BANK OF CANADA
BANQUE DU CANADA

Staff Discussion Paper/Document d'analyse du personnel — 2022-12

Last updated: May 9, 2022

Nowcasting Canadian GDP with Density Combinations

by Tony Chernis and Taylor Webley

Canadian Economic Analysis Department
Bank of Canada, Ottawa, Ontario, Canada K1A 0G9

tchernis@bankofcanada.ca, twebley@bankofcanada.ca

DOI: https://doi.org/10.34989/sdp-2022-12 | **ISSN** 1914-0568

Acknowledgements

The authors are grateful to Calista Cheung, Brigitte Desroches, Daniel de Munnik, Rodrigo Sekkel, Joshua Slive and Shaun Vahey for their comments and feedback through the development of this paper. We would also like to thank participants of the 2019 Forecasting at Central Banks Conference, the 2020 Joint Statistical Meetings and the 2020 IWH-CIREQ-GW Macroeconometric Workshop for their feedback and thoughts on the work, which will undoubtedly influence our future work on this subject. We are also thankful for the excellent research assistance of Daniel Friedland and Connor Neff, and the editorial work of Carole Hubbard and Maren Hansen. Any remaining errors are the responsibility of the authors.

Abstract

Assessing the state of the economy in real time is critical for policy-making, and understanding the risks to those assessments is equally important. Policy-makers are typically provided with point forecasts that contain insufficient information about risks. In contrast, predictive densities estimate the entire range of possible outcomes. This provides a method for quantifying not only the current state of the economy but also the degree of uncertainty, the tail risks and the overall balance of risks around that state. Accordingly, this paper extends the framework of Chernis and Sekkel (2018) to produce density nowcasts for Canadian real GDP growth. We compare several methods of combining predictive densities from 98 models representing four popular classes of nowcasting models. The performance of these combinations is then assessed in both real-time and pseudo-real-time out-of-sample exercises, with the limited sample real-time simulations reinforcing the importance of data revisions for nowcasting. We demonstrate that the combined densities are reliable and accurate tools for assessing the state of the economy and risks to the outlook. We highlight in particular risks at the start of the COVID-19 pandemic.

Topic: Econometric and statistical methods
JEL codes: C, C5, C52, C53, E, E3, E7

Résumé

L'élaboration de politiques exige de pouvoir évaluer la situation économique en temps réel, et il est tout aussi important de comprendre les risques entourant ces évaluations. En général, les décideurs reçoivent des prévisions ponctuelles qui ne contiennent pas assez d'information sur les risques. Or, les densités prédictives estiment, pour leur part, toute l'étendue des possibilités. Elles permettent ainsi de quantifier non seulement la situation actuelle de l'économie, mais aussi le degré d'incertitude, les risques extrêmes et la résultante de l'ensemble des risques concernant cette situation. Cette étude s'attache donc à développer le cadre conçu par Chernis et Sekkel (2018), qui visait à produire des prévisions par densité de la croissance du PIB réel canadien pour la période en cours. Nous comparons plusieurs méthodes consistant à combiner les densités prédictives de 98 modèles représentant quatre catégories courantes de modèles de prévision pour la période en cours. Les résultats de ces combinaisons sont ensuite évalués à l'aide d'exercices hors échantillon en temps réel et en temps quasi réel, les simulations en temps réel fondées sur un échantillon restreint faisant ressortir l'importance des révisions de données au regard des prévisions pour la période en cours. Nous démontrons que les densités combinées sont des outils fiables permettant d'évaluer avec justesse la situation économique et les risques pesant sur les perspectives. Nous abordons en particulier les risques présents au début de la pandémie de COVID-19.

Sujet : Méthodes économétriques et statistiques
Codes JEL : C, C5, C52, C53, E, E3, E7

1 Introduction

Assessing the state of the economy in real time is critical for policy-making. Significant effort and progress have been made toward finding the best models, though there is always uncertainty around the choice of model since no model for the growth of gross domestic product (GDP) is truly correctly specified. As a result, the relative performance of models often varies across time. In response, policy-makers often rely on the outputs of a variety of models, a strategy that the literature supports as a hedge against uncertainty (Coletti and Murchison 2002).

Hedging against uncertainty, however, is not always sufficient. From the perspective of a policy-maker, it is also desirable—and one could argue necessary—to characterize forecast uncertainty. This is because without a representation of uncertainty, the policy-maker is missing many important pieces of information, such as the quantification of tail risks, the degree of uncertainty or the balance of risks.

Standard approaches cannot provide much of this crucial information. A point prediction, by definition, contains no depiction of uncertainty unless paired with other information. And even though a predictive density provides a full accounting of the uncertainty, common approaches are seldom informative about these issues. We propose a density combination approach that could address the shortfalls of these standard approaches by creating densities that can be non-normal and vary depending on the underlying data. In addition, our goal is to find a density combination scheme that properly characterizes the probability of the events that it is predicting (i.e., its predictions are calibrated) and tightly targets Canadian real GDP growth (i.e., its predictions are sharp).

The case for combining point forecasts goes back to at least Bates and Granger (1969) and is lucidly explained in Timmermann (2006). A large body of theoretical and empirical literature (Clemen 1989; Granger and Ramanathan 1984; Timmermann 2006; Clark and McCracken 2010; Granziera, Luu and St-Amant 2013; and many others) supports the finding that combining forecasts generally improves accuracy (i.e., lower root mean forecast error). In practical terms, the success of forecast combinations is evidenced by their adoption by policy institutions, such as the Bank of England (Kapetanios, Labhard and Price 2007), the Norges Bank (Bjørnland et al. 2012; Aastveit, Gerdrup and Jore 2011) and the Bank of Canada (Chernis and Sekkel 2018), among others.

However, when density forecasts are combined, it is not always clear that the combination results in superior statistical properties (Ranjan and Gneiting 2010). Despite this, many studies have found benefits to combining density forecasts, and a substantial literature has recently developed. Over the past 20 years, researchers have laid the foundation for density

combinations in economics.[1] Several authors have shown that combining densities can make predictions more robust and improve them (Jore, Mitchell and Vahey 2010; Del Negro, Hasegawa and Schorfheide 2016), while others have worked on specifying optimal combination strategies from both frequentist (Conflitti, De Mol and Giannone 2015) and Bayesian perspectives (Geweke and Amisano 2011). More recent academic work focuses on modelling the dependence and correlation across forecasts and time variation in weights.[2] Similar to point forecast combinations, density combinations have also proven useful in central banks, but they have been less widely adopted.[3] Overall, the literature suggests that combined density forecasting provides both a rich characterization of forecast uncertainty and a robustness to the "uncertain instabilities" inherent in macroeconomic forecasting (for a thorough review of the evolution of density predictions in economics and its advantages, see Aastveit et al. 2018).

We add to this literature by studying the performance of many model-based density nowcasts and density combination methods for Canadian real GDP growth. To our knowledge, this is the first comprehensive study of density nowcasting in Canada. Moreover, we believe this is the first study to use real-time data in a Canadian nowcasting study (albeit on a shorter sample than the main analysis). In this paper we combine density nowcasts of Canadian real GDP growth from 98 models across four different classes: leading indicator models (ARX), dynamic factor models (DFM), mixed data sampling models (MIDAS) and Bayesian vector autoregressions (BVARs). Using a medium-sized dataset that encompasses the most commonly used variables for nowcasting Canadian GDP (Chernis and Sekkel 2018 and Binette and Chang 2013), we construct predictive densities from each of the 98 models. These predictions are then combined using both one- and two-step procedures (as in Aastveit, Gerdrup and Jore 2011 and Aastveit et al. 2014) using a variety of weighting schemes.

We evaluate the combined density nowcasts with two metrics: calibration and sharpness. We follow Gneiting and Raftery (2007, 359), who state that the best predictive densities will "maximize sharpness of the predictive distributions subject to calibration." A predictive density is calibrated when it properly characterizes the probability of the events that it is predicting. For example, events predicted to occur with a 20% probability should be observed in the data roughly 20% of the time. Sharpness refers to the concentration of the predictive density around the eventual outturn (i.e., how "accurate" it is) and can be thought of as the density forecasting analog of the root mean squared forecast error. We find that, overall, the combined density predictions produce consistently sharp (or accurate) forecasts when

[1] For example, see Wallis (2005); Hall and Mitchell (2007); Mitchell and Hall (2005); Bache et al. (2009).

[2] See Del Negro, Hasegawa and Schorfheide (2016), Billio et al. (2013), Aastveit, Ravazzolo and Van Dijk (2018) and McAlinn and West (2019).

[3] See Bjørnland et al. (2012) and Aastveit, Gerdrup and Jore (2011).

compared with individual models and classes. However, only the minorization-maximization (MM) algorithm produces consistently calibrated forecasts, which is a critical property of predictive densities in a policy-making environment.

The remainder of the paper proceeds as follows: Section 2 describes the dataset along with some unique characteristics of Canadian data. Section 3 describes the econometric framework of our paper and discusses the models and combination techniques. We then analyze the results of the nowcasting exercise in section 4. In section 5 we conduct our full exercise again using a shorter real-time dataset compiled internally at the Bank of Canada and compare the results with their pseudo-real-time counterparts. We then examine the real-time forecasting performance of our proposed platform through the second-quarter 2020 forecast cycle as an illustrative example of the predictive densities in practice. Section 6 concludes.

2 Data

The ideal time series for a nowcasting model meets three criteria. It is (i) relevant to the Canadian economy, (ii) released before GDP with little publication delay and (iii) updated frequently. We start with the dataset in Chernis and Sekkel 2018 that was constructed to meet these criteria by choosing variables that are followed by the market and, in many cases, reported on Statistics Canada's official release bulletin "The Daily." This results in a set of 35 variables, including 24 domestic indicators (quarterly GDP by expenditure, monthly GDP by industry and 22 other indicators), 7 US or international indicators and 4 financial variables.

The dataset includes a mix of hard and soft indicators as well as a high proportion of foreign variables. The hard indicators include commonly used macroeconomic time series such as retail trade, manufacturing sales and housing starts. Soft indicators include consumer confidence, the US purchasing managers' index (PMI), US consumer sentiment and Global PMI. Foreign data are included for several countries, with a focus on the United States because of its close trading relationship with Canada. A list of all data used can be found in Appendix A.1.

There are two notable features in the Canadian macroeconomic data worth highlighting before describing our exercise: first, Canadian GDP data are released with a larger delay relative to most other developed economies; second, Canada has a monthly GDP indicator.

The quarterly Canadian System of National Accounts data are released two months after the end of the reference quarter. This means that GDP data for the first quarter of a year are released at the end of May. This is quite different from other developed countries, such as the United States and the United Kingdom, for example, who release their first estimates of GDP four weeks after the reference period. Other European countries and Japan release their real GDP figures with a delay of only six weeks. Furthermore, many of the most commonly used and timely leading indicators, such as Industrial Production, are not available for Canada in a

timely manner. In the United States, Industrial Production is available within two weeks of the reference period, while in Canada it is reported as a special aggregation of monthly real GDP data roughly two months after the reference period.[4]

The monthly GDP data, while conceptually similar, are methodologically different from the quarterly GDP data.[5] The quarterly GDP data, or GDP at market prices, reflect GDP as measured by both expenditure and income, while monthly GDP is an estimate of the value-added production by industry. Consequently, both the level of and growth in monthly GDP expressed at quarterly rates differ from the quarterly growth rates observed in the quarterly GDP series. This makes monthly GDP growth a very strong, but imperfect, signal of quarterly real GDP growth. For a more in-depth discussion of these indicators and how they compare with similar data in other countries, see Chernis and Sekkel 2018.

All data are transformed to be stationary. Appendix A.1 lists the transformations by variable. In certain instances, the publishing agencies have rebased or made minor definitional changes to series. Where appropriate, we have spliced the most recent series with the corresponding older vintages and rebased them to a common year to extend the amount of historical data available for this study.

Our training sample runs from the first quarter of 1981 to the second quarter of 2000, after which we begin our out-of-sample forecasting exercise. This yields a total of 74 quarterly predictions running until the fourth quarter of 2019. In section 5, we examine the real-time forecasting performance of the platform proposed in this paper for the second-quarter 2020 observation of real GDP growth but leave an analysis of the impact of the pandemic on model coefficients and combination weights to future work.

3 Econometric framework

3.1 Models

The forecasts we produce are combinations of predictive densities from a set of commonly used nowcasting tools. The following section briefly describes the model classes and some of the specifics of our implementation. For a more detailed exposition of the models, see Chernis and Sekkel 2017 or Appendix A.3.

Leading indicator (LI) models: This simple model class is specified as an auto-regressive process for monthly GDP plus one additional independent indicator. The ragged edges for

[4] Canadian monthly real GDP is compiled on a by-industry basis, and industrial production is an aggregation of mining, quarrying and oil and gas extraction, utilities, manufacturing, and waste management services. In May 2020, Statistics Canada began publishing a monthly flash estimate of GDP by industry in each monthly release, reducing the publication delay by one month.

[5] Technically they are referred to as GDP at basic prices, by industry; but for simplicity they will be referred to as monthly GDP.

the leading indicator are filled in by an auxiliary AR(2) model. Since our dependent variable for this class is monthly GDP, we generate iterated monthly predictions until the end of the second quarter being forecast, aggregate those predictions to a quarterly frequency, and then map them into quarterly GDP growth using a simple ordinary least squares bridging model. These types of models have been used extensively at central banks (Angelini et al. 2011; Bundesbank 2013; and Bell et al. 2014), and although they are relatively simple, they have demonstrated good forecasting performance. There is one model for each selected monthly predictor, yielding a total of 24 LI models. The predictive densities are estimated via block wild bootstrap (Aastveit, Foroni and Ravazzolo 2017), first in the generation of the predictions for monthly GDP, and then again over the bridge equation mapping monthly GDP growth expressed at a quarterly rate into quarterly GDP growth.

Mixed data sampling models: MIDAS models are a popular class of models whose main feature is the ability to estimate using mixed frequency data (see Ghysels, Santa-Clara and Valkanov 2004). The MIDAS regression reduces the number of parameters to be estimated by assuming that the lag coefficients on the high-frequency variables can be approximated by a distribution. The distributional assumption reduces the number of parameters to be estimated to only those characterizing the weighting function of the distribution. In our application we use a beta distribution. We also include a lagged low-frequency quarterly autoregressive term. We use a bivariate MIDAS that includes monthly GDP as an additional high-frequency regressor. The same variables are used in the MIDAS class as in the leading-indicators model class, resulting in 24 unique MIDAS models. Predictive densities are estimated via the block wild bootstrap of Aastveit, Foroni and Ravazzolo (2017).

Dynamic factor models: Factor models have become popular among central banks because they have useful properties for nowcasting (Giannone, Reichlin and Small 2008). They utilize the co-movement of a large set of indicators to estimate common factors, which in turn represent the underlying dynamic of that set of variables. We use the factor model developed in Chernis and Sekkel 2017; however, we use Bayesian estimation techniques to facilitate quick and simple estimation for predictive densities. The model parameters are estimated using a Gibbs sampler, and the states are estimated using precision sampling (Chan and Jeliazkov 2009). We estimate two DFMs, using one and two factors extracted from the data. Mixed frequencies are modelled using the technique from Mariano and Murasawa 2003.

Bayesian vector autoregressions: Vector autoregressions and BVARs are among the most common models used by policy-makers for forecasting and economic analysis because of their simplicity and good forecasting performance. The BVAR differs from a VAR in that it uses prior assumptions on the parameters to reduce the problem of parameter proliferation. We use the standard Minnesota prior (see Koop and Korobilis 2010). This class includes 11 models with the set of variables always following this convention: monthly GDP (GDPBP), financial variable, another domestic real activity measure and a foreign activity variable. Since

the models are estimated using Bayesian methods, the predictive density is readily available (see Karlsson 2013).

As mentioned above, each model has a large set of possible specifications. We address the consequent "specification uncertainty" by systematically varying the models along a given dimension (lag length, data sample, variables to include, etc.). In addition, the time series may have structural changes, so for all model classes, except MIDAS, we use both rolling and recursive estimation schemes.[6] This results in a total of 98 different models for predicting Canadian GDP growth. Details on the methods used to derive our predictive densities can be found in Appendix A.3.

3.2 Combination methods

There are several approaches to combining density forecasts. We use the linear opinion pool since it is commonly used in the literature and previous papers find it performs well (Aastveit, Gerdrup and Jore 2011 and Aastveit et al. 2014, among others).[7] The linear opinion pool has the following functional form:

$$p(y_{\tau,h}) = \sum_{i=1}^{N} w_{i,\tau,h} f(y_{i,\tau,h} \mid I_{i,\tau})$$

where N denotes the number of models, and Ii,τ denotes the information set used by model i at time τ to produce the density forecast $f(y\tau,h)$ for variable y at forecasting horizon h, using non-negative combination weights wi,τ,h that sum to 1.[8]

The remaining question is how to specify the weights. There are many different ways to construct valid weights, and to simplify this process we choose weights that fit into three broad categories of increasing complexity. First, we use equal weights that do not leverage any information about relative forecast performance. Second, we weight the densities using scoring rules such that better-performing models get more weight, but we do not take into account correlation across forecasts. We use two scoring methods: log scores and the continuous rank probability score (CRPS). Third, we calculate optimal weights that maximize the log score of the combined forecasts. This is done using the MM method used in Conflitti, De Mol and Giannone (2015). Table 1 shows the expressions for the weights, and in the case of MM weights the objective function.

[6] We do not use rolling estimation with the MIDAS models because the computational intensity of maximum likelihood estimation significantly increases the time for backtesting.

[7] Another option would be the logistic pool. However, it imposes unimodality on the combined density, and we would prefer to be as agnostic as possible about the form of the final combined density.

[8] If the weights do not have these properties, the combined density will not be valid—that is, it will not integrate to 1.

Table 1: Combination methods

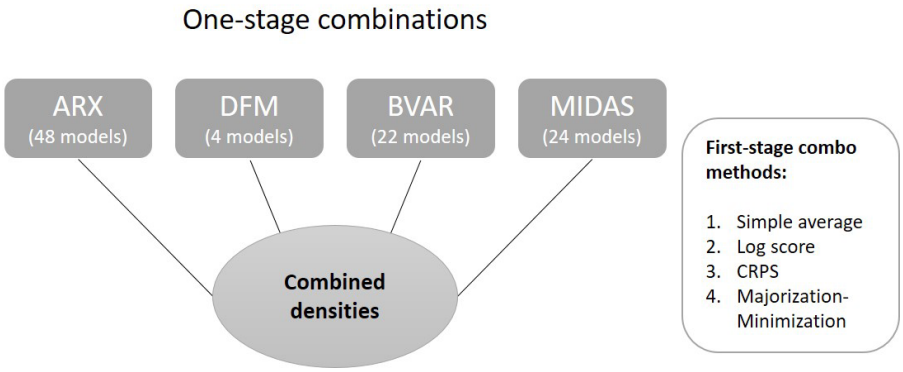
Equal Weights	$w_{i,\tau,h} = \frac{1}{N}$
Log Score	$w_{i,\tau,h} = \frac{\exp \ln(\hat{p}_{i,\tau,h}))}{\sum_{i=1}^{N} \exp \ln(\hat{p}_{i,\tau,h})}$
CRPS	$w_{i,\tau,h} = \frac{crps_{i,\tau,h}^{-1}}{\sum_{i=1}^{N} crps_{i,\tau,h}^{-1}}$
Optimal Weights	$\phi_{\lambda}(\mathbf{w}) = \max_{\mathbf{w}} \frac{1}{T-h} \sum_{t=1}^{T-h} \ln(\hat{p}_i w_i)_t - \lambda \sum_{i=1}^{N} w_i$

Note: $w_{i,\tau,h}$ are weights for model i, at time τ and horizon h. N is the number of models, $\hat{p}$ refers to the density forecast evaluated at the outturn, and λ is the Lagrange multiplier for the constraint that the vector of weights, **w**, sums to 1. CRPS stands for continuous rank probability score.

