DP2011/03

Evaluating density forecasts: Model combination strategies versus the RBNZ

Chris McDonald and Leif Anders Thorsrud

August 2011

JEL classification: C52 C53 E52

www.rbnz.govt.nz/research/discusspapers/

Discussion Paper Series

ISSN 1177-7567

DP2011/03

Evaluating density forecasts: Model combination strategies versus the RBNZ*

Chris McDonald and Leif Anders Thorsrud†

Abstract

Forecasting the future path of the economy is essential for good monetary policy decisions. The recent financial crisis has highlighted the importance of tail events, and that assessing the central projection is not enough. The whole range of outcomes should be forecasted, evaluated and accounted for when making monetary policy decisions. As such, we construct density forecasts using the historical performance of the Reserve Bank of New Zealand's (RBNZ) published point forecasts. We compare these implied RBNZ densities to similarly constructed densities from a suite of empirical models. In particular, we compare the implied RBNZ densities to combinations of density forecasts from the models. Our results reveal that the combined densities are comparable in performance and sometimes better than the implied RBNZ densities across many different horizons and variables. We also find that the combination strategies typically perform better than relying on the best model in real-time, that is the selection strategy.

* The Reserve Bank of New Zealand's discussion paper series is externally refereed. The views expressed in this paper are those of the author(s) and do not necessarily reflect the views of the Reserve Bank of New Zealand. We would like to thank colleagues at the Reserve Bank of New Zealand for helpful comments. Also, we are grateful for input from those at the New Zealand Association of Economists (NZAE) conference in 2010 and at the New Zealand Econometric Study Group meeting in January 2011. Two anonymous referees gave comments that greatly improved this paper.

† Address for correspondence: Leif Anders Thorsrud, Norwegian Business School (BI), 0442 Oslo, Norway. Tel: +47 98837976. *Email address:* leif.a.thorsrud@bi.no. Chris McDonald, Reserve Bank of New Zealand, 2 The Terrace, Wellington, New Zealand. Tel: +64 4 471 3634. *Email address:* chris.mcdonald@rbnz.govt.nz.

ISSN 1177-7567

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 transmission mechanisms and also to the way in which 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 highlighted the importance of having not only good point forecasts, but also a good assessment of the whole range of possible outcomes.

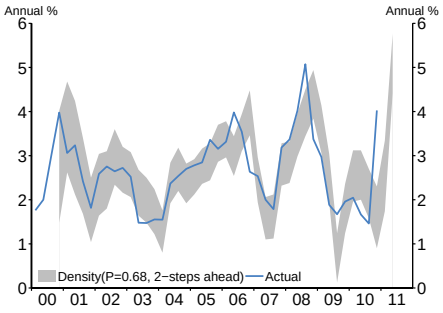
In this paper, we assess the performance of the Reserve Bank of New Zealand's (RBNZ) forecasts against a model combination approach. The densities used in this analysis have been constructed based on historical forecast errors and assuming normality, hence the forecasts are implied density forecasts. We evaluate the calibration of both a suite of models and the implied RBNZ density forecasts, and compare different density weighting schemes against each other.[1] In particular, we use the probability integral transform (PIT) to evaluate calibration, and derive recursive model weights using log scores and continuous ranked probability scores. The forecasting performance, in terms of highest log score, for four main macro variables is evaluated: 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. Thus, the uncertainty can change over time. For policy makers or forecasters, the amount of uncertainty is important since it affects the weight one puts on the most likely outcome and the risk of other possible outcomes. As Garratt et al (2003) states: "In general, 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

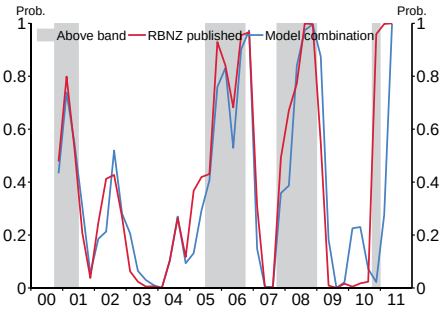
[1] Currently, point forecasts from the RBNZ's suite of empirical models are combined using a similar methodology to that described in this paper. These combined forecasts are presented as an alternative and robustness check to the central projection during the forecasting process at the RBNZ. Note that the published RBNZ forecasts apply judgement to forecasts from a model that includes an endogenous interest rate track. The empirical models we apply are run without judgement, and are thus purely statistical forecasts.

Figure 1: Historical densities

(a) Historical densities



(b) Probability of inflation above band



Notes: Figure 1a displays the historical inflation density forecast 2-steps ahead together with the outcome. Figure 1b displays the forecasted probability of inflation being above the target band (above three percent), 2-steps ahead. Grey areas show the quarters when the annual percent change of inflation was above this band.

ratio of taking a decision."

Figure 1 illustrates the usefulness of density forecasts with respect to understanding the uncertainty and probability of different outcomes. Figure 1a displays the actual annual inflation rate from 2000 to 2010 in addition to the combined density forecast from the models. The shaded area is the two quarters ahead 68 percent confidence interval forecasts given at each point in time.[2] As noted above, an important component of our model combination approach is that we allow for the forecast uncertainty to change over time. That is, the width of the density forecasts change over the evaluation period. 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 quarters ahead probability forecasts of such an event for the RBNZ and the combined model forecasts respectively. Generally, the forecast probabilities of inflation being above the band implied by RBNZ and by the model combination are similar. Given the fact that the policymakers had access to the model based forecasts in real-time, this is perhaps not so surprising.[3]

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 Tim-

[2] In this example, we have used log-score weights to combine the models. See section 2.1 for details.

[3] In section 4.5, we return to this discussion, and especially evaluate the forecasted probability of the 2008/09 decline in GDP.

mermann (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 and Watson (2004), giving a second argument for combining forecasts. A third argument in favor of combining forecasts is advocated by Timmermann (2006), who notes that the underlying forecasts may be based on different loss functions. If two forecasts produced under different asymmetric loss functions are combined, the combination may suit someone with a more symmetric loss function.[4]

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 authenticates 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.[5] 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

[4] There are of course 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.

[5] Many of these studies confirm empirically the theoretical advantages of a model combination approach compared to a model selection strategy.

Bjørnland et al (2011), which very much build on the findings in Jore et al (2010) and Hall and Mitchell (2007).[6]

Our results show that the suite of empirical models is able to generate density forecasts comparable in performance and calibration to densities based on the published RBNZ forecasts. 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. We generally find the RBNZ density forecasts do quite well for shorter horizons. Whereas the combination does especially well at longer horizons. The model combination strategy performs better than the model selection strategy for most variables and horizons, while equal and CRPS weighting methods tend to outperform the more volatile log score weights. Lastly, the PIT evaluation show that both the combined density forecasts and the implied published RBNZ forecasts have tended to have a negative bias for CPI and the exchange rate, while the other density forecasts seem to be well calibrated.

The rest of this paper is organized as follows: In section 2 we describe how we derive the individual model weights, how we produce the combined densities, and give a short description of the individual models. Section 3 outlines the real-time out-of-sample forecasting experiment and our evaluation criteria. We also outline how we produce density forecasts for each model given their historical forecast accuracy. In section 4 we present the results, while we in section 5 conduct some robustness checks with respect to how we have computed the individual model densities. Section 6 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 Hall and Mitchell (2007) and Timmermann (2006). As already mentioned, our approach follows Bjørnland et al (2011) closely.

[6] Further references of the use of density combination methods in economics and other sciences can be found in these papers.

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.

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:

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

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).

Recursive CRPS weights

Bjørnland et al (2011) 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 as 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.

