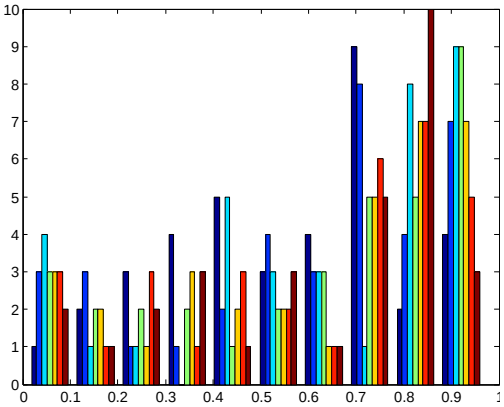
(b) *Interest rate: Model combination*



(c) *Exchange rate: RBNZ*



(d) *Exchange rate: Model combination*



4.2.1 Revisiting the performance of different weighing schemes

The different combination approaches examined in this paper differ markedly in performance across the different variables we are forecasting and compared to the published forecasts. It is important to understand why this is so. We propose two explanations for these differences.

Firstly, none of the individual models get all the weight for any of the variables we are forecasting. Lets assume that we knew the data generating process, $\mathbf{D}$. If one of the models in our model space $\mathbf{M_j} = \mathbf{D}$, this model would receive all the weight as $t \to \infty$ when evaluated using log score weights. This is clearly not the case, as illustrated in figure 5.[17]

Because we do not have the correct data generating process, that is the correct model, or even an obvious and consistent best model for most cases, the model combination approach performs better than the selection strategy. In the few cases where the best model is obvious, the selection strategy and a combination using log score weights tend to perform better than an equally weighted combination.

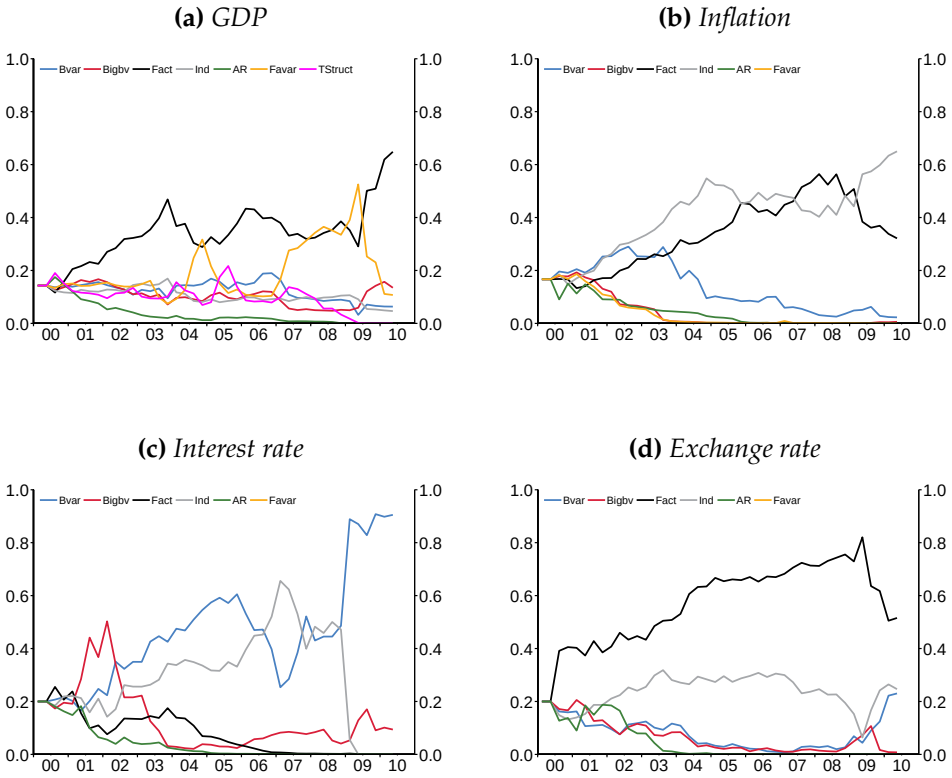
Still, the equal weighting strategy does relatively well for some variables and horizons, see for example table 2d. The difference between the CRPS strategy and the equal weighting strategy is however not large. As noted in section 2, the log score puts a harsh penalty for low probability events and is therefore highly sensitive to extreme cases. The CRPS weights are more forgiving, and thus more similar to the equal weighting strategy. The choice between using log score weights and CRPS weights is probably dependant on the problem at hand, that is the models entering the model suite and the variable being forecasted, as described in Bjørnland et al. (forthcoming).

