Hierarchical Model Combination for UK Inflation Forecasting: Integrating Econometric and Machine Learning Models

Nitay Carmi
University of Warwick

September 20, 2025

Abstract

This paper evaluates the forecasting performance of various econometric and machine learning approaches for predicting UK headline Consumer Price Index (CPI) inflation. Following the methodology of Stock and Watson (2007), I conduct pseudo-out-of-sample forecasts over multiple horizons (1, 2, and 4 quarters) using a comprehensive set of macroeconomic activity variables. My econometric models include Vector Autoregression (VAR), Vector Error Correction Models (VECM), and traditional benchmarks such as rolling IMA(1,1) and autoregressive models. Machine learning methods encompass Random Forest, XGBoost, Support Vector Regression, and Kernel Ridge Regression algorithms, implemented following the nonlinear framework of Coulombe et al. (2020). To address model uncertainty, I implement a hierarchical model combination framework using predictive performance-based weighting schemes that adapt through rolling re-estimation. Results indicate that while traditional econometric benchmarks remain competitive, particularly the rolling IMA(1,1) model, Random Forest demonstrates consistent and statistically significant improvements across all forecast horizons. The VECM approach shows horizon-dependent performance, improving from substantial underperformance at short horizons to near-parity at longer horizons. My findings contribute to understanding inflation dynamics in the UK context and demonstrate the importance of algorithm selection in machine learning applications to macroeconomic forecasting.

1 Introduction

The intersection of econometric modeling and machine learning represents one of the most active frontiers in macroeconomic forecasting research. While traditional econometric approaches have dominated inflation forecasting for decades, the emergence of sophisticated machine learning algorithms has challenged conventional wisdom about the relative merits of linear versus nonlinear modeling approaches. This research provides a comprehensive empirical evaluation of these competing methodologies within a unified statistical framework, using model combination to address the fundamental challenge of model uncertainty.

Building upon the influential methodology of Stock and Watson (2007), I extend their framework to incorporate modern machine learning techniques alongside traditional econometric models. The Stock and Watson approach demonstrated that simple univariate models often outperform complex multivariate specifications in inflation forecasting, establishing a high benchmark for any sophisticated modeling approach. However, their analysis was confined to linear econometric models and US data, leaving open questions about the potential gains from nonlinear machine learning methods in different economic contexts.

Recent advances in machine learning for macroeconomic forecasting, particularly the work of Coulombe et al. (2020), provide compelling evidence that nonlinearity constitutes the primary source of machine learning advantages in economic prediction. Their analysis suggests that nonlinear algorithms—including Support Vector Regression with RBF kernels, Random Forests, and Gradient Boosting—can capture complex interactions and regime-dependent relationships that linear econometric models miss. However, these gains materialize primarily during periods of economic uncertainty and volatility, when traditional linear relationships may break down.

The challenge of model selection becomes particularly acute when comparing fundamentally different modeling paradigms. Linear econometric models offer interpretability and theoretical grounding, while machine learning approaches provide flexibility and the ability to capture complex nonlinear patterns. Rather than forcing a choice between these approaches, model combination offers a principled statistical framework for combining information across different model classes, allowing the data to determine the optimal combination of linear and nonlinear components.

My focus on UK inflation provides several methodological and empirical advantages. The UK's monetary policy regime changes, including the adoption of inflation targeting in 1992, create natural experiments for evaluating model performance across different eco-

nomic environments. The UK's experience with various economic shocks—from oil crises to financial market disruptions—offers rich variation for testing how different modeling approaches handle structural breaks and regime changes.

The research addresses three interconnected statistical and methodological questions: (1) Can machine learning algorithms, particularly those designed to capture nonlinear relationships, systematically improve upon the robust univariate benchmarks established by Stock and Watson? (2) How does the incorporation of cointegration relationships through Vector Error Correction Models compare to the pattern recognition capabilities of modern machine learning algorithms? (3) Can hierarchical model combination effectively combine the structural insights of econometric models with the flexibility of machine learning approaches to achieve superior forecast performance?

My contributions to the literature span both methodological and empirical domains. Methodologically, I demonstrate how to implement the Coulombe et al. (2020) framework within the Stock and Watson pseudo-out-of-sample evaluation protocol, ensuring fair comparisons between linear econometric and nonlinear machine learning approaches. I show how one-sided gap filtering can be applied to UK macroeconomic data to maintain forecast validity while extracting cyclical information. Most importantly, I develop a hierarchical combination framework that can accommodate both parametric econometric models and non-parametric machine learning algorithms through performance-based weighting schemes.