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: empirical models that exploit statistical patterns in the data, and 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 models, which can be categorised into seven different types: Autoregressive (AR) models, vector autoregressive (VAR) 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 the Norges Bank (Norway), see Bjørnland et al (2009) for an overview. In this application, we use the output from the empirical 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.[7]

Each of the seven different model types may consist of one or more individual model of that type, with either different dynamic specifications or data. Thus, even though our model suite may seem rather limited, with "only" seven model types, compared to the model suite usually applied in model combination applications, the number of individual models applied is actually much larger.[8] The combination strategy in this paper is therefore actually a two step procedure. The individual models inside each of the seven different groups of models are first combined using different in-sample criteria.[9] The combined forecasts from each group are then combined into a single forecast using out of sample evaluation criteria, discussed in section 2.1.[10]

[7] Developing empirical models to forecast 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 estimation techniques and the dynamic specification of the models.

[8] As stressed by for example Jore et al (2010), the number of models in the model space can have very important implications for the properties of the combined forecast. Even though our model suite actually contains much more than seven models, the fact that we only combine at most seven different models will considerably restrict the ability of the combined density to approximate non-linear, non-Gaussian processes.

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

[10] A similar type of two step-procedure has been used previously by Garratt et al (2009), and called "grand ensemble". The paper by Bache et al (2009) explains the use of the ensemble terminology in more detail.

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 (2006) documents the factor model as well as the indicator models, Matheson (2007) outlines the FAVAR model, while Krippner and Thorsrud (2009) document the term structure model.

3 The experiment

The stylized recursive real-time forecasting experiment runs like this: All the models are estimated using information up to 1999Q4 and then forecast one to eight quarters ahead. One quarter of information is added to the information set, the models are re-estimated, and another vintage of out-of-sample forecasts are made. This procedure is repeated until 43 out-of-sample forecast vintages are generated. The real-time evaluation period runs from 2000Q1 to 2010Q3.[11]

The models are evaluated and weighted on a horizon and variable specific basis. The model weights are derived recursively using the available information set at each point in time. We use equal weights at the beginning of the evaluation period until forecasts can be scored, this affects one observation 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.[12]

3.1 Density forecasts

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. The forecast errors from which

[11] It should be noted that we have not recursively run all the models in the RBNZ model suite, but collected the forecasts from the RBNZ real-time forecasting data base. Thus, we cannot guarantee that none of the models have had slight modifications etc. during the evaluation period. Further, as discussed in section 3.1, because the RBNZ models have not been set up to produce density forecasts, the real-time data base do not contain densities.

[12] 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.

the densities are constructed are recursively updated as we move through the evaluation period, following the same structure as described above. Similar methods to this have been used by the Riksbank, the Bank of England and the National Institute of Economic and Social Research, as discussed in e.g. Wallis (1989) and Hall and Mitchell (2007).

In our main analysis, we have assumed that the historical errors are normally distributed. This can of course be a spurious assumption, which might have a material impact on our results. Further, some sort of training period is needed to determine the (initial) variance of the forecast errors. These choices impact on the specification of the derived forecast densities. In section 5.1 and 5.2 we thus investigate if the normality assumption is appropriate, and how different training periods might affect our results.

Clearly, the fact that the densities are implied density forecasts is an unsatisfactory feature of our analysis. Not only might the results be affected by the choices regarding the appropriate distribution and training period, but it does not allow for skewed and possibly multi modal distributions either. 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. Also, 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.

3.2 Data

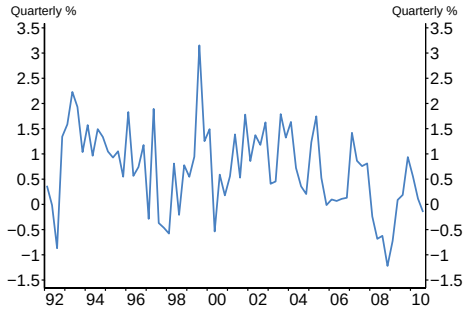
We forecast and evaluate four variables: GDP, headline inflation, the 90-day interest rate, and the exchange rate, see figure 2.[13]

For GDP we use total production GDP, seasonally adjusted and in real terms. The series is collected from the System of National Accounts. We use the headline consumer price index (CPI) as our measure of inflation. These series are released quarterly by Statistics New Zealand. Our measure of the exchange rate is the trade-weighted average value relative to trading partner currencies, while we use the 90-day bank bill interest rate as our measure of the interest rate. Both series are provided by the Reserve Bank of New Zealand. At each quarterly horizon, we evaluate the annual CPI inflation

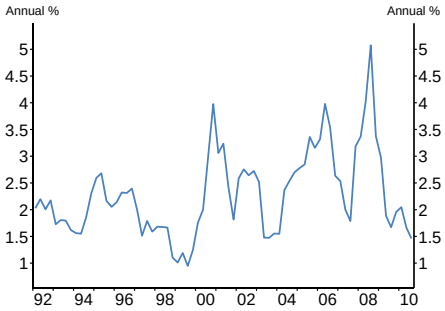
[13] Not all models forecast all the variables. For example, the term structure model only forecasts GDP. For some of the vintages, real-time forecasts have not been produced for some of the models. We have replaced these missing forecasts using forecasts from the BVAR model.

Figure 2: New Zealand data: 1992Q1 - 2010Q3

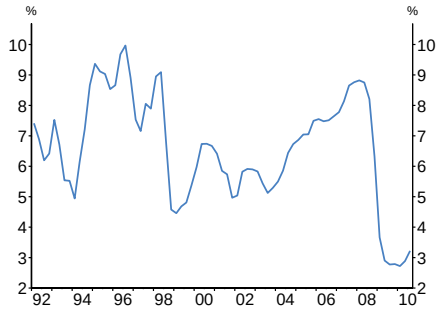
(a) Quarterly GDP growth



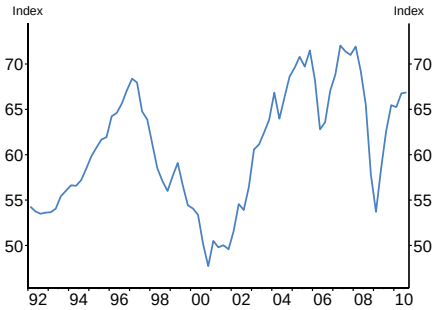
(b) Annual inflation



(c) Interest rate



(d) Exchange rate



forecasts and the quarterly percent change for GDP. This primarily reflects that CPI is not seasonally adjusted. 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, which have been collected 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 net revisions to quarterly GDP growth is 0.4 for period 2000Q1 to 2010Q3. This is large relative to other OECD countries.[14]

[14] See Sleeman (2006) 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.

3.3 Evaluation criteria

In this analysis we are primarily 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.[15] The log score is easy and fast to compute. It has some nice interpretations since it can viewed as a density forecast error, as described in section 2. Significant differences in the log scores are tested using the procedure suggested in Hall and Mitchell (2007).

To help us judge whether the densities are biased in a particular direction, and whether the width of the densities have been accurate, we use probability integral transforms (PITs). The PITs summaries the calibration of the densities. They are the ex-ante inverse predictive cumulative distribution, evaluated with ex-post actual observations. More simply, in our application, the PITs show the number of observations across the evaluation period in each 10 percent interval of the forecast densities. A well calibrated density should have 10 percent of the observations in each of these intervals. Of the around 40 real-time observations there should be roughly four in each interval.

4 Results

4.1 Log score performance

Table 1 summarises the log scores of the three different combination strategies described in section 2, the implied RBNZ density forecasts, and a selection strategy.[16] 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 the

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

[16] Log scores for the individual models are available in tables 7 and 8. Also, we have done a similar forecasting experiment evaluating only point forecasts 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. Results can be provided from the authors on request.

evaluation sample and for different horizons at each point in time.[17]

GDP: Looking at table 1a, the best combination strategy for GDP at almost every horizon is the CRPS weighting strategy. Only at the first forecasting horizon are log score weights better. Further, the combination forecasts perform better than the published forecasts at most horizons, and at the fourth forecasting horizon the CRPS and equal weighted combinations are significantly better than the published forecast. Compared to the selection strategy, the combination strategies are better at most forecasting horizons.

