EFL Economic & Financial Analysis Letters

Research Article
Araştırma Makalesi

Volume 1 • Issue 1 • 2026
e-ISSN: ____-____

When Do Tree-Based Models Fail? Evidence from Energy Inflation Forecasting in Emerging Markets

Erol Can İldeş

Research Assistant, Department of Economics, Faculty of Political Sciences, Ankara Yıldırım Beyazıt University, Esenboğa Campus, Çubuk, Ankara, Türkiye. erolcanildes@aybu.edu.tr • ORCID: 0009-0001-7367-1708

Received: 18.06.2026 • First decision: 01.07.2026 • Accepted: 04.08.2026

Abstract

This paper asks not whether tree-based machine-learning models beat a random walk in forecasting energy inflation, but under which conditions they fail to. We use monthly energy consumer price index data for Türkiye, South Africa, Mexico and Poland for 2002–2022, keeping the information set deliberately small and univariate. One-step-ahead forecasts from a random walk, an AR(1), two linear models (ordinary least squares and ridge) and two tree-based models (gradient boosting and random forest) are compared using the Diebold-Mariano and conditional Giacomini-White tests. On average no estimated model beats the random walk, and in some countries the tree-based models are significantly worse. More important, performance depends on the inflation regime: tree-based models lose accuracy against the random walk when inflation is high. The pattern runs in the same direction across countries and estimation windows, though its pooled significance is sensitive to how cross-country dependence is treated. The evidence fits a structural explanation: tree-based methods cannot predict outside the range of their training data, so their flexibility becomes a disadvantage once energy inflation rises far above past values. During inflation crises in emerging markets, simple benchmarks remain hard to beat, and models should be evaluated conditionally rather than on average.

Keywords: energy inflation; inflation forecasting; machine learning; tree-based models; conditional predictive ability; emerging markets

JEL classification: C53; E31; C52; Q43

Ağaç Tabanlı Modeller Ne Zaman Başarısız Olur? Gelişmekte Olan Piyasalarda Enerji Enflasyonu Tahmininden Bulgular

Öz

Bu çalışma, ağaç tabanlı makine öğrenmesi modellerinin enerji enflasyonu tahmininde rastgele yürüyüşü yenip yenemeyeceğini değil, hangi koşullarda yenmekte başarısız olduğunu sormaktadır. Türkiye, Güney Afrika, Meksika ve Polonya için 2002–2022 dönemine ait aylık enerji tüketici fiyat endeksi verileri kullanılmakta, bilgi kümesi bilinçli olarak dar ve tek değişkenli tutulmaktadır. Rastgele yürüyüş, AR(1), iki doğrusal model (en küçük kareler ve ridge) ve iki ağaç tabanlı model (gradyan artırma ve rastgele orman) ile üretilen bir ay ileriye dönük tahminler, Diebold-Mariano ve koşullu Giacomini-White testleriyle karşılaştırılmaktadır. Ortalamada tahmin edilen hiçbir model rastgele yürüyüşü yenememekte, bazı ülkelerde ağaç tabanlı modeller anlamlı biçimde daha kötü sonuç vermektedir. Daha önemlisi, performans enflasyon rejimine bağlıdır: enflasyon yüksek olduğunda ağaç tabanlı modeller rastgele yürüyüş karşısında doğruluk kaybetmektedir. Örüntünün yönü ülkeler ve tahmin pencereleri arasında tutarlıdır; ancak havuzlanmış anlamlılığı, ülkeler arası bağımlılığın nasıl ele alındığına duyarlıdır. Bulgular yapısal bir açıklamayla uyumludur: ağaç tabanlı yöntemler eğitim verisinin aralığı dışında tahmin üretemediğinden, enerji enflasyonu geçmiş değerlerinin çok üzerine çıktığında esneklikleri dezavantaja dönüşmektedir. Enflasyon krizlerinde basit ölçütleri yenmek güçtür ve modeller yalnızca ortalamada değil koşullu olarak da değerlendirilmelidir.