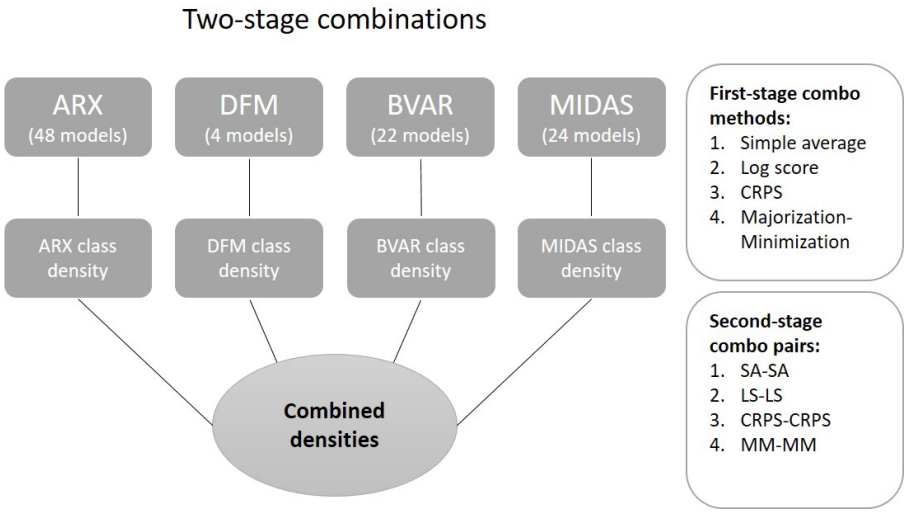
Using the above weights, densities are combined in both one- and two-step procedures. The one-step procedure combines densities across all model-level forecasts to create a final overall density. The two-step procedure first combines model-level forecasts within the same class and then joins these four combined forecasts across classes. The motivation behind the two-stage procedure is that the number of models in each class can differ greatly, and a one-stage combination method could overweight classes with many models. Figures 1a and 1b visualize the combination methods.

Figure 1: Illustration of combination methods

a. One-stage combination method



b. Two-stage combination method



3.3 Out-of-sample design

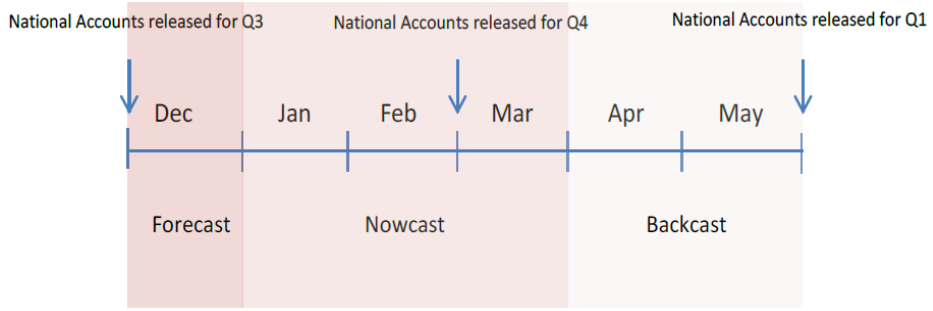
To evaluate a forecasting model, we need forecasts to evaluate. Given the limited availability of real-time macroeconomic data in Canada, we simulate the history of forecasts in pseudo real time over the past 20 years. This exercise aims to recreate the historical flow of data so that we can simulate what the forecasts would have been. We do this by recreating history in two-week blocks such that we simulate data availability over history. We create a forecast for the next two values of real GDP growth after each two-week block of new data. With each successive forecast, the models and combination weights are re-estimated. The exercise remains a pseudo out-of-sample exercise because we use final revised data and therefore do not consider data revisions. While this is less than ideal (see Croushore and Stark 2003; Stark and Croushore 2002; and Orphanides 2001), real-time data for Canada are severely limited. Internally at the Bank of Canada, real-time data have been collected for several years, and we leverage that data in section 5 to conduct our nowcasting exercise in true real time and compare the results with their pseudo-real-time counterparts.[9]

Figure 2 illustrates the timing of releases throughout the six-month forecast cycle. The model's performance is assessed 12 times over the six months, representing a prediction roughly every two weeks. This example shows a forecast cycle starting in December after the release of the third-quarter national accounts data targeting the first-quarter figures for the upcoming year. This exercise is designed to replicate the forecast cycle faced by a practitioner. The cycle starts in December, when the analyst is forecasting the first-quarter figures. Throughout the first quarter, the analyst is in the nowcast phase, and from April

[9] In 2018, Champagne, Poulin-Bellisle and Sekkel 2018 introduced a novel real-time dataset using Bank of Canada staff economic projection databases. This real-time dataset, however, contains only quarterly economic variables leveraged by the Bank's projection model infrastructure. It does not contain the key monthly indicators required for our analysis in this paper. We discuss the alternative real-time dataset used later in this paper in section 5.

through to the May national accounts data release, the analyst is backcasting the first-quarter figures while awaiting publication of the official figures.

Figure 2: Overview of the forecast cycle



Our training sample—the data used to estimate the models—begins in the second quarter of 1980 and ends in the first quarter of 2000. We do not include a training period for the weights. This is because we are combining forecasts in two stages, and having a training period for each stage would drastically reduce the sample size. However, we find that the weights stabilize very quickly for both the performance weighting methods and the optimal weights.[10] Section 4 shows the evolution of the model weights over time. As a result, our out-of-sample forecast exercise begins in the second quarter of 2000 and runs until the fourth quarter of 2019.

4 Results

This section examines the results of our out-of-sample forecasting exercise. In section 4.1, we first examine the calibration and sharpness results of our candidate densities. Section 4.2 then looks at the best candidate specification in detail. Section 5 showcases the results of our exercise when repeated on a limited subsample of real-time data.

4.1 Assessing nowcast accuracy: Pseudo real time

4.1.1 Calibration

Calibration (also referred to as absolute accuracy) is achieved when a predictive density properly characterizes the probability of the events that it is predicting. For example, events predicted to occur with a 20% probability should be observed in the data roughly 20% of the time. More formally, calibration refers to the statistical consistency between the predictive distributions and the observations of the data they are predicting (Gneiting and Raftery 2007). Calibration is a critical property in the context of using density predictions in a policy-making environment because it is necessary to make accurate probabilistic statements. In a practical

[10] The MM algorithm is designed to work with N > T, so the small sample size is not a computational issue.

sense, a prediction that real GDP has a 25% chance of contracting is useful only if over history that prediction comes true roughly 25% of the time (i.e., is calibrated). We assess calibration using probability integral transforms (PITs) (Diebold, Gunther and Tay 1998), using the tests proposed in Knüppel (2015) and in Rossi and Sekhposyan (2019).

In the end, we find that only predictive densities created using one-stage MM weights are calibrated. Table 2 shows the p-values from the PITs-based test proposed by Knüppel 2015, which is robust to serial dependence between the one- and multi-step-ahead density forecasts. We find solid evidence that the one-stage MM combination densities are well calibrated by not rejecting the null hypothesis of calibration at most forecast horizons. All other combination methods are found to be poorly calibrated, strongly rejecting the null hypothesis at the 5% of confidence level for most, if not all, forecast horizons. While evidence of calibration is not broad-based, the fact that the MM weighting method is found to be the superior method is unsurprising, given that they are algorithmically derived to optimize the log score of combined density, providing greater scope for the weights to reflect the best possible combination of the underlying densities. We also discuss in section 5 that we find broader evidence of calibration using a restricted sample from 2013 onward. Indeed, we find that the 2008–09 global financial crisis plays an important role in how calibrated the non-MM weighting methods are found to be. We view the robustness of the one-step MM weights' calibration to the global financial crisis as strong evidence of its usefulness as a tool in a policy-making environment.

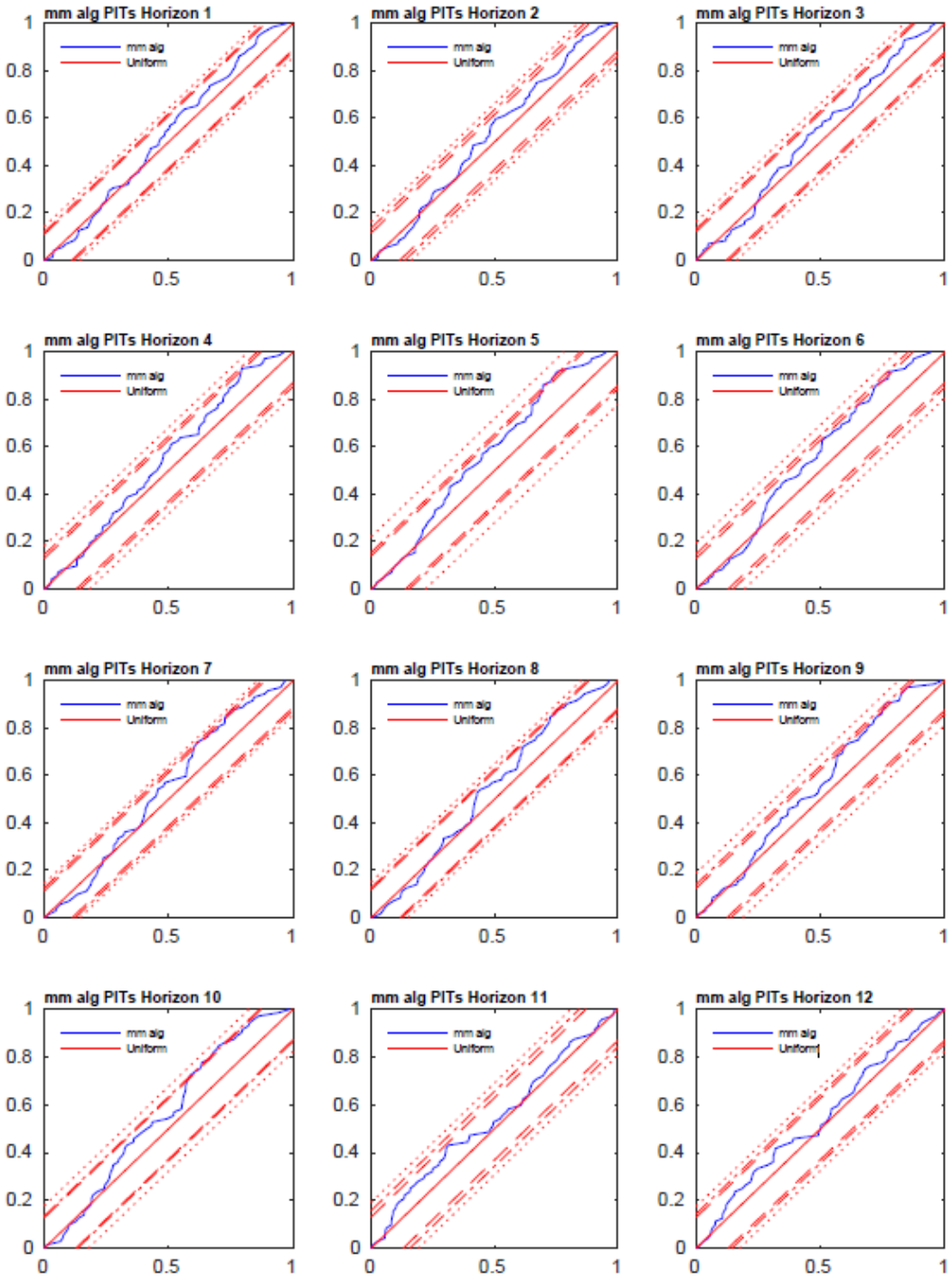
Visual inspection of the PITs illustrates the statistical results. Chart 1 shows the PITs at each out-of-sample forecast horizon for the densities produced using one-stage MM weights. Note that that when the PITs are uniformly distributed, the predictions are said to be calibrated. The dotted lines around the uniform benchmark are the 90%, 95% and 99% confidence intervals constructed using the test statistics proposed by Rossi and Sekhposyan (2019). The PITs for the remaining combination methods can be found in Appendix A.2. As suggested by the results of the Knüppel test, the PITs for the one-stage MM combination density are insignificantly different from the uniform benchmark. This is generally not true for the other combination methods. Overall, our finding that at least one combined predictive density is consistently calibrated is an important result in the context of our framework being suitable for forecasting in a policy-making environment.

Table 2: Calibration by combination methods: Knüppel (2015) p-values

		One-stage combinations				Two-stage combinations			
	Weeks until GDP	SA	CRPS	Log	MM	SA	CRPS	Log	MM
Forecast	24	0.07	0.07	0.12	0.15	0.04	0.04	0.04	0.13
	22	0.06	0.07	0.04	0.15	0.04	0.04	0.02	0.11
Nowcast	20	0.06	0.06	0.05	0.25	0.04	0.04	0.05	0.12
	18	0.04	0.04	0.02	0.13	0.01	0.01	0.01	0.07
	16	0.03	0.03	0.05	0.11	0.01	0.01	0.02	0.09
	14	0.05	0.04	0.04	0.04	0.02	0.02	0.04	0.08
	12	0.02	0.02	0.05	0.11	0.01	0.01	0.02	0.03
	10	0.02	0.02	0.03	0.12	0.01	0.01	0.03	0.05
Backcast	8	0.00	0.00	0.01	0.11	0.00	0.00	0.00	0.03
	6	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.01
	4	0.01	0.01	0.56	0.43	0.01	0.01	0.54	0.49
	2	0.01	0.01	0.69	0.60	0.02	0.01	0.64	0.70

Note: The abbreviations in the column headers are SA, simple average; CRPS, continuous rank probability score; MM, minorization-maximization.

Chart 1: Probability integral transforms for one-stage minorization-maximization combination density



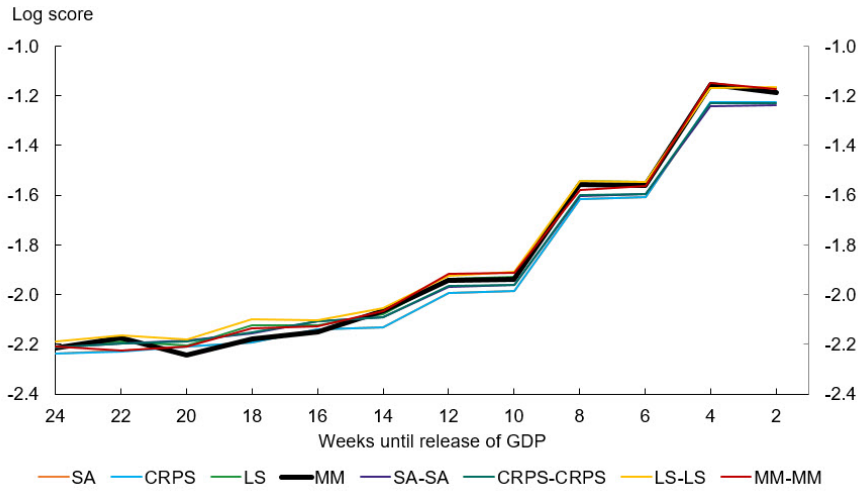
Note: These charts compare probability integral transforms (PITs, plotted in blue as "mm alg") at all forecast horizons as described in section 3.3. MM refers to minorization-maximization. The x-axis reflects values of the cumulative distribution function (CDF), while the y-axis shows the value of the PITs at each value of the CDF. The dotted lines around the uniform benchmark are the 90% (long-dashed line), 95% (short-dashed line) and 99% (dotted line) confidence intervals constructed using the test statistics proposed by Rossi and Sekhposyan (2019).

4.1.2 Sharpness

Sharpness (also known as relative accuracy) refers to the concentration of the predictive density around the eventual outturn. It can be thought of as the density forecasting analog of the root mean squared forecast error. Sharpness measures allow us to compare the relative accuracy of the predictive densities. We evaluate sharpness using two scoring rules: log scores and the CRPS. The log score is equal to the logarithm of a probability density function evaluated at the outturn of the forecast, while the CRPS is a measure of the integrated squared difference between the cumulative distribution function of the forecasts and the corresponding cumulative distribution function of the observations. Higher log scores and lower CRPS imply more accurate predictions. For a detailed discussion of scoring rules, see Gneiting and Raftery (2007) or Corradi and Swanson (2006).

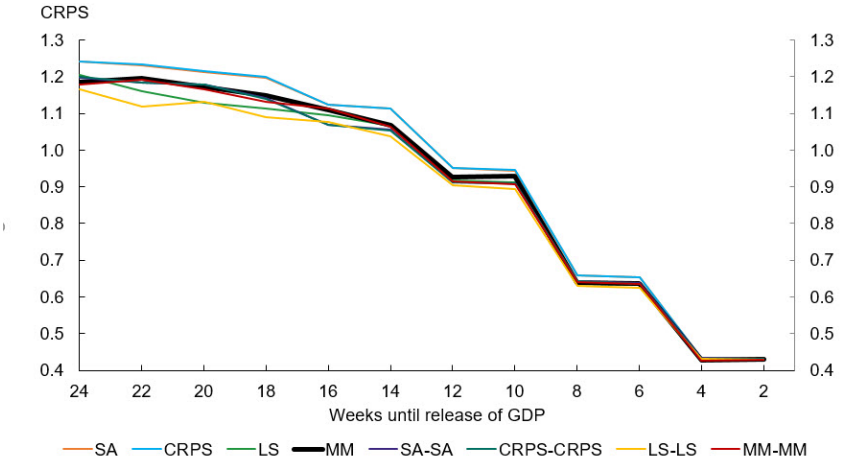
Charts 2 and 3 plot the average log scores and CRPS for all eight combination methods over the 12 out-of-sample forecasting periods in our exercise. Similar to the wider nowcasting literature, we find that forecast accuracy improves for all combination methods as the prediction is updated with new data releases, especially those for the reference quarter. Furthermore, we find that the largest gains in accuracy come from the release of GDP-by-industry data for the first and second months of the target quarter (eight and four weeks ahead of the release of the quarterly GDP figure, respectively), consistent with the findings of Chernis and Sekkel 2018 in the point-forecast setting.

Chart 2: Average log scores of the combined predictive densities



Note: The abbreviations in the legend are SA, simple average; CRPS, continuous rank probability score; LS, log score; MM, minorization-maximization. When paired, they refer to the two-stage combinations of each metric.

Chart 3: Continuous rank probability scores of the combined predictive densities



Note: The abbreviations in the legend are SA, simple average; CRPS, continuous rank probability score; LS, log score; MM, minorization-maximization. When paired, they refer to the two-stage combinations of each metric.

The sharpest predictions are generally made by the two-stage log score combination densities, with the two-stage MM combination sometimes outperforming them at shorter forecast horizons. However, as illustrated by charts 2 and 3, the sharpness of the different combination methods varies mostly in the first 8 weeks of the forecasting cycle, especially as measured by the CRPS. By about 12 weeks ahead of the release of GDP for the quarter being predicted, the accuracy of the eight combination forecasts is fairly similar.