Inflation: Table 1b displays the results from the inflation forecast evaluation. We see that the equal and the CRPS combinations do slightly better than the log score combination at most horizons. Also, at shorter horizons the published forecasts get better log scores than any of the combination strategies. Notably, the published forecasts are significantly better than the combination approaches at the first and second horizons. For forecasts beyond four quarters ahead though, the model combination approaches produce better log scores than the published forecast. Lastly, the model selection approach performs on par with the log score combination.

90-day interest rate: The results for the 90-day interest rate forecasts (see table 1c) show that at the longer forecasting horizons, equal and CRPS combinations perform slightly better than the log score approach. A finding similar to those above. At horizons 1-5 quarters ahead, the published forecasts generally get a better log score than the any of the combination strategies, suggesting that the RBNZ near-term forecasts are particularly good. As for the inflation evaluation, the selection strategy tends to produce relatively similar log scores to the log score combination.

Exchange rate: Finally, table 1d displays the log score evaluation for the exchange rate forecasts. The differences between the weighting strategies are not big in terms of log score, but generally the CRPS and equal weighting strategy perform better than the log score strategy on longer horizons, and vice versa. The combination strategies outperform the RBNZ forecasts for horizons greater than two quarters ahead. Both equal and CRPS combinations are significantly better than the RBNZ forecasts at the longest horizon. Again, the selection strategies do on average worse than the other combination approaches.

[17] 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.

Table 1: Average log scores. All forecasting horizons.

(a) GDP: model combinations and published

	1	2	3	4	5	6	7	8
equal	-0.71	-0.84	-0.96	**-0.99**	-1.09	-1.08	-1.14	-1.08
logScore	-0.61	-0.87	-1.00	-1.01	-1.14	-1.07	-1.18	-1.14
crps	-0.68	-0.83	-0.96	**-0.99**	-1.08	-1.07	-1.14	-1.08
RBNZ	-0.59	-0.88	-0.95	-1.09	-1.08	-1.13	-1.17	-1.14
bestlogscore	-0.65	-0.95	-0.99	-1.12	-1.21	-1.05	-1.22	-1.15

(b) Inflation: model combinations and published

	1	2	3	4	5	6	7	8
equal	-0.22	-0.96	-1.04	-1.19	-1.26	-1.26	-1.27	-1.32
logScore	-0.15	-0.98	-1.03	-1.21	-1.25	-1.29	-1.31	-1.32
crps	-0.14	-0.95	-1.02	-1.19	-1.26	-1.26	-1.26	-1.32
RBNZ	0.17	-0.63	-0.87	-1.10	-1.29	-1.30	-1.33	-1.33
bestlogscore	-0.60	-0.98	-1.03	-1.20	-1.26	-1.27	-1.30	-1.30

(c) Interest rate: model combinations and published

	1	2	3	4	5	6	7	8
equal	0.51	-0.85	-1.35	-1.67	-1.84	-1.99	-2.05	-2.09
logScore	0.75	-0.83	-1.33	-1.70	-1.82	-2.11	-2.21	-2.31
crps	0.71	-0.85	-1.34	-1.67	-1.83	-1.99	-2.06	-2.11
RBNZ	1.32	-0.52	-1.35	-1.62	-1.79	-1.98	-2.10	-2.17
bestlogscore	0.52	-0.89	-1.34	-1.67	-1.81	-2.09	-2.20	-2.36

(d) Exchange rate: model combinations and published

	1	2	3	4	5	6	7	8
equal	-1.48	-2.61	-2.85	-3.08	-3.27	-3.36	-3.44	**-3.42**
logScore	-1.26	-2.57	-2.87	-3.10	-3.28	-3.35	-3.47	-3.45
crps	-1.35	-2.61	-2.85	-3.09	-3.28	-3.36	-3.44	**-3.41**
RBNZ	-0.99	-2.46	-2.88	-3.17	-3.37	-3.52	-3.61	-3.59
bestlogscore	-2.08	-2.56	-2.86	-3.10	-3.25	-3.34	-3.52	-3.45

Notes: The columns display 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 baseline model for this table is the RBNZ forecasts. We test the significance of the log score differences using the technique described by Hall and Mitchell (2007). In this technique we regress the differences in the log scores on a constant and use HAC standard errors to determine its significance. A bolded number implies the combination strategy is better than the RBNZ forecasts, whereas a red number shows the combination is worse than the RBNZ forecasts at the 5 percent significance level.

Summarizing the results, a few main points stand out: The combination strategy performs well for horizons beyond a year ahead, and the equal and CRPS combination strategies tend to outperform the log score weighting strategy. For nearer horizons, the RBNZ forecasts did very well. For GDP and the exchange rate the combination strategies were on average slightly better. While the published forecasts were best for inflation and the interest rate. For only a few variables and horizons did the selection strategy do well. Typically, when the log score combination did well, so did the selection strategy.[18]

As it may be noted when looking at table 1, the differences in log scores are usually not large or significant. 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. In saying this, these results do tentatively suggest a weakness in the model suite. That is, the models' forecasts do not seem to capture the highest frequency data as well as the RBNZ published forecasts. Part of this may be because the models' forecasts are finalized a week or more before the published forecast. Still, the model combination strategy performs well at longer horizons compared to the published forecasts, which we think is encouraging.

4.2 Probability integral transforms

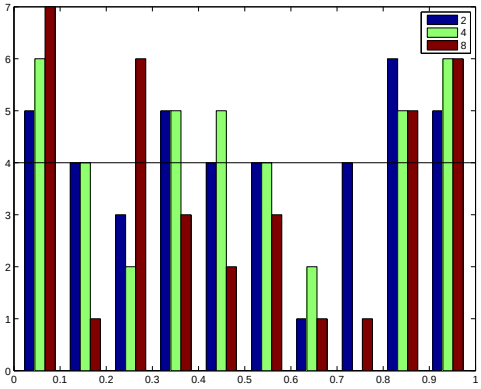
In this section, we compare the probability integral transforms (PITs) for the model combination density forecast and the implied published density forecast from the Reserve Bank of New Zealand.[19] Figures 3 and 4 plot the results for forecasting horizons two, four and eight quarters ahead for all the variables that we evaluate. As mentioned previously, the PITs show the number of observations in each 10 percent interval of the forecast densities, so a well calibrated density will have a similar number of observations in each of these intervals. The horizontal black line in each figure demonstrates the optimal height for each bar. We test the uniformity of the PITs using the chi-squared test, these results are available for all horizons in the appendix (table 9).

[18] See the appendix B for individual model scores.

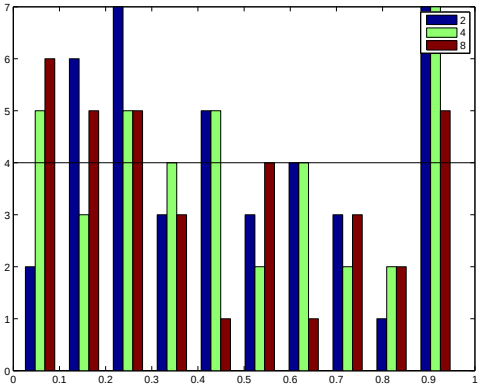
[19] For this section and those to come, we use log score weights to produce the model combination.

Figure 3: Probability integral transforms: GDP and inflation.