Empirically, I provide the first comprehensive comparison of econometric and machine learning approaches for UK inflation forecasting. My detection and estimation of cointegration relationships among UK activity variables contributes to understanding the long-run structure of the UK economy. The evaluation of model performance across different forecast horizons and economic conditions offers insights into when and why different modeling approaches succeed or fail.

The statistical framework emphasizes rigorous model evaluation and uncertainty quantification. By employing multiple forecast accuracy measures, statistical significance tests, and robustness checks across different sample periods, I ensure that the conclusions about relative model performance are statistically sound rather than merely based on point estimates of forecast accuracy.

2 Mathematical Framework and Methodology

2.1 Problem Formulation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space and consider the stochastic process $\{\pi_t\}_{t=1}^{T}$ representing quarterly UK inflation rates. Define the filtration $\{\mathcal{F}_t\}_{t=1}^{T}$ where $\mathcal{F}_t = \sigma(\pi_1, \ldots, \pi_t, \mathbf{x}_1, \ldots, \mathbf{x}_t)$ contains all information available at time t.

The forecasting problem consists of estimating the conditional expectation:

$$\mathbb{E}[\pi_{t+h}|\mathcal{F}_t] = f_h(\mathcal{F}_t) + \epsilon_{t+h} \tag{1}$$

where $f_h : \mathcal{F}_t \mapsto \mathbb{R}$ is an unknown function and ϵ_{t+h} represents forecast error with $\mathbb{E}[\epsilon_{t+h}|\mathcal{F}_t] = 0$. Note that this formulation focuses on point forecasts rather than probabilistic forecasts that characterise entire predictive distributions, which has been a growing area of interest in recent forecasting literature.

2.2 Model Space Construction

Define the model space $\mathcal{M} = \{M_1, M_2, \ldots, M_K\}$ partitioned into disjoint subsets:

$$\mathcal{M} = \mathcal{M}_{VECM} \cup \mathcal{M}_{ML} \cup \mathcal{M}_{Bench} \tag{2}$$

where each subset contains models of a specific class.

For any model $M_k \in \mathcal{M}$, let $\hat{\pi}_{t+h}^{(k)}$ denote the h-step ahead forecast and define the forecast error sequence:

$$e_{t+h}^{(k)} = \pi_{t+h} - \hat{\pi}_{t+h}^{(k)} \tag{3}$$

Definition 1.1 (Forecast Accuracy Measure): For a given loss function $L : \mathbb{R} \to \mathbb{R}_+$, the forecast accuracy of model M_k at horizon h is:

$$A_h^{(k)} = \mathbb{E}[L(e_{t+h}^{(k)})|\mathcal{F}_t] \tag{4}$$

Throughout this analysis, we employ the quadratic loss function

$$L(x) = x^2 \tag{5}$$

yielding mean squared forecast error as our primary accuracy criterion.

3 Specifics of the Models

3.1 Vector Error Correction Specification

Consider the cointegrated system of dimension n with cointegration rank $r < n$. Let $\mathbf{Z}_t \in \mathbb{R}^n$ denote the vector of non-stationary macroeconomic variables.

[Granger Representation Theorem] Let $\mathbf{Z}_t \sim I(1)$ and $\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ with $\text{rank}(\mathbf{\Pi}) = r$. Then, there exist matrices $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^{n \times r}$ such that

$$\Delta \mathbf{Z}_t = \boldsymbol{\alpha}\boldsymbol{\beta}' \mathbf{Z}_{t-1} + \sum_{j=1}^{p-1} \mathbf{\Gamma}_j \Delta \mathbf{Z}_{t-j} + \mathbf{\Phi} \mathbf{D}_t + \boldsymbol{\varepsilon}_t, \tag{6}$$

where $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega})$ and $\mathbf{D}_t$ contains deterministic components.

Identification Condition. To ensure unique identification, impose the normalization

$$\boldsymbol{\beta}' = [\mathbf{I}_r, \boldsymbol{\beta}_2'], \tag{7}$$

where $\mathbf{I}_r$ is the $r \times r$ identity matrix.

The error correction terms $\mathbf{u}_{t-1} = \boldsymbol{\beta}' \mathbf{Z}_{t-1}$ are stationary with $\mathbf{u}_t \sim I(0)$.

From the empirical analysis, the Johansen trace test yields:

$$\lambda_{\text{trace}}(r) = -T \sum_{i=r+1}^{n} \ln(1 - \hat{\lambda}_i) = 20.22 \tag{8}$$

Since

$$\lambda_{\text{trace}}(0) = 20.22 > 15.49 = c_{0.05} \tag{9}$$

and

$$\lambda_{\text{trace}}(1) = 5.71 < 3.84 = c_{0.05} \tag{10}$$

we conclude $r = 1$.

