Forecasting skill of a crowd-prediction platform: A comparison of exchange rate forecasts

Niklas V. Lehmann*

February 2025

Abstract

Open-online crowd-prediction platforms are increasingly used to forecast trends and complex events. In this analysis, exchange rate predictions made on Metaculus are compared to predictions made by the random-walk, a statistical model considered very hard-to-beat. The crowd-prediction proves to be less accurate than the random-walk. By using the random-walk as a benchmark, this analysis provides a rare comparison of online crowd-prediction platforms with traditional forecasting techniques.

Data Availability & Conflicts of interest: This paper contains all relevant information to replicate and build upon the results. Relevant data on the Metaculus prediction can be obtained via the API that Metaculus provides. I did not receive any funds or other benefits for the research presented and therefore have no conflicts of interest. None of the views presented here are attributable to Metaculus.

---

*Technical University Bergakademie Freiberg, Researcher at the Chair for General Economics and Macroeconomics. Contact: Niklas-Valentin.Lehmann@vwl.tu-freiberg.de
Schlossplatz 1, 09599 Freiberg, Fakultät 6, GERMANY

1   Introduction

Forecasts from crowd-prediction platforms - online platforms that allow anyone to predict outcomes of public questions - are increasingly seen as sources of foresight and evidence. Forecasts from open-online crowd-prediction platforms have been featured in European Central Bank reports (European Central Bank, 2021) and several news sites such as The Economist, Forbes, The Washington Post and Vox have started to incorporate crowd-predictions.[1] Moreover, it is widely accepted that crowd-prediction is a valuable tool that can support policy decisions (Tetlock et al., 2017). The US intelligence community, the Virginia Department of Health and Our World in Data have leveraged crowd-predictions to inform their research (Tetlock et al., 2014, Metaculus, 2022, Metaculus, 2023). Crowd-predictions have also been directly used for research purposes, e.g. to compare them to the prediction capabilities of large language models (Schoenegger and Park, 2023).

However, it is yet uncertain exactly how much confidence we should place in the predictions of crowd prediction platforms. Whilst there is an abundance of evidence that shows that they provide forecasts that are more accurate than random guessing (Petropoulos et al., 2022, Atanasov et al., 2017), there is little evidence on the comparative accuracy of crowd-predictions. Only a limited number of studies explore whether crowd-predictions are more accurate than traditional forecasting methods in relevant areas. Farrow et al. (2017) demonstrates that crowds produced superior short-term predictions for flu cases compared to statistical models. McAndrew, Majumder, et al. (2024) shows that crowds outperform statistical models in forecasting monkeypox outbreaks; however, statistical methods are more accurate in 1-week-ahead-forecasts. Meanwhile, McAndrew, Gibson, et al. (2024) indicates that combining crowd-predictions with statistical models can enhance the accuracy of epidemic forecasts. Additionally, Karvetski (2023) reveals that crowd-predictions provided more accurate interest rate forecasts than the CME FedWatch Tool.

This study assesses the accuracy of crowd-predictions from the Metaculus platform on questions related to exchange rates, where an objective and well-studied benchmark, the random-walk without drift, exists. Furthermore, the studied predictions had real-world importance. They were supposed to create decision support for operational needs of humanitarian agencies by identifying potential upcoming conflict zones and economic

---

[1]See e.g. ("What would humans do in a world of super-AI?", 2023)

crises around the world.[2] The result of this study is that the random-walk without drift provides significantly more accurate predictions than the crowd.

This paper proceeds as follows: The next section provides a literature review of both crowd-prediction in general as well as forecasting exchange rates. Section 3 describes the historical data and the crowd-prediction platform utilized in this study. Section 4 describes the studies methodology. Section 5 then presents results, which are briefly discussed in section 6. Section 7 concludes.

2 Literature Review

2.1 Crowd-prediction

There are many ways to elicit individual forecasts and combine them to form a consensus forecast, such as simple averages. Different ways of distilling the wisdom of crowds have been suggested and tested (Atanasov et al., 2022, Armstrong, 2001). Among the most well studied are prediction markets (Arrow et al., 2008, Hanson, 2003), forecasting tournaments, and prediction polls (Atanasov et al., 2017). Crowd-prediction, in the form of prediction markets, has been successfully employed at large companies to aid decision-making (Cowgill and Zitzewitz, 2015).

Open online crowd-prediction, hereinafter just 'crowd-prediction', is a type of forecast that results from combining predictions made by multiple forecasters via a shared online platform. Crowd-prediction platforms are similar to forecasting competitions in that they involve multiple participants, and while there are often monetary rewards involved, most forecasters are primarily driven by the desire to establish their reputation and prestige by winning competitions and demonstrating a history of accurate predictions (Mellers et al., 2014).

These crowd-prediction platforms have demonstrated impressive foresight across a wide range of questions in recent times (Tetlock et al., 2014, Tetlock et al., 2017). According to Nofer (2015), the online stock prediction community has consistently performed better than professional analysts when it comes to forecasting stock returns. Brown and Reade (2019) find that an online community of amateur tipsters outperformed bookmakers in real-money sports bets, when predictions of tipsters are properly combined. Additionally, Sjöberg (2009) shows that crowds have been successful in accurately predicting political and geopolitical events. This particular application

---

[2]See www.metaculus.com/questions/11505/economic-trouble-15-currency-depreciation/

of crowd-prediction is widely utilized due to the limited alternatives available for forecasting complex events. Moreover, the paper demonstrates that crowd-predictions are as accurate as expert predictions in this domain. Katz et al. (2017) find that crowds have been the best source of foresight regarding Supreme Court decisions in the United States.

2.2 Forecasting exchange rates

Throughout history, numerous endeavors have been made to forecast exchange rates. These endeavors have revealed the challenges associated with predicting exchange rates (Cornell and Dietrich, 1978, Giddy and Dufey, 1975). It has been observed that the *random-walk without drift* is not systematically outperformed by any other (more sophisticated) statistical model (Rossi, 2013). Predicting exchange rates is a challenging task due to the potential for profit if future exchange rate movements were known in advance. By buying currency at a low price and selling it at a high price, easily predictable movements should largely vanish. Consequently, information about future exchange rates should already be factored into current rates, leaving little room for predictable drift (Giddy and Dufey, 1975).

However, the behavior of exchange rates cannot be entirely explained by economic theory (Meese and Rogoff, 1983, Rossi, 2013). Despite the extensive research on the topic, there is still limited understanding of why exchange rates move in the manner they do and why they may not adhere to concepts such as interest parity (Chinn and Meredith, 2004, Kilian and Taylor, 2003, Engel and West, 2005). This phenomenon is commonly referred to as the Meese-and-Rogoff puzzle (Meese and Rogoff, 1983). Although there seem to have been modest successes at predicting exchange rates beyond random fluctuation (Li et al., 2015, Beckmann et al., 2020), researchers have struggled to reliably outperform the random-walk with statistical models. As a result, the prediction produced by the random-walk serves as a benchmark, representing the best-known approach.

Judgmental forecasting is also used to predict exchange rates, mostly in the form of surveys. These surveys, typically provided by firms such as Consensus Economics or FX4casts, collect forecasts from economists on a quarterly basis for specific questions such as: "What will be the value of the Euro (measured in USD) on January 1st, 2025?" Forecasters participating in these surveys provide a single point prediction for these exchange rate questions. MacDonald et al. (2009) find evidence suggesting that

3

certain forecasters possess valuable insights into future exchange rates, while others do not. Önkal et al. (2003) also observe that, on average, experts outperform amateurs in short-term exchange rate forecasts, whilst Leitner and Schmidt (2006) find the opposite.

3   This study

This study assesses the accuracy of crowd-predictions from the Metaculus platform on questions related to exchange rates, where the random-walk provides an objective, difficult-to-beat and well-studied benchmark that can be constructed post-hoc. The central research question is:

> **Is the crowd-prediction as accurate as the random-walk in forecasting exchange rates?**

Hypothesis 1 *The crowd-prediction and the random-walk produce the same error in forecasting exchange rates.*

Accuracy shall be measured via the *squared error*, often also known as the *brier score* (Brier, 1950). The squared error is used primarily because Metaculus forecasters were also assessed with the squared error. Furthermore, the squared error is a strictly proper scoring rule. Strictly proper scoring rules have the property that they are maximized in expectation by the true value. This means that forecasters have an incentive to carefully assess the question and provide their honest answer.[3] The squared error is described in equation 1 and 2. If the event occurred k is 1, otherwise k is 0. The prediction made shall be p_t.

$$y(k = 0) = p_t^2 \tag{1}$$

$$y(k = 1) = (1 - p_t)^2 \tag{2}$$

Hypothesis 2 *The crowd-prediction produces a lower error in forecasting exchange rates than random guessing.*

---

[3]This assumes linear von Neumann-Morgenstern utility functions (Gneiting and Raftery, 2007) and a host of other conditions that are not necessarily met at Metaculus. However, this is not a problem for our analysis because we just assess errors ex post.

Furthermore, we ask whether the crowd possesses any foresight at all. Therefore, we compare the crowd with an imaginary random guesser, who always predicts 50%. Such predictions are useless, but still yield a squared error of 0.25. We are also interested in systematic differences between the crowd-prediction and the random-walks prediction because such differences may reflect information that is contained in one forecast but not in the other. This is reasonable as the human forecasters have access to news and other sources. If so, we may be able to retrieve a more accurate forecast by combining the predictions of random-walk and the crowd, as McAndrew, Gibson, et al. (2024) do for epidemic forecasts.

Hypothesis 3 *The crowd-prediction is not systematically different from the random-walk prediction.*

4 Data

4.1 Metaculus forecasting platform

Metaculus is a crowd-prediction platform working to improve human reasoning and coordination on topics of global importance. As a Public Benefit Corporation, Metaculus largely provides forecasts publicly. Metaculus features questions on a wide range of topics. As of 2025, over ten thousand questions have been submitted, over half of which have been evaluated since the platforms inception in 2015. Over 2 million individual forecasts have been made on the platform by thousands of active users. This method for crowd-prediction is a forecasting tournament (Tetlock et al., 2014) and not to be confused with other ways of eliciting crowd opinion such as prediction markets. Metaculus publishes a combined forecast that uses the track record of forecasters to weight predictions, giving more weight to historically accurate forecasters. This forecast is called the *Metaculus prediction* and serves as the 'crowd-prediction' in this study. [4]

Forecasters submit predictions at their leisure, and can make as many predictions at any point in time as they like. Forecasters use sliders to report their prediction. The prediction interface is depicted in figure 2. Forecasters also have access to the median

