JEL Classification: C12, C53, F31
Keywords: exchange rate, path forecast, prediction region, family-wise prediction error rate (FWPER), simultaneous prediction regions (SPR), family-wise error rate joint prediction regions (FWEJPR)

Constructing Prediction Regions for Exchange Rate Path Forecasts: The Potential of Calibration

Filip OSTRIHOŇ - Institute of Economic Research of Slovak Academy of Sciences, Bratislava, Slovak Republic (filip.ostrihon@savba.sk) *corresponding author*

Boris FIŠERA - Webster Vienna Private University, Vienna, Austria & Institute of Economic Research of Slovak Academy of Sciences, Bratislava, Slovak Republic & Institute of Economic Studies, Charles University, Prague, Czech Republic

Abstract

We examine and compare the performance of two novel competing approaches - simultaneous prediction regions and bootstrap joint prediction regions - in constructing uncertainty bands for the consensus path forecasts of the EUR/USD exchange rate. The prediction regions are constructed using actual out-of-sample path-forecast errors computed based on historical EUR/USD exchange rate data. We also explore the potential to improve the simultaneous prediction regions by applying the calibration principle. We use the family-wise prediction error rate to measure the joint accuracy of individual per-period intervals, and the likelihood ratio tests for interval accuracy to assess the conditional coverages. We find that the bootstrap joint prediction regions outperform the simultaneous prediction regions on a small evaluation sample. While calibration can improve the performance of simultaneous prediction regions, additional robustness exercises reveal that bootstrap joint prediction regions are generally more reliable from the perspective of unconditional coverage. On the other hand, neither method properly accounts for the dependence in the EUR/USD exchange rate path forecasts.

1. Introduction

With regards to the plethora of powerful and innovative forecasting tools currently available, some central banks (e.g., Norges Bank, 2020; Sveriges Riksbank, 2020) opted to publish their exchange rate predictions as path forecasts, i.e., strings of successive individual (per-period) forecasts. Nonetheless, generating an exchange rate forecast itself remains notoriously difficult since it is not uncommon that prediction models based on economic fundamentals are outperformed by a random walk model (see, e.g., Ca'Zorzi et al., 2022; Hungnes, 2023; Meese and Rogoff,

https://doi.org/10.32065/CJEF.2024.04.03

The authors are grateful to Marián Vávra for proposing Loh's calibration principle as an approach for improving the coverage of simultaneous prediction region intervals, as well as to Štefan Lyócsa, Mária Širaňová, the participants of MIER 2022 and MIER 2023 conferences, and the two anonymous reviewers for helpful comments. Filip Ostrihoň acknowledges that this paper is the partial result of project VEGA 1/0476/21 *(Bootstrap Based Empirical Joint Prediction Regions for Path Forecasts)* and both authors appreciate the support from the grant APVV-20-0499 *(Follow the Money - Deciphering the Link between Shadow Banking Sector and the Illicit Financial Flows)*. The authors are also grateful to Refinitiv Eikon for providing the data on consensus exchange rate forecasts.

1983). However, Novotný and Raková (2011) report that consensus forecasts beat the random walk model in prediction accuracy for the EUR/USD exchange rate. Nevertheless, obtaining uncertainty (prediction) bands or regions for path forecasts, especially when derived from consensus forecasts, is remarkably less straightforward than in the case of a single per-period model-based predictions. Therefore, in this paper, we test the performance of two novel approaches for constructing such prediction regions for the EUR/USD exchange rate consensus forecasts: (i) family-wise error rate joint prediction regions (FWEJPRs) of Wolf and Wunderli (2015); and (ii) simultaneous prediction regions (SPRs) of Jordà et al. (2013).

Essentially, we revisit the "horse race" between the alternative approaches of FWEJPR and SPR already provided in Wolf and Wunderli (2015). However, while Wolf and Wunderli (2015) conduct their exercise for economic growth, which is a slow-moving macroeconomic variable, we use the exchange rate, which is a fast-moving financial market variable, as our forecasted variable. Additionally, instead of using a model generated path forecast, we use consensus forecasts as our path forecast for which we then generate the prediction regions using the FWEJPR and SPR approaches. The consensus forecasts of the EUR/USD exchange rate might serve as an interesting case for comparing these two competing approaches for several reasons: i) Path forecasts of exchange rates are commonly used by both central banks and financial market participants, and thus a practical application from this field could provide additional guidance on constructing prediction regions in such cases; ii) There is a considerable body of evidence on the underperformance of economic models in predicting exchange rates when compared to a simple random walk model (see, e.g., Hungnes, 2023), which implies potential non-stationarity and high temporal dependence of underlying data generating processes (DGPs) - making it interesting to study the differences between conditional and unconditional coverage in the case of exchange rate prediction regions; iii) The exchange rate consensus forecast are generated outside of our evaluation exercise set up, which could mitigate the effects of forecast model selection on the obtained results; iv) Christoffersen's LR tests, which we use to distinguish between conditional and unconditional coverage of prediction regions, are well-suited for application in the context of exchange rate prediction bands (see, e.g., Reeves, 2005; Lee and Scholtes, 2014).[2]

By performing another "horse race" between FWEJPRs and SPRs using the EUR/USD exchange rate consensus forecasts, we are able to delve deeper into the aspects of conditional coverage[3] of these prediction regions, which may be different from the unconditional coverage evaluated by Wolf and Wunderli (2015). Our paper thus extends the analysis of the aforementioned authors by i) adding the perspective of conditional coverage; ii) focusing on exchange rate path forecasts – with exchange rate being a more forward-looking and more volatile variable than GDP, which was previously used for the assessment of FWEJPRs and SPRs; and iii) using the pre-existing consensus forecasts, which require a model-free approach relying on actual (realized) forecast errors instead of model-based (expected) forecast errors –

[2] In fact, the first real-life illustration provided by Christoffersen (1998) was for interval forecasts of daily exchange rates of major currencies vis-à-vis the US dollar.
[3] The coverage of a prediction band conditional on its actual coverage in previous periods.

providing a more realistic outlook on the forecast performance.[4] Given the previous results of Wolf and Wunderli (2015), the SPR approach is clearly the "underdog" in this "horse race". Therefore, we also augment the analysis of Wolf and Wunderli (2015) by examining the potential to improve the SPR prediction regions by applying the calibration principle: We generate calibrated variants of SPRs using Loh's (1987) principle. Additionally, we also report the results for the test size and power in a smaller sample and case-specific set up of Monte Carlo simulations for Christoffersen's LR tests.

For a small evaluation sample of 40 consensus path forecasts of the EUR/USD exchange rate, we find that the FWEJPRs clearly outperform uncalibrated SPRs – confirming the findings of Wolf and Wunderli (2015). Calibration improves the performance of SPRs, although only the per-period stable SPR variants are able to match the performance of FWEJPRs. In this sense, the results indicate that the calibration provides a metaphoric "head start" for the per-period stable SPRs in our reenactment of the path-forecast prediction region "horse race". However, the robustness checks indicate that the overall ranking of different prediction regions is sensitive to the size of the evaluation sample and the specific characteristics of the pool of observations used for the assessment. Therefore, in general, the "winner" of the "horse race" depends on particular settings of the exercise used for the assessment of prediction regions. The robustness checks, nevertheless, do point out that the FWEJPRs are generally more reliable than the SPRs, with only the calibrated SPR-S (F) variant being able to rival them. Furthermore, robustness tests confirm that neither FWEJPRs nor SPRs (calibrated or uncalibrated) can consistently provide proper conditional coverage – opening an interesting avenue for future research.

The rest of the paper is organized as follows: Section 2 reviews the literature on exchange rate forecasting. Section 3 provides a detailed description of the methodology. Section 4 outlines our dataset, while section 5 reports our main results. Section 6 concludes the paper. Results of additional analyzes, Monte Carlo simulations, as well as of several robustness checks are provided in the Appendix.

2. Literature Review

The exchange rate is one of the most important economic and financial indicators, as exchange rate movements have significant macroeconomic consequences (Bussiere et al., 2020; Bruno and Shin, 2015; Georgiadis et al., 2024; Gopinath et al., 2020). As a result, the empirical literature has devoted much interest in both obtaining reliable forecasts of exchange rates (see, e.g., Beckmann and Schuessler, 2016; Ca'Zorzi and Rubaszek, 2020; Cheung et al., 2005; Cheung et al., 2019; Curran and Velic, 2019; Ferraro et al., 2015), as well as in constructing corresponding prediction bands (see, e.g., Beran and Ocker, 1999; Buncic, 2012; Cai et al., 2012; 2015; Islam and Hossain, 2021; Reeves, 2005; Wang and Wu, 2012; Wu, 2012; Zhang and Wan, 2006).

In spite of this attention, forecasting exchange rates remains difficult,

[4] There are also additional minor differences in the forecast horizon examined, as well as a transition from a time-series set up of observations to a panel set up, due to utilizing realized forecast errors instead of expected forecast errors.

primarily as a result of the nonlinearity and temporal instability of the associated DGP (Cai et al., 2015). The matter is also complicated from the theoretical perspective, as there is a well-known incongruity between economic fundamentals and exchange rates, dubbed the exchange rate disconnect puzzle. A seminal paper by Meese and Rogoff (1983) was among the first to provide evidence that a simple random walk model can match the accuracy of exchange rate forecasts based on economic fundamentals. Subsequent research has continued the investigation of exchange rate models based on economic fundamentals when compared to the simple random walk model (Cheung et al., 2005; Frankel and Rose, 1994; Wang and Wu, 2012; Westerlund and Basher, 2007; Wu, 2012; Hungnes, 2023). The exchange rate series, being a random walk process, poses considerable implications not only for the estimation of point forecasts but also for the construction of respective prediction bands, which was assessed by Lee and Scholtes (2014), who investigated the uncertainty with regard to unit root, as this assumption complicates the selection of suitable method for constructing associated intervals.

The exchange rate predictions provided by most empirical analyses are in the form of single, one to multiple-steps-ahead point forecasts. Corresponding prediction bands, if they are constructed, take, in most cases the form of marginal prediction bands (Lee and Scholtes, 2014), kernel density estimates based on the assumption of normality (Buncic, 2012), semiparametric intervals (Wang and Wu, 2012; Wu, 2012), nonparametric bootstrap intervals (Reeves, 2005; Wu, 2012), nonparametric empirical intervals (Lee and Scholtes, 2014; Wu, 2012), or intervals based on nonparametric quantile regression estimation (Cai et al., 2012). However, multi-period forecast trajectories (which are also referred to as “path forecast”) were also generated in a few cases. These utilized marginal prediction bands (Beran and Ocker, 1999) or fuzzy interval neural learning algorithms (Zhang and Wan, 2006). When considering exchange rate path forecasts, the consensus forecasts of exchange rates, synthesizing outlooks of a considerable number of institutions, should not be discounted, as there is evidence of consensus forecasts outperforming simple random walk forecasts (see, e.g., Novotný and Raková, 2011; Ince and Molodtsova, 2017).

In this regard, attention should also be drawn to the issue of the simultaneity of generating the per-period forecasts. This is manifested in the correlation between individual per-period forecasts or corresponding forecast errors, i.e., serial correlation of individual elements of a single path forecast (Jordà et al., 2013). This issue could be addressed by some of the methods mentioned above, which were already employed for the construction of prediction bands in the context of exchange rate forecasts. For instance, the application of bootstrapping, when parametric autoregressive correlated heteroscedasticity model would be used for producing the path-forecast realizations, as was suggested for exchange rate interval forecasts by Reeves (2005) (the method was also applied in a different context by Chudý et al., 2020). However, this approach would not be viable in cases where the DGP for the path forecasts is not known, or there is very little incentive to model it, such as the aforementioned consensus forecasts.

Due to the abovementioned difficulties with forecasting exchange rates, one may argue that available data on consensus forecasts of the EUR/USD exchange rate present an ideal historical "racing track” for the assessment of the performance of various methods for producing path-forecast prediction regions. The most intriguing

adepts in this regard are SPRs, proposed by Jordà et al. (2013) and bootstrap-based FWEJPRs, developed by Wolf and Wunderli (2015). These approaches have a common history in this regard, as they were already pitted against each other by Wolf and Wunderli (2015). For that exercise, the results indicated that the FEWJPRs provide superior empirical out-of-sample coverage for US log real GDP growth compared to SPRs. In our case, the prediction regions are not only applied to a different variable of interest (i.e., the EUR/USD exchange rate), but at the same time, the DGP is unknown, because of which all of the prediction regions are constructed based on realized past forecast errors over identical evaluation sample. To make this "rematch" of the two approaches even more interesting, the potential for improving the performance of SPRs by the application of the calibration principle in the manner suggested by Loh (1987) is also explored.

3. Empirical Methodology

From the technical perspective, the paper relies on methods for constructing prediction regions for arbitrary path forecasts. Specifically, these are the family-wise error rate joint prediction regions (FWEJPRs) proposed by Wolf and Wunderli (2015) and the simultaneous prediction regions (SPRs) developed by Jordà et al. (2013). As the first examined variant of FWEJPR, the two-sided (symmetric) FWEJPRs were computed as:

$$\left[Y_\tau(1) \pm d_{|\cdot|,1-\alpha}^{k-\max.*} \cdot \hat{\sigma}_\tau(1)\right] \times \cdots \times \left[Y_\tau(H) \pm d_{|\cdot|,1-\alpha}^{k-\max.*} \cdot \hat{\sigma}_\tau(H)\right] \tag{1}$$

On the other hand, the asymmetric family-wise error rate joint prediction regions (FWEJPR-A) were computed as an intersection of one-sided lower JPR and one-sided upper JPR proposed by Wolf and Wunderli (2015):

$$\begin{aligned}\left[Y_\tau(1) - d_{1-\alpha}^{k-\max.*} \cdot \hat{\sigma}_\tau(1), Y_\tau(1) - d_{1-\alpha}^{k-\min.*} \cdot \hat{\sigma}_\tau(1)\right] \times \cdots \\ \times \left[Y_\tau(H) - d_{1-\alpha}^{k-\max.*} \cdot \hat{\sigma}_\tau(H), Y_\tau(H) - d_{1-\alpha}^{k-\min.*} \cdot \hat{\sigma}_\tau(H)\right]\end{aligned} \tag{2}$$

In equations (1) and (2), the $Y_\tau(h)$ stands for a vector of forecasts of a random variable, d is an empirically determined multiplier[5] which is obtained using a block bootstrap, and $\hat{\sigma_\tau}(h)$ is prediction standard error for h = 1, ..., H at the time τ. The methodology of Wolf and Wunderli (2015) allows for obtaining generalized k-FWEJPR. Nevertheless, in our application, only the JPRs for k = 1 were computed, as these should correspond to the SPRs suggested by Jordà et al. (2013).

In the application presented by Wolf and Wunderli (2015), a bootstrap is used to generate a single path forecast based on modelled DGP. However, since the DGP is unknown in our case and the evaluation sample is larger than the one path forecast, we deviate from the approach of Wolf and Wunderli (2015) by using the block bootstrap to generate bootstraps of the evaluation sample, which then serve for the computation of standardized forecast errors. Subsequently, all standardized forecast

[5] For exact distinctions among specific multipliers $d_{|\cdot|}^{k\text{-max.}*}$, $d^{k\text{-max.}*}$, and $d^{k\text{-min.}*}$ please see Wolf and Wunderli (2015).

errors obtained through block bootstrap replications are ordered, to determine the corresponding quantile. The block bootstrap was set to 1000 replications in the main exercise with block lengths of 6, 12, and 24 path forecasts.

Their main competitors, the SPRs, were dubbed when they were presented by their creators, Jordà et al. (2013), as "Scheffé bands" due to being obtained by a version of Scheffé's projection. To avoid any confusion in their authorship, we refer to these bands simply as "SPRs". Jordà et al. (2013) also distinguish between general per-period SPRs and per-period stable SPRs (further designated by abbreviations "SPR-Gs" and "SPR-Ss", respectively). As presented by Jordà et al. (2013), SPR-Gs were computed as follows:

$$Y_{\tau,H} = Y_{\tau}(H) \pm |\mathbf{Q}| \sqrt{\frac{\delta^2}{H}} \mathbf{i}_H \tag{3}$$

while SPR-Ss were obtained using the formula:

$$Y_{\tau,H} = Y_{\tau}(H) \pm |\mathbf{Q}| \left[\sqrt{\frac{\delta_h^2}{h}} \right]_{h=1}^{H} \tag{4}$$

In equation (3) and (4), $Y_{\tau,H}$ stands for $H \times 1$ vector of realizations of random variable in 1-to-H steps ahead at time τ, $Y_{\tau}(H)$ for vector of forecasts of random variable in 1-to-H steps ahead at time τ, $\mathbf{Q}$ for lower triangular matrix obtained from the Cholesky decomposition of forecast-error covariance matrix, δ^2 and δ_h^2 stand for critical value and critical value calculated for forecast horizon $h = 1, ..., H$, respectively.

