Forecasting Inflation in Argentina: A Probabilistic Approach.

Tomas Marinozzi

Department of Economics UCEMA

August 2022

Abstract

Probability forecasts are gaining popularity in the macroeconomic discipline as point forecasts lack the ability to capture the level of uncertainty in fundamental variables like inflation, growth, exchange rate, or unemployment. This paper explores the use of probability forecasts to predict inflation in Argentina. Scoring rules are used to evaluate several autoregressive models relative to a benchmark. Results show that parsimonious univariate models have a relatively similar performance to that of the multivariate models around central scenarios but fail to capture tail risks, particularly at longer horizons.

Key Words: Probability Forecast, Forecasting, Inflation Forecast, Continuous Ranked Probability Scores.

1 Introduction

Forecasting has played an increasing role in the economic discipline, as easier access to data and faster computers has lowered the cost of making predictions (Agrawal, Gans, & Goldfarb, 2018). For macroeconomics forecasting in particular, the implementation of forecasting techniques plays a huge role in economic policy decisions, as well as in the process of anchoring expectations. Central banks in particular use forecasts as a way to predict future behaviour of the economic system and evaluate policy implementation. Effective forecasting allows policymakers to make appropriate policy decisions, build confidence, align expectations, and induce a forward looking perspective of the markets.

To this day, most "public" forecasts are presented in the form of baseline scenarios (point forecasts). Probability forecasting on the other hand attempts to quantify the uncertainty surrounding the projection of the target variable. The latter kind of forecasting has been used in other disciplines for quite some time, and along with it are techniques that allow the evaluation of probabilistic forecasts. Brier (1950), Winkler and Murphy (1968), Savage (1971) are some of the more renowned pioneers on the literature of the construction and evaluation of probability forecasting. Although probabilistic forecasts were most commonly seen in weather forecasting, over the last three decades the popularity and use of probabilistic forecasting has increased in disciplines like computational finance (Duffie & Pan, 1997), and macroeconomic forecasting (Garratt, Lee, Pesaran, & Shin, 2003). In finance in particular, the boom of financial risks management has accelerated the standard practice of probability forecasts in the field. In the field of macroeconomics, generally speaking, density forecasts are not standard practice but are gradually becoming more popular.

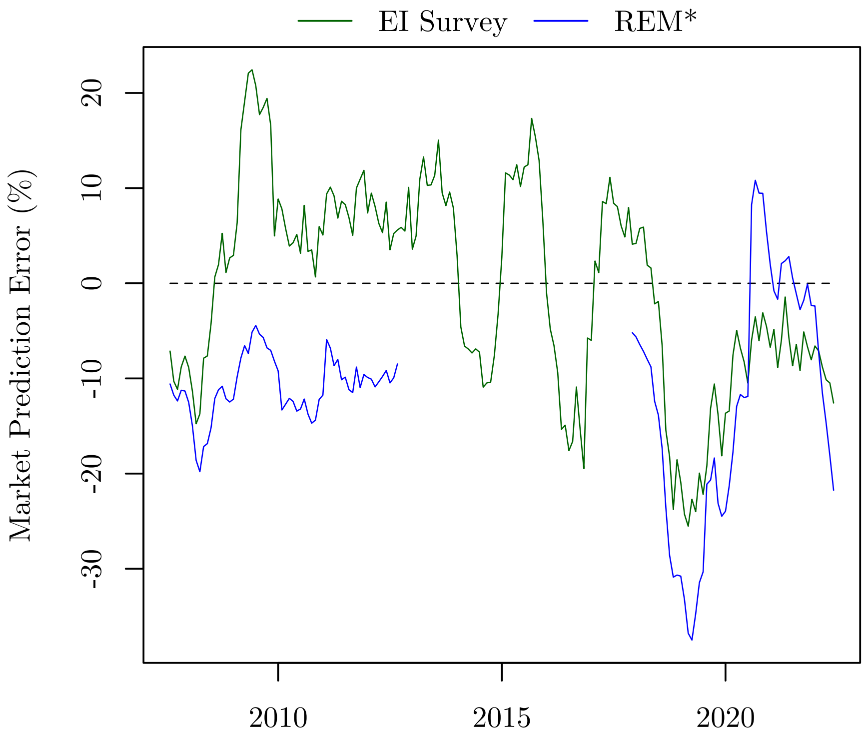
The objective of this paper is to explore some probability forecast models that could help quantify the uncertainty surrounding inflation in Argentina. A model of such nature could be useful for policy makers in planning economic decisions from a probabilistic perspective as well as setting contingent policy decisions. It could also improve contracts that use inflation forecasts as an input in their decisions making, like wage negotiations, bank rates and any type of investment decisions.

The paper is structured in six main sections. Section 2 discusses the need for probability forecasts in a context like Argentina. Section 3 provides a description of the structure of the models that are be used in the paper, including a set of different autoregressive models, variance treatments for simulations and a technique to combine models. Section 4 discusses the specific variables and treatments explicitly used for the ten selected models. Section 5 provides an analytical background on the evaluation strategy, describing some alternatives to evaluate probabilistic forecasts using scoring rules, as well as point forecast evaluation. Section 6 illustrates the results from the forecasting exercises, while section 7 provides concluding remarks.

2 The need for probabilistic forecasts

The fact that the "true" model is unknown implies that any economic forecaster faces uncertainties and has to accept some degree of inaccuracy. Modern forecasting techniques along with faster computational power allows forecasters to run hundreds of scenarios to better comprehend the probabilistic nature of the target variable. Surprisingly, despite the computational power to run these models, Central banks, the IMF, the World Bank, and other world-renowned institutions tend to public baseline scenarios (point forecast). In some cases, some institutions may even present a simple subset of alternative scenarios (upper/optimistic, lower/pessimistic), which lack applicability as the likelihood of occurrence is usually undisclosed. Sometimes quantile-based alternative scenarios are presented. But in many cases, they come from ad-hock distributions that were not necessarily tested for predictive accuracy (ex. assuming a normal-invariant shock distribution to be the appropriate distribution when in fact this might not be the case). Although central scenarios (mean or median for instance) are useful in providing guidance for the future, they provide very little understanding of the probabilistic likelihood of such events or the risks associated with extreme events (tail risks). Probabilistic forecasts are therefore encouraged as a way of embracing uncertainty and providing insightful information for asset pricing and scenario analysis.

Let us take inflation in Argentina as an example, which is characterized by high levels of inflation and has been particularly volatile over the past fifteen years. The chart below compares the year-on-year inflation prediction error derived from consumer inflation expectations (expected inflation survey from Di Tella University) and a survey of professional forecasters called the Market Expectations Survey (REM in Spanish) published by the Argentine Central Bank.

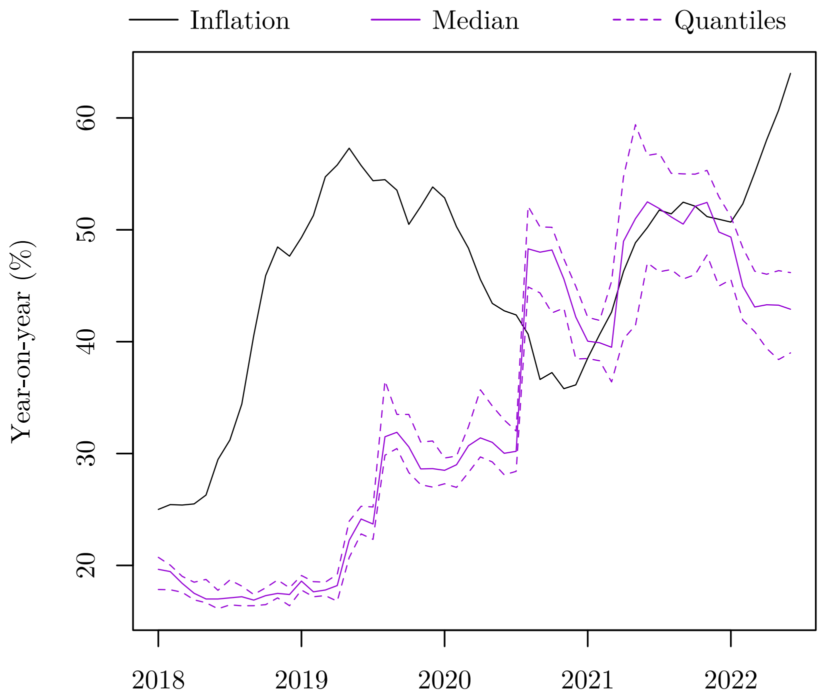


(∗) REM was not published between Sep. 2012 - May. 2016

Figure 1: 12M Inflation Forecast Errors. Source: UTDT & BCRA

Aside from any positive or negative trend-bias from both indicators, the margin of error is pretty substantial: notice the Mean Absolute Error (MAE) for the sample is around 10 percentage points. In a case like this, it is evident that a point forecast does not provide enough context for inflationary risks and the usefulness for the market is limited. Notice that this probably has more to do with the volatility that inflation exhibits, rather than "flaws" in the market's ability to appropriately forecast inflation. Therefore, the point can be made that if inflation exhibits such levels of uncertainty, then it should not be ignored but rather embraced and a way to do that is by using probabilistic models.

