ИКОНОМИЧЕСКИ ИЗСЛЕДВАНИЯ
ECONOMIC STUDIES
Volume 35(6), 2026

Dezdemona Gjylapi[1]
Alketa Hyso[2]
Filloreta Madani[3]

FORECASTING QUARTERLY GDP GROWTH AND THE GDP DEFLATOR IN ALBANIA UNDER DATA SCARCITY: A COMPARATIVE EVALUATION OF STATISTICAL AND MACHINE LEARNING MODELS[4]

The purpose of this paper is to evaluate forecasts of Albania's Real GDP Growth and GDP Deflator under real-time operational constraints by conducting an operational backtest using the leakage-free rolling origin approach and mixed-frequency predictors. The main difficulty is the small number of observations in the quarterly data set, partially observed predictors, and a 90-day publication delay for GDP. In this study, we evaluate 19 models grouped into six families: baseline univariate rules, classical time-series models, dynamic regression/SARIMAX specifications, regularised linear models, bridge/factor-augmented models, and selected nonlinear/probabilistic machine-learning models. These approaches were applied individually to Real GDP Growth and the GDP Deflator to generate point forecasts and forecast intervals for 1-, 2-, and 4-quarter-ahead horizons, using metrics such as MAE, RMSE, interval accuracy and coverage, interval width, and CRPS. Robustness was assessed by conducting Diebold-Mariano tests, applying the block bootstrap, and treating the period after 2016 as an out-of-sample period. The three highest-ranking models are univariate rules that converge toward the expanding-window in-sample mean. This suggests mean reversion in GDP growth and persistence in the GDP deflator. Among the structured models, Bridge_PCA_Ridge is the most consistent option at the 1- and 2-quarter horizons, while SARIMAX is the strongest structured specification at H=4. Selected nonlinear machine-learning models did not improve upon the regularised linear alternatives in this sample.

Keywords: mixed-frequency forecasting; quarterly GDP; rolling-origin backtesting; regularized regression; bridge pca
JEL: C22; C53; E37; O47

---

[1] *Dezdemona Gjylapi, Lecturer, Dr. (Doctor of Sciences) in Informatics, Department of Computer Science, "Ismail Qemali" University of Vlora, Albania, e-mail: dezdemona.gjylapi@univlora.edu.al, ORCID: 0009-0005-5186-3636.*
[2] *Alketa Hyso, Head of the Department of Computer Science, Dr. (Doctor of Sciences) in Computer Engineering, Department of Computer Science, "Ismail Qemali" University of Vlora, Albania, e-mail: alketa.hyso@univlora.edu.al, ORCID: 0009-0000-6256-9831.*
[3] *Filloreta Madani, Lecturer, Associate Professor, Department of Finance and Accounting, "Ismail Qemali" University of Vlora, Albania, e-mail: filloreta.madani@univlora.edu.al, ORCID: 0009-0000-8363-4743.*
[4] This paper should be cited as: *Gjylapi, D., Hyso, A., Madani, F. (2026). Forecasting Quarterly GDP Growth and the GDP Deflator in Albania under Data Scarcity: A Comparative Evaluation of Statistical and Machine Learning Models. – Economic Studies (Ikonomicheski Izsledvania), 35(6), pp. 107-131.*

1. Introduction

Gross Domestic Product (GDP) forecasting is important for fiscal planning and short-term macroeconomic assessment. In Albania, this task is especially challenging due to a clear institutional constraint. INSTAT releases quarterly GDP data with a 90-day lag after the reference quarter ends. No flash estimates are made available during this time (Çeliku, Kristo, 2022).

This means that policymakers and analysts lack any official GDP figures for an entire quarter. Meanwhile, the available information is limited and uneven. Hard data are scarce, mainly covering foreign trade, partial fiscal data, and electricity output. Other real-sector data appear close to the official GDP release, and there is no agricultural data. In this context, the Bank of Albania developed short-term GDP forecasting as a key part of its monetary policy. It uses bridge-type and factor-based models mixing monthly economic, fiscal, financial, and survey indicators to produce nowcasts and short-horizon forecasts before the quarterly GDP release (Çeliku, Kristo, Boka, 2009; Çeliku, Kristo, 2022).

Moreover, this problem is exacerbated by the evolution of the information setting. With few observations per quarter and structural breaks in the data, finding consistent predictors is challenging, particularly given the introduction of new predictors, such as business sentiment indices available from 2016 (Çeliku, Kristo, 2022). Consequently, achieving good performance is even more critical here compared to merely fitting the model, as overly complicated models may overfit the data (Pesaran, Timmermann, 2004; Harding, Hersh, 2018; Makridakis, Spiliotis, Assimakopoulos, 2018). Of note, there is the benefit of utilising mixed-frequency modelling, as monthly predictors can highlight both contemporaneous and future activity relative to quarterly GDP (Cascaldi-Garcia, Luciani, Modugno, 2023). Regarding Albania, prior studies have found that the informativeness of indicators varies over time: surveys predict GDP one to two quarters ahead, while economic and fiscal indicators lag (Çeliku, Kristo, Boka, 2009). Analysis of the error term suggests an increase in the error rate at a longer time horizon due to uncertainties associated with the predictors, and this calls for caution (Çeliku, Kristo, 2022). Another important issue that cannot be overlooked is data quality – vintage analysis shows that GDP was not far from the true value for comparison purposes, yet required considerable adjustments to investment and imports (Metani, Çeliku, 2022).

At this point, it is useful to clarify the scope of short-term GDP forecasting considered in this study. Specifically, this study aims to advance the practice of short-term GDP forecasting in Albania through testing various kinds of models (statistical and machine learning) using a leakage-free rolling-origin strategy. This paper deals with two kinds of forecast tasks: quarterly Real GDP Growth and GDP Deflator, both of which are separately modelled due to their distinct characteristics. While the first indicator exhibits high volatility and mean reversion, the latter is characterised by persistence and regime-dependence. As regards limitations, since in this case we consider only the Albanian data, the issues of concern include those related to a small sample size, mixed-frequency indicators, missing values, regime change, and the necessity to perform out-of-sample validation. This research continues the tradition at the Bank of Albania to evaluate the forecast accuracy of linear models, such as bridges, factors, and SUR system against simple SARIMA benchmarks in RMSE (Çeliku, Kristo, Boka, 2009; Çeliku, Kristo, 2022).

Having discussed the empirical environment and evaluation criteria employed in this study, it is reasonable to move to the research question. Which principles of forecasting and model combinations work better in the context of forecasting Real GDP Growth and GDP Deflator when taking into account the specificities of small samples, mixed-frequency predictors, and strict out-of-sample tests? Rather than simply comparing statistical and machine-learning models, the main issue is to identify which modelling principles remain reliable under small-sample and mixed-frequency constraints.

To provide an answer to the above question systematically, four research objectives are investigated. First, the forecasting database for Albania is created based on the quarterly GDP and multiple mixed-frequency predictors that encompass various economic, fiscal, financial, and survey data series. Second, a diverse set of model approaches, such as traditional univariate and classical time-series rules, regularisation techniques, bridge/factor-based models, as well as nonlinear approaches, are evaluated in comparison to the SARIMA reference benchmark by the BoA. Specifically, all approaches are independently applied to each target series. Third, the reliability of all forecast estimates is ensured through expanding window backtesting using strictly concurrent data to ensure there are no look-ahead biases (Hyndman, Athanasopoulos, 2021). Finally, the performance of both point and probabilistic forecasts of all approaches is assessed jointly in terms of MAE, RMSE, coverage rates, interval widths, and CRPS.

Together, these components shape what this paper offers in empirical terms: a systematic comparison of forecasting approaches across two Albanian GDP targets – Real GDP Growth and the GDP Deflator – conducted under a strictly leakage-free protocol and assessed not only through point error metrics but through the lens of predictive uncertainty as well. It is important to emphasise that this paper's insights pertain specifically to the Albanian experience and serve as a case study rather than a template for other small economies.

The paper is organised as follows. Section 2 reviews the literature. Sections 3 and 4 outline the data and framework. Section 5 describes the metrics. Sections 6 and 7 present and discuss results. Section 8 concludes.

2. Literature Overview

2.1. GDP forecasting under data scarcity

Quarterly GDP forecasting in small samples is a small-sample forecasting problem where parsimony of modelling and reliable evaluation matter more than goodness of fit in large samples. In such conditions, simple univariate models like ARIMA, SARIMA, and ETS remain important thanks to their lower informational requirements and ability to impose useful structures on persistence, trends, and seasonality of time series (Box et al., 2015; Hyndman, Athanasopoulos, 2021). They are also methodologically useful, as they can show how additional classes of models provide new information beyond GDP persistence and seasonal cycles. This is directly applicable to the Albanian case, where the quarterly GDP time series is short, and ARIMA/SARIMA-type specifications were a natural benchmark for evaluating other models (Çeliku, Kristo, Boka, 2009; Çeliku, Kristo, 2022).

