ISSN 1518-3548

BANCO CENTRAL DO BRASIL

Evaluation of Exchange Rate Point and Density Forecasts: an application to Brazil

Wagner Piazza Gaglianone and Jaqueline Terra Moura Marins

November, 2016

Working Papers

446

ISSN 1518-3548
CGC 00.038.166/0001-05

Working Paper Series	Brasília	n. 446	November	2016	p. 1-51

Working Paper Series

Edited by Research Department (Depep) – E-mail: workingpaper@bcb.gov.br

Editor: Francisco Marcos Rodrigues Figueiredo – E-mail: francisco-marcos.figueiredo@bcb.gov.br

Co-editor: João Barata Ribeiro Blanco Barroso – E-mail: joao.barroso@bcb.gov.br

Editorial Assistant: Jane Sofia Moita – E-mail: jane.sofia@bcb.gov.br

Head of Research Department: Eduardo José Araújo Lima – E-mail: eduardo.lima@bcb.gov.br

The Banco Central do Brasil Working Papers are all evaluated in double blind referee process.

Authorized by Carlos Viana de Carvalho, Deputy Governor for Economic Policy.

General Control of Publications

Banco Central do Brasil

Comun/Dipiv/Coivi

SBS – Quadra 3 – Bloco B – Edifício-Sede – 14º andar

Caixa Postal 8.670

70074-900 Brasília – DF – Brazil

Phones: +55 (61) 3414-3710 and 3414-3565

Fax: +55 (61) 3414-1898

E-mail: editor@bcb.gov.br

Citizen Service Division

Banco Central do Brasil

Deati/Diate

SBS – Quadra 3 – Bloco B – Edifício-Sede – 2º subsolo

70074-900 Brasília – DF – Brazil

Toll Free: 0800 9792345

Fax: +55 (61) 3414-2553

Internet: <http//www.bcb.gov.br/?CONTACTUS>

Evaluation of exchange rate point and density forecasts: an application to Brazil*

Wagner Piazza Gaglianone†

Jaqueline Terra Moura Marins‡

Abstract

The Working Papers should not be reported as representing the views of the Banco Central do Brasil. The views expressed in the papers are those of the author(s) and do not necessarily reflect those of the Banco Central do Brasil.

In this paper, we construct multi-step-ahead point and density forecasts of the exchange rate, from statistical or economic-driven approaches, using financial or macroeconomic data and using parametric or nonparametric distributions. We employ a set of statistical tools, from different strands of the literature, to identify which models work in practice, in terms of forecast accuracy across different data frequencies and forecasting horizons. We propose a novel full-density/local analysis approach to collect the many test results, and deploy a simple risk based decision rule to rank models. An empirical exercise with Brazilian daily and monthly data reveals that macro fundamentals matter when modeling the risk of exchange rate appreciation, whereas models using survey information or financial data are the best way to account for the depreciation risk. These findings have relevance for econometricians, risk managers or policymakers interested in evaluating the accuracy of competing exchange rate models.

Keywords: Density forecasts, Exchange rate, Risk, Model selection.

JEL codes: C14, C15, C53, E37, F31.

*An earlier version of this paper circulated in 2014 as "Risk Assessment of the Brazilian FX Rate", Working Paper n.344, Banco Central do Brasil. The authors thank Sergio Lago Alves, Waldyr Areosa, João Barata Barroso, Levent Bulut, Carlos Viana de Carvalho, Neil Ericsson, Aquiles de Farias, Osmani Guillén, João Victor Issler, Emanuel Kohlscheen, Eduardo Araújo Lima, Luiz Renato Lima, Michael Moore, Ricardo Schechtman and Jouko Vilmunen for their helpful comments and suggestions. We also benefited from comments given by the seminar participants of the IX Annual Seminar on Risk, Financial Stability and Banking of the Banco Central do Brasil (São Paulo, Brazil), the 35th International Symposium on Forecasting (Riverside, USA), the 9th International Conference on Computational and Financial Econometrics - CFE 2015 (London, UK) and the Workshop da Rede de Pesquisa do Banco Central 2016 (Brasília, Brazil). Gaglianone gratefully acknowledges the support from CNPq and INCT on different grants.

†Corresponding author. Research Department, Banco Central do Brasil. E-mail: wagner.gaglianone@bcb.gov.br

‡Research Department, Banco Central do Brasil. E-mail: jaqueline.terra@bcb.gov.br

1 Introduction

The foreign exchange (FX) rate market is one of the most important in the financial system. According to the report of the Bank for International Settlements (BIS, 2013), trading in foreign exchange markets averaged US$5.3 trillion per day in April 2013.[1] Besides its huge trading volume, it also represents the largest asset class in the world leading to high liquidity.[2] Other features of this market are the high volatility and the potential variety of factors that might affect exchange rates (e.g. economic fundamentals, speculative transactions and currency interventions, among many others).

Forecasting exchange rate is of great importance for economic agents, in particular, for investors and policy makers. Accurate forecasts of FX rates allow investors, for instance, to design adequate trading strategies and to hedge against market risk. On the other hand, central banks worldwide closely monitor the daily FX movements, since they impact future price dynamics and, thus, help setting the appropriate interest rate policy (Groen and Matsumoto, 2004). Besides, it is a useful information for central bankers to decide for interventions.

In practical terms, however, accurately forecasting the FX rate has proved to be a nontrivial exercise. The failure of standard economic theory to explain foreign exchange rate behavior using key economic fundamentals (such as the money supply, trade balance and national income) has prevailed in the international economics literature since the classical papers of Meese and Rogoff (1983a,b). The authors investigated the out-of-sample forecasting performance of standard exchange rate models during the post-Bretton Woods period and concluded that such models do not perform better than a naive random walk (RW) forecast.[3] Indeed, the macroeconomic theory has proposed several potential predictors of exchange rates (usually based upon the Purchasing Power Parity (PPP) hypothesis, the Uncovered Interest Rate (UIP) parity condition and the monetary model). However, the forecasting

[1] According to the same report, it is up from US$4.0 trillion in April 2010 and US$3.3 trillion in April 2007.

[2] Nonetheless, in the long run, the attractiveness of carry trade strategies relative to other investments is not clear. Indeed, there is a large literature that started with Burnside et al. (2006), which suggests that market frictions greatly reduce the profitability of currency speculation strategies.

[3] The random walk forecast is such that the (log) level of the nominal exchange rate is predicted to remain at the current (log) level (also known as the "no change" forecast).

contribution of such approaches has been under question since the highly influential findings of Meese and Rogoff. In this sense, Bacchetta and van Wincoop (2006) describe the RW paradigm as "*...the major weakness of international macroeconomics.*"[4]

Consequently, an extensive literature has studied the forecasting performance of empirical exchange rate models and several (potential) explanations have been put forward. Just to mention a few papers: Mark (1995) finds evidence of greater predictability of economic exchange rate models at longer horizons, although these findings have been questioned later by Kilian (1999). Kilian and Taylor (2003) argue that exchange rates can be predicted from economic models after taking into account the possibility of nonlinear exchange rate dynamics. Cheung et al. (2005) examine the out-of-sample performance of the interest rate parity, monetary, productivity-based and behavioral exchange rate models and conclude that (indeed) none of these models consistently beats the RW forecast at any horizon. The authors argue that even if a particular macroeconomic "fundamental" has some level of predictive power for a bilateral exchange rate (at a certain horizon), the same variable may show no predictive power at different horizons or for other bilateral exchange rates. On the other hand, Engel and West (2005) argue that it is not surprising that a random walk forecast outperforms fundamental-based models under some circumstances. The argument is based on the behavior of the exchange rate as an asset price within a rational expectation present-value (Taylor rule) model, among others, with a discount factor near one. Finally, there is a large and growing literature that aims at explaining currency movements in a cross-sectional rather than in a time-series framework (e.g. Burnside et al., 2011; Lustig et al., 2011; Menkhoff et al., 2012; Verdelhan, 2013). Its main findings have been used to address exchange rate predictability in a broadest sense based on multiple currencies.[5]

[4] See Rossi (2013a) for more on Meese and Rogoff and a review of the recent literature on exchange rate forecasting.

[5] As complementary lines of research, see also the following papers: Wu (2008) studies the importance of the order flows at short horizons, within the "microstructure approach". Engel et al. (2009) based on a panel of exchange rates argue that in the presence of stationary, but persistent, unobservable fundamentals, long-horizon predictability prevails in FX rate forecasting. Della-Corte et al. (2009) discuss the forward premium and its promising results in a portfolio allocation framework. Chen and Tsang (2009) find that cross-country yield curves are useful in predicting exchange rates. Molodtsova and Papell (2009) extend the standard set of exchange rate models by incorporating Taylor rule fundamentals. More recently, Fratzscher et al. (2012) investigate the scapegoat theory (as an attempt for explaining the poor performance of traditional models), and Morales-Arias and Moura (2013) explore forecast

In a distinct but complementary approach, several papers in the late 90s started investigating the random walk paradigm from a different view: out-of-sample density forecasting. For instance, Diebold, Hahn, and Tay (1999) use the RiskMetrics model to compute half-hour-ahead density forecasts for Deutschmark–dollar and yen–dollar returns. Christoffersen and Mazzotta (2005) construct option-implied density and interval forecasts for four major exchange rates. Clews et al. (2000) describes a nonparametric way to forecast risk neutral densities, from the smile interpolation of option prices. Boero and Marrocu (2004) obtain one-day-ahead density forecasts for euro nominal effective exchange rate using self-exciting threshold autoregressive (SETAR) models. Sarno and Valente (2005) use information from the term structure of forward premia to evaluate the FX rate density forecast performance of a Markov-switching vector error correction model (MS-VECM). Hong et al. (2007) construct half-hour-ahead density forecasts for euro-dollar and yen-dollar rates using a set of univariate time series models that capture fat tails, time-varying volatility and regime switches.

In general, these previous studies on exchange rate density forecasting use high frequency data, which are not available for most conventional economic fundamentals. In addition, these studies quite often do not consider multi-step-ahead forecasts and, generally, assume that conditional densities are analytically constructed (i.e. based on parametric densities). Wang and Wu (2012) tackle these issues by using a semiparametric method, applied to a group of exchange rate models, to generate out-of-sample exchange rate interval forecasts. The authors suggest that economic fundamentals might provide useful information in (out-of-sample) forecasting FX rate distributions. Based on forecast intervals for ten OECD exchange rates, the authors find that, in general, FX models generate tighter forecast intervals than the random walk, given that their intervals cover out-of-sample exchange rate realizations equally well. Moreover, the results suggest a connection between exchange rates and fundamentals: economic variables (indeed) contain information useful in forecasting distributions of exchange rates. In this sense, the Taylor rule model (Molodtsova and Papell, 2009) performs better than the monetary, PPP and forward premium models, and its advantages are more pronounced at longer horizons.

combination based on panel data and adaptive forecasting.

In this paper, we also go beyond point forecasting and follow the previous strand of literature focused on density forecasting. We address the subject by considering statistical approaches (such as GARCH), economic-driven models, and a financial data setup (treating the exchange rate as an asset price). We employ monthly data, as well as daily data, that enable us to investigate standard macroeconomic models for point and density forecast, constructed here from both parametric, nonparametric and semiparametric setups.

In addition, based on a set of density forecasts, generated for horizons from one to twelve months (or from one to twenty workdays), we go a step further and ask the following question: which is the best forecasting model for a given forecast horizon, and a given part of the conditional distribution of the FX rate? The objective here is to investigate a set of FX rate models and reveal which are more useful for point and/or density forecasting.

Moreover, we aim to increase our understanding of the exchange rate dynamics from a risk-analysis perspective. The main motivation is that macroeconomic fundamentals may vary in their predictive content at distinct parts of the distribution of the FX rate. In other words, our objective here is also to investigate risk measures of FX rate generated from distinct approaches, which may reveal potential links between exchange rates and economic fundamentals (or financial variables) that a simple point forecast evaluation might neglect.

This way, our main contribution is to bring together a whole set of statistical tools, from distinct strands of the literature (e.g. international economics, forecasting and risk management) to investigate the FX rate dynamics, in terms of point and density forecast, through the lens of competing models. In addition, we also conduct a local analysis of the competing density forecasts and use a simple decision rule for model ranking, employed for risk assessment purposes.

Why study the Brazilian case? Besides being one of the largest emerging economies, the Brazilian currency (Real) experienced a huge depreciation in the recent years, becoming one of the most volatile currencies among the emerging markets. For instance, among the BRICS countries, the Real depreciated an amount of 97% from March 2011 to March 2015 (only surpassed by the Russian Rublo, which devalu-

ated 106% in the same period, by comparing end-of-month figures).[6] This sharp devaluation of the domestic currency (compared to the U.S. dollar) has severe consequences for the Brazilian macroeconomic environment, for instance, by increasing the Brazilian inflation, with direct implications for monetary policy and market agents' expectations.

This paper is organized as follows: Section 2 presents the point and density forecast models, and the respective estimation schemes, as well as the adopted forecast evaluation tools. Section 3 presents our empirical exercise to investigate the Brazilian FX rate, based on a set of out-of-sample multi-step-ahead point and density forecasts. Section 4 concludes.

2 Methodology

2.1 Point and density forecast models

Along this paper, we investigate $m = 1, ..., 14$ models to construct the point (and density) forecasts for the nominal exchange rate (s_t) of the Brazilian Real with respect to the U.S. dollar (R$/US$).[7] The objective here is not to propose the best model to forecast the FX rate, but rather to evaluate a given set of available models to forecast the foreign exchange rate, within a range of forecast horizons. Following the notation of Wang and Wu (2012), a general setup of the (point forecast) model m takes the form of:

$$s_{t+h} - s_t = \mathbf{X}'_{m,t}\boldsymbol{\beta}_{m,h} + \varepsilon_{m,t+h} \tag{1}$$

in which $s_{t+h} - s_t$ is the h-periods change of the (log) exchange rate, $\mathbf{X}'_{m,t}$ is a vector with economic variables used in model m and $\varepsilon_{m,t+h}$ is the error term. Regarding multi-period ahead forecasts ($h > 1$), notice that we follow the "direct forecast" approach, in contrast to the "recursive (or iterated) forecast" route. See Marcellino, Stock and Watson (2006)[8] for a good discussion on this issue. We next briefly

[6] In the last sample observation (March 2015), the Real devaluated 11.5% compared to the previous month, which is the highest figure among the BRICS (the Russian Ruble: -4.6%; the Indian Rupee: 1.3%; the Chinese Yuan: -0.1%; and the South African Rand: 4.8%).

[7] The term "model" is used throughout this paper in a broad sense that includes forecasting methods.

[8] "Iterated" multi-period ahead time series forecasts are made using a one-period ahead model, iterated forward for the desired number of periods, whereas "direct" forecasts are made using a horizon-specific estimated model,

describe each model:

Model 1 (benchmark) is a standard random walk (RW) model without drift, coupled with a Gaussian distribution to generate the density forecast; in which the location of the distribution is the RW point forecast, and the variance of the distribution is given by the sample variance of past forecast errors.

Model 2 is a forward-looking approach based on financial data and the extraction of information from option prices.[9] It consists of two major steps: (i) obtaining risk-neutral densities (RND) and (ii) transforming these densities into real world densities (RWD). The RND for an asset price gives the set of probabilities that investors would attach to the future asset prices in a world in which they were risk-neutral. But if investors are risk-averse (as they usually are), risk premia will drive a wedge between the probabilities inferred from options (RND) and the true probabilities they attach to alternative values of the underlying asset price (RWD). See Appendix A for further details.

Model 3 is based on an AR(1)-GARCH(1,1)-Student's t-distribution model, with Monte Carlo simulation. It is a backward-looking approach, improved by variance reduction techniques employed over the traditional random sampling simulation method. After the estimation of different specifications[10], the one that better adjusted the data was the AR(1)-GARCH(1,1), with "Descriptive Sampling" as the simulation method. It can be represented as below ($h = 1$):

$$\Delta s_t = \alpha + \beta \Delta s_{t-1} + \eta_t \tag{2}$$

$$h_t^2 = \omega + \gamma h_{t-1}^2 + \delta \eta_{t-1}^2, \tag{3}$$

where the dependent variable is the multi-period ahead value being forecasted. Which approach is better is an empirical matter: in theory, iterated forecasts are more efficient if correctly specified, but direct forecasts are more robust to model misspecification.

[9] The main idea is that options are contracts giving the right (not the obligation) to buy or sell an asset at a given point in the future at a price set now (i.e. strike price). Options to buy (call options) are only valuable if there is a chance that when the option comes to be exercised, the underlying asset will be worth more than the strike price. Thus, if one considers options to buy a particular asset at a particular point in the future but at different strike prices, the prices at which such contracts are trade now provides some information about the market's view of the chances that the price of the underlying asset will be above the various strike prices. Therefore, options tell us something about the probability the market attaches to an asset being within a range of possible prices at some future date.

[10] AR(1)-GARCH(1,1) t-Student, random walk with drift and Gaussian white noise, random walk with Gaussian GARCH and random walk with t-Student GARCH. The sampling simulation methods combined with each one of these models were Simple Random Sampling, Simple Random Sampling with runs, Latin Hypercube and Descriptive Sampling.

where s_t is the log of the nominal exchange rate, h_t^2 is the conditional variance and η_t is the input variable of the simulation model assumed Student's t distributed and descriptive sampled instead of randomly sampled.[11]

Model 4 is the survey-based median forecast (from the "Focus" market survey, conducted by the Central Bank of Brazil), with Gaussian distribution based on past forecast errors. **Model 5** is also a forward-looking approach, based on the same previous survey median forecast, but employing a bias correction device, as proposed by Capistrán and Timmermann (2009). **Models 6-14** are economic-driven models following Molodtsova and Papell (2009), and Wang and Wu (2012). See Table 1 for further details.

The density forecast for models 5-14 is first constructed by using quantile regression (QR), as proposed by Gaglianone and Lima (2012).[12] The idea is to use a location-scale model to construct density forecasts from the covariate vector $\mathbf{X}'_{m,t}$, as it follows:

$$s_{t+h} - s_t = \mathbf{X}'_{m,t}\boldsymbol{\alpha}_{m,h} + \left(\mathbf{X}'_{m,t}\boldsymbol{\beta}_{m,h}\right)\zeta_{t+h} \tag{4}$$

where $\left(\zeta_{t+h}|\mathcal{F}_t\right) \sim F_{\zeta,h}\left(0,1\right)$; $F_{\zeta,h}\left(0,1\right)$ is some distribution with mean zero and unit variance, which depends on h but does not depend on the information set $\mathcal{F}_t$. $\mathbf{X}'_{m,t} \in \mathcal{F}_t$ is a $k \times 1$ vector of economic variables used in model m, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $k \times 1$ vectors of parameters, which include the intercepts α_0 and β_0. This class of data-generating process (DGP) is very broad and includes common volatility processes (e.g. ARCH, stochastic volatility).

Based on the previous model and using standard quantile regression techniques (see Koenker, 2005), the conditional quantiles of $(s_{t+h} - s_t)$ based on model m, are given by

$$Q_{m,\tau}(s_{t+h} - s_t \mid \mathcal{F}_t) = \mathbf{X}'_{m,t}\boldsymbol{\theta}_{m,h}(\tau) \tag{5}$$

where for a given quantile level $\tau \in [0;1]$, it follows that $\boldsymbol{\theta}_{m,h}(\tau)$ is a $k \times 1$ vector of parameters of the form $\theta_i(\tau) = \left(\alpha_i(\tau) + \beta_i(\tau)F^{-1}_{\zeta,h}\left(\tau\right)\right); i = 1, ..., k$. Given a family of estimated conditional quantiles $Q_{m,\tau}(\cdot)$, the conditional density of $(s_{t+h} - s_t)$ can

[11] For more details about "Descriptive Sampling" and other sampling methods for variance reduction, see Saliby (1989) and Glasserman (2004).

[12] The authors generate multi-step-ahead conditional density forecasts for the unemployment rate in the U.S. from (point) consensus forecasts and quantile regression; which is a setup that do not impose any parametric structure on the shape of the conditional distributions.

be estimated by using the Epanechnikov kernel, for instance, which is a weighting function that determines the shape of the bumps. To guarantee monotonicity of the conditional quantiles (and the validity of the related conditional distribution), by avoiding possible crossing of quantiles, some rearrangement procedure (e.g., He, 1997; Chernozhukov et al., 2010) could be further employed.[13] See Appendix C for further details on quantile regression.