In order to provide additional information about the inflation outlook, Central Banks often provide quantiles derived from market surveys or professional forecaster surveys. For instance, figure (2) shows actual inflation (12M lagged) against the median and quantiles derived from the survey conducted by the Central Bank. Although they are sometimes understood as a measure of inflation risks, that is not what those quantiles intended to showcase.



Note: Inflation was lagged 12 months. Upper and lower bounds represent quantile 0.25 and 0.75.

Figure 2: 12M REM inflation forecast. Source: BCRA

In this particular case, notice how in some periods, quantiles are extremely narrow with respect to the mean (particularly for the first 24 months of the chart). This is because a distribution of baseline projection does not necessarily reflect inflation uncertainty but rather the degree of dispersion in expectations. Although it is logical to assume that a more uncertain environment is generally associated with a higher degree of dispersion in expectations, this does not properly capture tail-risk events. To adequately capture risks, probabilistic exercises must be conducted.

3 Models structure

In line with the previous section, a probabilistic forecasting exercise was conducted as a way of finding modes that adequately capture inflation risks in Argentina. The following section describes the general structure and characteristics of the class of models tested. Naturally, every model has its own advantages and caveats. For instance, one could argue univariate models are unable to fully grasp the interactions driving the macroeconomic system. However, it is possible that complex multivariate models that include logical interactions among variables yield lower predictive performance than parsimonious models, as they might suffer over-specification causing overfitting. This paper in particular will focus on a selection of autoregressive models (both univariate and multivariate).

3.1 Conditional mean models

3.1.1 Univariate models (benchmark)

A series of conventional univariate models were tested to forecast inflation. Firstly, a random walk;

$$\pi_t = \pi_{t-1} + u_t \tag{1}$$

The random walk is perhaps the most commonly used benchmark in macro and finance literature, mostly due to its simplicity as well as its reasonable predictive ability. However, we also tested other univariate autoregressive specifications to have an alternative baseline.[1]

$$\pi_t = \rho_1 \pi_{t-1} + \cdots + \rho_p \pi_{t-p} + u_t \tag{2}$$

where $u_t \sim N(0, \sigma)$[2].

3.1.2 Phillips curve

A version of the "Hybrid New Keynesian Phillips Curve" (NKPC), originally proposed by Galí and Gertler, was included. The specification of the open economy version is an autoregressive version of the one presented in D'Amato, Aguirre, Garegnani, Krysa, and Libonatti (2018).

$$\pi_t = \phi_1 \pi_{t-1} + \phi_2 \mathrm{E}_{t-1}[\pi_t] + \delta x_{t-1} + \gamma \pi^*_{t-1} + \lambda \Delta e_{t-1} + u_t \tag{3}$$

where $\mathrm{E}_t[\pi_{t+1}]$ represents inflation expectations and x_t is the output gap. In the following specifications, exchange rate devaluation, Δe_t, and foreign inflation, π^*_t, have a direct effect on domestic inflation (Svensson, 2000).

3.1.3 Vector autoregressive (VAR)

Moving on to multivariate models, the vector autoregressive (VAR) model plays a significant role in the literature of economic and financial forecasting. They were first introduced by Manz and Sims Jr (1980) as a method for analysing macroeconomic data and they became popular because of their simplicity and their use as a flexible alternative to large scale econometric models.

[1] Because of the non-stationary nature of CPI, we treat the model in log-differences for univariate models and include a constant term when required.

[2] Further specifications for the variance will be discuss later in the section.

A set of endogenous variables, $\boldsymbol{Y}_t$, is represented as a linear function of its own p-lags. This assumes that the endogenous variables are treated symmetrically and that there is a feedback effect between them. The possibility of exogenous regressors, $\boldsymbol{X}_t$, that could affect the behaviour of the economy was not dismissed. This may be important as we are dealing with a small open economy with inflationary dynamics that are dependent on international commodity prices, fund flows, and global activity.

The VAR model can be expressed in its reduced form as

$$\boldsymbol{Y}_t = \boldsymbol{\nu} + \boldsymbol{A}_1\boldsymbol{Y}_{t-1} + \ldots + \boldsymbol{A}_p Y_{t-p} + \boldsymbol{B}_1\boldsymbol{X}_{t-1} + \ldots + \boldsymbol{B}_q\boldsymbol{X}_{t-q} + \boldsymbol{u}_t \tag{4}$$

$\boldsymbol{Y}_t$ is an $N \times 1$ dimensional vector of endogenous random variables; $\boldsymbol{X}_t$ is an $L \times 1$ dimensional vector of exogenous variables; $\boldsymbol{\nu}$ is a fixed $N \times 1$ vector of constants; $\mathbf{A}_i$ are the $N \times N$ coefficient matrices for the endogenous variables; $\mathbf{B}_j$ are the $N \times L$ coefficient matrices for the endogenous variables; $\mathbf{u}_t \sim (\mathbf{0}, \boldsymbol{\Sigma_t})$ is a $N \times 1$ vector of serially uncorrelated exogenous shocks ($\mathrm{E}\left[\mathbf{u}_t\mathbf{u}_s'\right] = \mathbf{0}\ \forall\ s \neq t$) with constant covariance matrix of size $N \times N$ and zero mean ($\mathrm{E}[\boldsymbol{u}_t] = 0\ \forall\ t$). The previous assumptions imply a conditional mean $\boldsymbol{\mu}_t$ and a constant covariance $\boldsymbol{\Sigma_t} = \boldsymbol{\Sigma_u}$

It is relevant to mention that it is assumed that exogenous variables follow a separate data generating process. This is an important point because for $h > 1$, there has to be a pre-determined parallel forecast for $\boldsymbol{X}_{t+h}$ feeding $\boldsymbol{Y}_{t+h}$. This might be a problem as the predictive ability of the regressor is strictly dependent on a parallel forecast. To put it simply, even if the true value of $\boldsymbol{X}$ has substantial predictive content, if the data generating process (DGP) is too complex to model, poor projections on the exogenous variables could actually worsen the predictive ability of the model[3].

3.1.4 Vector error correction (VEC))

Vector error correction (VEC) models are restricted versions of VAR models designed with the intention of dealing with a non-stationary series that follows a common deterministic trend and are known (or presumed) to be *co-integrated*. In essence, VEC models have co-integration relations that are specified so that it restricts the long-run behaviour of the endogenous variables to converge to their co-integrating relationships while allowing for short-run adjustment dynamics (often referred to as *Error Correction Term*).

For the VEC models, a Johansen test was conducted to specify the number of co-integrating relationships and estimate the relationships. That said, following the Johansen approach is not a necessary

[3]In this paper exogenous variables like U.S. CPI were modelled separately using tested versions of univariate models.

condition as, ultimately, models are judged solely on their predictive ability.

$$\Delta \boldsymbol{Y}_t = \boldsymbol{\nu} + \boldsymbol{\Pi} \boldsymbol{Y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \boldsymbol{Y}_{t-1} + ... + \boldsymbol{\Gamma}_{(p-1)} \Delta \boldsymbol{Y}_{t-(p-1)} + \boldsymbol{u}_t \tag{5}$$

where $\boldsymbol{\Pi} = -(\boldsymbol{I}_N - \boldsymbol{A}_1 - \cdots - \boldsymbol{A}_p)$ can also be written as $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$, where $\boldsymbol{\beta}$ is the "cointegration matrix" and $\boldsymbol{\alpha}$ represents the "loading matrix"; $\boldsymbol{\nu}$ represents the deterministic trend of the dynamic process[4].

3.1.5 Imposing long-run equilibrium

The notion of long-term equilibrium relationships derived from VEC models is a rather appealing concept for macroeconomic forecasting, as it might be a way to exploit theories within the models, especially since there is significantly more consensus over long-term equilibrium relationships rather than short-term dynamics. This approach has been used before by other authors (Garratt, Lee, Hashem Pesaran, & Shin, 2003; Schneider, Chen, & Frohn, 2008).

We tested two models with long-run relationships. The first one includes two exchange rate relations, Purchasing Power Parity (PPP) and Uncovered interest rate Parity (UIP),

$$\begin{aligned} PPI: \quad & P_t = E_t - P_t^* \\ UIP: \quad & \Delta E_t = i_t - i_t^* \end{aligned} \tag{6}$$

Where P_t is the domestic log-price, P_t^* is the foreign log-price, E_t is the exchange rate in logarithms, i_t is the domestic interest rate, and i_t^* is the foreign interest rate. The second model includes Money Neutrality (MN) and Real Wage Equilibrium (RWE).

$$\begin{aligned} MN: \quad & M_t - P_t = k \\ RWE: \quad & W_t - P_t = \delta \end{aligned} \tag{7}$$

where M_t and W_t represent money supply and nominal wages. k and δ represent constants guiding the long-term relationships[5].

[4] Read Lütkepohl (2005) on the different ways to model the deterministic trend, as well as the version with exogenous regressors.

[5] One could argue the validity of k and δ as constants. In this particular exercise, given the out-of-sample window, the models were tested with constant terms but the technology allows for a dynamic process for those variables.