Anahtar Kelimeler: enerji enflasyonu; enflasyon tahmini; makine öğrenmesi; ağaç tabanlı modeller; koşullu tahmin gücü; gelişmekte olan piyasalar

JEL Sınıflandırması: C53; E31; C52; Q43

1. Introduction

Energy prices are a large and volatile part of consumer inflation in emerging markets (EMs). They also pass through quickly to other prices, so energy inflation forecasts matter for central banks and forecasting units. The task becomes harder when prices move sharply, as they did during the pandemic and the 2021–22 energy crisis. A common way to study this issue is to compare machine learning (ML) methods with simple time-series models and ask which one gives the lowest forecast error. This paper takes a slightly different route. In inflation forecasting, naive benchmarks are often difficult to beat (Atkeson & Ohanian, 2001; Stock & Watson, 2007), and recent macroeconomic forecasting studies show that ML gains are usually modest (Medeiros et al., 2021; Goulet Coulombe et al., 2022). Therefore, another horse race is not very useful unless it explains why a model wins or fails. We ask when tree-based machine-learning methods fail against a random walk (RW), not simply whether they beat one.

Our hypothesis is simple. Tree-based (TB) methods, such as gradient boosting and random forests, cannot predict values outside the range of the training outcomes. This is not a problem when inflation remains close to its historical path. But if a shock pushes inflation far above past observations, tree forecasts are bounded by construction. In that case flexibility may become a disadvantage. If this mechanism matters, model performance should be regime-dependent: models should look similar in normal periods, while TB methods should fall behind during high-inflation episodes. The small information set is part of this design. The aim is not to build the best possible forecasting system by adding oil prices, exchange rates or policy variables. That is a different question. Here we keep the data seen by the models fixed, so that the comparison is mainly about how different functional forms behave when the future moves outside the range of the past.

We test this idea for Türkiye, South Africa, Mexico and Poland using monthly energy CPI data from 2002 to 2022. We did not choose these four countries because they are similar, but because they are structurally different. All four are EMs that target inflation and have their own currency, so the comparison is meaningful. But energy prices are formed in different ways. Türkiye depends heavily on energy imports and the exchange rate passes through quickly. Mexico is an oil producer and a net crude exporter, and it also smooths fuel prices with tax adjustments. Poland is an EU member with its own currency and mostly coal-based electricity, and its energy prices are set within a regulatory framework. In South Africa electricity tariffs and fuel prices are administered, but the currency is volatile. So the same global shock became a very different inflation experience in each country. In 2021–22 energy inflation rose above 170

percent in Türkiye, stayed close to its historical range in Mexico, and was between these two cases in South Africa and Poland. This variation is central to our design. If TB models fail because they cannot predict outside the training range, the failure should be larger when a shock pushes inflation further outside that range. So the differences between the countries are not a problem to remove. They are the variation that allows us to test the mechanism, and we treat the four countries as complementary cases rather than a representative sample of EMs. The information set is deliberately small and uses only each country's own inflation history. We compare a RW, an AR(1), ordinary least squares (OLS), ridge, gradient boosting (GBR) and random forest (RF). Forecast accuracy is evaluated unconditionally with the Diebold-Mariano (DM) test and conditionally with the Giacomini-White (GW) test, under recursive and rolling windows. The main finding is that no estimated model beats the RW on average, while TB models lose more when inflation is high. The contribution is therefore methodological: the paper shows that the ranking of forecasting models depends on the regime, and links this result to the extrapolation limit of TB methods.

We should be clear about the scope of this claim. Our evidence concerns two TB methods, gradient boosting and random forests, in a small univariate information set. It does not cover ML methods in general, and in particular it does not cover regularised linear learners, neural networks or models that use large information sets. The mechanism we study, namely the inability to predict outside the range of the training data, is a property of TB methods, so our conclusions should be read as applying to this model class

2. Conceptual framework

Energy inflation in EMs is difficult to forecast for both statistical and structural reasons. At the monthly frequency, year-on-year (YoY) inflation is highly persistent. A no-change forecast is therefore hard to improve, because an estimated model can help only by finding predictable movements around this persistence. When these movements are small compared with shocks, adding parameters may increase estimation variance more than it lowers bias. This is one reason why naive inflation forecasts remain strong benchmarks.

The problem is stronger in EMs. Energy prices are affected by exchange rate pass through, administered price changes, taxes and subsidies. These factors can create sudden jumps and make the stable estimation sample short. Türkiye, for example, saw YoY energy inflation rise above 170 percent after the 2021 currency crisis, while Mexico remained much

closer to its historical range. This difference is useful for our purpose because the same models face both moderate and extreme cases.