With the introduction of additional predictors, the main problem becomes one of dimensionality. Multivariate models tend to overfit quickly when the number of predictors is too high, they are highly correlated, and the dataset is short. Thus, shrinkage, factor compression, and state-space modelling emerge as the main strategies for overcoming potential estimation problems when data are scarce (Bai, Ng, 2002; Stock, Watson, 2002; Maccarrone, Morelli, Spadaccini, 2021). Empirical evidence from small European countries indicates that a parsimonious or well-regularised specification is generally preferred to larger alternatives when the quarterly sample is short and the predictor pool is large (Beņkovskis, 2008; Barhoumi et al., 2011; Tóth, 2017). For Albania, this means that having more predictors does not necessarily yield better forecasts, since out-of-sample performance matters.

2.2. Purely univariate versus models with other predictors included

The first group includes naïve benchmarks, AR models, ARIMA/SARIMA type models, and ETS models. The advantages of those models include the simplicity and transparency of the forecast, and they can perform surprisingly well when the sample is short and the signal-to-noise ratio is low. In particular, SARIMA models have been used in Bank of Albania studies as a reference model against which other models must demonstrate superior performance (Çeliku, Kristo, Boka, 2009; Çeliku, Kristo, 2022). Similarly, the earlier research on model comparison showed that the VAR model performed better than both ARIMA and bridge-type models during a pseudo-real-time evaluation horizon in Albania (Shahini, Haderi, 2013). The latter shows that model ranking depends on sample size and forecast horizon in Albania as well.

In addition to the purely autoregressive models mentioned above, Nerjaku and Sinaj applied machine learning techniques in their study of GDP forecasting in Albania. They used the unemployment rate as a regressor, thereby enriching existing knowledge of GDP forecasting using machine learning techniques in the Albanian context, yet lacking a leakage-proof multi-horizon validation process. Overall, the aforementioned Albanian evidence confirms the appropriateness of evaluating the performance of all types of models, from benchmarks to the application of machine learning approaches, as done in the current research (Nerjaku, Sinaj, 2024).

The second category comprises models that employ additional predictors, most often at different frequencies. These predictors can include indices of manufacturing output, trade, fiscal variables, interest rates, construction, and business/consumer sentiment surveys. The literature shows that such models tend to be especially effective for nowcasting and short-term forecasting, as they incorporate information that becomes available before the official release of GDP quarterlies (Arnostova et al., 2011; Stakėnas, 2012; Rusnák, 2013; Ljubišić, 2025). However, their effectiveness depends on parsimony or proper compression of the space of the predictors used. It cannot be assumed that large datasets are inherently beneficial for forecasting. In small-sample situations, smaller and better-structured models can prevail (Stakėnas, 2012; Feldkircher et al., 2015).

Regarding Albania, the second direction is especially relevant. According to Çeliku, Kristo, and Boka (2009), economic/financial/survey indicators play distinct roles in forecasting quarterly GDP: survey-based indicators enter earlier and provide leading information about near-term developments in GDP, whereas 'hard' economic data like fiscal expenditures, electricity generation, and construction permits usually enter with longer lags. Moreover, contemporary findings suggest that models utilising both types of indicators perform better than the baseline SARIMA model across various forecast horizons and can even outperform the naïve forecast, with the factorial model achieving the minimum RMSE during the 2015-2018 period (Çeliku, Kristo, Boka, 2009; Çeliku, Kristo, 2022). The horizon dependence of model ranking is an empirical fact that must be considered when designing the evaluation procedure for the current study.

2.3. Machine learning for macroeconomic forecasting

Machine learning methods have become common tools in macroeconomic forecasting thanks to their ability to detect non-linear relations, interactions, and regime-specific dynamics. However, their performance is highly sample-dependent. Using short time-series samples, high-dimensional models are prone to overfitting noise rather than stable structures in the economy, leading to poor out-of-sample predictive accuracy. The broader forecasting literature also indicates that more complicated machine learning models do not necessarily perform better than simple benchmark methods, especially in low-data environments (Harding, Hersh, 2018; Makridakis et al., 2018).

Regularised and parsimonious models are often more reliable in situations where macroeconomic models face limited sample sizes. Linear estimators subject to regularisation, i.e., Ridge (MF), Lasso (MF), and ElasticNet (MF), lower the variance of estimated coefficients by shrinking parameters when the number of predictors is large relative to the sample size (Hoerl, Kennard, 1970; Tibshirani, 1996; Zou, Hastie, 2005). Empirical studies on GDP nowcasts for similarly sized small economies – Serbia (Glišić, 2024), Croatia (Ljubišić, 2025), and Slovenia (Masten et al., 2008) – show that factor models and regularised regression outperform the more flexible ML methods, with possible gains being marginal and dependent on the evaluation period and sample size. Parsimonious models are more appropriate for small samples, and their value needs to be determined by out-of-sample results rather than being presumed a priori.

Probabilistic ML is another theme of recent works. The use of kernel functions in Gaussian Process Regression (GPR) incorporates regularisation into the procedure and provides the full probability distribution as a result. Therefore, this approach seems useful in situations where uncertainty matters as much as accuracy (Rasmussen, Williams, 2005; Görtler, Kehlbeck, Deussen, 2019). There have been encouraging applications of mixed-frequency GPR to GDP nowcast estimation (Hauzenberger et al., 2024). This makes GPR a theoretically intriguing choice to complement regularised linear methods in the present study.

2.4. Real-time data, data revisions, and probabilistic assessment

Another aspect highlighted by the literature on GDP forecasting is that assessments should replicate the conditions under which models are used in forecast production. Information about GDP publication delay, ragged edges due to missing values, and GDP revisions affects what information the forecaster knows at the time when the forecast is produced, and hence the relevance of ex-post assessment based solely on revisions of initial forecasts to the real situation (Rusnák, 2013; Franta, Havrlant, Rusnák, 2016; Pleşa, Sîrbu, Tănase, 2025). For Albania, Metani and Çeliku (2022) investigate quarterly GDP revisions for the period 2015-2020 and find that revisions for early vintages are typically insignificant for aggregates, thereby diminishing one source of uncertainty in evaluation. They also note that some expenditure components show more pronounced changes, which affect the choice of disaggregated predictors for bridge-type model structures.

Finally, the growing emphasis in recent literature on the quality of predictive distributions is another important aspect. For forecasts meant to guide policy, not only the accuracy of point prediction but also the credibility of the uncertainty measure are important, which has led to the increasing importance of coverage rates, proper scoring rules, and other aspects of predictive distributions (Gneiting, Raftery, 2007; Hyndman, Athanasopoulos, 2021). Existing empirical GDP nowcasting for Albania has focused exclusively on evaluating model accuracy using point error metrics such as RMSE and mean forecast errors (Çeliku, Kristo, Boka, 2009; Çeliku, Kristo, 2022).

2.5. Research gap and positioning

Against this background, three research gaps motivate the current research. First, there has been no systematic benchmarking exercise that would incorporate regularised and various nonlinear models, particularly GPR models (Çeliku, Kristo, Boka, 2009; Çeliku, Kristo, 2022). Secondly, the literature suggests that model performance varies across countries, underscoring the importance of an Albania-focused benchmarking study (Feldkircher et al., 2015). Thirdly, probabilistic evaluation has not been considered so far in Albania, although in a policy-oriented forecasting application, intervals need to be properly calibrated and sharp, regardless of RMSE or MAE (Gneiting, Raftery, 2007; Hyndman, Athanasopoulos, 2021).

Accordingly, this paper sets out to undertake four distinct aims. The first aim is to benchmark parsimonious univariate models that use historical data on the target variable. The second is to test the mixed-frequency models using the currently available monthly and quarterly predictors. The third is to examine regularised linear and selected non-linear models to find whether flexible models have forecasting ability when sample sizes are small. The fourth aim is to examine forecast accuracy in a rolling origin setup with no information leakage.

3. Data Description

This section describes the target variables, the explanatory predictors, and the preprocessing pipeline that integrates mixed-frequency information into the forecasting design. The section also explains the economic data environment in Albania that motivates the choices made.

3.1. Albania's Economic Data Environment

As already mentioned in the Introduction, INSTAT provides quarterly GDP statistics with a reporting delay of 90 days without providing any flash estimates; at the same time, the monthly data set contains trade, budget, and electricity series, but business sentiment indexes are only available starting from May 2016 (Çeliku, Kristo, 2022). This combination of various data frequencies and structural changes within the data set determines all aspects of model building.

3.2. Target Variables

We forecast two separate quarterly targets: real GDP growth and GDP deflator. Real GDP growth is primarily driven by demand-side and supply-side real activity, while the GDP deflator responds to monetary, cost-push, and external price dynamics (Çeliku, Kristo, 2022). Forecasting the two components separately enables a more economically interpretable evaluation and avoids the situation where a model that merely tracks a nominal price trend is rewarded as though it were capturing cyclical fluctuations.