Overall, our results ultimately point to a favourable trade-off between calibration and sharpness. The one-stage MM combination density, the only candidate specification with robust evidence of calibration, is never the most accurate density forecast. However, it is encouraging that over time its log score and CRPS converge toward the most accurate combination density, and by approximately halfway through the forecasting horizon its difference becomes negligible.

4.2 Examining the one-stage minorization-maximization combination weights

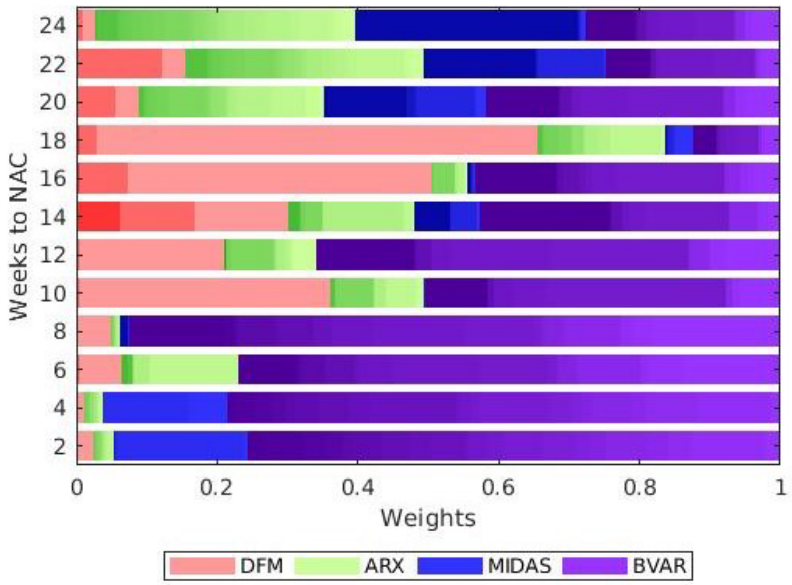
In this section, we examine the weights produced by the one-stage MM algorithm. We focus the discussion on the one-stage MM weights since, as the only combination method with robust evidence of calibration, the one-stage MM weights provide the only candidate specification that meets all the criteria for being useful in a policy-making environment.

Chart 4 shows the weights computed from the one-stage MM algorithm at the end of the backtest sample at all 12 horizons. Chart A-10, A-11 and A-12 in Appendix A.4 show the evolution of these weights at the first (i.e., forecast), seventh (i.e., nowcast), and twelfth (i.e., backcast) prediction horizons, respectively. Overall, we find that the one-stage MM algorithm

places roughly equal weights across the ARX, MIDAS and BVAR blocks for the initial forecast. Perhaps the most notable observation drawn from Chart 4 is the progressive dominance of the BVAR models, which see their class weight increase from 27% at the first horizon to 76% at the last. That said, 22 models receive above-average weights at the final horizon compared. However, the largest weight given to any one model at the last horizon is 18%, whereas the highest at the first horizon is 31%. Generally speaking, we find that the estimated weights can be divided into three distinct groups: before the global financial crisis, during the crisis and after the crisis. The weights in the first two periods (before and during the crisis) are fairly concentrated, with relatively fewer models receiving most of the weight. After the global financial crisis, however, weights become increasingly spread over many different models at most prediction horizons.

The increasing weight placed on various models in the BVAR block appears to be due to the performance of these models benefitting the most from the receipt of monthly GDP data, especially for the reference quarter during the backcast period of our prediction cycle. Indeed, inspection of the weights shows that the MM algorithm places much larger weights on the best performing models (i.e., those with better sharpness metrics) and also reacts quickly to changes in model performance. These are desirable characteristics for a nowcasting model, since we want to use a wide set of information but also emphasize the most useful information at any given time.

Chart 4: One-stage minorization-maximization combination weights, end-of-sample, all horizons



Note: The abbreviations in the legend are DFM, dynamic factor models; ARX, autoregressive leading indicator models; MIDAS, mixed-data sampling; BVAR, Bayesian vector autoregression. NAC on the vertical axis refers to "national accounts" (i.e., weeks until the release of GDP data).

4.2.1 An indicator of risk

Another aspect of a density nowcast that is useful to policy-makers is the ability to signal risks to the economy. As we argue throughout this paper, a prediction using only a measure of central tendency is insufficient for a policy-maker since it does not fully consider the uncertainty around possible outcomes. One of the most important pieces of information missing from this type of prediction is the quantification of tail risks. Standard approaches to creating predictive densities would not be informative about tail risks, which typically resemble a normal distribution. However, a density combination approach could address the shortfalls of these standard approaches by creating densities that can be non-normal and can vary depending on the underlying data. In this section we will study how well the combinations signal risk by creating measures of downside risk and studying them throughout the out-of-sample exercise.

We focus on downside risk following the results from Adrian, Boyarchenko and Giannone (2019), who show that downside risks tend to be more important over history compared with upside risks. Downside entropy quantifies downside uncertainty as the additional probability mass on the left of the predictive distribution relative to the unconditional density of GDP. More formally, we denote $\hat{g}_{y_{t+h}}$ as the unconditional distribution of GDP growth and $\hat{f}_{y_{t+h}}(y|x_t)$ as the predictive density at time *t* for predictive horizon *h*. $\hat{F}_{y_{t+h}}(y|x_t)$ is the cumulative distribution associated with $\hat{f}_{y_{t+h}}(y|x_t)$, and $\hat{F}_{y_{t+h}}(0.5|x_t)$ is the conditional median.

$$\mathcal{L}_t^D(\hat{f}_{y_{t+h}};\hat{g}_{y_{t+h}}) = -\int_{-\infty}^{\hat{F}_{y_{t+h}|x_t}(0.5|x_t)} (\log \hat{g}_{y_{t+h}}(y) - \log \hat{f}_{y_{t+h}}(y|x_t))\hat{f}_{y_{t+h}}(y|x_t)dy$$

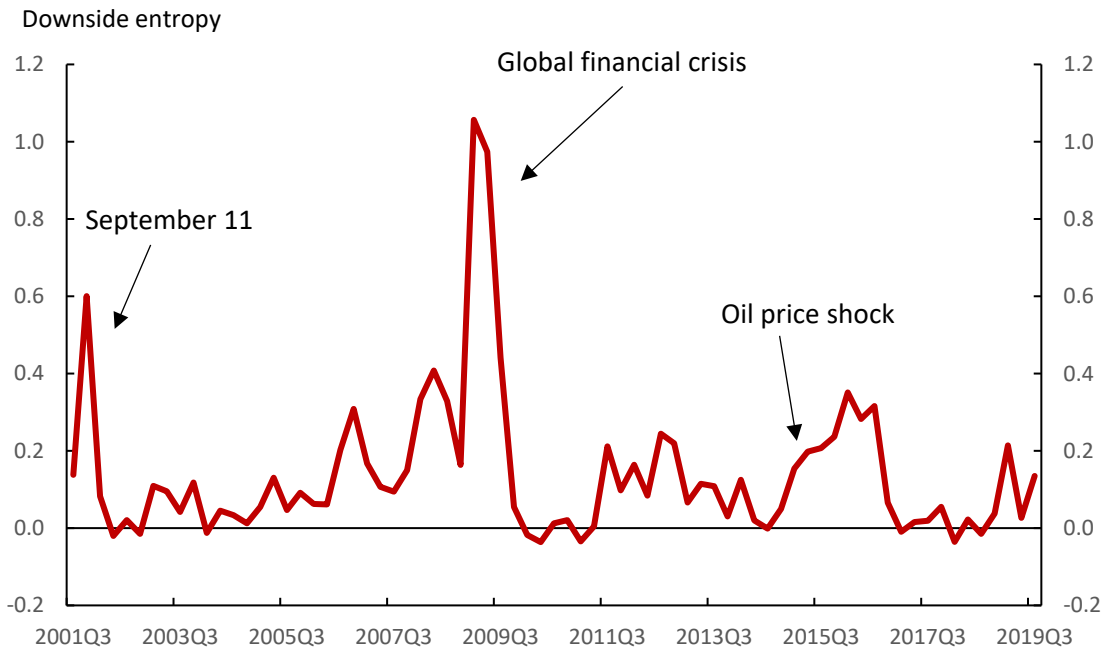
$\mathcal{L}_t^D$ is the measure of downside entropy that intuitively measures the divergence between the unconditional distribution and the conditional (predictive) distribution. When downside entropy is high, the conditional density assigns positive probability to more extreme left tail growth than the unconditional density does. Similarly, one could construct a measure for upside entropy. However, as noted, Adrian, Boyarchenko and Giannone (2019) find that upside risk to the conditional distribution of GDP tends to be less empirically important.

Chart 5 shows the downside entropy measure for the nowcast horizon (fifth forecast) from the one-stage MM algorithm combination.[11] The chart shows that the measure spikes before and during significant economic events. For example, the measure increases before the global financial crisis and the 2015 oil price shock and then peaks in the midst of these events. It is

[11] Other horizons show similar results.

important to note that this reflects both downside risk in the forecast from non-normality (skew) and a shift to the left in the location of the prediction conditional on the available data. Regardless, this suggests that predictions have valuable information about emerging risks to the economy.

Chart 5: Downside entropy rises during periods of economic stress



5 Real-time results

Until recently, real-time macroeconomic data have not been available at a monthly frequency in Canada. At the Bank of Canada, staff have been collecting real-time vintages for most major macroeconomic time series for several years. This allowed us to run a limited real-time experiment for this paper starting in the second quarter of 2013. This is the first such real-time evaluation done for nowcasting in Canada.[12] We then compare these results to a similar pseudo-real-time exercise. This exercise runs from the second quarter of 2013 to the fourth quarter of 2019 and includes all real-time data available for the series presented in Appendix A.1. For clarity, the calibration and sharpness metrics reported in this section are computed by evaluating the real-time density predictions against the first vintage of real GDP growth (i.e., the vintage consistent with the real-time data being used to generate the predictions).

We find similar results compared with the full-sample exercise in the sense that the MM algorithm is calibrated and there are no large differences in the scores across combinations.

[12] While other real-time data sources exist—for example, see Champagne, Poulin-Bellisle and Sekkel (2018)—this is the only one with monthly data that are critical for nowcasting.

However, we find that both the real-time and the pseudo-real-time results show more evidence of calibration, likely due to the subsample used. In both cases almost all combination methods show evidence of calibration. The results are not unambiguously better: the sharpness metrics deteriorate in real time, and the gains that coincide with the receipt of monthly GDP data for the reference quarter are meaningfully reduced. This is because revisions to the monthly GDP in real time can distort the indicator's signal for quarterly real GDP growth. We find some differences in the combination weights when estimated in real time. The weights are more evenly spread across models and increase the most in models that contain series that are either seldom or not revised, such as Labour Force Survey data and rig counts. Our real-time results highlight the importance of real-time data in model evaluation and reinforce the need to continue growing the availability of real-time data in Canada. We examine these findings in more depth below, and we conclude this section with a case study of the real-time performance of our proposed platform through the second-quarter 2020 forecast cycle.

5.1 Calibration

As mentioned above, we find that the MM algorithm and the other combinations show robust evidence of calibration in both real time and pseudo real time (see tables 3 and 4). Furthermore, the MM combination methods continue to show the strongest evidence of calibration (i.e., the highest p-values), though interestingly the two-stage MM combination method's calibration metrics in real time become very comparable to those of the one-stage MM combination. Comparing the pseudo and real-time results, we find relatively small differences in the p-values of the Knüppel test. this suggests that the use of pseudo real time in our main analysis is unlikely to be qualitatively affecting our finding that the MM algorithm provides the best and most calibrated predictive densities. Finally, we find more evidence of calibration in the post-2013 sample, which is notably different from the full-sample results, where we find calibration only for the MM algorithm. This is likely due to the combination's performance through the global financial crisis and the earlier periods.

Table 3: Knüppel (2015) p-values, real time, 2013Q2 to 2019Q4

		One-stage combinations				Two-stage combinations			
	Weeks until GDP	SA	CRPS	Log	MM	SA	CRPS	Log	MM
Forecast	24	0.21	0.22	0.19	0.17	0.16	0.16	0.16	0.20
	22	0.22	0.22	0.23	0.31	0.18	0.18	0.20	0.29
Nowcast	20	0.26	0.26	0.29	0.48	0.20	0.20	0.22	0.46
	18	0.23	0.24	0.25	0.39	0.17	0.17	0.16	0.42
	16	0.37	0.37	0.29	0.50	0.30	0.30	0.26	0.49
	14	0.38	0.37	0.26	0.45	0.31	0.31	0.25	0.46
	12	0.26	0.26	0.23	0.50	0.23	0.23	0.22	0.60
	10	0.25	0.25	0.21	0.51	0.21	0.21	0.19	0.43
Backcast	8	0.36	0.37	0.53	0.94	0.27	0.29	0.45	0.92
	6	0.35	0.35	0.53	0.75	0.26	0.28	0.44	0.84
	4	0.35	0.36	0.26	0.44	0.29	0.33	0.24	0.84
	2	0.33	0.34	0.24	0.41	0.26	0.29	0.22	0.44

Note: The abbreviations in the column headers are SA, simple average; CRPS, continuous rank probability score; MM, minorization-maximization.

Table 4: Knüppel (2015) p-values, pseudo real time, 2013Q2 to 2019Q4

		One-stage combinations				Two-stage combinations			
	Weeks until GDP	SA	CRPS	Log	MM	SA	CRPS	Log	MM
Forecast	24	0.22	0.23	0.16	0.17	0.19	0.19	0.15	0.18
	22	0.23	0.23	0.16	0.21	0.18	0.19	0.14	0.22
Nowcast	20	0.20	0.20	0.14	0.12	0.18	0.18	0.11	0.19
	18	0.18	0.18	0.12	0.20	0.16	0.17	0.10	0.32
	16	0.33	0.33	0.29	0.60	0.25	0.25	0.27	0.75
	14	0.33	0.33	0.26	0.42	0.29	0.28	0.24	0.64
	12	0.24	0.24	0.18	0.57	0.22	0.22	0.19	0.53
	10	0.23	0.23	0.15	0.55	0.20	0.20	0.16	0.37
Backcast	8	0.23	0.23	0.11	0.22	0.21	0.18	0.10	0.21
	6	0.23	0.23	0.10	0.22	0.21	0.21	0.10	0.21
	4	0.20	0.16	0.79	0.60	0.27	0.23	0.78	0.63
	2	0.21	0.17	0.75	0.58	0.27	0.24	0.74	0.39

Note: The abbreviations in the column headers are SA, simple average; CRPS, continuous rank probability score; MM, minorization-maximization.

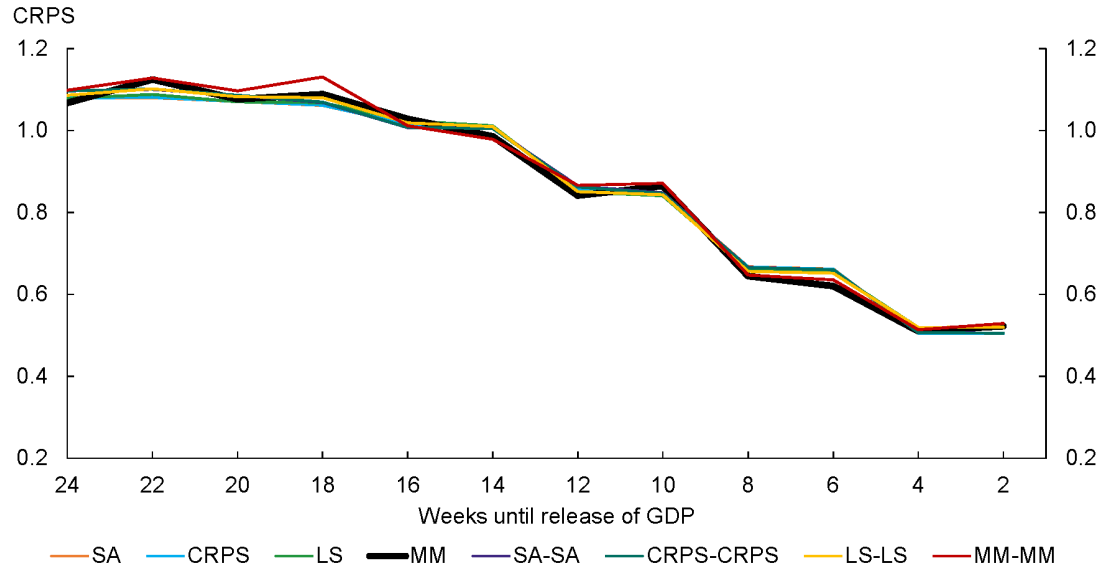
5.2 Sharpness

Chart 6 and Chart 8 show the CRPS and log scores from the real-time exercise. We find qualitatively similar results in the real-time exercise compared with the pseudo-real-time exercise. However, the sharpness metrics unambiguously deteriorate when we use real-time

data. Examining the ratio of real-time to pseudo-real-time scores for the CRPS (Chart 7) and log scores (Chart 9) shows that sharpness is worse at every forecast horizon for every combination method.[13] While the reduction in sharpness is small in the first eight horizons (generally between 1% and 10% worse), the deterioration becomes much more pronounced at the shortest prediction horizons when monthly GDP for the reference quarter is released (roughly 35% to 60% worse). While dramatic, these numbers are unsurprising given the importance of monthly GDP and the fact that monthly revisions to macroeconomic time series in real time can be significant between the initial release of the data and the publication of the national accounts (Rizzetto 2018). Consequently, our results highlight that the accuracy of real-time predictions suffers particularly when revisions are very large or when several series are revised at once, especially for highly informative series such as monthly GDP. Fortunately, revisions are generally unbiased; so, apart from reducing sharpness, the use of real-time data does not appear to present any additional obstacles to our framework.

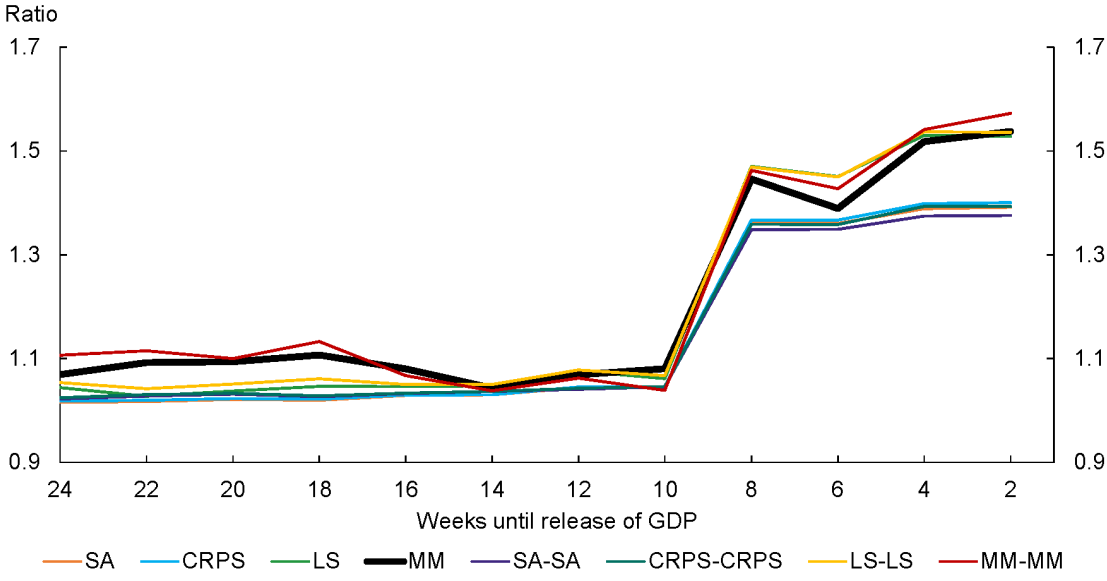
[13] A ratio greater than 1 signals that accuracy is worse in real time.