Table 1 - Models for the Exchange Rate (s_{t+h})

Model	Covariate Vector $X'_{m,t}$	Density
1) Random walk (without drift)	–	Gaussian
2) Option-implied (RND-RWD)	–	Nonparametric
3) GARCH - Monte Carlo	–	Student's t
4) Survey forecast	–	Gaussian
5) Survey forec. (bias-correct)	$[1; s^e_{t+1\|t}]$	QR or Gaussian
6) Taylor rule model	$[1; \pi_t - \pi^*_t; y^{gap}_t - y^{gap*}_t; q_t]$	QR or Gaussian
7) Taylor rule (PPP)	$[1; \pi_t - \pi^*_t; y^{gap}_t - y^{gap*}_t]$	QR or Gaussian
8) Taylor rule (PPP, smoothing)	$[1; \pi_t - \pi^*_t; y^{gap}_t - y^{gap*}_t; i_{t-1} - i^*_{t-1}]$	QR or Gaussian
9) Taylor rule (smoothing)	$[1; \pi_t - \pi^*_t; y^{gap}_t - y^{gap*}_t; q_t; i_{t-1} - i^*_{t-1}]$	QR or Gaussian
10) Absolute PPP model	$[1; q_t]$	QR or Gaussian
11) Relative PPP model	$[1; \Delta q_t]$	QR or Gaussian
12) Monetary model	$[1; s_t - ((m_t - m^*_t) - (y_t - y^*_t))]$	QR or Gaussian
13) Monetary model (weaker)	$[1; \Delta s_t - ((\Delta m_t - \Delta m^*_t) - (\Delta y_t - \Delta y^*_t))]$	QR or Gaussian
14) Forward premium model	$[1; i_t - i^*_t]$	QR or Gaussian

Notes: Covariate vectors shown for h=1. RND means risk-neutral density, RWD (real world density), QR (quantile regression). The $s^e_{t+1|t}$ term refers to the median survey forecast of the FX rate at period t+1 formed at period t, and the real exchange rate is defined as $q_t \equiv s_t + p^*_t - p_t$ in which $p_t(p^*_t)$ is the (log) consumer price index in the home (foreign) country. Models 6-14 are based on Molodtsova and Papell (2009) and Wang and Wu (2012), where $\pi_t(\pi^*_t)$ is the CPI inflation and $y^{gap}_t(y^{gap*}_t)$ is the output gap in the home (foreign) country, $i_t(i^*_t)$ is the short-term interest rate in the home (foreign) country, and $m_t(m^*_t)$ is the money supply and $y_t(y^*_t)$ is the output in the home (foreign) country.

Finally, models 5-14 are alternatively estimated by OLS, with the respective regressors shown in Table 1, coupled with a Gaussian distribution to generate the density forecast. The variance of the distribution is again given by the sample variance of past point forecast errors. This alternative method to quantile regression enables us to distinguish between whether it is the economic variables that affect forecast performance or whether it is how they are modelled.

[13] He (1997) argues that crossing problem occurs more frequently in multiple-variable regressions. Thus, we should not expect crossing to be an issue in our empirical exercise due to the reduced number of covariates in models 5-14.

2.2 Estimation schemes

Rossi (2013b) reviews the empirical evidence on forecasting in the presence of instabilities and concludes that the predictive content of several time series predictors is unstable over time in macroeconomics, finance and international finance. The author also argues that it is possible to exploit instabilities to improve the out-of-sample forecasting ability of existing models, for instance, by using methods that identify historic breaks (and impose them in the estimation) or consider time-varying parameter models. In this sense, several estimation procedures have been proposed in the literature to deal with unstable predictive content over time, such as rolling or recursive estimation schemes, discounted least squares, and exponential smoothing. Which approach should one use? According to Rossi (2013b): *"...rolling estimation is advantageous in the presence of big and recurrent breaks whereas recursive estimation is advantageous when such breaks are small or non-existent."*[14]

In this paper, we consider both recursive and rolling window estimation schemes in order to evaluate the predictive content of the competing point and density forecast models.

2.3 Point and density forecast evaluation

The forecast evaluation is conducted in this paper throughout distinct perspectives. First, we do a standard point forecast evaluation, focused on the forecast performance of the conditional mean. To do so, we compute the root mean squared error (RMSE)[15] and check whether it is possible to beat the random walk forecast for a given forecast horizon, based on the Diebold and Mariano (1995) and West (1996) tests and on the Giacomini and White (2006) predictive ability test. The directional change test of Pesaran and Timmermann (1992, 2009) is also conducted in order to verify whether a given model can correctly predict the directional change of the FX rate. Second, the density forecast evaluation is conducted along two dimensions:

(i) Full-density analysis, which is a shape evaluation based on the entire esti-

[14] However, rolling window estimation schemes could perform worse than recursive ones; even in the presence of breaks. See Pesaran and Timmermann (2005) and Morales-Arias and Moura (2013) for further details on rolling versus recursive estimation.

[15] See Gneiting (2011) for a detailed discussion on forecast evaluation and the choice of scoring function.

mated density. Following the literature on density forecast evaluation, we investigate: coverage rates, the density test of Berkowitz (2001), the density test of Knüppel (2015), the model ranking from the log predictive density scores (LPDS) and the test of Amisano-Giacomini (2007).

(ii) Local analysis, which evaluates specific parts of the densities, that is, the so-called Value-at-Risk (VaR) measures. To do so, we employ available risk management tools for VaR backtesting based on the tests of Kupiec (1995), Christoffersen (1998) and Gaglianone et al. (2011). See Nieto and Ruiz (2016) for a good review on the backtesting literature.[16]

3 Empirical exercise

3.1 Data

The exchange rate we investigate in this paper is the Brazilian Real (R$) in respect to the U.S. dollar (US$), that is, the price of one U.S. dollar in terms of the Brazilian Real, such that an increase of the exchange rate represents a depreciation of the Real currency. Figure 1 presents the behavior of the target variable, that is, the Brazilian FX rate along the considered sample. In the second semester of 2002, the exchange rate experienced a sharp increase due to (among other factors) the augment of investors uncertainty regarding the future of economic fundamentals after the presidential elections in October 2002. The FX rate had gradually appreciated in the following years up to the global crisis in 2008 and showed a depreciation trend after the mid-2011 (European crisis).

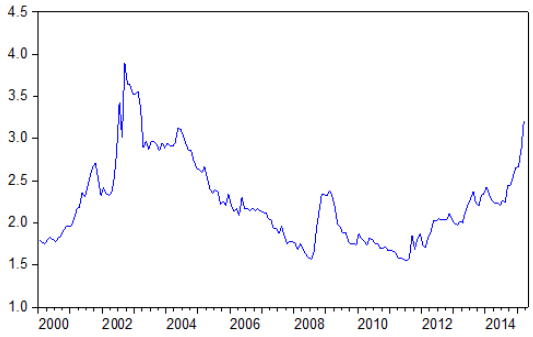
We employ two data frequencies: monthly and daily. All models (in both frequencies) are estimated by using recursive estimation (increasing sample size) as well as rolling window estimation[17] (with a fixed sample of six years = 72 months or 1,512 workdays).[18]

[16] Further details of such evaluation procedures are presented in Appendix B.

[17] Each model is initially estimated using the first 72 monthly (or 1,512 daily) observations and the one-period-ahead (up to the 12-months-ahead or 20-days-ahead) point and density forecasts are generated. We, then, drop the first data point, add an additional observation at the end of the sample, re-estimate the models and generate again out-of-sample forecasts. This process is repeated along the remaining data.

[18] As discussed in Pesaran and Timmermann (2005), the choice of the window size depends on the nature of the possible model instability and the timing of the breaks. A large window is preferable if the data generating

Figure 1 - Exchange rate R$/US$



This way, models are labeled "a" or "b" according to the sample used in estimation ("a" denotes the recursive estimation; "b" means a rolling window estimation). In the case of models 5-14 estimated with QR, we label them as "a" when using recursive estimation or "b" when employing rolling window estimation. The models 5-14 estimated with OLS (and Gaussian distribution) are labeled "c" when using recursive estimation or "d" when employing rolling window estimation.

Monthly frequency

The monthly data ranges from January 2000 through March 2015 (183 observations), covering the most recent period of floating exchange rate regime in Brazil, after the collapse of the fixed FX rate regime in 1999.[19]

For model estimation purposes (training sample), we use data over the period January 2000-December 2005 and reserve the remaining data for (pseudo) out-of-sample forecasting. We construct (point and density) forecasts for horizons $h = 1, ..., 12$ months. This way, the evaluation sample for $h = 1$ is January 2006-March 2015 (111 out-of-sample forecasts), whereas for $h = 12$ we have 100 out-of-sample forecasts.

process is stationary, but comes at the cost of lower power since there are fewer observations in the evaluation window. Similarly, a shorter window may be more robust to structural breaks, although it may not provide as precise estimation as larger windows if the data are stationary.

[19] The monthly (nominal) exchange rate is given by the sale rate (R$/US$) at the end of each month (Sisbacen PTAX800). The FX rate data is obtained from the website of the Central Bank of Brazil. For model 2, we use the BM&F's reference prices for dollar calls. For models 4-5, we employ survey-based (median) expectations from the Focus survey organized by the Central Bank of Brazil, which collects daily information on more than 100 institutions, including commercial banks, asset management firms, and non-financial institutions. For models 6-14, we also use data from the FRED dataset of the Federal Reserve Bank of St. Louis.

Daily frequency

The daily data (workdays) ranges from 3 January 2000 to 31 March 2015 (3,977 observations). The estimation sample ranges from 3 January 2000 to 30 December 2005. The (point and density) forecasts are constructed for horizons $h = 1, ..., 20$ workdays. The out-of-sample evaluation period for $h = 1$ ranges from 2 January 2006 to 31 March 2015 (2,412 out-of-sample forecasts), whereas for $h = 20$ we have 2,393 out-of-sample forecasts.

The following series are observed in daily frequency: exchange rate (sale rate R$/US$), the short-term interest rates Selic and Fed Funds (Brazil and U.S., respectively) and the median survey forecast of the FX rate (for end of month m formed at day d). The remaining variables (CPI inflation, output, output gap, money supply and real exchange rate), however, are only sampled in lower frequencies (weekly or monthly).

In particular, when estimating model 2 (options) with daily data, we proceed in the following way: we daily collect foreign exchange options data from January 2006 to March 2015. If the forecast horizon is lower than the amount of days until the first option maturity, we estimate the RND for that horizon by interpolating the RND estimated for the first maturity with a degenerated probability distribution, which takes as its only value the last occurred exchange rate. If the forecast horizon is above the amount of days until the first option maturity and below the amount of days until the second option maturity, we estimate the RND for that horizon by interpolating the RND estimated for the second maturity with the RND estimated for the first one. Proceeding this way, we have daily estimated RND's for each forecast horizon considered. All the rest remained the same in respect to the estimation using monthly data.

Regarding model 3 (GARCH) and daily data, we proceed in a similar way to the end-of-month data: we also estimate an AR(1)-GARCH(1,1) specification, with Descriptive Sampling as the variance reduction technique to improve the simulation process.

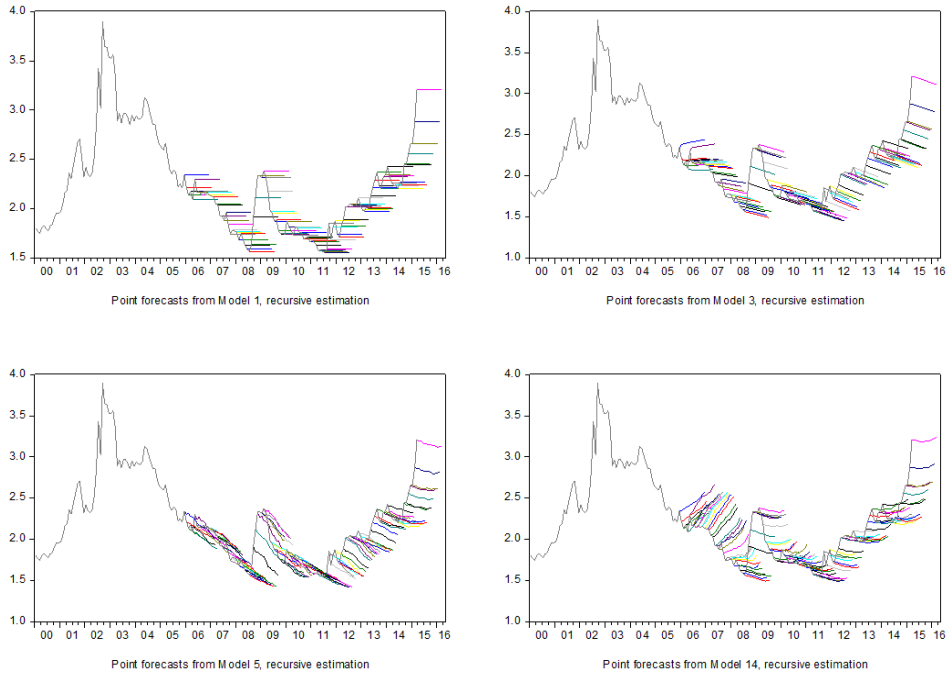
In respect to models 4-14, in order to overcome the lack of daily data regarding macro variables, we consider daily series with available information based on the

latest observed value. In other words, we take into account the actual release date of each variable, which varies across variables and usually changes across time. For example, the value of the CPI inflation series, in a given day d, contains information regarding the last available CPI figure; and such value is repeated over the following days until the release of the next CPI monthly data. This way, by using daily data we are able to enlarge the number of observations by roughly 20 times when compared to monthly data.[20]

3.2 Point forecasts

We start the model evaluation by investigating the performance of the exchange rate point forecast across the investigated models. We considered the conditional mean as the point forecast (in all models, data frequencies and estimation schemes).[21] Figure 2 presents the point forecasts of selected models for $h = 1, ..., 12$ constructed along the out-of-sample exercise (monthly frequency, recursive estimation).

Figure 2 - Point forecasts of selected models



[20] Alternatively, one could employ a more sophisticated method, for instance, by casting the daily model in a state-space approach (e.g. using the Kalman filter to deal with missing observations) or using a reverse MIDAS model, which incorporates low frequency information for predicting high frequency variables (see Foroni et al., 2015).

[21] For instance, in the case of model 6 estimated with quantile regression using recursive estimation (model 6a), we compute the conditional mean (see Koenker, 2005) as the average of the conditional quantiles estimated at the grid of quantile levels $\tau = \{0.25, 0.30, 0.35, ..., 0.75\}$. We proceed in the same way for all models.

Is it possible to beat the random walk? To tackle this question we first employ the Diebold-Mariano-West (DMW) test of equal accuracy in a recursive estimation scheme (Tables 2a-2b). The null hypothesis assumes equal RMSEs of two competing models.[22] Positive test statistics indicate that model $m \neq 1$ has a lower RMSE compared to the benchmark model (random walk).[23] We also investigate the DMW test modified by Harvey et al. (1997), which propose a hypothesis test more suitable to small samples. The results are similar (not presented here to save space). Regarding the rolling window estimation, we employ the Giacomini and White (2006) predictive ability test.

The monthly results indicate that the only model that exhibits a positive DMW test statistic is model 4 (for h between 5 and 12 months). In other words, only the model that embodies survey-based expectations (model 4) is able to present a lower RMSE compared to the random walk (gray cells on Table 2a).

Table 2a - Tests of equal forecast accuracy (monthly frequency)

Diebold-Mariano-West (1995, 1996): test statistic (p-value)

Model	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12a	13a	14a
h=1	-2.21	-1.74	-1.42	-1.32	-1.22	-0.85	-0.99	-1.70	-2.41	-1.46	-1.94	-1.49	-2.74
	(0.03)	(0.09)	(0.16)	(0.19)	(0.23)	(0.4)	(0.32)	(0.09)	(0.02)	(0.15)	(0.06)	(0.14)	(0.01)
h=2	-2.59	-1.66	-1.18	-1.36	-1.14	-0.65	-0.76	-1.54	-1.78	-0.30	-1.62	-1.70	-2.37
	(0.01)	(0.1)	(0.24)	(0.18)	(0.26)	(0.51)	(0.45)	(0.13)	(0.08)	(0.77)	(0.11)	(0.09)	(0.02)
h=3	-1.84	-1.79	-0.80	-1.50	-1.26	-0.83	-1.00	-1.42	-2.00	-1.24	-1.85	-1.51	-2.16
	(0.07)	(0.08)	(0.43)	(0.14)	(0.21)	(0.41)	(0.32)	(0.16)	(0.05)	(0.22)	(0.07)	(0.14)	(0.03)
h=4		-1.72	-0.26	-1.56	-1.30	-0.75	-0.87	-1.43	-2.13	-1.02	-2.23	-1.39	-1.94
		(0.09)	(0.8)	(0.12)	(0.2)	(0.45)	(0.39)	(0.16)	(0.04)	(0.31)	(0.03)	(0.17)	(0.06)
h=5		-1.60	0.40	-1.44	-1.12	-0.64	-0.76	-1.23	-2.00	-1.08	-2.46	-1.42	-1.67
		(0.11)	(0.69)	(0.15)	(0.27)	(0.52)	(0.45)	(0.22)	(0.05)	(0.28)	(0.02)	(0.16)	(0.1)
h=6		-1.53	0.82	-1.21	-1.11	-0.71	-0.82	-1.19	-1.86	-0.85	-2.53	-0.82	-1.52
		(0.13)	(0.41)	(0.23)	(0.27)	(0.48)	(0.41)	(0.24)	(0.07)	(0.4)	(0.01)	(0.41)	(0.13)
h=7		-1.52	1.00	-1.14	-1.33	-0.81	-0.97	-1.43	-1.76	-1.39	-2.24	-1.43	-1.52
		(0.13)	(0.32)	(0.26)	(0.19)	(0.42)	(0.33)	(0.16)	(0.08)	(0.17)	(0.03)	(0.16)	(0.13)
h=8		-1.52	0.99	-1.08	-1.24	-0.91	-1.15	-1.61	-1.53	-1.30	-1.88	-1.19	-1.49
		(0.13)	(0.32)	(0.28)	(0.22)	(0.36)	(0.25)	(0.11)	(0.13)	(0.2)	(0.06)	(0.24)	(0.14)
h=9		-1.54	1.20	-1.14	-1.34	-1.03	-1.29	-1.91	-1.46	-1.21	-1.63	-1.17	-1.47
		(0.13)	(0.23)	(0.26)	(0.18)	(0.3)	(0.2)	(0.06)	(0.15)	(0.23)	(0.11)	(0.24)	(0.15)
h=10		-1.53	0.93	-1.10	-1.42	-1.01	-1.23	-1.71	-1.38	-1.28	-1.45	-1.09	-1.39
		(0.13)	(0.35)	(0.28)	(0.16)	(0.31)	(0.22)	(0.09)	(0.17)	(0.2)	(0.15)	(0.28)	(0.17)
h=11		-1.52	0.74	-1.10	-1.51	-1.10	-1.27	-1.70	-1.34	-1.54	-1.35	-1.43	-1.38
		(0.13)	(0.46)	(0.27)	(0.13)	(0.27)	(0.21)	(0.09)	(0.18)	(0.13)	(0.18)	(0.15)	(0.17)
h=12		-1.58	0.53	-1.13	-1.69	-1.08	-1.20	-1.79	-1.31	-1.70	-1.32	-2.03	-1.32
		(0.12)	(0.6)	(0.26)	(0.09)	(0.28)	(0.23)	(0.08)	(0.19)	(0.09)	(0.19)	(0.05)	(0.19)

Notes: Recursive estimation. Table shows the test statistics (and p-values in parentheses). Gray cells denote positive test statistics (i.e. it means that a model has a lower RMSE in comparison to the RW).

[22] Tables in the Supplementary Appendix (not presented here to save space, but available upon request) show the Root Mean Squared Error (RMSE) in all cases.