The two target variables are: (i) Real GDP Growth, defined as the year-on-year percentage change in real quarterly GDP, which captures the cyclical and structural dynamics of output; and (ii) GDP Deflator, defined as the implicit price deflator derived from nominal and real GDP series, which summarises aggregate price dynamics. Both series are observed only at quarter-end months – March, June, September, and December – and the dataset construction preserves this property (no forward-filling into non-quarter months), thereby avoiding temporal leakage and keeping the forecasting setting consistent with real-world availability.

The sample covers data from Q2 2002 to Q4 2024, yielding 91 observations in total. After generating features for lagged variables and targets corresponding to horizons $H \in \{1, 2, 4\}$ quarters into the future, subject to basic validity conditions (one-quarter and four-quarter GDP lag variables should not be missing), we are left with 83-87 observations depending on the horizon and validity constraints.

3.3. Explanatory Variables

The explanatory side of the dataset follows a mixed-frequency design. Predictors are selected to cover the principal channels through which near-term activity information reaches the forecast origin, while remaining operationally available before the official quarterly GDP release. All predictors are sourced from INSTAT, the Ministry of Finance of Albania, and the Bank of Albania, and their availability is documented in the dataset feature log embedded in the forecasting pipeline.

The list of monthly predictors in the dataset includes: goods imports, goods exports, budget expenditures, the lending rate for EUR-denominated loans, and additional activity indicators. Quarterly construction permits are among the quarterly predictors. These predictors represent external demand, public sector activity, the financial environment, and investment activity.

An important feature of the data environment is the structural change in the availability of monthly predictors. While GDP and fiscal indicators span the full sample from 2002, business confidence survey indicators and broader monthly activity indicators became available only from May 2016 onward. This creates two distinct information regimes: a pre-2016 period in which only a limited set of monthly series is available, and a post-2016 period in which the full set of monthly predictors is usable. The forecasting pipeline handles this by explicitly encoding structured missingness (see Section 3.4) and including a regime dummy (post_2016m05) as a feature. This issue is examined through the post-2016 sub-sample robustness analysis described in Section 4.7 and reported in Section 6.8.

3.4. Preprocessing

Preprocessing integrates monthly information into the quarterly forecasting framework without imposing strong within-quarter aggregation assumptions and ensures that all transformations are leakage-free – computed using training-set statistics only, within each expanding-window fold.

Construction of features from monthly-to-quarterly data. Instead of aggregating monthly indicators into their quarterly averages or sums, we construct a lag-based mixed-frequency design. The complete history of each monthly indicator up to and including December 2024 is encoded into a grid. At each quarterly end observation, every monthly indicator is expanded into multiple lag features: mLag0 (value at the quarter-end month), mLag1 to mLag5 (values five months back from the quarter-end month), leading to a maximum of six lags per indicator. The number of lags per indicator is limited by the forecasting horizon *H* to avoid the look-ahead bias: for *H=1–2*, lags mLag0 to mLag2 are considered; for *H=4*, lags mLag0 to mLag5 are considered. A monthly missingness counter for each indicator is included as a feature to account for monthly data availability, which becomes critical for the pre-2016 timeframe, in which no monthly indicators were available. Quarterly series are constructed as lags of the target series: TARGET_lag_1Q, TARGET_lag_2Q, TARGET_lag_4Q. Finally, the regime variable post_2016m05 captures the regime where all monthly indicators become available.

Handling of the trend. Since both series have been seasonally adjusted by INSTAT, there is no need for additional seasonal adjustment in the forecasting pipeline. Both statistical time-series models are fit to the SA versions of the target series. Estimation of ETS includes additive error and additive trend, without a seasonal component (seasonal = None). Estimation of both ARIMA and SARIMA includes AIC-based (p,d,q) orders chosen within a non-seasonal orders' grid. The trend in machine-learning algorithms is captured using lagged target series and the post_2016 regime variable. Scaling and imputation leverage fold-specific training statistics to prevent leakage.

Leakage check. The pipeline has been equipped with automatic checks of potential internal leakage, i.e., whether (i) the target y_{t+h} correctly represents an h-step-ahead shift of the current-quarter target, and (ii) every mLag_k feature contains data points only at or before the forecast origin t. The dataset construction was designed to preserve correct target alignment and to ensure that all monthly lag features use only information available at or before the forecast origin. No outlier treatment is applied to the data: while we retain the unusual 2020 quarter related to COVID, these observations are essential for estimating uncertainty. The 90-day lag in GDP publication is accounted for through the rolling origin design, which ensures that each forecasting origin uses only GDP values known at least 90 days ago.

4. Methodology

This section presents the approach used to generate forecasts, along with the model classes involved, the hyperparameterisation strategy applied, and the robustness tests performed. All considerations have been limited to the empirical context of Albania without any attempts at generalisation to other economies.

4.1. Problem formulation and model evaluation

For the target variables Real GDP Growth and GDP Deflator, we consider forecasting horizons $H \in \{1, 2, 4\}$ without aggregation. For each horizon and each target variable, the forecasting task entails predicting y_{t+h}, where predictors have information up to the forecast origin t. As a result, we will evaluate 6 forecasts. Separate models will be estimated for each horizon using the direct strategy from Marcellino, Stock & Watson (2006) in order to avoid possible forecast error accumulation (Marcellino, Stock, Watson, 2006).

The performance of models will be measured using rolling-origin, expanding-window backtests to emulate the real-time setting. Training consists of all observations until time t_j in fold j, and testing occurs for the forecast origin with a horizon of H. A minimum training period of 24 quarters provides about 21 test origins per horizon on an overlapping test date schedule (Hyndman, Athanasopoulos, 2021; Bergmeir, Hyndman, Koo, 2018). Data preprocessing, including scaling, imputation, and rotation by PCA, is always performed exclusively on the training split in each fold.

4.2. Feature Design for Mixed Frequency GDP Forecasting

The forecasting task is a mixed-frequency one, with the target series quarterly and some predictors collected monthly within each quarter. Monthly data are used in the forecasting origin via the mLag_k lags, where each lag represents monthly data up to the point of the forecasting origin. The quarterly target series will be modelled using past values; hence, purely quarterly models may serve as possible baselines for comparison. Consistent with the approach adopted in mixed-frequency forecasting, high-frequency data will be used in the forecasting models via their lagged representations, rather than being aggregated over time periods (Ghysels, Santa-Clara, Valkanov, 2004).

4.3. The groups of models

The considered model comparison involves a wide range of models across two stages. At Stage 1, the complete set of models is included in the analysis. However, the selected models in the second stage constitute the final best-performing family subset, allowing for better interpretability and methodological variety.

Baseline forecasting rules. Persistence is a method in which the forecast is made equal to the previous observed target value. SeasonalNaive_4Q involves forecasting by repeating the target value observed in the same quarter of the previous year, but it has no forecasting value, since the data being forecast is already seasonally adjusted. RandomWalkDrift uses drift; HistoricalMean predicts the mean of the target values; and ThetaMethod is simple exponential smoothing with drift (Assimakopoulos and Nikolopoulos, 2000). AR4 adjusts the AR(p) order with the help of the AIC criterion.

Univariate statistical time series forecasting models. Because both targets are seasonally adjusted, all univariate models are estimated without any seasonal components. ETS uses additive error and trend with maximum likelihood estimation of the smoothing parameters (Hyndman et al., 2002). ARIMA uses the optimal (p, d, q) order based on AIC rather than non-seasonal orders (Box et al., 2015). SARIMA is used as a model-class identifier for comparison with earlier studies on Albanian macroeconomic data, but is estimated without seasonal components as a standard ARIMA on the seasonally adjusted data.

Time Series Dynamic Regression with ARIMA errors (SARIMAX). This model regresses the horizon-h forecast on PCA-reduced exogenous variables and models the residual serial dependence with an ARIMA process. The (p, d, q) order for an ARIMA model is optimised using the AIC criterion, without seasonal components. The exogenous variable features are defined as described in Section 3.3 and are reduced by PCA to at most min(10, $n_{train} - 2$) components. All models in this category are trained as horizon-specific models (Box et al., 2015; Marcellino, Stock, Watson, 2006).

Regularised Linear Models. Ridge(MF), Lasso(MF), and ElasticNet(MF) estimate a mapping from the mixed-frequency feature vector to y_{t+h} using ridge (L2), lasso (L1), and elasticnet (combination of L1 and L2) regularizers, respectively (Hoerl and Kennard, 1970; Tibshirani, 1996; Zou and Hastie, 2005). Quarterly-only versions of these methods − Ridge(Q), Lasso(Q), and ElasticNet(Q) − are estimated using only quarterly-lagged target variables.