The key point is that extreme observations matter differently across model classes. Regression trees divide the predictor space into regions and forecast the average outcome in each region. As a result, their forecasts are bounded by the range of the target variable in the training sample. This can be useful in normal times because trees avoid extreme extrapolation and can capture mild nonlinear patterns. But when inflation moves outside the training range, the same feature becomes a weakness. The tree forecast is pulled toward the old boundary, while a linear model can extrapolate and a RW automatically follows the latest observed level.

This argument should not be read as saying that TB methods are always weak. They can work well when the evaluation period looks like the training period, and they are often useful when the information set is rich. The point is narrower: in a short macroeconomic sample with a large out-of-support shock, the same flexibility that helps in ordinary times may not protect the model. The RW is crude, but after a large jump it immediately re-anchors to the new level.

This is why average forecast comparisons can be incomplete. Forecasting failures in macroeconomics are often related to structural breaks and parameter instability (Clements & Hendry, 2006; Rossi, 2013). When the data-generating process changes, models fitted closely to the old regime can suffer. A DM test averages losses across calm and crisis periods, so it may miss a state-dependent pattern. For this reason, we also use a conditional test and ask whether the loss difference between a model and the RW becomes larger when inflation is high.

3. Related literature

Many recent studies apply ML methods to macroeconomic forecasting. The general result is that ML can be useful in data-rich environments, but the gains over good and simple models are usually small and not stable (Medeiros et al., 2021; Goulet Coulombe et al., 2022). The reasons that are given most often are short samples, structural breaks, and the difficulty to separate the signal from the noise in aggregate series.

For inflation in particular, the literature has shown for a long time that simple forecasts are hard to beat. Atkeson and Ohanian (2001) found that a naive forecast was competitive with Phillips Curve models for US inflation, and Stock and Watson (2007) showed that inflation became harder to forecast because its predictable part became smaller. Survey studies reach a similar conclusion (Faust & Wright, 2013): the additional value of more complex models is

limited, and combinations and shrinkage methods often work better than a single complex model.

Energy inflation received much less direct attention than headline or core inflation, even if it is a large and volatile part of the consumer basket in EMs. The related literature on oil and commodity price forecasting also finds that simple benchmarks are hard to beat at short horizons (Alquist et al., 2013). This leaves a gap in the literature: the behaviour of ML models for energy inflation in EMs, and especially during the large shocks of these economies, is not well documented.

Regarding the evaluation, formal tests of predictive ability are now standard. The DM test (Diebold & Mariano, 1995), with its small-sample correction (Harvey et al., 1997), tests equal unconditional accuracy, and West (1996) gives the asymptotic theory. The Model Confidence Set (Hansen et al., 2011) compares many models at the same time, and for nested models, where one model is a special case of another, the Clark and West (2007) correction is preferred. Giacomini and White (2006) extended the framework to conditional predictive ability, so the comparison of two models can depend on the state of the economy. Even if these tools exist, many applied comparisons still report only the average accuracy, without a naive benchmark and without conditioning on the forecasting environment (Diebold, 2015).

Our paper is at the intersection of these branches of the literature. Compared to the ML literature, we add a naive RW benchmark and conditional tests. Compared to the inflation forecasting literature, we focus on energy inflation in EMs, where shocks outside the training range are common. And compared to both, we identify the mechanism, that is, the extrapolation failure, behind the regime-dependent underperformance of the flexible models. So our goal is not to find a new winner model, but to give a clearer explanation of when, and why, tree-based methods fail against a simple benchmark.

4. Data and method

4.1 Data

We used monthly energy consumer price indices for the four countries, taken from the FRED database of the Federal Reserve Bank of St. Louis, which distributes the OECD Main Economic Indicators series. Table 1 lists the exact series identifiers. All four series are index numbers with 2015 = 100, they are not seasonally adjusted, and they are published at monthly frequency. We use them in levels and construct inflation ourselves as in equations (1) and (2), so no seasonal adjustment or other transformation is applied by us. Our sample runs from January

2002 to October 2022 and contains no missing observations for any country. These series are subject to revision by the national statistical offices; the data used here were retrieved on 08.10.2025, and the exact vintage is included in the replication files.

$$\pi_t^{YoY} = 100 \times \left(\frac{P_t}{P_{t-12}} - 1 \right) \tag{1}$$

$$\pi_t^{MoM} = 100 \times \left(\frac{P_t}{P_{t-1}} - 1 \right) \tag{2}$$

where P_t is the energy CPI in month t.