[23] The variances entering the test statistics use the Newey-West estimator, with a bandwidth of 0 at the 1-month horizon and 1.5×horizon in the other cases, following Clark (2011, supplementary appendix) and Clark and McCracken (2012, p.61). This approach comes from past Monte Carlo assessments of the small-sample properties of the DM test from the referred authors. Nonetheless, it is worth mentioning that our results for the DM test are robust to different ways of computing the variance employed in the test statistic (e.g. using a rectangular kernel estimator of Hansen (1982), with lag length of $h-1$). The rounding off to an integer value is done upwards.

Table 2b - Tests of equal forecast accuracy (daily frequency)

Diebold-Mariano-West (1995, 1996): test statistic (p-value)

Model	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12a	13a	14a
h=1	-4.11	-0.95	-4.56	-2.61	-1.15	-1.00	-0.89	-1.45	-1.70	-0.88	-1.36	-1.02	-2.08
	(0)	(0.34)	(0)	(0.01)	(0.25)	(0.32)	(0.37)	(0.15)	(0.09)	(0.38)	(0.17)	(0.31)	(0.04)
h=2	-15.61	-0.74	-4.31	-2.12	-1.41	-1.31	-1.24	-1.69	-1.85	-0.68	-1.79	-1.03	-2.18
	(0)	(0.46)	(0)	(0.03)	(0.16)	(0.19)	(0.22)	(0.09)	(0.07)	(0.5)	(0.07)	(0.3)	(0.03)
h=3	-20.96	-1.62	-4.02	-1.45	-1.30	-1.27	-1.26	-1.52	-1.63	-1.58	-1.72	-1.47	-2.15
	(0)	(0.11)	(0)	(0.15)	(0.19)	(0.2)	(0.21)	(0.13)	(0.1)	(0.11)	(0.09)	(0.14)	(0.03)
h=4	-13.17	-1.77	-4.18	-1.31	-1.34	-1.26	-1.23	-1.53	-1.50	-1.91	-1.67	-1.71	-2.16
	(0)	(0.08)	(0)	(0.19)	(0.18)	(0.21)	(0.22)	(0.13)	(0.13)	(0.06)	(0.1)	(0.09)	(0.03)
h=5	-14.43	-2.21	-3.67	-1.37	-1.28	-1.18	-1.13	-1.50	-1.59	-2.19	-1.70	-2.20	-2.42
	(0)	(0.03)	(0)	(0.17)	(0.2)	(0.24)	(0.26)	(0.13)	(0.11)	(0.03)	(0.09)	(0.03)	(0.02)
h=6	-13.00	-2.67	-3.33	-1.26	-1.28	-1.13	-1.11	-1.48	-1.66	-2.86	-1.67	-2.99	-2.63
	(0)	(0.01)	(0)	(0.21)	(0.2)	(0.26)	(0.27)	(0.14)	(0.1)	(0)	(0.1)	(0)	(0.01)
h=7	-16.80	-2.36	-3.36	-1.12	-1.23	-1.10	-1.07	-1.43	-1.73	-2.25	-1.73	-2.35	-2.77
	(0)	(0.02)	(0)	(0.26)	(0.22)	(0.27)	(0.29)	(0.15)	(0.08)	(0.02)	(0.08)	(0.02)	(0.01)
h=8	-13.54	-2.15	-2.87	-1.01	-1.21	-1.09	-1.04	-1.41	-1.73	-1.88	-1.72	-2.18	-2.76
	(0)	(0.03)	(0)	(0.31)	(0.23)	(0.28)	(0.3)	(0.16)	(0.08)	(0.06)	(0.09)	(0.03)	(0.01)
h=9	-13.30	-2.04	-2.46	-0.76	-1.23	-1.09	-1.06	-1.44	-1.65	-1.79	-1.66	-2.11	-2.78
	(0)	(0.04)	(0.01)	(0.45)	(0.22)	(0.28)	(0.29)	(0.15)	(0.1)	(0.07)	(0.1)	(0.04)	(0.01)
h=10	-12.93	-2.03	-2.11	-0.51	-1.25	-1.10	-1.09	-1.46	-1.58	-1.58	-1.59	-1.89	-2.81
	(0)	(0.04)	(0.03)	(0.61)	(0.21)	(0.27)	(0.28)	(0.14)	(0.11)	(0.11)	(0.11)	(0.06)	(0.01)
h=11	-12.68	-1.97	-1.70	-0.35	-1.27	-1.11	-1.10	-1.49	-1.56	-1.80	-1.59	-2.16	-2.78
	(0)	(0.05)	(0.09)	(0.73)	(0.2)	(0.27)	(0.27)	(0.14)	(0.12)	(0.07)	(0.11)	(0.03)	(0.01)
h=12	-13.16	-1.93	-1.42	-0.18	-1.25	-1.11	-1.09	-1.46	-1.58	-1.89	-1.58	-2.43	-2.84
	(0)	(0.05)	(0.16)	(0.86)	(0.21)	(0.27)	(0.28)	(0.14)	(0.11)	(0.06)	(0.11)	(0.02)	(0)
h=13	-13.51	-1.85	-1.12	0.07	-1.25	-1.10	-1.07	-1.46	-1.59	-2.09	-1.58	-2.72	-2.87
	(0)	(0.06)	(0.26)	(0.95)	(0.21)	(0.27)	(0.29)	(0.14)	(0.11)	(0.04)	(0.12)	(0.01)	(0)
h=14	-12.94	-1.74	-0.82	0.34	-1.21	-1.07	-1.05	-1.42	-1.58	-2.14	-1.55	-2.77	-2.90
	(0)	(0.08)	(0.41)	(0.74)	(0.22)	(0.29)	(0.29)	(0.16)	(0.11)	(0.03)	(0.12)	(0.01)	(0)
h=15	-13.85	-1.71	-0.47	0.61	-1.18	-1.03	-1.02	-1.39	-1.58	-2.11	-1.54	-2.65	-2.93
	(0)	(0.09)	(0.64)	(0.55)	(0.24)	(0.3)	(0.31)	(0.16)	(0.11)	(0.04)	(0.12)	(0.01)	(0)
h=16	-11.95	-1.64	-0.14	0.81	-1.16	-1.01	-0.99	-1.36	-1.54	-2.40	-1.49	-2.92	-2.95
	(0)	(0.1)	(0.89)	(0.42)	(0.25)	(0.31)	(0.32)	(0.17)	(0.12)	(0.02)	(0.14)	(0)	(0)
h=17	-10.82	-1.61	0.19	1.05	-1.17	-1.01	-0.99	-1.39	-1.49	-2.37	-1.47	-2.65	-2.92
	(0)	(0.11)	(0.85)	(0.29)	(0.24)	(0.31)	(0.32)	(0.16)	(0.14)	(0.02)	(0.14)	(0.01)	(0)
h=18	-13.40	-1.67	0.47	1.23	-1.16	-0.98	-0.97	-1.38	-1.49	-2.46	-1.47	-2.87	-2.89
	(0)	(0.1)	(0.64)	(0.22)	(0.25)	(0.33)	(0.33)	(0.17)	(0.14)	(0.01)	(0.14)	(0)	(0)
h=19	-12.13	-1.62	0.72	1.47	-1.16	-0.98	-0.97	-1.39	-1.48	-2.25	-1.48	-2.67	-2.87
	(0)	(0.1)	(0.47)	(0.14)	(0.25)	(0.33)	(0.33)	(0.16)	(0.14)	(0.02)	(0.14)	(0.01)	(0)
h=20	-18.56	-1.57	0.93	1.65	-1.13	-0.94	-0.94	-1.36	-1.45	-2.06	-1.48	-2.54	-2.84
	(0)	(0.12)	(0.35)	(0.1)	(0.26)	(0.35)	(0.35)	(0.17)	(0.15)	(0.04)	(0.14)	(0.01)	(0)

Notes: Recursive estimation. Table shows the test statistics (and p-values in parentheses). Gray cells denote positive test statistics (i.e. it means that a model has a lower RMSE in comparison to the RW).

The results for a rolling window scheme (in the Supplementary Appendix) indicate that models 4, 6, 7 and 9 (for some horizons) show lower RMSE compared to the random walk, although these mentioned "gains" over the RW are not statistically significant (at the usual 5% level of significance).[24]

In respect to daily frequency, Table 2b reveals that, once again, only a few models (and for longer horizons) exhibit lower RMSEs in comparison to the RW (i.e. model 4 for $h > 16$ days and model 5 for $h > 12$ days); although these predictive gains are not statistically significant (at a 5% level). Nonetheless, model 5c (estimated with OLS, and coupled with a Gaussian density) is able to statistically beat the random walk (at a 5% significance level) for horizons $h = 15, ..., 20$ days (see Table 2.1.4 in

[24] The outcome is similar for models 5-14 estimated with OLS: models 5, 7, 11 and 13 are able to show lower RMSEs compared to the RW, in some horizons, for rolling window estimation (but these gains are not statistically significant at a 5% level). See Table 1.1.3 in Supplementary Appendix.

the Supplementary Appendix).

These results are in line with a vast literature reporting the practical difficulty on beating the naive random walk forecast in out-of-sample exercises (Mark, 1995).[25]

Now, we investigate a different empirical question: Can the competing models forecast the direction of change for the FX rate? The test of Pesaran and Timmermann (1992, 2009) is designed to answer this question. The results are presented on Table 3.[26]

Table 3 - Test of direction of change

Monthly frequency

Pesaran & Timmermann (1992, 2009) direction of change test (p-value)

Model	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12a	13a	14a
h=1	0.02	0.02	0.80	0.72	0.34	0.43	0.19	0.32	0.04	0.68	0.29	0.23	0.04
h=2	0.00	0.12	0.85	0.00	0.29	0.51	0.03	0.29	0.02	0.19	0.04	0.01	0.01
h=3	0.00	0.08	0.46	0.00	0.09	0.43	0.07	0.02	0.08	0.20	0.06	0.90	0.01
h=4		0.07	0.87	0.00	0.01	0.61	0.08	0.00	0.03	0.52	0.11	0.74	0.00
h=5		0.01	0.79	0.00	0.10	0.98	0.28	0.03	0.12	0.25	0.02	0.73	0.00
h=6		0.05	0.75	0.00	0.12	0.75	0.55	0.04	0.16	0.53	0.00	0.78	0.00
h=7		0.00	0.82	0.00	0.15	0.31	0.90	0.06	0.30	0.01	0.00	0.05	0.00
h=8		0.00	0.82	0.00	0.60	0.05	0.36	0.44	0.00	0.22	0.00	0.82	0.00
h=9		0.00	0.88	0.00	0.38	0.14	0.63	0.50	0.00	0.04	0.00	0.15	0.00
h=10		0.00	0.58	0.00	0.10	0.24	0.82	0.32	0.00	0.17	0.00	0.50	0.08
h=11		0.00	0.88	0.00	0.06	0.38	0.91	0.14	0.00	0.02	0.00	0.01	0.01
h=12		0.00	0.73	0.00	0.03	0.55	0.55	0.21	0.00	0.00	0.00	0.00	0.00

Daily frequency

Pesaran & Timmermann (1992, 2009) direction of change test (p-value)

Model	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12a	13a	14a
h=1	0.00	0.81	0.05	0.00	0.00	0.03	0.03	0.00	0.00	0.57	0.00	0.76	0.00
h=2	0.04	0.03	0.82	0.01	0.81	0.65	0.90	0.56	0.00	0.12	0.63	0.02	0.03
h=3	0.68	0.24	0.58	0.00	0.62	0.98	0.97	0.76	0.02	0.05	0.03	0.13	0.00
h=4	0.85	0.73	0.93	0.00	0.42	0.49	0.67	0.94	0.03	0.57	0.09	0.75	0.00
h=5	0.05	0.76	0.86	0.00	0.40	0.55	0.78	0.78	0.09	0.98	0.09	0.48	0.01
h=6	0.42	0.94	0.48	0.00	0.15	0.89	0.70	0.26	0.03	0.28	0.04	0.38	0.01
h=7	0.36	0.91	0.10	0.00	0.05	0.68	0.63	0.08	0.05	0.20	0.03	0.44	0.01
h=8	0.42	0.95	0.01	0.10	0.09	0.83	0.57	0.23	0.01	0.63	0.01	0.12	0.00
h=9	0.04	0.53	0.01	0.07	0.25	0.95	0.96	0.36	0.02	0.09	0.00	0.01	0.00
h=10	0.97	0.46	0.00	0.02	0.19	0.96	0.88	0.14	0.02	0.15	0.01	0.01	0.00
h=11	0.01	0.57	0.00	0.01	0.24	0.91	0.87	0.12	0.02	0.05	0.01	0.35	0.00
h=12	0.54	0.26	0.00	0.02	0.15	0.83	0.93	0.09	0.00	0.01	0.00	0.39	0.00
h=13	0.64	0.32	0.00	0.01	0.10	0.92	0.83	0.11	0.00	0.19	0.00	0.60	0.00
h=14	0.49	0.24	0.00	0.00	0.09	0.76	0.95	0.12	0.02	0.37	0.01	0.53	0.00
h=15	0.11	0.27	0.00	0.00	0.07	0.99	0.71	0.12	0.03	0.24	0.04	0.36	0.00
h=16	0.05	0.29	0.00	0.00	0.08	0.99	0.53	0.05	0.04	0.19	0.05	0.50	0.00
h=17	0.38	0.18	0.00	0.00	0.05	0.90	0.63	0.07	0.04	0.09	0.04	0.31	0.00
h=18	0.29	0.19	0.00	0.00	0.04	0.91	0.62	0.09	0.06	0.08	0.05	0.20	0.00
h=19	0.02	0.20	0.00	0.00	0.04	0.86	0.65	0.12	0.08	0.30	0.07	0.23	0.00
h=20	0.00	0.22	0.00	0.00	0.06	0.60	0.91	0.12	0.07	0.30	0.12	0.15	0.00

Note: Recursive estimation. The null hypothesis assumes that the model has no power in predicting the directional change of the FX rate. Table shows the p-values (blue cells indicate rejection of the null at a 5% level).

[25] Notice that given the very high inflation differential (between Brazil and the U.S.) a random walk with drift could possibly be even harder to beat.

[26] The directional forecasts are based on the conditional mean of the density forecast (in all models, data frequencies and estimation schemes).

In fact, many models are able to predict the correct sign of the FX rate in the future (see the blue cells on Table 3). In particular, model 2 (financial data), model 3 (GARCH), model 5 (the bias-corrected survey forecast) and several models based on macro fundamentals (in many horizons) can anticipate the monthly FX rate directional movement (i.e., increase, decrease or no-change), which is naturally easier to forecast compared to the magnitude of the exchange rate change h-periods ahead. This finding also holds for daily frequency, in which several models can correctly predict the exchange rate directional change (e.g., models 2 and 5-14, excepting models 11 and 13, for $h = 1$ day).

3.3 Density forecasts: Full-density analysis

Density forecast evaluation has become popular in the fields of time series forecasting and risk evaluation (Ko and Park, 2013) and related formal testing procedures have been developed by several studies.[27] Here, we estimate all the 14 models, in the 2 estimation schemes, both frequencies, and all considered forecast horizons, for a grid of 99 quantile levels τ, in which $\tau = \{0.01, 0.02, ..., 0.99\}$. We start the full-density evaluation by presenting in Figure 3, for illustrative purposes, the estimated conditional Probability Density Functions (PDFs) of the R$/US$ exchange rate at December 2014, constructed with different forecast horizons and monthly frequency.

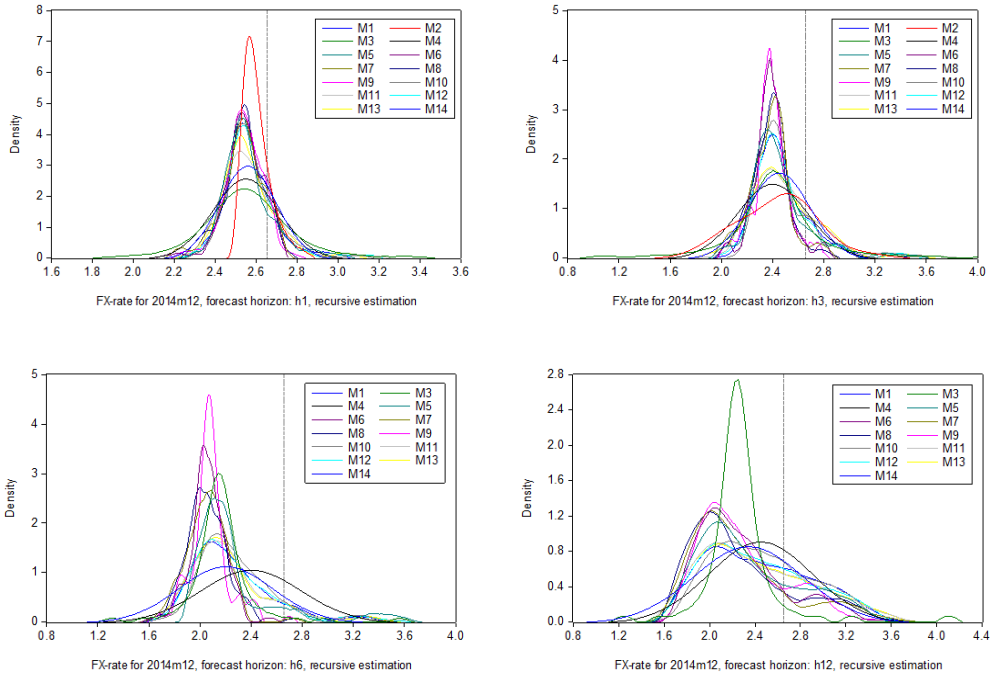
Note from Figure 3 that the variance monotonically increases as long as the forecast horizon augments (i.e. the average variance across the 14 models moves from 0.12 to 0.23, 0.27 and 0.41 for h =1, 3, 6, and 12, respectively). Moreover, note that the PDFs of models 1, 3 and 4 are symmetric (i.e. models 1 and 4 use a Gaussian distribution and model 3 employs a Student's t distribution) whereas the remaining PDFs clearly exhibit asymmetry and kurtosis (which is due to the nonparametric technique for extracting RND from option prices in model 2 and the quantile regression technique in models 5-14). On average, the 14 models exhibit a positive skewness (indicating that the conditional mean is not equal to the conditional median) and kurtosis above 3 (suggesting leptokurtic distributions, with fatter tails).[28]

[27] A good review of various testing methods in density forecasting is provided by Corradi and Swanson (2006).

[28] The average skewness (across the 14 models) is 0.46, 0.89, 0.97 and 0.73 for h =1, 3, 6, and 12, respectively. The average kurtosis (across the 14 models) is 4.66, 5.80, 5.66 and 3.69 for h =1, 3, 6, and 12, respectively.

These empirical findings can be related to the exchange rate dynamics observed in Brazil in the 2000-2015 period (see Figure 1), with several (and long) periods of gradual FX rate appreciation and a few (and short) periods of sharp currency depreciation.

Figure 3 - Conditional PDFs of R$/US$ at December 2014



Note: Recursive estimation, monthly frequency, forecast horizons h=1, 3, 6 and 12 months.

Vertical line denotes the actual FX rate at December 2014 (R$2.66/US$).

Regarding coverage rates for the 70% interval band (in the Supplementary Appendix), besides the relatively good result for several models, in many cases (in monthly and daily frequencies), the bias-corrected survey density forecast (model 5) is the only model not rejected at a 5% confidence level, in all horizons, both sampling schemes, and both density estimations (QR or OLS). On the other hand, the rolling window estimation scheme, in general, yields slightly more accurate interval forecasts (i.e., coverage rates closer to the 70% nominal rate) compared to the recursive estimation, in line with the previous findings of Clark (2011, p.336).[29] As a

[29] The referred author also argues that: "For a given model, differences in coverage across horizons likely reflect a variety of forces, making a single explanation difficult. One force is sampling error. Even if a model were correctly specified, random variation in a given data sample could cause the empirical coverage rate to differ from the nominal. Sampling error increases with the forecast horizon, due to the overlap of forecast errors for multistep

robustness check, we also perform the 50% and 90% interval bands, which (overall) point to similar conclusions.