Finally, the variables clearly differ in how difficult they are to forecast. For example, both inflation and GDP growth have roughly the same mean evaluated over our sample, but GDP is markedly more volatile. The model combination strategy, of course, does not do better than what the underlying model space allows. Compared with the published RBNZ forecasts, the statistical approach taken in this paper can probably be improved further by a careful extension of this model space. An obvious path for further development is to incorporate more high frequency data into the information set used by the models, for example monthly data.[18]

[17] The models getting a higher weight differs, as expected, between which variable is being forecasted.

[18] Krippner and Thorsrud (2009) have documented how important the use of timely data can be in a real time forecast evaluation for New Zealand GDP.

Figure 5
Model weights (two quarters ahead)

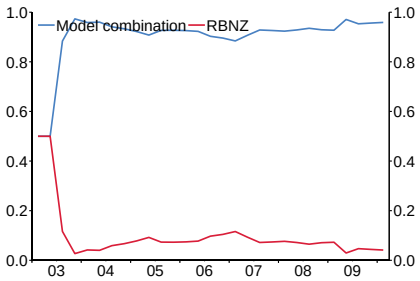


4.3 Weighting combination and published forecasts

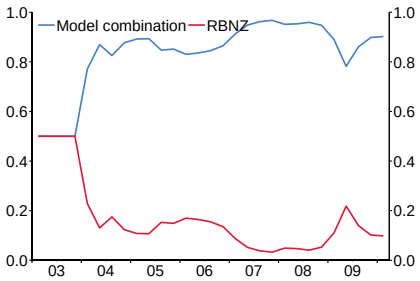
In this section, we evaluate the relative performance of the log score combination density forecasts and the RBNZ's published forecasts through time. We do this by combining these two forecasts, as with the individual models, and then by evaluating the log score weights. A larger weight implies a relatively better performance, and vice versa. An interesting aspect with this exercise is that we track the weights through time and answer the ex-post question: Who should we have trusted, the models or the published forecasts?

Figure 6
Time varying weights

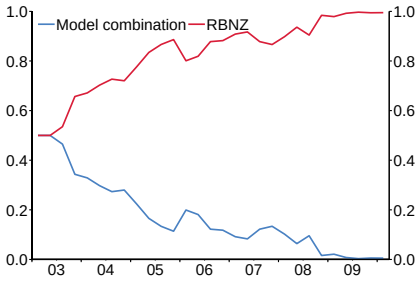
(a) *GDP; h=2*



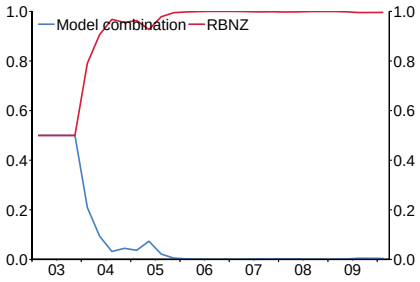
(b) *GDP; h=4*



(c) *Inflation; h=2*



(d) *Inflation; h=4*



As we saw in the previous sections, the combined density forecasts perform relatively well for GDP. As such, the log score weights on the combined forecasts are close to one for both the second and fourth horizons; thus the weights on the published forecasts are nearly zero, see figures 6a and 6b. Furthermore, the combined forecasts have received larger weights than the published forecasts for the majority of the sample period.[19]

Though the model combination does well for GDP, the RBNZ forecasts have tended to outperform the models for CPI inflation. Figures 6c and 6d show that the RBNZ forecasts get nearly all the weight at both the two and four quarter ahead forecasting horizons.

For both the 90-day interest rate and the exchange rate, the published

[19]Equal weights are assigned initially, as the first four-quarter ahead forecast to be evaluated was for 2001Q1. This is by construction due to the relatively short evaluation sample.

forecasts also tend to perform well across many horizons.[20] These results naturally confirm the average log score evaluation we reported in section 4.1.

5 Conclusion

The recent financial crisis has highlighted the importance of having not only good point forecasts, but also a good assessment of tail events. Assessing the mean or median projections is not enough. In this paper we have assessed the performance of the implied density forecasts of the Reserve Bank of New Zealand (RBNZ) against a density combination approach.

Our results reveal that the combined density forecasts from a suite of statistical models are comparable with the published forecasts. The density combination approach performs especially well relative to the published forecasts for GDP, while the RBNZ's published inflation forecasts outperform the combination forecasts on all horizons evaluated. For the 90-day interest rate and exchange rate forecasts the results are less clear. Further, we have shown that the density combination approach performs better than the selection strategy. We have evaluated three different weighting strategies: equal weighting, CRPS weighting and log score weighting. The empirical results do not give any clear indication on which weighting strategy should be preferred. As in Bjørnland et al. (forthcoming), the answer seem to be dependant on the problem at hand; the underlying model space and the properties of the variables being forecasted.