Chart 6: Real-time continuous rank probability score, 2013Q2 to 2019Q4



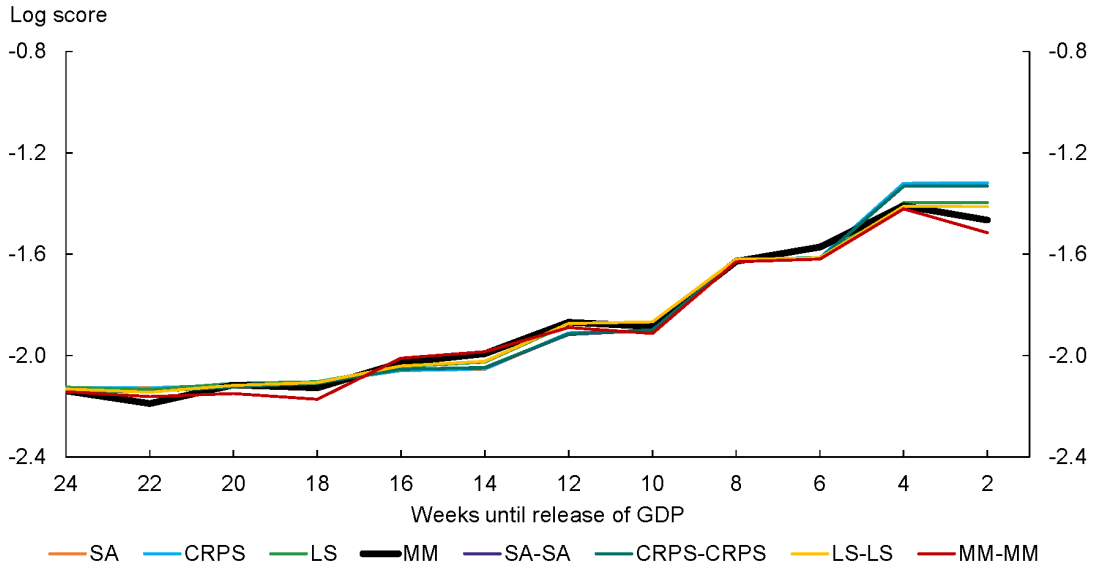
Note: The abbreviations in the legend are SA, simple average; CRPS, continuous rank probability score; LS, log score; MM, minorization-maximization. When paired, they refer to the two-stage combinations of each metric.

Chart 7: Ratio of real-time to pseudo-real-time continuous rank probability score, 2013Q2 to 2019Q4



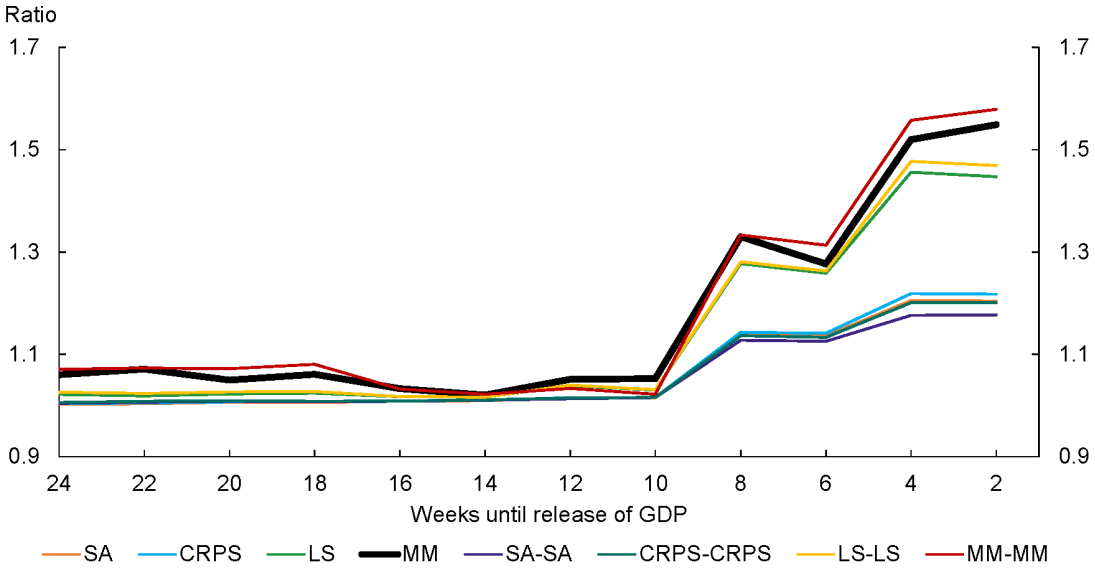
Note: The abbreviations in the legend are SA, simple average; CRPS, continuous rank probability score; LS, log score; MM, minorization-maximization. When paired, they refer to the two-stage combinations of each metric.

Chart 8: Real-time log score, 2013Q2 to 2019Q4



Note: The abbreviations in the legend are SA, simple average; CRPS, continuous rank probability score; LS, log score; MM, minorization-maximization. When paired, they refer to the two-stage combinations of each metric.

Chart 9: Ratio of real-time to pseudo-real-time log scores, 2013Q2 to 2019Q4



Note: The abbreviations in the legend are SA, simple average; CRPS, continuous rank probability score; LS, log score; MM, minorization-maximization. When paired, they refer to the two-stage combinations of each metric.

5.3 Weights

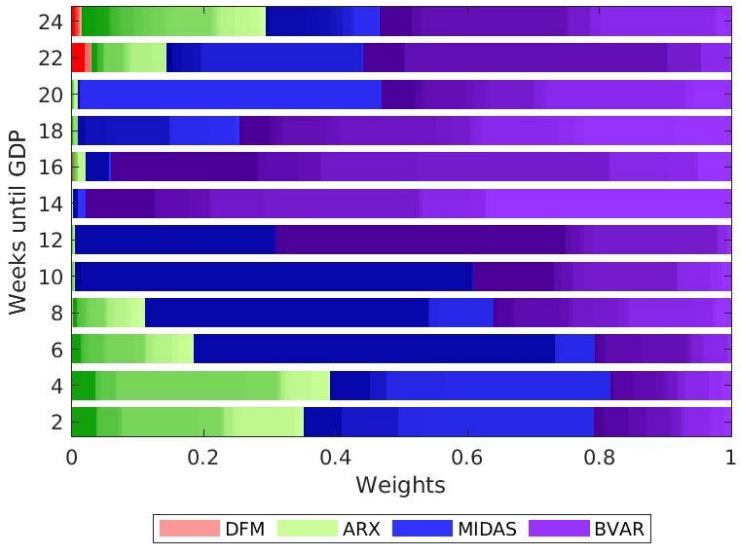
In this section we examine the weights calculated in real time. Overall, we find that the real-time estimation of our platform generates a wider distribution of weights, and models with unrevised data are given more weight. Charts 10 and 11 show the real-time and pseudo-real-time weights for all forecast horizons in the fourth quarter of 2019 (i.e., at the end of the out-of-sample exercise period). Appendix A.4.1 shows the evolution of the weights over the real-time subsample at two horizons, namely, 24 weeks and 2 weeks ahead of the release of GDP.

First, Chart 10 shows a fairly large pool of models receiving non-negligible weights in our shortened sample relative to those in pseudo real time (Chart 11). For example, in the first forecast horizon in the real-time estimation, 15 models are given an above-average weight in the predictive density (i.e., greater than 1). Contrast this with the pseudo-real-time weights, which show fewer models getting substantial weight, and almost all of those are BVARs. Only 7 models in the pseudo-real-time estimation get more than 1/98 weight. This difference increases as the prediction horizons progress: by the final forecast horizon, 23 models receive above-average weights in the real-time estimation, compared with only 8 models in pseudo real time.

Moreover, these figures reinforce our earlier finding that the BVAR block is favoured by the MM algorithm in pseudo real time, though this result is true across all 12 forecast horizons rather than 4 to 6 horizons in the full-sample estimation. This appears to be driven by the fact that the BVAR's sharpness tends to benefit more from the use of revised data than the other classes and, in turn, is likely due to a larger reliance on the monthly GDP data, whose signal deteriorates in real time because of revisions. The more even distribution of weights across all model classes in our full-sample pseudo-real-time results suggests to us that the complete dominance of the BVAR block shown in Chart 11 is largely a function of our subsample. That said, as noted in Section 4.2, in the final four forecasting periods, the BVAR block receives approximately 75% of the weight in the predictive densities by the end of the forecasting exercise in the fourth quarter of 2019. Our real-time results suggest that these weights could be somewhat biased upward due to the use of pseudo-real-time data.

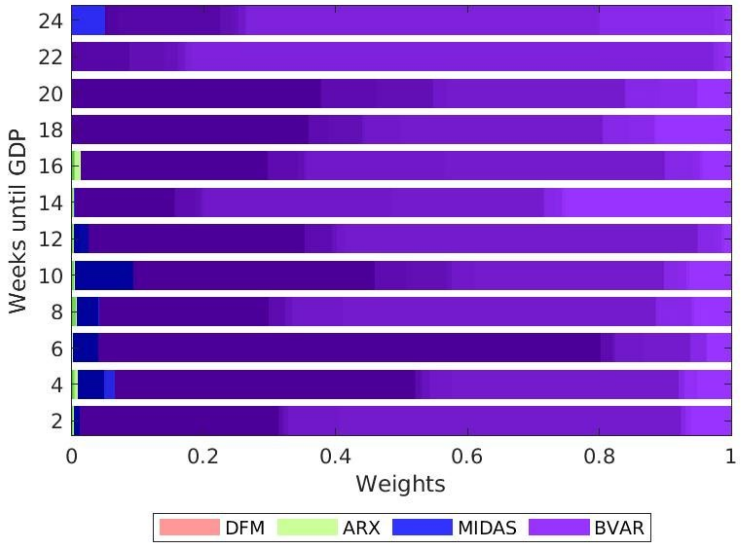
Interestingly, several of the models that see their weights increase significantly in the real-time estimation are those that leverage indicators that are seldom revised or not revised at all, for example Labour Force Survey data (which are revised following the release of final population estimates after each census) or unrevised data such as Baker Hughes' drilling rig counts, the US PMI, and stock market data. A possible explanation is that the pseudo-real-time estimation is boosting weights on models that rely on data that provide a good signal after revisions but not in real time. However, we also find that the weight increases on several leading-indicator and MIDAS models that use data subject to revisions, namely, data on manufacturing sales, retail sales (both Canadian and US), terms of trade and US industrial production.

Chart 10: Real-time one-stage minorization-maximization combination weights, end-of-sample, all horizons



Note: The abbreviations in the legend are DFM, dynamic factor models; ARX, autoregressive leading indicator models; MIDAS, mixed-data sampling; BVAR, Bayesian vector autoregression.

Chart 11: Pseudo-real-time one-stage minorization-maximization combination weights, end-of-sample, all horizons



Note: The abbreviations in the legend are DFM, dynamic factor models; ARX, autoregressive leading indicator models; MIDAS, mixed-data sampling; BVAR, Bayesian vector autoregression.

5.4 Case study: Real-time predictions for the second quarter of 2020

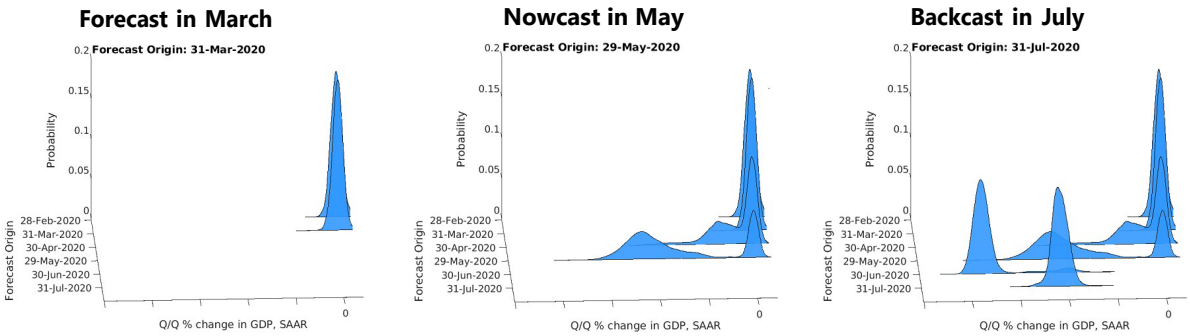
This section studies predictions during the COVID-19 pandemic.[14] The pandemic presents an excellent opportunity to study the predictions from the model in a time of extreme uncertainty and economic importance. Chart 12 shows that despite the historic nature of the 38.7% (quarterly, annualized rate) contraction in real GDP observed in the second quarter of 2020, the density combination platform presented in this paper did a remarkably good job of reacting to the data and generating predictions equally as historic as the eventual outturn of GDP growth.[15]

The first panel of Chart 12 illustrates the predictive densities generated at the end of February and March 2020. At that time, few hard data were available for the first quarter of 2020, let alone the second quarter; as a result, the density predictions remained centred just above zero despite the imposition of containment measures in many major jurisdictions across Canada. Of course, the deterioration in the densities resulting from the weakness observed in the more timely data received in March and April was not sufficient for the models to predict the full magnitude of the contraction in GDP that was to come . That said, by April 29, 2020—a day before the release of February's monthly GDP data—the growing left tail of the predictive density suggested an 80% probability of real GDP contracting. Moreover, it had already pulled down its mean prediction to -9.7%, which at that time already marked the sharpest quarterly decline in real GDP ever recorded. Over the course of May 2020, the ongoing deterioration across virtually all indicators moved the densities into unprecedented negative territory. First was the Labour Force Survey data released on May 8, 2020, which pulled the mean of the densities down to -18.5%. This was followed by the Monthly Survey of Manufacturing data released on May 14, which resulted in a mean of -28%, a median of -37% and a probability of contraction of 80%—and all of this even before the release of March's initial contraction in real GDP. The nowcast for the second quarter of 2020 as of May 29, 2020 (as shown in the centre panel of Chart 12) shows the density prediction after incorporating March's observation for monthly GDP, which recorded an historic decline of 7.3% (monthly, unannualized rate). This observation edged the mean of the predictive density down further to -32%. More importantly, however, it shifted the entire mass of the predictive distribution into negative territory, predicting that a contraction in real GDP was essentially certain, all without having any GDP data for the actual reference quarter, the second quarter of 2020.

[14] So far, we have omitted analysis of the pandemic since it features large outliers that, beyond worsening the sharpness results, do not substantively change the results.

[15] For the remainder of this section, all growth rates are expressed at quarterly annualized rates unless otherwise noted.

Chart 12: Predictive densities for 2020Q2



Note: Q/Q refers to quarter-over-quarter percent change. SAAR refers to seasonally adjusted at annual rates.

Of course, judgement would allow practitioners to anticipate the weakness in these variables in an even more timely fashion than the models through the use of news-based information (e.g., announcements of containment measures) and high-frequency data (such as the daily or weekly transaction-level spending data being leveraged by several commercial banks). Nonetheless, the real-time performance of the judgement-free density combinations in the second quarter of 2020, even without these sources of news, speaks to the potential for this as a nowcasting tool. However, this episode does highlight the importance of off-model information and unconventional data, which could be an avenue to improve the combined predictive densities.

6 Conclusion

In this paper, we propose a new platform for generating and combining predictive densities for Canadian real GDP growth. In addition to the combined predictive densities being accurate (sharp), we also find a combination method that produces robustly calibrated densities—the one-stage MM weighting. This is an important result because it confirms that density nowcasting provides a viable and, we argue, far superior alternative to point forecasts in a policy-making setting. The robustness of our findings is reinforced by evaluating the framework on a novel real-time dataset (a first for Canada), which also reinforces the importance of data revisions to nowcasting accuracy. Calibrated predictive densities mean that policy-makers can confidently leverage the information from the predictive density to evaluate uncertainty, which, in turn, can help identify and quantify tail risks. None of this is possible with traditional point forecasting tools. Overall, we view our findings as an important step forward in the nowcasting of Canadian real GDP growth. We believe that monitoring the evolution of these predictive densities will prove a useful tool for policy-makers in Canada.