In respect to the Berkowitz (2001) test, Table 4 suggests that, for $h = 1$, monthly frequency, all economic-driven approaches (models 6-14), excepting model 9, present an adequate density forecast at the usual 5% significance level. For medium-term horizons ($h = 2, 3$) there are some models not rejected by the Berkowitz test (e.g. model 10, in both horizons). For longer horizons ($h > 3$), with very few exceptions (e.g. models 6 and 9, in some horizons), there is no predominant model to properly forecast the FX rate density. A similar result is obtained for rolling window estimation, where economic-driven models, overall, are not rejected by the Berkowitz test at short horizons ($h = 1$ to 3 months). Regarding the alternative models 5-14 estimated with OLS, it seems that the Gaussian density, coupled with OLS estimation, does not provide a good density forecast based on the Berkowitz test. On the other hand, the results for the daily frequency indicate p-values below 0.01 in all considered cases (and, thus, are not reported).

Table 4 - Berkowitz (2001) density test

Berkowitz test (p-value)

Model	1a	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12a	13a	14a
h=1	0.00	0.00	0.00	0.00	0.03	0.05	0.90	0.06	0.02	0.48	0.65	0.52	0.32	0.18
h=2	0.00	0.00	0.00	0.00	0.04	0.00	0.92	0.00	0.00	0.62	0.05	0.43	0.01	0.48
h=3	0.00	0.00	0.00	0.00	0.00	0.55	0.01	0.00	0.34	0.14	0.00	0.04	0.00	0.01
h=4	0.00		0.00	0.00	0.00	0.06	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00
h=5	0.00		0.00	0.00	0.00	0.10	0.01	0.13	0.09	0.00	0.00	0.00	0.00	0.00
h=6	0.00		0.00	0.00	0.00	0.00	0.02	0.00	0.82	0.00	0.00	0.00	0.00	0.00
h=7	0.00		0.00	0.00	0.00	0.00	0.22	0.14	0.19	0.00	0.00	0.00	0.00	0.00
h=8	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00
h=9	0.00		0.00	0.00	0.00	0.67	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00
h=10	0.00		0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
h=11	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00
h=12	0.00		0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Notes: Table shows p-values from recursive estimation with monthly frequency. The p-values for recursive estimation with daily frequency are not reported since they are all below the 1% significance level. Gray cell indicates p-value>0.05.

horizons (effectively reducing the number of independent observations relative to the one-step horizon). Of course, an increased sampling error across horizons will translate into reduced power to detect departures from accurate coverage."

Table 5 - Knüppel (2015) density test

Monthly frequency

Knüppel (2015): test statistic (p-value)

Model	1a	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12a	13a	14a
h=1	**19.02**	**21.74**	**20.29**	**13.56**	4.98	**11.43**	3.04	**10.13**	**13.34**	4.80	1.11	5.49	1.87	5.49
	(0)	(0)	(0)	(0.01)	(0.29)	(0.02)	(0.55)	(0.04)	(0.01)	(0.31)	(0.89)	(0.24)	(0.76)	(0.24)
h=2	**16.78**	**17.01**	8.93	**12.85**	2.95	**9.62**	4.68	**10.36**	**10.35**	4.27	2.00	3.73	1.38	4.54
	(0)	(0)	(0.06)	(0.01)	(0.57)	(0.05)	(0.32)	(0.03)	(0.03)	(0.37)	(0.74)	(0.44)	(0.85)	(0.34)
h=3	**14.38**	**13.76**	5.97	**10.85**	3.59	5.05	2.41	5.72	4.93	5.62	2.67	7.03	3.77	5.83
	(0.01)	(0.01)	(0.2)	(0.03)	(0.47)	(0.28)	(0.66)	(0.22)	(0.29)	(0.23)	(0.62)	(0.13)	(0.44)	(0.21)
h=4	**12.53**		5.00	5.29	2.48	7.77	5.88	7.40	6.14	2.22	2.40	3.61	2.47	5.14
	(0.01)		(0.29)	(0.26)	(0.65)	(0.1)	(0.21)	(0.12)	(0.19)	(0.69)	(0.66)	(0.46)	(0.65)	(0.27)
h=5	6.25		6.98	4.48	2.55	8.38	5.39	7.78	7.39	3.38	1.56	4.91	1.59	4.35
	(0.18)		(0.14)	(0.35)	(0.64)	(0.08)	(0.25)	(0.1)	(0.12)	(0.5)	(0.82)	(0.3)	(0.81)	(0.36)
h=6	5.27		7.46	4.46	2.87	6.84	3.75	5.98	6.38	1.62	2.22	3.39	2.13	4.03
	(0.26)		(0.11)	(0.35)	(0.58)	(0.14)	(0.44)	(0.2)	(0.17)	(0.81)	(0.7)	(0.5)	(0.71)	(0.4)
h=7	5.59		7.89	3.95	3.00	6.80	4.14	7.06	4.81	2.51	1.97	3.84	1.87	4.54
	(0.23)		(0.1)	(0.41)	(0.56)	(0.15)	(0.39)	(0.13)	(0.31)	(0.64)	(0.74)	(0.43)	(0.76)	(0.34)
h=8	5.47		7.71	4.20	2.71	7.70	3.59	5.37	5.58	2.45	1.37	3.93	1.35	4.96
	(0.24)		(0.1)	(0.38)	(0.61)	(0.1)	(0.46)	(0.25)	(0.23)	(0.65)	(0.85)	(0.42)	(0.85)	(0.29)
h=9	6.12		6.47	4.95	2.95	6.24	6.40	4.97	4.51	2.89	1.87	3.62	2.15	5.16
	(0.19)		(0.17)	(0.29)	(0.57)	(0.18)	(0.17)	(0.29)	(0.34)	(0.58)	(0.76)	(0.46)	(0.71)	(0.27)
h=10	3.89		6.10	4.18	3.51	4.59	5.94	4.74	3.81	3.18	2.24	3.57	1.89	4.78
	(0.42)		(0.19)	(0.38)	(0.48)	(0.33)	(0.2)	(0.32)	(0.43)	(0.53)	(0.69)	(0.47)	(0.76)	(0.31)
h=11	2.81		6.09	4.83	4.36	3.99	3.40	4.12	4.74	3.61	3.23	4.04	2.76	3.36
	(0.59)		(0.19)	(0.31)	(0.36)	(0.41)	(0.49)	(0.39)	(0.32)	(0.46)	(0.52)	(0.4)	(0.6)	(0.5)
h=12	3.88		5.91	5.60	4.44	5.33	3.70	3.92	3.66	4.49	4.05	5.11	4.51	2.98
	(0.42)		(0.21)	(0.23)	(0.35)	(0.25)	(0.45)	(0.42)	(0.45)	(0.34)	(0.4)	(0.28)	(0.34)	(0.56)

Daily frequency

Knüppel (2015): test statistic (p-value)

Model	1a	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12a	13a	14a
h=1	**114.73**	8.23	**163.73**	**19.68**	**53.95**	**10.45**	9.03	4.84	**12.30**	**35.54**	7.95	**14.93**	**10.26**	5.74
	(0)	(0.08)	(0)	(0)	(0)	(0.03)	(0.06)	(0.3)	(0.02)	(0)	(0.09)	(0)	(0.04)	(0.22)
h=2	**93.82**	8.39	**108.30**	**19.68**	**46.06**	**15.69**	5.48	4.79	**28.11**	**31.39**	5.21	**20.04**	6.77	**12.79**
	(0)	(0.08)	(0)	(0)	(0)	(0)	(0.24)	(0.31)	(0)	(0)	(0.27)	(0)	(0.15)	(0.01)
h=3	**80.76**	**9.89**	**49.07**	**19.67**	**38.75**	**13.43**	2.62	8.56	**22.98**	**27.47**	3.83	**17.14**	4.52	**14.99**
	(0)	(0.04)	(0)	(0)	(0)	(0.01)	(0.62)	(0.07)	(0)	(0)	(0.43)	(0)	(0.34)	(0)
h=4	**81.87**	**13.09**	**42.23**	**19.65**	**37.09**	**11.91**	0.69	8.03	**19.68**	**20.17**	4.79	**10.25**	5.74	**12.56**
	(0)	(0.01)	(0)	(0)	(0)	(0.02)	(0.95)	(0.09)	(0)	(0)	(0.31)	(0.04)	(0.22)	(0.01)
h=5	**76.88**	**21.03**	**40.63**	**19.64**	**28.60**	8.70	0.48	8.08	**13.01**	**15.05**	2.89	7.07	3.40	9.50
	(0)	(0)	(0)	(0)	(0)	(0.07)	(0.98)	(0.09)	(0.01)	(0)	(0.58)	(0.13)	(0.49)	(0.05)
h=6	**75.08**	**26.62**	**37.96**	**19.62**	**23.67**	8.40	0.94	7.10	**11.36**	**14.16**	3.79	6.68	4.60	8.20
	(0)	(0)	(0)	(0)	(0)	(0.08)	(0.92)	(0.13)	(0.02)	(0.01)	(0.43)	(0.15)	(0.33)	(0.08)
h=7	**75.54**	7.47	**35.77**	**19.59**	**20.68**	8.95	1.49	7.71	**12.20**	**13.18**	5.64	6.27	6.43	7.60
	(0)	(0.11)	(0)	(0)	(0)	(0.06)	(0.83)	(0.1)	(0.02)	(0.01)	(0.23)	(0.18)	(0.17)	(0.11)
h=8	**70.96**	**12.17**	**31.71**	**19.56**	**20.57**	7.91	1.54	7.64	**10.80**	**10.48**	6.51	5.08	7.35	7.96
	(0)	(0.02)	(0)	(0)	(0)	(0.1)	(0.82)	(0.11)	(0.03)	(0.03)	(0.16)	(0.28)	(0.12)	(0.09)
h=9	**69.36**	**17.18**	**29.37**	**19.52**	**20.05**	7.69	1.04	7.98	**9.69**	8.71	6.63	5.64	8.17	8.56
	(0)	(0)	(0)	(0)	(0)	(0.1)	(0.9)	(0.09)	(0.05)	(0.07)	(0.16)	(0.23)	(0.09)	(0.07)
h=10	**63.31**	**16.98**	**21.86**	**19.51**	**20.88**	7.29	0.65	8.68	**10.15**	9.21	6.03	6.92	7.47	8.90
	(0)	(0)	(0)	(0)	(0)	(0.12)	(0.96)	(0.07)	(0.04)	(0.06)	(0.2)	(0.14)	(0.11)	(0.06)
h=11	**60.97**	**26.61**	**20.21**	**19.49**	**22.79**	8.12	0.49	8.93	**11.06**	9.26	5.61	7.44	6.81	8.18
	(0)	(0)	(0)	(0)	(0)	(0.09)	(0.97)	(0.06)	(0.03)	(0.05)	(0.23)	(0.11)	(0.15)	(0.09)
h=12	**57.19**	**29.93**	**18.08**	**19.58**	**22.26**	8.54	0.43	**9.35**	**11.03**	8.94	5.29	7.17	6.47	8.22
	(0)	(0)	(0)	(0)	(0)	(0.07)	(0.98)	(0.05)	(0.03)	(0.06)	(0.26)	(0.13)	(0.17)	(0.08)
h=13	**53.82**	**14.62**	**16.08**	**19.56**	**25.00**	9.61	0.67	9.08	**12.98**	8.93	5.81	6.86	7.11	7.58
	(0)	(0.01)	(0)	(0)	(0)	(0.05)	(0.96)	(0.06)	(0.01)	(0.06)	(0.21)	(0.14)	(0.13)	(0.11)
h=14	**53.07**	**14.03**	**15.89**	**19.56**	**26.29**	**10.86**	0.92	**9.51**	**13.87**	9.66	4.54	7.07	5.58	7.55
	(0)	(0.01)	(0)	(0)	(0)	(0.03)	(0.92)	(0.05)	(0.01)	(0.05)	(0.34)	(0.13)	(0.23)	(0.11)
h=15	**50.96**	**9.94**	**13.95**	**19.55**	**25.49**	**10.05**	0.52	**9.31**	**13.29**	9.40	4.39	7.28	5.32	6.87
	(0)	(0.04)	(0.01)	(0)	(0)	(0.04)	(0.97)	(0.05)	(0.01)	(0.05)	(0.36)	(0.12)	(0.26)	(0.14)
h=16	**48.28**	8.96	**13.26**	**19.53**	**24.37**	9.70	0.45	7.77	**13.86**	9.27	4.22	7.24	5.45	5.89
	(0)	(0.06)	(0.01)	(0)	(0)	(0.05)	(0.98)	(0.1)	(0.01)	(0.05)	(0.38)	(0.12)	(0.24)	(0.21)
h=17	**47.56**	**10.77**	**13.03**	**19.54**	**25.10**	9.00	0.37	7.72	**13.44**	9.17	3.64	7.40	4.50	6.52
	(0)	(0.03)	(0.01)	(0)	(0)	(0.06)	(0.98)	(0.1)	(0.01)	(0.06)	(0.46)	(0.12)	(0.34)	(0.16)
h=18	**47.95**	**11.28**	**12.53**	**19.49**	**23.41**	8.81	0.30	7.59	**12.31**	**9.79**	3.46	7.07	4.24	7.10
	(0)	(0.02)	(0.01)	(0)	(0)	(0.07)	(0.99)	(0.11)	(0.02)	(0.04)	(0.48)	(0.13)	(0.37)	(0.13)
h=19	**45.64**	**13.94**	**12.87**	**19.49**	**20.92**	7.73	0.39	7.16	**11.66**	8.72	3.00	6.48	3.69	6.88
	(0)	(0.01)	(0.01)	(0)	(0)	(0.1)	(0.98)	(0.13)	(0.02)	(0.07)	(0.56)	(0.17)	(0.45)	(0.14)
h=20	**45.17**	**12.35**	**14.62**	**19.46**	**20.17**	8.69	0.49	7.59	**12.91**	8.39	2.73	6.79	3.19	7.68
	(0)	(0.01)	(0.01)	(0)	(0)	(0.07)	(0.97)	(0.11)	(0.01)	(0.08)	(0.6)	(0.15)	(0.53)	(0.1)

Note: Recursive estimation. Null hypothesis assumes correct calibration of the density forecast. Tables show the test statistic (and p-values in parentheses). We employ the first four raw moments to build the test statistic. Bold values highlight rejection of the test at a 5% significance level.

Regarding the Knüppel (2015) test, Table 5 reveals that no model is rejected for $h > 4$ months. In addition, models 5, 7 and 10-14 indeed are not rejected in any horizon (a similar result holds for rolling window estimation). The results for daily frequency show a slightly different picture, where only models 7 and 11 are not rejected (at 5% level) in any horizon.

Table 6 - Ranking of density models according to the LPDS

Monthly frequency

Rank of models based on LPDS

Model	1a	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12a	13a	14a
h=1	7	13	2	10	1	12	6	8	14	11	3	9	4	5
h=2	6	11	4	8	3	13	5	12	14	10	1	9	2	7
h=3	5	10	4	6	3	13	9	12	14	11	2	8	1	7
h=4	4		3	6	5	13	9	11	12	8	1	10	2	7
h=5	4		5	6	3	13	7	9	12	10	2	11	1	8
h=6	4		6	5	3	12	7	9	13	11	2	10	1	8
h=7	4		7	5	3	12	6	10	13	9	2	11	1	8
h=8	3		10	4	5	11	6	8	9	12	2	13	1	7
h=9	4		10	3	8	11	6	7	9	12	2	13	1	5
h=10	5		12	4	8	11	1	7	9	10	3	13	2	6
h=11	6		11	2	5	12	1	8	10	9	4	13	3	7
h=12	5		13	1	2	12	6	9	11	7	3	10	4	8

Daily frequency

Rank of models based on LPDS

Model	1a	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12a	13a	14a
h=1	12	14	7	13	1	11	5	6	10	8	3	9	4	2
h=2	12	14	1	13	2	10	6	7	11	9	3	8	4	5
h=3	12	14	1	13	2	11	5	7	10	9	3	8	4	6
h=4	12	14	1	13	2	10	6	7	11	8	3	9	4	5
h=5	12	14	1	13	2	11	5	7	10	8	3	9	4	6
h=6	12	14	2	13	1	10	6	7	11	9	3	8	4	5
h=7	12	14	4	13	1	10	6	8	11	9	2	7	3	5
h=8	12	14	5	13	1	10	6	7	11	9	2	8	3	4
h=9	12	14	4	13	3	10	6	9	11	8	2	7	1	5
h=10	12	14	4	13	3	11	6	9	10	8	1	7	2	5
h=11	10	14	4	13	3	12	6	9	11	8	1	7	2	5
h=12	10	14	5	13	3	11	6	9	12	7	1	8	2	4
h=13	10	14	6	13	3	11	5	9	12	7	1	8	2	4
h=14	11	14	6	13	3	10	4	7	12	9	1	8	2	5
h=15	10	14	6	13	3	11	4	7	12	9	1	8	2	5
h=16	7	14	6	13	3	12	5	8	11	9	1	10	2	4
h=17	7	14	6	13	1	11	4	8	12	9	2	10	3	5
h=18	8	14	6	12	1	11	5	9	13	7	3	10	2	4
h=19	7	14	6	11	1	12	5	10	13	9	2	8	3	4
h=20	6	14	7	9	1	12	5	11	13	8	3	10	2	4

Note: Recursive estimation. The best three models according to the LPDS rank ordering (i.e. higher LPDS figures) are highlighted in yellow for each horizon.

On the other hand, the LPDS ranking shown on Table 6 indicates, in general, model 5 (at short to medium horizons) and models 11 and 13 (almost all horizons) as the best-ranked models at monthly frequency. The results from rolling window (in the Supplementary Appendix) point out to models 1 and 4 (excepting some short horizons) as the best ones in the LPDS sense, whereas the OLS-Gaussian density estimation seems to improve the LPDS of models 6-9, for many horizons, *vis-à-vis* models 5 and 10-14. The daily frequency indicates a similar outcome, with models

5 and 11 in the Top3-ranking, for all horizons, followed by models 3 and 13, which also entered the best-three group of models for several horizons.

Now, we turn to an interesting empirical question, by investigating the RW paradigm from a density forecast perspective: Is it possible to beat the RW density forecast? To answer this question, we use the Amisano-Giacomini (2007) test, which compares the log score distance between two competing models. Because the theoretical setup of the test proposed by Amisano and Giacomini requires estimates with rolling samples of data, we only apply the test to the models estimated with the rolling window scheme.[30]

The null hypothesis assumes equal LPDS between model 1 (RW) and model $m \neq 1$. A negative test statistic indicates a higher LPDS of model m in comparison to the RW approach (i.e. model m is better than the RW, in the LPDS sense). In Table 7, the negative figures are highlighted in gray, whereas the green cells indicate negative values which are also statistically significant (at a 5% significant level); that is, green values indicate those cases where the density forecast has a LPDS statistically higher when compared to the random walk-based density forecast.

In the monthly frequency, note that the random walk density approach is overwhelmed in several cases (i.e. negative test statistics on Table 7, highlighted in gray). Models 3, 4, 7, 11 and 14 showed a relatively superior performance in respect to the RW in some horizons, although the LPDS difference (in all cases) is not statistically significant at the usual 5% significance level.[31] Regarding the daily frequency, the green cells depicted on Table 7 indicate the many cases which statistically defeat the random walk density forecast (e.g. all models, excepting models 2 and 4, for horizons up to 6 days; and model 11 for $h = 1, ..., 18$ days). Table 8 summarizes the results of the full-density analysis by presenting the recommended models based on recursive estimation.[32]

[30] We use the unweighted version of the Amisano-Giacomini (2007) test since, among others, Diks et al. (2011) noted that the weighted version of the test is improper, meaning that it can assign a higher average score to an incorrect density forecast than to the true conditional density, which is undesirable.

[31] The difficulty of the competing density forecasts to outperform the random walk approach on monthly frequency should not be a surprise given the probable low power of the considered evaluation methods due to a relatively short sample size (around 100 out-of-sample observations) to perform forecast comparison. In contrast, empirical exercises usually reported in the empirical finance literature (e.g. using daily returns) are based on sample sizes of thousands of observations.

[32] A given model is only presented on Table 8 if it is recommended by a given evaluation test/procedure.