Another distinction in the methodologies suggested by Jordà et al. (2013) is whether the empirical distribution of forecast errors was used for the construction of prediction regions or if somewhat conventional theoretical distributions were employed instead.

With regards to the empirical distribution, the authors propose two approaches for obtaining the Mahalanobis distance depending on the assumption of forecast unbiasedness (Jordà et al., 2013). As the first option, forecast-error Mahalanobis distances based on orthogonalized path-forecast errors are used for the computation of SPRs based on the empirical distribution (further distinguished with the suffix "Emp I"):

$$W_{\tau} = \sqrt{\left(U_{\tau}(H) - \bar{U}(H)\right)' \Omega_H^{-1} \left(U_{\tau}(H) - \bar{U}(H)\right)'} \tag{5}$$

In equation (5), W_{τ} stands for the forecast-error Mahalanobis distance at time τ, $U_{\tau}(H)$ for the vector of forecasts error path in 1-to-H steps ahead at time τ, $\bar{U}(H)$ for the vector of sample means of forecasts errors in 1-to-H steps, and Ω_H is $H \times H$ forecast error covariance matrix.

Alternatively, ordinary path-forecast errors are used for the computation of SPRs based on the empirical distribution (henceforth designated with the suffix "Emp II"):

$$W_\tau = \sqrt{\left(Y_\tau(H) - Y_{\tau,H}\right)' \Omega_H^{-1} \left(Y_\tau(H) - Y_{\tau,H}\right)'} \quad (6)$$

As a theoretical analogue, the authors suggested the application of standard chi-square or F distribution. All of the abovementioned variations in the methodology developed by Jordà et al. (2013) were examined in our analysis.

As an avenue to potentially improve the coverage of SPRs in terms of the family-wise prediction error rate (FWPER), the calibration approach proposed by Loh (1987) was explored. The author suggested a rather straightforward calibration of normal-theory intervals using a procedure of one-step calibration plus a linear interpolation:

$$\gamma_1 = \begin{cases} {\gamma_0}^2 \hat{\gamma}_n^{-1}, & \hat{\gamma}_n \geq \gamma_0 \\ \gamma_0 + (1-\gamma_0)(\gamma_0 - \hat{\gamma}_n)(1-\hat{\gamma}_n)^{-1}, & \hat{\gamma}_n < \gamma_0 \end{cases} \quad (7)$$

In equation (7), γ_0 is the desired coverage, $\hat{\gamma_n}$ is the computed actual coverage, and γ_1 is the corresponding calibrated coverage.

According to Loh (1987), the methodology can be used for any interval based on an estimate and its standard error. Although the examined SPRs are not explicitly constructed based on standard errors of predictions, they still utilize the Cholesky decomposition of the forecast error covariance matrix (Jordà et al., 2013).

Following the example of Jordà et al. (2013) for Greenbook forecasts, marginal prediction bands (see, e.g., Ravishanker et al. 1991) based on the assumption of normality were used as a benchmark for comparison. Similarly, marginal bands constructed using Bonferroni's adjustment (Lehmann and Romano, 2005), which are hereafter referred to as Bonferroni bands, were generated to provide additional benchmark since they are meant to also accommodate the desired proportion of realization trajectories simultaneously contained by the prediction region.

To statistically test the coverage of the examined prediction regions, the methodology proposed by Christoffersen (1998) for evaluating the unconditional and conditional coverage of arbitrary prediction bands was employed. Christoffersen (1998, p. 842) states as part of the motivation for his tests, the desire to distinguish an interval[6] forecast, which neglects the underlying dynamics as it "may be correct on average (have correct unconditional coverage), but in any given period it will have incorrect conditional coverage characterized by clustered outliers." The author developed a series of likelihood ratio (LR) tests, which sequentially test the unconditional validity of the nominal coverage rate (LR_{UC}), the independence assumption (LR_{IN}), and, subsequently, the joint independence and coverage hypothesis, further referred to as the conditional coverage (LR_{CC}). All of the tests carried out in the analysis were performed at a 5 per cent significance level, i.e., using 5 per cent critical values of the corresponding chi-square distribution the tests

[6] Although Christoffersen (1998) used the term "interval", this was due to him primarily examining one-step-ahead (per-period) forecasts. However, since we consider "interval" to have only one dimension, we relegate this term to describing a single per-period interval forecast and otherwise prefer to use the terms "band" or "region" instead.

should converge to. Since the units of a particular test, as well as the corresponding critical value, do not change with the variant of the prediction region examined, we graphically distinguish the insignificant results with a single grey band in the figures in the Appendix.

The LR tests described above were originally applied by Christoffersen (1998) to evaluate the conditional coverage of the exchange rate prediction intervals with a forecast horizon of one period. In order to adjust their implementation for the purposes of our analysis, the metric for assessing band coverage examined throughout the paper is in line with Jordà et al. (2013) the FWPER. This metric considers a realization to be covered by path-forecast bands if all of the elements of the realization throughout the entire trajectory are within corresponding per-period intervals. Put differently, the entire realization is deemed not to be covered by the prediction region if at least one of its elements is outside of the associated per-period interval.

Owing to the reliance on the FWPER, the assessment is based on a sequence of prediction region coverage indicator variables, like the one assumed by Christoffersen (1998) for one-step-ahead interval forecasts. As a result, our analysis utilizes the LR tests only to explore the dependence between the path forecasts, while the dependence within path forecasts plays a role solely during the construction of prediction regions. The dependence in the coverage between consecutive prediction regions, as a potential result of a partial overlap of path forecasts, is, thus, meant to be empirically tested by the corresponding LR_{IN} test. Following the original application of Christoffersen (1998), all of the considered prediction bands were computed for a range of desired nominal coverage levels from 0.5 to 0.95 using an increment of 0.01, with additional descriptive statistics regarding the geometric average width and actual coverage rate of the examined prediction regions for desired nominal coverage levels, is provided in Section 5 in Tab. 2 and 3, respectively.

4. Data

Our empirical analysis is carried out on the basis of pre-existing consensus forecasts of the EUR/USD exchange rate, which were acquired from the Refinitiv Eikon database. These consensus forecasts represent an aggregation of individual predictions produced by various institutions. For the purposes of our analysis, only the mean of consensus forecasts was used as the point estimate of the most likely realization of the future exchange rate. The corresponding statistic was available for forecasts with different horizons produced over the period January 1999 – November 2022, representing in total 287 entries.

In terms of the forecast horizon, 1-month-ahead, 3-months-ahead, 6-months-ahead, and 12-months-ahead forecasts were used. The availability of consecutive future realization forecasts with different time horizons was exploited by rearranging multiple forecasts produced at the same time into path forecasts with a length of 4 elements. Although the actual time interval between two consecutive elements varies throughout the trajectory, there still appears to be evidence of serial correlation of the forecast error as well as of the underlying forecast (see Tab. 1). By pooling all available path forecasts, a panel with 1148 (partially) overlapping individual (per-

period) forecasts is obtained.
In order to obtain the realized forecast errors, the EUR/USD exchange rate consensus forecasts were matched with historical data (i.e., the actual realizations of the EUR/USD exchange rate) by shifting the forecasts into the future based on the corresponding forecast horizon (i.e., the 1-month-ahead forecast was shifted to the end of the following month, 3-months-ahead forecast to the end of the third month from the date of the forecast, etc.). The number of path forecasts used was subsequently diminished when searching for a common sample with historical reference variables.
Historical data of closing daily market positions for the EUR/USD exchange rate reported by Yahoo Finance were used as the reference variable. These were available in daily frequency from 01/12/2003 to 30/11/2023. When constructing the corresponding path-forecast errors, only the closing exchange rates for the last day of the month were used. If the resulting date ended up being a weekend or a holiday, the last historical value reported for the month was inputted for the missing date to provide a reference in terms of the actual realization.
Additionally, average EUR/USD exchange rates published by Eurostat as "Euro/ECU exchange rates - daily data [ERT_BIL_EUR_D]" were obtained. From available data, the average exchange rate with the daily frequency over the period 01/01/1999-30/11/2023 was used as an alternative reference. Applying an analogous approach to the case of closing exchange rates, overlapping path-forecast errors were constructed for average exchange rates as a historical reference.
After these adjustments, the number of actually available observations (presented in Tab. 1 as the "Consensus Forecast of EUR/USD (mean)") is based on 229/287 path forecasts; each consisting of 4 elements with monthly frequency, thus, is in total 916/1148 for historical reference of closing/average EUR/USD exchange rate, described below.

Table 1 Descriptive Statistics

Variable	*N*	*H*	*Obs*	*Mean*	*Std. Dev.*	*Min*	*Max*	*AC-1*
Reference: closing EUR/USD exchange rate								
Historical data	229	4	916	1.244	0.129	0.983	1.576	0.863
Consensus Forecast of EUR/USD (mean)	229	4	916	1.240	0.118	0.971	1.573	0.992
Forecast error	229	4	916	-0.004	0.083	-0.267	0.277	0.694
Reference: average EUR/USD exchange rate								
Historical data	287	4	1148	1.194	0.161	0.842	1.581	0.916
Consensus Forecast of EUR/USD (mean)	287	4	1148	1.194	0.146	0.844	1.573	0.994
Forecast error	287	4	1148	0.000	0.088	-0.272	0.297	0.735

Notes: The table presents descriptive statistics for the number of path forecasts (N), path-forecast horizon (H), no. of observations (Obs), mean (Mean), standard deviation (Std. Dev.), minimum (Min), maximum (Max), and within path-forecast serial correlation of order one (AC-1)

Regarding the utilization of described path-forecast errors, Jordà et al. (2013) used 40 path forecasts as the evaluation sample for their Monte Carlo simulations and for their practical application of the SPRs. We followed this example and also used the evaluation sample of 40 path forecasts. Similarly to Jordà et al. (2013), since we use the FWPER as a measure of coverage, each observation in our setting translates into a single path forecast. Therefore, 40 path forecasts were used as the evaluation sample for the construction of all examined prediction regions as well as for the purposes of calibration. Following Jordà et al. (2013) and Wolf and Wunderli (2015), the rolling window of evaluation samples is in each iteration used to generate a single path-forecast band, i.e., each prediction region is constructed by utilizing only the most recently available data.

To more closely resemble the real-life process of constructing the path-forecast bands, a gap of 11 periods (months) is inserted between the last period of the evaluation sample and the first period of the prediction region. This is to mimic the time lag with which an analyst is able to evaluate 12-month-ahead forecasts. To illustrate the described process, 40 path forecasts starting with the initial path forecast of the period from 31/12/2003 to 30/11/2004 and ending with the path forecast of the period from 31/3/2007 to 29/2/2008 were used to produce the prediction region for path forecast of period from 31/3/2008 to 28/2/2009 after which the evaluation sample of 40 path forecasts was shifted by one path forecast, i.e., the initial path forecast of the evaluation sample was shifted to period from 31/1/2004 to 31/12/2004 and the ending path forecast of the evaluation sample was shifted to period from 30/4/2007 to 31/3/2008, which were used for constructing the prediction region for path forecast of period from 30/4/2008 to 31/3/2009. This process was repeated for 229-40-11=178 path forecasts in the case of closing exchange rate reference variable, which served as individual observations for the statistical analysis of the coverage.

For the alternative historical reference of average exchange rate when the evaluation sample of 40 path forecasts was used, the initial path forecast of the evaluation sample covers periods from 28/2/1999 to 31/1/2000 and the last path forecast of the evaluation sample spanned the period from 31/5/2002 to 30/4/2003, which were used to generate prediction region for path forecast from 31/5/2003 to 30/4/2004. Analogously, there are 287-40-11=236 path forecasts in the case of the average exchange rate that serve as a source of out-of-sample observations, which are available for assessment of the coverage.

However, Lee and Scholtes (2014) recommend for the construction of empirical intervals an evaluation sample with a size of at least 120 observations. This recommendation also appears to match the backtest exercise performed by Wolf and Wunderli (2015), who used 120 periods for their evaluation. Nevertheless, one should be reminded that Wolf and Wunderli (2015) constructed the prediction regions in their exercise using expected forecast errors, which do not translate to 120 path forecasts. Therefore, we performed additional auxiliary analyses on evaluation samples of 80 and 120 path forecasts in an effort to gain insight into the sensitivity of the results to the size of the evaluation sample. Outcomes of these exercises are further analyzed as part of the robustness checks of the main results discussed in section 6.

In any case, the number of available out-of-sample observations used for assessment appears to be at odds with Christoffersen (1998), who suggests at least

2,000 out-of-sample observations based on the results of the power plot performed for the aforementioned LR tests. Nevertheless, since the initial application by Christoffersen (1998), other studies applied identical LR test methodology with a number of out-of-sample observations substantially smaller than the suggested 2,000 observations. For reference, Reeves (2005) used at most 600 out-of-sample observations for the assessment of conditional coverages of exchange rate prediction intervals, while Lee and Scholtes (2014) had between 183 and 317 out-of-sample observations available for their own assessment of the exchange rate prediction intervals based on aforementioned LR tests.

In order to obtain further insight into the empirical small sample properties of these LR tests, we have performed a series of Monte Carlo simulations by generating samples of $T \in \{50, 100, 150, 200, 250\}$ observations. Each generated sample contained a sequence of a binary variable $I \in \{0, 1\}$, simulating the coverage of an arbitrary prediction band. We have controlled each sequence to exhibit preset attributes in terms of empirical coverage $\gamma \in \{0.55, 0.65, 0.75, 0.85, 0.95\}$, order of autocorrelation $p \in \{1, 2, 3, 4\}$, and magnitude of autocorrelation $\rho \in \{0, 0.1, 0.25, 0.5, 0.75\}$, using the following equation, which was inspired by the approach for simulating autocorrelated data presented in a discussion on Statalist - Stata Forum:[7]

$$i_t = 1[d_t \geq \rho]u_t + 1[d_t < \rho]i_{t-p} \tag{8}$$

In equation (8), the i_t stands for the random variable underlying the coverage variable I_t, d_t is a randomly generated dummy variable with uniform distribution $d_t \sim U(0,1)$, and u_t is a randomly generated variable with uniform distribution $d_t \sim U(0,1)$. In order to ensure that the LR tests will be viable for most of the generated samples (i.e., avoid the situation that it is not possible to compute the LR independence test due to missing elements of the associated transition probability matrix), first four observations of i_t in each sample were "seeded" with the following values $i_1=0$, $i_2=1$, $i_3=1$, $i_4=0$.

$$I_t = \begin{cases} 0, & i_t > \gamma \\ 1, & i_t \leq \gamma \end{cases} \tag{9}$$

The equation (9) represents the rule for generating the coverage variable I_t based on the underlying variable i_t and set empirical coverage level γ.

The resulting LR test size / test power statistics were determined using 100,000 repetitions[8] for simulations carried out in the Stata environment while assuming for each setting nominal coverage levels $\gamma_0 \in \{0.55, 0.65, 0.75, 0.85, 0.95\}$. Selected results of these simulations are presented and commented upon in Appendix, part 2.

5. Results

In the next section, we first provide a simple illustration to outline the

[7]For more details, see https://www.statalist.org/forums/forum/general-stata-discussion/general/1480300-simulating-autocorrelated-data-via-monte-carlo.