Using probability integral transforms, we show that both the published and model combination forecasts have tended to have a negative bias and, if anything, the densities under-estimate the amount of uncertainty. In addition, the PITs suggest the combination densities were slightly better calibrated for GDP and 90-day interest rate forecasts. While, for inflation and the exchange rate forecasts the differences were less clear.

Our results are hampered by two facts. Firstly, our evaluation sample is rather short. Many of the forecasts and observations are highly affected by the dramatic turning point the New Zealand economy experienced during the financial crisis. Further, the densities we evaluate are, as already noted, derived on past forecasting performance. Since we only have a short history of past forecasting performance available, the earliest densities are constructed using very few observations, and thus may not be representa-

[20] We do not display these graphs.

tive.

This paper has also documented the methodology used by the RBNZ when making statistical model forecasts. Continuously tracking forecasting performance should be an important task for policy makers in central banks and forecasters in general. As more and more real time forecasts become available, the robustness of similar studies to this should increase. Given the setup of the forecasting system at the RBNZ, such an analysis can be conducted in real time.

A Individual model scores

Table 2
Average log scores. All forecasting horizons.

(a) *GDP*

	1	2	3	4	5	6	7	8
Bvar	-1.23	-1.44	-1.59	-1.66	-1.97	-1.98	-2.09	-2.11
Factor	-1.24	-1.38	-1.60	-1.77	-2.03	-2.10	-2.22	-2.17
Indicator	-1.25	-1.44	-1.64	-1.88	-2.10	-2.27	-2.37	-2.33
tstruct	-1.38	-1.74	-1.98	-2.15	-2.14	-2.05	-2.01	-1.96
favar	-1.07	-1.45	-1.64	-1.89	-2.10	-2.14	-2.30	-2.29
Bigbvar	-1.24	-1.37	-1.58	-1.70	-1.89	-2.08	-2.14	-2.09
AR	-1.43	-1.79	-2.01	-2.27	-2.26	-2.33	-2.31	-2.20

(b) *Inflation*

	1	2	3	4	5	6	7	8
Bvar	-0.13	-0.87	-1.15	-1.26	-1.39	-1.28	-1.40	-1.40
Factor	0.13	-0.79	-1.05	-1.16	-1.30	-1.18	-1.54	-1.63
Indicator	0.09	-0.78	-1.04	-1.20	-1.41	-1.43	-1.55	-1.59
favar	-0.73	-1.12	-1.28	-1.37	-1.38	-1.35	-1.28	-1.23
Bigbvar	-0.52	-0.90	-1.17	-1.23	-1.36	-1.22	-1.30	-1.30
AR	-0.67	-1.15	-1.25	-1.40	-1.44	-1.41	-1.41	-1.47

(c) *Interest rate*

	1	2	3	4	5	6	7	8
Bvar	0.78	-0.84	-1.43	-1.73	-1.97	-2.12	-2.25	-2.25
Factor	0.45	-0.96	-1.56	-1.88	-2.01	-2.18	-2.29	-2.40
Indicator	0.44	-0.90	-1.48	-1.84	-2.04	-2.11	-2.18	-2.12
Bigbvar	0.73	-0.77	-1.48	-1.85	-2.05	-2.14	-2.23	-2.35
AR	-0.80	-1.85	-1.93	-1.97	-2.09	-2.19	-2.16	-2.22

(d) *Exchange rate*

	1	2	3	4	5	6	7	8
Bvar	-0.98	-2.51	-2.96	-3.22	-3.39	-3.39	-3.42	-3.46
Factor	-1.27	-2.55	-3.09	-3.30	-3.37	-3.36	-3.37	-3.34
Indicator	-1.27	-2.59	-3.04	-3.26	-3.32	-3.34	-3.39	-3.42
Bigbvar	-1.83	-2.60	-3.04	-3.20	-3.30	-3.30	-3.40	-3.39
AR	-2.51	-3.00	-3.23	-3.36	-3.39	-3.44	-3.45	-3.38

Notes: The columns displays the forecasting horizon, and the rows the weighting strategy. The log scores are averages over the whole real time sample. A high log score is better than a low log score. See section 2.3 for a description of the different individual models.

References

Malin Adolfson, Michael K. Andersson, Jesper Lindé, Mattias Villani, and Anders Vredin. Modern forecasting models in action: Improving macroeconomic analyses at central banks. *International Journal of Central Banking*, 3(4):111–144, December 2007.

Marco Aiolfi and Allan Timmermann. Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics*, 135(1-2):31–53, 2006.