Tree-based Ensembles, Gradient Boosting Trees, and Gaussian Process Regression. XGBoost is used to estimate a boosted ensemble of regression trees trained using the mixed-frequency feature vector, with hyperparameters selected via inner-time-series cross-validation (Chen, Guestrin, 2016). Random Forest trains an ensemble of bagged decision trees on the same input (Breiman, 2001). GPR(RBF) and GPR(Matern) use Gaussian process regression with radial basis function and Matérn covariance kernels, respectively, and offer probabilistic uncertainty quantification (Rasmussen, Williams, 2005). XGBoost serves as a representative model from this family in the main results tables, while Random Forest and GPR models are retained as candidate models due to their higher computational costs.

Two-Stage Bridge/Factor-Augmented Models. BridgePCA_Ridge and BridgePCA_ElasticNet transform mixed-frequency monthly indicators using PCA to construct quarterly factors (using either post-2016 data points or the entire training fold if the number of post-2016 observations is less than eight), then regress quarterly GDP on the PCA factors using the ridge (MF) and elasticnet (MF) estimators, respectively (Stock, Watson, 2002; Bai, Ng, 2002).

4.4. Hyperparameter Selection

All model hyperparameters are tuned via inner-time-series cross-validation on the training partitions, without any information from out-of-sample test observations. The inner split uses TimeSeriesSplit with n_{splits} = min(5, max(2, n_{train} / 12)). Ridge(MF) and Ridge(Q) select α from a logarithmically spaced grid between [10^{-8}, 10^{4}] of 120 points. Lasso(MF), Lasso(Q), ElasticNet(MF), and ElasticNet(Q) choose α and ρ for the L1/L2 penalty from the corresponding 120-point grids. For bridge models, the number of PCA factors varies between [1, min(8,$n_{features}$)]. For the ETS, ARIMA, and SARIMAX models, optimal orders are selected using AIC criteria.

4.5. Probabilistic Forecasts Assessment Based on Conformal Prediction

To avoid potential confounding effects arising from differences in the construction of prediction intervals across model families, the conformal prediction approach is used to estimate probability levels across all models (Vovk, Gammerman, Shafer, 2022). For each model and fold, a calibration set is generated from the last eight observations of the training sample within an adaptively bounded region. A quantile value $\hat{q}_{1-\alpha}$ of per-fold conformal residuals allows one to define a prediction interval [$\hat{y}_{t+h} - \hat{q}$, $\hat{y}_{t+h} + \hat{q}$]. The conformal approach is consistently utilised across all model types: statistical, regularised linear, tree-based, and Gaussian Process regressions. The interval probability level is set to 80% (α = 0.20).

4.6. Formal Tests of Point Forecast Accuracy

Diebold-Mariano tests corrected for small sample sizes are applied to detect statistically significant forecast superiority among top-ranking models and baseline benchmarks (Harvey, Leybourne, Newbold, 1997). For each forecasting horizon *H*, the DM statistic is computed between the squared errors produced by each of the top-ranked models and those of Persistence. If applicable, the comparison also includes squared errors of the second-best alternative. Test results are reported as two-sided p-values at the 10% significance level. It should be noted that the small number of outer test folds (about 21) will reduce the power of the DM test. Therefore, DM tests need to be combined with the difference in actual performance values.

4.7. Sub-Sample Robustness Tests

Training and testing sets include data from 2016 to examine whether models' ranks remain unchanged when the full monthly indicator set becomes available. The post-2016 analysis tests whether model rankings change when the full monthly indicator set is available throughout the evaluation period.

For each model and horizon, Moving Block Bootstrap (MBB) resampling is performed over the test-fold error sequence, using B = 1,000 replications and a block length of l = 4 quarterly observations, corresponding to one year. This provides 90% confidence intervals for RMSE and MAE as measures of the sensitivity of the combined measure to particular test folds. Bootstrap RMSE distribution confidence-interval plots for selected models are provided in Figure 4.

4.8. Stage-2 Ensemble

To serve as an additional robustness check, an equally weighted two-model ensemble was created using the top two models for each forecasting horizon identified in Stage 1 analysis. The ensemble point forecast equals the arithmetic mean of the forecasts from the two individual models. Conformal prediction intervals were constructed based on ensemble residuals, as in Section 4.5. The ensemble forecast was calculated test-fold by test-fold, using the same rolling-origin test points as the individual models, with matching train-test splits.

5. Evaluation Metrics

Each model was evaluated using a rolling origin, expanding-window backtest for $H \in \{1, 2, 4\}$ quarters into the future. Thus, for each target variable, we obtained roughly 21 strict out-of-sample predictions. All measures were computed using these test fold results. Both point forecast accuracy and probabilistic properties of a prediction interval are reported, as both types of forecast contribute equally to decision-making in a policy context (Gneiting, Raftery, 2007).

5.1. Point Forecast Accuracy

The accuracy of point forecasts was assessed using two measures: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). MAE represents average absolute deviation, while RMSE penalises larger deviations and, consequently, is a measure of tail-risk performance. RMSE serves as the main criterion for model comparison and MAE as a tie-breaking rule in favour of more robust performance in small samples (Gneiting and Raftery, 2007). The percentage improvement in RMSE over the persistence baseline is defined as (RMSE_Persistence − RMSE_Model) / RMSE_Persistence × 100.

5.2. Prediction Interval Quality

Coverage is defined as the empirical probability of observing test-fold realisation values in the corresponding prediction intervals, which should be as close as possible to 80% at the nominal significance level $\alpha = 0.20$. Under-coverage indicates an overly optimistic assessment of uncertainty; over-coverage via very wide intervals renders a forecasting tool useless in practice. The sharpness of intervals is measured by their mean, median, and 90th percentile widths. Coverage and width are always reported in tandem to avoid misinterpretation: the superiority of a forecast model that guarantees high coverage through exceedingly wide intervals over another model with moderate coverage and narrow intervals cannot be easily argued (Gneiting, Raftery, 2007).

5.3. Continuous Ranked Proba bility Score

The Continuous Ranked Probability Score (CRPS) summarises calibration and sharpness jointly as a proper scoring rule (Gneiting, Raftery, 2007; Hersbach, 2000). Lower CRPS indicates a better-calibrated and sharper predictive distribution. It is computed using a Gaussian approximation based on the conformal sigma estimate, ensuring a consistent uncertainty representation across all model classes.

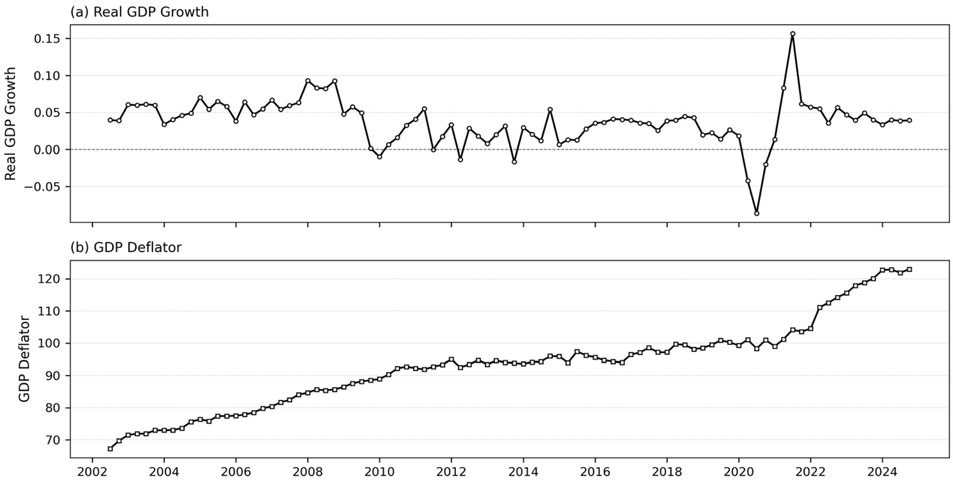
6. Results

In this section, we present the strict out-of-sample results obtained using the rolling-origin expanding-window backtesting procedure outlined in Section 4. For each forecast horizon $H \in \{1, 2, 4\}$, the model is trained on all observations up to the forecast origin t, and the forecast evaluation is done with respect to y_{t+h}, which yields 21 folds of size *n* per target series and per forecast horizon. Throughout the work, Real GDP Growth and the GDP Deflator are considered separate forecasting tasks.

The ranking of models is based on RMSE, whereas MAE was used as the tie-breaking criterion. All prediction intervals are derived using the conformal prediction framework, as discussed in Section 4.5, ensuring a consistent basis for comparing the coverage probability and length of prediction intervals across the different groups of models. The proper score for CRPS accounts for both miscalibration and interval length. Figure 1 illustrates the time evolution of both target series over the entire estimation period from 2002 Q2 to 2024 Q4, showing mean-reverting behaviour of the GDP growth rate and slow dynamics of the GDP deflator during the inflation era from 2021 to 2023.