---

[4]Metaculus chooses to disclose how exactly their aggregation mechanism works. The Metaculus prediction is usually hidden for as long as questions are still open for predictions. The community prediction *can* also be hidden for some period after the opening of a new question. This feature may limit early groupthink, yet disadvantages predictors which predict on questions early.
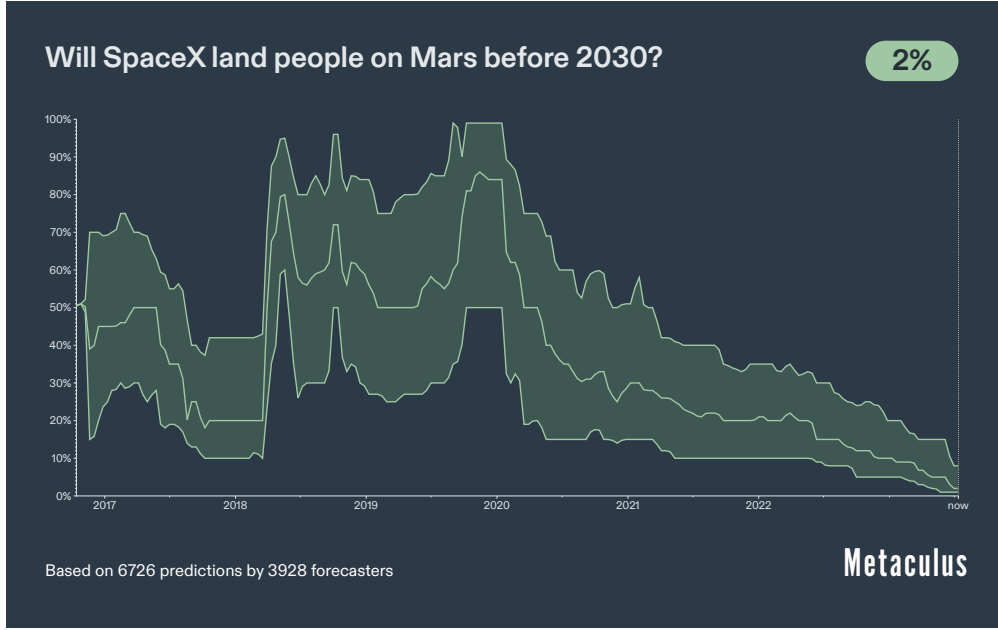
Figure 1: Metaculus interface to explore how predictions on a question have evolved

prediction as well as quartiles. Figure 1 shows how forecasters see other forecasters predictions.

4.2 Forecasting question series on exchange rates

I compiled all exchange rate questions on Metaculus that can be directly compared to a prediction made by the random-walk. These include 12 questions from a question series that ran between June and December 2022, as well as 2 questions regarding a potential British Pound parity with the dollar. Table 1 provides a comprehensive list of the relevant exchange rate questions that were featured on Metaculus.

All exchange rates in this study are based on the US dollar as the reference currency. Question 12 opened on 2016-07-09 and question 13 opened on 2022-09-29. The questions in this study were classified as resolved 'Yes' ($k = 1$) if the depreciation threshold was reached, or 'No' ($k = 0$) if the depreciation threshold was not reached by the time the questions closed on December 31st 2022. Over a period of six months, a total of 61 amateur forecasters participated in the question series, collectively submitting 2453 individual predictions.[5] On average, there were 144 total predictions per question.[6] As

---

[5]Since this study is purely observational and not conducted in a lab-setting, no additional information about forecasters is available.
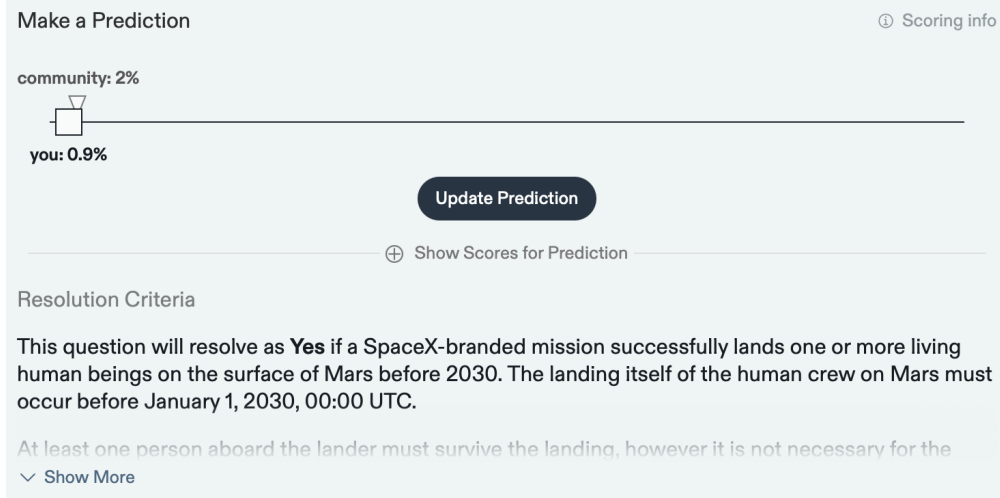
[6]The question series can be explored online at

Figure 2: Metaculus prediction interface for the SpaceX-question from figure 1

1	Euro to depreciate by $> 15\%$ in 2022?	$k = 0$
2	Indonesian Rupiah to depreciate by $> 15\%$ in 2022?	$k = 0$
3	Thai Baht to depreciate by $> 15\%$ in 2022?	$k = 0$
4	Russian Ruble to depreciate by $> 15\%$ in 2022?	$k = 1$
5	Turkish Lira to depreciate by $> 15\%$ in 2022?	$k = 0$
6	Polish zloty to depreciate by $> 15\%$ in 2022?	$k = 0$
7	Brazilian Real to depreciate by $> 15\%$ in 2022?	$k = 0$
8	Mexican Peso to depreciate by $> 15\%$ in 2022?	$k = 0$
9	Indian Rupee to depreciate by $> 15\%$ in 2022?	$k = 0$
10	Pakistani Rupee to depreciate by $> 15\%$ in 2022?	$k = 0$
11	Chinese Yuan Renminbi to depreciate by $> 15\%$ in 2022?	$k = 0$
12	British Pound to reach market parity with the dollar by 2017?	$k = 0$
13	British Pound to reach market parity with the dollar by 2023?	$k = 0$

Table 1: Studied questions from the crowd-forecasting platform

an incentive for their participation, forecasters received a small monetary compensation based on the accuracy of their predictions. The forecasters were rewarded based on their squared error. A 2500$ prize pool was awarded to accurate forecasts across a larger set of 64 questions, which included the questions from 1 to 12. The average amount of money awarded per predictor was below 10$. However, only the most accurate forecasters could claim payments. The predictions on questions 12 and 13 were not monetarily incentivized.

5   Methodology

In order to compare the Metaculus prediction, which is discussed in the prior section, to the random-walks predictions, we need to generate the latter. Since this study seeks to collect predictions for whether a currency will depreciate to a certain value by a certain time, a Monte Carlo simulation was used to collect these predictions. The random-walk, described in equation 3, is a model that simply projects the normal distribution with the historic standard deviation into the future.

$$x_{t+1} = x_t + \epsilon_t, \quad \text{with } \epsilon_t \sim \mathcal{N}(0, \sigma_h^2) \tag{3}$$

x_t shall be the value of a currency at time t, as measured in US-dollars. The next time steps value is defined as the current value plus some random change (ϵ), which is normally distributed with the historic standard deviation of the exchange rate series.

The financial data on exchange rates was gathered from Yahoo Finance using the quantmod-package in R.[7] Exchange rate data was collected on a daily basis from January 5, 2022, until December 31, 2022, which encompassed the resolve time of the questions. There was an exception for the questions specific to the British Pound, which opened earlier. For these cases, exchange rate data was gathered starting from January 1, 2015.

---

www.metaculus.com/questions/11505/economic-trouble-15-currency-depreciation/.

[7]The data used to resolve the crowd-prediction questions was sourced from xe.com. In other words, the crowd-prediction is based on other financial data then the random-walk-prediction. The difference between the two is small and negligible in comparison to the exchange rate movements and uncertainty in exchange rate expectations displayed.

Algorithm 1 describes how the predictions were generated.[8]

Algorithm 1: The random-walks prediction

Input: Historical exchange rate time series; Date from which to forecast
Output: Probability of currency depreciation
Step 1: The historical variance σ_h^2 of the exchange rate is calculated based on the available data for that specific day. That is, if the probability of the Euro depreciating below the specified threshold is to be predicted for Oct. 30th 2022, then the historical standard deviation is calculated using the exchange rate values (variance) from January 5th on up to the 29th of October.;
Step 2: Ten thousand paths of the required length (up until December 31) are generated using a Monte-Carlo simulation. The next days value is determined by equation 1, whereas ϵ is a random draw from a normal distribution with a variance equal to the historical variance σ_h^2 computed in the previous step.;
Step 3: To derive the probability that a currency will depreciate below the specified threshold, the number of simulated paths where this occurs (at some point), is divided by 10000 (the total number of paths).;

Most importantly, the approach arrives at a probability or forecast that anyone would have been able to easily generate in real time. Only information available at the specific day is used to make forecasts. The forecasts reflect pseudo-out-of-sample-performance of the random-walk. Thereby, the forecasts generated by this method provide a fair benchmark for the crowd-prediction.

In order to evaluate the performance of the crowd-prediction and the random-walk model, and test hypothesis 1, forecasts were compared on a question-by-question basis. For each individual question, and corresponding exchange rate time series, the squared error of the crowd-prediction and the random-walk model's predictions were computed. Errors of the crowd-prediction and random-walk were then summed up across questions and divided by the number of questions to arrive at the mean squared error.[9] The method with the lower mean squared error is more accurate. In order to assess whether differences in accuracy between the two methods are merely a chance result, a Diebold-Mariano test is deployed (Diebold and Mariano, 2002). The Diebold-Mariano test takes the difference between the errors produced by either forecasting technique and uses a simple z-test to check whether the difference in error can be explained by

---
[8]Exchange rate fluctuations are not observed on weekends - a fact that is brushed over in the Monte-Carlo-Simulation. This is insofar relevant as currencies cannot depreciate below the threshold on weekends. If a currency is close to the depreciation barrier, but there are no more days in which the market is open in the year, the algorithm would still assume a potential depreciation of the currency on each day. However, this situation has not turned up in this analysis.

[9]Since some questions resolved early, the number of questions changes across time.

variation around 0. This plain version of the test would assume a stationary time se-
ries, i.e. that the differences in error are normally distributed draws with a mean of
0 and an estimated variance that is constant. The situation here is different, since
errors are clearly autocorrelated. Therefore, this analysis employs heteroskedasticity-
and-autocorrelation-robust-standard-error estimates, as suggested by Diebold (2015).
Furthermore, we test the hypothesis 2, that the Metaculus prediction is no more accu-
rate than a random guesser, by taking the difference between the mean squared error
of the crowd and 0.25 (the imaginary random guessers error) and deploying a Diebold-
Mariano test. If this test turns out to be negative, we can conclude that the crowd
does possess foresight regarding exchange rate movements. To test hypothesis 3 a sim-
ple linear regression was performed, where the crowd-prediction is regressed on the
random-walks prediction. If the crowd-prediction is not significantly different from the
random-walks prediction, i.e. hypothesis 3 holds, than the two predictions should be
perfectly correlated and the slope of the regression line should be 1 and the intercept 0.
Equation 4 describes the regression and x corresponds to the prediction made by the
random-walk.

$$x = \beta_0 + \beta_1 \text{crowd} + \psi \tag{4}$$

6   Results

Figure 3 plots the accuracy of the two methods over time, as measured by the mean
squared error. The black line represents the evolving error of crowd-predictions over
time (x-axis), while the red line shows the error of the random-walk model. The dashed
green line represents the error that a random guesser would achieve. Initially, the error
of crowd-predictions fluctuates substantially due to the limited number of predictions
available at the start, resulting in a small sample size for the forecast combination
method. As more predictions are received, the combined predictions and the error tend
to change considerably. As the figure 3 already shows, the random-walk makes more
accurate predictions on average. Table 2 contains the full results of the test. The
average error of the crowd-prediction is 0.0725 and the average error of the random-
walks prediction is 0.0421. A relatively large difference in error of around 0.0304 results.
The Diebold-Mariano test informs us that this difference is extremely unlikely to be a
chance result. Therefore, we need to reject hypothesis 1 and conclude that the random-

walk provided far more accurate predictions.