Table 7 - Amisano-Giacomini (2007) test applied to average LPDS

Monthly frequency - rolling window

Amisano-Giacomini (2007): test statistic (p-value)

Model	2b	3b	4b	5b	6b	7b	8b	9b	5b	11b	12b	13b	14b
h=1	0.18	-0.05	0.01	0.04	0.09	-0.03	0.07	0.20	0.13	-0.01	0.02	0.03	-0.05
	(0.01)	(0.48)	(0.88)	(0.68)	(0.2)	(0.73)	(0.47)	(0.03)	(0.12)	(0.9)	(0.82)	(0.77)	(0.41)
h=2	0.20	0.03	0.04	0.14	0.25	0.05	0.25	0.45	0.14	0.13	0.19	0.17	0.10
	(0)	(0.67)	(0.01)	(0.32)	(0.07)	(0.59)	(0.08)	(0.01)	(0.21)	(0.37)	(0.09)	(0.26)	(0.34)
h=3	0.10	0.10	0.02	0.01	0.20	-0.03	0.20	0.55	0.15	-0.03	0.19	0.01	0.19
	(0.2)	(0.43)	(0.33)	(0.91)	(0.06)	(0.57)	(0.23)	(0)	(0.29)	(0.76)	(0.12)	(0.9)	(0.18)
h=4		0.15	0.01	0.02	0.34	0.15	0.33	0.44	0.30	0.19	0.39	0.17	0.36
		(0.35)	(0.5)	(0.86)	(0.01)	(0.17)	(0.04)	(0)	(0.11)	(0.27)	(0.05)	(0.23)	(0.09)
h=5		0.20	-0.01	0.03	0.49	0.01	0.33	0.51	0.31	0.02	0.52	0.04	0.37
		(0.31)	(0.75)	(0.84)	(0.01)	(0.91)	(0.14)	(0.01)	(0.1)	(0.88)	(0.01)	(0.73)	(0.06)
h=6		0.26	0.01	0.05	0.26	0.04	0.31	0.37	0.38	0.09	0.61	0.16	0.45
		(0.22)	(0.57)	(0.66)	(0)	(0.56)	(0.01)	(0.01)	(0.13)	(0.48)	(0.02)	(0.27)	(0.04)
h=7		0.27	-0.03	0.00	0.30	0.05	0.50	0.37	0.44	0.04	0.63	0.10	0.36
		(0.18)	(0.46)	(0.97)	(0.01)	(0.69)	(0.07)	(0.05)	(0.11)	(0.75)	(0.03)	(0.52)	(0.14)
h=8		0.32	-0.04	0.06	0.19	0.16	0.41	0.33	0.47	0.05	0.67	0.05	0.52
		(0.14)	(0.41)	(0.67)	(0.1)	(0.36)	(0.02)	(0.08)	(0.14)	(0.69)	(0.02)	(0.67)	(0.07)
h=9		0.66	-0.04	0.18	0.22	0.27	0.48	0.14	0.51	0.30	0.70	0.33	0.48
		(0.06)	(0.35)	(0.29)	(0.04)	(0.28)	(0.06)	(0.23)	(0.13)	(0.17)	(0.03)	(0.2)	(0.1)
h=10		0.99	-0.04	0.07	0.18	0.30	0.52	0.08	0.54	0.17	0.83	0.15	0.51
		(0.03)	(0.35)	(0.56)	(0.05)	(0.3)	(0.06)	(0.4)	(0.13)	(0.27)	(0.04)	(0.3)	(0.08)
h=11		1.26	-0.05	0.04	0.35	0.26	0.57	0.17	0.51	0.12	0.65	0.22	0.57
		(0.02)	(0.29)	(0.77)	(0.08)	(0.35)	(0.04)	(0.04)	(0.17)	(0.39)	(0.08)	(0.19)	(0.05)
h=12		1.45	-0.02	0.10	0.36	0.14	0.60	0.29	0.53	0.09	0.79	0.14	0.54
		(0.01)	(0.23)	(0.52)	(0.07)	(0.14)	(0.02)	(0.03)	(0.15)	(0.3)	(0.08)	(0.26)	(0.02)

Daily frequency - rolling window

Amisano-Giacomini (2007): test statistic (p-value)

Model	2b	3b	4b	5b	6b	7b	8b	9b	10b	11b	12b	13b	14b
h=1	1.84	-0.07	0.63	-0.10	-0.09	-0.10	-0.11	-0.10	-0.10	-0.11	-0.10	-0.12	-0.13
	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
h=2	2.20	-0.13	0.46	-0.11	-0.10	-0.11	-0.10	-0.10	-0.09	-0.11	-0.10	-0.11	-0.11
	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
h=3	2.90	-0.13	0.37	-0.10	-0.08	-0.10	-0.09	-0.10	-0.09	-0.11	-0.10	-0.11	-0.11
	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
h=4	1.72	-0.11	0.31	-0.09	-0.08	-0.09	-0.09	-0.09	-0.08	-0.11	-0.09	-0.10	-0.10
	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
h=5	1.54	-0.11	0.25	-0.11	-0.08	-0.10	-0.09	-0.08	-0.09	-0.12	-0.07	-0.12	-0.11
	(0)	(0)	(0)	(0)	(0.01)	(0)	(0)	(0.01)	(0)	(0)	(0.01)	(0)	(0)
h=6	1.55	-0.09	0.22	-0.09	-0.06	-0.07	-0.08	-0.07	-0.07	-0.10	-0.07	-0.10	-0.10
	(0)	(0)	(0)	(0)	(0.03)	(0.01)	(0.01)	(0.03)	(0.01)	(0)	(0.01)	(0)	(0)
h=7	1.77	-0.07	0.19	-0.09	-0.05	-0.06	-0.07	-0.06	-0.07	-0.09	-0.07	-0.10	-0.10
	(0)	(0.01)	(0)	(0)	(0.11)	(0.07)	(0.06)	(0.11)	(0.01)	(0)	(0.03)	(0)	(0)
h=8	1.54	-0.06	0.17	-0.08	-0.04	-0.06	-0.06	-0.05	-0.05	-0.09	-0.06	-0.09	-0.09
	(0)	(0.03)	(0)	(0)	(0.25)	(0.1)	(0.1)	(0.14)	(0.09)	(0)	(0.05)	(0)	(0)
h=9	1.45	-0.04	0.15	-0.08	-0.03	-0.05	-0.05	-0.03	-0.04	-0.08	-0.04	-0.08	-0.08
	(0)	(0.12)	(0)	(0)	(0.49)	(0.15)	(0.22)	(0.45)	(0.13)	(0)	(0.18)	(0)	(0.01)
h=10	1.34	-0.03	0.13	-0.07	-0.03	-0.05	-0.05	-0.03	-0.04	-0.08	-0.05	-0.08	-0.07
	(0)	(0.23)	(0)	(0)	(0.36)	(0.15)	(0.15)	(0.44)	(0.15)	(0)	(0.13)	(0)	(0.01)
h=11	1.32	-0.02	0.11	-0.07	-0.01	-0.04	-0.03	-0.01	-0.02	-0.07	-0.03	-0.08	-0.06
	(0)	(0.53)	(0)	(0.01)	(0.74)	(0.24)	(0.45)	(0.88)	(0.47)	(0)	(0.44)	(0)	(0.04)
h=12	1.43	-0.01	0.09	-0.06	-0.02	-0.03	-0.03	0.01	-0.03	-0.08	-0.03	-0.08	-0.06
	(0)	(0.75)	(0.02)	(0.04)	(0.65)	(0.46)	(0.5)	(0.83)	(0.41)	(0)	(0.32)	(0)	(0.04)
h=13	1.38	0.00	0.07	-0.06	-0.01	-0.02	-0.01	0.03	-0.03	-0.07	-0.02	-0.07	-0.06
	(0)	(0.95)	(0.07)	(0.07)	(0.87)	(0.58)	(0.77)	(0.59)	(0.37)	(0)	(0.67)	(0)	(0.05)
h=14	1.40	0.00	0.05	-0.07	0.02	-0.02	0.00	0.04	-0.02	-0.07	-0.01	-0.07	-0.06
	(0)	(0.99)	(0.23)	(0.06)	(0.73)	(0.64)	(0.93)	(0.47)	(0.61)	(0.01)	(0.79)	(0.01)	(0.12)
h=15	1.16	0.01	0.02	-0.06	0.00	-0.02	0.00	0.05	-0.02	-0.08	-0.01	-0.08	-0.07
	(0)	(0.79)	(0.63)	(0.14)	(0.92)	(0.7)	(0.93)	(0.42)	(0.62)	(0)	(0.75)	(0)	(0.07)
h=16	1.22	0.02	0.00	-0.07	0.02	-0.01	0.01	0.07	-0.02	-0.08	-0.01	-0.08	-0.06
	(0)	(0.67)	(1)	(0.16)	(0.74)	(0.88)	(0.93)	(0.27)	(0.73)	(0.02)	(0.83)	(0.02)	(0.15)
h=17	1.31	0.04	-0.01	-0.07	0.05	0.01	0.03	0.11	0.00	-0.07	-0.01	-0.07	-0.06
	(0)	(0.41)	(0.83)	(0.21)	(0.37)	(0.8)	(0.61)	(0.12)	(0.98)	(0.04)	(0.89)	(0.06)	(0.22)
h=18	1.32	0.03	-0.04	-0.08	0.05	0.00	0.04	0.11	0.00	-0.07	-0.01	-0.08	-0.07
	(0)	(0.44)	(0.51)	(0.2)	(0.41)	(0.94)	(0.6)	(0.14)	(0.97)	(0.04)	(0.89)	(0.04)	(0.15)
h=19	1.47	0.04	-0.05	-0.08	0.06	0.01	0.04	0.12	0.00	-0.07	0.00	-0.08	-0.06
	(0)	(0.38)	(0.39)	(0.2)	(0.35)	(0.82)	(0.58)	(0.12)	(0.98)	(0.07)	(0.94)	(0.06)	(0.19)
h=20	1.43	0.05	-0.06	-0.08	0.05	0.02	0.05	0.13	0.01	-0.07	0.00	-0.08	-0.06
	(0)	(0.27)	(0.3)	(0.24)	(0.4)	(0.75)	(0.48)	(0.12)	(0.85)	(0.08)	(0.95)	(0.06)	(0.21)

Note: Null hypothesis of zero average difference in LPDS between model 1 (benchmark) and model $m \neq 1$. Similar to Clark (2011), the p-values are computed by regressions of differences in log scores (time series) on a constant, using the Newey-West estimator of the variance of the regression constant (with a bandwidth of 0 at the 1-month horizon and 1.5$\times$horizon for other cases). Gray cells denote that model m is better than the RW in the LPDS sense. Green cells indicate m is statistically better (at a 5% level) than the RW.

Table 8 - Selected models - Full-density analysis

Monthly frequency

Horizon (months)	Coverage rate	LPDS	Knüppel test	Berkowitz test
1	5,7,10,11,12,13,14	3,5,11	5,7,10,11,12,13,14	6,7,8,10,11,12,13,14
2	3,5,7,11,13	5,11,13	3,5,7,10,11,12,13,14	7,10,11,12,14
3	3,5,7,11,12,13	5,11,13	3,5,6,7,8,9,10,11,12,13,14	6,9,10
6	5,7,10,11,12,13,14	5,11,13	1,3,4,5,6,7,8,9,10,11,12,13,14	9
9	5,6,7,8,9,10,11,12,14	4,11,13	1,3,4,5,6,7,8,9,10,11,12,13,14	6
12	5,6,7,8,9,10,12,14	4,5,11	1,3,4,5,6,7,8,9,10,11,12,13,14	-

Daily frequency

Horizon (days)	Coverage rate	LPDS	Knüppel test	AG test
1	7,11,13	5,11,14	2,7,8,11,14	3,5,6,7,8,9,10,11,12,13,14
2	11,13	3,5,11	2,7,8,11,13	3,5,6,7,8,9,10,11,12,13,14
3	2,7,11,13	3,5,11	7,8,11,13	3,5,6,7,8,9,10,11,12,13,14
5	3,7,11,13	3,5,11	6,7,8,11,12,13	3,5,6,7,8,9,10,11,12,13,14
10	3,7,11,13	5,11,13	6,7,8,10,11,12,13,14	5,11,13,14
20	7,11,13	5,11,13	6,7,8,10,11,12,13,14	-

Notes: Column 2 shows the models that presented a p-value above 0.05 in the coverage rate analysis. Column of LPDS shows the best 3 models according to the LPDS ranking. Columns of Knüppel and Berkowitz exhibit models not rejected (p-value>5%) in the respective tests. Column of AG test presents the models that statistically beat the RW density forecast (at a 5% significance level) in the Amisano-Giacomini (2007) test. No model is selected in the AG test at monthly frequency as well as in the Berkowitz (2001) test at daily frequency.

By comparing the results based on PITs (i.e. Berkowitz, 2001; and Knüppel, 2015) and the LPDS ranking shown on Table 8, note that the Top3 forecasts according to the LPDS ranking also belong to the set of density forecasts recommended by the Knüppel test (i.e. p-value>5%) for all cases with monthly frequency (excepting model 3, for $h = 1$); and often belong to the set of forecasts with good coverage rates (70% interval band). However, besides model 11 (for $h = 1, 2$), the mapping between the results on the LPDS and the Berkowitz test is less clear when compared to the one from Knüppel test; probably due to the fact that the former test is originally designed to test densities only for $h = 1$, whereas the latter test is constructed to properly deal with $h \geq 1$ forecast horizons.[33] Regarding daily frequency, the direct link between first-ranked LPDS forecasts and well-calibrated PITs (or good coverage rates) seems to be weaker compared to the results from monthly frequency.

On the other hand, also note from Table 8 the non-trivial amount of models (for several horizons at daily frequency) able to generate statistically better density

[33] In theory, we expect densities with higher LPDS to be better calibrated (according to the PITs); although in finite samples (for misspecified densities) the rankings can be distorted. Moreover, a well-calibrated density should be preferred by all loss functions (see Diebold et al., 1998) and indeed ranked first in any local evaluation too.

forecasts compared to the random walk approach (in AG test); which does not happen at monthly frequency.

3.4 Density forecasts: Local analysis

Now, we investigate the predictive accuracy of the density models under a local analysis approach. The idea is to check the performance of distinct parts of the conditional distribution, estimated through different approaches. A given model to generate the whole conditional density of the variable of interest might produce, for instance, an "adequate" risk measure for the left tail of the distribution (i.e., at low percentiles) but, at the same time, can generate "poor" risk measures at the central part (or at the right tail) of the distribution. For this reason, we next analyze the density models through the lens of their respective performance along a grid of selected quantile levels $\tau = \{0.1, 0.2, ..., 0.9\}$, in order to cover the key parts of the conditional distribution.[34]

A percentile of the conditional distribution, called here simply as a "conditional quantile", can also be viewed as a Value-at-Risk (VaR) measure (see Christoffersen et al., 2001). As pointed out by Wang and Wu (2012), the VaR is a prevalent risk management tool used by investors. It is essentially a one-sided forecast interval measuring downside risks. For this reason, the forecast evaluation of the selected "slices" of the distribution can naturally be conducted by using the many statistical tests available in the risk management literature, also known as "backtests" (see Jorion (2007) and Crouhy et al. (2001) for a good review). In this paper, we employ four procedures to conduct the local analysis: Local Forecast Coverage Rate, Kupiec (1995) test, Christoffersen (1998) test, and VQR test (see Appendix B3 for more details); although many more tests are currently available in the literature.[35] The results are shown on Table 9 (for $h = 1$, recursive estimation, monthly frequency).[36]

[34] It is worth mentioning that an analysis at extreme quantile levels (e.g., $\tau = 0.995$) is possible within our framework, although it would require a much higher number of observations in order to generate significant model estimates.

[35] Such as the nonparametric test of Crnkovic and Drachman (1997), the duration approach of Christoffersen and Pelletier (2004), the CAViaR setup of Engle and Manganelli (2004) and the Ljung-Box type-test of Berkowitz et al. (2008), among many others.

[36] See the Supplementary Appendix for further results on monthly frequency and all results based on daily frequency.

Table 9 - Local coverage rates and backtests for selected percentiles

Monthly frequency, recursive estimation, h=1

Forecast coverage rates: % of actual outcomes below the nominal quantile level (tau)

h=1	Model													
tau	1a	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12a	13a	14a
0.1	0.01	0.01	0.01	0.00	0.06	0.18	0.14	0.20	0.18	0.12	0.09	0.12	0.09	0.16
0.2	0.05	0.02	0.11	0.03	0.14	0.28	0.22	0.24	0.24	0.22	0.20	0.18	0.20	0.24
0.3	0.12	0.03	0.23	0.06	0.22	0.32	0.29	0.31	0.32	0.31	0.30	0.29	0.30	0.31
0.4	0.28	0.03	0.33	0.26	0.32	0.42	0.37	0.44	0.40	0.39	0.38	0.38	0.41	0.41
0.5	**0.57**	**0.04**	**0.50**	**0.51**	**0.43**	**0.50**	**0.47**	**0.50**	**0.50**	**0.49**	**0.47**	**0.49**	**0.47**	**0.50**
0.6	0.73	0.11	0.65	0.72	0.52	0.58	0.59	0.60	0.58	0.60	0.60	0.62	0.62	0.62
0.7	0.79	0.35	0.71	0.81	0.64	0.66	0.66	0.66	0.66	0.68	0.71	0.68	0.70	0.70
0.8	0.87	0.71	0.84	0.90	0.73	0.73	0.76	0.75	0.71	0.77	0.78	0.77	0.77	0.78
0.9	0.94	0.95	0.92	0.95	0.91	0.85	0.86	0.87	0.85	0.84	0.89	0.86	0.89	0.87

Kupiec (1995) test

h=1	p-value for each model													
tau	1a	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12a	13a	14a
0.1	0.00	0.00	0.00	0.00	0.17	0.01	0.14	0.00	0.01	0.56	0.72	0.56	0.72	0.04
0.2	0.00	0.00	0.01	0.00	0.07	0.05	0.67	0.27	0.27	0.67	0.96	0.60	0.96	0.27
0.3	0.00	0.00	0.12	0.00	0.05	0.58	0.79	0.88	0.58	0.88	0.95	0.79	0.95	0.88
0.4	0.01	0.00	0.15	0.00	0.06	0.62	0.51	0.38	0.94	0.79	0.64	0.64	0.91	0.76
0.5	**0.15**	**0.00**	**0.92**	**0.78**	**0.15**	**0.92**	**0.51**	**0.92**	**0.92**	**0.78**	**0.51**	**0.78**	**0.51**	**0.92**
0.6	0.00	0.00	0.29	0.01	0.10	0.62	0.91	0.94	0.62	0.94	0.94	0.64	0.64	0.64
0.7	0.03	0.00	0.79	0.01	0.17	0.34	0.34	0.34	0.34	0.73	0.79	0.58	0.95	0.95
0.8	0.04	0.03	0.31	0.00	0.07	0.07	0.27	0.18	0.03	0.38	0.67	0.38	0.38	0.67
0.9	0.17	0.08	0.49	0.03	0.72	0.08	0.24	0.38	0.08	0.04	0.78	0.14	0.78	0.38

Christoffersen (1998) test

h=1	p-value for each model													
tau	1a	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12a	13a	14a
0.1	0.00	0.00	0.00	0.00	0.14	0.02	0.06	0.00	0.02	0.39	0.16	0.77	0.59	0.05
0.2	0.00	0.00	0.03	0.00	0.20	0.11	0.91	0.52	0.52	0.73	0.98	0.80	0.93	0.38
0.3	0.00	0.00	0.25	0.00	0.14	0.73	0.90	0.77	0.73	0.81	0.61	0.56	0.61	0.54
0.4	0.02	0.00	0.21	0.00	0.17	0.63	0.60	0.45	0.60	0.95	0.23	0.84	0.33	0.72
0.5	**0.29**	**0.00**	**0.69**	**0.01**	**0.32**	**0.80**	**0.78**	**0.80**	**0.80**	**0.77**	**0.78**	**0.77**	**0.17**	**0.36**
0.6	0.01	0.00	0.44	0.00	0.09	0.63	0.55	0.82	0.63	0.85	0.97	0.90	0.63	0.84
0.7	0.04	0.00	0.43	0.00	0.33	0.63	0.63	0.63	0.63	0.86	0.43	0.73	0.61	0.61
0.8	0.07	0.08	0.47	0.00	0.04	0.14	0.25	0.37	0.06	0.09	0.29	0.09	0.22	0.11
0.9	0.00	0.13	0.18	0.00	0.12	0.20	0.50	0.66	0.21	0.10	0.32	0.16	0.32	0.41

VQR (2011) test

h=1	p-value for each model													
tau	1a	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12a	13a	14a
0.1	0.00	0.00	0.00	0.00	0.01	0.03	0.55	0.06	0.01	0.17	0.10	0.28	0.10	0.04
0.2	0.00	0.00	0.06	0.00	0.14	0.45	0.50	0.32	0.18	0.62	0.88	0.40	0.53	0.42
0.3	0.00	0.00	0.51	0.00	0.19	0.55	0.44	0.58	0.56	0.74	0.98	0.78	0.92	0.80
0.4	0.20	0.00	0.64	0.25	0.39	0.51	0.49	0.64	0.79	0.94	0.89	0.91	0.83	0.99
0.5	**0.63**	**0.00**	**0.93**	**0.33**	**0.48**	**0.67**	**0.69**	**0.60**	**0.75**	**0.98**	**0.98**	**0.97**	**0.95**	**0.97**
0.6	0.00	0.00	0.78	0.04	0.57	0.83	0.74	0.69	0.82	0.83	0.99	0.86	0.93	0.93
0.7	0.00	0.00	0.60	0.00	0.13	0.28	0.19	0.23	0.20	0.64	0.54	0.65	0.42	0.66
0.8	0.01	0.13	0.72	0.00	0.25	0.17	0.23	0.11	0.28	0.34	0.99	0.46	0.95	0.93
0.9	0.04	0.54	0.45	0.00	0.92	0.13	0.31	0.40	0.23	0.33	0.26	0.43	0.23	0.50

A rejection of a given model for a selected horizon and a percentile of the distribution suggests the need for local improvement on the density model (in order to eliminate, for instance, a wrong coverage rate, a clustering behavior or even a poor time-dynamics). We next summarize the local analysis in terms of the three considered backtests (Kupiec, Christoffersen and VQR).