The rest of the results section is organised as follows. Sections 6.1 and 6.2 provide leaderboards for Stage 1 of Real GDP Growth and the GDP Deflator. We interpret the obtained model rankings in two aspects: (i) insights on the respective time series from the model order, and (ii) the structured model that provides the best forecast. In Section 6.3, we present the results of the Stage-2 family-curated analysis, with full interval diagnostics, enabling us to interpret the model family's performance beyond its raw leaderboard rank. Visualisations of Stage-2 *H=1* predictive intervals for the Real GDP Growth target are given in Section 6.4. Finally, Sections 6.5 to 6.9 discuss the robustness analyses: Diebold-Mariano tests, block bootstrap intervals, sensitivity analysis to training window length, post-2016 subsample analysis, and equally weighted ensembles.

Figure 1. Real GDP Growth and GDP Deflator in Albania, 2002 Q2 – 2024 Q4 (quarterly, seasonally adjusted)



Source: INSTAT Albania; authors' calculations.

6.1. Stage-1 Results: Real GDP Growth

Table 1 reports the top-ranked models and the Persistence baseline across all three horizons for Real GDP Growth (n = 21 strictly out-of-sample test folds).

Table 1. Stage-1 leaderboard – Real GDP Growth (n = 21 folds; 80% coverage). Units: decimal y/y growth; 0.01 corresponds to 1 percentage point. Baseline rows shaded

Model	H	MAE	RMSE	Coverage	CRPS
RandomWalkDrift / HistoricalMean / ThetaMethod	1	0.0273	0.0390	0.619	0.0228
Bridge_PCA_Ridge	1	0.0295	0.0464	0.667	0.0254
ElasticNet_Q	1	0.0308	0.0507	0.714	0.0282
Persistence (baseline)	1	0.0369	0.0574	0.667	0.0326
RandomWalkDrift / HistoricalMean / ThetaMethod	2	0.0272	0.0390	0.619	0.0231
Bridge_PCA_Ridge	2	0.0279	0.0438	0.667	0.0248
Bridge_PCA_ElasticNet	2	0.0299	0.0468	0.667	0.0267
Persistence (baseline)	2	0.0426	0.0674	0.619	0.0380
RandomWalkDrift / HistoricalMean / ThetaMethod	4	0.0264	0.0389	0.667	0.0230
SARIMAX	4	0.0318	0.0415	0.714	0.0242
Ridge (MF)	4	0.0304	0.0451	0.762	0.0252
Persistence (baseline)	4	0.0479	0.0705	0.619	0.0406

The three highest-ranking models – RandomWalkDrift, HistoricalMean, and ThetaMethod – produce identical RMSE and CRPS values across all forecast horizons. This is not a pipeline problem. Under the expanding-window design, the estimated drift becomes negligible, and the trend component of the ThetaMethod tends toward zero. As a result, the three specifications behave like near-mean forecasts for a mean-reverting series.

Among the structured models using external information, the one achieving the best performance at *H=1–2* in this evaluation setting is Bridge_PCA_Ridge, followed by Bridge_PCA_ElasticNet. For *H=4*, the most competitive structured model is SARIMAX; yet part of the latter's superiority over the Persistence benchmark stems from Persistence's inferiority at longer horizons, rather than the strength of SARIMAX's specifications.

6.2. Stage-1 Results: GDP Deflator

Table 2 presents the corresponding leaderboard for the GDP Deflator. Just as before, the univariate models dominate across H, although here the comparison between structured models bears important economic implications, as the regime shift occurs right in the middle of the sample – in 2021-2023, when inflation accelerates.

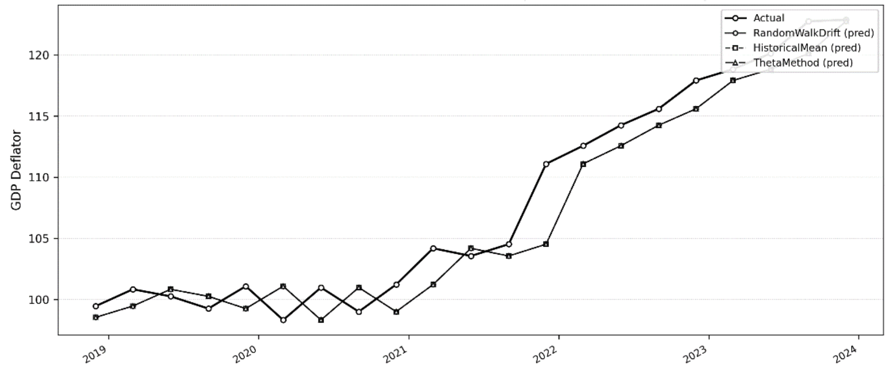
Table 2. Stage-1 leaderboard – GDP Deflator (n = 21 folds; 80% coverage). Units: pp of implicit price index. Baseline rows shaded

Model	H	MAE	RMSE	Coverage	CRPS
RandomWalkDrift / HistoricalMean / ThetaMethod	1	1.820	2.244	0.714	1.375
SARIMAX	1	1.971	2.480	0.714	1.456
Bridge_PCA_ElasticNet	1	1.849	2.552	0.857	1.405
Persistence (baseline)	1	2.583	3.361	0.667	1.947
RandomWalkDrift / HistoricalMean / ThetaMethod	2	1.822	2.245	0.714	1.377
Lasso (MF)	2	2.006	2.341	0.762	1.498
ElasticNet (MF)	2	2.224	2.547	0.667	1.689
Persistence (baseline)	2	3.805	4.519	0.571	2.936
RandomWalkDrift / HistoricalMean / ThetaMethod	4	1.815	2.240	0.714	1.372
SARIMAX	4	2.044	2.659	0.857	1.454
ElasticNet (MF)	4	2.340	3.537	0.714	1.954
Persistence (baseline)	4	5.803	7.082	0.619	4.468

Figure 2 shows actual versus predicted GDP Deflator at *H=1* for the three top-ranked models.

Three nearly identical univariate benchmarks produce almost identical forecast paths. They capture the overall level of the GDP deflator, but respond with a visible lag to the sharp price increase observed during 2021–2023, thereby underestimating the peak of the inflationary episode. Among the structured models, SARIMAX does not outperform the three univariate benchmarks in aggregate RMSE at *H=1*. However, it achieves a lower RMSE than Bridge_PCA_ElasticNet and appears more responsive during the inflationary period. This suggests that, while the univariate rules dominate the full-window ranking, structured models may still provide useful information in periods of regime change.

Figure 2. Actual versus predicted GDP Deflator, H=1 (test folds only)



Compared with the Persistence model, SARIMAX performs 26% better when *H=1* and 62% better when *H=4*, due to its ability to estimate the long-term dynamics between the predictor and the slow-drifting variable.

6.3. Stage-2 Family-Curated Comparison: Point and Interval Diagnostics

Table 3 presents the best models for each family, with full interval diagnostics, allowing one to compare the performance of models across classes beyond the leaderboard positions. Nominal coverage: 80% (α = 0.20).

Table 3. Stage-2 family-curated comparison – Real GDP Growth. 80% coverage. $\bar{W}$ = mean width; W(p90) = 90th pct. Read coverage and width jointly

Model	H	MAE	RMSE	Cov.	CRPS	$\bar{W}$ (mean)	W(p90)
RandomWalkDrift	1	0.0273	0.0390	0.619	0.0228	0.0858	0.133
Bridge_PCA_Ridge	1	0.0295	0.0464	0.667	0.0254	0.0825	0.114
ElasticNet_Q	1	0.0308	0.0507	0.714	0.0282	0.0927	0.142
SARIMAX	1	0.0367	0.0613	0.762	0.0325	0.1233	0.172
ARIMA/SARIMA	1	0.0371	0.0603	0.667	0.0342	0.1255	0.199
XGBoost	1	0.0300	0.0512	0.667	0.0279	0.0905	0.135
Persistence	1	0.0369	0.0574	0.667	0.0326	0.1253	0.200
RandomWalkDrift	2	0.0272	0.0390	0.619	0.0231	0.0912	0.133
Bridge_PCA_Ridge	2	0.0279	0.0438	0.667	0.0248	0.0873	0.124
Ridge_Q	2	0.0294	0.0469	0.714	0.0260	0.0854	0.127
SARIMAX	2	0.0359	0.0558	0.762	0.0311	0.1219	0.167
Persistence	2	0.0426	0.0674	0.619	0.0380	0.1266	0.203
RandomWalkDrift	4	0.0264	0.0389	0.667	0.0230	0.0989	0.133
SARIMAX	4	0.0318	0.0415	0.714	0.0242	0.1023	0.132
Ridge (MF)	4	0.0304	0.0451	0.762	0.0252	0.1021	0.142
Bridge_PCA_ElasticNet	4	0.0280	0.0459	0.714	0.0254	0.0821	0.106
Persistence	4	0.0479	0.0705	0.619	0.0406	0.1332	0.200

Bridge_PCA_Ridge provides the lowest CRPS amongst all structured models (0.0254), together with narrower mean intervals (0.0825) than those of RandomWalkDrift (0.0858). Higher coverage (0.714-0.762) is achieved by SARIMAX with approximately 50% wider intervals.