Table 2: Diebold-Mariano test regarding hypothesis 1

	Estimate	Std. Error	t-value	p-value
avg. error difference	0.0304168***	0.0048398	6.2847	$< 10^{-6}$

However, when compared with the strategy of random guessing, the crowd-prediction is significantly more accurate. The results regarding hypothesis 2 are printed in table 3. The average error difference of 0.1775 arises as the difference between the average error of the crowd and 0.25. The hypothesis 2 is confirmed.

Table 3: Diebold-Mariano test regarding hypothesis 2

	Estimate	Std. Error	t-value	p-value
avg. error difference	0.1775***	0.0078074	22.734	$< 10^{-6}$

Given that the crowd-prediction is less accurate than the random-walk, we may investigate hypothesis 3 and ask: Do both methods generate systematically different predictions? Table 4 contains results of the regression analysis described in the previous section. From them we conclude that the random-walks predictions are significantly different from the crowd-predictions. Specifically, the crowd-predictions are far higher on average. The random-walk assigns on average only 84% of the probability crowd-prediction to the currency depreciating, as the slope coefficient *crowd* is 0.84. The intercept implies that the crowd-prediction is additionally unconditionally higher than the random-walks prediction, roughly by 1 percentage point.

Table 4: Regression results regarding hypothesis 3

	Estimate	Std. Error	t-value	p-value
Intercept	-0.01199**	0.00370	-3.242	0.00121
crowd	0.84076***	0.01248	-12.75962	$< 10^{-6}$
R^2	0.6998			

7   Discussion

Using crowd-prediction for forecasting exchange rates should be avoided, if the random-walk is available, because the latter yields more accurate predictions in this study. This
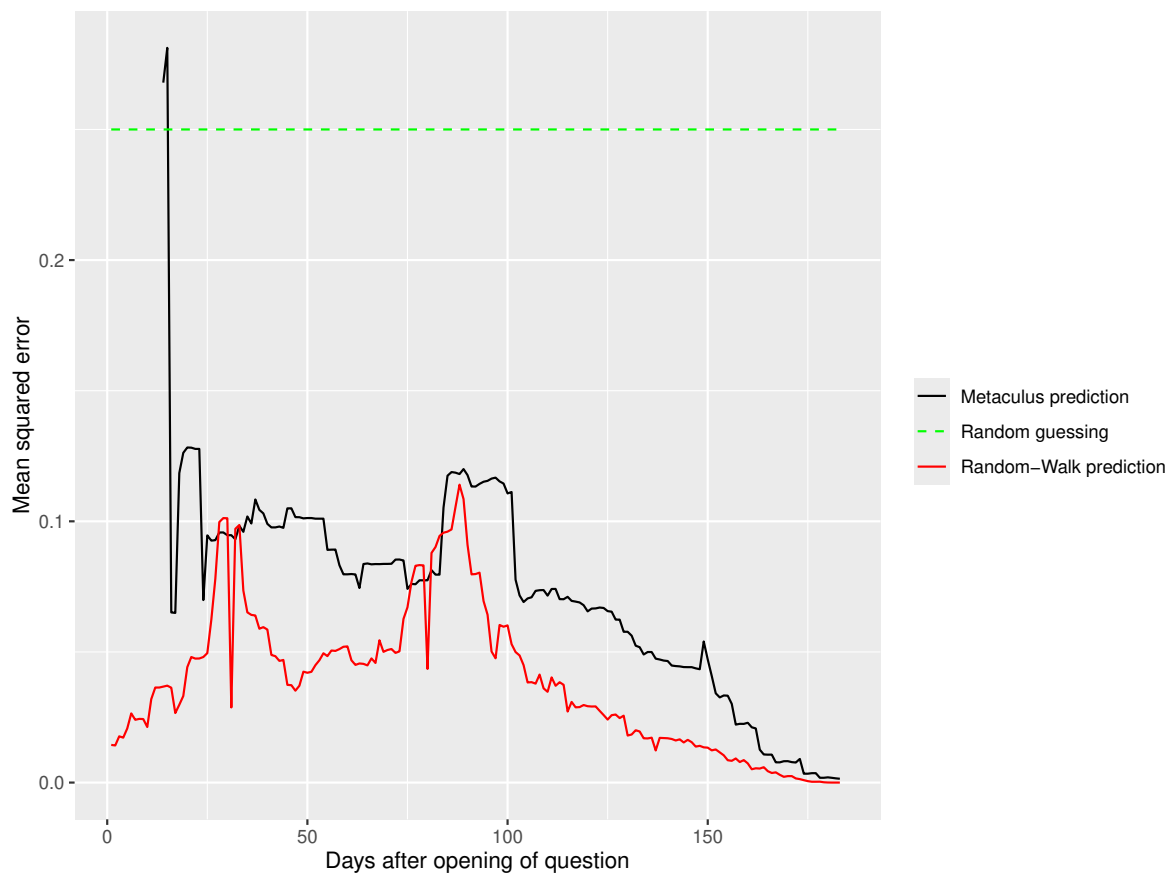
Figure 3: Accuracy of the Metaculus Prediction compared to the predictions made by the random-walk

study documents that the forecasting error of the crowd-prediction is a whopping 72% higher, on average, when compared to the random-walk.

Crowd-predictions are increasingly incorporated as a decision-support, which has tremendous potential, but we should carefully assess whether better alternatives exist before doing so. In the realm of exchange rates, a statistical forecasting technique fares better. A lot of literature that touts the benefits of crowd-prediction, mostly along the line of research pursued by Tetlock (see e.g. Tetlock and Gardner, 2016). This result is not in contrast with this research, as the crowd did successfully forecast exchange rates, significantly outperforming random guessing. Yet, the random-walk is still better.

Should we expect the crowd and the random-walk to produce forecasts that can be combined to yield a better forecast? Probably not. The crowd mostly seems to have produced forecasts that are mostly less certain and less responsive to exchange rate movements. Even on a question-by-question basis, the crowd is often clearly less accurate than the random-walk. The graphs in the appendix, which detail all predictions, provide great visual evidence of this.

The studies results are to be interpreted in light of the setting. Since the forecasts were only 6-month ahead (at maximum), the prediction accuracy on long-term outcomes might be different. Furthermore, since substantial currency depreciation is a rare phenomenon, this study involves some low probability forecasts, which are more difficult to interpret given the limited sample of events. However, this is not very relevant, as these predictions do not contribute much to the mean squared error and thus do not affect the outcome of this study much. We have no information regarding the forecasters knowledge of exchange rates, and thus it is not implausible to conjecture that a set of experts could have provided more accurate forecasts.

8   Conclusion

As crowd-prediction platforms increasingly inform public and private expectations regarding future events, their reliability and comparative efficacy warrants examination. This paper compared the accuracy of exchange rate predictions from the Metaculus platform to a well-studied statistical benchmark, the random-walk without drift, which provided mixed evidence. The crowd did possess considerable foresight but the random-walk without drift provided significantly more accurate predictions. Undoubtedly, the increasing use of crowd-predictions can improve decision-making in countless areas, and

is thus a great development. However, this analysis shows that simple statistical methods can be far more accurate than the crowd-prediction. Therefore, when it comes to decision-making, it is imperative to exercise caution in the use and interpretation of crowd-predictions and to evaluate whether more effective alternatives are available. To better understand what level of confidence we should place into crowd-prediction, and when to utilize crowd-prediction over other methods, we need more research. We should e.g. analyze which factors predict accuracy, so that we can know when to expect reliable forecasts. While this study specifically examines exchange rate forecasts, extending similar analyses to other domains would reveal how crowd-prediction fares relative to other forecasting techniques. Currently, this line of research is inhibited by a limited sample of available crowd-predictions. Thus, another avenue for future research is to collect crowd-predictions on events that allow a rigorous comparison to other means of collecting forecasts.

References

Armstrong, J. S. (2001). *Combining forecasts*. Springer.

Arrow, K. J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J. O., Levmore, S., Litan, R., Milgrom, P., Nelson, F. D., et al. (2008). The Promise of Prediction Markets. *Science*, *320*(5878), 877–878.

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2017). Distilling the Wisdom of Crowds: Prediction markets vs. Prediction polls. *Management Science*, *63*(3), 691–706.

Atanasov, P., Witkowski, J., Mellers, B., & Tetlock, P. (2022). Crowd Prediction Systems: Markets, Polls, and Elite Forecasters. *Proceedings of the 23rd ACM Conference on Economics and Computation*, 1013–1014.

Beckmann, J., Koop, G., Korobilis, D., & Schüssler, R. A. (2020). Exchange rate predictability and dynamic Bayesian learning. *Journal of Applied Econometrics*, *35*(4), 410–421.

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

Brown, A., & Reade, J. J. (2019). The wisdom of amateur crowds: Evidence from an online community of sports tipsters. *European Journal of Operational Research*, *272*(3), 1073–1081. https://doi.org/https://doi.org/10.1016/j.ejor.2018.07.015

Chinn, M. D., & Meredith, G. (2004). Monetary policy and long-horizon uncovered interest parity. *IMF staff papers*, *51*(3), 409–430.

Cornell, W. B., & Dietrich, J. K. (1978). The Efficiency of the Market for Foreign Exchange Under Floating Exchange Rates. *The Review of Economics and Statistics*, 111–120.

Cowgill, B., & Zitzewitz, E. (2015). Corporate prediction markets: Evidence from google, ford, and firm x. *The Review of Economic Studies*, *82*(4), 1309–1341.

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business & Economic Statistics*, *33*(1), 1–1.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, *20*(1), 134–144.

Engel, C., & West, K. D. (2005). Exchange Rates and Fundamentals. *Journal of Political Economy*, *113*(3), 485–517.

European Central Bank. (2021). *Rotation towards normality – the impact of COVID-19 vaccine-related news on global financial markets*. Retrieved November 13, 2023, from https://www.ecb.europa.eu/pub/economic-bulletin/html/eb202101.en.html

Farrow, D. C., Brooks, L. C., Hyun, S., Tibshirani, R. J., Burke, D. S., & Rosenfeld, R. (2017). A human judgment approach to epidemiological forecasting. *PLoS computational biology*, *13*(3). https://doi.org/10.1371/journal.pcbi.1005248

Giddy, I. H., & Dufey, G. (1975). The Random Behavior of Flexible Exchange Rates: Implications for Forecasting. *Journal of International Business Studies*, 1–32.

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

Hanson, R. (2003). Combinatorial information market design. *Information Systems Frontiers*, *5*, 107–119.

Karvetski, C. (2023). Superforecasting the fed's target range [Accessed: 2024-12-13].

Katz, D. M., Bommarito II, M. J., & Blackman, J. (2017). Crowdsourcing accurately and robustly predicts Supreme Court decisions. *arXiv preprint arXiv:1712.03846*.