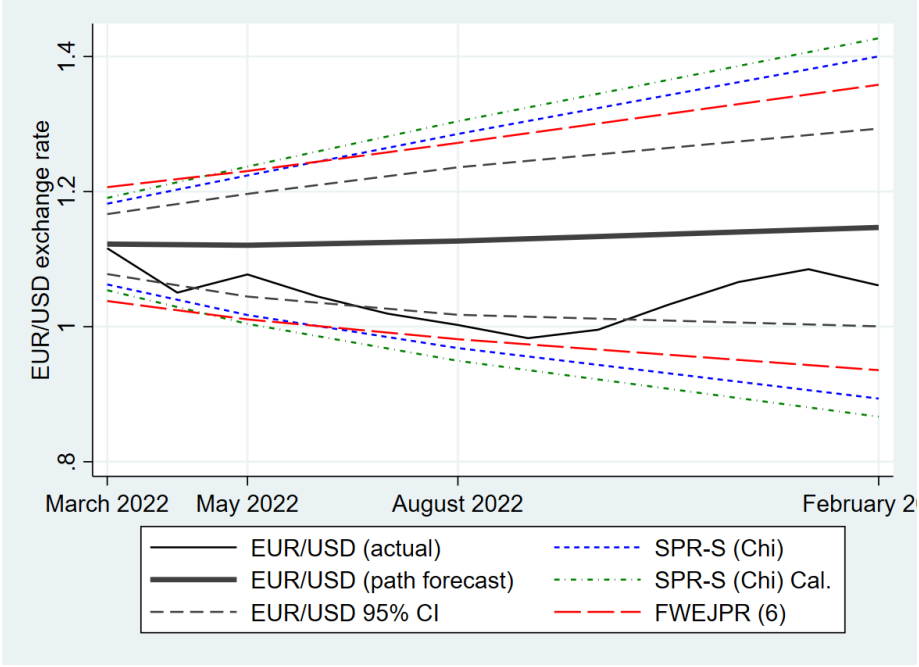
[8]Vast majority of all of these runs were successful. However, some of the repetitions failed to produce desired LR test statistics.

performance of the SPR and FWEJPR prediction regions during the highly uncertain period around the onset of the full-scale Russian invasion of Ukraine. Next, we present our main results on the comparison of the relative performance of SPR and FWEJPR prediction regions. Finally, we report the results of various robustness checks to verify the robustness of our main findings.

5.1 Illustrative Exercise

To provide an illustration of the performance of SPR and FWEJPR prediction regions, we used the two methods to generate out-of-sample prediction regions for the EUR/USD consensus exchange rate path forecast from just before the outbreak of the full-scale Russian invasion of Ukraine in February 2022. Fig. 1 compares the SPR-S (Chi), calibrated SPR-S (Chi) and FWEJPR (6) prediction regions with the actual EUR/USD exchange rate during the forecast period.

Figure 1 EUR/USD Exchange Rate Path Forecast and Prediction Regions in the Aftermath of the Russian Invasion of Ukraine



Notes: The figure depicts the path consensus forecast of the EUR/USD exchange rate from February 2022 for 1-month ahead (March 2022), 3-months ahead (May 2022), 6-months ahead (August 2022), and 12-month ahead (February 2023) periods. For comparison, we also plot the development of the average actual end-of-month EUR/USD exchange rate during the selected time frame. For the list of abbreviations used for individual prediction regions, see the note to Tab. 2 below.

During the highly volatile period in the months following the initiation of the Russian invasion, the US dollar appreciated significantly beyond the consensus forecast. Interestingly, the US dollar appreciation was so substantial that the actual exchange rate appreciated even beyond the level of 2 standard deviations from the consensus forecast (EUR/USD 95% CI) – indicating that during the periods of

increased market volatility, relying on the confidence levels from the consensus path forecasts might not be sufficient.

On the other hand, both the SPR and FWEJPR prediction regions[9] perform well even during volatile periods, as the actual EUR/USD exchange rate remains within both the SPR and FWEJPR prediction regions for the entire year that is covered by the path forecast.

This simple exercise thus illustrates the benefits of relying on prediction regions: Prediction regions provide a more accurate assessment of the uncertainty about the future path of the exchange rate by providing a credible range within which the exchange rate will remain during the entire period being forecasted with priory set likelihood (nominal coverage level).

Furthermore, Fig. 1 indicates that the FWEJPR has more attractive properties than either uncalibrated or calibrated SPRs. Namely, the FWEJPR prediction region turns relatively narrower than the SPR prediction regions - while supposedly maintaining the same nominal coverage rate – thus providing a more nuanced prediction band, which may be valuable for potential users of the forecast (e.g., FX analysts).

5.2 Main Results

We report the actual (unconditional) coverages of SPRs, FWEJPRs, and some commonly used alternative prediction regions[10] for EUR/USD consensus forests in Tab. 2 for different nominal coverage levels.[11] Despite using a different forecasted variable, our results for 90% nominal coverage level are broadly similar to those of Wolf and Wunderli (2015), with our marginal prediction bands and joint prediction regions for block bootstrap of length 6 [FWEJPR (6)] performing rather close to the empirical out-of-sample coverages reported by the aforementioned authors. The performance of our SPRs was somewhat better than that of Wolf and Wunderli (2015), especially in the case of SPR-S (Chi), for which we obtained coverage higher by almost nine percentage points (p.p.) compared to the aforementioned authors.

Another tendency consistent with Wolf and Wunderli (2015) is that even in cases when an SPR is able to almost match the width[12] of an FWEJPR (e.g., SPR-S (Chi) and FWEJPR (6) for 90% nominal coverage level, the measured actual coverage is considerably worse (83.7%[13] to 87.6%, respectively). While the FWEJPR prediction regions generally outperform the SPR prediction regions, both these methods outperform the commonly used alternative methods for generating the

[9] For this illustration, we have selected as the representatives of SPR and FWEJPR prediction regions, (calibrated / uncalibrated) SPR-S (Chi) and FWEJPR (6), respectively. The SPR-S (Chi) was selected due to this variant of SPR also being scrutinized by Wolf and Wunderli (2015) and FWEJPR (6) for performing sufficiently similar to the alternative FWEJPR (12) and FWEJPR (24).

[10] Marginal prediction bands and Bonferroni prediction bands.

[11] Nominal coverage level quantifies the probability that the prediction region contains all actual realizations of a forecasted variable in all of the periods of the path forecast.

[12] The geometric average width of all examined prediction regions is reported in Tab. A1 in App., part 1.

[13] The actual (empirical) coverage quantifies the proportion of future path forecasts (in this case EUR/USD exchange rates) which are captured in the prediction regions. For prediction regions to accurately represent the uncertainty in the forecasts, this proportion should be as close to the desired (nominal) coverage level as possible.

prediction regions: Marginal and Bonferroni prediction bands.
Regarding the rest of examined nominal coverage levels, SPRs in most cases tend to undercover set nominal levels, while FWEJPRs appear to cover more future realizations than desired, especially for nominal levels below 70%. Moreover, for these lower nominal levels, the calibration of SPRs appears to be more valuable as it brings the uncalibrated SPR variants relatively closer to their coverage targets.

Table 2 Actual (Empirical) Coverage of Examined Prediction Regions

Nominal coverage level \ Prediction region	***0.5***	***0.55***	***0.6***	***0.65***	***0.7***	***0.75***	***0.8***	***0.85***	***0.9***	***0.95***
Marg	0.118	0.146	0.197	0.287	0.382	0.466	0.522	0.584	0.646	0.792
Bonf	0.466	0.506	0.522	0.573	0.584	0.618	0.646	0.708	0.792	0.854
SPR-G (Chi)	0.489	0.522	0.573	0.601	0.612	0.640	0.669	0.702	0.758	0.826
SPR-G (Chi) Cal.	0.522	0.551	0.584	0.601	0.624	0.669	0.691	0.770	0.826	0.871
SPR-S (Chi)	0.404	0.438	0.483	0.511	0.601	0.640	0.685	0.747	0.837	0.910
SPR-S (Chi) Cal.	0.522	0.556	0.590	0.607	0.624	0.697	0.742	0.815	0.860	0.910
SPR-G (F)	0.494	0.528	0.590	0.601	0.624	0.646	0.669	0.713	0.775	0.843
SPR-G (F) Cal.	0.528	0.556	0.584	0.601	0.635	0.674	0.730	0.787	0.837	0.899
SPR-S (F)	0.421	0.444	0.489	0.534	0.612	0.646	0.702	0.764	0.848	0.910
SPR-S (F) Cal.	0.522	0.562	0.590	0.607	0.640	0.697	0.753	0.820	0.854	0.910
SPR-G (Emp I)	0.494	0.534	0.551	0.584	0.635	0.640	0.657	0.691	0.770	0.826
SPR-G (Emp I) Cal.	0.506	0.556	0.584	0.612	0.635	0.674	0.725	0.770	0.809	0.860
SPR-G (Emp II)	0.500	0.522	0.579	0.612	0.629	0.652	0.657	0.719	0.775	0.820
SPR-G (Emp II) Cal.	0.528	0.556	0.590	0.601	0.624	0.674	0.725	0.770	0.820	0.865
SPR-S (Emp)	0.388	0.421	0.466	0.506	0.579	0.629	0.685	0.764	0.843	0.899
SPR-S (Emp) Cal.	0.500	0.551	0.590	0.624	0.635	0.697	0.747	0.815	0.860	0.910
FWEJPR (6)	0.601	0.612	0.629	0.669	0.697	0.742	0.809	0.837	0.876	0.938
FWEJPR-A (6)	0.573	0.601	0.629	0.657	0.697	0.753	0.803	0.843	0.882	0.910
FWEJPR (12)	0.607	0.618	0.635	0.669	0.719	0.764	0.826	0.865	0.893	0.949
FWEJPR-A (12)	0.579	0.607	0.652	0.669	0.697	0.753	0.798	0.820	0.876	0.921
FWEJPR (24)	0.607	0.624	0.635	0.657	0.680	0.747	0.781	0.837	0.882	0.910
FWEJPR-A (24)	0.556	0.584	0.624	0.657	0.674	0.708	0.747	0.781	0.826	0.899

Notes: Actual coverage (i.e., the empirical coverage measured by FWPER) obtained for evaluation sample of 40 path forecasts and different nominal coverage levels (i.e., levels of desired coverage) of examined regions is reported. The examined individual prediction regions include: marginal prediction bands [Marg], Bonferroni prediction bands [Bonf], ordinary general per-period simultaneous prediction regions based on the theoretical chi-square distribution [SPR-G (Chi)] and their calibrated counterparts [SPR-G (Chi) Cal.], ordinary per-period stable simultaneous prediction regions based on the theoretical chi-square distribution [SPR-S (Chi)] and their calibrated counterparts [SPR-S (Chi) Cal.], ordinary general per-period simultaneous prediction regions based on the theoretical F distribution [SPR-G (F)] and their calibrated counterparts [SPR-G (F) Cal.], ordinary per-period stable simultaneous prediction regions based on the theoretical F distribution [SPR-S (F)] and their calibrated counterparts [SPR-S (F) Cal.], two variants for computing (for further information on the distinction between them see section 2) ordinary general per-period simultaneous prediction regions based on empirical distribution [SPR-G (Emp I) and SPR-G (Emp II)] and their two calibrated counterparts [SPR-G (Emp I) Cal. and SPR-G (Emp II) Cal.], ordinary general per-period stable simultaneous prediction regions based on the empirical distribution [SPR-S (Emp)] and their calibrated counterparts [SPR-S (Emp) Cal.], family-wise error rate joint prediction regions [FWEJPR] and asymmetric family-wise error rate joint prediction regions [FWEJPR-A], which are computed with block bootstrap of length 6 [FWEJPR (6) and FWEJPR-A (6)], 12 [FWEJPR (12) and FWEJPR-A (12)], and 24 [FWEJPR (24) and FWEJPR-A (24)] path forecasts.

As our analysis aims to go beyond the mere comparison of the obtained actual coverages, we statistically test the difference between the actual and desired coverage with Christoffersen's (1998) LR tests. We report the main findings of this test in Tab. 3. For more detailed results (also featuring more granular nominal levels and other prediction region variants not presented in Tab. 3) of these tests, see Fig. A1-A3 in the Appendix.

The aforementioned tests allow us to statistically test whether the observed difference between actual and desired coverage can be attributed to a mere random error or whether there is a noteworthy shortcoming on the part of the particular prediction region. This distinction might be of interest to the analyst using the forecasts as a statistically significant difference in the unconditional coverage, which indicates that the prediction region either over- or under-estimates the actual uncertainty of the future realization of the forecasted process. In any case, the notion of the uncertainty held by the analyst would be biased, which is why the analyst should always strive to obtain prediction regions corresponding to the coverage rate set in advance.

This assessment is also further complicated by the potential difference in conditional coverage from unconditional coverage, which should be identical only when the coverage is independent across time. However, if the independence assumption is not met, the period-specific coverage obtained can deviate from the general measure of unconditional coverage and, thus, in time, result in lower reliability of the prediction region.

Table 3 The p-values of LR Tests of Selected Prediction Regions Based on an Evaluation Sample of 40 Path Forecasts

Nominal coverage level \ ***Prediction region***	***0.5***	***0.55***	***0.6***	***0.65***	***0.7***	***0.75***	***0.8***	***0.85***	***0.9***	***0.95***
Marg: UC	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Marg: IN	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Bonf: UC	0.368	0.235	**0.036**	**0.034**	**0.001**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Bonf: IN	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
SPR-S (Chi): UC	**0.011**	**0.003**	**0.002**	**0.000**	**0.005**	**0.001**	**0.000**	**0.000**	**0.010**	**0.027**
SPR-S (Chi): IN	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
SPR-S (Chi) Cal.: UC	0.549	0.868	0.783	0.230	**0.029**	0.108	0.059	0.199	0.088	**0.027**
SPR-S (Chi) Cal.: IN	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
FWEJPR (6): UC	**0.007**	0.093	0.424	0.603	0.922	0.796	0.763	0.633	0.310	0.485
FWEJPR (6): IN	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.001**

Notes: The table reports the p-values of LRUC (rows designated with the abbreviation "UC", H0: empirical coverage is equal to nominal coverage), LRIN (rows designated with the abbreviation "IN", H0: coverage is the same regardless of the particular observations assessed) tests for the evaluation sample of 40 path forecasts when closing exchange rate is used as historical reference. Statistically significant results (results of p-value < 0.05) are highlighted in bold. For the list of abbreviations used for individual prediction regions, see the note to Tab. 2.

This may, in fact, be the case for all of the prediction regions presented in Tab. 3. As we may see, the independence assumption is being uniformly rejected by the LR independence test (LR_{IN} abbreviated "IN" in the given table rows), regardless of the unconditional coverage, set nominal level, or any additional calibration. Nevertheless, the tested unconditional coverage starkly differs between the presented SPR-S (Chi) and FWEJPR (6). The results of the LR unconditional coverage test (LR_{UC} abbreviated "UC" in the given table rows) in Tab. 3 indicate that the SPR-S (Chi) fails to achieve the set nominal coverage, in a statistical sense, for the majority of examined nominal levels, while FWEJPR (6) is not able to pass this test only for a minority of examined nominal levels.

The calibration appears to be beneficial for SPR-S (Chi) from the perspective of unconditional coverage, as the number of set nominal levels for which the calibrated SPR-S (Chi) passes the LR unconditional coverage test is able to almost match the results for FWEJPR (6). However, based on the results of the independence test, the conditional coverage test (the results of which are not presented in Tab. 3 but are depicted in Fig. A3 in the Appendix) rejects the notion of actual conditional coverage meeting the set nominal coverage, in a statistical sense for each prediction region variant examined, regardless of set nominal level or whether calibration was applied.

Our results thus confirm the suspicion of Wolf and Wunderli (2015) that the assessment of SPRs and FWEJPRs based on historical data should be mindful of the dependence in future realizations. Furthermore, our results support the claim of the aforementioned authors that they produced a "fair" assessment, given that, in our case, neither of the methods was able to account for the dependence in forecasts and provide valid (adequate) conditional coverage.

5.3 Robustness Checks

In the following sub-section, we describe several robustness checks conducted to verify the stability of our results. We initiated these exercises by utilizing the average EUR/USD exchange rate for the last day of the month instead of the closing EUR/USD exchange rate on the last day of the month as the historical reference for the computation of forecast errors.[14] In this manner, we aim to investigate to what extent are the main results driven by our particular choice of historical reference. The results of these robustness checks indicate that the main conclusions remain stable. However, the calibrated SPR-S (F) and SPR-S (Emp) are able to pass the unconditional coverage test from 2.6 to 34.6 % more frequently than the presented SPR-S (Chi) (and occasionally even more often than competing FWEJPR). Thus, these particular SPR variants might be more appropriate for constructing prediction

[14] We also gave additional attention to the difference in periods for which these two alternative historical references are observed. Specifically, the average EUR/USD exchange rate for the last day of the month is also available for the sub-period from 01/01/1999 to 30/11/2003, for which the closing EUR/USD exchange rate on the last day of the month was not reported by Yahoo Finance. We, therefore, assessed the performance of prediction regions while matching the average EUR/USD exchange rate period with the period of closing EUR/USD exchange rate (which we further referred to as the restricted data span) or all available data are exploited instead (which we further referred to as the unrestricted data span).

regions based on small evaluation samples than SPR-S (Chi).