In this sense, we aggregate the results in respect to the lower quantiles $\tau = \{0.1, 0.2, 0.3\}$ or higher quantiles $\tau = \{0.7, 0.8, 0.9\}$. The results are shown on Table

10, where a given model is only shown at a given cell (e.g. monthly frequency, $h = 1$) if it has a p-value>5% in all 9 possible cases (i.e. 3 percentiles x 3 tests). Indeed, at monthly frequency and $h = 1$ only a few models are not rejected (at the same time) in the three statistical tests (and in the three selected percentiles). For $h > 1$ not a single model would be selected. This way, for $h > 1$ the adopted criteria to select models is weakened as long as the forecast horizon increases. Regarding the daily frequency, we only use the VQR test to select models, since the other two considered backtests reject all models (in all cases) at a 5% significance level.

Table 10 - Selected Models - Local Analysis

Monthly frequency

Horizon (months)	Selection Criteria	Lower Quantiles (τ = 0.1, 0.2, 0.3)	Higher Quantiles (τ = 0.7, 0.8, 0.9)
1	Kupiec, Christoffersen, VQR	7,10,11,12,13	3,6,7,8,11,12,13,14
2	Kupiec, VQR	7,11,13	3,11,13,14
3	Kupiec	7,13	1,3,5,11,13,14
6	Kupiec	3,7	1,10,11,13,14
9	Kupiec	14	7,9,10,11,13,14
12	Kupiec	7,8,14	6,7,8,12

Daily frequency

Horizon (days)	Selection Criteria	Lower Quantiles (τ = 0.1, 0.2, 0.3)	Higher Quantiles (τ = 0.7, 0.8, 0.9)
1	VQR	6,7	-
2	VQR	6,9	11
3	VQR	7	3
5	VQR	7	-
10	VQR	-	11,13
20	VQR	10	11,14

Note: Recursive estimation results. A model is selected at monthly frequency only if it is recommended (i.e. p-value>5%) by a given criteria (1, 2 or 3 backtests); and at daily frequency only if it shows a p-value>5% in the VQR test (on at least two quantile levels, out of the three considered levels for each tail).

Note that the models more suitable to forecast quantiles related to the lower part of the exchange rate distribution (focused on the appreciation of the domestic currency *vis-à-vis* the U.S. dollar) might not serve to properly account for the other tail of the distribution (especially, at the daily frequency).

By comparing Tables 8 and 10 (that is, the full-density analysis with the local analysis), one can clearly note which models work in practice, in terms of forecasting accuracy, and for which purpose, data frequency and/or forecast horizon.

A closer look at those tables reveal that, indeed, some models recommended by the local analysis also exhibit a good performance according to the full-density investigation. For instance, at monthly frequency and $h = 1$, among the many models indicated by Table 10, in the local analysis (at both lower and higher quantile levels), note models 11 and 13, which also belong to Table 8, due to their well-calibrated densities, according to the coverage rate, LPDS ranking (model 11 only), and Knüppel and Berkowitz tests. For the longest horizon at monthly frequency ($h = 12$), note models 7 and 8, which belong to Table 10 in both tails (lower and higher quantiles) and are also recommended by Table 8, according to the coverage rate and Knüppel criteria.

At daily frequency, one can also identify a few models that produce good density forecasts from both full-density and local analysis points-of-view. For example, in all horizons (excepting $h = 20$), all models indicated by Table 10, at lower or higher quantile levels, also belong to the AG test's column on Table 8, at respective horizons. In particular, model 7, which is indicated to properly forecast the lower quantiles in the local analysis (Table 10, for $h = 1$, 3 and 5 days) is also recommended on Table 8, at respective horizons, according to the coverage rate, and the AG and Knüppel tests. In respect to the higher quantiles, the local analysis at daily frequency suggests, for instance, model 11 for $h = 2$, 10 and 20 days; whereas this same model shows a well-calibrated density according to the coverage rate, LPDS ranking, Knüppel test and AG test (excepting $h = 20$).

On the other hand, based on the set of results obtained from both full-density and the local analysis, we are also able to distinguish between whether it is the economic variables that affect forecast performance or how they are modelled. In this sense, by comparing the forecast performance of models 5-14 estimated with QR *versus* OLS, at daily frequency, it seems that the QR approach is quite often more indicated; whereas, at monthly frequency, the OLS estimation (in several cases) is the best one.[37]

[37] For example, at daily frequency, the QR-based models produce more well-calibrated forecasts (compared to the OLS ones) according to the following criteria: Giacomini-White (2006) test (for shorter horizons), FX rate direction of change test (in general), coverage rates (70% interval band), Knüppel (20015) test, Amisano-Giacomini (2007) test, and in the local analysis (for both lower and higher quantile levels). OLS is recommended at daily frequency (instead of QR) only in the following cases: Diebold-Mariano (1995) test, Giacomini-White (2006) test (only model 5, for longer horizons) and the FX rate direction of change test (for rolling window and longer horizons). In respect

3.5 Model ranking from risk analysis

According to Crouhy et al. (2001), backtests provide a key check of model accuracy and robustness, by considering *ex-ante* risk measure forecasts and comparing it to *ex-post* realized outcomes. In other words, the backtesting procedures (used in previous section) are designed to test whether a particular forecasting approach provides an accurate conditional quantile (i.e. VaR) forecast. However, instead of only checking the performance of a single model, one might be interested in discriminating among competing models and deciding which of them is best for risk analysis purposes (see Nieto and Ruiz, 2016).

With this goal, Lopez (1999) proposes the selection of the forecast procedure that minimizes the loss $L(m) = \sum_{t=T+1}^{T+P} L_{m,t}$, with

$$L_{m,t} = \begin{cases} f(s_t; VaR_{m,\tau,t}) & \text{if } s_t < VaR_{m,\tau,t} \\ g(s_t; VaR_{m,\tau,t}) & \text{if } s_t \geq VaR_{m,\tau,t} \end{cases} \tag{6}$$

where $t = T+1, ..., T+P$ describes the out-of-sample forecast evaluation period, the index (m) stands for forecasting model m of the target variable s_t, τ is the quantile level of interest and f and g are functions such that $f(x; y) \geq g(x; y)$. The idea is to measure the conditional coverage distance of a VaR from its nominal benchmark. According to the author, a single numerical score could reflect regulatory concerns and provide a measure of relative performance to compare competing VaR models across time and institutions.

Nonetheless, the loss functions proposed by Lopez (1999) are not able to distinguish properly between the true data generating process and alternative models for forecasting the VaR (e.g. as consistent-scoring functions; see Gneiting and Raftery (2007) and Gneiting (2011) for further details).

In this sense, we adopt here the following predictive quantile loss function pro-

to the monthly frequency, the OLS approach seems to be more suitable, in general, for estimation of models 5-14. For instance, OLS produces lower RMSEs compared to the QR-based forecasts (and, thus, OLS shows a better performance, compared to QR, according to DM and GW tests vis-à-vis the random walk forecast). OLS is also better according to the coverage rates, the AG test, and along the local analysis at higher quantile levels (results for lower quantiles indicate a relatively similar performance between OLS and QR). Nonetheless, the FX rate direction of change test, based on recursive estimation, is more favorable to the QR-based point forecasts. In addition, the QR-based density forecasts, at monthly frequency, generate more well-calibrated PITs, compared to OLS, according to the tests of Berkowitz and Knüppel.

posed by Giacomini and Komunjer (2005):

$$L_{m,t} = \left[\tau - 1_{s_t < VaR_{m,\tau,t}}\right] \left[s_t - VaR_{m,\tau,t}\right] \tag{7}$$

where $VaR_{m,\tau,t} = Q_{m,\tau}(s_t \mid \mathcal{F}_{t-h})$ is the conditional quantile of s_t at quantile level τ (conditioned on the information set $\mathcal{F}_{t-h}$). This simple decision rule for model selection, based on the loss of Giacomini and Komunjer (2005), allows us to rank the competing models according to risk events of interest (e.g. what is the best model to forecast the exchange rate at a given quantile of the FX rate conditional distribution?)

Note that the adopted decision rule for model selection does not require knowledge of the underlying density model or, if the model is known, it does not restrict attention to a specific estimation procedure, since it only relies on the conditional quantile forecasts over a selected quantile level τ.

By establishing $\underline{\tau} = 0.1$ we are able to rank the forecasting models according to their ability of providing good value-at-risk measures (or conditional quantiles) related to the left tail of the conditional distribution of the R$/US$ exchange rate (that is, to properly account for the valuation risk of the R$ currency in respect to the US$). In the same way, we also set $\overline{\tau} = 0.9$ to investigate the best models to account for the devaluation risk of the FX rate. Table 11 shows the model rankings for both tails, monthly and daily frequencies, and recursive estimation.

Firstly, note that fundamentals matter for the risk of FX rate valuation. The Top5 group of models that produce better forecasts at $\tau = 0.1$ (i.e. left tail of the FX rate distribution) belongs to the economic-driven set of models 6-14, in both frequencies. The Taylor rule with interest rate smoothing (model 9) is the best one at monthly frequency, whereas the absolute PPP model (model 10) is the best one at daily frequency.

On the other hand, the Top5 best models to deal with devaluation risk, at the right tail of the FX rate distribution (i.e. $\tau = 0.9$), come from the set of models 1-5, also in both frequencies. Indeed, the best VaR forecast of FX rate devaluation is the survey-based forecast (model 4) at monthly frequency, whereas the financial-data (option-implied) model 2 is the best one at daily frequency.

Table 11 - Model Ranking for Risk Assessment ($h = 1$)

Model	Devaluation (M)	Devaluation (D)	Valuation (M)	Valuation (D)
1) Random walk (without drift)	3	4	12	11
2) Option-implied (RND-RWD)	2	1	14	14
3) GARCH - Monte Carlo	4	3	11	12
4) Survey forecast	1	2	13	13
5) Survey forec. (bias-correct)	5	5	10	10
6) Taylor rule model	13	11	2	4
7) Taylor rule (PPP)	10	7	7	8
8) Taylor rule (PPP, smoothing)	8	9	4	6
9) Taylor rule (smoothing)	12	10	1	3
10) Absolute PPP model	14	14	3	1
11) Relative PPP model	6	8	8	7
12) Monetary model	11	13	6	5
13) Monetary model (weaker)	7	6	9	9
14) Forward premium model	9	12	5	2

Notes: (M) means monthly frequency and (D) denotes daily frequency. The devaluation risk refers to $\tau = 0.9$ whereas valuation refers to $\tau = 0.1$. The Giacomini-Komunjer (2005) predictive quantile loss function is adopted.

Also note that the best models from the risk analysis exercise often does not belong to the set of models previously recommended in the Local Analysis, as presented on Table 10. This results is probably because the Local Analysis considers a range of quantile levels (with different backtesting results from distinct quantile levels), whereas the risk analysis is focused on a single (and extreme) quantile of interest.

Moreover, the models that produced the best point forecasts (based on RMSE and, thus, focused on the central part of the FX rate conditional distribution) are not necessarily the same that showed the best forecasting performance for risk assessment purposes. This empirical finding can be attributed, for instance, to an asymmetric response of the exchange rate in respect to the macroeconomic fundamentals.

Finally, it is worth mentioning that tail risk in Brazil can also be affected by official interventions in the FX market, which are not properly captured by any of the investigated models here. In this case, a different setup with intra-day data, beyond the scope of this paper, could be further explored (see Kohlscheen and Andrade, 2013).[38]

[38] The authors investigate official interventions in the Brazilian FX market (i.e. currency swap auctions, which are focused on providing hedge to economic agents, liquidity to domestic FX market and reducing excessive market volatility) based on high-frequency data.

3.6 Recursive estimation *versus* rolling window

A careful investigation of the tables presented in the Supplementary Appendix reveals mixed results regarding recursive estimation and rolling window estimation schemes. On the monthly frequency, in general, the rolling window estimation scheme seems to produce slightly better (point and density) forecasts compared to the recursive estimation. For instance, rolling window generates more forecasts with lower RMSEs compared to the random walk; more density forecasts with adequate 70% coverage rates (and not rejected at the Berkowitz (2001)'s test); and more conditional quantiles recommended in the local analysis on the left tail of the FX rate distribution (i.e. lower quantile levels). In contrast, recursive estimation produces slightly more forecasts able to correctly predict the FX rate direction of change based on the Pesaran and Timmermann (1992, 2009)'s test; and more value-at-risk measures to properly deal with FX rate upside risks (i.e. at higher quantile levels). The Knüppel (2015) test shows similar outcomes for both estimation schemes.

On the other hand, the forecast comparison at the daily frequency seems to be slightly more favorable to the recursive estimation, for example, which generates more forecasts with lower RMSEs compared to the random walk (among the models 5-14 estimated with OLS), more forecasts that correctly predict the FX rate direction of change and more density forecasts not rejected in the Knüppel test. In turn, the rolling window (at daily frequency) produces more forecasts with lower RMSEs compared to the random walk (among the models 5-14 estimated with QR), more density forecasts with adequate 70% coverage rates and more adequate VaRs for the FX rate upside risk (i.e. higher quantile levels).

These empirical findings (recursive estimation for daily data; and rolling scheme for monthly data) point out to possibly few (or none) structural breaks in the FX rate dynamics on a daily basis but, at the same time, to some probable breaks on a monthly basis.[39] This outcome is in line with the fact that statistical relationships among macro variables, including the FX rate, potentially change over time, which is most likely to happen under a lower frequency and longer time span.

[39] Of course a formal investigation of structural breaks in the FX rate dynamics (beyond the scope of this paper) would require adequate statistical tests, instead of a simple forecast comparison of different sampling estimation schemes, as discussed here.

4 Conclusion

This article has examined several models of the Brazilian foreign exchange rate (R$/US$) through the lens of forecast evaluation tools. We follow a strand of literature that goes beyond the conditional mean analysis and focus on the density forecast of the FX rate.

Our contribution is, thus, to provide a toolkit to evaluate available FX rate models according to its point and density forecast performance; bridging the gap between distinct strands of the literature on international economics, forecasting and financial risk analysis. In this paper, we put together distinct techniques to construct FX rate models, for instance, based on statistical or economic-driven approaches; using financial data information as well as macroeconomic variables; and employing parametric (e.g. Gaussian, t-Student) or nonparametric distributions or quantile regression techniques.

In order to evaluate such forecasts, we use standard point forecast evaluation tools; and propose a full-density/local analysis approach, which can reveal the suitable models for a determined forecasting goal. Finally, we employ a simple decision rule for model selection, which allows one to rank models according to different risk events of interest. Such a tool can be useful for econometricians, risk managers or policy makers interested in evaluating competing models and selecting those that historically provide more accurate predictions of risk events.

Overall, the results for Brazil point out that no single model properly accounts for the entire density in all forecast horizons, at least at conventional levels of significance. In fact, the choice of a density forecast model for the FX rate depends on the part of the conditional distribution of interest as well as on the forecast horizon. The reason is that some models are more prone to produce good forecasts at high (or low) percentiles of the FX rate density, which is in line with an asymmetric response of covariates (e.g. macro fundamentals) in respect to the exchange rate conditional distribution. In other words, a given macroeconomic fundamental, for instance, which might be useless to explain the conditional mean exchange rate dynamics (as widely reported in the literature), might be adequate to explain the upside (or downside) risk of FX rate at a particular horizon. By focusing on the accuracy of

density models in predicting the likelihood of a sharp valuation/devaluation event, we are also able to select models to be used for risk management purposes.

What are the lessons from the empirical investigation of the Brazilian FX rate that could be potentially used to other currencies? The results for the FX rate point forecast in Brazil corroborate previous findings of the literature, such as the difficulty on beating the random walk forecast at monthly frequency, although the random walk paradigm can statistically be broken, in some cases, at daily frequency[40]; fundamental relationships (e.g. parity conditions) hold better in the long-run (Mark, 1995); economic-fundamental-based exchange rate models perform better than the random walk in predicting tighter forecast intervals, especially at monthly frequency and long horizons (Wang and Wu, 2012); correct FX rate "direction prediction" is found in many cases and appears to cluster at longer horizons (Cheung, Chinn and Pascual, 2005); option data-implied forecasts provide relatively accurate forecasts at monthly frequency and short horizons (Christoffersen and Mazzotta, 2005).

Regarding density forecasts, we compile a set of empirical findings (stylized facts) - that are not present in the existing literature - which may provide some guidance to academics, policy makers and market practitioners who are interested in forecasting the full density of exchange rate returns. The majority of models showed an adequate coverage rate in many horizons and both frequencies. Nonetheless, the density forecasts only survive the PIT-based test of Berkowitz (2001) at monthly frequency for short horizons (e.g. the economic-driven and the bias-corrected survey forecasts), with a very few exceptions at longer horizons (and no model survived the test at daily frequency). The test of Knüppel (2015) also based on PITs and using raw moments allows us to discriminate the competing density forecasts along the considered horizons and both frequencies. On the other hand, the density forecast comparison based on the test of Amisano and Giacomini (2007) reveals that the random walk approach is overwhelmed in several cases (at a 5% significant level, daily frequency, for several models and horizons).

In respect to the local analysis, we find that those models more suitable to forecast quantiles related to the lower part of the exchange rate distribution (focused

[40] Indeed, it is well documented in the literature of density forecasts that statistical and financial models can beat the random walk in forecasting exchange rates at higher frequencies (e.g. intra-day data). See Diebold, Hahn, and Tay (1999), Christoffersen and Mazzotta (2005) and Sarno and Valente (2005).

on the appreciation of the domestic currency *vis-à-vis* the U.S. dollar) might not serve to properly account for the other tail of the distribution. Nonetheless, some models recommended by the local analysis also exhibit a good performance according to the full-density investigation.

By ranking the competing models based on risk analysis, we find that models that produced the best point forecasts (based on RMSE and, thus, focused on the central part of the FX rate conditional distribution) are not necessarily the same that showed the best forecasting performance for risk assessment. In addition, macro fundamentals matter for valuation risk of the FX rate, in both frequencies (e.g. the Taylor rule with interest rate smoothing is the best one at monthly frequency, whereas the absolute PPP model is the best one at daily frequency). In contrast, regarding devaluation risk, the best models are the survey-based at monthly frequency and the financial data model (options) at daily frequency.

On the other hand, taking into account all the evaluation procedures used in this paper, and by comparing the forecast performance of economic-driven models estimated with QR *versus* OLS, at daily frequency, it seems that the QR approach (quite often) is more indicated; whereas at monthly frequency the OLS estimation (in several cases) is the best one.