6.4. Visualisation of Predictive Intervals at H=1 (Real GDP Growth)

Figure 3. 80% conformal intervals for Real GDP Growth, H=1: RandomWalkDrift, Bridge_PCA_Ridge, and SARIMAX.

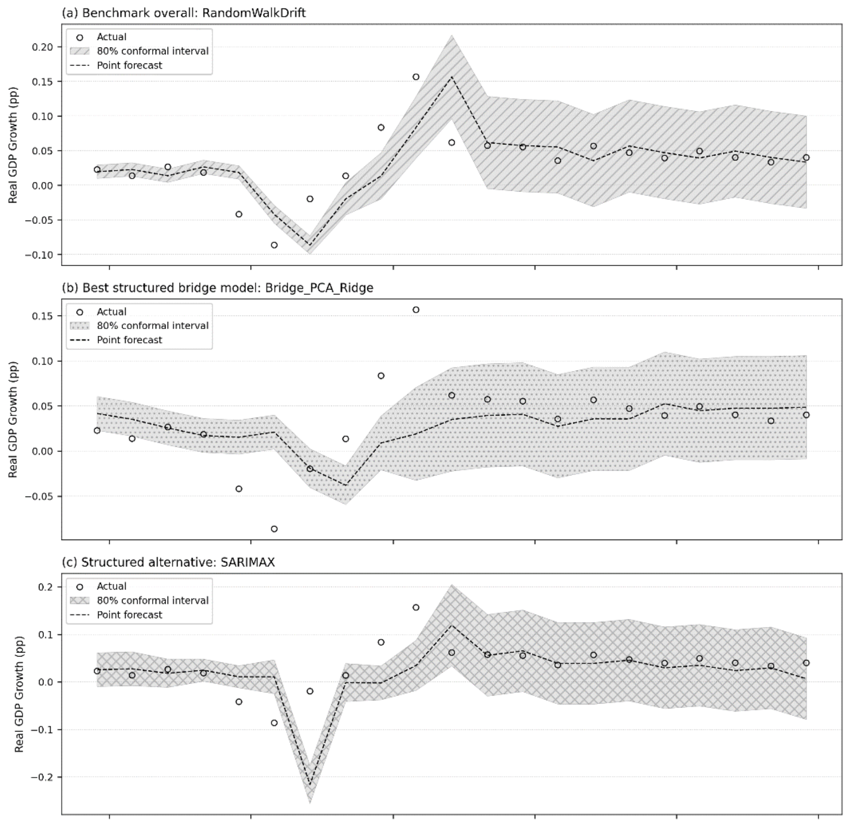


Figure 3 shows conformal prediction intervals at *H=1* for Real GDP Growth for the selected models: RandomWalkDrift (benchmark for the overall leaderboard), Bridge_PCA_Ridge (the leader among structured models at *H=1–2*), and SARIMAX (a wider-interval alternative). While achieving tighter intervals, Bridge_PCA_Ridge maintains relatively high interval coverage. On the contrary, SARIMAX obtains better coverage, but at the cost of larger intervals. The 2019–2020 pandemic is the period during which all models incur their largest errors.

6.5. Comparisons of Models' Forecasting Accuracy via Formal Tests

Table 4 includes Diebold-Mariano test statistics (with the correction from Harvey, Leybourne, and Newbold for small samples, 1997) for comparisons of RandomWalkDrift and Persistence forecasts for Real GDP Growth (n = 21 folds; approximately follows t(20) distribution).

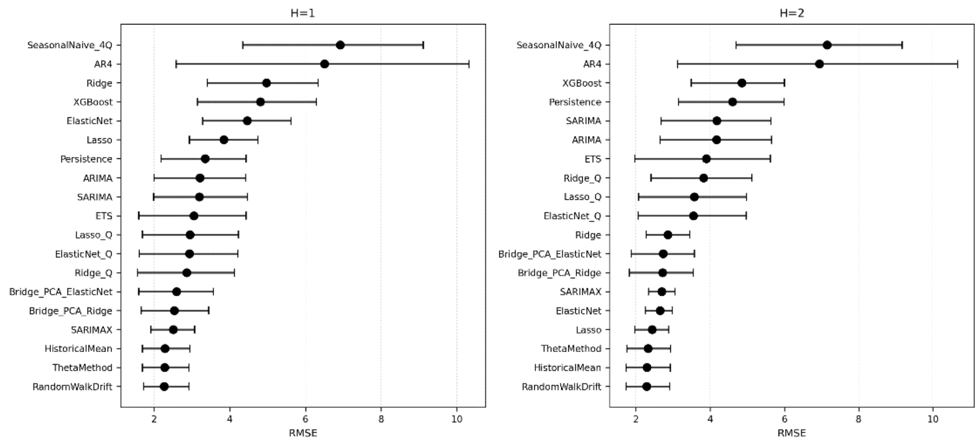
Table 4. Diebold-Mariano test: RandomWalkDrift vs Persistence, Real GDP Growth. HLN small-sample correction; two-sided; n = 21 folds. * p < 0.10; ** p < 0.05

Best overall model vs Persistence	H	DM stat	p-value	Significance
RandomWalkDrift vs Persistence	1	2.029	0.062	* (10% level)
RandomWalkDrift vs Persistence	2	2.324	0.043	** (5% level)
RandomWalkDrift vs Persistence	4	2.250	0.076	* (10% level)

RandomWalkDrift forecast significantly dominates Persistence forecast at the 5% significance level at *H=2* (p = 0.043), as well as at *H=1* and *H=4* at the 10% significance level. With only 21 test folds, statistical power is rather weak, and results must be considered in conjunction with RMSE values.

6.6. Block Bootstrap Confidence Intervals

Figure 4. Block bootstrap 90% CI for RMSE — Real GDP Growth, H=1 and H=2



In Figure 4, we see 90% confidence intervals for the RMSE of key models for Real GDP Growth at *H=1* and *H=2*. Confidence intervals for bootstrap samples confirm that the superiority to Persistence does not depend on which test sample was used. At *H=2*, the confidence intervals of Bridge_PCA_Ridge and Persistence have no intersection, whereas at *H=1*, the intervals overlap more but are directionally consistent with the Diebold-Mariano test outcome.

6.7. Sensitivity to Training Window Length

Table 5. Sensitivity to training window — RMSE range across 50/60/70% fractions, Real GDP Growth, H=1

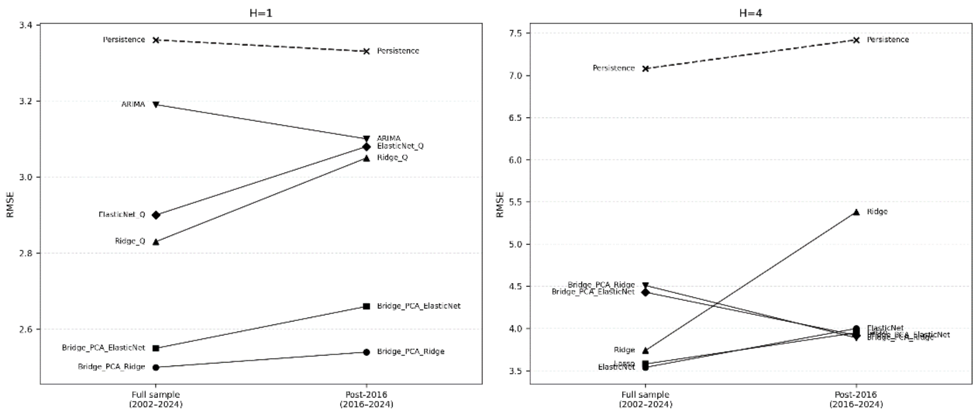
Model	RMSE min	RMSE max	Range
Bridge_PCA_Ridge	0.0353	0.0438	0.009
Bridge_PCA_ElasticNet	0.0347	0.0485	0.014
Ridge_Q	0.0373	0.0470	0.010
ElasticNet_Q	0.0351	0.0470	0.012
ARIMA	0.0392	0.0623	0.023
Ridge (MF)	0.0339	0.0699	0.036
Persistence (baseline)	0.0416	0.0650	0.023

From Table 5, we conclude that Bridge_PCA_Ridge has the smallest RMSE range (0.009) and thus the largest robustness to training window length, compared to other models. Ridge (MF) has the highest sensitivity (0.036). Most structured models remain below or close to Persistence across the tested training-window fractions, while Bridge_PCA_Ridge is the most stable specification and Ridge (MF) is the most sensitive.

6.8. Post-2016 Sub-Sample Analysis

Figure 5 and Table 6 show RMSE for GDP Deflator for the full sample and post-2016 sub-samples at *H=1* and *H=4*.