In the next set of robustness checks, we increased the size of the evaluation sample, i.e., the number of path forecasts used for constructing prediction regions. Instead of 40 path forecasts used for the main results, we subsequently utilized 80 and 120 path forecasts. We carried out this set of checks in order to assess the sensitivity of the main results to the choice of evaluation sample size. Interestingly, with larger evaluation samples, we found that, on average, SPRs satisfy the unconditional coverage likelihood ratio tests more frequently than FWEJPRs. This is particularly the case for the evaluation sample of 120 path forecasts - regardless of the historical reference used - although the contrast is most clear for the average EUR/USD exchange rate while utilizing the unrestricted data span.

Additionally, since the SPRs can provide appropriate unconditional coverage in many instances by themselves, their calibration is scarcely beneficial, and in most cases, it only leads to an increase in the frequency of not passing the unconditional coverage test. Occasionally, some instances of SPRs and FWEJPRs do pass the conditional coverage test. However, such instances are only a small minority of examined cases. Similarly, the results obtained for the evaluation sample of 80 path forecasts also deviate from the main results presented in Tab. 3 to some extent. The exception in this sense are test results achieved with average EUR/USD exchange rate (for a sample of 80 path forecasts based on the unrestricted data span), which, in contrast to those for 120 path forecasts, corroborate the same conclusions that can be drawn based on Tab. 3. We discuss potential reasons why these distinct properties are observed later in this section.

To gain further insight into the observed differences between the main results and the robustness checks, the data used for the LR test assessment were sequentially evaluated with a moving window limited to 50[15] observations at a time. Each window of observations was used for the computation of the corresponding empirical (actual) coverages. The variation in obtained actual coverages was then separately evaluated via normalized root mean square error (NRMSE) statistic for each nominal (desired) coverage level.

The NRMSE exercise (the results for closing EUR/USD exchange rate are reported in Tab. A2 – A4 in the appendix, with alternative results for average EUR/USD exchange rate also being available from the authors upon request) has shown that the error in actual coverage is not uniformly distributed across nominal levels, nor is it stable with respect to the size of the evaluation sample used for the prediction regions construction. In general, the prediction regions are far less deviating from the desired nominal coverage when higher levels are set, e.g., 95%, 90%, 85%, than the NRMSE for the lowest examined nominal levels, e.g., 50%, 55%, 60%. Additionally, the size of the evaluation sample used for the construction of prediction regions may aid in decreasing the error, although such reduction is rather sensitive to the choice of prediction region variant.

This finding indicates that the differences in the results discussed above are mostly driven by the specific features of the period recorded for the average

[15] The number of 50 observations was selected because this was the lowest number of observations for which we performed simulations in terms of LR test power. For more details, see the Appendix.

EUR/USD exchange rate but not available for the closing EUR/USD exchange rate, namely the time span from 01/01/1999 to 30/11/2003. Although this period might not be assessed from the perspective of prediction region coverage per se, its availability allows for other subsequent periods to be featured in the assessment and affects the width of regions for corresponding path forecasts.

The calibration of the SPR appears to mitigate the differences in coverage error caused by the size of the evaluation sample used for the construction of prediction regions. However, it does not alleviate the coverage errors stemming from the choice of the set nominal level. Put differently, the calibration can help to produce more consistent empirical coverage regardless of the size of the evaluation sample, although that may manifest as a worsening of the performance of some of the uncalibrated SPRs for lower nominal coverage levels (e.g., 50%, 55%, 60%) in case of certain evaluation sample sizes.

Regarding the sensitivity of the variants SPR-S (Chi) and FWEJPR (6) presented in Tab. 3, the NRMSE exercise has shown that FWEJPR (6) appears to be comparably more stable at higher nominal coverage levels (e.g., 95%, 90%, 85%) for smaller evaluation samples. The calibration reduces such errors, although, in terms of sensitivity, the calibrated SPR-S (Chi) still underperforms the FWEJPR (6). From this perspective, there are no signs of any of the explored SPRs being able to compete with the FWEJPR (6), and the results are rather robust in this sense.

If all nominal coverage levels and evaluation sample sizes are taken into account, then the stability of FWEJPR (6) is on par with that of FWEJPR (24) and calibrated SPR-S (F) while being outmatched by that of FWEJPR (12) for evaluation sample of 40 path forecasts; surpassing that of calibrated SPR-S (F), equaling that of FWEJPR (12), and falling behind that of FWEJPR (24) for evaluation sample of 80 path forecasts; and outperforming that of FWEJPR (12) and that of FWEJPR (24), while underperforming to that of calibrated SPR-S (F) for evaluation sample of 120 path forecasts. All of the mentioned variants can be, therefore, considered as the best performing, depending on the evaluation sample size used. Concerning the coverage independence and associated conditional coverage, the main results presented in Tab. 3 appear to be robust both when the average EUR/USD exchange rate is used instead of the closing exchange rate as well as when larger evaluation samples are used.

The understanding of such occurrences may be further expanded by taking into account the insight from the aforementioned NRMSE exercise with a moving window of assessed observations. Since the error in the actual coverage tends to increase with lower nominal levels, one may expect that any temporal dependence is more discernable at these coverage levels. Therefore, the independence LR test may be far more likely to pick up any underlying dependence in the EUR/USD exchange rate path forecast at lower nominal levels than at higher nominal levels.

We, therefore, conclude that the NRMSE descriptive statistics support the notion of general non-independent coverage of all examined prediction region variants for the EUR/USD exchange rate path, which further corroborates the robustness of our main results. However, the remarkable consistency of some of the prediction region variants at higher nominal levels may produce independent coverage, at least in a statistical sense (i.e., from the perspective of LR independence test).

6. Conclusions

In this paper, we have followed Wolf and Wunderli (2015) and conducted a second round of prediction band "horse race", this time based on the consensus EUR/USD exchange rate forecasts from the Eikon database. As in the case of the aforementioned authors, the contenders were again two novel approaches to constructing prediction regions for path forecasts: The rather analytical SPRs proposed by Jordà et al. (2013) and the bootstrap-based FWEJPRs introduced by Wolf and Wunderli (2015). To provide a statistical assessment of this additional run of horse race, we have adapted the LR tests of Christoffersen (1998) to the utilization of FWPER as a path forecast performance measure to allow for the comparison of the two competing methods also from the perspective of accounting for the dependence in the path-forecast band coverage. In addition to examining the performance of the two methods in terms of producing 90% nominal coverage bands, which was previously investigated by Wolf and Wunderli (2015), we also extend our analysis to the range of nominal coverage levels from 55 to 95%

Our main results based on closing EUR/USD exchange rates with an evaluation sample of 40 path forecasts confirm the findings of Wolf and Wunderli (2015) and overwhelmingly favor FWEJPRs compared to the alternative SPRs. The results also show that the calibration does provide a head start for this second round of horse races, which SPRs, in case of a small evaluation sample of EUR/USD exchange rate path forecasts, desperately need. Nevertheless, even after the calibration, only the per-period stable variants – i.e., SPR-S (Chi), SPR-S (F), and SPR-S (Emp) – are feasibly able to compete with alternative FWEJPRs from the perspective of width, empirical coverage, and satisfaction of the Christoffersen's unconditional coverage test. However, the described outcome is not robust for larger evaluation samples, and both the superiority of FWEJPRs and the potential benefits of calibrating SPRs would have been questioned if an evaluation sample of size 80 or 120 path forecasts were used instead. The differences in this qualitative outcome were consequently attributed to the specificities of observations (path forecasts) used for the assessment of prediction region performance via Christoffersen's LR test methodology.

Our second related finding indicates that neither the FWEJPRs nor SPRs are able to reliably accommodate the serial dependence, which is rather inherent to EUR/USD exchange rate path forecasts used for the horse race. This finding was further validated by the NRMSE exercise performed as part of the robustness checks, thus confirming the suspicion of Wolf and Wunderli (2015) that the illustrative exercise they carried out might have disregarded the role of dependence in their empirical coverage assessment. Nevertheless, both SPRs and FWEJPRs appear to be equally inept in dealing with this issue. Thus, our results indicate that the assessment of the aforementioned authors remains valid. This finding also has its practical implications. If conditional coverage is of the highest concern to the analyst aiming to construct prediction regions for EUR/USD exchange rate path forecasts (or path forecast of other highly dependent variables), then neither of the two competing general methods (FWEJPR and SPR – calibrated or uncalibrated) can be recommended for the prediction region construction.

The NRMSE exercise has also shown that despite the FWEJPR not providing

stellar performance for larger evaluation samples examined as part of the robustness checks, its variants are still among the most reliable methods for constructing prediction regions from the perspective of NRMSE regardless of evaluation sample size. Additionally, the SPR-S (F) can provide reliable performance in terms of NRMSE in empirical coverage once calibrated, thus showing that some of the SPRs can rival the FWEJPR when provided a metaphoric head start in our reenactment of the "horse race". However, from a practical perspective, the FWEJPRs tend to be much more consistent. So, if an analyst is deciding on which variant of prediction regions to use for the EUR/USD path forecast regardless of the conditional coverage validity, the FWEJPRs appear as the most reliable option.

Lastly, our paper provides the results of Monte Carlo simulations for small pools of observations used for LR test assessment, which can prove useful for future empirical research utilizing Christoffersen's LR test methodology.

APPENDIX

1. Additional Results

Table A1 Geometric Average Width of Examined Prediction Regions

Nominal coverage level \ *Prediction region*	*0.5*	*0.55*	*0.6*	*0.65*	*0.7*	*0.75*	*0.8*	*0.85*	*0.9*	*0.95*
Marg	0.099	0.110	0.123	0.137	0.152	0.168	0.187	0.210	0.240	0.287
Bonf	0.168	0.177	0.187	0.198	0.210	0.224	0.240	0.260	0.287	0.328
SPR-G (Chi)	0.194	0.204	0.213	0.224	0.234	0.246	0.260	0.276	0.296	0.327
SPR-G (Chi) Cal.	0.194	0.205	0.218	0.231	0.246	0.262	0.281	0.302	0.328	0.369
SPR-S (Chi)	0.163	0.177	0.193	0.209	0.227	0.246	0.269	0.296	0.330	0.383
SPR-S (Chi) Cal.	0.194	0.205	0.218	0.234	0.253	0.273	0.297	0.318	0.345	0.383
SPR-G (F)	0.196	0.206	0.216	0.227	0.239	0.251	0.266	0.284	0.307	0.343
SPR-G (F) Cal.	0.195	0.206	0.220	0.233	0.249	0.267	0.287	0.310	0.338	0.386
SPR-S (F)	0.164	0.179	0.195	0.212	0.230	0.251	0.274	0.302	0.339	0.397
SPR-S (F) Cal.	0.195	0.206	0.219	0.236	0.255	0.276	0.300	0.320	0.347	0.389
SPR-G (Emp I)	0.194	0.203	0.211	0.222	0.233	0.242	0.254	0.270	0.293	0.319
SPR-G (Emp I) Cal.	0.192	0.204	0.218	0.230	0.245	0.262	0.281	0.299	0.319	0.345
SPR-G (Emp II)	0.200	0.208	0.216	0.225	0.234	0.244	0.256	0.273	0.294	0.320
SPR-G (Emp II) Cal.	0.198	0.207	0.218	0.228	0.244	0.263	0.284	0.301	0.320	0.346
SPR-S (Emp)	0.160	0.173	0.189	0.205	0.225	0.245	0.266	0.293	0.326	0.382
SPR-S (Emp) Cal.	0.191	0.207	0.222	0.237	0.252	0.273	0.300	0.316	0.342	0.389
FWEJPR (6)	0.208	0.220	0.232	0.244	0.258	0.272	0.289	0.308	0.332	0.370
FWEJPR-A (6)	0.202	0.213	0.223	0.235	0.249	0.264	0.280	0.299	0.323	0.360
FWEJPR (12)	0.213	0.225	0.238	0.251	0.264	0.279	0.296	0.317	0.343	0.384
FWEJPR-A (12)	0.205	0.215	0.226	0.238	0.251	0.267	0.285	0.306	0.331	0.371
FWEJPR (24)	0.209	0.221	0.233	0.246	0.258	0.270	0.286	0.307	0.332	0.365
FWEJPR-A (24)	0.198	0.208	0.218	0.229	0.240	0.255	0.271	0.291	0.313	0.347

Notes: Geometric average width for evaluation sample of 40 path forecasts and desired nominal coverage levels of following examined prediction regions is reported: marginal prediction bands [Marg], Bonferroni prediction bands [Bonf], ordinary general per-period simultaneous prediction regions based on the theoretical chi-square distribution [SPR-G (Chi)] and their calibrated counterparts [SPR-G (Chi) Cal.], ordinary per-period stable simultaneous prediction regions based on the theoretical chi-square distribution [SPR-S (Chi)] and their calibrated counterparts [SPR-S (Chi) Cal.], ordinary general per-period simultaneous prediction regions based on the theoretical F distribution [SPR-G (F)] and their calibrated counterparts [SPR-G (F) Cal.], ordinary per-period stable simultaneous prediction regions based on the theoretical F distribution [SPR-S (F)] and their calibrated counterparts [SPR-S (F) Cal.], two variants for computing (for further information on the distinction between them see section 2) ordinary general per-period simultaneous prediction regions based on empirical distribution [SPR-G (Emp I) and SPR-G (Emp II)] and their two calibrated counterparts [SPR-G (Emp I) Cal. and SPR-G (Emp II) Cal.], ordinary general per-period stable simultaneous prediction regions based on the empirical distribution [SPR-S (Emp)] and their calibrated counterparts [SPR-S (Emp) Cal.], family-wise error rate joint prediction regions [FWEJPR] and asymmetric family-wise error rate joint prediction regions [FWEJPR-A], which are computed with block bootstrap of length 6 [FWEJPR (6) and FWEJPR-A (6)], 12 [FWEJPR (12) and FWEJPR-A (12)], and 24 [FWEJPR (24) and FWEJPR-A (24)] path forecasts.

Table A2 Results of Normalized Relative Mean Square Error (NRMSE) Exercise for Evaluation Sample of 40 Path Forecasts

Nominal coverage level* \ *Prediction region	***0.5***	***0.55***	***0.6***	***0.65***	***0.7***	***0.75***	***0.8***	***0.85***	***0.9***	***0.95***
Marg	2.399	2.040	1.438	0.864	0.502	0.351	0.300	0.249	0.221	0.108
Bonf	0.215	0.183	0.148	0.146	0.131	0.122	0.125	0.105	0.088	0.065
SPR-G (Chi)	0.235	0.210	0.199	0.157	0.118	0.114	0.109	0.115	0.107	0.098
SPR-G (Chi) Cal.	0.264	0.236	0.194	0.154	0.128	0.102	0.101	0.085	0.073	0.061
SPR-S (Chi)	0.211	0.195	0.177	0.169	0.128	0.114	0.097	0.088	0.079	0.043
SPR-S (Chi) Cal.	0.274	0.238	0.197	0.166	0.130	0.124	0.099	0.079	0.061	0.043
SPR-G (F)	0.242	0.215	0.206	0.157	0.116	0.112	0.109	0.108	0.099	0.083
SPR-G (F) Cal.	0.274	0.236	0.194	0.154	0.129	0.104	0.094	0.086	0.072	0.043
SPR-S (F)	0.206	0.187	0.176	0.160	0.118	0.112	0.093	0.091	0.085	0.043
SPR-S (F) Cal.	0.274	0.245	0.197	0.166	0.133	0.124	0.097	0.083	0.072	0.043
SPR-G (Emp I)	0.243	0.227	0.202	0.167	0.128	0.114	0.119	0.128	0.094	0.092
SPR-G (Emp I) Cal.	0.259	0.230	0.194	0.152	0.126	0.100	0.097	0.070	0.074	0.065
SPR-G (Emp II)	0.248	0.223	0.193	0.166	0.129	0.110	0.125	0.104	0.096	0.096
SPR-G (Emp II) Cal.	0.266	0.240	0.199	0.167	0.130	0.102	0.092	0.076	0.073	0.064
SPR-S (Emp)	0.236	0.231	0.185	0.175	0.136	0.127	0.113	0.097	0.081	0.051
SPR-S (Emp) Cal.	0.235	0.226	0.194	0.177	0.134	0.117	0.087	0.092	0.067	0.043
FWEJPR (6)	0.321	0.267	0.222	0.194	0.163	0.136	0.139	0.100	0.065	0.046
FWEJPR-A (6)	0.331	0.296	0.273	0.238	0.213	0.180	0.145	0.125	0.089	0.054
FWEJPR (12)	0.321	0.268	0.219	0.200	0.179	0.154	0.141	0.107	0.065	0.046
FWEJPR-A (12)	0.359	0.314	0.288	0.249	0.214	0.183	0.148	0.113	0.091	0.054
FWEJPR (24)	0.329	0.270	0.219	0.191	0.161	0.147	0.123	0.097	0.065	0.038
FWEJPR-A (24)	0.338	0.297	0.267	0.236	0.202	0.166	0.131	0.104	0.075	0.058

Notes: Results of normalized relative mean square error (NRMSE) for actual coverage (i.e., the empirical coverage measured by FWPER) obtained for evaluation sample of 40 path forecasts when the closing exchange rate is used as a historical reference, with an average value of actual coverage used for normalization. Different nominal coverage levels (i.e., levels of desired coverage) of examined regions (for the list of abbreviations used for individual prediction regions see note to Tab. A1) is reported.