3.2 Conditional volatility and non-parametric innovations

So far, a parametric (in particular a white noise) process has been assumed for the residuals, that is, zero mean with constant variance. The paper also explores non-parametric distributions based on re-sampling techniques as well as conditional volatility models.

The key behind these type of models is that σ_t^2 is conditional on past information $\mathcal{F}_{t-1}$. Assume z_t as a white noise series with zero mean and constant unit variance, and the conditional variance σ_t^2 is modeled by

$$u_t = \sigma_t z_t$$

where z_t is a sequence of independent and identically distributed random variables with mean zero and unit variance.

3.2.1 GARCH innovations

Autoregressive conditional heteroskedasticity (ARCH) models describe the current variance as a function of the square of the previous periods' error terms. They were first developed by Engle (1982) and then evolved to a generalized version (GARCH) first introduced by Bollerslev (1986). The generalized version includes ARCH process with additional lag versions of the variance. Following the conventional GARCH specification, the conditional heteroskedasticity is assumed to be;

$$\sigma_t^2 = \gamma + \sum_{i=1}^{m} \alpha_i u_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2 \tag{8}$$

It is assumed that m and s are non negative integers, where $\gamma > 0, \alpha_i \geq 0, \beta_j \geq 0$ for all $i > 0$ and $j > 0$ and $\sum_{i=1}^{m} \alpha_i + \sum_{j=1}^{s} \beta_j \leq 1$. The ARCH component of the model is written as $\sum_{j=1}^{s} \beta_j \sigma_{t-j}^2$.

An issue when modelling conditional variance, in addition to the conditional mean, is that the number of parameters escalates and could ultimately cause *overfitting*, particularly short samples like in our case. To avoid, significant number of parameters, GARCH(1,1) specification is followed. The logic is partially backed by empirical evidence. (Hansen & Lunde, 2005) compare 330 different volatility models, in their case using daily exchange rate data, and they conclude that there was not significant improvement by using a forecast model different than GARCH(1,1). Although the original study focused on equity volatility this particular specification is common practice for forecasters.

It is possible to derive a multivariate extension of the GARCH model (MGARCH), to allow the covariance matrix of the dependent variables to follow a flexible dynamic structure conditional on past information. However, for this specific analysis, these types of models are not considered, as the number of parameters grows exponentially with the number of variables. Because of the nature of the frequency of the data and the short sample used, a variation of this model was attempted but quickly discarded as it led to overfitting and extremely poor performance, particularly in the early periods of the out-of-sample testing. These type of multivariate GARCH models will not be included in the analysis. To preserve covariance structure in multivariate models in addition to a GARCH process, bootstrapping is used to ensure that the simulated GARCH errors preserve a joint distribution.

3.2.2 Bootstrap innovations

Bootstrapping is another technique used in this paper to introduce shock innovation derived from non-parametric distributions. Bootstrapping is essentially achieved by repeatedly sampling (with replacement) the model residuals to create simulated shock. The methodology used here is similar to the one originally proposed by (Efron, 1992) Efron.

Let $X = \{X_1, \cdots, X_n\}$ be the residuals from a stationary process estimated by the model. Because the paper deals with multivariate structures it should be clarified that X_j is a tuple of multivariate residuals such that $X_j = \left\{u_j^1, \cdots, u_j^s\right\}$ for any j in $1 \leq j \leq n$, where s stands for the number of variables in the model. Finally, the exercise consists in a simple random sample drawn with replacement from X creating an innovations matrix of size $s \times r \times h$ where r stands for the number of simulations and h the number of horizons for the s variables.

3.3 Mixture Models

In practice, is common for a forecast combination, of "best" performing models, to yield an even better performance than an individual model (Clemen, 1989). In this case, a combination of probabilistic distributions is used, or what is the same as a *mixture* distributions. In order to represent this mathematically, it is assumed that the target variable y is generated by a latent variable z. In this context, z is considered to be an unobserved variable named the *mixture component*. Formally, $p(z)$ is a multinomial distribution, while $p(y|z)$ can take a variety of parametric forms. We can compute the probability density function over y by marginalising out z in the following way

$$p(y) = \sum_{i=1}^{K} P[Z = z_i]\, p(y|z = z_i)$$

It is important to distinguish between a mixture of distributions and a weighted average of the distributions. In practice, averaging two equal size distributions corresponds to a component by component weighted average. A mixture on the other hand, draws samples from the predictive distributions i according to z_i which occurs with frequency $p(z)$. An important feature is that a mixture of two Gaussian distributions with different means will not generate another Gaussian distribution. Figure (3) shows an example of two arbitrary (Gaussian) distributions for the same random variable y (black and red densities). Notice that averaging both distributions gives an in-between (Gaussian) distribution, while the mixture generates a non-Gaussian distribution which contemplates aspects of both distributions.

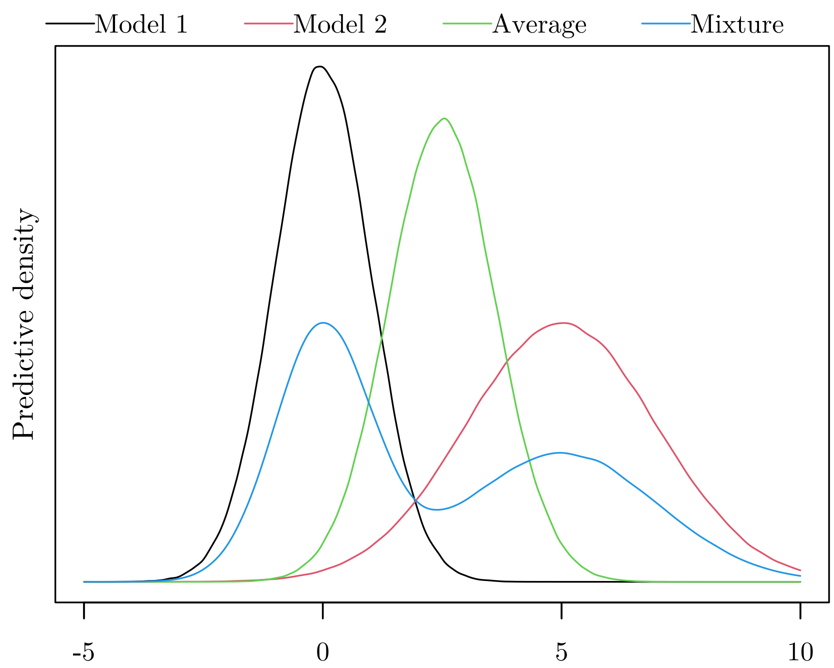


Figure 3: 12M inflation Forecast errors. Source: UTDT & BCRA

Why is it important to make this distinction between average and mixture? Imagine if the two distributions came from two different forecasters (F_1 and F_2). They clearly disagree on the most likely outcome. While F_1 says y will take a value closer to zero, F_2 says it will be closer to 5. If both distributions were averaged, the result would be the in between distribution (green density) with the most likely outcomes at around 2.5. Notice however, even when both forecasters disagree on the most likely outcome, they both agree that the chances of variable y turning out to be around 2.5 are low. That is why from a probabilistic forecasting stance, averaging distributions might not be necessarily the best approach as it may assign a high probability to unlikely scenarios.

4 Selected models and variables included

As explained in section 1, thirty different models were tested. However, for illustration purposes, only a selection of ten models (and the benchmark) will be displayed. This section describes the selected models and the variables included in those models. The data set starts in February 2004 and ends in December 2019. Variables included in selected models:

- CPI: Argentina Consumer Price Index[6].
- Expectations: Twelve months ahead mean inflation expectations. Source: Di Tella University.
- EMAE: A Monthly Economic Activity Indicator. Source: INDEC.
- Interest Rate: 30 to 59 day fixed deposit rates. Source: Central Bank of Argentina (BCRA).
- Wages: Mean wage of registered private sector workers. Source: Ministry of Labour.
- Money: A proxi consisted of money base and central bank short-term liabilities. Source: Central Bank of Argentina (BCRA).
- ARS/USD: Bilateral nominal exchange rate. Source: Central Bank of Argentina (BCRA).
- U.S. CPI: U.S. Consumer Price Index. Source: U.S. Bureau of Economic Analysis (BEA).
- U.S. Interest Rate: 3-Month treasury bill, market rate. Source: Board of Governors of the Federal Reserve System.

Table 1 (see below) clarifies which were the selected models, the variables used and the specific variance treatment to generate the random shocks behind the simulations.

[6]The price index was constructed combining the index from the National Institute of Statistics and Census of Argentina (INDEC) and the Price Index of the City of Buenos Aires and San Luis prices index. The methodology is identical to the one used by the University of CEMA (see https://ucema.edu.ar/cea_vce/serie).