The four series are very different in their volatility. In the sample, YoY energy inflation reaches about 172% in Türkiye, and between 28% and 39% in the other three countries. This difference makes the cross-country comparison useful, because the same models meet both moderate and extreme conditions.

Table 1. Data Description[1]

Country	FRED series ID	Units	Adjustment	Frequency
Türkiye	TURCPIENGMINMEI	Index 2015 = 100	Not seasonally adjusted	Monthly
South Africa	ZAFCPIENGMINMEI	Index 2015 = 100	Not seasonally adjusted	Monthly
Mexico	MEXCPIENGMINMEI	Index 2015 = 100	Not seasonally adjusted	Monthly
Poland	POLCPIENGMINMEI	Index 2015 = 100	Not seasonally adjusted	Monthly

4.2 Models and predictors

We forecast YoY energy inflation one month ahead, using only information that is available at the time of the forecast. We keep this information set deliberately small. The predictor vector contains a linear trend, a pair of sine and cosine terms for seasonality, the YoY inflation at lags of one, three, six and twelve months, and the most recent MoM inflation rate, as shown in equation (3):

$$x_t = \left(t, \sin\left(2\pi m_t/12\right), \cos\left(2\pi m_t/12\right), \pi_t, \pi_{t-2}, \pi_{t-5}, \pi_{t-11}, \pi_t^{MoM} \right)^{\top} \tag{3}$$

The last term is the MoM inflation rate defined in equation (2), and the seasonal terms use the calendar month of the forecast. The information set used for forecasting is the one generated by this predictor vector,

[1] *These series are sourced by OECD Main Economic Indicators and they are published in FRED with a heading that is "Consumer Price Indices: Energy".*

$$I_t = \sigma(x_t) \tag{4}$$

and all the models below use exactly the same information set. So the differences between them come only from the functional form, and not from the data that they see.

The RW forecast is simply the last observed value,

$$\hat{\pi}_{t+1}^{RW} = \pi_t \tag{5}$$

The autoregressive benchmark is a first order autoregression,

$$\pi_{t+1} = c + \varphi\, \pi_t + \varepsilon_{t+1} \tag{6}$$

which gives the forecast

$$\hat{\pi}_{t+1}^{AR} = \hat{c} + \hat{\varphi}\, \pi_t \tag{7}$$

The linear model uses the full predictor vector,

$$\pi_{t+1} = \beta_0 + \beta^{\top} x_t + \varepsilon_{t+1} \tag{8}$$

and its parameters are estimated by OLS,

$$(\hat{\beta}_0, \hat{\beta}) = arg\, min_{\beta_0, \beta} \sum_{t}^{T} \left(\pi_{t+1} - \beta_0 - \beta^{\top} x_t\right)^2 \tag{9}$$

$$(\hat{\beta}_0, \hat{\beta})^{Ridge} = arg\, min_{\beta_0, \beta} \left\{ \sum_{t}^{t} \left(\pi_{t+1} - \beta_0 - \beta^{\top} x_t\right)^2 + \lambda \sum_{j}^{t} \beta_j^{\,2} \right\} \tag{10}$$

Ridge regression uses the same predictors, but it adds a penalty on the size of the coefficients, where the penalty parameter is selected by expanding window cross validation, and the intercept is not penalised. A larger penalty shrinks the coefficients towards zero, which can reduce the variance of the forecast.

For GBR and RF, the forecast is a nonlinear function of the same predictors,

$$\hat{\pi}_{t+1} = \hat{f}(x_t) \tag{11}$$

where this function is estimated from the training data. For the RF, the function is the average of many regression trees (Breiman, 2001),

$$\hat{f}^{RF}(x_t) = \frac{1}{B} \sum_{b=1}^{B} T_b(x_t) \tag{12}$$

so the prediction is the mean of the B individual tree predictions. For GBR, the function is a sum of successive regression trees (Friedman, 2001),

$$\hat{f}^{GB}(x_t) = \sum_{m=1}^{M} \nu\, h_m(x_t) \tag{13}$$

where each new tree is fitted to the part that the previous trees did not yet explain, and the learning rate ν controls how much of each tree is added. The hyperparameters of Ridge, GBR and RF are all selected by expanding-window cross-validation on the training sample, and the search grids are reported in Appendix A.6.