Table A3 Results of Normalized Relative Mean Square Error (NRMSE) Exercise for Evaluation Sample of 80 Path Forecasts

Nominal coverage level \ Prediction region	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
Marg	1.521	1.119	0.804	0.328	0.221	0.208	0.180	0.088	0.065	0.045
Bonf	0.254	0.216	0.166	0.159	0.156	0.132	0.112	0.123	0.090	0.044
SPR-G (Chi)	0.315	0.274	0.270	0.221	0.201	0.165	0.125	0.098	0.058	0.032
SPR-G (Chi) Cal.	0.331	0.330	0.282	0.254	0.212	0.188	0.149	0.115	0.079	0.042
SPR-S (Chi)	0.207	0.149	0.159	0.206	0.164	0.156	0.140	0.119	0.079	0.044
SPR-S (Chi) Cal.	0.309	0.313	0.293	0.255	0.217	0.197	0.167	0.129	0.089	0.042
SPR-G (F)	0.315	0.275	0.270	0.221	0.204	0.165	0.125	0.098	0.065	0.032
SPR-G (F) Cal.	0.331	0.333	0.282	0.255	0.214	0.188	0.149	0.119	0.079	0.042
SPR-S (F)	0.207	0.149	0.159	0.206	0.164	0.165	0.149	0.119	0.079	0.044
SPR-S (F) Cal.	0.318	0.313	0.293	0.258	0.215	0.197	0.167	0.129	0.089	0.042
SPR-G (Emp I)	0.305	0.269	0.269	0.228	0.200	0.161	0.112	0.086	0.076	0.032
SPR-G (Emp I) Cal.	0.354	0.334	0.291	0.264	0.212	0.177	0.148	0.126	0.079	0.042
SPR-G (Emp II)	0.335	0.322	0.278	0.228	0.182	0.159	0.120	0.087	0.067	0.032
SPR-G (Emp II) Cal.	0.349	0.334	0.292	0.247	0.210	0.177	0.149	0.115	0.079	0.042
SPR-S (Emp)	0.190	0.149	0.137	0.143	0.164	0.130	0.112	0.119	0.079	0.044
SPR-S (Emp) Cal.	0.338	0.312	0.265	0.251	0.222	0.205	0.167	0.129	0.089	0.042
FWEJPR (6)	0.320	0.318	0.270	0.241	0.231	0.219	0.183	0.139	0.092	0.044
FWEJPR-A (6)	0.366	0.334	0.327	0.291	0.259	0.232	0.182	0.138	0.092	0.044
FWEJPR (12)	0.332	0.311	0.281	0.251	0.223	0.219	0.183	0.139	0.092	0.044
FWEJPR-A (12)	0.350	0.327	0.289	0.280	0.251	0.232	0.186	0.140	0.092	0.044
FWEJPR (24)	0.332	0.311	0.270	0.241	0.201	0.194	0.173	0.139	0.092	0.044
FWEJPR-A (24)	0.352	0.330	0.271	0.248	0.253	0.222	0.186	0.142	0.094	0.044

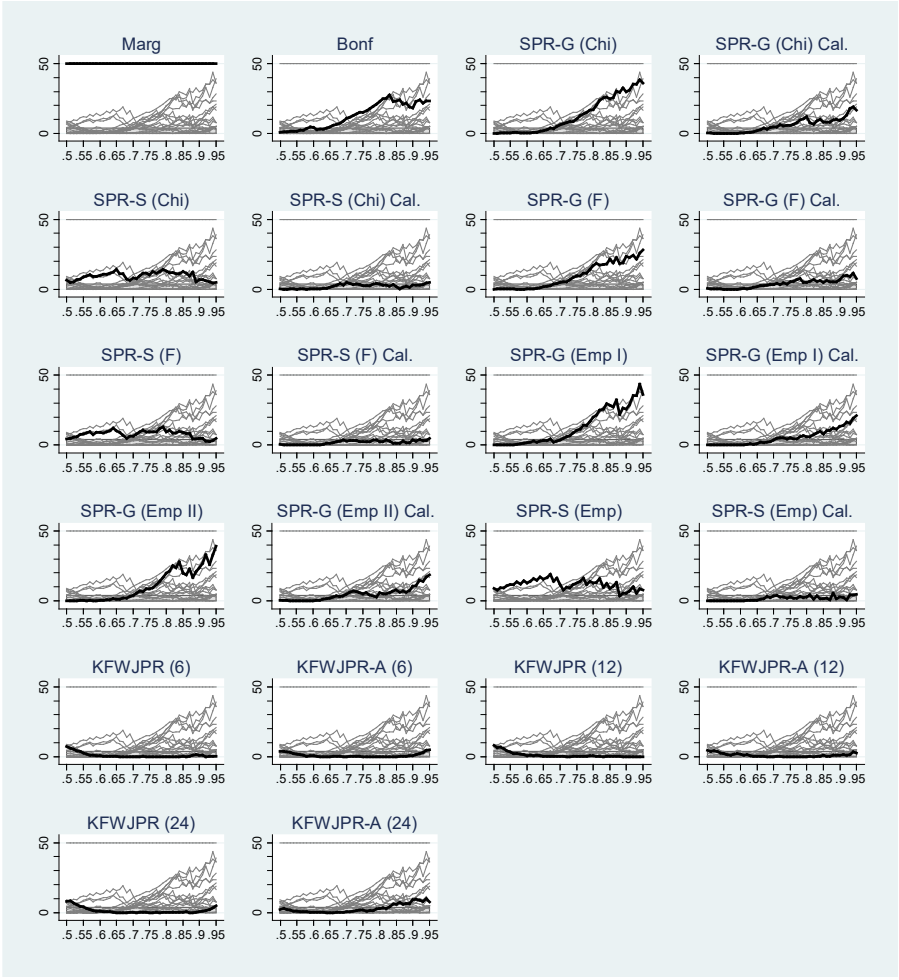
Notes: Results of normalized relative mean square error (NRMSE) for actual coverage (i.e., the empirical coverage measured by FWPER) obtained for evaluation sample of 80 path forecasts when the closing exchange rate is used as a historical reference, with an average value of actual coverage used for normalization. Different nominal coverage levels (i.e., levels of desired coverage) of examined regions (for the list of abbreviations used for individual prediction regions see note to Tab. A1) is reported.

Table A4 Results of Normalized Relative Mean Square Error (NRMSE) Exercise for Evaluation Sample of 120 Path Forecasts

Nominal coverage level \ Prediction region	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
Marg	0.870	0.639	0.540	0.274	0.229	0.159	0.104	0.097	0.071	0.053
Bonf	0.315	0.257	0.249	0.235	0.219	0.169	0.156	0.121	0.080	0.045
SPR-G (Chi)	0.350	0.316	0.289	0.262	0.230	0.198	0.163	0.112	0.076	0.033
SPR-G (Chi) Cal.	0.344	0.330	0.303	0.278	0.251	0.215	0.163	0.126	0.076	0.036
SPR-S (Chi)	0.272	0.225	0.216	0.202	0.195	0.198	0.163	0.126	0.085	0.039
SPR-S (Chi) Cal.	0.335	0.310	0.294	0.284	0.262	0.215	0.177	0.126	0.085	0.050
SPR-G (F)	0.350	0.316	0.289	0.262	0.230	0.198	0.163	0.112	0.076	0.037
SPR-G (F) Cal.	0.344	0.335	0.303	0.262	0.251	0.215	0.163	0.126	0.076	0.036
SPR-S (F)	0.272	0.227	0.216	0.202	0.213	0.198	0.163	0.126	0.085	0.039
SPR-S (F) Cal.	0.335	0.319	0.294	0.284	0.262	0.215	0.177	0.126	0.085	0.050
SPR-G (Emp I)	0.315	0.277	0.289	0.246	0.224	0.203	0.163	0.112	0.084	0.037
SPR-G (Emp I) Cal.	0.323	0.343	0.302	0.261	0.255	0.215	0.163	0.135	0.085	0.038
SPR-G (Emp II)	0.318	0.265	0.277	0.255	0.230	0.190	0.163	0.112	0.075	0.037
SPR-G (Emp II) Cal.	0.328	0.325	0.304	0.279	0.238	0.215	0.163	0.125	0.085	0.038
SPR-S (Emp)	0.220	0.181	0.163	0.154	0.170	0.168	0.137	0.112	0.085	0.049
SPR-S (Emp) Cal.	0.351	0.296	0.296	0.266	0.228	0.204	0.177	0.126	0.085	0.050
FWEJPR (6)	0.355	0.350	0.320	0.286	0.258	0.219	0.170	0.124	0.090	0.050
FWEJPR-A (6)	0.350	0.342	0.328	0.294	0.258	0.220	0.170	0.132	0.089	0.045
FWEJPR (12)	0.355	0.350	0.320	0.287	0.258	0.219	0.170	0.124	0.090	0.050
FWEJPR-A (12)	0.364	0.334	0.314	0.295	0.258	0.219	0.170	0.124	0.089	0.045
FWEJPR (24)	0.362	0.350	0.327	0.275	0.258	0.219	0.170	0.124	0.090	0.050
FWEJPR-A (24)	0.363	0.332	0.314	0.305	0.254	0.219	0.170	0.124	0.086	0.045

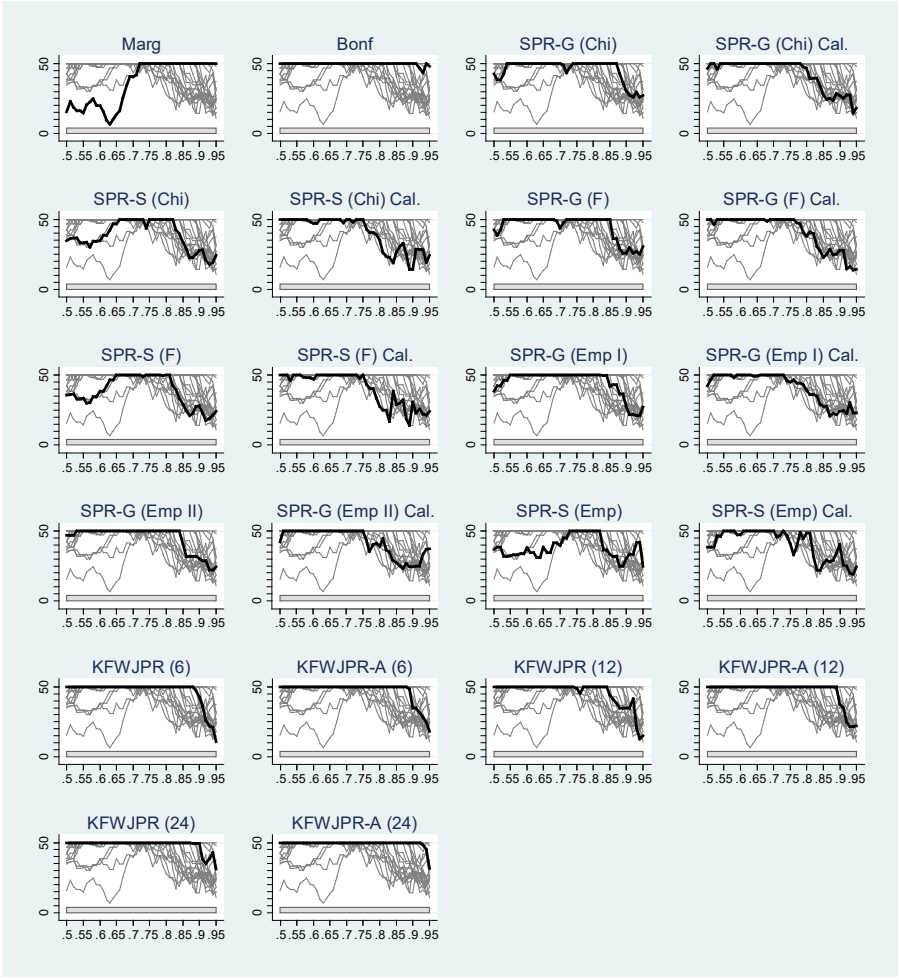
Notes: Results of normalized relative mean square error (NRMSE) for actual coverage (i.e., the empirical coverage measured by FWPER) obtained for evaluation sample of 120 path forecasts when the closing exchange rate is used as a historical reference, with an average value of actual coverage used for normalization. Different nominal coverage levels (i.e., levels of desired coverage) of examined regions (for the list of abbreviations used for individual prediction regions see note to Tab. A1) is reported.

Figure A1 Results of the Unconditional Coverage Test for Evaluation Sample of 40 Path Forecasts



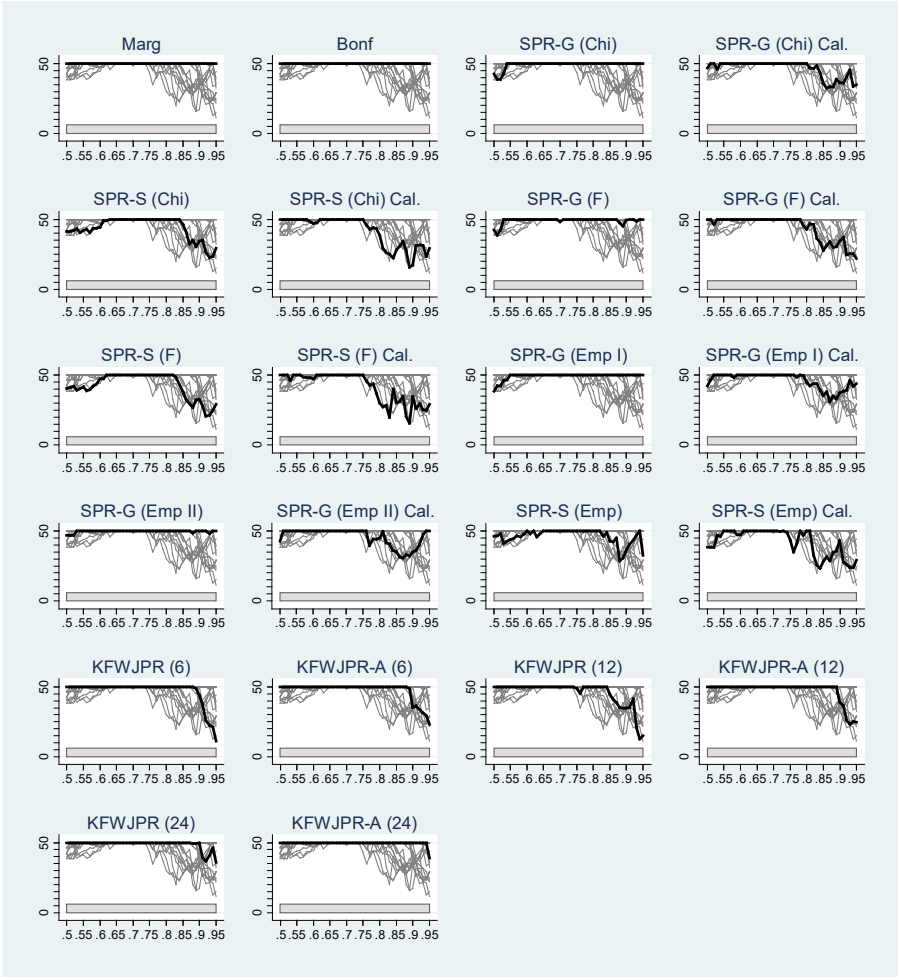
Notes: Results of the LR_{UC} test for unconditional coverage for examined prediction regions for the evaluation sample of 40 path forecasts when closing exchange rate is used as a historical reference, each region variant in its own panel with corresponding LR_{UC} statistics highlighted with a bold line. All of the panels depict nominal coverage levels on the horizontal axis and values of LR_{UC} statistic on the vertical axis. Areas filled with grey color at the bottom of each panel represent regions for which the LR_{UC} tests are statistically insignificant at the 5 per cent level. That is, if the obtained LR_{UC} statistic exceeds the grey band, the unconditional coverage of associated prediction regions is significantly different from the corresponding desired nominal coverage. For the list of abbreviations used for individual prediction regions see note to Tab. A1.