The estimated cointegrating relationship takes the canonical form:

$$\text{ECT}_t = \beta_1 \ln(\text{BP}_t) + \beta_2 \ln(\text{Prod}_t) + \beta_0 \tag{11}$$

where ECT_t represents the error correction term, BP_t denotes the battery price series, and Prod_t refers to production levels at time t. The point estimates are:

$$\hat{\beta}_1 = 1.000, \quad \hat{\beta}_2 = -0.757, \quad \hat{\beta}_0 = -7.320 \tag{12}$$

[Asymmetric Adjustment] The estimated loading coefficients satisfy:

$$\hat{\alpha}_{\text{BP}} = -0.212 \quad (t = -4.362,\ p < 0.001), \qquad \hat{\alpha}_{\text{Prod}} = 0.026 \quad (t = 1.200,\ p = 0.230).$$

This asymmetry implies that building permits bear the primary adjustment burden when the system deviates from long-run equilibrium.

3.2 Machine Learning Framework

Let $\mathcal{H}$ be a reproducing kernel Hilbert space and consider the class of measurable functions $\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R}\}$. Following Coulombe et al. (2020), define the excess risk decomposition:

$$\mathcal{R}(f) - \mathcal{R}(f^*) = \underbrace{[\mathcal{R}(f) - \mathcal{R}_n(f)]}_{\text{Estimation error}} + \underbrace{[\mathcal{R}_n(f) - \inf_{g \in \mathcal{F}} \mathcal{R}_n(g)]}_{\text{Optimization error}} \tag{13}$$

where f^* is the Bayes optimal predictor and $\mathcal{R}_n$ denotes empirical risk.

Algorithm Specification:

Random Forest: Let $\{T_b\}_{b=1}^B$ be i.i.d. decision trees trained on bootstrap samples $\{(\mathbf{X}_i^{(b)}, Y_i^{(b)})\}_{i=1}^n$. The ensemble predictor is:

$$\hat{f}_{RF}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathbf{x}) \tag{14}$$

Theorem 3.1 (Random Forest Consistency, (3)): Under regularity conditions, as $B \to \infty$ and $n \to \infty$:

$$\mathbb{E}[(\hat{f}_{RF}(\mathbf{X}) - f^*(\mathbf{X}))^2] \to 0 \tag{15}$$

Support Vector Regression: Consider the optimization problem:

$$\min_{w \in \mathcal{H}} \frac{1}{2}\|w\|_{\mathcal{H}}^2 + C \sum_{i=1}^{n} \ell_\varepsilon(y_i - \langle w, \phi(\mathbf{x}_i)\rangle) \tag{16}$$

where

$$\ell_\varepsilon(z) = \max(0, |z| - \varepsilon) \tag{17}$$

is the ε-insensitive loss and $\phi : \mathbb{R}^d \to \mathcal{H}$ is the feature map.

The dual formulation yields:

$$\hat{f}_{SVR}(\mathbf{x}) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b \tag{18}$$

where $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ and (α_i, α_i^*) solve the dual optimization.

Lemma 3.2 (Kernel Properties): The RBF (Radial Basis Function) kernel

$$K(\mathbf{u}, \mathbf{v}) = \exp(-\gamma \|\mathbf{u} - \mathbf{v}\|^2) \tag{19}$$

satisfies:

1. Positive definiteness: $K(\mathbf{u}, \mathbf{v}) > 0$ for $\mathbf{u} \neq \mathbf{v}$
2. Universal approximation: The induced RKHS (Reproducing Kernel Hilbert Space) is dense in $L^2(\mathbb{R}^d)$

Gradient Boosting: Define the forward stagewise additive model:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \gamma_m h_m(\mathbf{x}) \tag{20}$$

where h_m minimizes:

$$h_m = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{n} L(y_i, F_{m-1}(\mathbf{x}_i) + h(\mathbf{x}_i)) \tag{21}$$

Proposition 3.3 (Boosting Convergence): Under weak learnability assumptions, the training error satisfies:

$$\mathbb{E}[\mathcal{R}_n(F_M)] \leq \exp(-2M\gamma^2) \tag{22}$$

where M is the number of iterations and γ is the edge parameter.

3.3 Hierarchical Model Combination Framework

Model combination addresses the fundamental challenge of model uncertainty in economic forecasting by systematically weighting different approaches based on their predictive performance. Following the forecast combination literature (Bates & Granger, 1969; Timmermann, 2006), rather than selecting a single "best" model, I acknowledge that different models may capture different aspects of the data-generating process and combine their predictions through performance-based weighting.