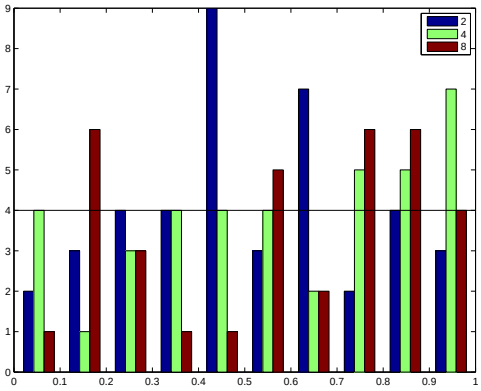
(a) GDP: RBNZ



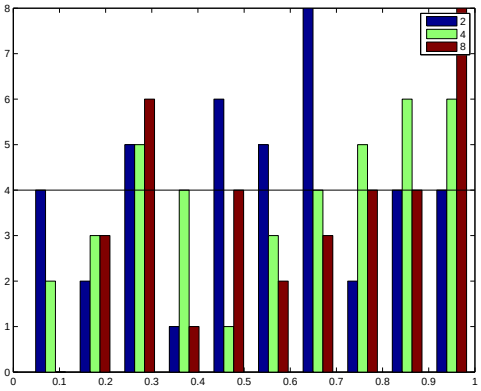
(b) GDP: Model combination



(c) Inflation: RBNZ



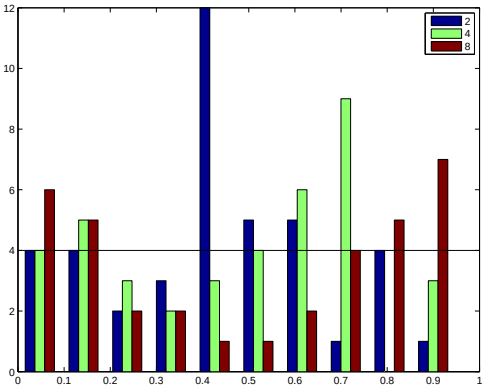
(d) Inflation: Model combination



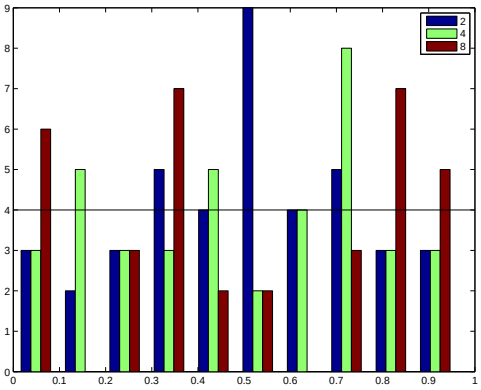
Notes: The bars show forecasting horizons two, four and eight quarters ahead. Each bar color relates to one horizon. A well specified density should have a uniform distribution, close to the horizontal black line. The model combination density forecasts have been derived using log score weights.

Figure 4: Probability integral transforms: Interest rate and exchange rate

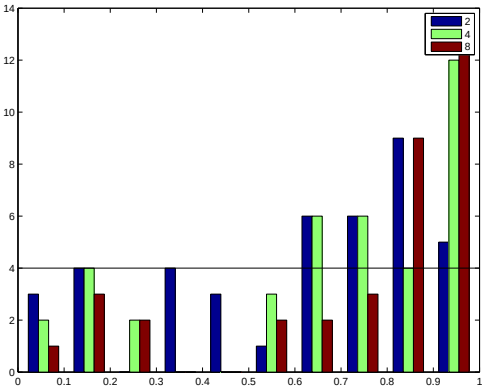
(a) Interest rate: RBNZ



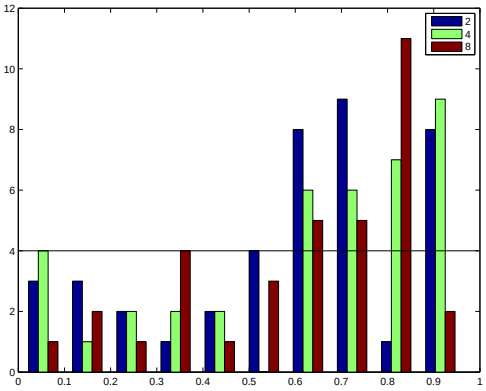
(b) Interest rate: Model combination



(c) Exchange rate: RBNZ



(d) Exchange rate: Model combination



Notes: The bars show forecasting horizons two, four and eight quarters ahead. Each bar color relates to one horizon. A well specified density should have a uniform distribution, close to the horizontal black line. The model combination density forecasts have been derived using log score weights.

GDP: The RBNZ forecasts for GDP, figure 3a, have tended to underestimate the uncertainty at the 8 quarter ahead horizon. Too many observations end up in the tail of the densities. The combined forecasts seem to have a slight positive bias, evident by the downward sloping nature of the bars in figure 3b. However, the chi-squared tests for uniformity, table 9a, show that we cannot reject uniformity for both sets of GDP forecasts at every horizon.

Inflation: Figures 3c and 3d display the PITs for the inflation forecasts. The difference between the published and combined inflation forecasts is quite small, and there is a tendency for both forecasts to underestimated the inflation pressure, particularly at longer horizons. The PIT uniformity tests show that for horizons greater than 3 quarters ahead the p-values are small, see table 9b. Thus, both sets of forecasts seem to be less uniform at longer forecasting horizons, though we cannot reject uniformity at the 95 percent significance level.

90-day interest rate: Over the evaluation sample, the RBNZ densities for the 90-day interest rate have been too narrow for long horizon forecasts. Figure 4a shows that the long-run forecasts too often ended in the tails of the distribution, portrayed by the U-shape of the red bars. For the combined forecasts the PITs appear more uniform. The chi-squared tests on the other hand, suggest we cannot reject uniformity for either the RBNZ or the combined 90-day rate forecasts.

Exchange rate: As figure 4c reveals, too many observations ended up in the upper end of the forecast densities. Neither the combined density forecast nor the RBNZ forecast densities seem well calibrated, and both possess a negative bias. As such, at almost every horizon the chi-squared test rejects a uniform PIT distribution.

Summarizing the results: The performance in terms of PITs between the published density forecasts and the model combination densities is very similar. We can seldom reject uniformity of the PIT distributions for the GDP and interest rate forecasts, while the inflation and exchange rate forecasts both possess a negative bias.

4.3 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, see table 1. It is important to understand why this is so. We propose two explanations for these differences.

First, 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.[20]

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. For the few horizons where a particular model is consistently better than others, the selection strategy can perform well. Likewise, the combination approach using log score weights does well where a particular model is consistently the best. As noted in section 2, the log score puts a harsh penalty for low probability events and is therefore highly sensitive to extreme cases. Because of this, the log score combination is more similar to model selection approach than the other weighting schemes.

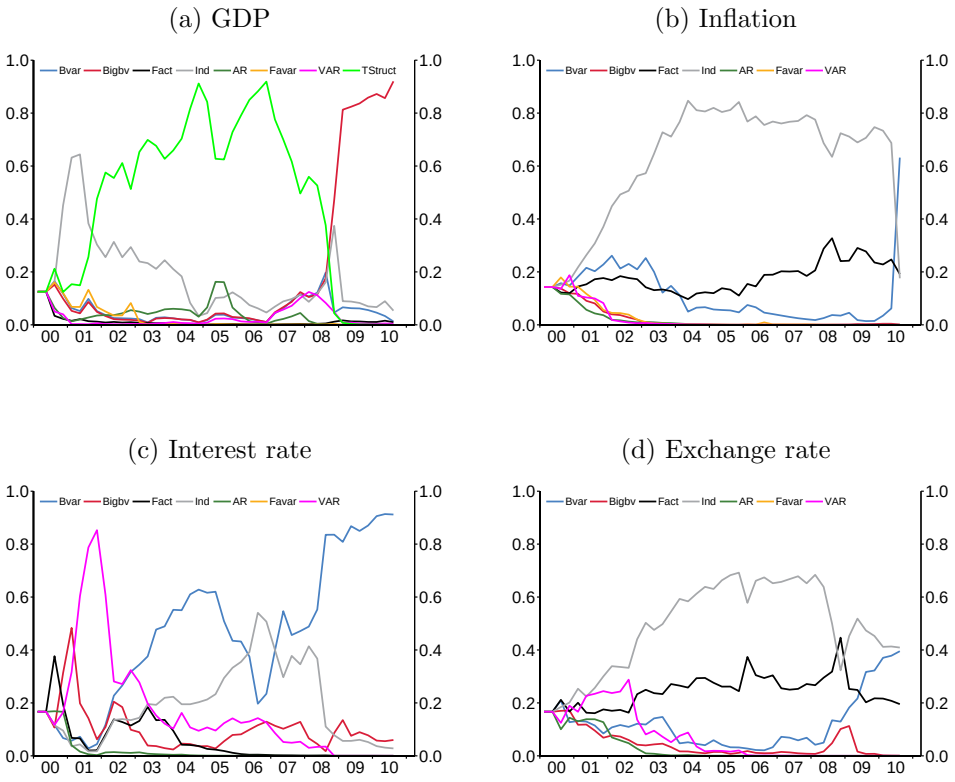
Consistent with the model suite lacking a consistent best model for most variables and horizons, the equal and CRPS weighting strategies do relatively well for many horizons, see table 1d. The difference between the CRPS strategy and the equal weighting strategy is however not large. The CRPS weights are quite forgiving, and are more similar to the equal weighting strategy than log score weights. 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 (2011).