4.3 Estimation and evaluation

The predictors are on very different scales, which matters for ridge regression because the penalty is not scale invariant. For this reason the linear and ridge models are estimated on standardized predictors. The standardisation is part of the estimation step, not a preliminary transformation of the whole sample: at every forecast origin the means and standard deviations are computed only from the observations in the estimation window that ends at that origin, and they are then applied to the predictor vector of the forecast month. No information after the forecast origin enters the scaling, so the exercise remains a genuine out-of-sample experiment. The intercept is not penalised. Gradient boosting and random forests are invariant to monotone rescaling of the predictors, so they are estimated on the raw predictors.

Models are trained on 2002–2018 and evaluated out of sample from January 2019 to October 2022, giving 46 one-step-ahead forecasts. The main scheme is recursive: each month, the model is re-estimated using all data available up to that month. As a robustness check, we also use a fixed 120-month rolling window. Accuracy is reported using mean absolute error (MAE) and root mean squared error (RMSE), with RMSE also shown relative to the RW.

We test average accuracy with the DM test and the Harvey et al. (1997) small-sample correction. Since the RW is nested in the AR and linear models, we also report the Clark West test (CW). To test whether performance changes with the inflation regime, we use the GW conditional test with lagged YoY inflation as the conditioning variable. We also split the test sample into normal and high-inflation regimes, where high inflation means that lagged YoY inflation is above the 90th percentile of the training sample. The four countries are pooled with country fixed effects and within-country standardisation for the main conditional test.

$$d(k,t) = \alpha + \beta \cdot z(k,t) + \Sigma(j)\, c(j) \cdot D(j,k) + \mu(k,t) \tag{14}$$

where d(k,t) is the difference between the squared forecast error of the model and the squared forecast error of the random walk for country k in month t, z(k,t) is lagged YoY

inflation, which is also the random-walk forecast, D(j,k) are country dummies, and β is the coefficient of interest. Both d and z are standardized within each country before pooling, so that countries with more volatile energy inflation do not dominate the estimate, and the country dummies absorb differences in the average loss level. A positive β means that the model loses more to the random walk as inflation rises.

Inference in this pooled regression needs care. The four countries were hit by the 2021–22 energy shock at roughly the same time, so the residuals are likely to be correlated across countries within a month as well as over time. Newey-West standard errors deal with serial correlation but not with this contemporaneous cross-country dependence. We therefore report Driscoll-Kraay standard errors, which are robust to both, alongside the Newey-West ones. Because the panel contains only four cross-sectional units, the asymptotic justification for either correction is weak, and we treat the pooled p-values as indicative rather than decisive.

5. Results

5.1 No model beats the random walk on average

Table 2 reports the out-of-sample RMSE. The RW is best in Türkiye and Poland. The linear model is slightly better in South Africa and clearly better in Mexico. The TB models are never the best, and random forest performs especially poorly in Türkiye and Poland.

The statistical tests make this ranking less clear than the raw RMSE numbers suggest. The DM test shows that no estimated model beats the RW at the 5% level. The only p values below 10% are for random forest in Türkiye and Poland, where the RW is more accurate. The Model Confidence Set gives the same message: at the 90% level, all six models remain in the set for every country. Thus, average accuracy alone does not provide strong evidence against the RW.

This is important for the interpretation of the paper. We do not claim that the RW is always the best possible model. Rather, the results show that, with this information set and this short crisis-heavy evaluation window, more flexible estimated models do not produce a reliable average improvement over a very simple benchmark.

This result has an economic reading. In these economies the monthly movement of energy inflation is driven largely by common shocks that are hard to anticipate, such as global energy prices, exchange-rate pass-through and administered-price adjustments, so the component that can be predicted from the country's own past is small. In such an environment

most of the information in the past is concentrated in the most recent observation, and this is the only information the random walk uses. The extra parameters of the estimated models add estimation uncertainty in a short sample, and this cost is not offset by the small predictable signal.