In respect to the adopted estimation scheme, we find that, in general, the rolling window estimation scheme at monthly frequency seems to produce slightly better point and density forecasts compared to the recursive estimation, whereas the forecast comparison at the daily frequency seems to be slightly more favorable to the recursive estimation. These findings point out to probably few (or none) structural breaks in the FX rate dynamics on a daily basis, but to possible breaks on a monthly basis.

Possible extensions of this research include: (i) other covariates to explain FX dynamics in the long-run (e.g., commodity price index, as suggested by Kohlscheen, 2013); (ii) additional density models (e.g., GARCH-in-mean); (iii) density forecast combination (Hall and Mitchell, 2007; Jore et al., 2010; Kascha and Ravazzolo, 2010; Gaglianone and Lima, 2014); (iv) risk assessment based on alternative risk measures (Artzner et al., 1999); or (v) microstructure approach based on intra-day data.

References

[1] Adolfson, M., Linde, J., Villani, M., 2005. Forecasting Performance of an Open Economy Dynamic Stochastic General Equilibrium Model. Sveriges Riksbank Working Paper n.190.

[2] Amisano, G., Giacomini, R., 2007. Comparing Density Forecasts via Weighted Likelihood Ratio Tests. *Journal of Business and Economic Statistics* 25, 177-190.

[3] Artzner, P., Delbaen, F., Eber, J., Heath, D., 1999. Coherent Measures of Risk. *Mathematical Finance* 9, 203-228.

[4] Bacchetta, P., van Wincoop, E., 2006. Can Information Heterogeneity Explain the Exchange Rate Determination Puzzle? *American Economic Review* 96, 552-576.

[5] Berkowitz, J., 2001. Testing Density Forecasts, With Applications to Risk Management. *Journal of Business and Economic Statistics* 19, 465-474.

[6] Berkowitz, J., Christoffersen, P., Pelletier, D., 2008. Evaluating Value-at-Risk Models with Desk-Level Data. Mimeo. Available at: http://www4.ncsu.edu/~dpellet/papers/BCP_23Jun08.pdf

[7] BIS - Bank for International Settlements, 2013. Triennial Central Bank Survey - Foreign exchange turnover in April 2013: preliminary global results. Available at http://www.bis.org/publ/rpfx13fx.pdf

[8] Black, F., 1976. The Pricing of Commodity Contracts. *Journal of Financial Economics*, 3, 167-179.

[9] Boero, G., Marrocu, E., 2004. The Performance of SETAR Models: A Regime Conditional Evaluation of Point, Interval and Density Forecasts. *International Journal of Forecasting* 20, 305–20.

[10] Breeden, D., Litzenberger, R., 1978. Prices of state-contingent claims implicit in option prices, *Journal of Business*, 51, 621-51.

[11] Burnside, C., Eichenbaum, M., Kleshchelski, I., Rebelo, S., 2006. The Returns to Currency Speculation. NBER Working Paper 12489.

[12] Burnside, C., Eichenbaum, M., Kleshchelski, I., Rebelo, S., 2011. Do Peso Problems Explain the Returns to the Carry Trade? *Review of Financial Studies* 24(3), 853-891.

[13] Capistrán, C., Timmermann, A., 2009. Forecast Combination with Entry and Exit of Experts. *Journal of Business and Economic Statistics* 27, 428–40.

[14] Chen, Y., Tsang, K.P., 2009. What Does the Yield Curve Tell Us about Exchange Rate Predictability? Manuscript, University of Washington, and Virginia Tech.

[15] Chernozhukov, V., Fernandez-Val, I., Galichon, A., 2010. Quantile and Probability Curves without Crossing. *Econometrica* 78, 1093-1125.

[16] Cheung, Y.-W., Chinn, M.D., Pascual, A.G., 2005. Empirical exchange rate models of the nineties: are any fit to survive? *Journal of International Money and Finance* 24, 1150–1175.

[17] Christoffersen, P.F., 1998. Evaluating Interval Forecasts. *International Economic Review* 39, 841–862.

[18] Christoffersen, P.F., Hahn, J., Inoue, A., 2001. Testing and Comparing Value-at-Risk Measures. *Journal of Empirical Finance* 8, 325-342.

[19] Christoffersen, P.F., Mazzotta, S., 2005. The Accuracy of Density Forecasts from Foreign Exchange Options. *Journal of Financial Econometrics* 3, 578–605.

[20] Christoffersen, P.F., Pelletier, D., 2004. Backtesting Value-at-Risk: A Duration-Based Approach. *Journal of Financial Econometrics* 2 (1), 84-108.

[21] Clark, T.E., 2011. Real-Time Density Forecasts from Bayesian Vector Autoregressions with Stochastic Volatility. *Journal of Business and Economic Statistics* 29, 327-341.

[22] Clark, T.E., McCracken, M.W., 2012. Advances in Forecast Evaluation. Federal Reserve Bank of St. Louis, Working Paper Series 2011-025B. Available at: http://research.stlouisfed.org/wp/2011/2011-025.pdf

[23] Clements, M.P., 2004. Evaluating the Bank of England Density Forecasts of Inflation. *Economic Journal* 114, 844-866.

[24] Clements, M.P., 2005. Evaluating econometric forecasts of economic and financial variables. Palgrave Macmillan, first edition.

[25] Clews, R., Panigirtzoglou, N., Proudman, J., 2000. Recent developments in extracting information from options markets. Bank of England, Quarterly Bulletin: February 2000.

[26] Corradi, V., Swanson, N., 2006. Predictive density evaluation. In Handbook of economic forecasting, vol. 1, 197-284.

[27] Crnkovic, C., Drachman, J., 1997. Quality Control in VaR: Understanding and Applying Value-at-Risk. London: Risk Publications.

[28] Crouhy, M., Galai, D., Mark, R., 2001. Risk Management. McGraw-Hill.

[29] Della-Corte, P., Sarno, L., Tsiakas, I., 2009. An Economic Evaluation of Empirical Exchange Rate Models. *Review of Financial Studies* 22, 3491–530.

[30] Diebold, F.X., Gunther, T.A., Tay, A.S., 1998. Evaluating Density Forecasts, with Applications to Financial Risk Management. *International Economic Review* 39(4), 863-883.

[31] Diebold, F.X., Hahn, J., Tay, A.S., 1999. Multivariate Density Forecast Evaluation and Calibration in Financial Risk Management: High-Frequency Returns of Foreign Exchange. *Review of Economics and Statistics* 81, 661–73.

[32] Diebold, F.X., Mariano, R.S., 1995. Comparing Predictive Accuracy. *Journal of Business and Economic Statistics* 13, 253-263.

[33] Diks, C., Panchenko, V., van Dijk, D., 2011. Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics* 163(2), 215-230.

[34] Engel, C., Wang, J., Wu, J.J., 2009. Long-Horizon Forecasts of Asset Prices when the Discount Factor is close to Unity. Globalization and Monetary Policy Institute Working Paper No. 36.

[35] Engel, C., West, K.D., 2005. Exchange rate and fundamentals. *Journal of Political Economy* 113, 485–517.

[36] Engle, R.F., Manganelli, S., 2004. CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles. *Journal of Business and Economic Statistics* 22 (4), 367-381.

[37] Foroni, C., Guérin, P., Marcellino, M., 2015. Using low frequency information for predicting high frequency variables. Norges Bank Research 13-2015.

[38] Fratzscher, M., Sarno, L., Zinna, G., 2012. The Scapegoat Theory of Exchange Rates: The First Tests. European Central Bank (ECB) Working Paper Series 1418.

[39] Gaglianone, W.P., Lima L.R., Linton, O., Smith D.R., 2011. Evaluating Value-at-Risk Models via Quantile Regressions. *Journal of Business and Economic Statistics* 29, 150-160.

[40] Gaglianone, W.P., Lima, L.R., 2012. Constructing density forecasts from quantile regressions. *Journal of Money, Credit and Banking* 44(8), 1589-1607.

[41] Gaglianone, W.P., Lima, L.R., 2014. Constructing optimal density forecasts from point forecast combinations. *Journal of Applied Econometrics* 29(5), 736-757.

[42] Giacomini, R., Komunjer, I., 2005. Evaluation and Combination of Conditional Quantile Forecasts. *Journal of Business and Economic Statistics* 23(4), 416-431.

[43] Giacomini, R., White, H., 2006. Tests of Conditional Predictive Ability. *Econometrica* 74, 1545-1578.

[44] Glasserman, P., 2004. Monte Carlo Methods in Financial Engineering. Springer.

[45] Gneiting, T., 2011. Making and Evaluating Point Forecasts. *Journal of the American Statistical Association* 106(494), 746-762.

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

[47] Groen, J.J., Matsumoto, A., 2004. Real Exchange Rate Persistence and Systematic Monetary Policy Behavior. Bank of England Working Paper No. 231.

[48] Hall, S.G., Mitchell, J., 2007. Combining density forecasts. *International Journal of Forecasting* 23, 1-13.

[49] Harvey, D., Leybourne, S., Newbold, P., 1997. Testing the equality of prediction mean squared errors. *International Journal of Forecasting* 13(2), 281–291.

[50] Hansen, L.P., 1982. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica* 50, 1029-1054.

[51] He, X., 1997. Quantile curves without crossing. *The American Statistician* 51, 186-92.

[52] Hong, Y., Li, H., Zhao, F., 2007. Can the Random Walk Be Beaten in Out-of-Sample Density Forecasts: Evidence from Intraday Foreign Exchange Rates. *Journal of Econometrics* 141, 736–76.

[53] Jore, A.S., Mitchell, J., Vahey, S.P., 2010. Combining Forecast Densities from VARs with Uncertain Instabilities. *Journal of Applied Econometrics* 25(4), 621-634.

[54] Jorion, P., 2007. Value-at-risk: The new benchmark for managing financial risk. McGraw Hill, 3rd edition.

[55] Kascha C., Ravazzolo, F., 2010. Combining Inflation Density Forecasts. *Journal of Forecasting* 29, 231-250.

[56] Kilian, L., 1999. Exchange rates and monetary fundamentals: what do we learn from long-horizon regressions? *Journal of Applied Econometrics* 14, 491-510.

[57] Kilian, L., Taylor, M., 2003. Why Is It So Difficult to Beat the Random Walk Forecast of Exchange Rate? *Journal of International Economics* 60, 85-107.

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

[59] Ko, S., Park, S.Y., 2013. Multivariate density forecast evaluation: A modified approach. *International Journal of Forecasting* 29, 431-441.

[60] Koenker, R., 2005. Quantile Regression. Cambridge University Press, Cambridge.

[61] Koenker, R., Bassett, G., 1978. Regression Quantiles. *Econometrica* 46, 33-49.

[62] Kohlscheen, E., 2013. Long-Run Determinants of the Brazilian Real: a Closer Look at Commodities. Banco Central do Brasil. Working Paper no. 314.

[63] Kohlscheen, E., Andrade, S.C., 2013. Official Interventions through Derivatives: Affecting the Demand for Foreign Exchange. Banco Central do Brasil. Working Paper no. 317.

[64] Kupiec, P., 1995. Techniques for Verifying the Accuracy of Risk Measurement Models. *Journal of Derivatives* 3, 73–84.

[65] Lopez, J.A., 1999. Methods for Evaluating Value-at-Risk Estimates. Federal Reserve Bank of San Francisco, Economic Review 2, 3-17.

[66] Lustig, H., Roussanov, N., Verdelhan, A., 2011. Common Risk Factors in Currency Markets. *Review of Financial Studies* 24(11), 3731-3777.

[67] Marcellino, M., Stock, J., Watson, M., 2006. A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics* 135, 499-526.

[68] Mark, N.C., 1995. Exchange Rates and Fundamentals: Evidence on Long-Horizon Predictability. *The American Economic Review* 85(1), 201-218.

[69] Meese, R., Rogoff, K., 1983a. Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of International Economics* 14, 3-24.

[70] Meese, R., Rogoff, K., 1983b. The Out-of-Sample Failure of Empirical Exchange Rate Models: Sampling Error or Misspecification. In J.A. Frenkel, (ed.) Exchange Rates and International Macroeconomics. University of Chicago Press.

[71] Menkhoff, L., Sarno, L. Schmeling, M., Schrimpf, A., 2012. Carry Trades and Global Foreign Exchange Volatility. *Journal of Finance* 67(2), 681-718.

[72] Molodtsova, T., Papell, D.H., 2009. Out-of-Sample Exchange Rate Predictability with Taylor Rule Fundamentals. *Journal of International Economics* 77, 167-180.

[73] Morales-Arias, L., Moura, G.V., 2013. Adaptive forecasting of exchange rates with panel data. *International Journal of Forecasting* 29, 493-509.

[74] Newey, W.K., West, K.D., 1987. A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55(3), 703-708.

[75] Nieto, M.R., Ruiz, E., 2016. Frontiers in VaR forecasting and backtesting. *International Journal of Forecasting* 32(2), 475-501.

[76] Pesaran, M.H., Timmermann, A., 1992. A Simple Nonparametric Test of Predictive Performance. *Journal of Business and Economic Statistics* 10(4), 461-465.

[77] Pesaran, M.H., Timmermann, A., 2005. Small Sample Properties of Forecasts from Autoregressive Models under Structural Breaks. *Journal of Econometrics* 129, 183-217.

[78] Pesaran, M.H., Timmermann, A., 2009. Testing Dependence Among Serially Correlated Multicategory Variables. *Journal of the American Statistical Association* 104(485), 325-337.

[79] Rossi, B., 2013a. Exchange Rate Predictability. CEPR Discussion Papers no. 9575.

[80] Rossi, B., 2013b. Advances in Forecasting under Instability. Handbook of Economic Forecasting, volume 2B, chapter 21, 1203-1324.

[81] Saliby, E., 1989. Repensando a simulação: a Amostragem Descritiva. Ed. Atlas.

[82] Sarno, L., Valente, G., 2005. Empirical exchange rate models and currency risk: some evidence from density forecasts. *Journal of International Money and Finance* 24(2), 363-385.

[83] Shimko, D., 1993. Bounds of Probability. RISK 6, 33-37.

[84] Verdelhan, A., 2013. The Share of Systematic Risk in Bilateral Exchange Rates. MIT mimeo.

[85] Vincent-Humphreys, R., Noss, J., 2012. Estimating probability distributions of future asset prices: empirical transformations from option-implied risk-neutral to real-world density functions. Working Paper 455. Bank of England.

[86] Wang, J., Wu, J., 2012. The Taylor Rule and Forecast Intervals for Exchange Rates. *Journal of Money, Credit and Banking* 44, 103–144.

[87] West, K.D., 1996. Asymptotic Inference About Predictive Ability. *Econometrica* 64, 1067-1084.

[88] Wu, T.Y., 2008. Order Flow in the South: Anatomy of the Brazilian FX Market. Department of Economics, University of California, Santa Cruz. Mimeo.

Appendix

A - Additional details on model 2

The first step follows Shimko (1993) which proposes a nonparametric technique for extracting RND from option prices based on the construction of an implied volatility curve for the option via interpolation of its strike prices (smile volatility curve). Shimko's method was developed for stock option prices and we adapted it for exchange rates, by using the Black Model for pricing future price options (Black, 1976).[41] Breeden and Litzenberger (1978) derived an explicit relationship between the risk-neutral density of an asset and the price of the option on that asset, as follows:

$$\frac{\partial^2 C_t}{\partial K_t^2} = e^{-r_t T} f(s_t), \tag{8}$$

in which C_t is the (call) option price of an underlying asset s_t, K_t is the respective exercise (strike) price of the referred option, r_t is the risk-free interest rate, T denotes the number of days to maturity, and $f(s_t)$ is the risk-neutral probability density

[41] If the underlying asset of the future contract is the exchange rate, the Black Model becomes equivalent to the Garman-Kohlagen Model for pricing exchange rate options.

(RND) of the underlying asset s_t. Shimko obtained the densities from this formula by interpolating the calculated implicit volatilities for the same maturity and different exercise prices. To do so, one must generate an entire continuum of values for the relation of the option price *versus* its exercise price, given that only a few points of this curve are indeed known.[42]

The second step follows Vincent-Humphreys and Noss (2012). Instead of the commonly used method of applying utility-function transformations to the RND, these authors propose an empirical and less restrictive methodology by using a Beta distribution function to calibrate the difference between RND and RWD. According to the authors, although the Beta distribution is parsimonious, as it depends on only two parameters, it nests many simple forms of transformation, such as mean shift, mean-preserving changes in variance and changes involving mean, variance and skewness.

B - Further details on forecast evaluation

B1 - Point forecast

Diebold and Mariano (1995) propose a test for non-nested models comparison that allows for a wide variety of forecast accuracy measures and relies on assumptions made directly on the forecast error loss differential (e.g. the loss differential be covariance stationary). Let $L(e_t)$ be the loss function associated with forecast error e_t. For example, a quadratic loss would be $L(e_t) = (e_t)^2$. The time-t loss differential between forecasts 1 and 2 is then $d_{12t} = L(e_{1t}) - L(e_{2t})$. The null hypothesis of equal predictive accuracy corresponds to $\mathbb{E}(d_{12t}) = 0$, in which the DM test statistic asymptotically converges to a standard normal distribution:

$$DM = \frac{\overline{d}_{12}}{\widehat{\sigma}_{\overline{d}_{12}}} \xrightarrow{d} N(0;1) \tag{9}$$

where $\overline{d}_{12} = T^{-1} \sum_{t=1}^{T} d_{12t}$ is the sample mean loss differential and $\widehat{\sigma}_{\overline{d}_{12}}$ is a consistent estimate of the standard deviation of $\overline{d}_{12}$. **West (1996)** allows for estimation uncertainty within this setup (non-nested models only) and provides conditions for the t-type statistics be asymptotically distributed as $N(0;1)$.

Giacomini and White (2006) propose a framework for out-of-sample predictive ability testing and forecast selection (nested or non-nested models). The GW asymptotics provide a way of testing whether forecasts are equally accurate, in which coefficients include parameter estimation error, but applies only under a rolling window estimation scheme. Besides, the null hypothesis of the GW test is conditional on the forecasts (constructed using estimated parameters), rather than unconditional. Suppose one wants to compare the accuracy of two competing forecasts $f_t(\beta_1)$ and $g_t(\beta_2)$ for the h-step ahead variable Y_{t+h} using a loss function

[42] In this sense, Shimko proposes a quadratic interpolation of the implied volatilities associated with each existing exercise price. From this new curve of implied volatilities, the continuum of values for the option price is obtained, allowing the calculation of second derivatives and, thus, the respective densities. In this paper, risk neutral densities for future exchange rates were generated only for one, two and three months ahead ($h = 1, 2, 3$), due to the low liquidity of exchange rate-based options (and the lack of available data) for longer maturities.

$L_{t+h}(\cdot)$. The null hypothesis of the previous test (DMW) is the following

$$DMW \quad Ho : \mathbb{E}\left[L_{t+h}\left(Y_{t+h}, f_t\left(\beta_1^*\right)\right) - L_{t+h}\left(Y_{t+h}, g_t\left(\beta_2^*\right)\right)\right] = 0, \tag{10}$$

where β_1^* and β_2^* are population values. The DMW null is a statement about the *forecasting models* (i.e. the two models are equally accurate on average). According to Giacomini and White (2006), a key feature of West's (1996) test of Ho is the recognition and accommodation of the fact that, although Ho concerns population values, the actual forecasts that appear in the test statistic depend on estimated parameters. The central idea of the GW test is to consider a null hypothesis that differs from the DMW test in two aspects: (i) the losses depend on estimates $\widehat{\beta}_{1t}$ and $\widehat{\beta}_{2t}$, rather than on their probability limits; and (ii) the expectation is conditional on some information set $\mathcal{F}_t$.

$$GW \quad Ho : \mathbb{E}\left[L_{t+h}\left(Y_{t+h}, f_t\left(\widehat{\beta}_{1t}\right)\right) - L_{t+h}\left(Y_{t+h}, g_t\left(\widehat{\beta}_{2t}\right)\right) \mid \mathcal{F}_t\right] = 0. \tag{11}$$

Now, the focus on parameter estimates makes the null a statement about the *forecasting methods*, which includes the models as well as the estimation procedures and the possible choices of estimation window (since two forecasts may use different estimation windows).