Hilde C. Bjørnland, Karsten Gerdrup, Anne Sofie Jore, Christie Smith, and Leif Anders Thorsrud. Does forecast combination improve Norges Bank inflation forecasts? Working Paper 2009/01, Norges Bank, Jan 2009.

Hilde C. Bjørnland, Karsten Gerdrup, Anne Sofie Jore, Christie Smith, and Leif Anders Thorsrud. Weights and Pools for a Norwegian Density Combination. *North American Journal of Economic and Finance*, forthcoming.

Chris Bloor. The use of statistical forecasting models at the Reserve Bank of New Zealand. *Reserve Bank of New Zealand: Bulletin*, 72(2), June 2009.

Chris Bloor and Troy Matheson. Real-time conditional forecasts with bayesian vars: An application to New Zealand. Reserve Bank of New Zealand Discussion Paper Series DP2009/02, Reserve Bank of New Zealand, April 2009.

R. T. Clemen. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5, 1989.

M. P. Clements. Evaluating the Bank of England density forecasts of inflation. *Economic Journal*, 114, 2004.

Francis X. Diebold and Jose A. Lopez. Forecast evaluation and combination. Technical report, 1995.

Francis X. Diebold and Peter Pauly. Structural change and the combination of forecasts. *Journal of Forecasting*, 6:21–40, 1987.

Francis X. Diebold and Peter Pauly. The use of prior information in forecast combination. *International Journal of Forecasting*, 6(4):503–08, December 1990.

Jana Eklund and Sune Karlsson. Forecast combination and model averaging using predictive measures. *Econometric Reviews*, 26(2–4):329–363, 2007.

R. Elder, G. Kapetanios, T. Taylor, and T. Yates. Assessing the MPC's fan charts. *Bank of England Quarterly Bulletin*, 45(3):326–348, 2005.

S. Figlewski and S. Urich. Optimal aggregation of money supply forecasts: Accuracy, profitability and market efficiency. *Journal of Finance*, (28):695–710, 1983.

Anthony Garratt, Kevin Lee, M. Hashem Pesaran, and Yongcheol Shin. Forecast uncertainties in macroeconomic modeling: An application to the UK economy. *Journal of the American Statistical Association*, 98(464): 829–38, December 2003.

Christian Genest and James V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1): 114–148, 1986.

Karsten R. Gerdrup, Anne Sofie Jore, Christie Smith, and Leif Anders Thorsrud. Evaluating ensemble density combination - forecasting GDP and inflation. Working Paper 2009/19, Norges Bank, November 2009.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, March 2007.

Jan J.J. Groen, George Kapetanios, and Simon Price. A real time evaluation of Bank of England forecasts of inflation and growth. *International Journal of Forecasting*, 25(1):74–80, 2009.

Stephen G. Hall and James Mitchell. Combining density forecasts. *International Journal of Forecasting*, 23(1):1–13, 2007.

David F. Hendry and Michael P. Clements. Pooling of forecasts. *Econometrics Journal*, 5:1–26, 2002.

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

H. Kang. Unstable weights in the combination of forecasts. *Management Science*, 32:683–695, 1986.

Christian Kascha and Francesco Ravazzolo. Combining inflation density forecasts. *Journal of Forecasting*, 29(1-2):231–250, 2010.

Leo Krippner and Leif Anders Thorsrud. Forecasting New Zealand's economic growth using yield curve information. Discussion Paper DP2009/18, Reserve Bank of New Zealand, 2009.

Spyros Makridakis. Why combining works. *International Journal of Forecasting*, 5:601–603, 1989.

Troy Matheson. Factor model forecasts for New Zealand. Discussion Paper DP2005/1, 2005.

Troy Matheson. An analysis of the informational content of New Zealand data releases: the importance of business opinion surveys. Reserve Bank of New Zealand Discussion Paper Series DP2007/13, Reserve Bank of New Zealand, September 2007.

James Mitchell and K.F. Wallis. Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness. NIESR Discussion Papers 320, National Institute of Economic and Social Research, August 2008.

James Mitchell, A. S. Jore, and S. P. Vahey. Combining forecast densities from VARs with uncertain instabilities. NIESR Discussion Papers 303, National Institute of Economic and Social Research, January 2008.

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

Christina D. Romer and David H. Romer. The FOMC versus the staff: Where can monetary policymakers add value? 98(2):230–35, May 2008.

James H Stock and Mark W Watson. Combining forecasts of output growth in seven-country data set. *Journal of Forecasting*, 23:405–430, 2004.

A. Timmermann. *Forecast combinations*, volume 1. Elsevier Science B.V., 2006.

Y. Yang. Combining forecasts procedures: Some theoretical results. *Econometric Theory*, 20:176–190, 2004.