Second, the variables clearly differ in how difficult they are to forecast. For example, quarterly GDP growth is quite volatile over our sample compared to the other variables. 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. As already mentioned, 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.[21]

[20] The models getting higher weights differ, as expected, for each variable that is being forecast.

[21] 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: Individual model weights



Notes: The lines are two quarters ahead log score model weights.

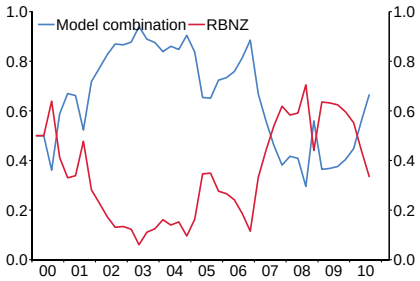
4.4 Weighting combination and published forecasts

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 using log score weights, as we did with the individual models. 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?

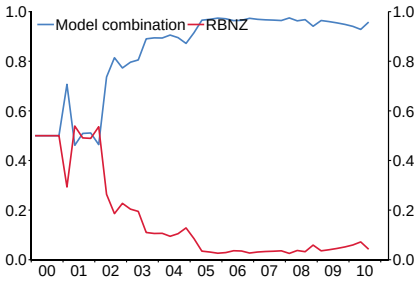
As we saw in the previous sections, the combined forecasts perform relatively well for GDP. As such, the log score weights on the combined forecasts are large for both the second and fourth horizons, see figures 6a and 6b. Further, the combined forecasts have received larger weights than the published

Figure 6: RBNZ and model combination weights

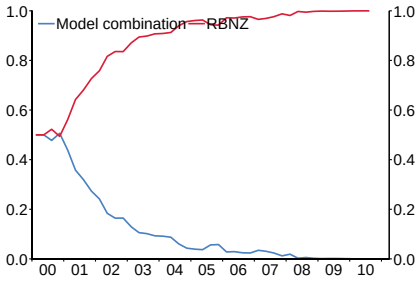
(a) GDP; h=2



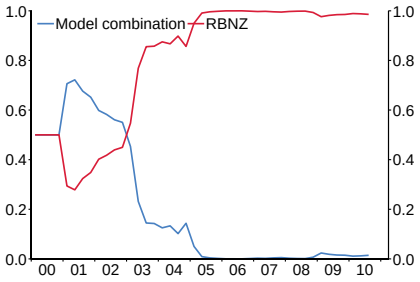
(b) GDP; h=4



(c) Inflation; h=2



(d) Inflation; h=4



Notes: The time varying "model" weights have been derived using the log score weighting criteria.

forecasts for the majority of the sample period.[22] At the two quarter ahead horizon the RBNZ forecasts did well in 2007 and 2008. Since then the model combination has performed better and again receives the majority of the weight.

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 forecasts perform well at the short horizons, while the combination approach

[22] 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.

gets higher weight on the longer forecasting horizons.[23]

Thus, these results confirm the average log score evaluation we reported in section 4.1. It also adds to that analysis by displaying the time varying performance of the two density forecasts, which is particularly evident in figure 6a. For GDP, the published forecasts do not add much value compared to the purely statistical models and the combination approach. For inflation, the RBNZ forecasts are generally better and probably adds value, while the results are horizon dependant for the interest rate and the exchange rate.

4.5 The usefulness of density forecasts

We started this paper by stating that the whole range of outcomes should be forecasted, evaluated and accounted for when making monetary policy decisions. As was shown in figure 1, in the introduction, density forecasts give policy makers a better characterization of all the possible future outcomes, and the probability of different outcomes can be assessed. The density combination approach naturally makes this uncertainty time varying, by applying time varying model weights. In our view a desirable property.

As seen in figure 1b, the combined and the published RBNZ probability forecasts do not differ much. Both forecasts show marked increases in the probability of inflation being above the target band two quarters ahead of time during the episodes around 2005 and 2007. Further, although the RBNZ inflation forecasts generally seem better than the model combination forecasts in terms of log score rankings (see section 4.1), there is not much difference between the two when it comes to the probability forecasts. In contrast to the results shown in section 4.4, this suggests that the published forecasts do not add much information to the purely model based forecasts.

A natural questions is how well the proposed methodology can forecast tail events, such as the recent financial crisis and subsequent huge fall in economic activity. In a similar manner to figure 1b, which displayed the probability of inflation being above its target band, figure 7 shows the probability forecast for negative growth in GDP four quarters and eight quarters ahead. In each of the two subfigures, we consider the probability of this event according to the implied RBNZ forecasts and the combination forecasts using log score weights.

As the figure shows, the RBNZ densities imply a higher probability of GDP

[23] For brevity, we do not display these graphs.

falling throughout the forecast period, particularly at the two year ahead horizon. As such, the RBNZ densities put a greater probability of GDP declining going into 2008/09, but this is to be expected when it always puts a higher probability on negative GDP changes, even in periods that were not. The figure also shows that the size of the revisions to the RBNZ forecasts were larger than for the combined forecasts.

While the model combination did not forecast the GDP decline in 2008/09 with any great probability, the model combination suggested two years ahead of time that a fall in GDP was more likely in 2008 than it had been for the four years prior.[24] This period, around 2006, was towards the end of a long upswing in the New Zealand economy and when interest rates were at quite a high level.

The key models responsible for the changes in the probability of negative GDP outturns were the autoregressive model, the term structure model, and the BVAR. These models' probability forecasts are shown in figure 8.

The models' forecasts for the probability of GDP declining rose for forecasts of 2008. Most notably, the BVAR predicted the decline in GDP to be more than twice as likely in 2008/09 than it had been for the six years prior. Also, the shape of the term structure model seems positively correlated with lagged short term interest rates. As such, this model suggests that the probability of GDP falling has been very low since early 2010, consistent with interest rates also being low.

These results are of course all rather informal. Still, the results displayed in figures 1 and 7 are in our view supportive of the ability of the density combination approach to provide informative assessments of the probability of bad events and/or tail events. The fact that revisions to the models forecasts are gradual and quite small (at least for the long run forecasts) makes these probabilities even more interesting than the implied RBNZ probability forecasts.[25] That said, neither the individual models or the model combination framework produce overly good probability forecasts, at least for GDP. An extension of the mode suite with models that are more tailored to this type of forecasting would most likely be a positive contribution, and improve the

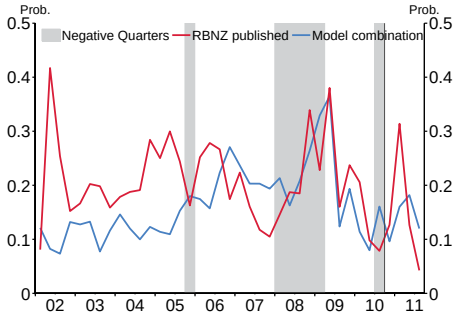
[24] The unconditional probability of GDP being negative is around 20 to 25 percent, depending on the observation period.