A Appendix

A.1 Data appendix

Table A-1: Data table

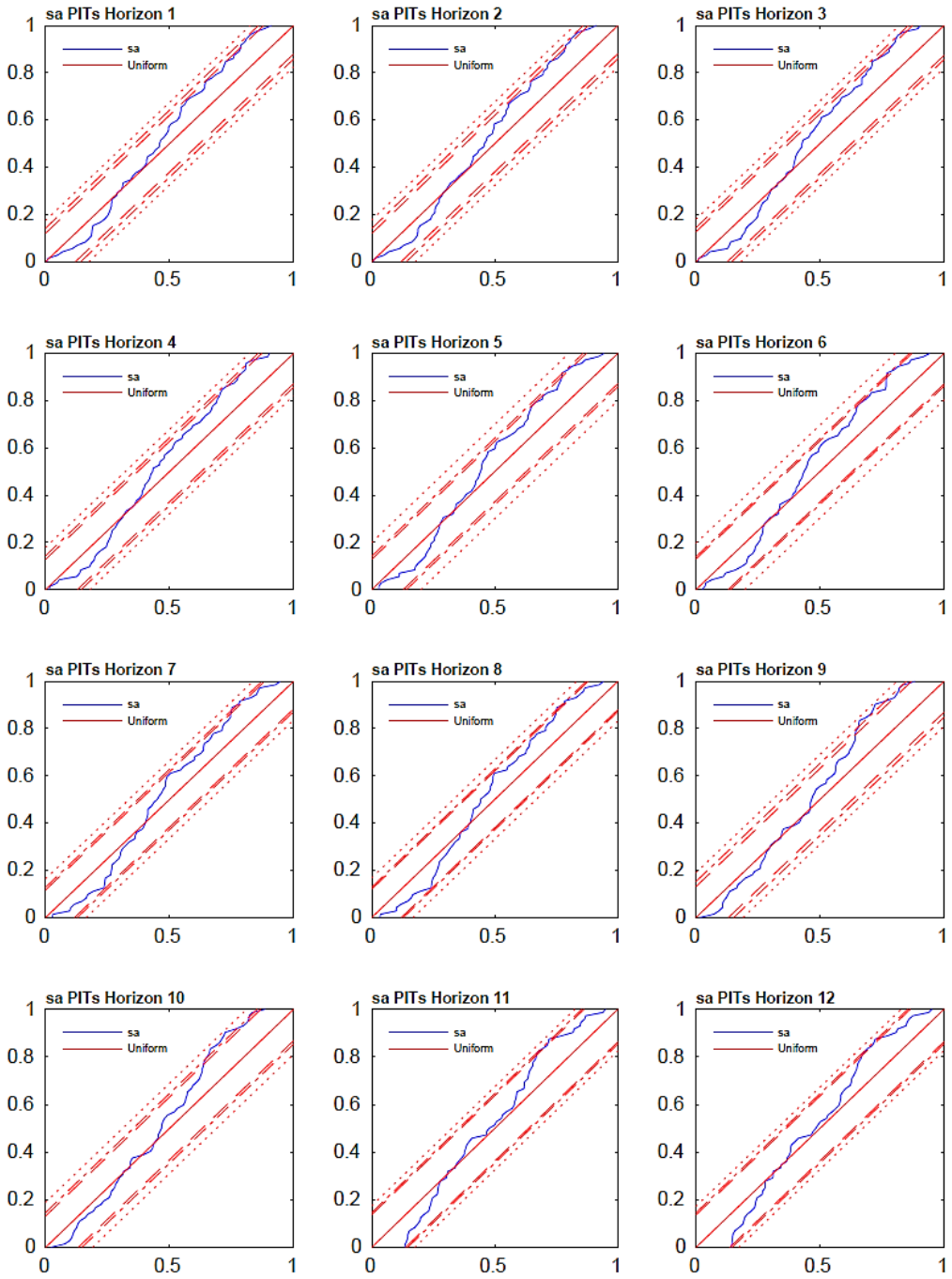
Series	Release	Start date	Transformation*
GDP at market prices (quarterly)	End	Q2-61	4
Monthly GDP	End	Jan-81	3
Canadian motor vehicle production	Middle	Jan-93	5
Finished goods price index	Beginning	Jan-72	3
Canadian terms of trade	Beginning	Jan-81	3
Manufacturing new orders	Middle	Jan-81	5
Railway carloadings	End	Jan-70	5
US industrial production	Middle	Jan-21	3
US retail sales	Middle	Jan-67	3
Wholesale trade	Middle	Jan-81	3
Total actual hours worked	Beginning	Jan-76	3
Retail trade	End	Jan-81	3
Canadian car sales	Beginning	Jan-46	3
Housing starts	Middle	Jan-56	3
US housing starts	Middle	Jan-59	3
Building permits	Beginning	Jan-48	5
Merchandise exports	Beginning	Jan-68	3
Merchandise imports	Beginning	Jan-68	3
Toronto Stock Exchange	Beginning	Jan-56	3
Consumer confidence index	End	Mar-61	3
Global purchasing managers' index	Beginning	Jan-98	3
US car sales	Beginning	Jan-76	5
Food services and drinking places prices	End	Jan-81	5
Employment rate—Labour Force Survey	Beginning	Jan-76	2
Chicago Fed National Activity Index	Beginning	Mar-67	0
West Texas Intermediate crude oil prices	End	Jan-72	3
Baker Hughes rig count	Middle	Jan-68	5
Baltic Dry Index	Beginning	Jan-85	3
Manufacturing Sales Indicator	Middle	Jan-97	5
US Purchasing Managers' Index	Beginning	Jan-48	3
Monthly residential unit sales—MLS	Middle	Jan-80	3

* The transformation applied to each of the series: 0 = no change, 1 = log level, 2 = M/M difference, 3 = M/M log difference, 4 = Q/Q log difference, 5 = seasonally adjusted (x12) M/M log difference, 6 = term spread, 7 = two-month moving average

A.2 Results appendix

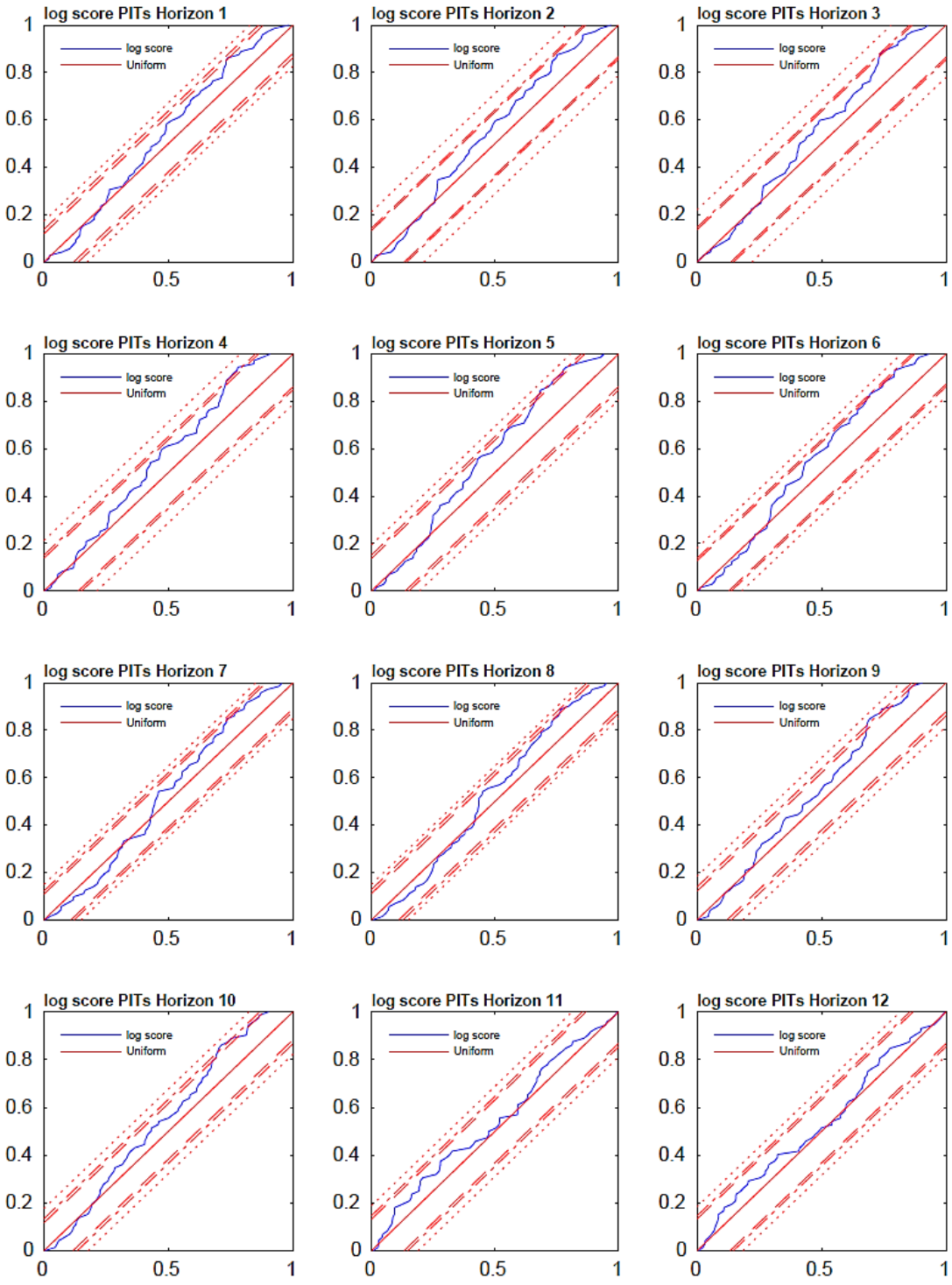
A.2.1 Probability integral transforms for the cumulative distribution functions of the combined densities with Rossi and Sekhposyan (2019) confidence intervals

Chart A-1: Probability integral transforms—simple average combinations



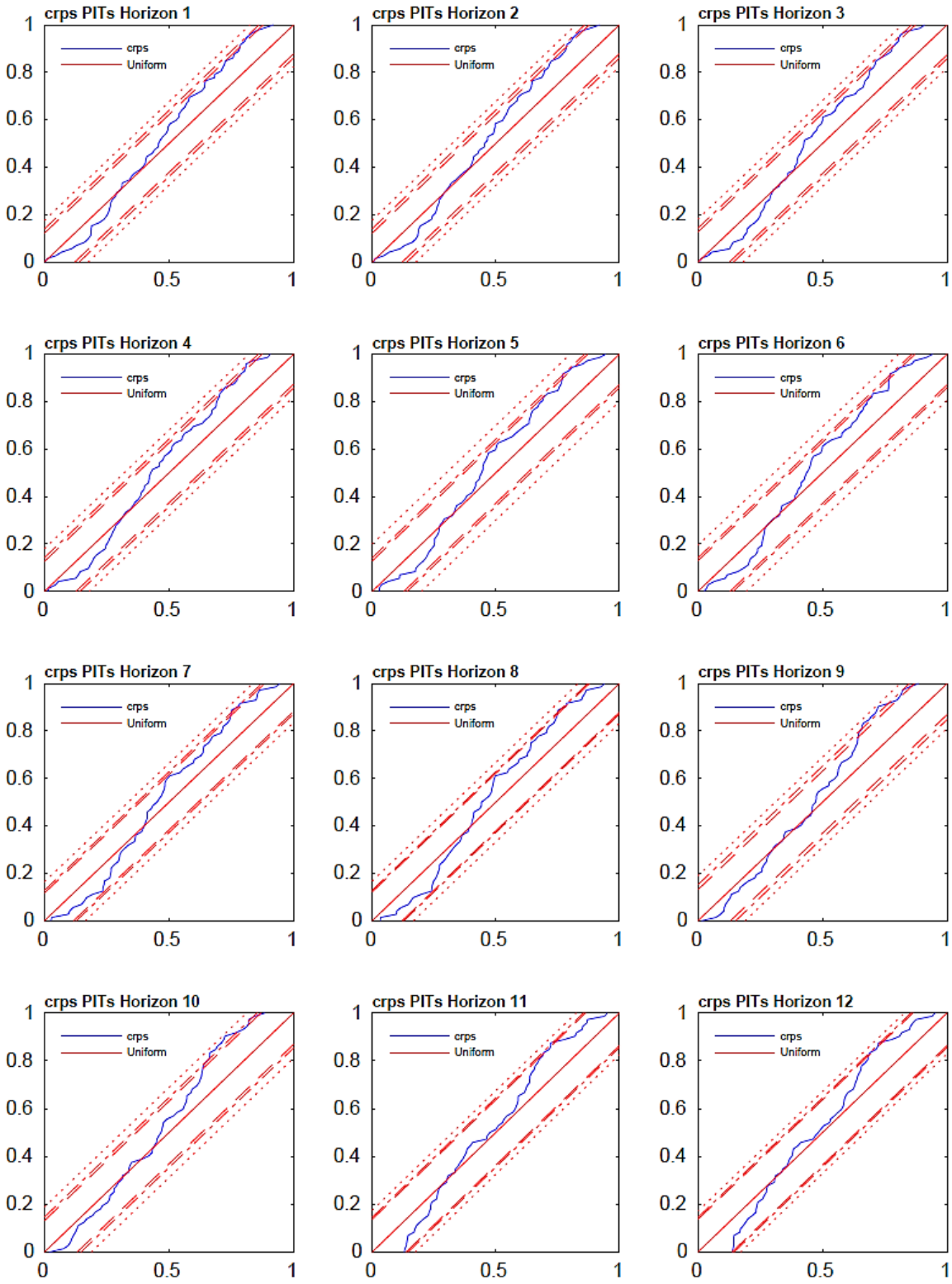
Note: These charts compare probability integral transforms (PITs) for the stated combination method (plotted in blue) against the uniform benchmark at all forecast horizons as described in section 3.3. MM refers to minorization-maximization. The x-axis reflects values of the cumulative distribution function (CDF), while the y-axis shows the value of the PITs at each value of the CDF. The dotted lines around the uniform benchmark are the 90% (long-dashed line), 95% (short-dashed line) and 99% (dotted line) confidence intervals constructed using the test statistics proposed by Rossi and Sekhposyan (2019).

Chart A-2: Probability integral transforms—one-stage log score combinations



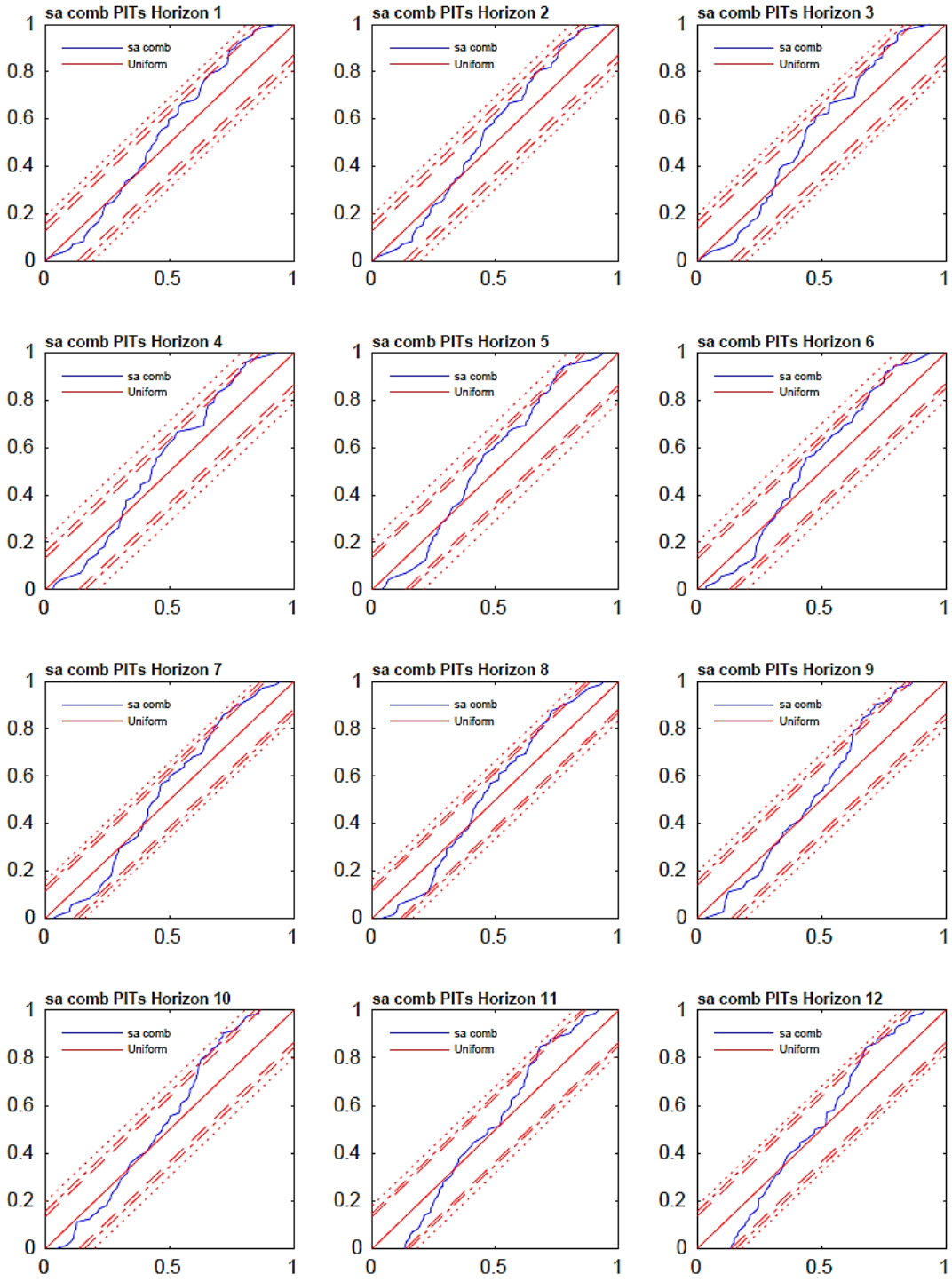
Note: These charts compare probability integral transforms (PITs) for the stated combination method (plotted in blue) against the uniform benchmark at all forecast horizons as described in section 3.3. MM refers to minorization-maximization. The x-axis reflects values of the cumulative distribution function (CDF), while the y-axis shows the value of the PITs at each value of the CDF. The dotted lines around the uniform benchmark are the 90% (long-dashed line), 95% (short-dashed line) and 99% (dotted line) confidence intervals constructed using the test statistics proposed by Rossi and Sekhposyan (2019).

Chart A-3: Probability integral transforms—one-stage continuous rank probability score combinations



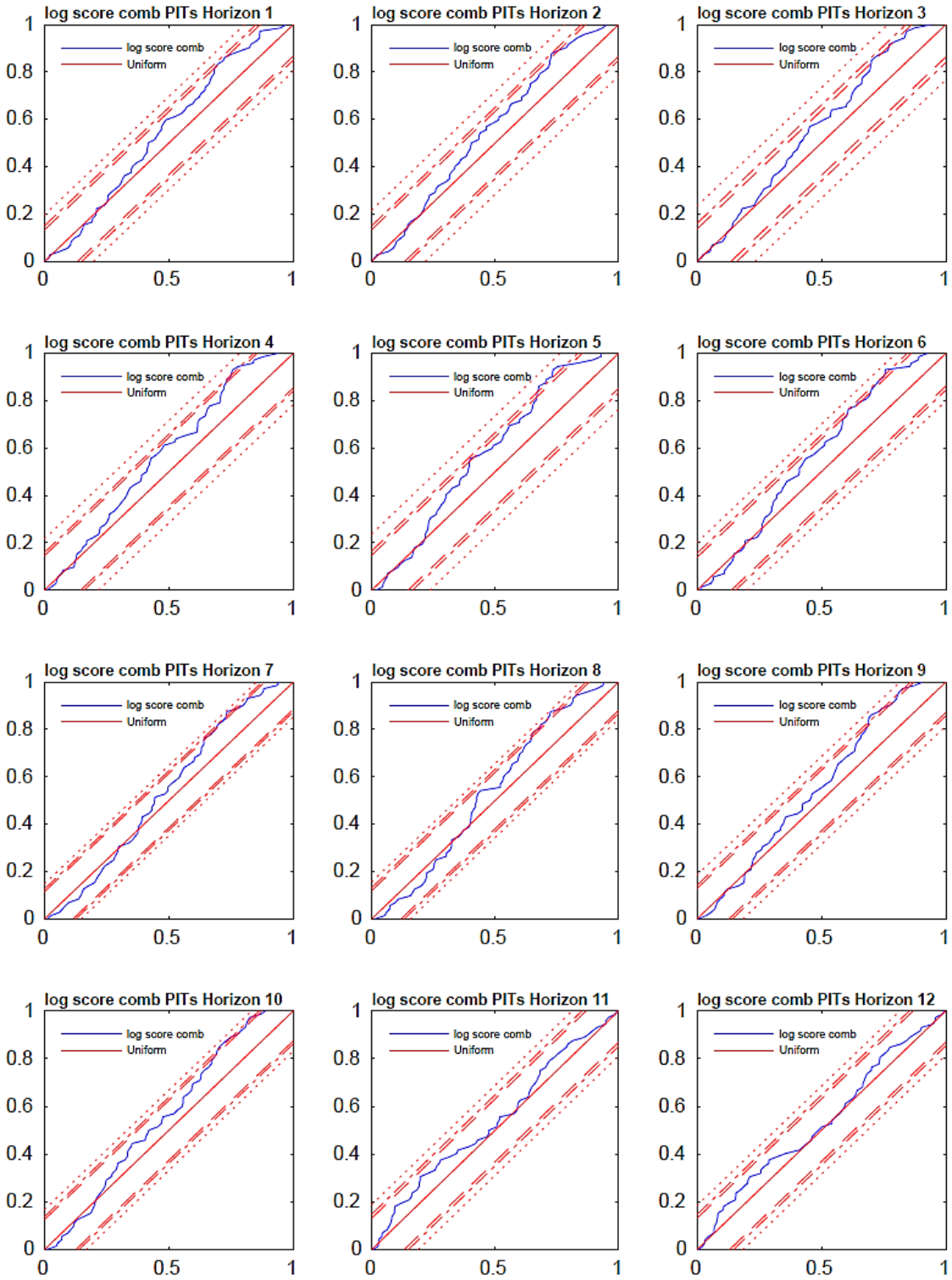
Note: These charts compare probability integral transforms (PITs) for the stated combination method (plotted in blue) against the uniform benchmark at all forecast horizons as described in section 3.3. MM refers to minorization-maximization. The x-axis reflects values of the cumulative distribution function (CDF), while the y-axis shows the value of the PITs at each value of the CDF. The dotted lines around the uniform benchmark are the 90% (long-dashed line), 95% (short-dashed line) and 99% (dotted line) confidence intervals constructed using the test statistics proposed by Rossi and Sekhposyan (2019).