Kilian, L., & Taylor, M. P. (2003). Why is it so difficult to beat the random walk forecast of exchange rates? *Journal of International Economics*, *60*(1), 85–107.

Leitner, J., & Schmidt, R. (2006). A systematic comparison of professional exchange rate forecasts with the judgemental forecasts of novices. *Central European Journal of Operations Research*, *14*, 87–102. https://doi.org/10.1007/s10100-006-0161-x

Li, J., Tsiakas, I., & Wang, W. (2015). Predicting Exchange Rates Out of Sample: Can Economic Fundamentals Beat the Random Walk? *Journal of Financial Econometrics*, *13*(2), 293–341.

MacDonald, R., Menkhoff, L., & Rebitzky, R. R. (2009). Exchange Rate Forecasters' Performance: Evidence of Skill? *CESifo Working Paper Series*.

McAndrew, T., Gibson, G. C., Braun, D., Srivastava, A., & Brown, K. (2024). Chimeric forecasting: An experiment to leverage human judgment to improve forecasts of infectious disease using simulated surveillance data. *Epidemics*, *47*, 100756. https://doi.org/https://doi.org/10.1016/j.epidem.2024.100756

McAndrew, T., Majumder, M. S., Lover, A. A., Venkatramanan, S., Bocchini, P., Besiroglu, T., Codi, A., Dempsey, G., Abbott, S., Chevalier, S., et al. (2024). Assessing Human Judgment Forecasts in the Rapid Spread of the Mpox Outbreak: Insights and Challenges for Pandemic Preparedness. *arXiv preprint arXiv:2404.14686*.

Meese, R. A., & Rogoff, K. (1983). Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of International Economics*, *14*(1-2), 3–24.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., et al. (2014). Psychological Strategies for Winning a Geopolitical Forecasting Tournament. *Psychological Science*, *25*(5), 1106–1115.

Metaculus. (2022). *The Keep Virginia Safe Tournament 2021/22 Project Summary*. Retrieved June 21, 2023, from https://www.metaculus.com/notebooks/11162/the-keep-virginia-safe-tournament-202122-project-summary/

Metaculus. (2023). *Forecasting Our World in Data: The Next 100 Years*. Retrieved June 21, 2023, from https://www.metaculus.com/notebooks/14965/forecasting-our-world-in-data-the-next-100-years/

Nofer, M. (2015). Are Crowds on the Internet Wiser than Experts?–The Case of a Stock Prediction Community. *The Value of Social Media for Predicting Stock Returns: Preconditions, Instruments and Performance Analysis*, 27–61.

Önkal, D., Yates, J., Simga-Mugan, C., & Öztin, Ş. (2003). Professional vs. amateur judgment accuracy: The case of foreign exchange rates. *Organizational Behavior and Human Decision Processes*, *91*(2), 169–185. https://doi.org/https://doi.org/10.1016/S0749-5978(03)00058-X

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., et al. (2022). Forecasting: Theory and practice. *International Journal of Forecasting*, *38*(3), 705–871.

Rossi, B. (2013). Exchange Rate Predictability. *Journal of Economic Literature*, *51*(4), 1063–1119.

Schoenegger, P., & Park, P. S. (2023). Large Language Model Prediction Capabilities: Evidence from a Real-World Forecasting Tournament. *arXiv preprint arXiv:2310.13014*.

Sjöberg, L. (2009). Are all crowds equally wise? A comparison of political election forecasts by experts and the public. *Journal of Forecasting*, *28*(1), 1–18.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The Art and Science of Prediction*. Random House.

Tetlock, P. E., Mellers, B. A., Rohrbaugh, N., & Chen, E. (2014). Forecasting Tournaments: Tools for Increasing Transparency and Improving the Quality of Debate. *Current Directions in Psychological Science*, *23*(4), 290–295.

Tetlock, P. E., Mellers, B. A., & Scoblic, J. P. (2017). Bringing probability judgments into policy debates via forecasting tournaments. *Science*, *355*(6324), 481–483.

What would humans do in a world of super-AI? (2023, May). Retrieved October 25, 2023, from www.economist.com/finance-and-economics/2023/05/23/what-would-humans-do-in-a-world-of-super-ai