[25] As noted by one of the referees, more formal tests of the models ability to forecast tail events can be carried out. Specifically, the same PIT testing framework that was applied in section 4.2, could have been tailored to specific events of interest. We very much acknowledge these comments, but leave it to future work to do a more thorough examination of this issue.

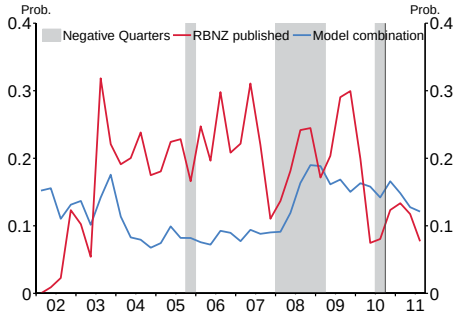
usefulness of the model combination framework applied here.

Figure 7: Probability of negative GDP growth: Combination versus RBNZ

(a) One year ahead



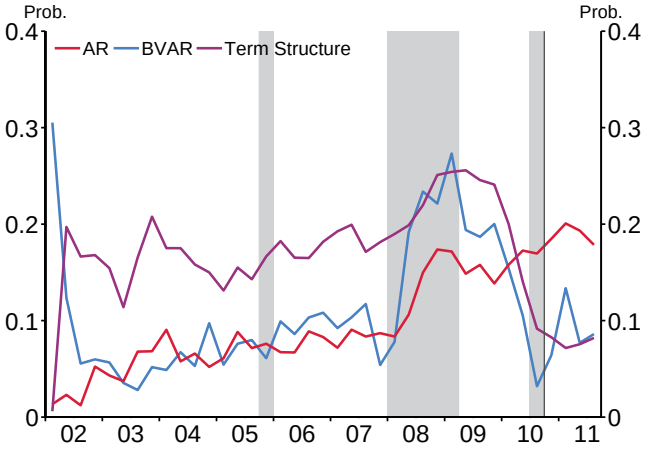
(b) Two years ahead



Notes: The lines show the probability that GDP will be negative at a particular forecast horizon. E.g. the two year ahead forecast probability made in March 2006 would be plotted above March 2008. The lines are forecasts, thus the lines continue beyond the last GDP outturn (2010Q3). Grey areas show the quarters when the quarterly percent change of GDP was negative. Black vertical line indicates the last GDP outturn.

Figure 8: Probability of negative GDP changes: Individual models

(a) Two years ahead



Notes: See notes for figure 7.

5 Robustness

As described in section 3.1, we have derived our implied density forecasts based on historical forecast errors, normality assumptions, and a predetermined training period. Both of these choices can have a material impact on our results. In this section we investigate wether the normality assumption is appropriate, and how different training periods might affect the results.

5.1 Are forecast errors normal?

Table 2 reports the mean and standard deviation of forecast errors along with measures of kurtosis and skewness, estimated over the period 2000Q1 to 2010Q3. We have used the Jarque-Bera (JB) test to check if the errors are normally distributed. Results from this test are shown in the form of p-values. We can reject that the errors come from a normal distribution when the p-value is less than 0.05.

For GDP, we find that using a normal distribution seems appropriate. Whereas for CPI, the forecast errors have a large amount of kurtosis, though we cannot reject a normal distribution.

An alternative to using a normal distribution is a student's t-distribution. This would give slightly fatter tails and may improve the performance of density forecasts where we have a small sample or when kurtosis is a problem. To test this, we compare the performance of the individual models when we produce densities using the student's t and normal distributions. To evaluate the density forecasts, we show the average log scores across all models for the period from 2000Q1 to 2010Q3, in the same manner as in section 4.1.

As seen in table 3, the log scores are generally not significantly different. If anything, for GDP the normal densities are slightly better, while for CPI the t-distributed densities are better, particularly at shorter horizons. These results are consistent with the extent that kurtosis is a problem. The same analysis for the exchange rate and interest rate suggests a similar relationship, high kurtosis coinciding with the t-distributed densities performing better (see tables; 5, 6a and 6b in appendix A).

All in all, our choice of using the normality assumption to construct implied density forecasts has an impact on our results. Even better density forecasts, in terms of log score and PIT evaluation, may have been obtained for some

variables if we had chosen to use for example the t-distribution.[26] That said, the number of observations available to us for this analysis are rather limited, so we would not put to much weight on the distributional statistics reported above. It would certainly not be a desirable strategy to tailor the construction of the implied densities to the sample at hand. The best option would be to have simulated model densities in real-time. Without this, we have chosen to stick to the simplest of assumptions; normality. Also, we do not think this assumption has a material impact on the qualitative results reported in section 4.1, as it affects all of the models.

5.2 Training period

The training period used to determine the variance of forecast errors impacts on the specification of the forecast densities. Real-time forecasts for each model are only available for a short period of time, since 2000. For the models, we use all the available forecasts to train the densities. That is, we use all the forecasts from 2000Q1. Further, we update the data used to derive the densities as we move through the evaluation period. The training period for density forecasts produced in 2010 are based on nearly ten years of forecast errors.

For the RBNZ forecasts, we can use a much longer period to evaluate the forecast errors. As such, in table 4 we compare the performance of the density forecasts when we make the training period start in 1995Q1 and 2000Q1.

The training period starting in 2000Q1 tends to produce better log scores. Underlying this, the forecast errors prior to 2000Q1 were larger, therefore when the densities were trained using data back to 1995Q1 they were wider. With smaller forecast errors occurring since then, the narrower densities (where the training period started in 2000Q1) did better.

These results highlight the sensitivity of the analysis to the training period assumptions. Unfortunately, the rather short sample of pure real-time out-of-sample forecasts does not leave us with much leeway. As noted in section 3.1 and 5.1, the best option would have been to have the models produce density forecasts directly, e.g. through simulations. This would have made the choices regarding both training period and distributions less critical.

[26] Of course, other distributional assumptions could also have been appropriate.

Table 2: Forecast Error Statistics

(a) GDP: models average

Horizon	No. Obs	Mean	Std	Skew	Kurtosis	JB test - P value
1	42	0.09	0.55	-0.08	2.52	0.39
2	41	0.04	0.66	-0.05	2.71	0.46
4	39	0.11	0.71	0.13	2.69	0.48
8	35	0.21	0.75	0.05	2.38	0.49

(b) Inflation: models average

Horizon	No. Obs	Mean	Std	Skew	Kurtosis	JB test - P value
1	42	-0.02	0.48	-1.32	9.13	0.08
2	41	-0.09	0.73	-0.5	4.53	0.29
4	39	-0.24	0.88	-0.02	2.74	0.46
8	35	-0.4	0.93	-0.3	2.6	0.38

Notes: Each row shows the results for forecast errors at a different horizon. The results are the average of the individual models. A negative mean implies a negative bias. The kurtosis value should be close to 3, higher values indicating too many observations in the tails. Skew should be close to zero, a positive value indicators a positive skew in the forecast errors and vice versa. The maximum p-value for the JB test is 0.5.

Table 3: Log scores for density forecasts, comparing student's t and normal distributions.