	Variance Treatment	Expectations	EMAE	ARS/USD	Wages	Money Supply	Interest Rate	U.S. CPI
(0) RW	Garch (1,1)							
(1) AR(1)	Parametric							
(2) AR(2)	Garch (1,1)							
(3) AR(4)	Parametric							
(4) VAR(2)	Garch (1,1)		X		X		X	
(5) VAR(2)	Parametric		X		X		X	
(6) VEC(4)	Garch (1,1)			X	X	X		
(7) VEC(4)	Bootstrap			X	X	X		
(8) PC	Bootstrap	X	X	X				X
(9) Long-Run	Bootstrap	X	X		X	X		
(10) Mixture	-	X	X	X	X	X	X	X

Table 1: Models and variables included

For obvious reasons, all models use CPI and therefore the variable was excluded from the table[7]. The long-run model in the table includes money neutrality and constant real wages, the alternative long-run model (with PPP and UIP) was not included due to its lower performance. The number to the left represents the ID number for the models. From now on, referring to the models by the name or by their ID number will be indifferent. A couple of mixtures were tested, but the mixture selected in the summary was a combination of multivariate models 6, 8, 9. The remaining models and variables used are displayed in Table 4 within the appendix.

5 Evaluation strategy

In order to provide a comparison of the models, a recursive out-of-sample evaluation was conducted. The parameters were recursively estimated over the out-of-sample stage using all the observations available until the time of the forecast (Rossi, 2014). After the recursive estimation, different accuracy measures were applied to evaluate the models' predictive ability. For probability forecasts, Continuous Ranked Probability Scores (CRPS) and Quantile Scores (QS) were used. Point forecasts were also derived from the distributions and compared using Root Mean Square Error (RMSE) and Mean Percentage Errors (MPE) to check model bias. Inference on probabilistic ability was computed for selected models using the Diebold-Mariano (DM) test. Finally, a probability integral transforma-

[7]Note that U.S. Interest rate was also excluded from the table because none of the models above use it as part of their input.

tion (PIT) approach was used to evaluate the specification on the top probabilistic model.

5.1 Point forecasts evaluation

When running an out-of-sample evaluation for a point-forecast, it is necessary to introduce some sort of performance measure to compare simulations. Usually, this refers to a loss function that maps the forecast deviation from the actual realisation across the out-of-sample window, at horizon, h. Therefore, for a specific model m and the specified loss function L, the average score Π will be defined as

$$\Pi_h = \frac{1}{T}\sum_{t=1}^{T} L(\hat{y}_{t+h}, y_{t+h}) \tag{9}$$

where $\hat{y}_{t+h}$ is the forecast produced h periods prior, and y_{t+h} is the observed value. The most conventional loss function to evaluating point forecast is the *Root Mean Square Error* (RMSE),

$$RMSE_h = \sqrt{\frac{1}{T}\sum_{t=1}^{T} (\hat{y}_{t+h} - y_{t+h})^2} \tag{10}$$

Mean percentage error (MPE) will also be used to visualise if the models exhibit a bias across the horizons,

$$MPE_h = \frac{1}{T}\sum_{t=1}^{T} \frac{\hat{y}_{t+h} - y_{t+h}}{\hat{y}_{t+h}} \tag{11}$$

because actual rather than absolute values of the forecast errors are used in the formula, positive and negative forecast errors should be offset in the absence of bias.

5.2 Probabilistic forecasts evaluation

Similarly to the concept of loss functions described in point forecast evaluations, *scoring rules* are generally used as a summary measure for the evaluation of probabilistic predictions or forecasts.

Definition 1. ***(Scoring rule)*** *Given a forecaster predicted cumulative distribution function, $F \in \mathcal{F}$, for a random variable Y, a scoring rule S is a map such that $S : \mathcal{F} \times \mathbb{R} \to \mathbb{R}$. Specifically, the scoring rule assigns a numerical score $S(F, y) \in \mathbb{R}$ to F after evaluating its performance relative to the actual observation y.*

Analogous to point forecasts, when using scoring rules a forecaster should try to minimise the expected score. Suppose that the agent believes the true distribution is G, then the expected score should be

$$\min_F \mathrm{E}_G S(F, y) = \min_F \sum_y q(y) S(F, y)$$

where q represents a probability. In this context it is important to recognise “fair” scoring rules that reward forecasters that seek the true distribution.

Definition 2. ***(Proper scoring rule)*** *A scoring rule S is proper (with respect to class $\mathcal{F}$) if the expected loss is minimized at the true CDF. i.e. if $Y \sim G$ then*

$$\mathbb{E}_G S(G, Y) \leq \mathbb{E}_G S(F, Y), \quad \forall F \in \mathcal{F}$$

A scoring rule is **strictly proper** if its expected value is *uniquely minimized* by the true probability distribution. Improper rules should be avoided as they could encourage the forecaster to present predictions that are believed by the forecaster to be incorrect. A detailed review of this topic can be found in Gneiting and Raftery, (2007) and Bröcker and Smith, (2007).

5.2.1 Probability score

The *Brier Score* (BS), first introduced by Brier (1950), is a type of proper scoring rule that evaluates forecast accuracy based on the Euclidean distance between the true likelihood of a binary observation (around a threshold) and the predicted probability assigned to the outcome to that observation. Vaguely speaking, Brier scores, also known as *probability scores*, showcase the predictive distribution’s ability to capture the true probability of an event’s occurrence. Formally, the evaluation of the predictive likelihood of a discrete event $Y \in A$ with $p = P_F[Y \in A]$ is characterised by

$$BS^A = (p - \mathbb{I}[y \in A])^2$$

where $\mathbb{I}$ is a $[0, 1]$ binary distribution that assigns probability 0 to events which did not occur and 1 to those that did. In the context of probabilistic models, the forecast seeks to find the models ability of capturing the likelihood of $y \leq z$ where z is an arbitrary threshold. The mapping is relatively straightforward as any density forecast f induces a probability forecast for the binary event $Y \leq z$ via the value of the associated cumulative distribution function (CDF),

$$F(z) = \int_{-\infty}^{z} f(y)dy$$

at the threshold z. Therefore, the Brier Score can be re-written as

$$BS_{t,h}^{z}(F_{t+h}(z), y_{t+h}) = (p_{t+h} - \mathbb{I}[y_{t+h} \leq z])^2 \tag{12}$$

where $p_{t+h} = P[Y_{t+h} \leq z] = F_{t+h}(z)$ for all $t = 0, 1, 2, \dots, T$. For uni-dimensional predictions, the Brier score is the probabilistic version of the squared error used for point forecast evaluation[8].

5.2.2 Quantile score

If F is a monotonically increasing cumulative distribution function, then it is possible to define a unique inverse function F^{-1}, often referred to as a *quantile* function. Quantile functions allow forecasters to asses the performance of the predictive distribution across quantiles (this is particularly relevant when assessing the ability of a model to predict tail-risk events). For this purpose, the most conventional (strictly proper) scoring rule is the *quantile score* (QS)[9] (Koenker & Bassett Jr, 1978). Formally, the quantile score is defined as

$$\text{QS}_{t,h}^{\alpha}(F_{t+h}^{-1}(\alpha), y_{t+h}) = 2(\mathbb{I}\{y_{t+h} < q\} - \alpha)(q - y_{t+h}) \tag{13}$$

where $q = F_{t+h}^{-1}(\alpha)$ for a quantile $\alpha \in (0, 1)$.

5.2.3 Continuous ranked probability score (CRPS)

The scoring rules analysed so far evaluate a specific portion of the distribution, either a probability region or quantile of the distribution. *Continuous ranked probability score* (CRPS)[10] allows forecasters to assess the predictive performance of the distribution as a whole. Formally,

$$\text{CRPS}_{t,h}(F_{t+h}, y_{t+h}) = \int_{-\infty}^{\infty} (F_{t+h}(x) - \mathbb{I}\{y_{t+h} \leq x\})^2 \, \mathrm{d}x \tag{14}$$

in the context of CRPS, $\mathbb{I}$ is a Heaviside step function that takes the value of 0 for any value below the true value and 1 for any value equal or above the true value (Matheson & Winkler, 1976).

[8] Some other types of scoring rules are the spherical score, logarithmic score, zero-one score.

[9] also known as the pinball score, and the asymmetric piece-wise linear score

[10] This is sometimes referred to as the *Stochastic euclidean error distance* presented by Diebold and Shin (2017).

One could also split the original integral into two integrals on the critical threshold $y_{t+h} = x$ to simplify the Heaviside step function,

$$\mathrm{CRPS}_{t,h}(F_{t+h}, y_{t+h}) = \int_{-\infty}^{y_{t+h}} F_{t+h}(x)^2 dx + \int_{y_{t+h}}^{\infty} (F_{t+h}(x) - 1)^2 dx$$

In practice, because F_{t+h} is an empirical distribution, there are only a finite number of points to evaluate, meaning the integrals can be turned into discrete finite sums that are computationally feasible.