Appendix

According to figure A11 the Russian Ruble triggers the depreciation threshold very early. There are two important remarks to be made here: First, the data from the quantmod-package records a downward spike at around day 20 that is not present in the data from xe.com and therefore would not have triggered the question to resolve positively at that time. However, the data from xe.com records the Russian Ruble crossing the depreciation threshold around day 100, whereas the Ruble barely touches the threshold in the quantmod-data. Secondly, both databases record the Ruble going below the depreciation threshold on day 13. However, predictions were just collected from day 14 onwards. Therefore, the question did not resolve as 'Yes' $k = 1$ on day 13, as the tournament had not really started at that point.
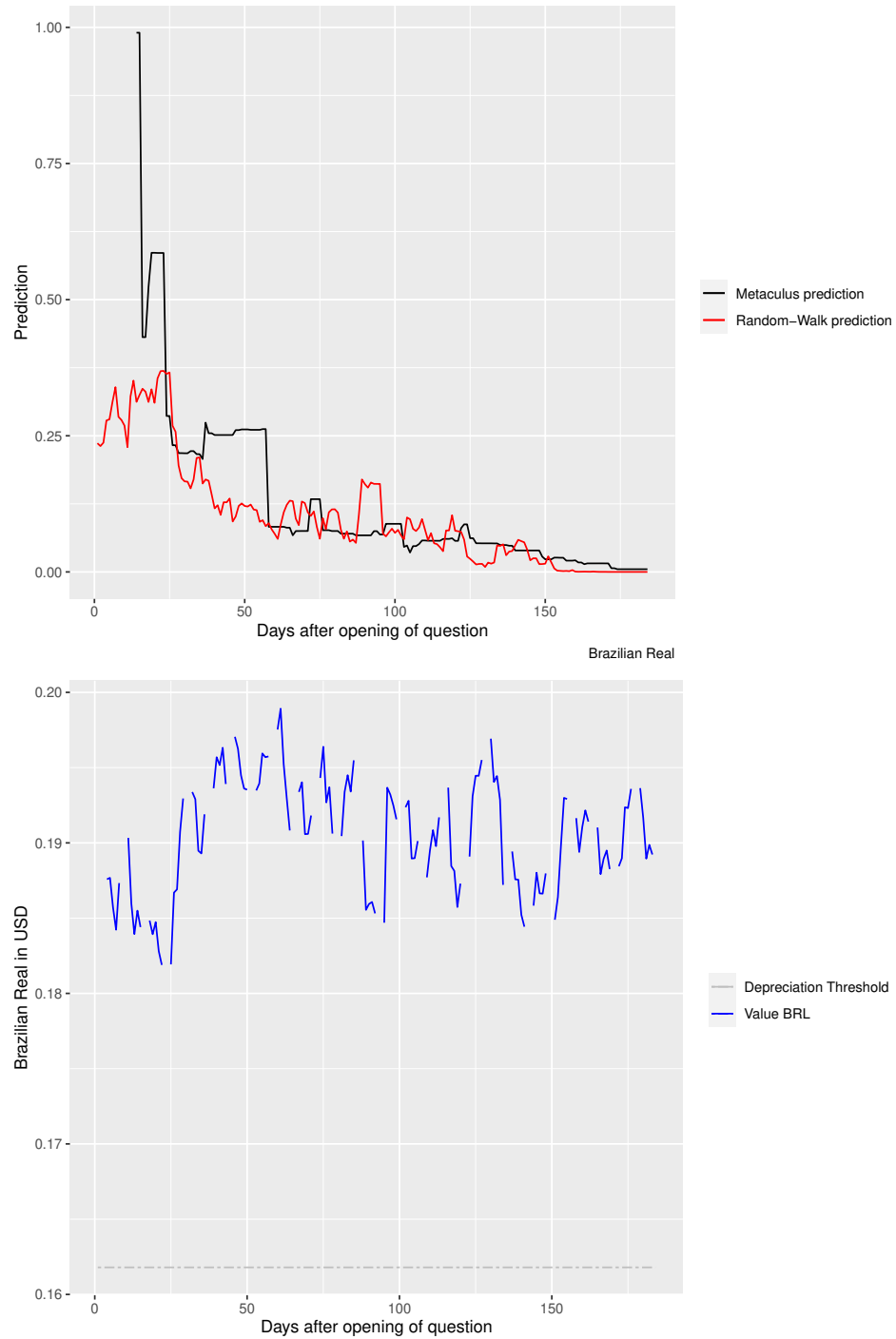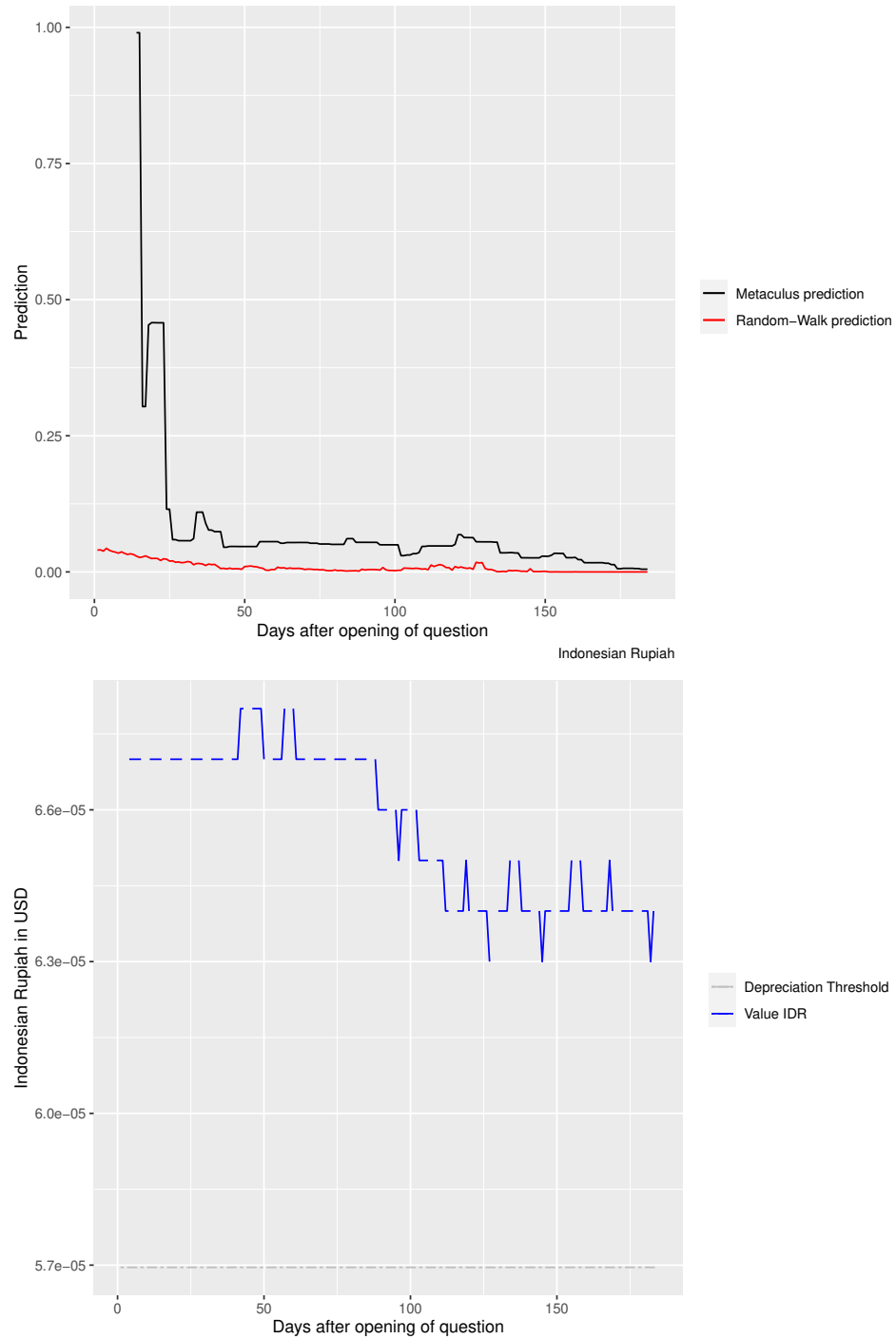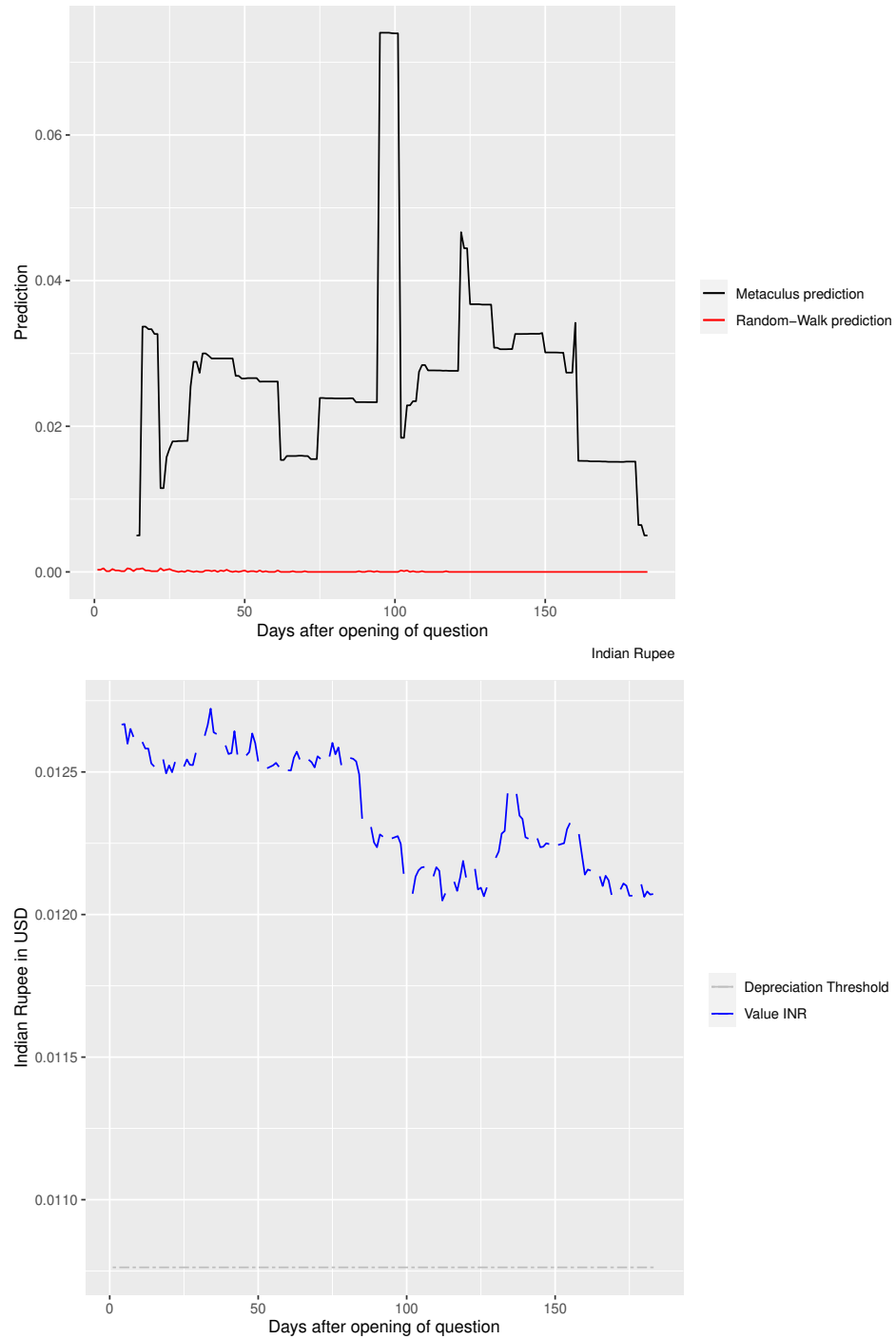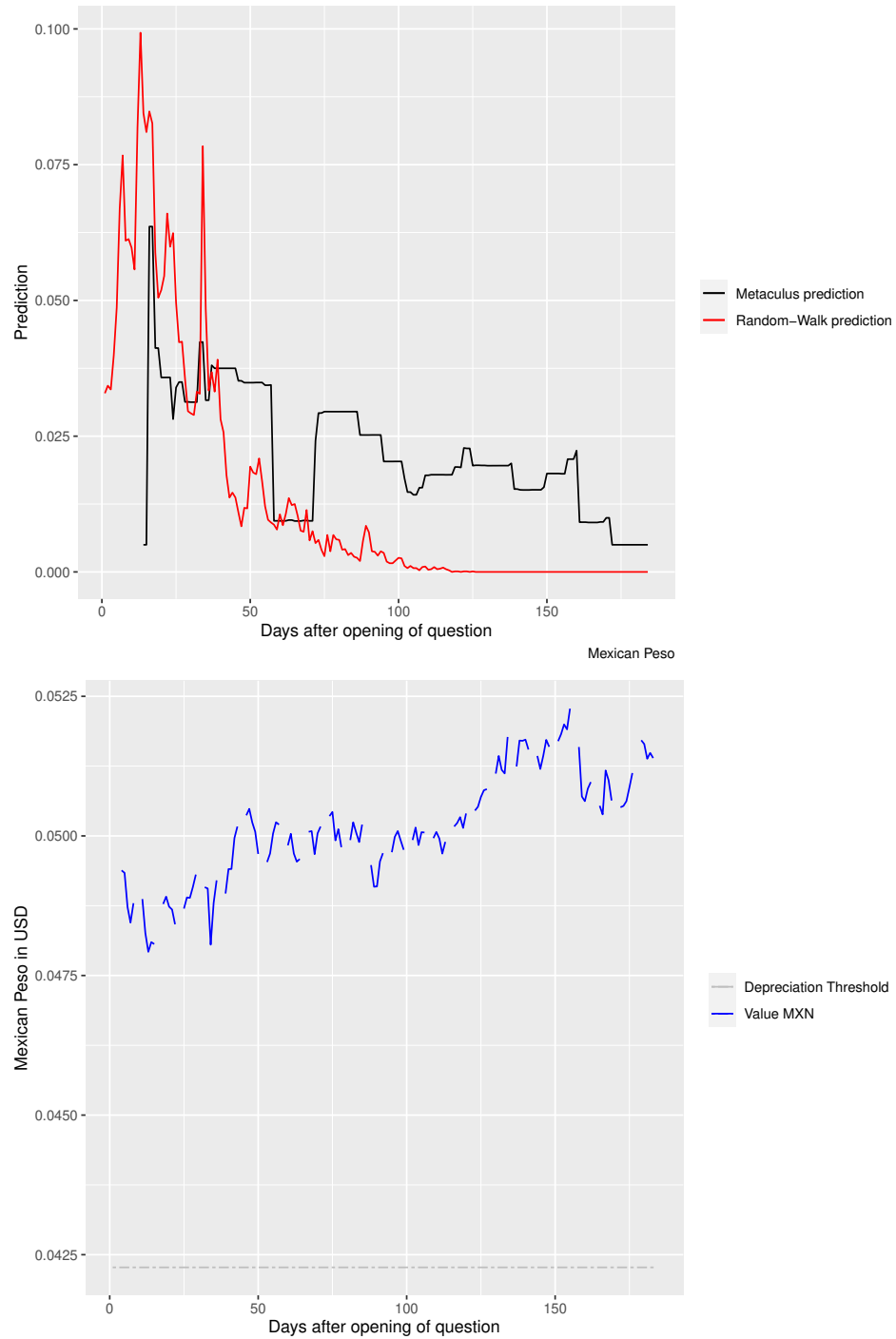
Figure A1: Top picture: Metaculus Prediction compared to the predictions made by the random-walk
Bottom picture: Exchange rate across the same time
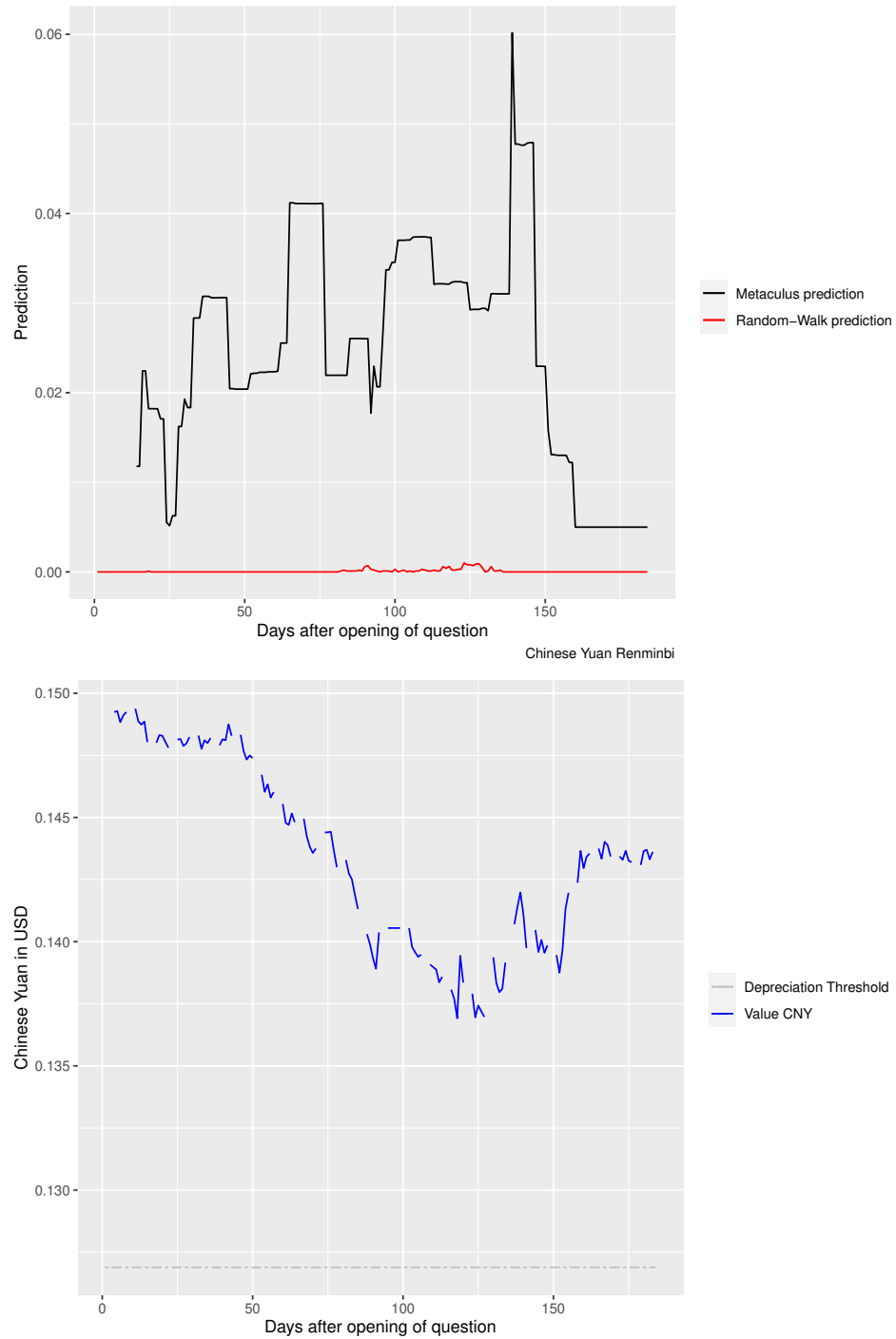
Figure A2: Top picture: Metaculus Prediction compared to the predictions made by the random-walk
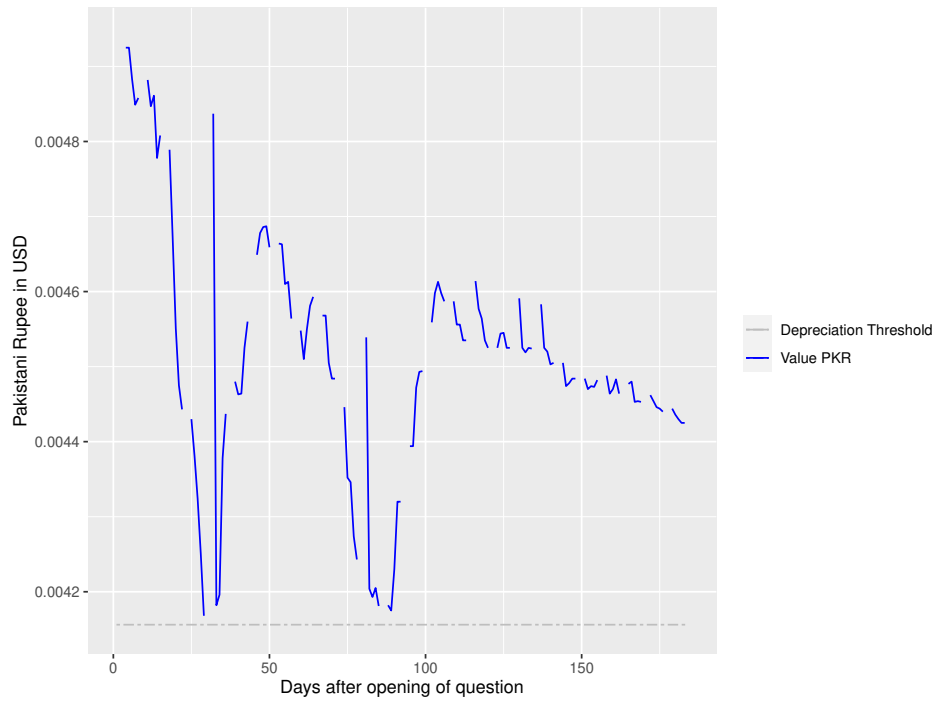Bottom picture: Exchange rate across the same time
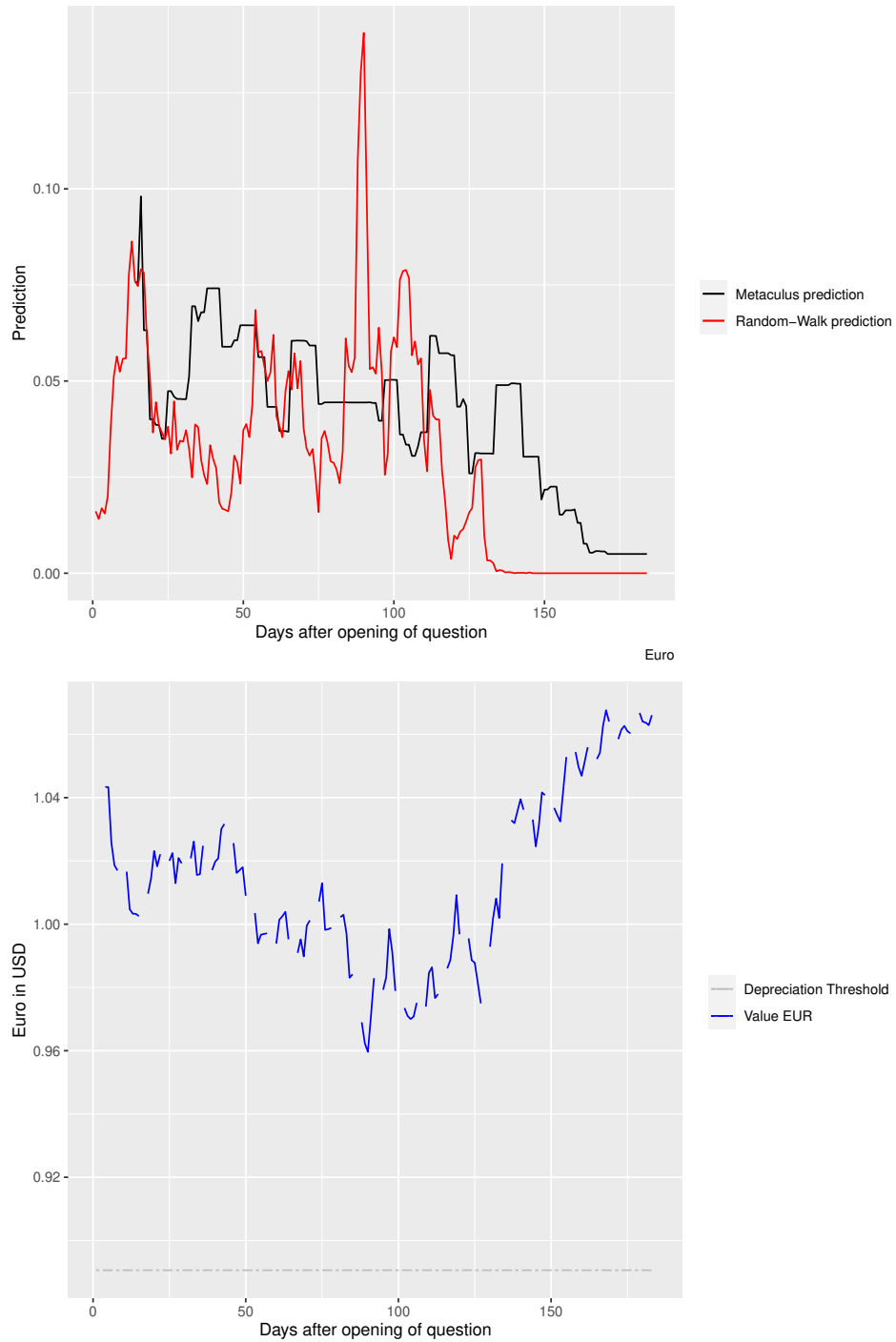
Figure A3: Top picture: Metaculus Prediction compared to the predictions made by the random-walk