Table 2. Out-of-sample RMSE by model, recursive estimation, 2019:M01–2022:M10.

Country	RW	AR(1)	Linear	Ridge	GBR	RF
Türkiye	**11.17**	12.22	12.80	12.80	12.37	15.55
South Africa	4.25	4.23	**4.20**	4.20	4.51	4.33
Mexico	3.89	3.83	**3.55**	3.57	3.94	3.83
Poland	**2.95**	3.00	3.09	3.09	3.31	3.50

Note: *The best model in each row is in bold. RW = random walk; GBR = gradient boosting; RF = random forest.*

5.2 The estimation scheme drives the “simple beats complex” result

The estimation scheme is important for interpreting the tree models. Appendix Table A1 compares the recursive scheme with a frozen scheme, where each model is estimated only once on 2002–2018. Under the frozen scheme, the tree models collapse in Türkiye: gradient boosting has an RMSE of about 51, compared with about 12.5 for the linear model. Under recursive re-estimation, the gradient boosting RMSE falls to about 12.4.

Figure 1 explains why. In the frozen case, the tree forecasts stay near the initial training-sample maximum of about 34 percent, while actual energy inflation rises above 170 percent. When the surge months enter the recursive training window, the tree forecasts recover. This does not prove that linearity is generally more robust. It shows that part of the 'linear beats trees' result can come from the estimation scheme and from the inability of tree-based models to extrapolate beyond the range of their training data.

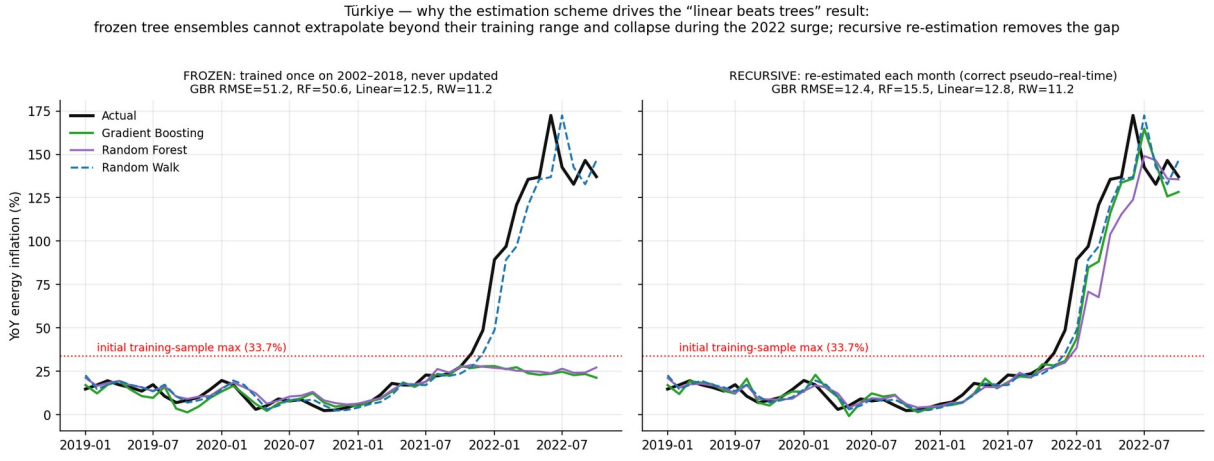

Figure 1. Tree-model forecasts under frozen and recursive estimation, Türkiye.

Note: *Under the frozen scheme (left) the tree models flatten* near the training-sample maximum (about 34%) while actual energy inflation rises above 170%; under recursive re-estimation (right) they recover. Trees cannot predict outside the range of their training data.

5.3 Flexibility loses in the high-inflation regime

The main result appears when performance is evaluated conditionally. Table 3 shows the pooled GW test. All the slopes are positive, which means that every estimated model tends to lose more to the random walk as inflation rises. With Newey-West standard errors, the slope for the tree class is significant under both the recursive and the rolling scheme. When we use Driscoll-Kraay standard errors, which also allow for dependence across countries within the same month, the point estimates are unchanged but the evidence becomes weaker: the pooled tree result is no longer significant in the recursive scheme and only marginally significant in the rolling scheme, while gradient boosting on its own remains significant in the rolling scheme and marginal in the recursive one. We read this as follows. The direction of the effect is stable across specifications and estimation windows, but the pooled significance is fragile, which is what one would expect with four countries hit by a common shock at the same time. The country-level evidence in Appendix Table A3, especially for South Africa, and the regime split below do not depend on pooling.