Finally, notice that there is a strong link between the three scoring rules discussed so far. In fact, the first two are equivalent to CRPS when aggregated across the distribution. That is ,

$$\mathrm{CRPS}_{t,h}(F_{t+h}, y_{t+h}) = \int_{-\infty}^{\infty} BS^{z}_{t,h}(F_{t+h}(z), y_{t+h})\, \mathrm{d}z = \int_{0}^{1} \mathrm{QS}^{\alpha}_{t,h}(F^{-1}_{t+h}(\alpha), y_{t+h})\, \mathrm{d}\alpha \tag{15}$$

In the end, the performance of the scoring rule S at the horizon h was averaged across the sample to obtain an average score of Π_h,

$$\Pi_h = \frac{1}{T}\sum_{t=1}^{T} S_{t,h}$$

5.3 Testing for equal predictive performance

For a given loss function, two competing models (say i and j) may be tested to see if they have equal predictive performance using a *Diebold-Mariano* (DM) test. The formal test of equal forecast performance can be based on the statistic,

$$t_h = \sqrt{T}\,\frac{\Pi^i_h - \Pi^j_h}{\hat{\sigma}^2_h} \tag{16}$$

where

$$\hat{\sigma}^2_h = \frac{1}{T}\sum_{i=1}^{T}\left(S^i_{t,h} - S^j_{t,h}\right)^2 \tag{17}$$

is an estimate of the variance of the score differential. The DM test does not require any specific behaviour for individual scores, it does however, assume that the score differential is *covariance stationary*.

5.4 PIT scores

The evaluation methods described so far are only useful for relative comparison against a benchmark (or other models) as there is no standard measure of an "appropriate" CRPS value. To provide a notion of an "absolute" rather than a "relative" evaluation measure of predictive performance, forecaster commonly take a calibration analysis based on the use of probability integral transform (PIT) (Diebold, Gunther, & Tay, 1997).

A probability integral transform (PIT) is the cumulative probability evaluated at the actual, realised value of the target variable. It measures the likelihood of observing a value less than the actual realised value, where the probability is measured by the density forecast. According to (Diebold et al., 1997), a density forecast is correctly specified if 1) the probability integral transforms of the realisations are uniformly distributed over the interval $(0, 1)$, 2) for one-step-ahead forecasts[11], the PITs also display independence (meaning no auto-correlation).

6 Results

As mentioned in previous sections, thirty different models were tested using a recursive out-of-sample estimation across twelve horizons. The evaluation starts in January 2012 until December 2019, splitting the date approximately 50% for in-sample vs out-of-sample estimation. For illustration purposes ten models were selected (plus the benchmark) to be include in the charts and tables[12].

For a target variable like inflation, it is not obvious what transformation of prices index is more appropriate for a forecasting evaluation. For instance, it is very common for forecaster to forecast year-on-year inflation, but its also possible that forecaster are more interested in forecasting, monthly, quarterly, year-end or average year inflation. Because of this fact, the paper evaluates performance of the price index itself rather than a specific transformation of the data[13]. That said, as yearly inflation is a very conventional transformation, special attention will be paid to the twelve horizon ($h = 12$) when displaying some fixed horizon charts.

[11] In practice forecaster tend to test calibration on on-step-ahead forecasts. Although there is literature on multi-step-ahead forecasts, they yield serially correlated forecast errors, and, thus, serially correlated PITs, complicating the analysis (Knüppel, 2015). The calibration of multi-step-forecast goes beyond the scope of this paper.

[12] CRPS results for the rest of the models can be found in Table 5 within the appendix.

[13] Due to the non-stationary of CPI, the data was generally transformed to percentage differences or log differences during the estimation and forecasting process but was later reverted to CPI for comparison.

6.1 Out-of-sample testing

In general, metrics including Continuous Ranked Probability Scores (CRPS), Quantile Scores (QS), Root Mean Square Errors (RMSE) were displayed in relative terms with respect to the benchmark. Readers should note that a lower CRPS is desired, therefore a lower relative performance to the benchmark actually means predictive gains versus the benchmark (a lower relative CRPS / QS / RMSE is desired). Mean Percentage Errors (MPE) was also used, in order to check for bias across point forecasts models[14].

Figure (4) shows the relative CRPS performance of the models. Notice there is a subset of models which, for the given data sample and selected out-of-sample dates, outperform the benchmark (lower relative CRPS) across all horizons, in contrast with other models which only outperform the benchmark at some horizons. For instance, the selected long-run model, which has two long-run relationships, under-performs the benchmark at shorter horizons (1-5 horizons) but it outperforms at further horizons.

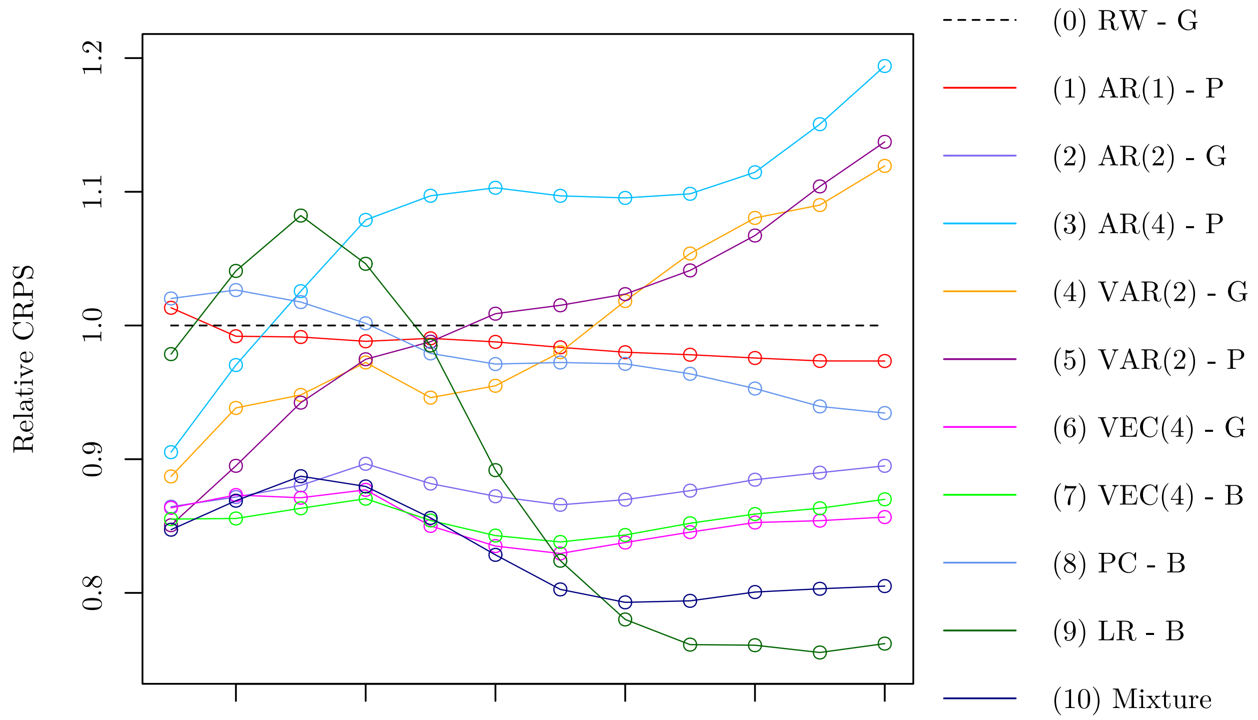


Figure 4: Relative CRPS by horizon

[14]Note that in this case the performance is not compared relative to the benchmark as it could have been possible that the benchmark exhibit a substantial bias, if so, a relative comparison would be erroneous.

On the other hand, the selected VAR models tend to outperform in shorter horizons but fail to capture longer-term dynamics. In general, outperforming models have between 5% and 15% gains when compared to the random walk, while the long-run model has clearly a better performance relative to all the other models on a 9-12 horizon (20-25% above the benchmark).

Figure (5) shows QS at horizon 12. It is evident from the chart that the performance was very different across quantiles. Notice that in general, models tend to exhibit a rather similar performance at the median.