Figure A2 Results of the Independence Test for Evaluation Sample of 40 Path Forecasts



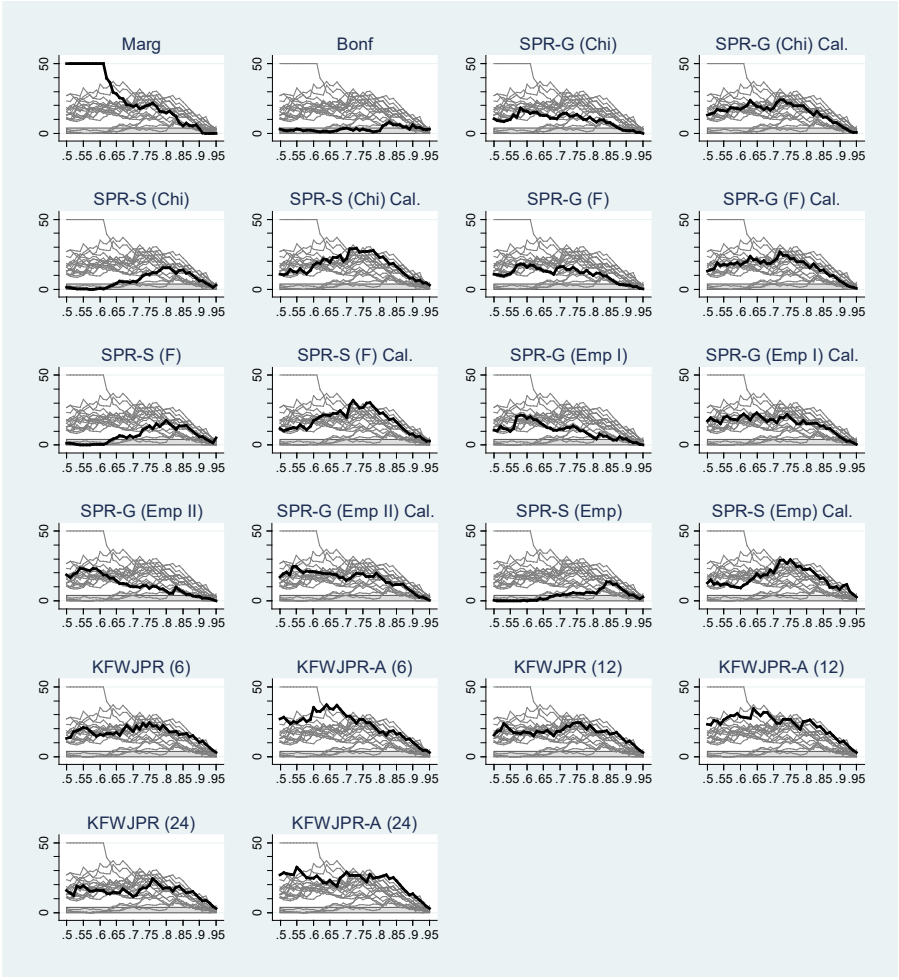
Notes: Results of the LR_{IN} test for independence hypothesis for examined prediction regions for the evaluation sample of 40 path forecasts when closing exchange rate is used as a historical reference, each region variant in its own panel with corresponding LR_{IN} statistics highlighted with a bold line. All of the panels depict nominal coverage levels on the horizontal axis and values of LR_{IN} statistic on the vertical axis. Areas filled with grey color at the bottom of each panel represent regions for which the LR_{IN} tests are statistically insignificant at the 5 per cent level. That is, if the obtained LR_{IN} statistic exceeds the grey band, the unconditional coverage of associated prediction regions is significantly different from the corresponding desired nominal coverage. For the list of abbreviations used for individual prediction regions see note to Tab. A1.

Figure A3 Results of the Conditional Coverage Test for Evaluation Sample of 40 Path Forecasts



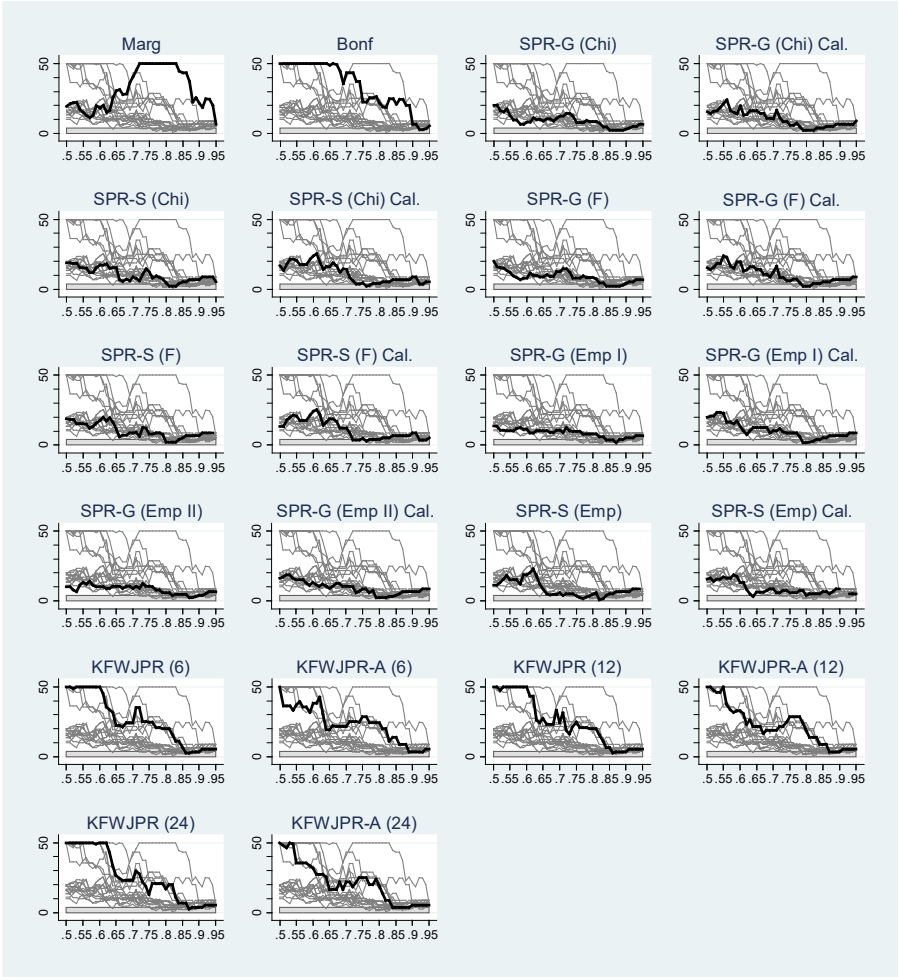
Notes: Results of the LR_{CC} test for conditional coverage for examined prediction regions for the evaluation sample of 40 path forecasts when closing exchange rate is used as a historical reference, each region variant in its own panel with corresponding LR_{CC} statistics highlighted with a bold line. All of the panels depict nominal coverage levels on the horizontal axis and values of LR_{CC} statistic on the vertical axis. Areas filled with grey color at the bottom of each panel represent regions for which the LR_{CC} tests are statistically insignificant at the 5 per cent level. That is, if the obtained LR_{CC} statistic exceeds the grey band, the unconditional coverage of associated prediction regions is significantly different from the corresponding desired nominal coverage. For the list of abbreviations used for individual prediction regions see note to Tab. A1.

Figure A4 Results of the Unconditional Coverage Test for Evaluation Sample of 80 Path Forecasts



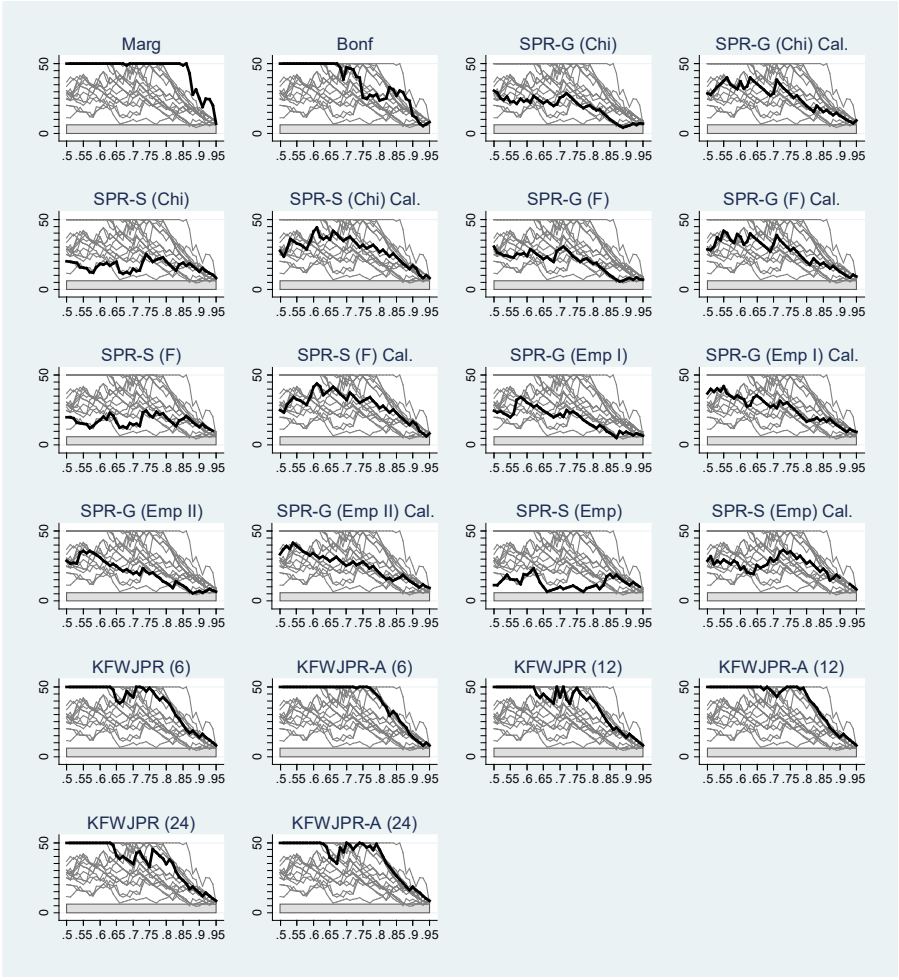
Notes: Results of the LR_{UC} test for unconditional coverage for examined prediction regions for the evaluation sample of 80 path forecasts when closing exchange rate is used as a historical reference, each region variant in its own panel with corresponding LR_{UC} statistics highlighted with a bold line. All of the panels depict nominal coverage levels on the horizontal axis and values of LR_{UC} statistic on the vertical axis. Areas filled with grey color at the bottom of each panel represent regions for which the LR_{UC} tests are statistically insignificant at the 5 per cent level. That is, if the obtained LR_{UC} statistic exceeds the grey band, the unconditional coverage of associated prediction regions is significantly different from the corresponding desired nominal coverage. For the list of abbreviations used for individual prediction regions see note to Tab. A1.

Figure A5 Results of the Independence Test for Evaluation Sample of 80 Path Forecasts



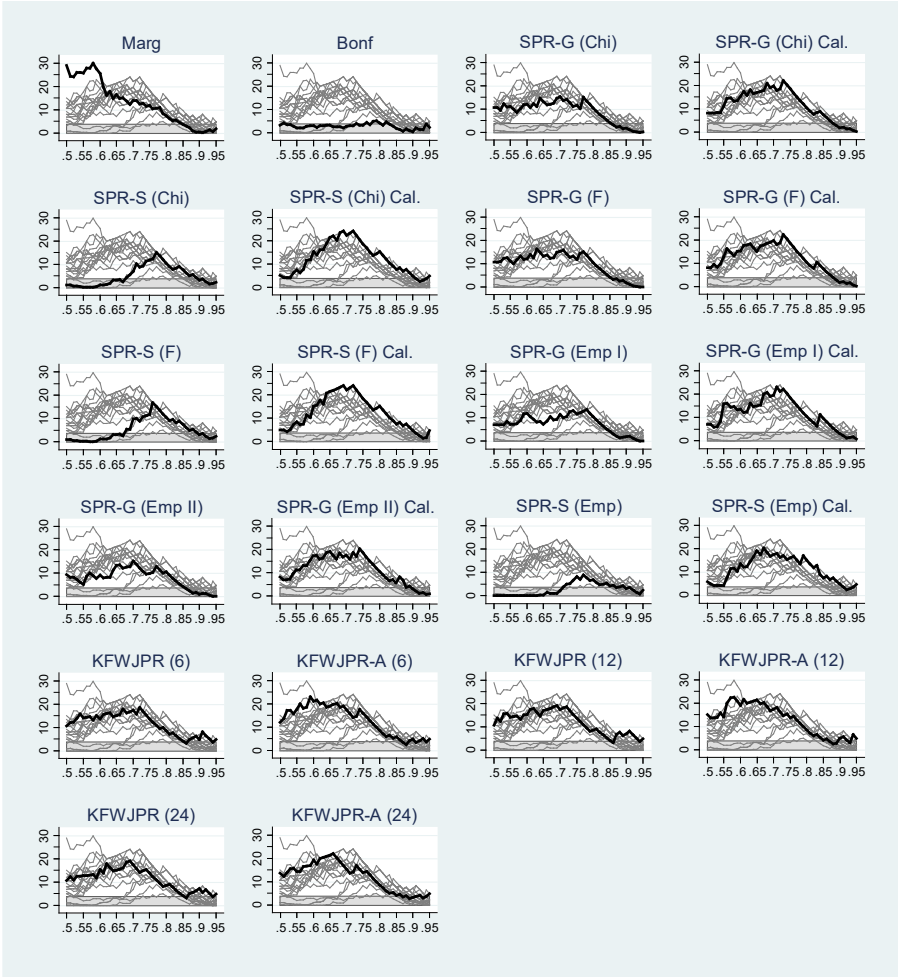
Notes: Results of the LR_{IN} test for independence hypothesis for examined prediction regions for the evaluation sample of 80 path forecasts when closing exchange rate is used as a historical reference, each region variant in its own panel with corresponding LR_{IN} statistics highlighted with a bold line. All of the panels depict nominal coverage levels on the horizontal axis and values of LR_{IN} statistic on the vertical axis. Areas filled with grey color at the bottom of each panel represent regions for which the LR_{IN} tests are statistically insignificant at the 5 per cent level. That is, if the obtained LR_{IN} statistic exceeds the grey band, the unconditional coverage of associated prediction regions is significantly different from the corresponding desired nominal coverage. For the list of abbreviations used for individual prediction regions see note to Tab. A1.