Table 3. GW conditional test, pooled across countries with country fixed effects

Model vs. random walk	Slope (rec.)	p HAC (rec.)	p DK (rec.)	Slope (roll.)	p HAC (roll.)	p DK (roll.)
Linear	+0.19	0.054	**0.032**	+0.16	0.089	0.062
Ridge	+0.15	0.227	0.229	+0.14	0.260	0.278
Gradient boosting	+0.19	0.088	0.061	+0.22	**0.014**	**0.022**
Random forest	+0.17	0.206	0.303	+0.18	0.161	0.239
Trees (GBR + RF)	+0.18	**0.040**	0.154	+0.20	**0.011**	0.092

Note: *Entries are the slope of the model-versus-random-walk squared-error loss differential on lagged YoY inflation, pooled across countries with country fixed effects and within-country standardisation. "HAC" are Newey-West standard errors with three lags; "DK" are Driscoll-Kraay standard errors with three lags, which are also robust to contemporaneous dependence across countries. A positive slope means the model loses more to the random walk as inflation rises. Values in bold are significant at the 5% level. Rec. = recursive estimation; roll. = rolling window.*

The economic interpretation of this pattern is straightforward. The high-inflation months in our sample are the months in which the jump in global energy prices coincided with rapid currency depreciation and administered-price adjustments, most visibly in Türkiye after 2021, and this combination pushed inflation far outside the range covered by the training sample. In

those months the tree-based models implicitly predict a return towards the historical average: their forecasts are pulled towards the old boundary while inflation is still accelerating. The random walk instead assumes full persistence and re-anchors immediately on the new level, which is closer to the behaviour of the series during a crisis. The Mexican exception is consistent with the same mechanism, because there the shock remained within past experience, so historical patterns kept their informational value and conditional-mean models could still help.

5.4 Robustness and the cross-country picture

Several checks support this interpretation. The CW test (Table 4) shows that OLS and ridge beat the RW in Mexico and are marginally better in South Africa, but not in Türkiye or Poland. This suggests that estimated models add value when the shock is moderate, not when inflation moves far outside its past range. Forecast encompassing tests also point in the same direction: in Türkiye and Poland, the RW encompasses the tree forecasts, while the reverse does not hold. Rolling window results are close to the recursive results, and both are very different from the frozen scheme.

These robustness checks are useful because they do not add new predictors. They ask whether the same message survives when the comparison method changes. The answer is broadly yes: the RW remains hard to beat on average, while the weakness of TB methods appears mainly in the high-inflation part of the sample.

The cross-country evidence should still be read carefully. The sample contains only four countries, and the 2021–22 shock affected them at roughly the same time, so the pooled observations are not fully independent. Türkiye is the clearest visual example, but the conditional result is not driven only by Türkiye; it is statistically strongest for South Africa, while Mexico works as a useful counterexample. For this reason, we read the findings as evidence about a mechanism, not as a universal ranking of forecasting methods. In economic terms, in Türkiye and Poland the forecasts of the flexible models carry no usable information beyond what is already contained in the last observation.

Table 4. Clark–West test for nested comparisons (the random walk nested in the larger model), recursive estimation

Country	RW vs AR(1)	RW vs Linear	RW vs Ridge
Türkiye	0.28 (0.39)	0.03 (0.49)	0.03 (0.49)
South Africa	1.16 (0.12)	1.54 (0.06)	1.54 (0.06)
Mexico	1.08 (0.14)	**2.35 (0.01)**	**2.09 (0.02)**
Poland	−1.30 (0.90)	−0.19 (0.58)	−0.19 (0.58)

Note: *Each cell reports the one-sided CW statistic and its p-value; a positive, significant value means the larger model beats the random walk. Significant values ($p < 0.05$) in bold.*

6. Conclusion

This paper asked under which conditions tree-based models fail to beat a RW in forecasting EM energy inflation. The answer is regime-dependent. On average, no estimated model is significantly better than the RW. During high-inflation episodes, however, TB models fall behind because their forecasts are bounded by the training range. The evidence is therefore consistent with an extrapolation-based explanation.