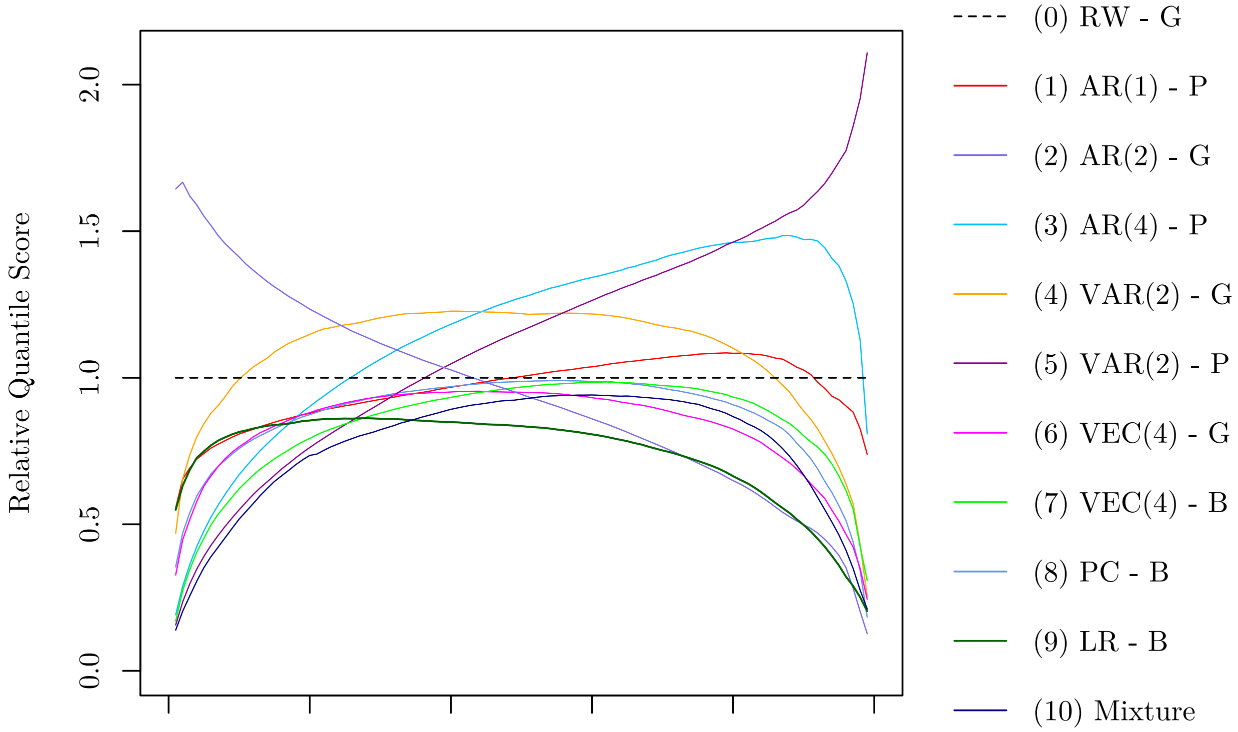


Figure 5: QS by quantiles for $h = 12$

However, there are models that are better than the benchmark at predicting one of the tail. For instance, models 3 and 4 have better performance at the left-tail, while others have better performance only at the right-tail (high-inflation-risks), like model 2. Some of the models have similar performance at the median but out-perform at both tails.

Figure (6) shows the cumulative CRPS differentials. This metric helps to understand evolution of the model performance (relative to the benchmark) across the sample. It should be noted that the metric can not be expressed as a percentage because in early periods the cumulative CRPS scores are approximately zero, causing instability in the performance making it impossible to interpret. Therefore the relative performance is showcased as the cumulative differences of the CRPS levels with respect to the bench mark.

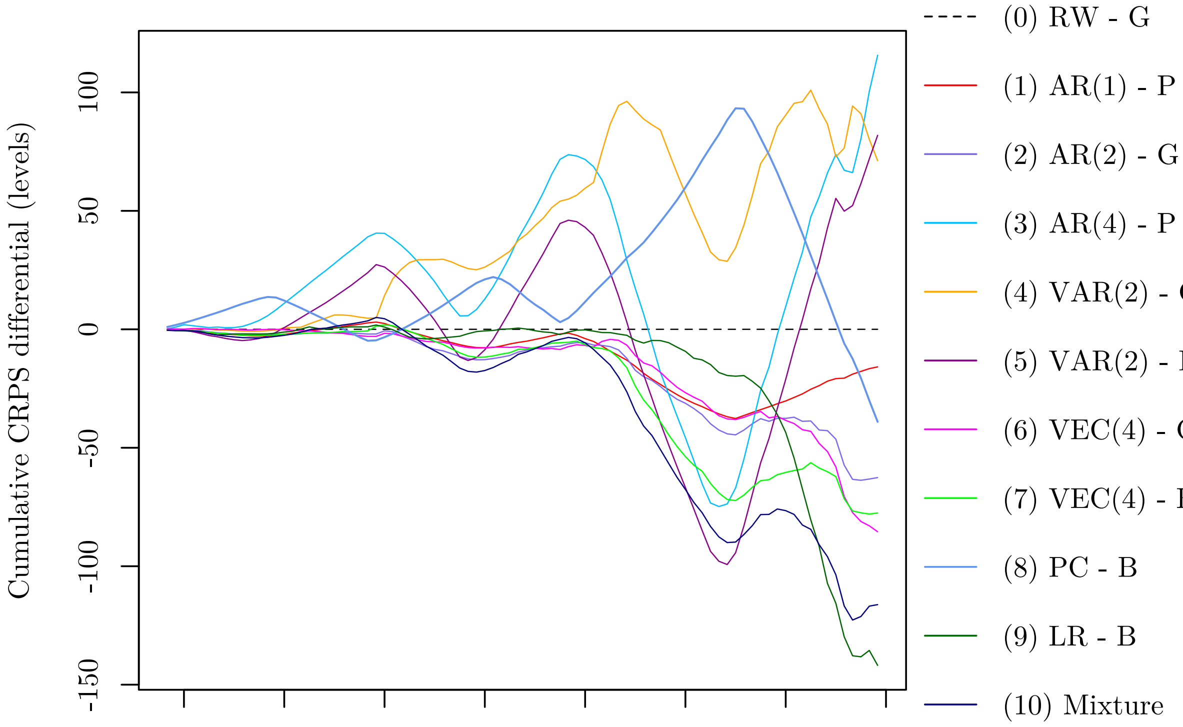


Figure 6: CCRPS across out-of-sample window for $h = 12$

Interestingly, most models had a similar performance to the benchmark until 2016-2017, then the performance of most models deviated significantly. In general, it can be argued that the benchmark failed to capture the accelerating inflation risks from 2016-2019 versus other models with long-run relationships like the VEC models or the Long-run model. Lastly, notice that despite the fact that the Long-run model outperformed the rest of the models at horizon twelve, the performance of this model only improved notoriously over the last 24 months. Although a model like such is an option to consider, in practice, it is perhaps more appropriate to seek for models with consistently better performance across the sample as opposed to some specific periods in time. In this case, the mixture model, despite having a lower final CRPS score than the Long-run model, has a consistently better cumulative performance across the whole sample with the exception of the last periods. This result highlights the attractiveness of model combinations as they may have better but also more stable performance than a single model.

A point forecast evaluation was also conducted by taking the median of the probabilistic forecast. Figure (7) shows the MPE across horizons. The results indicate that three of the selected models

(models 3, 4 and 5), presented noticeable bias on their point forecasts at longer horizons. The rest of the models exhibited a bias of less than $\pm 1\%$.

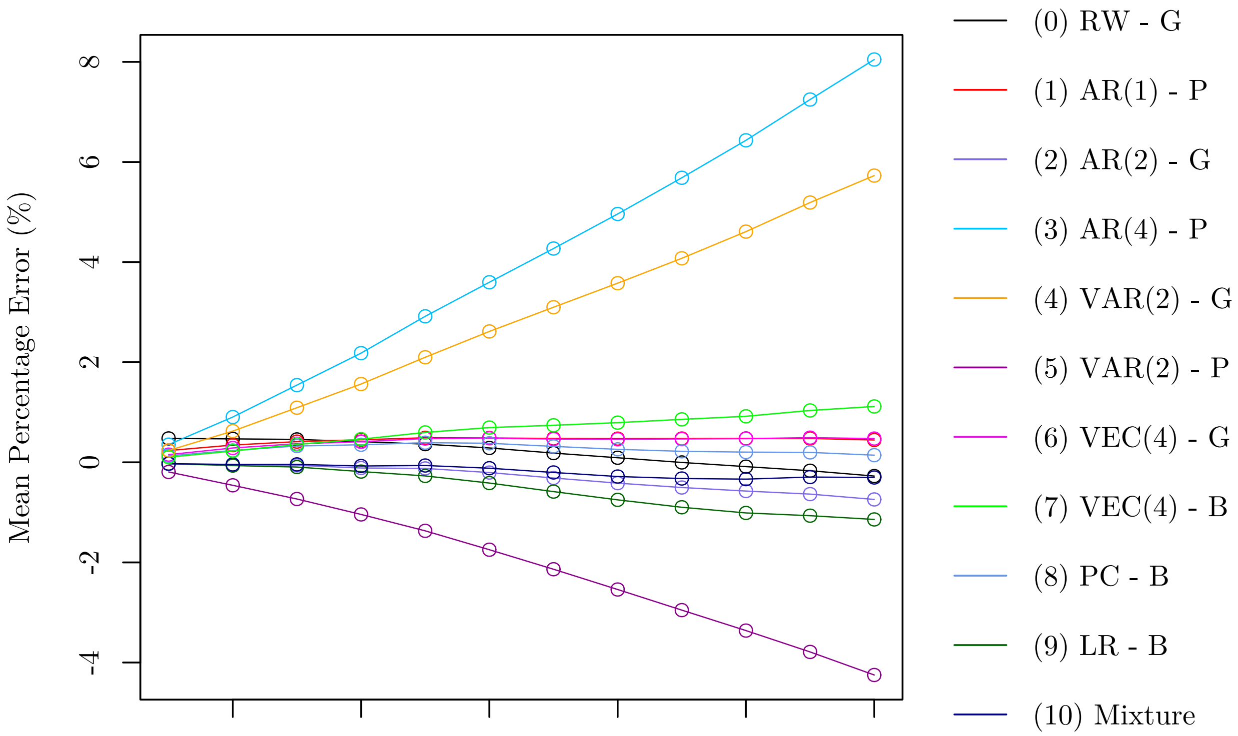


Figure 7: MPE by horizon

Figure (8) compares the RMSE of the point forecast. Notice that this metric unveils slightly different results than the CRPS analysis. For instance, notice that the performance with respect to the benchmark worsened significantly for models 3, 4 and 5. This is associated with the fact that the median of these models exhibited notorious bias at longer horizons, yet the modes displayed some improvements in the tails improving the overall CRPS score. Notice however, with the exception of the long-run model, the relative gains of the rest of the models narrowed remarkably in contrast to the results shown by the CRPS.