Figure A6 Results of the Conditional Coverage Test for Evaluation Sample of 80 Path Forecasts



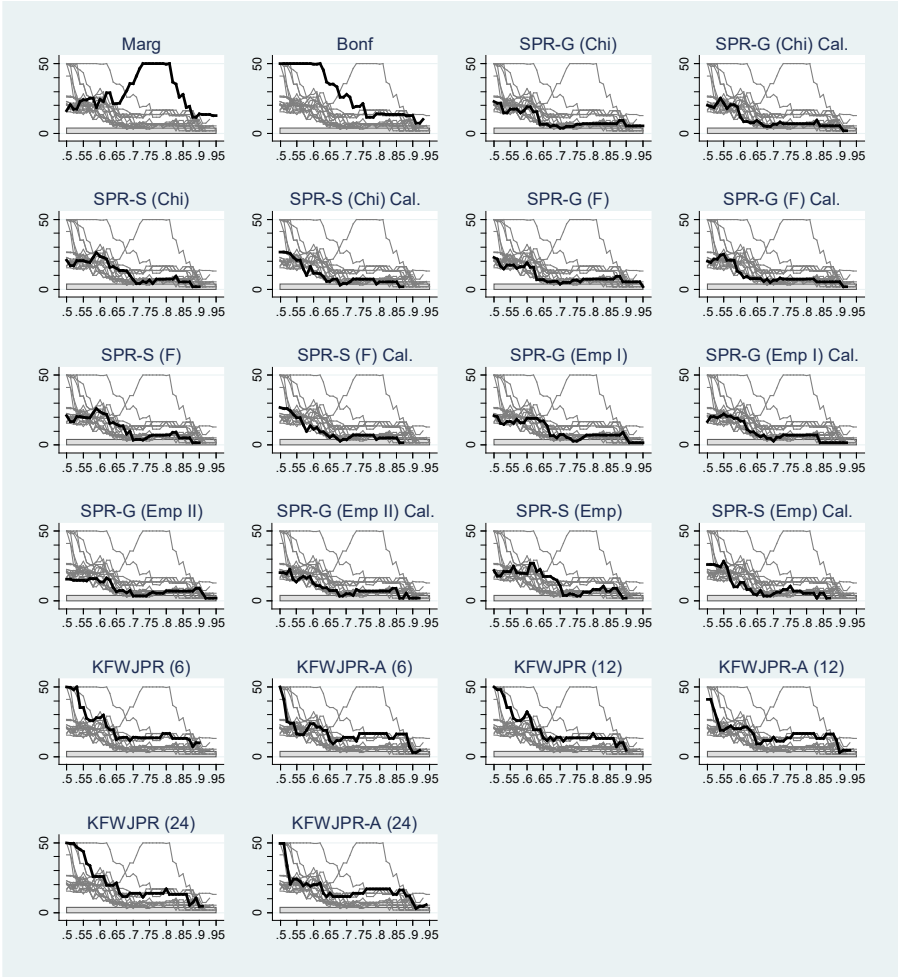
Notes: Results of the LR_{CC} test for conditional coverage for examined prediction regions for the evaluation sample of 80 path forecasts when closing exchange rate is used as a historical reference, each region variant in its own panel with corresponding LR_{CC} statistics highlighted with a bold line. All of the panels depict nominal coverage levels on the horizontal axis and values of LR_{CC} statistic on the vertical axis. Areas filled with grey color at the bottom of each panel represent regions for which the LR_{CC} tests are statistically insignificant at the 5 per cent level. That is, if the obtained LR_{CC} statistic exceeds the grey band, the unconditional coverage of associated prediction regions is significantly different from the corresponding desired nominal coverage. For the list of abbreviations used for individual prediction regions see note to Tab. A1.

Figure A7 Results of the Unconditional Coverage Test for Evaluation Sample of 120 Path Forecasts



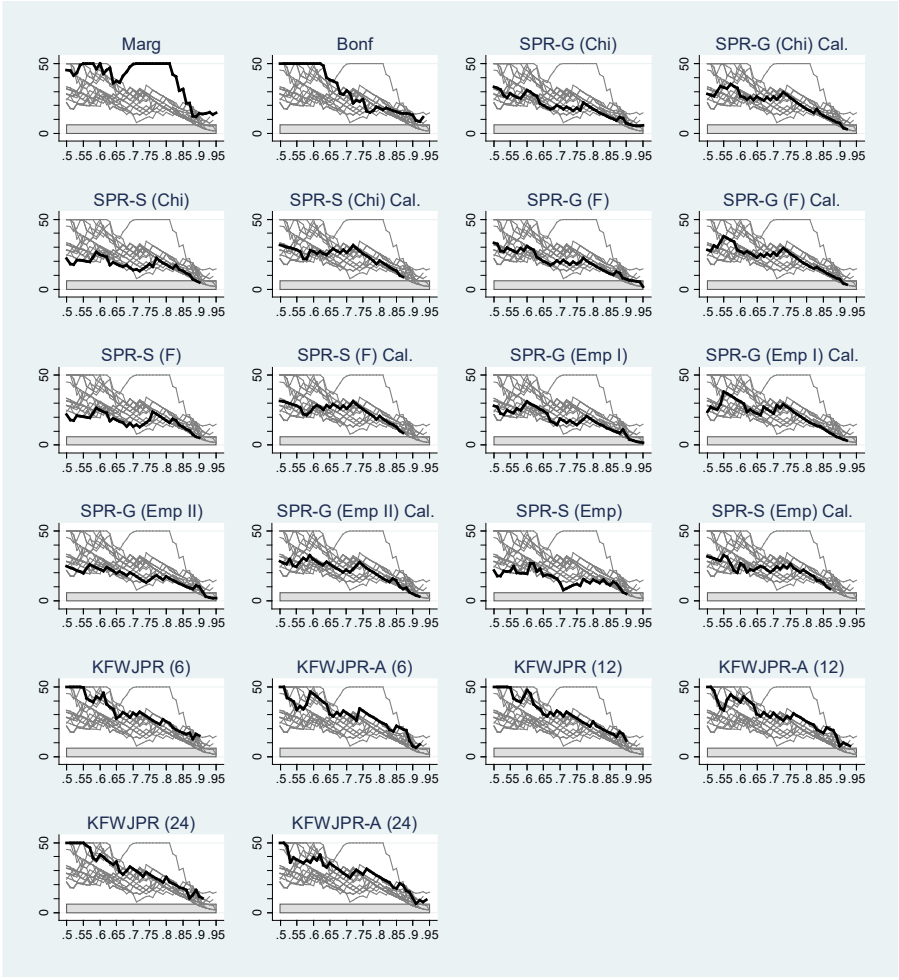
Notes: Results of the LR_{UC} test for unconditional coverage for examined prediction regions for the evaluation sample of 120 path forecasts when closing exchange rate is used as a historical reference, each region variant in its own panel with corresponding LR_{UC} statistics highlighted with a bold line. All of the panels depict nominal coverage levels on the horizontal axis and values of LR_{UC} statistic on the vertical axis. Areas filled with grey color at the bottom of each panel represent regions for which the LR_{UC} tests are statistically insignificant at the 5 per cent level. That is, if the obtained LR_{UC} statistic exceeds the grey band, the unconditional coverage of associated prediction regions is significantly different from the corresponding desired nominal coverage. For the list of abbreviations used for individual prediction regions see note to Tab. A1.

Figure A8 Results of the Independence Test for Evaluation Sample of 120 Path Forecasts



Notes: Results of the LR_{IN} test for independence hypothesis for examined prediction regions for the evaluation sample of 120 path forecasts when closing exchange rate is used as a historical reference, each region variant in its own panel with corresponding LR_{IN} statistics highlighted with a bold line. All of the panels depict nominal coverage levels on the horizontal axis and values of LR_{IN} statistic on the vertical axis. Areas filled with grey color at the bottom of each panel represent regions for which the LR_{IN} tests are statistically insignificant at the 5 per cent level. That is, if the obtained LR_{IN} statistic exceeds the grey band, the unconditional coverage of associated prediction regions is significantly different from the corresponding desired nominal coverage. For the list of abbreviations used for individual prediction regions see note to Tab. A1.

Figure A9 Results of the Conditional Coverage Test for Evaluation Sample of 120 Path Forecasts



Notes: Results of the LR_{CC} test for conditional coverage for examined prediction regions for the evaluation sample of 120 path forecasts when closing exchange rate is used as a historical reference, each region variant in its own panel with corresponding LR_{CC} statistics highlighted with a bold line. All of the panels depict nominal coverage levels on the horizontal axis and values of LR_{CC} statistic on the vertical axis. Areas filled with grey color at the bottom of each panel represent regions for which the LR_{CC} tests are statistically insignificant at the 5 per cent level. That is, if the obtained LR_{CC} statistic exceeds the grey band, the unconditional coverage of associated prediction regions is significantly different from the corresponding desired nominal coverage. For the list of abbreviations used for individual prediction regions see note to Tab. A1.

2. Monte Carlo Simulations for Small Sample Rejection Rate of the Utilized LR Tests

The presented simulations indicate that there is a substantial distortion of the unconditional coverage LR_{UC} test when the autocorrelation of the generated coverage sequence is non-zero. The intensity of the size distortion increases with the magnitude of autocorrelation, while the particular order of the autocorrelation appears to be inconsequential for the size distortion to occur. This size distortion is also not mitigated by increasing the sample size in the limited scope that was examined, as the simulations appear to suggest that it is rather persistent. As for the test power, that is to a high degree dependent on the difference between the actual (empirical) coverage and set (desired) nominal coverage, as greater difference improves the test power. The test power is also improved by the sample size. However, the power appears to generally deteriorate with the increasing magnitude of the autocorrelation, again regardless of the order of the autocorrelation.

The size of the LR_{IN} independence test is far more stable and appears to be substantially distorted only in cases of smaller sample sizes. However, in terms of the test power, the test appears to be substantially weakened by the order of the autocorrelation of the coverage sequence, with higher orders substantially diminishing its power, regardless of the autocorrelation magnitude. The larger sample size does improve the power compared to smaller samples. Nevertheless, the gains are very small, apart from the cases of autocorrelation of order one, in which case even an increase from a sample size of 50 to 100 can make a stark difference in the test power.

Regarding the conditional coverage LR_{CC} test, the test size appears rather stable, apart from size distortions observed for empirical coverage of 0.95. Since the joint test is an amalgamation of the previous two individual tests, the conditional coverage test appears to be far more robust in terms of power than its two components. However, for combinations of multiple factors which diminish the power of its components, such as higher orders of autocorrelation, the small magnitude of autocorrelation, the small difference between the actual (empirical) coverage and set nominal coverage and small sample sizes, its power can be as low as 0.065.

Regarding the practical implications of these simulations for the obtained results, since the utilization of the restricted compared to unrestricted period used for assessment leads to additional 57 to 58 observations available to the independence and unconditional coverage test, respectively, the extension of a sample by additional 50 observations alone does not result in dramatic differences in terms of test size and test power. Potentially further complicating the assessment, the additional observations may exhibit different properties in terms of autocorrelation structure, which is why it is not possible to determine the robustness of the results based on the sample size alone.

Table B1 Monte Carlo Simulations of LR Test Rejection Rate for Actual Coverage of 0.55

$\gamma = 0.55$	***T***	***50***					***150***					***250***				
	γ_0	0.55	0.65	0.75	0.85	0.95	0.55	0.65	0.75	0.85	0.95	0.55	0.65	0.75	0.85	0.95
$\rho = 0$	UC	**0.037**	0.211	0.759	0.994	1.000	**0.046**	0.664	0.999	1.000	1.000	**0.054**	0.895	1.000	1.000	1.000
	IN	**0.046**	**0.039**	**0.028**	**0.027**	**0.114**	**0.051**	**0.047**	**0.042**	**0.034**	**0.057**	**0.052**	**0.048**	**0.047**	**0.041**	**0.055**
	CC	**0.045**	0.173	0.676	0.987	1.000	**0.049**	0.553	0.997	1.000	1.000	**0.050**	0.816	1.000	1.000	1.000
$\rho = 0.1$	UC	**0.059**	0.240	0.741	0.988	1.000	**0.070**	0.653	0.998	1.000	1.000	**0.082**	0.871	1.000	1.000	1.000
	IN	0.055	0.060	0.075	0.098	0.204	0.181	0.183	0.194	0.224	0.286	0.298	0.311	0.319	0.337	0.375
	CC	0.065	0.217	0.692	0.982	1.000	0.152	0.638	0.996	1.000	1.000	0.249	0.876	1.000	1.000	1.000
$\rho = 0.25$	UC	**0.106**	0.288	0.719	0.973	1.000	**0.121**	0.634	0.992	1.000	1.000	**0.134**	0.837	1.000	1.000	1.000
	IN	0.253	0.268	0.298	0.315	0.386	0.814	0.807	0.787	0.750	0.651	0.968	0.964	0.952	0.921	0.804
	CC	0.244	0.414	0.786	0.985	1.000	0.752	0.925	0.999	1.000	1.000	0.945	0.995	1.000	1.000	1.000
$\rho = 0.5$	UC	**0.228**	0.376	0.697	0.933	0.998	**0.248**	0.621	0.966	1.000	1.000	**0.266**	0.778	0.997	1.000	1.000
	IN	0.858	0.841	0.804	0.719	0.691	1.000	1.000	0.999	0.989	0.906	1.000	1.000	1.000	1.000	0.972
	CC	0.846	0.891	0.968	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$\rho = 0.75$	UC	**0.435**	0.528	0.700	0.875	0.981	**0.453**	0.643	0.901	0.992	1.000	**0.468**	0.730	0.968	1.000	1.000
	IN	0.985	0.965	0.929	0.886	0.902	1.000	1.000	1.000	0.995	0.981	1.000	1.000	1.000	1.000	0.996
	CC	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Notes: Rejection rate for Christoffersen's unconditional coverage (UC), independence (IN), and conditional coverage (CC) tests based on Monte Carlo simulations for actual coverage $\gamma = 0.55$ and order of autocorrelation $p = 1$, T denoting the number of generated path-forecasts assessed by the aforementioned LR tests, ρ the magnitude of autocorrelation, and γ_0 the desired (nominal) coverage levels. The values represent the power of the corresponding LR test unless assumptions associated with the test are met, in which case the rejection rate is highlighted in bold and represents the size of the corresponding LR test.

Table B2 Monte Carlo Simulations of LR Test Rejection Rate for Actual Coverage Level 0.65

$\gamma = 0.65$	T	*50*					*150*					*250*				
	γ_0	0.55	0.65	0.75	0.85	0.95	0.55	0.65	0.75	0.85	0.95	0.55	0.65	0.75	0.85	0.95
$\rho = 0$	UC	0.296	**0.044**	0.256	0.858	1.000	0.702	**0.045**	0.719	1.000	1.000	0.906	**0.046**	0.907	1.000	1.000
	IN	**0.046**	**0.038**	**0.029**	**0.027**	**0.114**	**0.050**	**0.048**	**0.042**	**0.035**	**0.058**	**0.050**	**0.047**	**0.048**	**0.041**	**0.054**
	CC	0.261	**0.042**	0.142	0.705	0.999	0.629	**0.047**	0.587	1.000	1.000	0.842	**0.047**	0.851	1.000	1.000
$\rho = 0.1$	UC	0.309	**0.068**	0.281	0.839	0.999	0.684	**0.071**	0.706	1.000	1.000	0.881	**0.070**	0.884	1.000	1.000
	IN	0.056	0.061	0.073	0.096	0.203	0.179	0.183	0.196	0.222	0.286	0.301	0.310	0.317	0.340	0.378
	CC	0.283	0.073	0.202	0.738	0.998	0.685	0.161	0.683	0.999	1.000	0.883	0.253	0.906	1.000	1.000
$\rho = 0.25$	UC	0.330	**0.116**	0.323	0.808	0.997	0.650	**0.122**	0.685	0.998	1.000	0.839	**0.119**	0.851	1.000	1.000
	IN	0.253	0.270	0.293	0.318	0.388	0.814	0.807	0.786	0.749	0.650	0.968	0.964	0.952	0.923	0.806
	CC	0.448	0.263	0.408	0.830	0.999	0.917	0.753	0.942	1.000	1.000	0.991	0.939	0.997	1.000	1.000
$\rho = 0.5$	UC	0.373	**0.242**	0.412	0.768	0.983	0.608	**0.248**	0.659	0.984	1.000	0.768	**0.247**	0.788	0.999	1.000
	IN	0.856	0.843	0.803	0.717	0.691	1.000	1.000	0.999	0.988	0.907	1.000	1.000	1.000	0.999	0.972
	CC	0.871	0.837	0.893	0.978	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$\rho = 0.75$	UC	0.483	**0.451**	0.556	0.756	0.946	0.596	**0.452**	0.661	0.934	0.999	0.690	**0.450**	0.734	0.982	1.000
	IN	0.984	0.966	0.931	0.886	0.901	1.000	1.000	1.000	0.995	0.981	1.000	1.000	1.000	1.000	0.996
	CC	0.999	0.999	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Notes: Rejection rate for Christoffersen's unconditional coverage (UC), independence (IN), and conditional coverage (CC) tests based on Monte Carlo simulations for actual coverage $\gamma = 0.65$ and order of autocorrelation $p = 1$, T denoting the number of generated path-forecasts assessed by the aforementioned LR tests, ρ the magnitude of autocorrelation, and γ_0 the desired (nominal) coverage levels. The values represent the power of the corresponding LR test unless assumptions associated with the test are met, in which case the rejection rate is highlighted in bold and represents the size of the corresponding LR test.