Figure 5. Full-sample vs post-2016 RMSE — GDP Deflator, H=1 and H=4



For GDP Deflator, Bridge_PCA_Ridge is the most stable model across the regimes, ranking first post-2016 at *H=4* and among the top two at *H=1*. Mixed-frequency regularised estimators perform significantly better in the post-2016 period due to the higher informativeness of the data and are moving up substantially (e.g., from 7th position for *H=1* to 1st position).

Table 6. Full-sample vs post-2016 RMSE and ranks among selected structured and benchmark models – GDP Deflator, H=1 and H=4. Units: pp of implicit price index

Model	H	RMSE (full)	RMSE post-2016	Rank (full)	Rank post-2016
Bridge_PCA_Ridge	1	2.50	2.54	1	2
Bridge_PCA_ElasticNet	1	2.55	2.66	2	4
Lasso (MF)	1	3.87	2.35	7	1
Persistence	1	3.36	3.33	6	9
Bridge_PCA_Ridge	4	4.51	3.89	5	1
Bridge_PCA_ElasticNet	4	4.43	3.92	4	2
ElasticNet (MF)	4	3.54	4.00	1	4
Persistence	4	7.08	7.42	9	9

6.9. Stage-2 Ensemble Results

Table 7 reports the two-model equal-weight ensemble results at each horizon for Real GDP Growth, together with the best single-model RMSE for reference.

Table 7. Equal-weight ensembles – Real GDP Growth. Ensemble vs best single-model RMSE per horizon

Ensemble composition	H	Ensemble RMSE	Best single RMSE
RandomWalkDrift + Bridge_PCA_Ridge	1	0.0390	0.0390
RandomWalkDrift + Bridge_PCA_Ridge	2	0.0390	0.0390
RandomWalkDrift + SARIMAX	4	0.0389	0.0389

Equal-weight averaging did not improve the results for Real GDP Growth. We also checked the same approach for the GDP Deflator, but the results are not reported separately because the gain was small, limited to *H=2*, and did not affect the overall interpretation. For this reason, such a method cannot be regarded as an optimal forecasting approach.

7. Discussion

7.1. The Three Top Models: Only Univariate Rules, No Structural Predictors

One interpretation is that the top three performing models are exclusively univariate, i.e., based solely on historical data of the target variable without accounting for any other predictor variables. This does not imply that structured models are unhelpful. Rather, their value becomes clearer in regime-departure periods, such as the COVID shock, the post-2020 recovery, and the 2021–2023 inflation episode, where external information may help models respond better than purely univariate rules.

7.2. Why Bridge Models Perform Strongest Among Structured Forecasters in This Sample

Bridge_PCA_ElasticNet and Bridge_PCA_Ridge emerge as the most robust models utilising exogenous information in this particular application. The two-step nature of this approach is likely to provide the necessary flexibility for dimensionality reduction and handling mixed-frequency data. Sensitivity testing supports this interpretation: Bridge_PCA_Ridge has the lowest RMSE across different fractions of the training set (0.009), considerably better than ARIMA (0.023) or Ridge (MF) (0.036). However, differences between leading structured models should be viewed as indicative rather than definitive, given the limited number of test folds.

7.3. Probabilistic Quality and the Coverage-Sharpness Trade-off

The opportunity to directly compare the qualities of the intervals using the unified conformal approach highlights another interesting fact. The first observation from Table 3 is that high coverage quality does not automatically imply good quantification of uncertainty. Specifically, when *H=1*, the SARIMAX interval coverage (0.762) is higher than the Bridge_PCA_Ridge one (0.667); however, SARIMAX is nearly double as large (the mean interval width being 0.123 vs 0.082). This trade-off becomes very obvious when we consider CRPS: the latter measure shows that Bridge_PCA_Ridge is much more efficient (0.025) than SARIMAX (0.033). As far as practical implications are concerned, narrower intervals that properly account for uncertainty are a clear plus for constructing scenario forecasts for decision-making purposes. The undercoverage is likely due to the small size of the conformal calibration sample.

7.4. Nonlinear Machine Learning: Consistent Underperformance Relative to Regularised Linear Models

XGBoost, Random Forest, and Gaussian Process versions generally rank behind any regularised linear methods across all horizons and targets considered. A plausible explanation is that the effective fold-specific training sample remains limited relative to the mixed-frequency feature space, while the strict evaluation window contains only about 21 test folds. Under these conditions, higher-capacity nonlinear models are more exposed to overfitting and instability.

7.5. Practical Implications for Albanian GDP Forecasting

The results suggest several practical implications for Albanian GDP forecasting:

- Bridge_PCA_Ridge emerges as the best structured model in the sample overall, in particular, at horizons *H=1* and *H=2*, while SARIMAX performs relatively better at *H=4*.
- Rankings must be seen as indicative, given the 21-fold testing period.
- CRPS and interval widths should be included in the forecast report along with point metrics, since the latter measure misses an essential piece of information in predictive uncertainty.
- The findings from post-2016 subset analysis indicate that models like Lasso (MF) and ElasticNet (MF) should not be evaluated only based on rankings in the full sample.

- Lastly, equally weighted averaging is not recommended as a default ensemble approach.

These findings complement rather than challenge existing BoA work. What this study brings into play is a strict leakage-safe evaluation design, extended model comparison, and the introduction of probabilistic assessment, which demonstrate that the unconditional mean is actually a more robust aggregate benchmark than usually perceived in operational practice.

7.6. Limitations

The first limitation arises from the low number of folds for testing; thus, the statistical power is insufficient to rank the best-structured models. The second limitation lies in the low coverage below nominal, which is explained by the use of a small conformal set rather than by any model specification problems. The third limitation is the restriction to the Albanian context; therefore, validity needs to be demonstrated elsewhere. Finally, better performance at *H=4* is partly due to the benchmark's instability at this horizon.

8. Conclusions

This paper investigates quarterly forecasts of Real GDP Growth and the GDP Deflator in Albania under real-time constraints, using a leakage-free, expanding-origin, rolling-backtest approach and a comparative analysis of 19 models spanning six families. Real GDP Growth and the GDP Deflator are analysed separately as forecasting problems, with their performance assessed based on point forecast accuracy, probabilistic performance, and robustness tests.

The three best models are purely univariate approaches – RandomWalkDrift, HistoricalMean, and ThetaMethod – where, under expanding windows, the mean forecasts converge. The result stems from the mean-reverting behaviour of GDP Growth and persistence in the Deflator, yet with some lagged reaction during the inflationary period of 2021-2023. Among the structured approaches, Bridge_PCA_Ridge shows the most consistent performance in our analysis, achieving the highest joint calibration/sharpness, while SARIMAX performs relatively better on horizon *H=4*. Regularised linear regressors outperform nonlinear machine-learning methods in the current context.

Probabilistic metrics highlight discrepancies that RMSE cannot distinguish. Greater coverage mainly comes from larger prediction intervals, whereas CRPS shows that Bridge_PCA_Ridge provides the most practical density predictions. The below-nominal coverage should be interpreted cautiously, as the conformal calibration sets are small and may not fully capture the interval width needed in a volatile macroeconomic sample. DM tests reveal significant statistical distinctions between RandomWalkDrift and Persistence at *H=2*, whereas at *H=1* and *H=4*, there is less evidence of significance, albeit in the same direction. Block-bootstrap confidence intervals confirm the robustness of the performance discrepancy against Persistence, especially for Bridge_PCA_Ridge at *H=2*. Post-2016 outcomes demonstrate marked improvement for models affected by sparse initial data, such as Lasso (MF) and ElasticNet (MF).

In practical terms, the results support a simple and cautious forecasting strategy. In this dataset, the strongest structured option is not a large nonlinear model, but a bridge model in which monthly information is first reduced through PCA and then used in a regularised regression. Even so, the model should be judged by its behaviour in rolling-origin tests, not by how well it fits the available sample. This is why MAE and RMSE should be reported alongside CRPS: forecast accuracy matters, but so does the quality of the forecast's uncertainty. Additional indicators should therefore be included only when they improve out-of-sample performance, not simply because they make the model appear more informative.

Contribution of individual authors

Dezdemona Gjylapi led the study process, encompassing research design, methodology, complete implementation, testing, data analysis, and manuscript writing.

Alketa Hyso provided technical assistance in verifying and assessing the robustness of the model and manuscript revision for the methods and results sections.

Filloreta Madani offered domain expertise concerning the interpretation of the macroeconomic findings and the manuscript review in the field of economic policy.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Grammarly to support language editing and improvement of academic phrasing. After using this tool, the authors carefully reviewed, verified, and edited all content as needed and take full responsibility for the final content of the publication.

References

Arnostova, K. et al. (2011). Short-Term Forecasting of Czech Quarterly GDP Using Monthly Indicators. – Czech Journal of Economics and Finance (Finance a uver), 61(6), pp. 566-583.

Assimakopoulos, V., Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. – International Journal of Forecasting, 16(4), pp. 521-530. Available at: https://doi.org/10.1016/S0169-2070(00)00066-2.