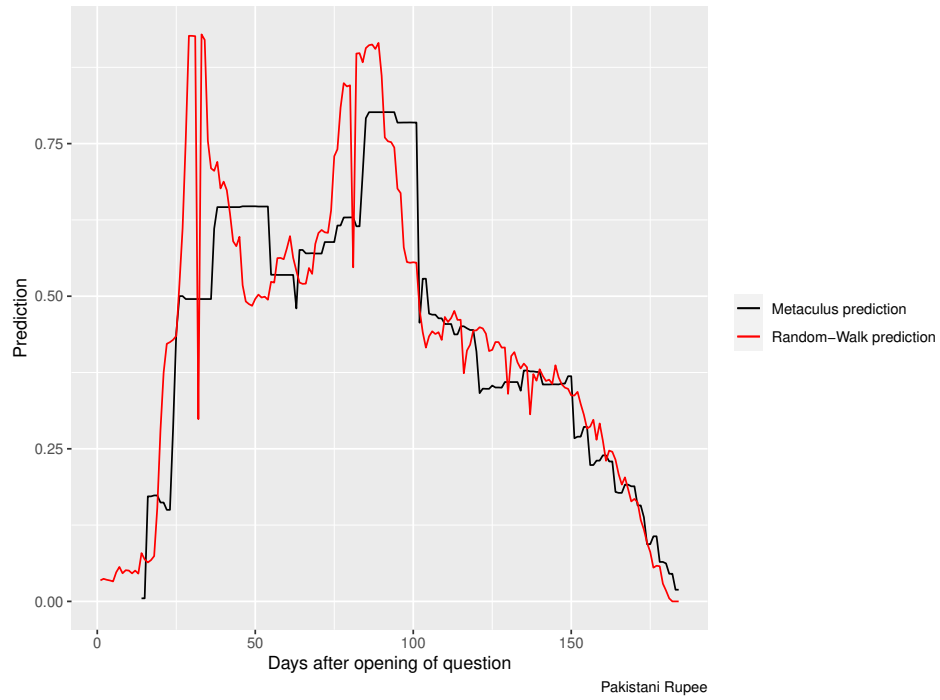Bottom picture: Exchange rate across the same time
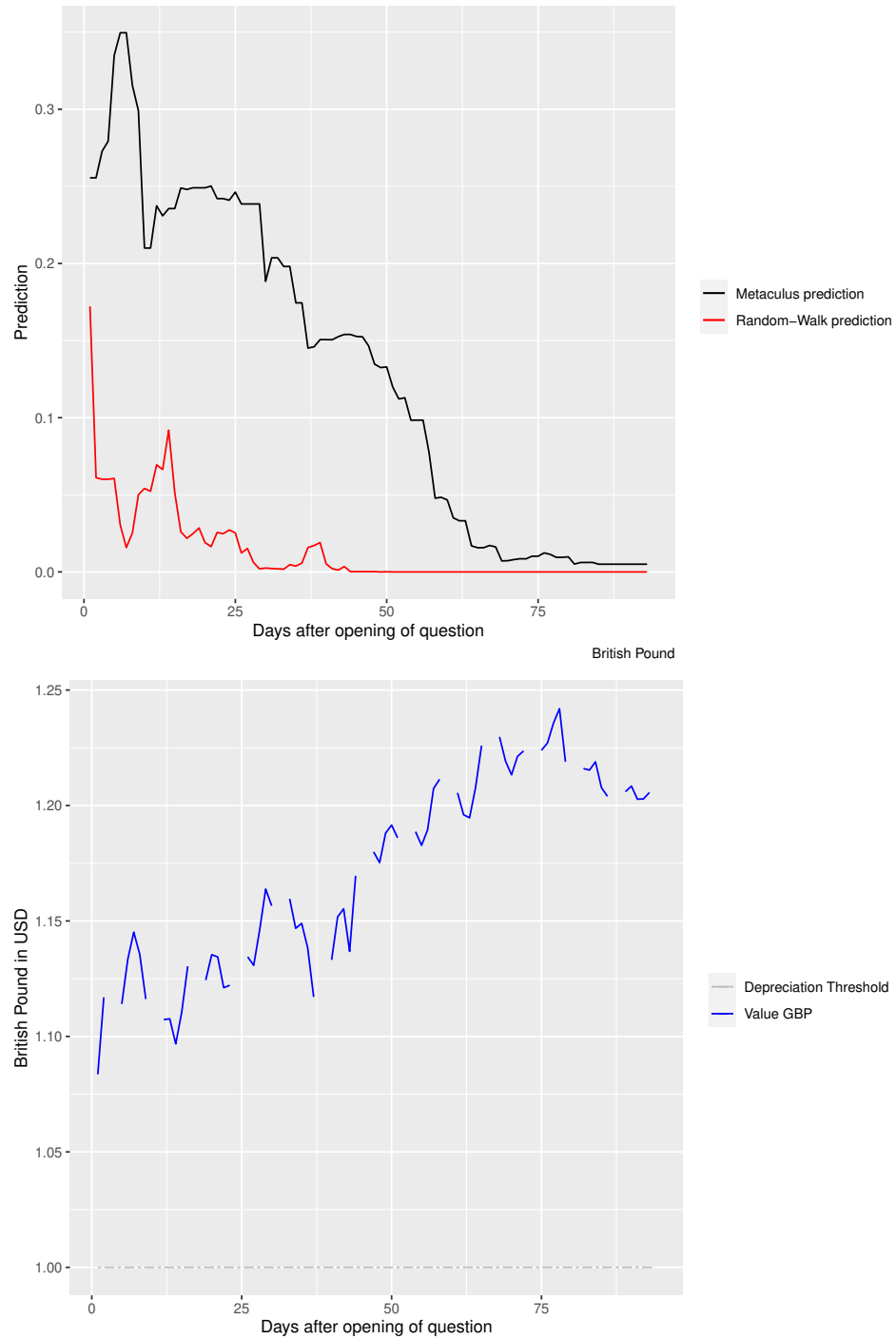
Figure A4: Top picture: Metaculus Prediction compared to the predictions made by the random-walk
Bottom picture: Exchange rate across the same time
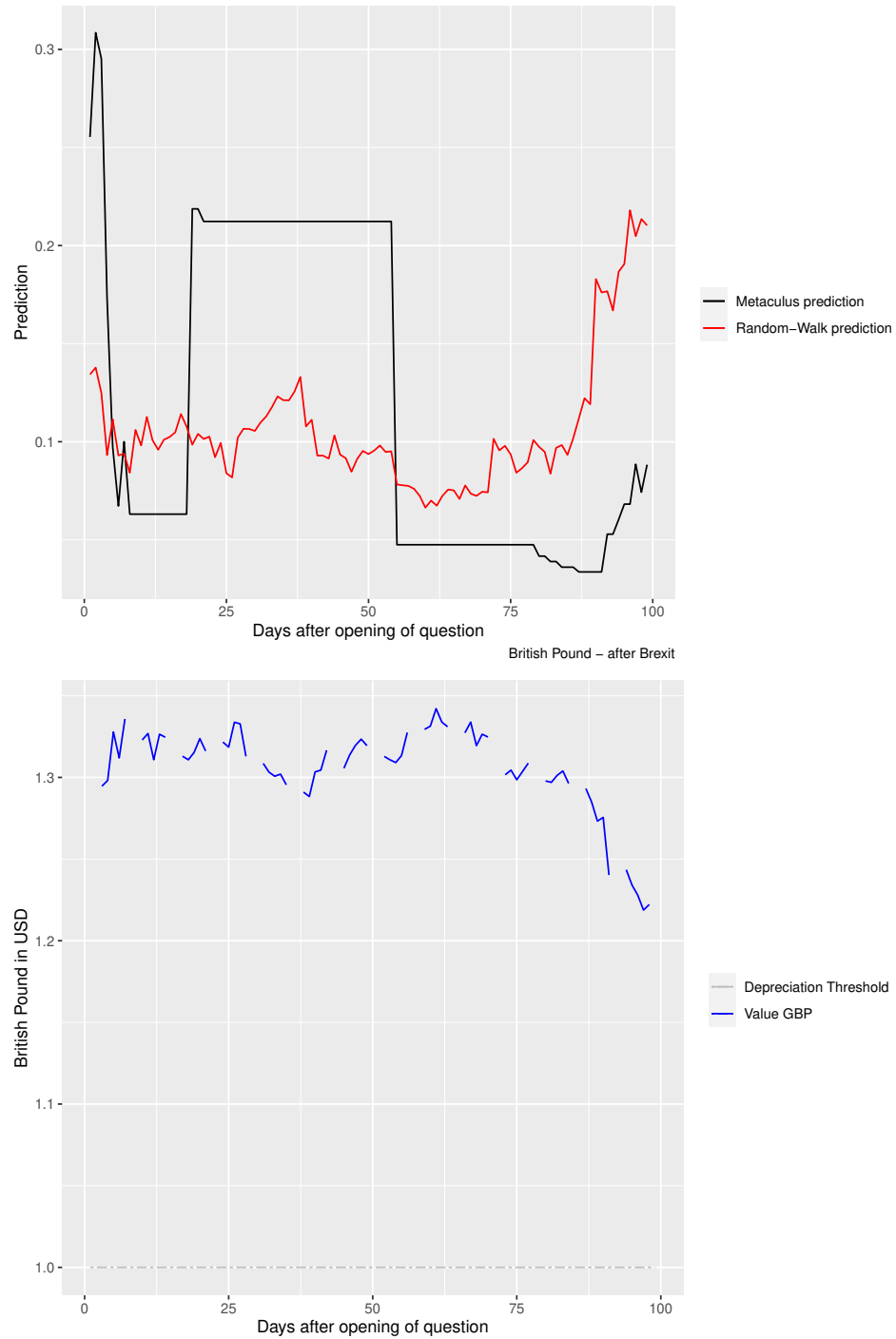
Figure A5: Top picture: Metaculus Prediction compared to the predictions made by the random-walk
Bottom picture: Exchange rate across the same time