Table B3 Monte Carlo Simulations of LR Test Rejection Rate for Actual Coverage Level 0.75

$\gamma = 0.75$	***T***	***50***					***150***					***250***				
	γ_0	0.55	0.65	0.75	0.85	0.95	0.55	0.65	0.75	0.85	0.95	0.55	0.65	0.75	0.85	0.95
$\rho = 0$	UC	0.892	0.446	**0.053**	0.160	0.922	1.000	0.787	**0.045**	0.801	1.000	1.000	0.942	**0.046**	0.968	1.000
	IN	**0.045**	**0.040**	**0.028**	**0.026**	**0.112**	**0.050**	**0.048**	**0.042**	**0.034**	**0.057**	**0.051**	**0.048**	**0.047**	**0.040**	**0.052**
	CC	0.788	0.299	**0.033**	0.104	0.835	0.999	0.720	**0.045**	0.690	1.000	1.000	0.902	**0.048**	0.929	1.000
$\rho = 0.1$	UC	0.868	0.442	**0.075**	0.190	0.904	0.998	0.761	**0.068**	0.780	1.000	1.000	0.920	**0.072**	0.951	1.000
	IN	0.056	0.060	0.072	0.096	0.203	0.179	0.186	0.195	0.221	0.285	0.304	0.313	0.317	0.340	0.377
	CC	0.771	0.323	0.070	0.182	0.879	0.997	0.749	0.169	0.770	1.000	1.000	0.921	0.264	0.959	1.000
$\rho = 0.25$	UC	0.823	0.445	**0.116**	0.240	0.875	0.994	0.721	**0.119**	0.753	1.000	1.000	0.883	**0.123**	0.923	1.000
	IN	0.255	0.269	0.293	0.316	0.386	0.815	0.806	0.787	0.748	0.650	0.968	0.965	0.951	0.920	0.805
	CC	0.802	0.480	0.251	0.396	0.936	0.998	0.919	0.725	0.959	1.000	1.000	0.991	0.926	0.999	1.000
$\rho = 0.5$	UC	0.737	0.443	**0.225**	0.348	0.835	0.965	0.661	**0.244**	0.712	0.999	0.996	0.807	**0.250**	0.866	1.000
	IN	0.856	0.841	0.804	0.718	0.693	1.000	1.000	0.999	0.988	0.906	1.000	1.000	1.000	0.999	0.972
	CC	0.924	0.831	0.777	0.857	0.991	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$\rho = 0.75$	UC	0.649	0.494	**0.427**	0.530	0.825	0.876	0.621	**0.450**	0.700	0.980	0.958	0.718	**0.452**	0.795	0.998
	IN	0.984	0.966	0.929	0.887	0.901	1.000	1.000	1.000	0.995	0.981	1.000	1.000	1.000	1.000	0.996
	CC	0.995	0.992	0.993	0.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Notes: Rejection rate for Christoffersen's unconditional coverage (UC), independence (IN), and conditional coverage (CC) tests based on Monte Carlo simulations for actual coverage $\gamma = 0.75$ and order of autocorrelation $p = 1$, T denoting the number of generated path-forecasts assessed by the aforementioned LR tests, ρ the magnitude of autocorrelation, and γ_0 the desired (nominal) coverage levels. The values represent the power of the corresponding LR test unless assumptions associated with the test are met, in which case the rejection rate is highlighted in bold and represents the size of the corresponding LR test.

Table B4 Monte Carlo Simulations of LR Test Rejection Rate for Actual Coverage Level 0.85

$\gamma = 0.85$	***T***	***50***					***150***					***250***				
	γ_0	0.55	0.65	0.75	0.85	0.95	0.55	0.65	0.75	0.85	0.95	0.55	0.65	0.75	0.85	0.95
$\rho = 0$	UC	0.999	0.962	0.623	**0.080**	0.323	1.000	1.000	0.911	**0.054**	0.968	1.000	1.000	0.988	**0.059**	0.999
	IN	**0.047**	**0.039**	**0.028**	**0.026**	**0.114**	**0.051**	**0.047**	**0.042**	**0.034**	**0.057**	**0.050**	**0.047**	**0.048**	**0.040**	**0.053**
	CC	0.996	0.908	0.455	**0.045**	0.337	1.000	1.000	0.857	**0.046**	0.922	1.000	1.000	0.971	**0.047**	0.998
$\rho = 0.1$	UC	0.998	0.944	0.606	**0.103**	0.354	1.000	1.000	0.888	**0.077**	0.953	1.000	1.000	0.980	**0.082**	0.998
	IN	0.055	0.061	0.074	0.095	0.204	0.178	0.186	0.195	0.222	0.285	0.298	0.311	0.319	0.341	0.377
	CC	0.993	0.891	0.483	0.102	0.417	1.000	1.000	0.866	0.205	0.962	1.000	1.000	0.971	0.286	0.999
$\rho = 0.25$	UC	0.992	0.909	0.579	**0.147**	0.405	1.000	0.999	0.847	**0.126**	0.925	1.000	1.000	0.959	**0.131**	0.993
	IN	0.253	0.266	0.297	0.318	0.384	0.814	0.806	0.785	0.748	0.651	0.968	0.965	0.951	0.923	0.806
	CC	0.986	0.884	0.581	0.285	0.579	1.000	0.999	0.941	0.699	0.995	1.000	1.000	0.992	0.893	1.000
$\rho = 0.5$	UC	0.958	0.826	0.539	**0.252**	0.505	1.000	0.987	0.768	**0.247**	0.871	1.000	0.999	0.901	**0.253**	0.967
	IN	0.856	0.843	0.805	0.716	0.690	1.000	1.000	0.999	0.988	0.907	1.000	1.000	1.000	0.999	0.972
	CC	0.980	0.921	0.799	0.692	0.854	1.000	1.000	0.998	0.996	1.000	1.000	1.000	1.000	1.000	1.000
$\rho = 0.75$	UC	0.851	0.713	0.551	**0.467**	0.689	0.990	0.921	0.683	**0.451**	0.818	0.999	0.981	0.791	**0.452**	0.906
	IN	0.984	0.965	0.929	0.886	0.902	1.000	1.000	1.000	0.995	0.982	1.000	1.000	1.000	1.000	0.996
	CC	0.985	0.968	0.949	0.947	0.983	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Notes: Rejection rate for Christoffersen's unconditional coverage (UC), independence (IN), and conditional coverage (CC) tests based on Monte Carlo simulations for actual coverage $\gamma = 0.85$ and order of autocorrelation $p = 1$, T denoting the number of generated path-forecasts assessed by the aforementioned LR tests, ρ the magnitude of autocorrelation, and γ_0 the desired (nominal) coverage levels. The values represent the power of the corresponding LR test unless assumptions associated with the test are met, in which case the rejection rate is highlighted in bold and represents the size of the corresponding LR test.

Table B5 Monte Carlo Simulations of LR Test Rejection Rate for Actual Coverage level 0.95

$\gamma = 0.95$	***T***	***50***					***150***					***250***				
	γ_0	0.55	0.65	0.75	0.85	0.95	0.55	0.65	0.75	0.85	0.95	0.55	0.65	0.75	0.85	0.95
$\rho = 0$	UC	1.000	1.000	0.995	0.841	**0.079**	1.000	1.000	1.000	0.995	**0.063**	1.000	1.000	1.000	1.000	**0.075**
	IN	**0.046**	**0.039**	**0.027**	**0.027**	**0.114**	**0.051**	**0.047**	**0.043**	**0.035**	**0.056**	**0.053**	**0.049**	**0.048**	**0.039**	**0.053**
	CC	1.000	1.000	0.989	0.763	**0.064**	1.000	1.000	1.000	0.991	**0.083**	1.000	1.000	1.000	1.000	**0.065**
$\rho = 0.1$	UC	1.000	1.000	0.991	0.810	**0.098**	1.000	1.000	1.000	0.991	**0.083**	1.000	1.000	1.000	1.000	**0.100**
	IN	0.056	0.060	0.074	0.096	0.203	0.178	0.185	0.198	0.222	0.285	0.301	0.309	0.317	0.340	0.378
	CC	1.000	1.000	0.984	0.765	0.144	1.000	1.000	1.000	0.989	0.299	1.000	1.000	1.000	0.999	0.347
$\rho = 0.25$	UC	1.000	0.999	0.978	0.764	**0.125**	1.000	1.000	1.000	0.978	**0.117**	1.000	1.000	1.000	0.998	**0.144**
	IN	0.254	0.270	0.296	0.315	0.387	0.813	0.806	0.786	0.750	0.651	0.968	0.965	0.952	0.921	0.804
	CC	1.000	0.999	0.975	0.774	0.260	1.000	1.000	1.000	0.987	0.612	1.000	1.000	1.000	0.999	0.755
$\rho = 0.5$	UC	0.999	0.989	0.924	0.672	**0.164**	1.000	1.000	0.999	0.930	**0.192**	1.000	1.000	1.000	0.987	**0.253**
	IN	0.857	0.843	0.805	0.720	0.690	1.000	1.000	0.999	0.988	0.907	1.000	1.000	1.000	0.999	0.971
	CC	0.999	0.992	0.947	0.782	0.386	1.000	1.000	1.000	0.991	0.885	1.000	1.000	1.000	1.000	0.980
$\rho = 0.75$	UC	0.968	0.909	0.784	0.540	**0.182**	1.000	0.999	0.977	0.807	**0.381**	1.000	1.000	0.998	0.919	**0.448**
	IN	0.984	0.965	0.930	0.886	0.902	1.000	1.000	1.000	0.995	0.981	1.000	1.000	1.000	1.000	0.996
	CC	0.985	0.956	0.884	0.717	0.340	1.000	1.000	0.999	0.993	0.980	1.000	1.000	1.000	1.000	1.000

Notes: Rejection rate for Christoffersen's unconditional coverage (UC), independence (IN), and conditional coverage (CC) tests based on Monte Carlo simulations for actual coverage γ = 0.95 and order of autocorrelation p = 1, T denoting the number of generated path-forecasts assessed by the aforementioned LR tests, ρ the magnitude of autocorrelation, and γ_0 the desired (nominal) coverage levels. The values represent the power of the corresponding LR test unless assumptions associated with the test are met, in which case the rejection rate is highlighted in bold and represents the size of the corresponding LR test.

REFERENCES

Beckmann J, Schuessler (2016): Forecasting Exchange Rates under Parameter and Model Uncertainty. Journal of International Money and Finance, 60:267-288.

Beran J, Ocker D (1999): SEMIFAR Forecasts, with Applications to Foreign Exchange Rates. Journal of Statistical Planning and Inference, 80:137-153. https://doi.org/10.1016/S0378-3758(98)00247-X

Bruno V, Shin H-S (2015): Cross-Border Banking and Global Liquidity. Review of Economic Studies, 82(2):535-564.

Buncic D (2012): Understanding Forecast Failure of ESTAR Models of Real Exchange Rates. Empirical Economics, 43:399–426.

Bussiere M, Gaulier G, Steingress W (2020): Global Trade Flows: Revisiting the Exchange Rate Elasticities. Open Economies Review, 31:25-78.

Cai N, Cai Z, Fang Y, Xu Q (2015): Forecasting Major Asian Exchange Rates Using a New Semiparametric STAR Model. Empirical Economics, 48:407–426.

Cai Z, Chen L, Fang Y (2012): A New Forecasting Model for USD/CNY Exchange Rate. Studies in Nonlinear Dynamics & Econometrics, 16(3).

Ca'Zorzi M, Rubaszek M (2020): Exchange Rate Forecasting on a Napkin. Journal of International Money and Finance, 104: 102168.

Ca'Zorzi M, Cap A, Mijakovic A, Rubaszek M (2022): The Reliability of Equilibrium Exchange Rate Models: A Forecasting Perspective. International Journal of Central Banking, 18: 229-280.

Cheung Y-W, Chinn MD, Pascual AG (2005): Empirical Exchange Rate Models of the Nineties: Are Any Fit to Survive? Journal of International Money and Finance, 24:1150-1175. https://doi.org/10.1016/j.jimonfin.2005.08.002

Cheung Y-W, Chinn MD, Pascual AG, Zhang Y (2019): Exchange Rate Prediction Redux: New Models, New Data, New Currencies. Journal of International Money and Finance, 95: 332-362.

Christoffersen PF (1998): Evaluating Interval Forecasts. International Economic Review, 39(4):841-862.

Chudý M, Karmakar S, Wu WB (2020): Long-Term Prediction Intervals of Economic Time Series. Empirical Economics, 58:191–222

Curran M, Velic A (2019): Real Exchange Rate Persistence and Country Characteristics: A Global Analysis. Journal of International Money and Finance, 97: 35-56.

Ferraro K, Rogoff K, Rossi B (2015): Can Oil Prices Forecast Exchange Rates? An Empirical Analysis of the Relationship Between Commodity Prices and Exchange Rates. Journal of International Money and Finance, 54: 116-141.

Frankel JA, Rose AK, (1994): A Survey of Empirical Research on Nominal Exchange Rates. NBER Working Paper, 4865.

Georgiadis G, Mueller GJ, Schumann B (2024): Global Risk and the Dollar. Journal of Monetary Economics, 144: 103549.

Gopinath G, Boz E, Casas C, Diez FJ, Gourinchas PO, Plagborg-Moller M (2020): Dominant Currency Paradigm. American Economic Review, 110(3):677-719.

Hungnes, H. (2023). Predicting the Exchange Rate Path: The Importance of Using Up-to-Date Observations in the Forecasts. In: Valenzuela, O., Rojas, F., Herrera, L.J., Pomares, H., Rojas, I. (eds) Theory and Applications of Time Series Analysis and Forecasting. ITISE 2021. Contributions to Statistics. Springer, Cham.

Ince O, Molodtsova T (2017): Rationality and Forecasting Accuracy of Exchange Rate Expectations: Evidence from Survey-Based Forecasts. Journal of International Financial Markets, Institutions and Money, 47: 131-151.

Islam MS, Hossain E (2021): Foreign Exchange Currency Rate Prediction Using a GRU-LSTM Hybrid Network. Soft Computing Letters, 3:100009.

Jordà Ò, Knüppel M, Marcellino M (2013): Empirical Simultaneous Prediction Regions for Path-Forecasts. International Journal of Forecasting, 29:456–468.

Lee YS, Scholtes S (2014): Empirical Prediction Intervals Revisited. International Journal of Forecasting, 30:217–234.

Lehmann EL, Romano JP (2005): Testing Statistical Hypotheses. Springer, New York.

Loh W-Y (1987): Calibrating Confidence Coefficients. Journal of the American Statistical Association, 82(397):155-162.

Meese RA, Rogoff K (1983): Empirical Exchange Rate Models of the Seventies: Do They Fit Out-Of-Sample? Journal of International Economics, 14:3–24.

Norges Bank (2020), 'Monetray Policy Report', Norges Bank, Monetary Policy Report 1/2020. URL: https://www.norges-bank.no/en/news-events/news-publications/Reports/Monetary-Policy-Report-withfinancial-stability-assessment/2020/mpr-12020

Novotný F, Raková M (2011): Assessment of Consensus Forecasts Accuracy: The Czech National Bank Perspective. Czech Journal of Economics and Finance, 61(4): 348-366.

Ravishanker N, Shiao-Yen Wu L, Glaz J (1991): Multiple Prediction Intervals for Time Series: Comparison of Simultaneous and Marginal Intervals. Journal of Forecasting, 10:445-463.

Reeves JJ (2005): Bootstrap Prediction Intervals for ARCH Models. International Journal of Forecasting, 21:237– 248.

Sveriges Riksbank (2020), 'Monetary Policy Report - February 2020', Monetary Policy Report. URL: https://www.riksbank.se/en-gb/monetary-policy/monetary-policy-report/2020/monetary-policy-reportfebruary-2020/

Wang J, Wu JJ (2012): The Taylor Rule and Interval Forecast for Exchange Rates. Journal of Money Credit and Banking, 44(1):103-144. http://www.jstor.org/stable/41336817

Westerlund J, Basher SA (2007): Can Panel Data Really Improve the Predictability of the Monetary Exchange Rate Model? Journal of Forecasting, 26:365-383.

Wolf M, Wunderli D (2015): Bootstrap Joint Prediction Regions. Journal of Time Series Analysis, 36:352-376.

Wu JJ (2012): Semiparametric Forecast Intervals. Journal of Forecasting, 31:189–228.

Zhang Y-Q, Wan X (2006): Statistical fuzzy interval neural networks for currency exchange rate time series prediction. Applied Soft Computing, 7:1149–1156.