Chart A-4: Probability integral transforms—two-stage simple average combinations



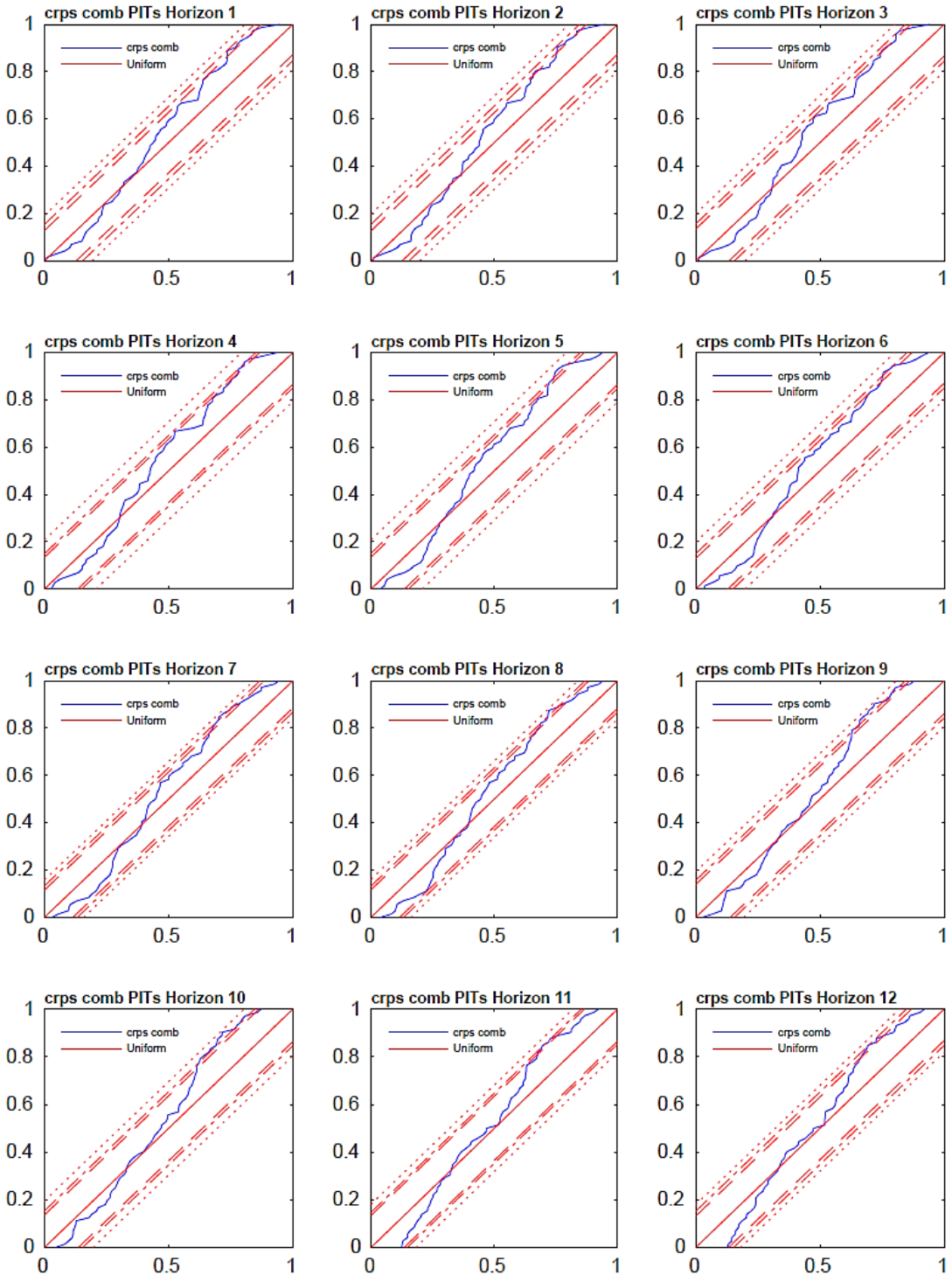
Note: These charts compare probability integral transforms (PITs) for the stated combination method (plotted in blue) against the uniform benchmark at all forecast horizons as described in section 3.3. MM refers to minorization-maximization. The x-axis reflects values of the cumulative distribution function (CDF), while the y-axis shows the value of the PITs at each value of the CDF. The dotted lines around the uniform benchmark are the 90% (long-dashed line), 95% (short-dashed line) and 99% (dotted line) confidence intervals constructed using the test statistics proposed by Rossi and Sekhposyan (2019).

Chart A-5: Probability integral transforms—two-stage log score combinations



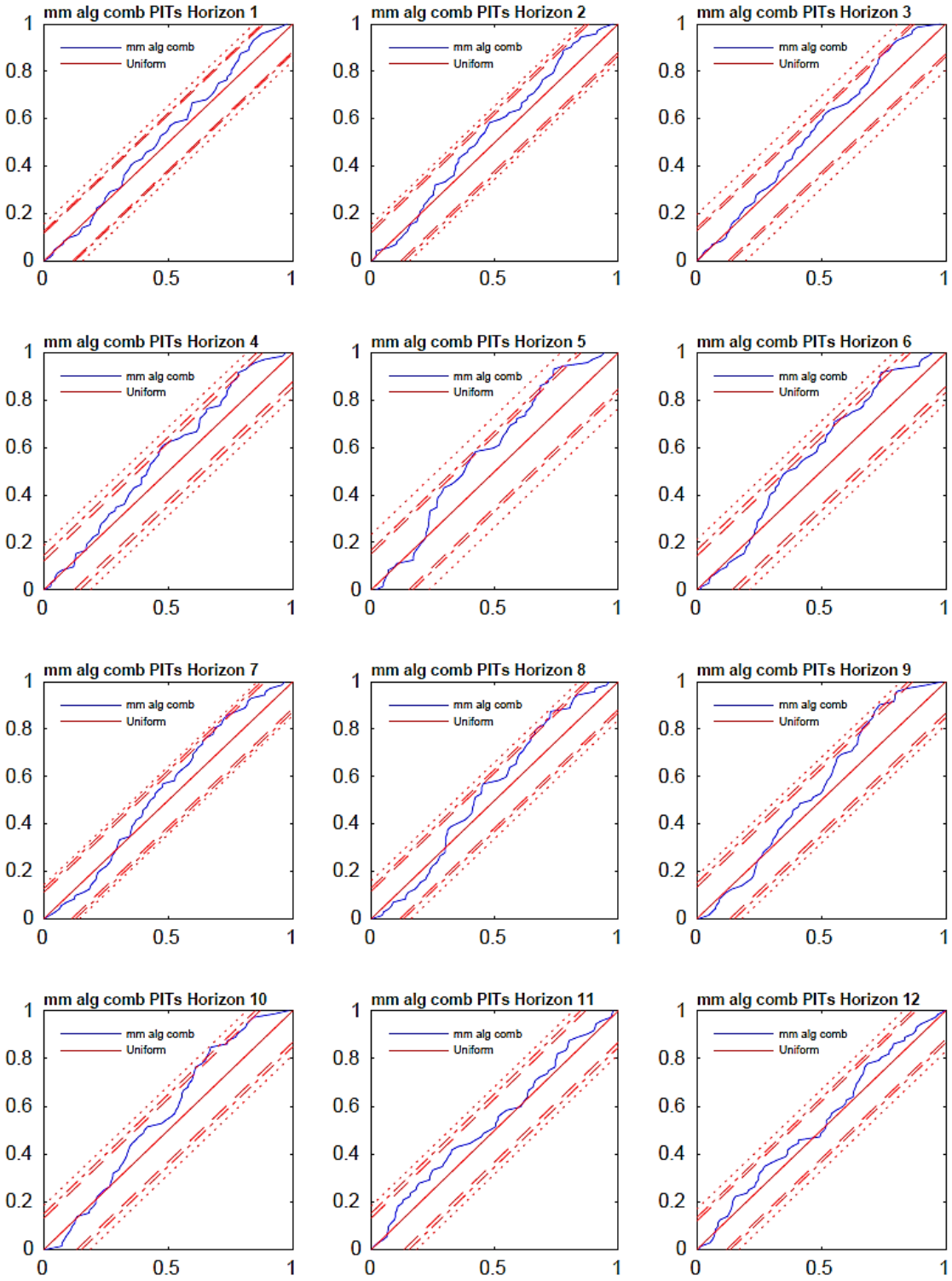
Note: These charts compare probability integral transforms (PITs) for the stated combination method (plotted in blue) against the uniform benchmark at all forecast horizons as described in section 3.3. MM refers to minorization-maximization. The x-axis reflects values of the cumulative distribution function (CDF), while the y-axis shows the value of the PITs at each value of the CDF. The dotted lines around the uniform benchmark are the 90% (long-dashed line), 95% (short-dashed line) and 99% (dotted line) confidence intervals constructed using the test statistics proposed by Rossi and Sekhposyan (2019).

Chart A-6: Probability integral transforms—two-stage continuous rank probably score combinations



Note: These charts compare probability integral transforms (PITs) for the stated combination method (plotted in blue) against the uniform benchmark at all forecast horizons as described in section 3.3. MM refers to minorization-maximization. The x-axis reflects values of the cumulative distribution function (CDF), while the y-axis shows the value of the PITs at each value of the CDF. The dotted lines around the uniform benchmark are the 90% (long-dashed line), 95% (short-dashed line) and 99% (dotted line) confidence intervals constructed using the test statistics proposed by Rossi and Sekhposyan (2019).

Chart A-7: Probability integral transforms—two-stage minorization-maximization combinations



Note: These charts compare probability integral transforms (PITs) for the stated combination method (plotted in blue) against the uniform benchmark at all forecast horizons as described in section 3.3. MM refers to minorization-maximization. The x-axis reflects values of the cumulative distribution function (CDF), while the y-axis shows the value of the PITs at each value of the CDF. The dotted lines around the uniform benchmark are the 90% (long-dashed line), 95% (short-dashed line) and 99% (dotted line) confidence intervals constructed using the test statistics proposed by Rossi and Sekhposyan (2019).

A.2.2 Comparing combination methods and individual models

As discussed in section 1, both theoretical motivations and empirical evidence support combining density forecasts. This section examines the benefits of combining predictions in the Canadian context by comparing the performance of the combined density predictions with those from the individual models.

Chart A-9 plots the continuous rank probability score (CRPS) of the one-stage minorization-maximization (MM) combination (solid line) against the range of CRPS from the individual models (shaded area). The MM algorithm predictions, across all forecast horizons, are in the bottom 20% of CRPS scores, and for 60% of the time they are in the bottom 10%. This indicates that the MM combination methods create relatively accurate predictions.

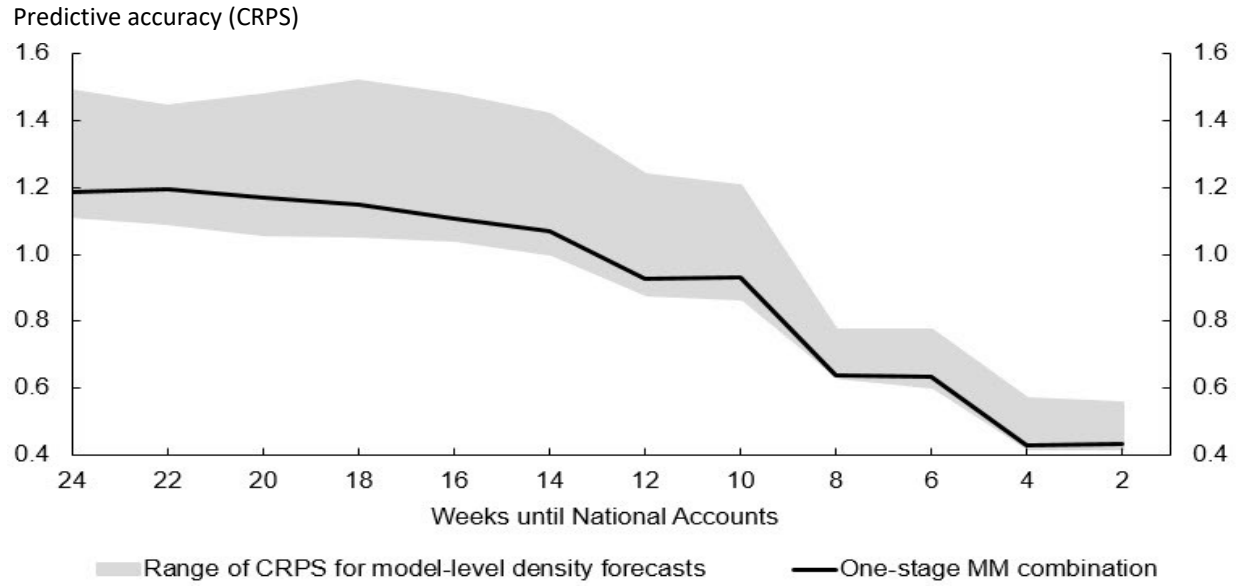
While some individual models are relatively more accurate than the one-stage MM combination density, generally speaking the most accurate models are not calibrated. To illustrate the point, Table A-2 displays the best-performing model at a given prediction horizon and the associated Knüppel test results. The far-right column shows the CRPS ratio relative to the one-stage MM algorithm. These results suggest that the best performing models could outperform the MM by at most 12%, but the null hypothesis of calibrated densities is rejected in most cases at the 5% level.

A complication with this analysis is that at any forecast horizon there are a handful of models that may outperform the one-stage MM algorithm, and we are only comparing against the best. Closer inspection reveals that there are 135 occurrences of a model outperforming the combination over all the forecast horizons, but many of these models only outperform the MM slightly and for a single horizon. A different way to make this comparison would be to sum the CRPS across all forecast horizons and compare models against combination methods. Doing this, we find only three of the four dynamic factor models (DFMs) outperform the one-stage MM algorithm—and by a maximum of 6%—and that the models do not show evidence of calibration over most forecast horizons.[1]

These results suggest that the strategy of combining density forecasts is a reliable method of producing sharp and calibrated predictions. The one-stage MM algorithm does not always produce the sharpest predictions, but they are among the sharpest while maintaining calibration. In short, the one-stage MM algorithm is the method that maximizes sharpness subject to calibration. Moreover, the trade-off for achieving calibration is extremely small (amounting to between one-tenth and one-hundredth of a percentage point in annualized real GDP growth).

[1] Over the 36 forecast horizons, there are only 8 cases where the Knüppel test is passed at 5% and only 4 cases where it is passed at 10%.

Chart A-9: Continuous rank probability score—one-stage minorization-maximization combination compared with range of model



Note: CRPS refers to continuous rank probability score and MM refers to minorization-maximization. The shaded area shows the range of CRPS computed from each of the 98 models underpinning the one-stage MM combination, shown in black.

Table A-2: Relative performance of minorization-maximization algorithm

	Weeks until GDP	Best Model	Knupple p-value	CRPS ratio to MM
Forecast	24	BDFM3	0.01	1.07
	22	BDFM3	0.02	1.10
Nowcast	20	MIDAS20	0.10	1.12
	18	BDFM4	0.01	1.10
	16	MIDAS20	0.07	1.07
	14	BDFM3	0.05	1.08
	12	BDFM3	0.05	1.06
	10	BDFM3	0.03	1.09
Backcast	8	BVAR5	0.52	1.03
	6	BDFM4	0.00	1.07
	4	BVAR2	0.07	1.05
	2	BVAR13	0.07	1.05

Note: Red cells highlight p-values that reject the null hypothesis of calibration at the 5% confidence level. Yellow cells highlight p-values between 5% and 10%, and green cells highlight p-values that do not reject calibration at the 10% significance level.

A.3 Model appendix

A.3.1 Leading indicators

A commonly used nowcasting model is the leading indicator or bridge model. These simple models are an autoregression plus an additional indicator. More formally,

$$y_{t_m+h} = \alpha + \sum_{i=1}^{p} \rho_i y_{t_{m-i}} + \beta_1 x_{i,t+h_m} + \varepsilon_{t_m}, \tag{1}$$

where $x_{i,t+h}$ are the remaining monthly indicators.

Density forecasts are produced with a block wild bootstrap procedure. For each bootstrap iteration, the following procedure is followed.

First, the explanatory variables are forecast out to the required horizon. This process starts by estimating the following equation:

$$x_{t_m+1} = \alpha + \sum_{i=1}^{p} \rho_i y_{t_{m-i}} + \epsilon_t \tag{2}$$

Using the model parameters, generate the bootstrap data $\hat{x}_t$ with the following equation:

$$\hat{x}_t = \hat{\alpha} + \sum_{i=1}^{p} \hat{\rho}_i x_{t_{m-i}} + \hat{\epsilon}_t, \tag{3}$$

where $\hat{\epsilon}_t$ is drawn from the vector of residuals (ϵ) in blocks of size h, and each element is multiplied by a variable drawn from the Rademacher distribution $f(k)$.

$$f(k) = \begin{cases} 1/2, & \text{if } k = -1 \\ 1/2, & \text{if } k = +1 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Using the bootstrapped data $\hat{x}_t$, re-estimate the auxiliary equation (2) $\ddot{\alpha}$, $\ddot{\rho}$, $\ddot{\sigma}_\epsilon$.

Retrieve the fitted values $\ddot{x}_t$ for $t = 1 \ldots T \ldots T + h$ from an equation using the bootstrapped parameters,

$$\ddot{x}_t = \ddot{\alpha} + \sum_{i=1}^{p} \ddot{\rho}_i x_{t_{m-i}} + \ddot{\epsilon}_t \tag{5}$$

where $\ddot{\epsilon}_t$ is drawn from the vector of residuals (ϵ) in blocks of size h, and each element is multiplied by a variable drawn from the Rademacher distribution $f(k)$.

Equation 1 is estimated, and parameters $\hat{\alpha}, \hat{\rho}, \hat{\beta}$ and residuals $\hat{\varepsilon}_{t_m}$ are retrieved.

Using these data, the bootstrap dataset $\hat{y}_t$ is constructed from the formula

$$\hat{y_{t_m+h}} = \hat{\alpha} + \sum_{i=1}^{p} \hat{\rho}_i y_{t_{m-i}} + \hat{\beta}_1 \ddot{x}_{i,t+h_m} + \hat{\varepsilon}_{t_m}, \tag{6}$$

where $\hat{\varepsilon}_{t_m}$ is drawn from the vector of residuals (ε) in blocks of size h, and each element is multiplied by a variable drawn from the Rademacher distribution $f(k)$.

Using the bootstrapped data $\hat{y}_t$ and $\ddot{x}_{t+h_m}$, the bootstrapped model parameters $\ddot{\alpha}, \ddot{\rho}, \ddot{\beta}$ and residuals $\ddot{\varepsilon}_{t_m}$ are retrieved.