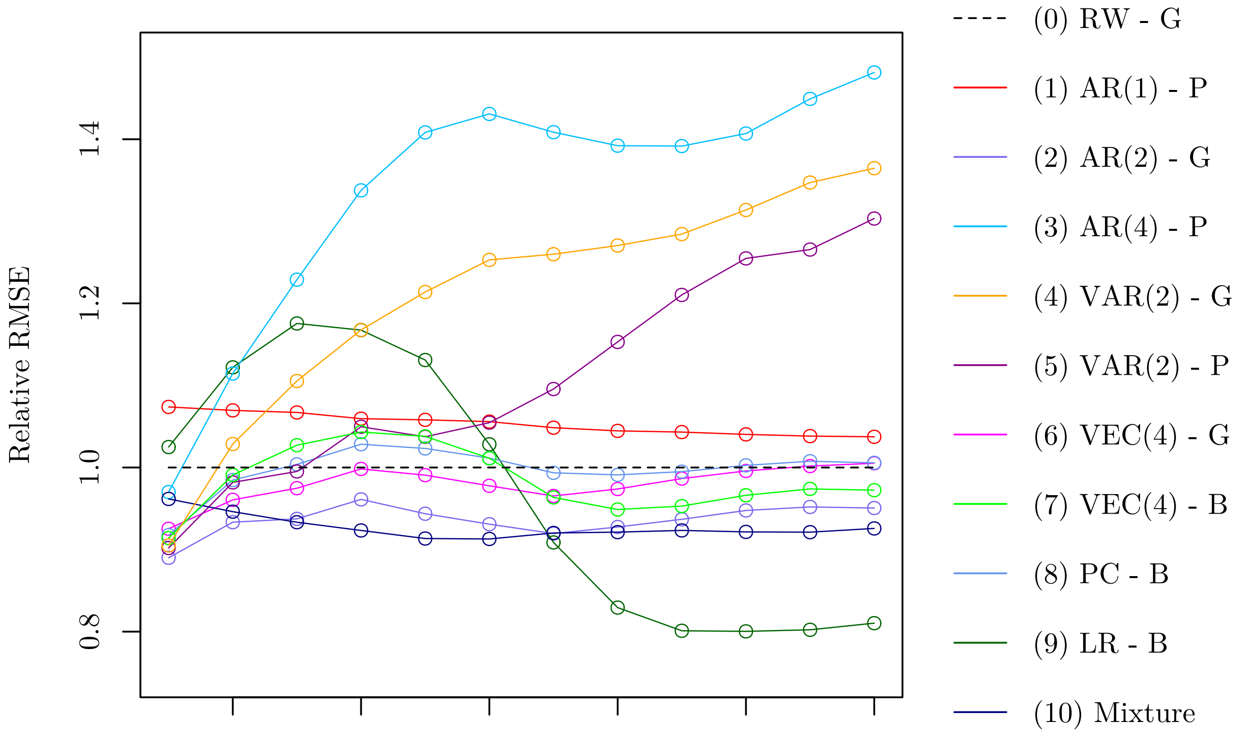


Figure 8: Relative RMSE by horizon

Although this was somewhat expected, as the quantile score showed that models tend to have better performance at the tails than at the median[15] the implications remain meaningful. One could argue that in general the difference in performance between multivariate models with parsimonious models, like random walk and AR models, might not be as evident in point forecasts. However, there is a much greater opportunity to exploit multivariate models, including models with direct links to economic theory, in probability forecast as they might be able to capture other embedded dynamics that are not present in regular (base case) scenarios.

6.2 DM test results

A Diebold-Mariano (DM) test was used to formally evaluate the predictive performance of the probabilistic models. We specifically chose to test the predictive ability of two models, the mixture model and the AR(2) against the Random Walk. The selected univariate model was chosen because it was the best univariate model across horizons. The mixture model, on the other hand, was not superior at all horizons but was selected among all the multivariate models because it yielded the best CRPS

[15] Although this was only check for the twelve horizon, it is not ridicule to assume a similar situation for the other horizons.

score averaged across horizons. Predictive ability was tested for both models versus the Random Walk but also against each other. As it was mentioned before, the test is only valid in those cases where the CRPS differential is stationary. An Augmented Dickey-Fuller (ADF) test was applied on the CRPS differentials across horizons (see Table 2). Horizons that failed to reject the unit root hypothesis were discarded for the DM test.

Model 1 - Model 2	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}	h_{12}
AR(2) - RW	0.01^*	0.02^*	0.01^*	0.03^*	0.06^*	0.11	0.08^*	0.06^*	0.10^*	0.9^*	0.16	0.28
Mixture - RW	0.21	0.32	0.06^*	0.08^*	0.05^*	0.05^*	0.9^*	0.8^*	0.11	0.07^*	0.09^*	0.16
Mixture - AR(2)	0.01^*	0.01^*	0.12	0.32	0.23	0.09^*	0.09^*	0.10^*	0.8^*	0.05^*	0.06^*	0.05^*

Horizons that exhibited a p-value ≤ 0.1 were represented by $(*)$.

Table 2: Augmented Dickey-Fuller (ADF) test for stationary (p-values)

From Table 2, the accepted horizons were marked and the DM test was enforced using the following premise.

Null Hypothesis: Model 1 and Model 2 have equal predictive ability.

Alt. Hypothesis: Model 1 has a superior predictive ability than Model 2.

Model 1 - Model 2	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}	h_{12}
AR(2) - RW	0.01^*	0.01^*	0.02^*	0.02^*	0.03^*	-	0.16	0.19	0.17	0.37	-	-
Mixture - RW	-	-	0.30	0.14	0.07^*	0.03^*	0.02^*	0.02^*	-	0.03^*	0.01^*	-
Mixture - AR(2)	0.27	0.30	-	-	-	0.22	0.16	0.28	0.33	0.09^*	0.4^*	0.01^*

Horizons that exhibited a p-value ≤ 0.1 were represented by $(*)$.

Table 3: Diebold-Mariano test for predictive accuracy (p-values)

Table 3 illustrates the p-values of the DM test between the random walk versus the selected AR model and the mixture model. The mixture model failed to reject the null hypothesis at shorter horizons (h_3, h_4), while the AR model failed to reject the hypothesis at longer horizons (h_7, h_8, h_9, h_{10}). When compared to each other, the mixture model outperformed the AR model at (h_{10}, h_{11}, h_{12}). The results are in line with some of the common premises in the macroeconomic forecasting literature; 1) Random walks or some other types of parsimonious models tend to be as good as multivariate models in shorter horizons but may under-perform in longer horizons. 2) Using multivariate models with greater level of sophistication may be more effective at forecasting longer horizons.

6.3 PIT evaluation results

A PIT evaluation was conducted on the mixture model in order to asses the calibration of the model. Once again, the mixture model was chosen from all the models as it had the best performance averaged across all horizons. Following (Rossi, 2014), a histogram and ACF plot of the PITs was taken. Figure (9) does not show any signs of model miss specification. PITs do not exhibit auto-correlation suggesting independence and the histogram revels a uniform distribution. This is a good sign in fact as it suggest a correct calibration.

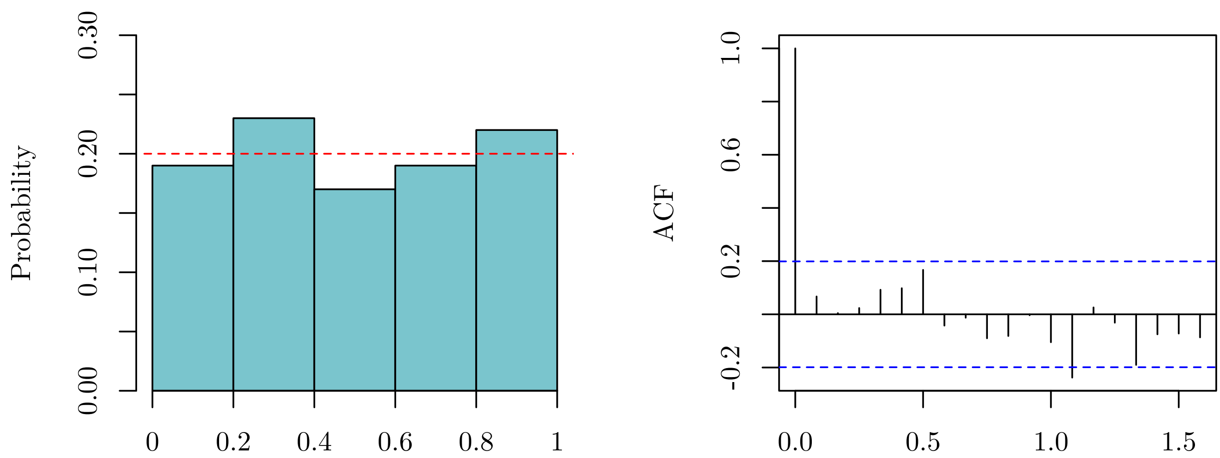


Figure 9: PIT evaluation results for one-step-ahead forecasts

If the forecasts lacked calibration, the shape of the PIT histogram would reveal the nature of the misspecification. For instance, a $\bigcup$-shaped is a sign of underdispersion as many observations are considered to be too extreme when in fact they are more common in practice suggesting that the predictive density is too narrow. Conversely, over dispersion is reflected in a hump or $\bigcap$-shape as the distributions are too wide. Bias causes a inclinations or triangular shapes towards an extreme, generally a “L” or “J” shape, depending on the direction of the bias.