(a) GDP: models average

	1	2	3	4	5	6	7	8
Normal	-0.79	-0.94	-1.03	-1.03	-1.12	-1.11	-1.16	-1.13
Student's t-dist	-0.81	-0.95	-1.03	-1.06	-1.15	-1.13	-1.19	-1.14

(b) Inflation: models average

	1	2	3	4	5	6	7	8
Normal	-0.66	-1.07	-1.13	-1.27	-1.32	-1.36	-1.35	-1.39
Student's t-dist	-0.51	-1.03	-1.13	-1.29	-1.34	-1.36	-1.36	-1.39

Notes: We show the mean log score across the individual models, these log scores are means of all available forecasts for that horizon. We test the significance of the log score differences using the technique described by Hall and Mitchell (2007). In this technique we regress the differences in the log scores on a constant and use HAC standard errors to determine its significance. The baseline model for this table is the normal model. A bolded number implies that using the t-distribution strategy is better than using the normality assumption, whereas a red number shows that the t-distribution strategy is worse at the 5 percent significance level. We show the mean log score across the individual models, these log scores are means of all available forecasts for that horizon.

Table 4: Log scores of the RBNZ densities using different training windows

(a) GDP

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
2000Q1	**-0.59**	**-0.88**	-0.95	-1.09	-1.08	-1.13	-1.17	-1.14
1995Q1	-0.75	-1.02	-1.05	-1.15	-1.14	-1.17	-1.20	-1.21

(b) Inflation

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
2000Q1	**0.17**	-0.63	-0.87	-1.10	-1.29	-1.30	-1.33	-1.33
1995Q1	-0.31	-0.68	-0.95	-1.18	-1.36	-1.38	-1.39	-1.35

(c) Interest rate

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
2000Q1	**1.32**	-0.52	-1.35	-1.62	-1.79	-1.98	-2.10	-2.17
1995Q1	0.96	-0.82	-1.40	-1.71	-1.91	-2.04	-2.17	-2.25

(d) Exchange rate

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
2000Q1	-0.99	-2.46	-2.88	-3.17	-3.37	-3.52	-3.61	-3.59
1995Q1	-0.98	-2.50	-2.94	-3.24	-3.42	-3.54	-3.62	-3.64

Notes: We test the significance of the log score differences using the technique described by Mitchell and Hall (2005). In this technique we regress the differences in the log scores on a constant and use HAC standard errors to determine its significance. The baseline model used for statistical tests uses the training period from 1995Q1. A bolded number implies the shorter training period is better than using the training period starting in 1995, where as a red number shows that the longer training period performs better at the 5 percent significance level.

6 Conclusion

The recent financial crisis has highlighted that assessing the central projection path is not enough. The whole range of future outcomes should be forecasted, evaluated and accounted for when making monetary policy decisions. Accordingly, 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 when forecasting GDP and the exchange rate. In particular, the combination forecasts have performed well forecasting a year or more ahead of time. The RBNZ's forecasts have performed well for inflation and the 90-day interest rate. Also, the RBNZ

forecasts appear to make better use of high frequency data and have produced preferred near-term forecasts.

Overall, the forecast densities seem well calibrated. However, both the published and model combination forecasts have tended to have a negative bias for inflation and the exchange rate. For GDP though, the bias is if anything positive. In saying this, the PITs are usually not significantly different from uniform, except in the exchange rate case.

We have evaluated three different weighting strategies: equal weighting, CRPS weighting and log score weighting. The empirical results suggest that most often equal or CRPS weights perform better than log score weights. Further, the density combination approach typically performs better than the selection strategy. The log score and selection strategies tend to do well for the same variables and horizons, these are typically when a particular model does best consistently. Which weighting strategy to use probably depends on the problem at hand; the underlying model space and the properties of the variables being forecasted, as in Bjørnland et al (2011).

Our results are hampered by two facts. First, 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. Second, the densities we evaluate are derived on past forecasting performance, and based on specific distributional assumptions. Both factors will affect our results, but we do not think that the qualitatively difference between the RBNZ and the model combination forecasts would change dramatically.

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.

Appendix

A The distribution of forecast errors

Table 5: Forecast Error Statistics

(a) Exchange rate: models average

Horizon	No. Obs	Mean	Std	Skew	Kurtosis	JB test - P value
1	42	-0.19	1.61	-0.05	4.2	0.25
2	41	-0.73	3.7	0.73	3.66	0.2
4	39	-1.64	6.3	0.93	4.01	0.06
8	35	-3.33	7.59	0.5	2.87	0.31

(b) Interest rate: models average

Horizon	No. Obs	Mean	Std	Skew	Kurtosis	JB test - P value
1	42	0.04	0.27	0.87	7.03	0.1
2	41	0.09	0.73	1.6	7.39	0.01
4	39	0.3	1.54	1.06	4.4	0.05
8	35	0.46	2.38	0.71	2.71	0.14

Notes: Each row shows the results for forecast errors at a different horizon. The results are the average of the individual models. A negative mean implies a negative bias. The kurtosis value should be close to 3, higher values indicating too many observations in the tails. Skew should be close to zero, a positive value indicators a positive skew in the forecast errors and vice versa. The maximum p-value for the JB test is 0.5. Less than 0.05 denotes that we can reject the null hypothesis. The null hypothesis is that the errors come from a normal distribution.

Table 6: Log scores for density forecasts, comparing student's t and normal distributions.

(a) Exchange rate: models average

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Normal	-1.83	-2.69	-2.97	-3.19	-3.34	-3.43	-3.51	-3.50
Student's t-dist	-1.73	-2.69	-2.97	-3.20	-3.36	-3.46	-3.53	-3.53

(b) Interest rate: models average

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Normal	0.10	-1.04	-1.52	-1.82	-1.96	-2.09	-2.17	-2.21
Student's t-dist	0.18	-0.97	-1.44	-1.75	-1.93	-2.07	-2.14	-2.18

Notes: We show the mean log score across the individual models, these log scores are means of all available forecasts for that horizon. We test the significance of the log score differences using the technique described by Mitchell and Hall (2005). In this technique we regress the differences in the log scores on a constant and use HAC standard errors to determine its significance. The baseline model for this table is the normal model. A bolded number implies that using the t-distribution strategy is better than using the normality assumption, where as a red number shows that the t-distribution strategy is worse at the 5 percent significance level. We show the mean log score across the individual models, these log scores are means of all available forecasts for that horizon.

B Individual model log scores

Table 7: Average log scores for each horizon.

(a) GDP

	1	2	3	4	5	6	7	8
Bvar	**-0.65**	-0.89	-0.98	-1.08	-1.11	-1.06	-1.16	-1.13
Factor	**-0.65**	-0.90	-0.93	-1.00	-1.09	-1.11	-1.12	-1.10
Indicator	**-0.62**	-0.85	**-0.96**	**-0.99**	-1.17	-1.15	-1.22	-1.21
Tstruct	-1.03	-1.01	-1.12	-1.05	-1.09	-1.05	-1.10	-1.09
Favar	**-0.57**	-1.06	-1.03	-1.06	-1.21	-1.16	-1.19	-1.17
Bigbvar	**-0.69**	-0.78	-0.98	-0.95	-1.08	-1.00	-1.14	-1.11
AR	-1.08	-1.09	-1.15	-1.11	-1.13	-1.16	-1.17	-1.08
VAR	-1.02	-0.92	-1.10	-1.03	-1.13	-1.17	-1.17	-1.12

(b) Inflation

	1	2	3	4	5	6	7	8
Bvar	**-0.10**	**-0.93**	-1.08	-1.24	-1.33	-1.37	-1.35	-1.33
Factor	-0.60	**-0.96**	-1.05	-1.18	-1.28	-1.41	-1.48	-1.56
Indicator	-0.53	**-0.96**	**-0.97**	**-1.15**	-1.28	-1.39	-1.46	-1.44
Favar	-0.66	-1.16	-1.21	-1.40	-1.42	-1.44	-1.32	-1.38
Bigbvar	-0.96	-1.09	-1.10	-1.26	-1.40	-1.40	-1.31	-1.41
AR	-0.88	-1.22	-1.28	-1.38	-1.33	-1.31	-1.38	-1.40
VAR	-0.86	-1.18	-1.23	-1.28	**-1.18**	-1.21	-1.17	**-1.21**

Notes: The columns display the forecasting horizon, and the rows the individual models. 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. We test the significance of the log score differences using the technique described by Mitchell and Hall (2005). In this technique we regress the differences in the log scores on a constant and use HAC standard errors to determine its significance. The baseline model is the AR model, bold implies the corresponding model's forecasts have been statistically (at the 5 percent significance level) better than the AR, while red implies worse.