Pesaran and Timmermann (1992) develop a non-parametric procedure for testing the accuracy of forecasts when the focus is on the correct prediction of the direction of change in the variable of interest. Let $I(A)$ be an indicator function that takes the value of unity if $A > 0$ and zero otherwise. Suppose one is interested in testing whether a binary variable $x_t = I(X_t)$ is related to another binary variable $y_t = I(Y_t)$ using a sample of observations $(y_1, x_1), ..., (y_T, x_T)$. Now, let $\widehat{P}$ be the so-called hit rate (i.e. the proportion of periods where Y_t and X_t fall in the same category, that is, have the same sign), while $\widehat{P}_*$ is the hit rate expected under the null hypothesis of independence between x_t and y_t. The PT test statistic is given by

$$PT = \frac{\widehat{P} - \widehat{P}_*}{\left[\widehat{V}(\widehat{P}) - \widehat{V}(\widehat{P}_*)\right]^{1/2}}, \tag{12}$$

where $\widehat{P} = T^{-1}\sum_{t=1}^{T} I(Y_t X_t)$; $\widehat{P}_* = \overline{yx} + (1-\overline{y})(1-\overline{x})$; $\widehat{V}(\widehat{P}) = T^{-1}\widehat{P}_*(1-\widehat{P}_*)$; $\widehat{V}(\widehat{P}_*) = T^{-1}\left(2\overline{y}-1\right)^2\overline{x}(1-\overline{x}) + T^{-1}\left(2\overline{x}-1\right)^2\overline{y}(1-\overline{y}) + 4T^{-2}\overline{yx}(1-\overline{y})(1-\overline{x})$; $\overline{y} = T^{-1}\sum_{t=1}^{T} y_t$; $\overline{x} = T^{-1}\sum_{t=1}^{T} x_t$. Under the null hypothesis of independence between x_t and y_t (i.e. x_t has no power in predicting y_t), the PT statistic is asymptotically distributed as a standard normal distribution. Since this setup only covers the case without serial dependence (e.g. $h = 1$), **Pesaran and Timmermann (2009)** extended it, among others, to allow for such dependencies by using an OLS regression of y_t on x_t and an intercept.

B2 - Full-density forecast

Coverage rates

According to Clark (2011, p.336): "*In light of central bank interest in uncertainty surrounding forecasts, confidence intervals, and fan charts, a natural starting point for forecast density evaluation is interval forecasts - that is, coverage rates.*" In this sense, a necessary (but not sufficient) condition for a "good" density model is to produce a conditional density with an adequate coverage rate.[43] The objective here is to check whether the model departures from a given nominal coverage rate (e.g. 70%) appear to be statistically meaningful. In practice, one needs to compute the frequency of observations of Y_{t+h} that have fallen inside the forecast interval. In our case, we adopt the 70% interval band, which leads to a forecast interval based on the conditional quantiles $Q_{m,\tau}(Y_{t+h} \mid \mathcal{F}_t)$ of model m, ranging from quantile level $\underline{\tau} = 0.15$ to $\overline{\tau} = 0.85$. Then, a simple statistical test verifies the equality between the frequency of observations that have fallen in the forecast interval (nominal coverage) and the true coverage. The main drawback is that coverage rates ignore time dependence and cluster behavior.

Probability Integral Transform (PIT)

The coverage rates although providing an initial approach to analyze density models can be viewed as unconditional tests, since they ignore potential cluster behavior (along the sample size) of a given percentile of the estimated density and, thus, do not take into account time dependence. We next investigate the density forecast models based on a broader measure of density calibration: the probability integral transform (PIT). The PIT of the realization of the variable with respect to the density forecast is given by

$$z_{t+h} = \int_{-\infty}^{Y_{t+h}} \widehat{f}_{t+h,t}(u)du \equiv \widehat{F}_{t+h,t}(Y_{t+h}), \tag{13}$$

where $\widehat{F}_{t+h,t}(Y_{t+h})$ is the probability of the variable of interest not exceeding the observed value Y_{t+h}, and $\widehat{f}_{t+h,t}$ is the density forecast of a given model m with forecast horizon h. The main idea is that, under correct model specification for $h = 1$, the PIT yields independent and uniformly distributed random variable. In this case, when the forecast density $\widehat{f}_{t+1,t}$ equals the true density, it follows that $z_{t+1} \sim i.i.d.\ U(0,1)$, where $U(0,1)$ is the uniform distribution over the interval $(0,1)$. However, in the case of $h > 1$, one should no longer expect well-calibrated densities to deliver $i.i.d.$ PITs, due to the serial correlation[44] of the corresponding probability integral transforms.

Berkowitz (2001) develops tests to evaluate the conditional density based on the normality of the normalized forecast errors that have better power than tests

[43] Coverage rates reveal the difference between the probability that realizations fall into the forecasted intervals and the respective nominal coverage.

[44] It is well known that optimal (i.e. MSE loss function) h-step ahead point forecasts lead to forecast errors following a moving-average (MA) process of order $h-1$. According to Clements (2005, p.7): "When the forecast horizon, h, exceeds the frequency at which forecasts are made ... forecasts will overlap in the sense of being made before the realization paired to the previous forecast is known." See also Knüppel (2015) for further details.

based on the uniformity of the PITs. The normalized forecast error is defined as $\widetilde{z}_{t+1} \equiv \Phi^{-1}(z_{t+1})$, where z_{t+1} denotes the PIT of a 1-step ahead forecast error and Φ^{-1} is the inverse of the standard normal distribution. Under the null, it follows that $\widetilde{z}_{t+1} \sim i.i.d.\ N(0,1)$.[45] See Clements (2004), Jore et al. (2010) and Clark (2011) for further details.

Knüppel (2015) proposes a testing approach based on raw moments designed to handle multi-step-ahead densities and the overlapping nature of the PITs. Under the null hypothesis, there is correct calibration of the forecast density *vis-à-vis* the true density; in the sense that the true density shows statistically the same moments compared to the ones from the forecast density. Let the variable of interest be denoted by x_t and u_t be the respective PIT computed from the density forecast of interest. Denote the transformed PIT by $y_t = H(u_t)$, where (for example) $H(u_t) = \Phi^{-1}(u_t)$ would yield standard normally distributed variables y_t. Also let the r-th raw moment of y_t be denoted as $m_r = \mathbb{E}(y_t^r)$ and define the vector $\widehat{\mathbf{D}}_{r_1,r_2,...,r_N} = [\widehat{m}_{r_1} - m_{r_1}; \widehat{m}_{r_2} - m_{r_2}; ...; \widehat{m}_{r_N} - m_{r_N}]'$ as the difference between the N empirical raw moments of interest $(\widehat{m}_{r_1}, \widehat{m}_{r_2}, ..., \widehat{m}_{r_N})$, where $r_1 < r_2 < ... < r_N$, and the corresponding expected raw moments of y_t, where $\widehat{m}_{r_i} = T^{-1}\sum_{t=1}^{T} y_t^{r_i}$ for $i = 1, 2, ..., N$. The test statistic is given by $\widehat{\alpha}_{r_1,r_2,...,r_N} = T\widehat{\mathbf{D}}'_{r_1,r_2,...,r_N}\widehat{\Omega}^{-1}_{r_1,r_2,...,r_N}\widehat{\mathbf{D}}_{r_1,r_2,...,r_N}$, where $\widehat{\Omega}_{r_1,r_2,...,r_N}$ is the long-run covariance matrix of the vector series $d_t = [y_t^{r_1} - m_{r_1}; y_t^{r_2} - m_{r_2}; ...; y_t^{r_N} - m_{r_N}]'$. Assuming the Central Limit Theorem (CLT) holds for d_t, the test statistic $\widehat{\alpha}_{r_1,r_2,...,r_N}$ asymptotically converges to a $\chi^2_{(N)}$ distribution under the null.

Log Predictive Density Score (LPDS)

Another useful indicator of the calibration of density forecasts is given by the log predictive density score (LPDS). This approach allows one to rank the investigated models, for each forecast horizon, according to their log-scores. The LPDS of model m and forecast horizon h is defined in the following way:

$$LPDS_{m,h} = T^{-1}\sum_{t=1}^{T} \ln\left(\widehat{f}^{m}_{t+h,t}\left(Y_{t+h}\right)\right) \tag{14}$$

where $\widehat{f}^{m}_{t+h,t}$ is the density of the variable of interest estimated from model m and based on the information set available at period t. The referred density is evaluated at the observed value Y_{t+h} and (log) averaged along the out-of-sample observations. A higher score implies a better model (see Adolfson et al., 2005).

Amisano and Giacomini (2007) propose a test for comparing the out-of-sample accuracy of competing density forecasts.[46] The authors restrict attention to the logarithmic scoring rule (LPDS) and propose an out-of-sample 'weighted likelihood ratio' test that compares weighted averages of the scores for the competing forecasts.

[45] For h=1 the test statistic (jointly) assumes independence and standard normality for $\widetilde{z}_{t+1}$ and (under the null) it converges to a $\chi^2_{(3)}$ distribution. For h>1, one can adopt a modified version of the test (see Jore et al., 2010) using a two degrees-of-freedom variant (without a test for independence).

[46] The test is valid under general conditions (i.e. forecasts can be based on nested or non-nested parametric models or produced by semiparametric, non-parametric or Bayesian estimation techniques).

For a given weight function $w(\cdot)$ and two conditional density forecasts f and g for Y_{t+1}, let $WLR_{m,t+1} \equiv w(Y_{t+1}^{st}) \left(\ln \left(\widehat{f}_{m,t} \left(Y_{t+1} \right) \right) - \ln \left(\widehat{g}_{m,t} \left(Y_{t+1} \right) \right) \right)$, where $Y_{t+1}^{st} = \left(Y_{t+1} - \widehat{\mu}_{m,t} \right) / \widehat{\sigma}_{m,t}$ is the realization of the variable at time $t+1$, standardized using estimates of the unconditional mean and standard deviation of Y_t, $\widehat{\mu}_{m,t}$ and $\widehat{\sigma}_{m,t}$, computed on the same sample on which the density forecasts are estimated (where m is finite and denotes the most recent observations).

The AG null hypothesis is $Ho : \mathbb{E}\left(WLR_{m,t+1} \right) = 0$. Similar to the GW test, the AG's null hypothesis depends on parameter estimates (rather than on population values). The AG test is also applicable to $h \geq 1$ but (as in GW) models must be estimated with a rolling window estimation scheme. The test statistic is given by $t_{m,n} = \frac{\overline{WLR}_{m,n}}{\widehat{\sigma}_n / \sqrt{n}}$, where $\overline{WLR}_{m,n} = n^{-1} \sum_{t=m}^{T-1} WLR_{m,t+1}$ and $\widehat{\sigma}_n^2$ is a heteroskedasticity and autocorrelation consistent (HAC) estimator of the asymptotic variance $\sigma_n^2 = var\left(\sqrt{n} \overline{WLR}_{m,n} \right)$. Under the null hypothesis, the test statistic asymptotically converges to a standard normal distribution.

B3 - Local-density analysis

Local Forecast Coverage Rate: $LFCR_{m,h,\tau}$ of model m and horizon h at quantile level τ. Similar to the coverage rate, we now compute (for all out-of-sample observations) the percentage of outcomes below a given nominal quantile level τ. Ideally, the empirical $LFCR_{m,h,\tau}$ should be as close as possible to one minus the nominal level τ.

$$\widehat{LFCR}_{m,h,\tau} = T^{-1} \sum_{t=1}^{T} H_{t+h} \tag{15}$$

where $H_{t+h} = \begin{cases} 1 \text{ ; if } Y_{t+h} > \widehat{Q}_{m,\tau}(Y_{t+h} \mid \mathcal{F}_t) \\ 0 \text{ ; if } Y_{t+h} \leq \widehat{Q}_{m,\tau}(Y_{t+h} \mid \mathcal{F}_t) \end{cases}$. The statistical significance of $LFCR_{m,h,\tau} - (1-\tau)$ is checked via the Kupiec (1995) backtest.

Kupiec (1995): It is a nonparametric test (also known as the unconditional coverage test) based on the proportion of violations H_{t+h}, in which the null hypothesis assumes that:

$$Ho : \mathbb{E}(H_{t+h}) = (1-\tau) \tag{16}$$

The probability of observing N violations, in which $Y_{t+h} > \widehat{Q}_{m,\tau}(Y_{t+h} \mid \mathcal{F}_t)$, over a sample size of T is driven by a Binomial distribution. This way, the null can be tested through a standard likelihood ratio (LR) test of the form:

$$LR_{uc} = 2 \ln \left(\frac{\left(\widehat{LFCR}_{m,h,\tau} \right)^N \left(1 - \widehat{LFCR}_{m,h,\tau} \right)^{T-N}}{(1-\tau)^N (\tau)^{T-N}} \right), \tag{17}$$

which follows (under the null) a $\chi^2_{(1)}$.

Christoffersen (1998): The unconditional coverage test does not provide any information about the temporal dependence of observed violations. In this sense, Christoffersen (1998) extends the previous test to incorporate an evaluation of time

independence of the referred violations. To do so, define T_{ij} as the number of days in which a state j occurred in one day, while it was at state i the previous day. The test statistic also depends on π_i, which is defined as the probability of observing a violation, conditional on state i the previous day. The author assumes that the H_{t+h} stochastic process follows a first order Markov sequence. This way, under the null hypothesis of independence it follows that $\pi = \pi_0 = \pi_1 = (T_{01} + T_{11})/T$, and the complementary test statistic can be constructed, as it follows.

$$LR_{ind} = 2 \ln \left(\frac{(1-\pi_0)^{T_{00}} \pi_0^{T_{01}} (1-\pi_1)^{T_{10}} \pi_1^{T_{11}}}{(1-\pi)^{(T_{00}+T_{10})} \pi^{(T_{01}+T_{11})}} \right). \tag{18}$$

The conditional coverage test of Christoffersen (1998) has the following joint statistic of unconditional coverage and independence: $LR_{cc} = LR_{uc} + LR_{ind}$. The joint test statistic LR_{cc} is asymptotically distributed as $\chi^2_{(2)}$.

Value-at-Risk test based on Quantile Regression (VQR test): The previous test has a restrictive feature, since it only takes into account the autocorrelation of order 1 in the violation sequence. Moreover, it clearly ignores the magnitude of violations $\left| Y_{t+h} - \widehat{Q}_{m,\tau}(Y_{t+h} \mid \mathcal{F}_t) \right|$ when comparing the observed figures of Y_{t+h} with the estimated conditional quantile $\widehat{Q}_{m,\tau}(Y_{t+h} \mid \mathcal{F}_t)$. To overcome these features, Gaglianone et al. (2011) propose the VQR test to evaluate the predictive performance of the estimated Value-at-Risk measure $V_{t+h} \equiv \widehat{Q}_{m,\tau}(Y_{t+h} \mid \mathcal{F}_t)$. The VQR test is simply a Wald test based on the following quantile regression:

$$Q_\tau \left(Y_{t+h} \mid \widetilde{\mathcal{F}}_t \right) = \alpha_0(\tau) + \alpha_1(\tau) V_{t+h} \; ; \; \tau \in (0;1) \tag{19}$$

Under the null hypothesis that V_{t+h} is indeed the τ-level conditional quantile of Y_{t+h}, it follows that $V_{t+h} = Q_\tau \left(Y_{t+h} \mid \widetilde{\mathcal{F}}_t \right)$, which can be verified through the joint coefficient test: $Ho : \alpha_0(\tau) = 0$ and $\alpha_1(\tau) = 1$.

C - Further details on quantile regression

Let $\{y_{t+h}\}$ be some stationary univariate time series and assume one is interested in forecasting y_{t+h} given the information available at time t, $\mathcal{F}_t$. We denote the conditional distribution of y_{t+h} given $\mathcal{F}_t$ as $F_{t+h,t}$, and the conditional density as $f_{t+h,t}$. Assume the data generating process (DGP) with conditional mean and variance dynamics is defined as[47]

$$\begin{aligned} y_{t+h} &= X'_{t+h,t}\alpha + \left(X'_{t+h,t}\gamma \right) \eta_{t+h}, \\ \left(\eta_{t+h} | \mathcal{F}_t \right) &\sim i.i.d.\ F_{\eta,h}(0,1), \end{aligned} \tag{20}$$

where $F_{\eta,h}(0,1)$ is some distribution with mean zero and unit variance, which depends on h but does not depend on $\mathcal{F}_t$, $X_{t+h,t} \in \mathcal{F}_t$ is a $k \times 1$ vector of covariates that can be predicted using information available at time t, and α and γ are $k \times 1$ vectors

[47] This class of DGPs is very broad and includes most common volatility processes (e.g. ARCH and stochastic volatility). The important thing to notice is that no parametric structure is placed on $F_{\eta,h}$ and that covariates affect here the location as well as the scale of the distribution.

of parameters.[48] For simplicity (and to avoid the curse-of-dimensionality problem) assume that $X'_{t+h,t} = (1, x_t)$. Thus, it follows that $\alpha = (\alpha_0, \alpha_1)'$ and $\gamma = (\gamma_0, \gamma_1)'$. Based on this DGP, one can identify a family of conditional quantiles $Q_\tau(y_{t+h} \mid \mathcal{F}_t)$, $\tau \in (0,1)$, as follows:

$$Q_\tau(y_{t+h} \mid \mathcal{F}_t) = \theta_0(\tau) + \theta_1(\tau) x_t \tag{21}$$

where $\theta_0(\tau) = (\alpha_0 + \gamma_0\gamma_h)$, $\theta_1(\tau) = (\alpha_1 + \gamma_1\gamma_h)$ and $\gamma_h = F^{-1}_{\eta,h}(\tau)$ for $\tau \in (0,1)$. Equation (21) says that we can identify the conditional quantiles of y_{t+h} through a quantile regression of y_{t+h} on the single covariate x_t and an intercept. The quantile regression estimation[49] within this setup involves the solution to the problem

$$\min_{\{\theta \in R^2\}} \sum_{t=1}^{T} \rho_\tau(y_{t+h} - \theta_0 - \theta_1 x_t), \tag{22}$$

where ρ_τ is defined as in Koenker and Basset (1978) by $\rho_\tau(u) = \begin{cases} \tau u, u \geq 0 \\ (\tau - 1) u, u < 0 \end{cases}$. Thus, conditional on x_t and estimated parameters $\widehat{\theta}$, one can compute forecasts for the conditional quantiles. For instance, the forecast for the conditional median[50] is given by $\widehat{Q}_{\tau=0.5}(y_{t+h} \mid \mathcal{F}_t) = \widehat{\theta}_0(0.5) + \widehat{\theta}_1(0.5) x_t$. Finally, given a family of estimated conditional quantiles $\widehat{Q}_\tau(y_{t+h} \mid \mathcal{F}_t)$, for a grid of k quantile levels $\tau \in [\tau_1, \tau_2, ..., \tau_k]'$, it is straightforward to estimate the conditional density forecast through the formula (see Koenker, 2005):

$$\widehat{f}_{t+h,t} = \frac{(\tau_i - \tau_{i-1})}{\widehat{Q}_{\tau_i}(y_{t+h} \mid \mathcal{F}_t) - \widehat{Q}_{\tau_{i-1}}(y_{t+h} \mid \mathcal{F}_t)}.$$

The conditional densities can alternatively be estimated (for instance) by using the Epanechnikov kernel, which is a weighting function that determines the shape of the bumps. The latter approach is often preferred (especially in short sample sizes) because it generates smoother densities.

[48] See Gaglianone and Lima (2014) for further details.

[49] The quantile regression method is robust in distributional assumptions, a property that is inherited from the robustness of the ordinary sample quantiles. In addition, it is not the magnitude of the dependent variable that matters in quantile regression, but its position relative to the estimated hyperplane. As a result, the estimated coefficients are less sensitive to outlier observations than, for example, the standard OLS estimator. This superiority over OLS estimator is, in fact, common to any M-estimator.

[50] According to Koenker (2005, p.302), integrating the conditional quantile function $Q_\tau(y_{t+h} \mid \mathcal{F}_t)$ over the entire domain $\tau \in (0,1)$ yields the mean of y_{t+h} conditional on $\mathcal{F}_t$. In practice, one can compute the conditional mean as the average of $\widehat{Q}_\tau(y_{t+h} \mid \mathcal{F}_t)$ over a grid of quantile levels τ.