7 Conclusion

This paper explores the use of probability forecasts to predict inflation in Argentina using a range of autoregressive models. Different metrics were used to asses the performance of the point forecast but also the entire distribution across different horizons. A Diebold-Mariano (DM) test was applied to selected models to test predictive ability. For the mixture model, a PIT evaluation was conducted and the qualitative interpretations suggest the model was correctly calibrated.

The results show that some of the models statistically outperform the benchmark at particular horizons, but there is no unique model that outperforms the benchmark at every horizon. In general, models with structure (either VEC models or theory-related models linked to wages and money growth) have a better performance.

A key feature to take from this forecasting exercise is that although some models may be better at forecasting central events (mean or median values), they may not be able to appropriately capture other moments of the distribution. For instance, the performance between the random walk and the mixture model is relatively similar at the median, however, the mixture model is significantly better at capturing tail risk events.

Equally weighted mixture models were used as a way of exploring forecast combinations. Because of the short nature of the sample, the use of other types of modelling techniques was limited. Further research should incorporate dynamic combinations like Bayesian model averaging or Dynamic model averaging techniques (Koop & Korobilis, 2012). Perhaps, a DMA combination of models with different theory-related structures could allow forecasters to extend the sample backwards to capture shifts in regimes (such as the hyper-inflationary phase in the 80's, or the hard peg exchange rate policy in the 90's). On the other hand, the evaluation techniques described in this paper may very well be replicated for DSGE models often used by central banks.

8 Appendix

	Var. Treatment	Exp.	EMAE	Ex. rate	Wage	MS	Int. rate	U.S. CPI	U.S. Int. rate
(11) AR(1)	Garch (1,1)								
(12) AR(2)	Parametric								
(13) AR(3)	Parametric								
(14) AR(3)	Garch (1,1)								
(15) AR(4)	Garch (1,1)								
(16) VAR(2)	Bootstrap		X		X		X		
(17) VAR(3)	Parametric	X		X		X	X		
(18) VAR(3)	Garch (1,1)	X		X		X	X		
(19) VAR(3)	Bootstrap	X		X		X	X		
(20) VAR(4)	Parametric	X	X		X		X		X
(21) VAR(4)	Garch (1,1)	X	X		X		X	X	
(22) VAR(4)	Bootstrap	X	X		X		X	X	
(23) VEC(2)	Parametric		X			X			
(24) VEC(2)	Garch (1,1)		X			X			
(25) VEC(3)	Bootstrap			X		X			
(26) VEC(3)	Garch (1,1)			X		X			
(27) VEC(4)	Parametric	X		X			X	X	X
(28) VEC(4)	Garch (1,1)	X		X			X	X	X
(29) PC	Parametric	X	X	X				X	
(30) Long-Run Alt	Bootstrap		X	X			X	X	X

Table 4: The rest of the models and variables included

Model	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}	h_{12}
Random Walk	0.622	1.181	1.745	2.304	2.863	3.399	3.964	4.574	5.193	5.81	6.398	7.30
Model 11	0.741	1.651	2.451	3.136	3.806	4.603	5.226	6.002	6.729	7.541	8.744	10.094
Model 12	0.749	1.683	2.494	3.208	3.92	4.781	5.473	6.351	7.233	8.18	9.668	11.415
Model 13	0.576	1.17	1.828	2.573	3.199	3.766	4.266	4.768	5.289	5.955	6.764	7.66
Model 14	0.591	1.233	1.91	2.596	3.213	3.861	4.582	5.427	6.345	7.361	8.392	9.432
Model 15	0.665	1.447	2.282	3.117	3.969	4.772	5.609	6.625	7.702	8.659	9.479	10.159
Model 16	0.567	1.211	1.871	2.543	3.123	3.803	4.544	5.405	6.382	7.402	8.417	9.412
Model 17	0.549	1.144	1.803	2.511	3.209	3.917	4.564	5.311	6.108	6.994	7.916	8.724
Model 18	0.628	1.279	1.941	2.641	3.264	4.016	4.779	5.767	6.878	8.033	9.178	10.237
Model 19	0.798	1.678	2.371	3.353	4.092	4.975	5.708	6.613	7.66	8.542	9.911	11.395
Model 20	0.580	1.212	1.877	2.562	3.162	3.737	4.243	4.805	5.427	6.166	6.977	7.843
Model 21	0.694	1.417	2.153	2.974	3.537	4.185	4.724	5.217	5.991	6.904	8.011	9.256
Model 22	0.717	1.541	2.398	3.264	4.077	5.003	5.733	6.372	7.367	8.625	10.117	11.832
Model 23	0.689	1.479	2.291	3.257	3.988	4.711	5.388	6.132	7.106	8.293	9.454	10.702
Model 24	0.739	1.588	2.374	3.423	4.171	4.959	5.673	6.449	7.366	8.467	9.576	10.646
Model 25	0.536	1.068	1.66	2.267	2.847	3.452	4.044	4.701	5.427	6.239	7.103	7.909
Model 26	0.755	1.599	2.34	3.343	4.343	5.252	6.114	7.201	8.398	9.683	11.247	12.979
Model 27	0.944	1.933	2.557	3.215	3.812	4.338	4.969	5.737	6.122	6.942	8.146	9.307
Model 28	0.924	1.884	2.594	3.188	3.867	4.531	5.125	5.888	6.657	7.406	8.215	9.184
Model 29	0.764	1.613	2.35	3.026	3.861	4.608	5.426	6.527	7.576	8.759	9.975	11.316
Model 30	1.013	2.145	2.835	3.479	3.883	4.122	4.599	5.284	6.037	6.306	6.905	8.052

Table 5: CRPS by horizon for the remaining models

References

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, *31*(3), 307–327.

Bollerslev, T., Engle, R. F., & Wooldridge, J. M. (1988). A capital asset pricing model with time-varying covariances. *Journal of political Economy*, *96*(1), 116–131.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1–3.

Bröcker, J., & Smith, L. A. (2007). Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, *22*(2), 382–388.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, *5*(4), 559–583.

Cordeiro, C., & Neves, M. (2006). The bootstrap methodology in time series forecasting. *Proceedings of CompStat2006"(J. Black and A. White, Eds.), Springer Verlag*, 1067–1073.

D'Amato, L., Aguirre, M. G., Garegnani, M. L., Krysa, A., & Libonatti, L. (2018). *Forecasting inflation in argentina: A comparison of different models* (Tech. Rep.). Economic Research Working Papers.

Diebold, F. X., Gunther, T. A., & Tay, A. (1997). *Evaluating density forecasts.* National Bureau of Economic Research Cambridge, Mass., USA.

Diebold, F. X., & Shin, M. (2017). Assessing point forecast accuracy by stochastic error distance. *Econometric Reviews*, *36*(6-9), 588–598.

Duffie, D., & Pan, J. (1997). An overview of value at risk. *Journal of derivatives*, *4*(3), 7–49.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics* (pp. 569–593). Springer.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the econometric society*, 987–1007.

Engle, R. F., & Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, 251–276.

Garratt, A., Lee, K., Hashem Pesaran, M., & Shin, Y. (2003). A long run structural macroeconometric model of the uk. *The Economic Journal*, *113*(487), 412–455.

Garratt, A., Lee, K., Pesaran, M. H., & Shin, Y. (2003). Forecast uncertainties in macroeconomic modeling: An application to the uk economy. *Journal of the American Statistical Association*, *98*(464), 829–838.

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*, 125–151.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, *102*(477), 359–378.

Granger, C. W. (1981). Some properties of time series data and their use in econometric model specification. *Journal of econometrics*, *16*(1), 121–130.

Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a garch (1, 1)? *Journal of applied econometrics*, *20*(7), 873–889.

Knüppel, M. (2015). Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *Journal of Business & Economic Statistics*, *33*(2), 270–281.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Koop, G., & Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, *53*(3), 867–886.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Manz, C. C., & Sims Jr, H. P. (1980). Self-management as a substitute for leadership: A social learning theory perspective. *Academy of Management review*, *5*(3), 361–367.

Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management science*, *22*(10), 1087–1096.

Rossi, B. (2014). Density forecasts in economics, forecasting and policymaking.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*(336), 783–801.

Schneider, E., Chen, P., & Frohn, J. (2008). A long-run structural macroeconometric model for germany: An empirical note. *Economics*, *2*(1).

Stenberg, E. (2016). *On the autoregressive conditional heteroskedasticity models.*

Svensson, L. E. (2000). Open-economy inflation targeting. *Journal of international economics*, *50*(1), 155–183.

Winkler, R. L., & Murphy, A. H. (1968). "good" probability assessors. *Journal of Applied Meteorology and Climatology*, *7*(5), 751–758.