Figure A6: Top picture: Metaculus Prediction compared to the predictions made by the random-walk
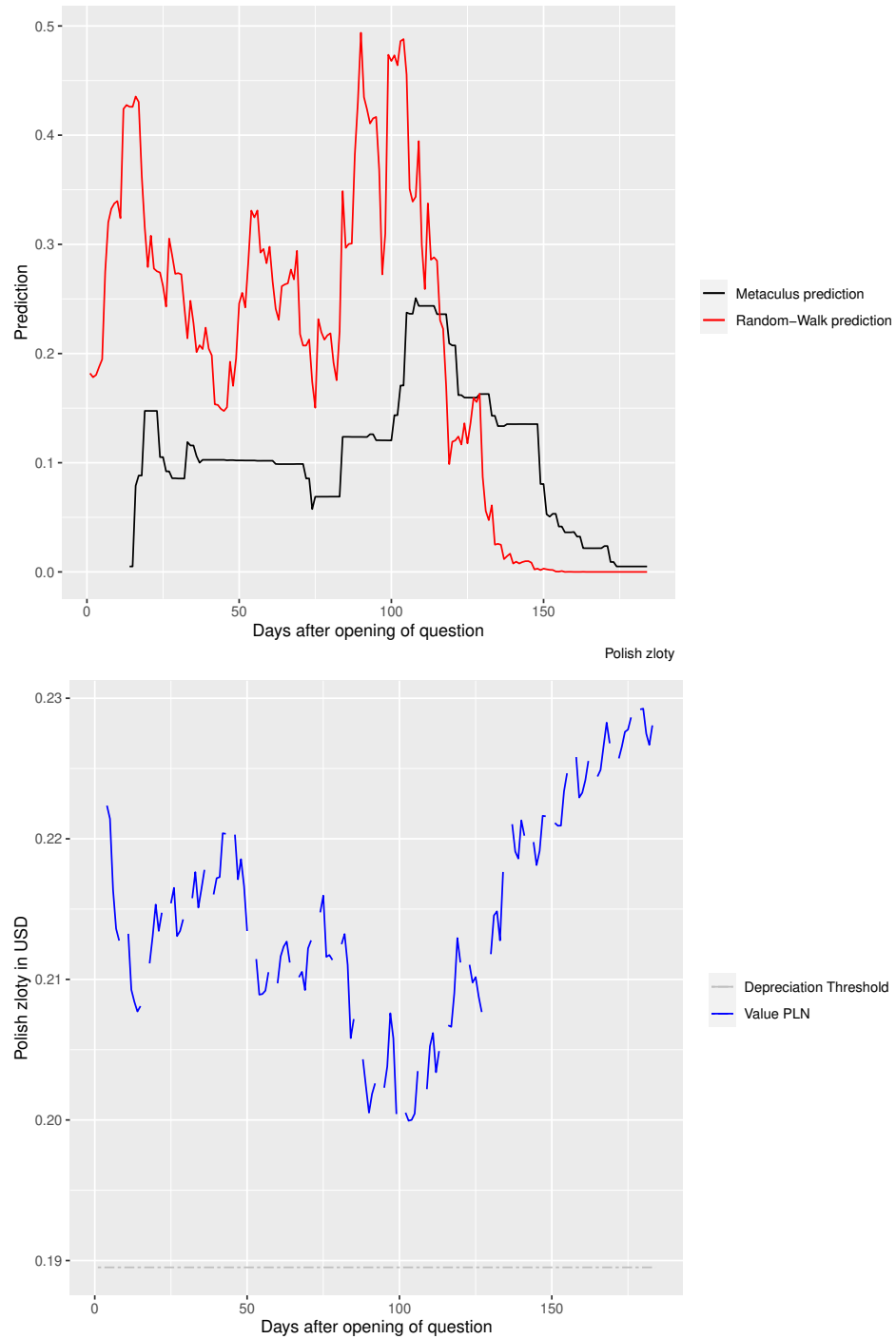Bottom picture: Exchange rate across the same time