The forecast is then

$$\ddot{y}_{t_m+h} = \ddot{\alpha} + \sum_{i=1}^{p} \ddot{\rho}_i y_{t_{m-i}} + \ddot{\beta}_1 \ddot{x}_{i,t+h_m} + \hat{\varepsilon}_{t_m} \tag{7}$$

We then convert the iterated monthly forecasts for y_{t_m} into quarterly predictions for real GDP growth. Using equation 8 below, we follow the same block wild bootstrap procedure discussed above. This is necessary since monthly GDP growth expressed at quarterly rates is not identicalto quarterly real GDP growth as published in the national accounts (see section 2 for additional details).

$$GDP_{quarterly,t} = \alpha_{GDPquarterly} + \beta_{GDPmonthly} GDP_{monthly,t} + \varepsilon_{GDPquarterly,t} \tag{8}$$

A.3.2 Mixed data sampling (MIDAS) models

A more modern benchmark model is the MIDAS regression (Ghysels, Santa-Clara and Valkanov 2004). The defining feature of MIDAS models is the way they deal with mixed frequencies. These models use a polynomial weighting function to link high-frequency regressors onto a low-frequency regressand.This makes the MIDAS regression a direct forecasting tool, which does not explicitly model the dynamics

of the indicator. Instead, the MIDAS directly relates future quarterly GDP to present and lagged high-frequency indicators. This necessitates a model for each forecast horizon.

The basic model for forecasting h_q quarters ahead with $h_q = h_m/3$ is:

$$y_{t_q+h_q} = y_{t_m+h_m} = \beta_0 + \beta_1 b(L_m, \theta) x^{(3)}_{t_m+w} + \varepsilon_{t_m+h_m}, \tag{9}$$

where y_{tm} is GDP growth and $x^{(3)}_{t_m}$ is the corresponding skip-sampled monthly indicator, L_m is the monthly lag operator, and $w = T_x - T_y$. The lag polynomial $b(L_m, \theta)$ is defined as:

$$b(L_m, \theta) = \sum_{k=0}^{K} c(k; \theta) L_m^k. \tag{10}$$

The parsimonious parametrization of the lagged coefficients $c(k; \theta)$ is one of the key features of MIDAS models. While there are several common ways to parameterize the lagged coefficients, we choose the so-called "Beta Lag":

$$c(k, \theta_1, \theta_2) = \frac{f(\frac{k}{K}, \theta_1; \theta_2)}{\sum_{k=1}^{K} f(\frac{k}{K}, \theta_1; \theta_2)}, \tag{11}$$

where $f(k, \theta_1, \theta_2) = \frac{k^{\theta_1-1}(1-k)^{\theta_2-1}\Gamma(\theta_1+\theta_2)}{\Gamma(\theta_1),\Gamma(\theta_2)}$, $\Gamma(\theta) = \int_0^\infty e^{-x} x^{\theta-1} dx$ and parameters θ_1 and θ_2 govern the shape of the distribution. This parametrization is quite general and can take various shapes with only a few parameters. These include increasing, decreasing or hump-shaped patterns. Furthermore, we restrict the last lag to be equal to zero.

The MIDAS model is estimated using nonlinear least squares (NLS) in a regression of y_t onto $x^{(3)}_{t-h}$ for each forecast horizon $h = 1, \ldots, H$.

The direct forecast is given by the conditional expectation:

$$\hat{y}_{T_y+h|T_x} = y_{t_m+h_m} = \hat{\beta}_0 + \hat{\beta}_1 b(L_m, \theta) x^{(3)}_{t_{m+w}}, \quad (12)$$

where $T_x = T_y + \omega$ is such that the most recent observations of the indicator are included in the conditioning set of the projection. For example, if we were trying to forecast GDP for the second quarter and the July Purchasing Managers Index was available, the regression would include a lead of our indicator.

Since Canada has a monthly GDP measure, it is necessary to extend the basic MIDAS model to have multiple explanatory variables. Furthermore, we include a low-frequency AR term. The forecasting model then becomes

$$y_{t_q+h_q} = y_{t_m+h_m} = \beta_0 + \beta_1 b(L_m, \theta_1) x^{(3)}_{1,t_{m+w-h_m}} + \beta_2 b(L_m, \theta_2) x^{(3)}_{2,t_{m+w-h_m}} + \lambda y_{t_m} + \varepsilon_{t_m+h_m} \quad (13)$$

with $x_1^{(3)}$ being monthly GDP, and $x_2^{(3)}$ an additional leading indicator. As in the bridge models, we take the same set of leading indicators, create a model with each and average the individual forecasts with equal weights to create the MIDAS class forecast.

To generate predictive densities for the MIDAS models, we employ the same block wild bootstrapping algorithm discussed in subsection A.3.1 above. The one exception is that, since the MIDAS model forecasts quarterly GDP growth directly, the final step of bootstrapping over the monthly-to-quarterly bridge equation is not necessary: the imperfect mapping between them is covered in the use of monthly GDP as an indicator variable in the MIDAS regression.

A.3.3. Bayesian vector autoregressions

Bayesian vector autoregressions (BVARs) are a staple workhorse model used in economics for both structural analysis and, more importantly for this paper's purposes, forecasting. In a macroeconometric setting, practitioners are generally faced with a trade-off between comprehensiveness and parsimony: while one wants to include enough data to properly capture the economic mechanisms being modelled, one also wants to avoid over-parameterization. A key advantage of Bayesian estimation of VARs is that it allows the user to work with a larger set of explanatory variables by limiting the number of parameters to be estimated and helping avoid over-parameterization.

In this paper, we estimate VAR(4) models. Each of the VARs is estimated twice, once using the full sample of data available at any given point in the out-of-sample evaluations, and another using a rolling window. We

present an overview of the model structure below and refer the reader to Koop and Korobilis (2010) and Karlsson (2013) for more detailed expositions of BVARs, the Minnesota prior and the sampling algorithms.

Our VARs follow the traditional form:

$$Y_t = X_t A + E. \tag{14}$$

Let $\hat{A}$ be the OLS estimates of A and $S = (Y - X\hat{A})'(Y - X\hat{A})$ then $\hat{\Sigma} = S/(T - K)$.

Or, stacking the variables in vector notation,

$$y_t = (I_M \otimes X_t)a + \epsilon\,, \epsilon \sim N(0, \Sigma \otimes I_t), \tag{15}$$

where $t = 1$ to T, $M = 6$ and $\rho = 4$, and $K = 1 + M\rho$ is the number of coefficients in each equation of the VAR.

We also use the Minnesota prior, which results in a posterior of the following form:

$$a \mid y \sim N(\bar{a}_{Mn}, \bar{V}_{Mn}), \tag{16}$$

where

$$\bar{V}_{Mn} = [\underline{V}_{Mn}^{-1} + (\hat{\Sigma}^{-1} \otimes (X'X))]^{-1} \tag{17}$$

and

$$\bar{a}_{Mn} = \bar{V}_{Mn}[\underline{V}_{Mn}^{-1}\underline{a}_{Mn} + (\hat{\Sigma}^{-1} \otimes X)'y], \tag{18}$$

with the prior mean $\underline{a}_{Mn}$ (set equal to 0) and the Minnesota prior covariance matrix $\underline{V}_{Mn}$.

Following Karlsson (2013), we employ algorithm 1 to sample from predictive distribution (treating Σ as known) and algorithm 22 to sample a. We generate iterated monthly predictionsfor the variables contained in each BVAR until the end of the second reference quarter being forecast. Finally, we convert these monthly GDP forecast profiles into quarterly predictive distributions for annualized real GDP growth.

A.3.4 Dynamic factor models

Model set-up

The model we are estimating is, in general terms, of this form:

$$y_{0t} = \lambda_0 f_t + \varepsilon_0 t \tag{19}$$

$$y_{1t} = \lambda_1 f_t + \upsilon_t + \varepsilon_1 t \tag{20}$$

$$y_{2t} = 1/3\lambda_1 f_t + 2/3\lambda_1 f_{t-1} + \lambda_1 f_{t-2} + 2/3\lambda_1 f_{t-3} + 1/3\lambda_1 f_{t-4} \tag{21}$$

$$+ 1/3\upsilon_t + 2/3\upsilon_{t-1} + \upsilon_{t-2} + 2/3\upsilon_{t-3} + 1/3\upsilon_{t-4} + \varepsilon_2 t, \tag{22}$$

where y_{0t} is a vector of n observables; y_{1t} and y_{2t} are closely related variables (different measures of GDP), one quarterly and one monthly; λ_0 is a vector of q factor loadings and f_t is a vector of q factors, while $\varepsilon_t \sim \mathsf{N}(0, \sigma_\varepsilon)$ and υ_t is a common factor.

The factor process is:

$$f_t = Af_{t-1} + \ldots + Af_{t-p} + \eta_t \tag{23}$$

$$\upsilon \sim \mathsf{N}(0, \sigma_\upsilon). \tag{24}$$

f_t is a vector of q factors, which follows a VAR process of order p, with $\eta_t \sim \mathsf{N}(0, \Omega_\eta)$.

Mixed frequencies

Since this model is used for nowcasting GDP, which is a quarterly frequency, and most of the predictors we use are monthly, we need to modify the model. The technique we use is from Mariano and Murasawa (2003).

Quarterly series are incorporated into the model by expressing them in terms of their partially observed monthly counterparts. Quarterly variable, like GDP (GDP_t^Q), are expressed as the sum of their unobserved monthly contributions (GDP_t^M):

$$GDP_t^Q = GDP_t^M + GDP_{t-1}^M + GDP_{t-2}^M \tag{25}$$

for $t = 3, 6, 9, \ldots$. define $Y_t^Q = 100 \times \log(GDP_t^Q)$ and $Y_t^M = 100 \times \log(GDP_t^M)$

To link y_t with the observed quarterly GDP series, we construct a partially observed monthly series:

$$y_t^Q = \begin{cases} y_t^Q - y_{t-3}, & t = 3, 6, 9 \\ unobserved, & otherwise \end{cases}$$

and use the approximation of Mariano and Murasawa (2003) to obtain:

$$\begin{aligned} y_t^Q = Y_t^Q - Y_{t-3}^Q &\approx (Y_t^M + Y_{t-1}^M + Y_{t-2}^M) - (Y_{t-3}^M + Y_{t-4}^M + Y_{t-5}^M) \\ &= y_t + 2y_{t-1} + 3y_{t-2} + 2y_{t-3} + y_{t-4}. \end{aligned} \tag{26}$$

Estimation

The model is estimated using a Markov chain Monte Carlo scheme iterated 15,000 times, discarding the first 5,000 draws as burn-in.

1. Sample $(f|y, v, \theta)$ using a precision sampler.
2. Sample $(v|y, f, \theta)$ using a precision sampler.

3. Sample $(\theta|y, f, \upsilon)$.

Matrix notation

Before we explain the precision sampler, we set up the model in matrix form. We begin by writing the joint sampling density of $f(y|\theta, f, \upsilon)$ by stacking equation 22 over the T time periods:

$$Y = \Lambda F + H_{\upsilon}\upsilon + \varepsilon, \tag{27}$$

where

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_T \end{bmatrix}, F = \begin{bmatrix} F_1 \\ \vdots \\ F_T \end{bmatrix}, \upsilon = \begin{bmatrix} \upsilon_1 \\ \vdots \\ \upsilon_T \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{bmatrix}$$

with $\Sigma_{\varepsilon} \sim \mathsf{N}(0, I_T \otimes \varepsilon)$. Y is a $TN \times 1$ matrix; Λ is a $TN \times TQ$ matrix; F is a $TQ \times 1$ matrix; H_{υ} is a $TN \times T$ banded matrix, which includes the mixed frequency approximation of Mariano and Murasawa; υ is a $T \times 1$ matrix, and ε is a $TN \times 1$ matrix. The coefficient matrix Λ is a banded matrix stacking the coefficients and the lagged coefficients, taking into account the Mariano and Murasawa mixed-frequency approximation.

Onto the state equation (equation 23),

$$f_t = af_{t-1} + \ldots + af_{t-p} + \eta_t. \tag{28}$$

We rewrite this using banded matrices. For simplicity, we assume p = 1.

$$H_f F = \eta_t, \tag{29}$$

where

$$H_f = \begin{bmatrix} I_q & & & & \\ -a & I_q & & & \\ & -a & I_q & & \\ & & \ddots & \ddots & \\ & & & -a & I_q \end{bmatrix}, \Sigma_\eta = \begin{bmatrix} \Omega_\eta & & & & \\ & \Omega_\eta & & & \\ & & \Omega_\eta & & \\ & & & \ddots & \\ & & & & \Omega_\eta \end{bmatrix}$$

and $F \sim \mathsf{N}(0, K^{-1})$ where $K = H_f' \Sigma_\eta H_f$.

Sampling the states

We use a precision sampler to sample the latent state variables (Chan and Jeliazkov 2009).

First, the vector of factors, $(f|y, \upsilon, \theta)$, is sampled from a normal distribution with mean D_f and variance P_f:

$$P_f = K + \Lambda' \Sigma_\varepsilon^{-1} \Lambda \qquad D_f = P_f^{-1}[\Lambda' \Sigma_\varepsilon^{-1}(Y - H_\upsilon \upsilon)]. \tag{30}$$

Similarly, we sample the common error υ:

$$P_\upsilon = \Sigma_\upsilon + H_\upsilon' \Sigma_\varepsilon^{-1} H_\upsilon \qquad D_\upsilon = P_\upsilon^{-1}[H_\upsilon' \Sigma_\varepsilon^{-1}(Y - \Lambda F)]. \tag{31}$$

Sampling state parameters

The state equation is a BVAR with an independent, normal-Wishart prior, and the parameters are estimated via Gibbs sampling. The coefficients (a) are estimated with the standard linear regression results. F_t is a vector of f_t stacked over T (subscripts are included to signify lags).

$$(a \mid F_t, \Omega_\eta) \sim N(\hat{a}, K_a^{-1}), \tag{32}$$

where

$$K_a = V_a^{-1} + F_{t-1}'(I_T \otimes \Omega_\eta^{-1}) F_{t-1}, \qquad \hat{a} = K_a^{-1}(V_a^{-1} a_0 + F_{t-1}'(I_T \otimes \Omega_\eta^{-1}) F_t. \tag{33}$$

V_a is the prior on the covariance of a, and a_0 is the prior mean.The covariance matrix is sampled from:

$$(\Sigma_\eta \mid F, \alpha) \sim IW(v_0 + T, S_0 + \sum_{t=1}^{T}(f_t - f_{t-1}\alpha)(f_t - f_{t-1}\alpha)') \tag{34}$$

with v_0 as the prior degrees of freedom and S_0 as the prior scale matrix.

Sampling observation equation parameters

In this section, we discuss the estimation of the observation equation parameters. Since the equations for y_0 are independent, λ_0 and $\sigma_{\varepsilon}0$ can be sampled with Bayesian linear regression.

However, some adjustments have to be made for λ_1 and $\sigma_{\varepsilon}1,2$. Firstly, we estimate

$\lambda_1|y, f, \{\theta|\lambda_1 \notin \theta\}$.

In matrix form, we can write the equations as $\tilde{Y} = Fr\lambda + \tilde{v}r + \tilde{\varepsilon}$, where

$$\tilde{Y} = \begin{bmatrix} Y_{1t} \\ Y_{2t} \end{bmatrix}$$

$$F = \begin{bmatrix} F_t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & F_t & F_{t-1} & F_{t-2} & F_{t-3} & F_{t-4} \end{bmatrix}$$

$$\tilde{v} = \begin{bmatrix} v_t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & v_t & v_{t-1} & v_{t-2} & v_{t-3} & v_{t-4} \end{bmatrix}$$

$$\tilde{\varepsilon} = \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1T} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2T} \end{bmatrix}$$

where Y_{1t} and Y_{2t} are vectors of y_1t and y_2t (the two measures of GDP) stacked over T, and F_t is a vector

of f_t stacked over T (subscripts are included to signify lags). The same is true for υ and ε.

The mixed-frequency restrictions from Mariano and Murasawa (2003) in the case where q=1 are reflected in the following matrix:

$$r' = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/3 & 2/3 & 1 & 2/3 & 1/3 \end{bmatrix}$$

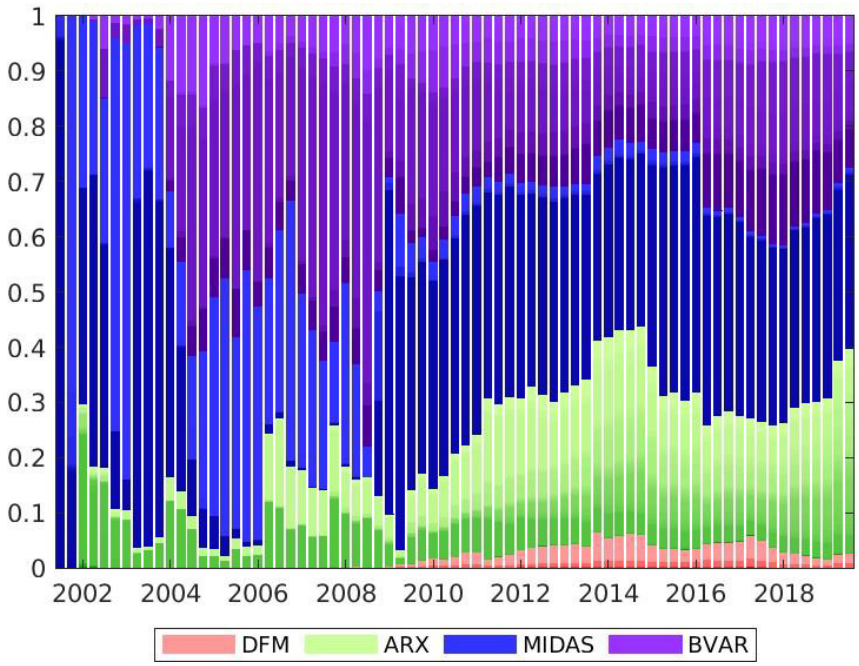
Since $cov(\boldsymbol{f}_t, \upsilon_t) = 0$ and we define $Fr = \tilde{F}$ then we can write $(y \mid \theta, \boldsymbol{f}) \sim N(\tilde{F}\lambda, S)$, where

$S = \Sigma_\varepsilon + H_\upsilon \Sigma_\upsilon H_\upsilon^I$. This means we now have an equation that can be estimated with Bayesian linear regression techniques. The variances σ_1 and σ_2 are sampled from an inverse Wishart restricted to be diagonal.