Bai, J., Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. – Econometrica, 70(1), pp. 191-221. Available at: https://doi.org/10.1111/1468-0262.00273.

Barhoumi, K. et al. (2011). Short-Term Forecasting of Bulgarian GDP Using a Large Dataset. – Discussion Paper 86. Sofia, Bulgaria: Bulgarian National Bank. Available at: https://www.bnb.bg/bnbweb/groups/public/documents/bnb_publication/discussion_2011_86_en.pdf.

Beņkovskis, K. (2008). Short-Term Forecasts of Latvia's Real Gross Domestic Product Growth Using Monthly Indicators. – Working Paper 5. Riga, Latvia: Bank of Latvia. Available at: https://monetas.bank.lv/public_files/images/img_lb/izdevumi/english/citas/wp_2008-5_benkovsky.pdf.

Bergmeir, C., Hyndman, {Rob J.}, Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. – Computational Statistics and Data Analysis, 120, pp. 70-83. Available at: https://doi.org/10.1016/j.csda.2017.11.003.

Box, G.E.P. et al. (2015). Time Series Analysis: Forecasting and Control. 5th ed. Wiley.

Breiman, L. (2001). Random Forests. – Machine Learning, 45(1), pp. 5-32. Available at: https://doi.org/10.1023/A:1010933404324.

Cascaldi-Garcia, D., Luciani, M., Modugno, M. (2023). Lessons from Nowcasting GDP across the World. – International Finance Discussion Papers 1385. Washington, DC: Board of Governors of the Federal Reserve System. Available at: https://doi.org/10.17016/IFDP.2023.1385.

Çeliku, E., Kristo, E. (2022). Vlerësimi i performancës parashikuese të modeleve afatshkurtra të PBB-së. – Tiranë: Banka e Shqipërisë (Material studimor, 4 (84)).

Çeliku, E., Kristo, E., Boka, M. (2009). Modelimi i PBB-së tremujore: roli i treguesve ekonomikë dhe atyre të vrojtimeve. Tiranë: Banka e Shqipërisë.

Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery (KDD '16), pp. 785–794. Available at: https://doi.org/10.1145/2939672.2939785.

Feldkircher, M. et al. (2015). Bridging the information gap: small-scale nowcasting models of GDP growth for selected CESEE countries. – Focus on European Economic Integration, None(2), pp. 56-75.

Franta, M., Havrlant, D., Rusnák, M. (2016). Forecasting Czech GDP Using Mixed-Frequency Data Models. – Journal of Business Cycle Research, 12(2), pp. 165-185. Available at: https://doi.org/10.1007/s41549-016-0008-z.

Ghysels, E., Santa-Clara, P., Valkanov, R. (2004). The MIDAS Touch: Mixed Data Sampling Regression Models. – Journal of Econometrics, 121(1-2), pp. 89-129. Available at: https://doi.org/10.1016/j.jeconom.2003.11.002.

Glišić, I. (2024). A Comparison of Using MIDAS and LSTM Models for GDP Nowcasting. – Working Paper Bulletin 3. Belgrade, Serbia: National Bank of Serbia. Available at: https://www.nbs.rs/export/sites/NBS_site/documents-eng/publikacije/wp_bulletin/wp_bulletin_03_24_1.pdf.

Gneiting, T., Raftery, A.E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. – Journal of the American Statistical Association, 102(477), pp. 359-378. Available at: https://doi.org/10.1198/016214506000001437.

Görtler, J., Kehlbeck, R., Deussen, O. (2019). A Visual Exploration of Gaussian Processes. – Distill [Preprint]. Available at: https://doi.org/10.23915/distill.00017.

Harding, M., Hersh, J. (2018). Big Data in Economics. – IZA World of Labor, 2018(451), pp. 1-10. Available at: https://doi.org/10.15185/izawol.451.

Harvey, D., Leybourne, S., Newbold, P. (1997). Testing the equality of prediction mean squared errors. – International Journal of Forecasting, 13(2), pp. 281-291. Available at: https://doi.org/10.1016/S0169-2070(96)00719-4.

Hauzenberger, N. et al. (2024). Nowcasting with Mixed Frequency Data Using Gaussian Processes. – arXiv preprint, arXiv:2402.10574. Available at: https://doi.org/10.48550/arXiv.2402.10574.

Hersbach, H. (2000). Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. – Weather and Forecasting, 15(5), pp. 559-570. Available at: https://doi.org/10.1175/1520-0434(2000)015%253C0559:DOTCRP%253E2.0.CO;2.

Hoerl, A. E., Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. – Technometrics, 12(1), pp. 55-67. Available at: https://doi.org/10.1080/00401706.1970.10488634.

Hyndman, R. J. et al. (2002). A State Space Framework for Automatic Forecasting Using Exponential Smoothing Methods. – International Journal of Forecasting, 18(3), pp. 439-454. Available at: https://doi.org/10.1016/S0169-2070(01)00110-8.

Hyndman, R. J., Athanasopoulos, G. (2021). Forecasting: Principles and Practice. 3rd ed. Melbourne, Australia: OTexts. Available at: https://otexts.com/fpp3/.

Ljubišić, M. (2025). Guiding through uncertainty: nowcasting the GDP of Croatia. – Public Sector Economics, 49(2), pp. 185-211. Available at: https://doi.org/10.3326/pse.49.2.1.

Maccarrone, G., Morelli, G., Spadaccini, S. (2021). GDP Forecasting: Machine Learning, Linear or Autoregression?. – Frontiers in Artificial Intelligence, 4, p. 757864. Available at: https://doi.org/10.3389/frai.2021.757864.

Makridakis, S., Spiliotis, E., Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. – PLOS ONE, 13(3), pp. 1-26. Available at: https://doi.org/10.1371/journal.pone.0194889.

Marcellino, M., Stock, J. H., Watson, M. W. (2006). A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series. – Journal of Econometrics, 135(1–2), pp. 499-526.

Masten, A. B. et al. (2008). Forecasting Macroeconomic Variables in Slovenia Using Dynamic Factor Models. – IMAD Working Paper Series 9/2008. Ljubljana, Slovenia: Institute of Macroeconomic Analysis and Development. Available at: https://umar.gov.si/fileadmin/user_upload/publikacije/dz/2008/dz09-08.pdf.

Metani, I., Çeliku, E. (2022). Rishikimet e PBB-së tremujore në Shqipëri: vlerësime nga një përqasje vintage. – Tiranë: Banka e Shqipërisë (Material studimor, 3 (84)).

Nerjaku, S., Sinaj, V. (2024). The Use of Quantitative and Intelligent Methods for Forecasting GDP Based on the Unemployment Rate: The Case of Albania.

Pesaran, M. H., Timmermann, A. (2004). Small Sample Properties of Forecasts from Autoregressive Models under Structural Breaks. Available at: https://doi.org/10.17863/CAM.5119.

Pleşa, G., Sîrbu, D., Tănase, A. (2025). Taking into Account Data Revisions when Nowcasting Romanian GDP. – Occasional Papers 39. Bucharest, Romania: National Bank of Romania. Available at: https://www.bnr.ro/uploads/2025-06-12occasionalpapersno.39_documentpdf_545_1749722322.pdf.

Rasmussen, C., Williams, C. (2005). Gaussian Processes for Machine Learning. Available at: https://doi.org/10.7551/mitpress/3206.001.0001.

Rusnák, M. (2013). Nowcasting Czech GDP in Real Time. Working Papers 2013/06. Czech National Bank, Research and Statistics Department. Available at: https://ideas.repec.org/p/cnb/wpaper/2013-06.html.

Shahini, L., Haderi, S. (2013). Short Term Albanian GDP Forecast: One Quarter to One Year Ahead. – European Scientific Journal, 9(34), pp. 198-208.

Stakėnas, J. (2012). Generating Short-Term Forecasts of the Lithuanian GDP Using Factor Models. – Working Paper Series 13. Vilnius, Lithuania: Bank of Lithuania.

Stock, J. H., Watson, M. W. (2002). Forecasting Using Principal Components from a Large Number of Predictors. – Journal of the American Statistical Association, 97(460), pp. 1167-1179. Available at: https://doi.org/10.1198/016214502388618960.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. – Journal of the Royal Statistical Society: Series B (Methodological), 58(1), pp. 267-288.

Tóth, P. (2017). Nowcasting Slovak GDP by a small dynamic factor model. – Ekonomicky casopis, 65, pp. 107-124.

Vovk, V., Gammerman, A., Shafer, G. (2022). Algorithmic Learning in a Random World. 2nd ed. Cham: Springer. Available at: https://doi.org/10.1007/978-3-031-06649-8.

Zou, H., Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. – Journal of the Royal Statistical Society Series B: Statistical Methodology, 67(2), pp. 301-320. Available at: https://doi.org/10.1111/j.1467-9868.2005.00503.x.