Figure A7: Top picture: Metaculus Prediction compared to the predictions made by the random-walk
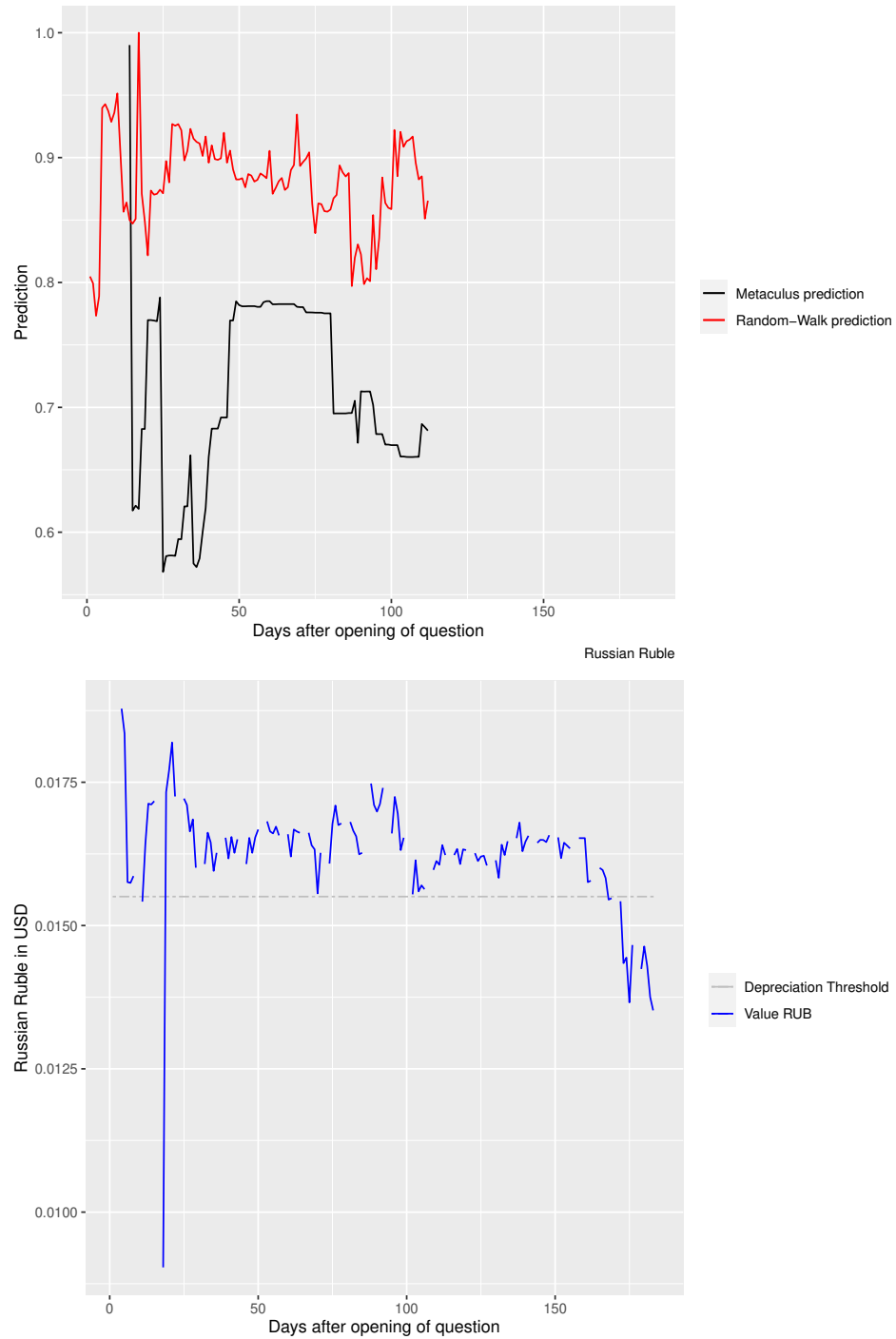Bottom picture: Exchange rate across the same time

Figure A8: Top picture: Metaculus Prediction compared to the predictions made by the random-walk
Bottom picture: Exchange rate across the same time

Figure A9: Top picture: Metaculus Prediction compared to the predictions made by the random-walk
Bottom picture: Exchange rate across the same time

Figure A10: Top picture: Metaculus Prediction compared to the predictions made by the random-walk
Bottom picture: Exchange rate across the same time

Figure A11: Top picture: Metaculus Prediction compared to the predictions made by the random-walk
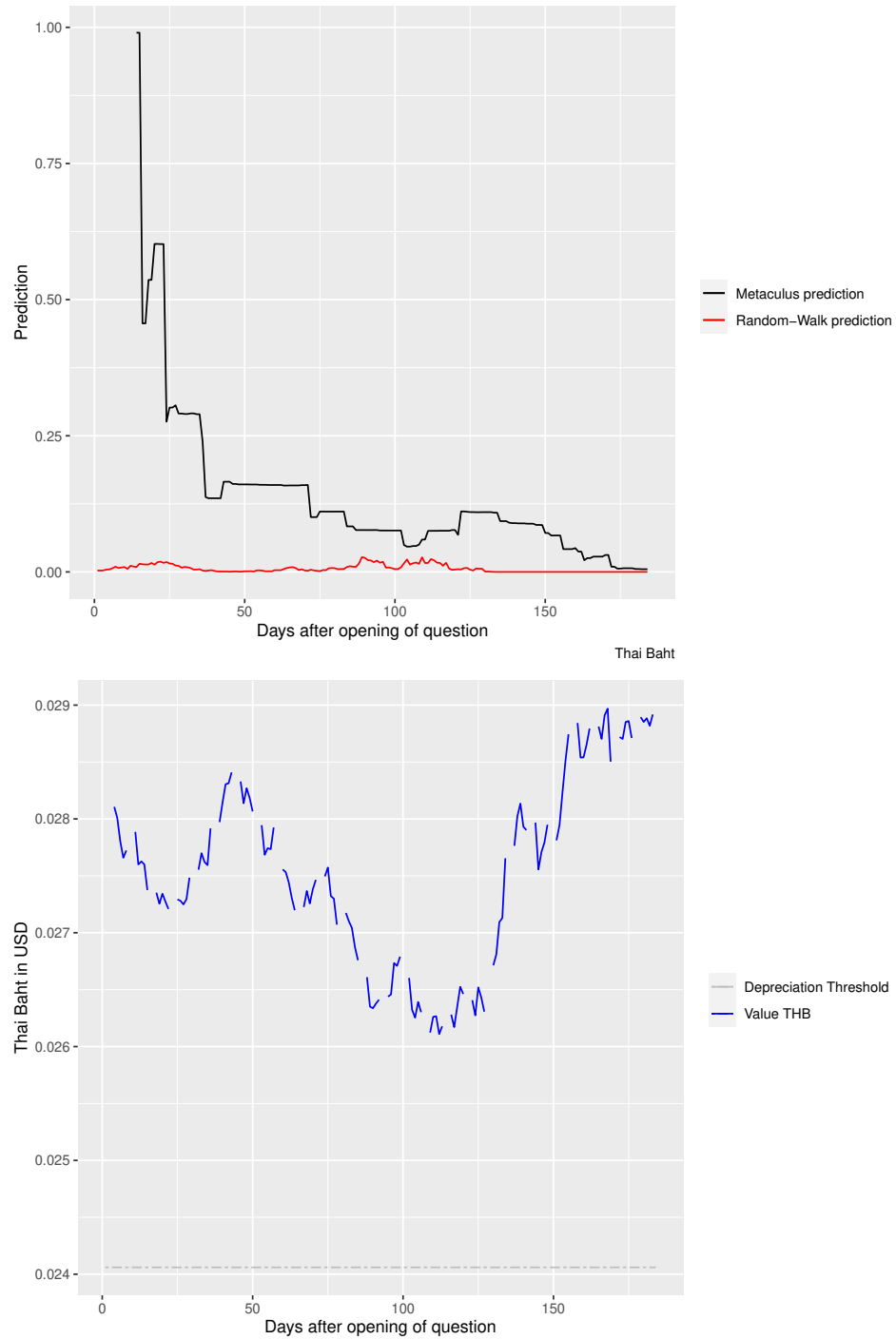Bottom picture: Exchange rate across the same time

Figure A12: Top picture: Metaculus Prediction compared to the predictions made by the random-walk
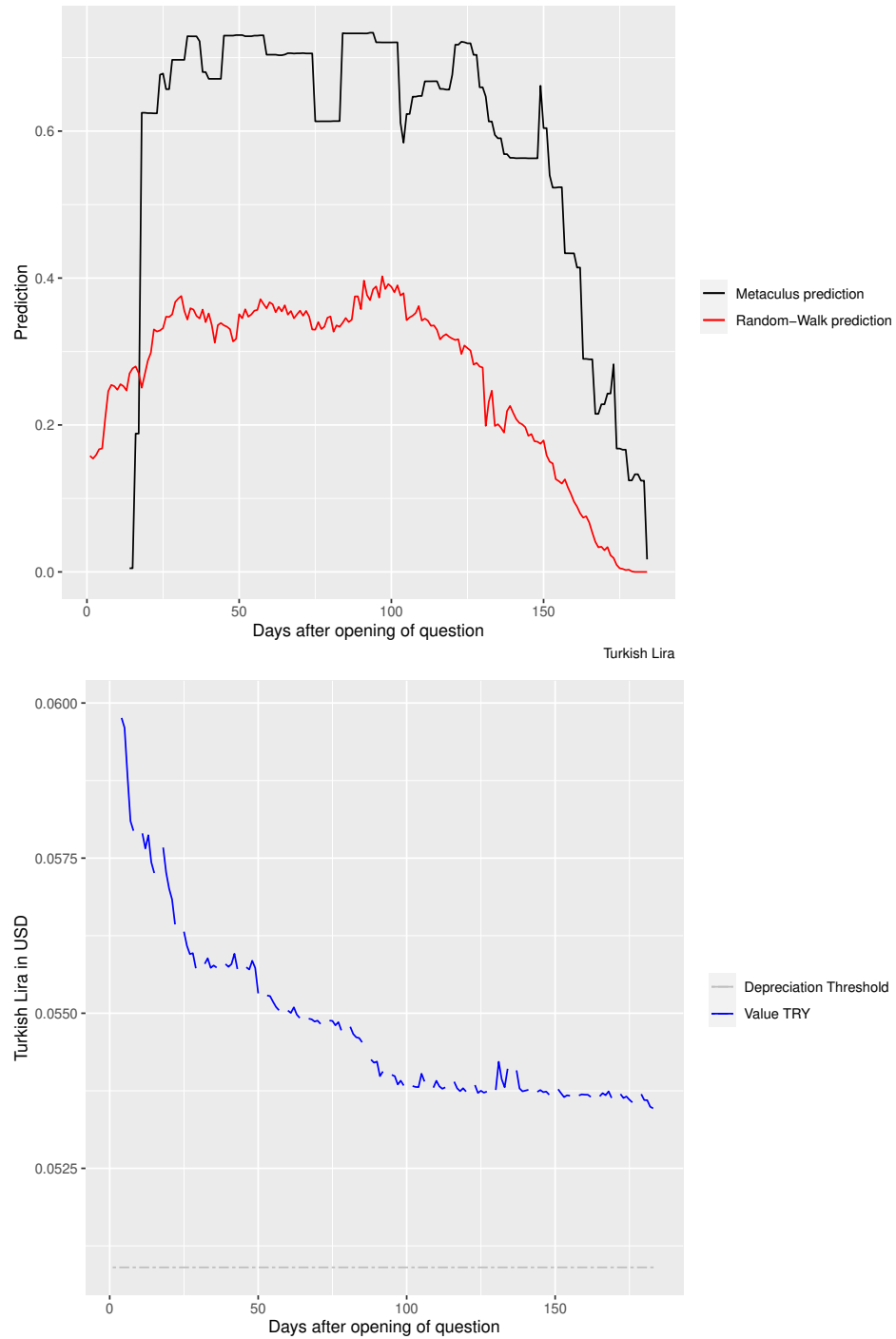Bottom picture: Exchange rate across the same time

Figure A13: Top picture: Metaculus Prediction compared to the predictions made by the random-walk
Bottom picture: Exchange rate across the same time