A.4 Weights appendix

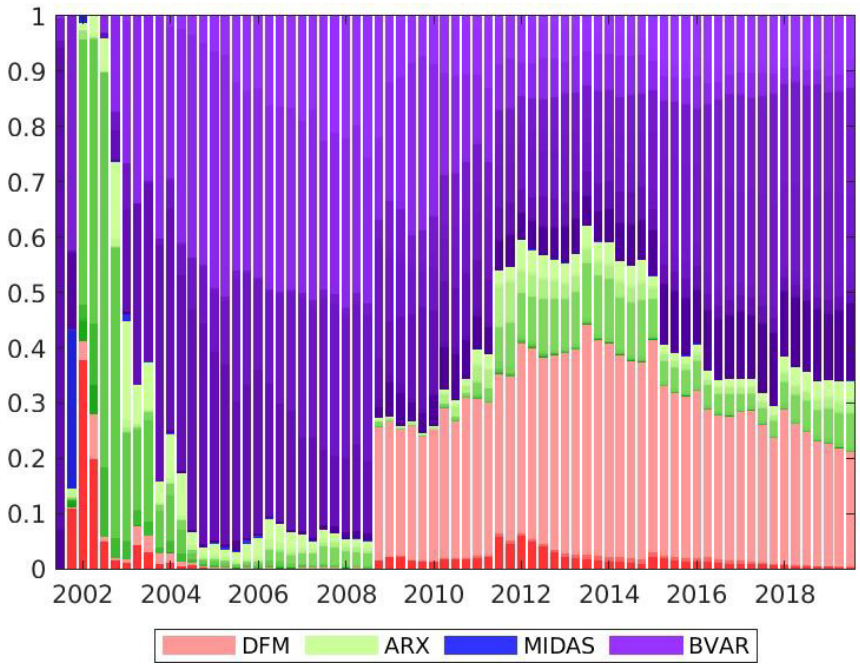
To aid the visualization of 98 weights in the following figures, models within each class are shaded using a gradient of the same colour. (That is, there are 22 BVAR models represented by 22 shades of blue, 4 DFMs represented by 4 shades of red, and so forth.)

Chart A-10: One-stage minorization-maximization combination weights, 24 weeks from release of GDP



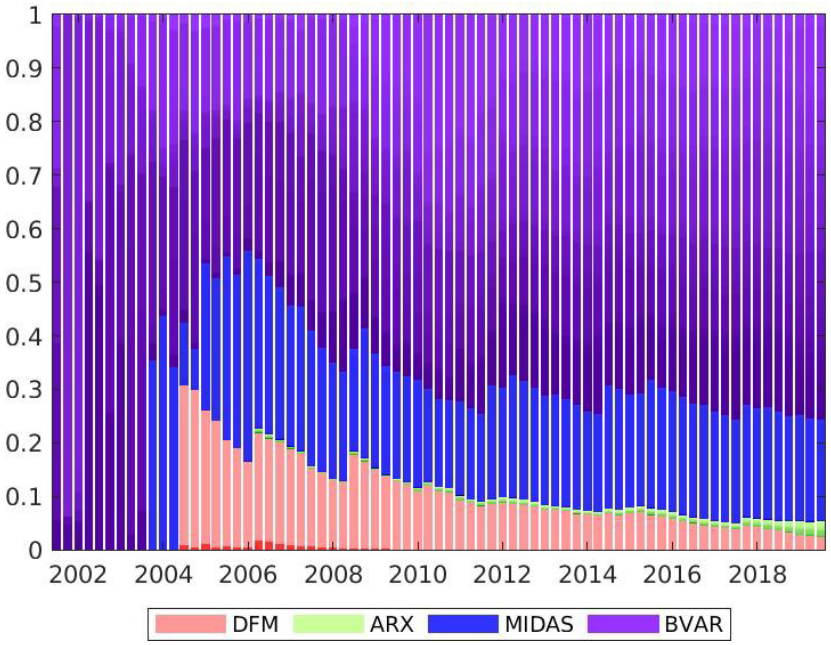
Note: Weights, shown on the y-axis, sum to 1. The abbreviations in the legend are DFM, dynamic factor models; ARX, autoregressive leading indicator models; MIDAS, mixed-data sampling; BVAR, Bayesian vector autoregression.

Chart A-11: One-stage minorization-maximization combination weights, 12 weeks from release of GDP



Note: Weights, shown on the y-axis, sum to 1. The abbreviations in the legend are DFM, dynamic factor models; ARX, autoregressive leading indicator models; MIDAS, mixed-data sampling; BVAR, Bayesian vector autoregression.

Chart A-12: One-stage minorization-maximization combination weights, 2 weeks from release of GDP

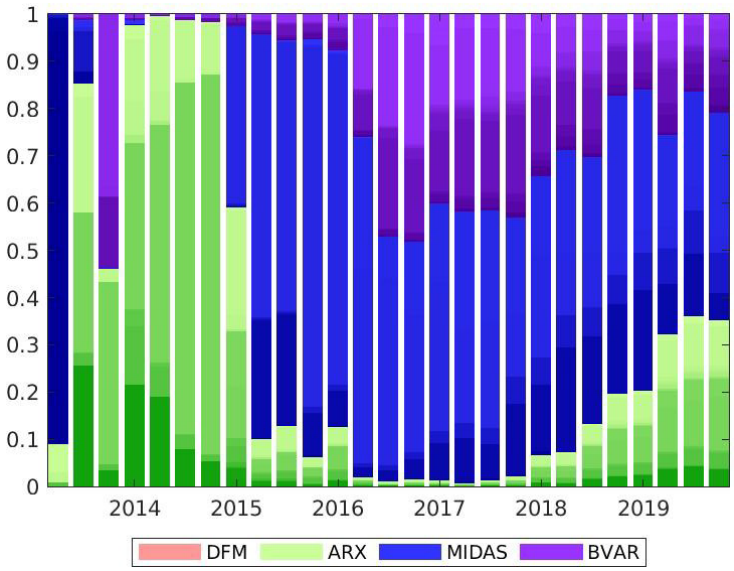


Note: Weights, shown on the y-axis, sum to 1. The abbreviations in the legend are DFM, dynamic factor models; ARX, autoregressive leading indicator models; MIDAS, mixed-data sampling; BVAR, Bayesian vector autoregression.

A.4.1 Real-time weights appendix

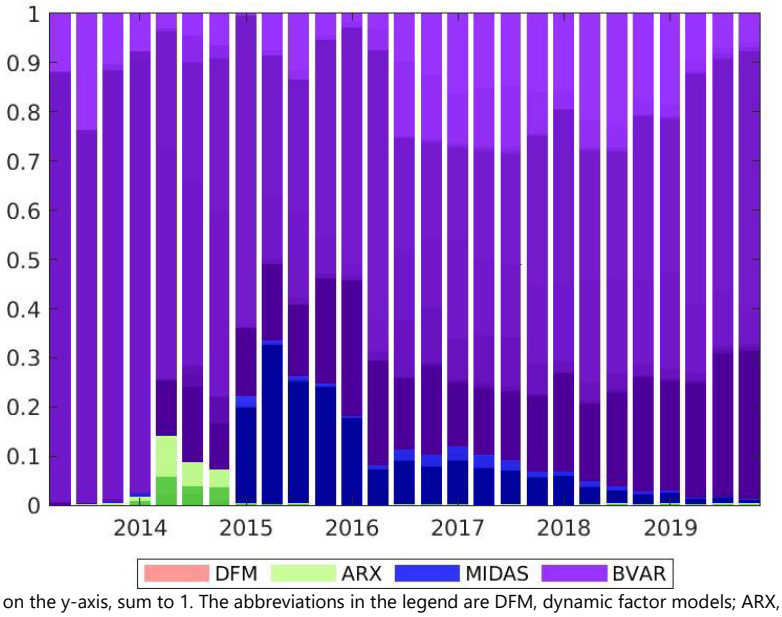
Charts A-13 and A-14 show the evolution of the one-stage MM weights over the course of the limited sample at the first forecast horizon (i.e., 24 weeks ahead of the release of GDP data). Charts A-15 and A-16 show the same weights at the final forecast horizon (i.e., 2 weeks ahead of the GDP data).

Chart A-13: Real-time one-stage minorization-maximization combination weights, 24 weeks from release of GDP



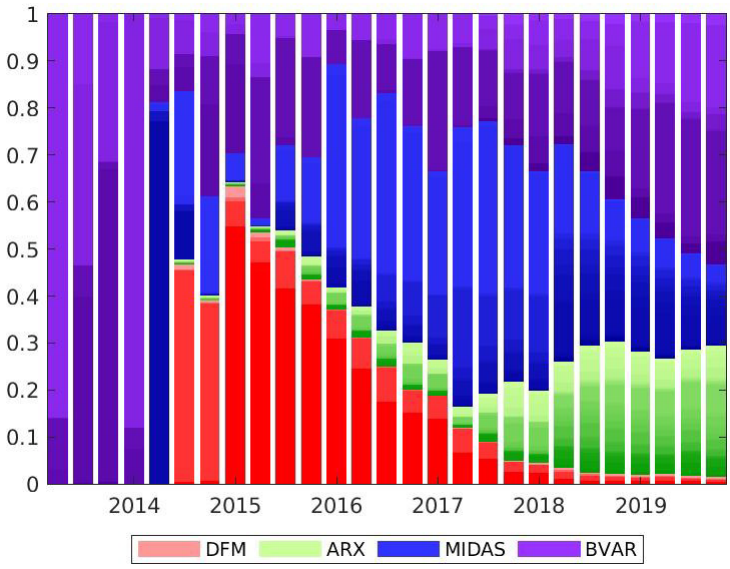
Note: Weights, shown on the y-axis, sum to 1. The abbreviations in the legend are DFM, dynamic factor models; ARX, autoregressive leading indicator models; MIDAS, mixed-data sampling; BVAR, Bayesian vector autoregression.

Chart A-14: Pseudo-real-time one-stage minorization-maximization combination weights, 24 weeks from release of GDP



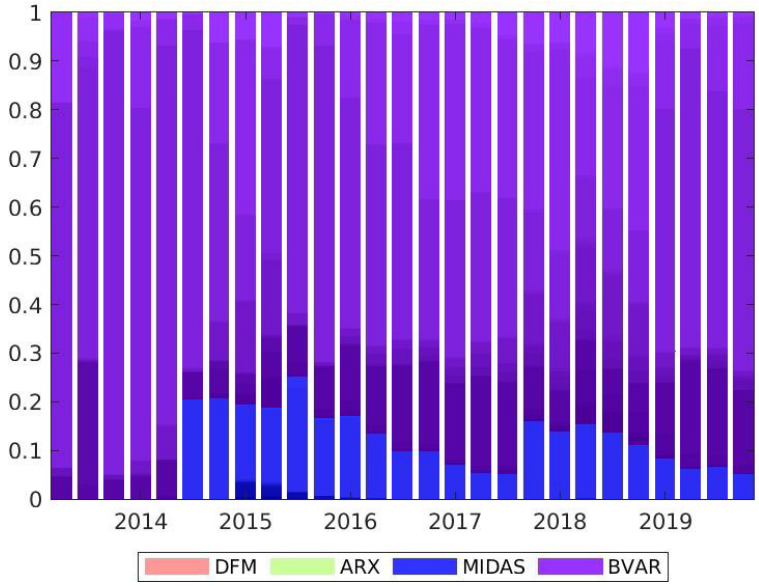
Note: Weights, shown on the y-axis, sum to 1. The abbreviations in the legend are DFM, dynamic factor models; ARX, autoregressive leading indicator models; MIDAS, mixed-data sampling; BVAR, Bayesian vector autoregression.

Chart A-15: Real-time one-stage minorization-maximization combination weights, 2 weeks from release of GDP



Note: Weights, shown on the y-axis, sum to 1. The abbreviations in the legend are DFM, dynamic factor models; ARX, autoregressive leading indicator models; MIDAS, mixed-data sampling; BVAR, Bayesian vector autoregression.

Chart A-16: Pseudo-real-time one-stage minorization-maximization combination weights, 2 weeks from release of GDP



Note: Weights, shown on the y-axis, sum to 1. The abbreviations in the legend are DFM, dynamic factor models; ARX, autoregressive leading indicator models; MIDAS, mixed-data sampling; BVAR, Bayesian vector autoregression.

References

Aastveit, K. A., C. Foroni and F. Ravazzolo. 2017. "Density Forecasts with MIDAS Models." *Journal of Applied Econometrics* 32 (4): 783–801.

Aastveit, K. A., K. R. Gerdrup and A. S. Jore. 2011. "Short-Term Forecasting of GDP and Inflation in Real-time: Norges Bank's System for Averaging Models." Norges Bank Staff Memo No. 09.

Aastveit, K. A., K. R. Gerdrup, A. S. Jore and L. A. Thorsrud. 2014. "Nowcasting GDP in Real Time: A Density Combination Approach." *Journal of Business & Economic Statistics* 32 (1): 48–68.

Aastveit, K. A., J. Mitchell, F. Ravazzolo and H. van Dijk. 2018. "The Evolution of Forecast Density Combinations in Economics." Tinbergen Institute Discussion Paper No. 2018-069/III.

Aastveit, K. A., F. Ravazzolo and H. K. van Dijk. 2018. "Combined Density Nowcasting in an Uncertain Economic Environment." *Journal of Business & Economic Statistics* 36 (1): 131–145.

Adrian, T., N. Boyarchenko and D. Giannone. 2019. "Vulnerable Growth." *American Economic Review* 109 (4): 1263–1289.

Angelini, E., G. Camba-Mendez, D. Giannone, L. Reichlin and G. Rünstler. 2011. "Short-Term Forecasts of Euro Area GDP Growth." *Econometrics Journal* 14 (1): C25–C44.

Bache, I. W., J. Mitchell, F. Ravazzolo and S. P. Vahey. 2009. "Macro Modelling with Many Models." Norges Bank Working Paper No. 15.

Bates, J. M. and C. W. J. Granger. 1969. "The Combination of Forecasts." *Journal of the Operational Research Society* 20 (4): 451–468.

Bell, V., L. W. Co, S. Stone and G. Wallis. 2014. "Nowcasting UK GDP Growth." Bank of England *Quarterly Bulletin* 2014 Q1.

Billio, M., R. Casarin, F. Ravazzolo and H. K. van Dijk. 2013. "Time-Varying Combinations of Predictive Densities Using Nonlinear Filtering." *Journal of Econometrics* 177 (2): 213–232.

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

Bjørnland, H. C., K. Gerdrup, A. S. Jore, C. Smith and L. A. Thorsrud. 2012."Does Forecast Combination Improve Norges Bank Inflation Forecasts?" *Oxford Bulletin of Economics and Statistics* 74 (2): 163–179.

Champagne, J., G. Poulin-Bellisle and R. Sekkel. 2018. "Evaluating the Bank of Canada Staff Economic Projections Using a New Database of Real-Time Data and Forecasts." Bank of Canada Staff Working Paper No. 2018-52.

Chan, J. C. and I. Jeliazkov. 2009. "Efficient Simulation and Integrated Likelihood Estimation in State Space Models." *International Journal of Mathematical Modelling and Numerical Optimisation* 1 (1/2): 101–120.

Chernis, T. and R. Sekkel. 2017. "A Dynamic Factor Model for Nowcasting Canadian GDP Growth." *Empirical Economics* 53 (1): 217–234.

Chernis, T. and R. Sekkel. 2018. "Nowcasting Canadian Economic Activity in an Uncertain Environment." Bank of Canada Staff Discussion Paper No. 2018-9.

Clark, T. E. and M. W. McCracken. 2010. "Averaging Forecasts from VARs with Uncertain Instabilities." *Journal of Applied Econometrics* 25 (1): 5–29.

Clemen, R. T. 1989. "Combining Forecasts: A Review and Annotated Bibliography." *International Journal of Forecasting* 5 (4): 559–583.

Coletti, D. and S. Murchison. 2002. "Models in Policy-Making." *Bank of Canada Review* (Summer): 19–26.

Conflitti, C., C. De Mol and D. Giannone. 2015. "Optimal Combination of Survey Forecasts." *International Journal of Forecasting* 31 (4): 1096–1103.

Corradi, V. and N. Swanson. 2006. "Predictive Density Evaluation." In *Handbook of Economic Forecasting*, vol. 1, edited by G. Elliott, C. Granger and A. Timmermann, 197–284. Amsterdam: Elsevier.

Croushore, D. and T. Stark. 2003. "A Real-Time Data Set for Macroeconomists: Does the Data Vintage Matter?" *Review of Economics and Statistics* 85 (3): 605–617.

Del Negro, M., R. B. Hasegawa and F. Schorfheide. 2016. "Dynamic Prediction Pools: An Investigation of Financial Frictions and Forecasting Performance." *Journal of Econometrics* 192 (2): 391–405.

Deutsche Bundesbank. 2013. "Forecasting Models in Short-Term Business Cycle Analysis – A Workshop Report." Monthly report (September): 69–83.

Diebold, F. X., T. A. Gunther and A. S. Tay. 1998. "Evaluating Density Forecasts with Applications to Financial Risk Management." *International Economic Review* 39 (4): 863–883.

Geweke, J. and G. Amisano. 2011. "Optimal Prediction Pools." *Journal of Econometrics* 164 (1): 130–141.

Ghysels, E., P. Santa-Clara and R. Valkanov. 2004. "The MIDAS Touch: Mixed Data Sampling Regression Models." CIRANO Working Paper No. 2004s-20.

Giannone, D., L. Reichlin and D. Small. 2008. "Nowcasting: The Real-Time Informational Content of Macroeconomic Data." *Journal of Monetary Economics* 55 (4): 665–676.

Gneiting, T. and A. E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102 (477): 359–378.

Granger, C. W. J. and R. Ramanathan. 1984. "Improved Methods of Combining Forecasts." *Journal of Forecasting* 3 (2): 197–204.

Granziera, E., C. Luu and P. St-Amant. 2013. "The Accuracy of Short-Term Forecast Combinations." *Bank of Canada Review* (Summer): 13–21.

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

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.

Kapetanios, G., V. Labhard and S. G. Price. 2007. "Forecast Combination and the Bank of England's Suite of Statistical Forecasting Models." Bank of England Working Paper No. 323.

Karlsson, S. (2013): "Forecasting with Bayesian Vector Autoregression." In H*andbook of Economic Forecasting*, vol. 2, edited by G. Elliott, C. Granger and A. Timmermann, 791–897. Amsterdam: Elsevier.

Knüppel, M. 2015. "Evaluating the Calibration of Multi-Step-Ahead Density Forecasts Using Raw Moments." *Journal of Business & Economic Statistics* 33 (2): 270–281.

Koop, G. and D. Korobilis. 2010. "Bayesian Multivariate Time Series Methods for Empirical Macroeconomics." *Foundations and Trends(R) in Econometrics* 3 (4): 267–358.

Mariano, R. S. and Y. Murasawa. 2003. "A New Coincident Index of Business Cycles Based on Monthly and Quarterly Series." *Journal of Applied Econometrics* 18 (4): 427–443.

McAlinn, K. and M. West. 2019. "Dynamic Bayesian Predictive Synthesis in Time Series Forecasting." *Journal of Econometrics* 210 (1): 155–169.

Mitchell, J. and S. 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.

Orphanides, A. 2001. "Monetary Policy Rules Based on Real-Time Data." *American Economic Review* 91 (4): 964–985.

Ranjan, R. and T. Gneiting. 2010. "Combining Probability Forecasts." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 72 (1): 71–91.

Rizzetto, P. 2018. "GDP by Industry in Real Time: Are Revisions Well Behaved?" Bank of Canada Staff Analytical Note No. 2018-40.

Rossi, B. and T. Sekhposyan. 2019. "Alternative Tests for Correct Specification of Conditional Predictive Densities." *Journal of Econometrics* 208 (2): 638–657.

Stark, T. and D. Croushore. 2002. "Forecasting with a Real-Time Data Set for Macroeconomists." *Journal of Macroeconomics* 24 (4): 507–531.

Timmermann, A. 2006. "Forecast Combinations." In *Handbook of Economic Forecasting*, vol. 1, edited by G. Elliott, C. Granger and A. Timmermann, 135–196. Amsterdam: Elsevier.

Wallis, K. F. 2005. "Combining Density and Interval Forecasts: A Modest Proposal." *Oxford Bulletin of Economics and Statistics* 67 (s1): 983–994.