The results have a simple lesson for applied forecasting. During inflation crises, simple benchmarks should not be treated as weak competitors. Forecasting models should be checked against a RW, tested formally, and evaluated separately across regimes. The estimation scheme should also be stated clearly, because frozen and recursive schemes can give very different impressions of TB models. For policy users, this means that a sophisticated model should not be trusted only because it is flexible; it should be checked precisely in the periods where the economy leaves its usual range.

These findings can be beneficial suggestions for policy institutions. First, central banks and forecasting units should report model-based energy inflation forecasts together with a benchmark. They also should monitor performance separately across regimes rather than only on average. Second, the distance between current inflation and the range covered by the estimation sample can be used as a simple early-warning indicator: when inflation approaches or exceeds that range, the reliability of flexible models falls, so their weight should be reduced and the models should be re-estimated more frequently. Third, since model uncertainty is larger precisely during crises, forecast combinations and scenario-based communication are preferable to a single point forecast in those periods. Fourth, the estimation scheme, that is whether the model is frozen, rolling or recursive, should be stated explicitly in official reporting, and forecasting infrastructure should be stress-tested against out-of-range scenarios before a crisis rather than during one.

The study has limits: four countries, one common shock and a one-month horizon. Future work can extend the sample, include more EMs, study longer horizons and compare robust forecast combinations, while keeping the same parsimonious design.

Declarations

Funding. The author did not receive support from any organization for the submitted work.

Conflict of interest. The author declares no conflict of interest and no financial interests.

Data and code availability. The datasets and Python codes used in this study are available from the corresponding author upon reasonable request.

Use of generative AI. During the preparation of this manuscript the author used GPT-4o and QuillBot to improve language and readability. The author subsequently reviewed and revised the text and accepts full responsibility for the published content.

Ethics approval. The study uses secondary, publicly available macroeconomic data and did not require ethics committee approval.

Authorship. Single-authored study.

References

Alquist, R., Kilian, L., & Vigfusson, R. J. (2013). Forecasting the price of oil. In G. Elliott & A. Timmermann (Eds.), Handbook of economic forecasting (Vol. 2A, pp. 427–507). Elsevier.

Atkeson, A., & Ohanian, L. E. (2001). Are Phillips curves useful for forecasting inflation? Federal Reserve Bank of Minneapolis Quarterly Review, 25(1), 2–11.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. Journal of Econometrics, 138(1), 291–311.

Clements, M. P., & Hendry, D. F. (2006). Forecasting with breaks. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), Handbook of economic forecasting (Vol. 1, pp. 605–657). Elsevier.

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), 1–9.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics, 13(3), 253–263.

Faust, J., & Wright, J. H. (2013). Forecasting inflation. In G. Elliott & A. Timmermann (Eds.), Handbook of economic forecasting (Vol. 2A, pp. 2–56). Elsevier.

Federal Reserve Bank of St. Louis. (2024). Consumer price index: Energy (OECD main economic indicators) [Data set]. FRED. https://fred.stlouisfed.org

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

Giacomini, R., & Rossi, B. (2010). Forecast comparisons in unstable environments. Journal of Applied Econometrics, 25(4), 595–620.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. Econometrica, 74(6), 1545–1578.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., & Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? Journal of Applied Econometrics, 37(5), 920–964.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. Econometrica, 79(2), 453–497.

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

Hendry, D. F. (2006). Robustifying forecasts from equilibrium-correction systems. Journal of Econometrics, 135(1–2), 399–426.

Medeiros, M. C., Vasconcelos, G. F. R., Veiga, Á., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. Journal of Business & Economic Statistics, 39(1), 98–119.

Rossi, B. (2013). Advances in forecasting under instability. In G. Elliott & A. Timmermann (Eds.), Handbook of economic forecasting (Vol. 2B, pp. 1203–1324). Elsevier.

Stock, J. H., & Watson, M. W. (2007). Why has U.S. inflation become harder to forecast? Journal of Money, Credit and Banking, 39(s1), 3–33.

West, K. D. (1996). Asymptotic inference about predictive ability. Econometrica, 64(5), 1067–1084.