Table 8: Average log scores for each horizon.

(a) Interest rate

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Bvar	**0.79**	**-0.76**	-1.37	-1.72	-1.91	-2.08	-2.23	-2.28
Factor	**0.33**	-0.97	-1.56	-1.86	-2.06	-2.15	-2.17	-2.21
Indicator	**0.32**	**-0.85**	-1.68	-1.86	-2.05	-2.10	-2.13	-2.20
Bigbvar	**0.66**	**-0.83**	-1.53	-1.87	-1.99	-2.17	-2.23	-2.27
AR	-0.91	-1.56	-1.70	-1.96	-2.01	-2.06	-2.18	-2.18
VAR	**-0.61**	-1.31	**-1.27**	**-1.62**	**-1.73**	-1.99	-2.06	-2.13

(b) Exchange rate

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Bvar	**-1.22**	**-2.55**	**-2.84**	-3.18	-3.39	-3.50	-3.63	-3.72
Factor	**-1.44**	**-2.57**	-2.96	-3.24	-3.38	-3.43	-3.44	-3.33
Indicator	**-1.45**	**-2.55**	**-2.83**	-3.04	**-3.21**	-3.27	-3.35	-3.40
Bigbvar	-2.17	-2.72	-2.99	-3.23	-3.39	-3.53	-3.65	-3.68
AR	-2.38	-2.88	-3.12	-3.24	-3.36	-3.45	-3.48	-3.43
VAR	-2.33	-2.88	-3.07	-3.22	-3.33	-3.42	-3.50	-3.46

Notes: See notes for table 7.

C Probability integral transforms

Table 9: PITs tests - Chi-squared test (p-values)

(a) GDP: model combinations and published

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Model combination	0.54	0.27	0.34	0.42	0.19	0.41	0.32	0.24
RBNZ	1.00	0.88	0.53	0.42	0.33	0.25	0.74	0.61

(b) Inflation: model combinations and published

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Model combination	0.22	0.43	0.21	0.08	0.10	0.07	0.10	0.13
RBNZ	0.76	0.64	1.00	0.26	0.10	0.25	0.18	0.06

(c) Interest rate: model combinations and published

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Model combination	0.76	0.43	0.75	0.87	0.75	0.87	0.50	0.87
RBNZ	0.76	0.04	0.34	0.87	0.52	0.41	0.50	0.87

(d) Exchange rate: model combinations and published

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Model combination	0.06	0.00	0.00	0.01	0.01	0.01	0.10	0.01
RBNZ	1.00	0.04	0.01	0.00	0.00	0.00	0.00	0.00

Notes: The null hypothesis is that the densities are well specified, or that the PITs are uniform. A value below 0.05 implies that we can reject the null with 95 percent confidence.

References

Adolfson, M, M K Andersson, J Lindé, M Villani, and A Vredin (2007), "Modern forecasting models in action: Improving macroeconomic analyses at central banks," *International Journal of Central Banking*, 3(4), 111–144.

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

Bache, I W, J Mitchell, F Ravazzolo, and S P Vahey (2009), "Macro modelling with many models," *Norges Bank*, *Working Paper*, 2009/15.

Bjørnland, H C, K Gerdrup, A S Jore, C Smith, and L A Thorsrud (2009), "Does forecast combination improve Norges Bank inflation forecasts?" *Norges Bank*, *Working Paper*, 2009/01.

Bjørnland, H C, K Gerdrup, A S Jore, C Smith, and L A Thorsrud (2011), "Weights and pools for a Norwegian density combination," *The North American Journal of Economics and Finance*, 22(1), 61–76.

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

Bloor, C and T Matheson (2009), "Real-time conditional forecasts with Bayesian VARs: An application to New Zealand," *Reserve Bank of New Zealand*, *Reserve Bank of New Zealand Discussion Paper Series*, DP2009/02.

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

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

Diebold, F X and J A Lopez (1995), "Forecast evaluation and combination," *Federal Reserve Bank of New York*, *Research Paper*, 9525.

Diebold, F X and P Pauly (1987), "Structural change and the combination of forecasts," *Journal of Forecasting*, 6, 21–40.

Diebold, F X and P Pauly (1990), "The use of prior information in forecast combination," *International Journal of Forecasting*, 6(4), 503–08.

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

Elder, R, G Kapetanios, T Taylor, and T Yates (2005), "Assessing the

MPC's fan charts," *Bank of England Quarterly Bulletin*, 45(3), 326–348.

Figlewski, S and T Urich (1983), "Optimal aggregation of money supply forecasts: Accuracy, profitability and market efficiency," *Journal of Finance*, 38(3), 695–710.

Garratt, A, K Lee, M H Pesaran, and Y Shin (2003), "Forecast uncertainties in macroeconomic modeling: An application to the UK economy," *Journal of the American Statistical Association*, 98(464), 829–38.

Garratt, A, J Mitchell, and S P Vahey (2009), "Measuring output gap uncertainty," *Reserve Bank of New Zealand*, *Reserve Bank of New Zealand Discussion Paper Series*, DP2009/15.

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

Gneiting, T and A E Raftery (2007), "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, 102, 359–378.

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

Hall, S G and J Mitchell (2007), "Combining density forecasts," *International Journal of Forecasting*, 23(1), 1–13.

Hendry, D F and M P Clements (2002), "Pooling of forecasts," *Econometrics Journal*, 5, 1–26.

Hoeting, J A, D Madigan, A E Raftery, and C T Volinsky (1999), "Bayesian model averaging: A tutorial," *Statistical Science*, 14(4), 382–417.

Jore, A S, J Mitchell, and S P Vahey (2010), "Combining forecast densities from VARs with uncertain instabilities," *Journal of Applied Econometrics*, 25(4), 621–634.

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

Kascha, C and F Ravazzolo (2010), "Combining inflation density forecasts," *Journal of Forecasting*, 29(1-2), 231–250.

Krippner, L and L A Thorsrud (2009), "Forecasting New Zealand's economic growth using yield curve information," *Reserve Bank of New Zealand*, *Discussion Paper*, DP2009/18.

Makridakis, S (1989), "Why combining works," *International Journal of Forecasting*, 5, 601–603.

Matheson, T (2007), "An analysis of the informational content of New

Zealand data releases: The importance of business opinion surveys," *Reserve Bank of New Zealand*, *Reserve Bank of New Zealand Discussion Paper Series*, DP2007/13.

Matheson, T D (2006), "Factor model forecasts for New Zealand," *International Journal of Central Banking*, 2(2).

Mitchell, J and S G Hall (2005), "Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR 'fan' charts of inflation," *Oxford Bulletin of Economics and Statistics*, 67(s1), 995–1033.

Mitchell, J and K Wallis (2008), "Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness," *National Institute of Economic and Social Research*, *NIESR Discussion Papers*, 320.

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

Romer, C D and D H Romer (2008), "The FOMC versus the staff: Where can monetary policymakers add value?" *American Economic Review*, 98(2), 230–35.

Sleeman, C (2006), "Analysis of revisions to quarterly GDP - a real-time database," *Reserve Bank of New Zealand Bulletin*, 69, 44p.

Stock, J H and M W Watson (2004), "Combining forecasts of output growth in seven-country data set," *Journal of Forecasting*, 23, 405–430.

Timmermann, A (2006), *Forecast Combinations*, vol 1 of *Handbook of Economic Forecasting*, chap 4, 135–196, Elsevier.

Wallis, K F (1989), "Macroeconomic forecasting: A survey," *Economic Journal*, 99(394), 28–61.

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