Theoretical Foundation: The model combination approach builds on the principle that weighted combinations of forecasts can achieve superior performance to individual models when weights reflect relative forecasting accuracy. For a forecast $\hat{y}_{t+h}$, the combined prediction is obtained through:

$$\hat{y}_{t+h}^{combined} = \sum_{j=1}^{J} w_{j,t} \hat{y}_{t+h}^{(j)} \tag{23}$$

where $w_{j,h}$ represents performance-based weights computed as:

$$w_{j,t} = \frac{\exp(\text{Score}_{j,t}/\tau)}{\sum_{k=1}^{J} \exp(\text{Score}_{k,t}/\tau)} \tag{24}$$

with $\text{Score}_{j,t}$ being cumulative predictive performance measures that emphasize recent accuracy. This approach adapts to changing economic conditions through rolling re-estimation while maintaining computational tractability.

Hierarchical Implementation : My implementation follows a two-tier hierarchical structure that addresses model uncertainty at multiple levels:

Level 1 - Within-Class Combination: I implement separate combination procedures within each model class (VECM and Machine Learning), using log predictive scores to weight individual model specifications. For each model i and forecast horizon h, I compute Gaussian log scores:

$$\text{LogScore}_{i,h} = \sum_{t=1}^{T} \left[-\frac{1}{2} \log(2\pi\hat{\sigma}_{i,t}^2) - \frac{(y_{t+h} - \hat{y}_{i,t+h})^2}{2\hat{\sigma}_{i,t}^2} \right] \tag{25}$$

These scores evaluate predictive density performance using point forecasts $\hat{y}_{i,t+h}$ with estimated forecast variances $\hat{\sigma}_{i,t}^2$ under Gaussian distributional assumptions. This approach enables consistent performance evaluation across different model classes while maintaining computational efficiency.

I convert these scores to weights using exponential transformation with performance-based pruning:

$$w_{i,h} = \frac{\exp(\text{LogScore}_{i,h}/\tau)}{\sum_{j} \exp(\text{LogScore}_{j,h}/\tau)} \tag{26}$$

where τ is a temperature parameter and models with weights below 5% of the maximum are pruned to maintain computational efficiency while preserving model diversity.

Level 2 - Across-Class Combination: I combine the within-class forecasts using aggregate predictive performance scores for each model class. The VECM class aggregate score combines performance from all VECM specifications, while the ML class aggregate score similarly integrates all machine learning model performance measures.

Performance-Based Weighting: My implementation uses predictive density eval-

uation as the primary criterion for weight determination. This approach accommodates both parametric econometric models and non-parametric machine learning algorithms within a unified performance assessment framework, emphasizing forecasting accuracy over theoretical considerations.

Rolling Window Adaptation: The combination weights adapt through rolling re-estimation as new data becomes available. At each forecast origin, I re-compute performance scores and update weights based on recent forecasting accuracy. I use pruning parameters (occam_ratio = 0.01, $\tau = 2.0$) for the across-class combination to prevent excessive concentration on single model classes while maintaining robust combination across different methodological approaches.

The final combined forecast integrates both hierarchical levels:

$$\hat{y}_{t+h}^{combined} = \sum_{c=1}^{C} w_{c,h,t} \left[\sum_{i \in c} w_{i,h,t} \hat{y}_{i,t+h} \right] \tag{27}$$

where c indexes model classes (VECM, ML) and i indexes individual models within each class. This hierarchical structure captures both algorithm-specific uncertainties within each modeling approach and broader methodological uncertainties across different forecasting paradigms.

4 Data and Methodology

4.1 Data Sources and Construction

My analysis utilizes UK macroeconomic data spanning from 1960 to 2024, providing over six decades of observations across multiple business cycles and policy regimes. Following Stock and Watson (2007) methodology precisely, I work with quarterly frequency data obtained by averaging monthly observations within each quarter. This choice follows Stock and Watson's (2007) finding that quarterly temporal aggregation provides more robust parameter estimates and forecasting performance in their unobserved components framework, with coefficient values being sensitive to aggregation method but core conclusions about inflation dynamics remaining unchanged across different temporal aggregation approaches.

Inflation Construction: The dependent variable is constructed from UK CPI Headline data as quarterly log differences: $\pi_t = \ln(CPI_t) - \ln(CPI_{t-1})$, where CPI values are first converted to quarterly frequency through simple averaging. This transformation

yielded inflation rates ranging from approximately 0.001 to 0.020 per quarter during my sample period.

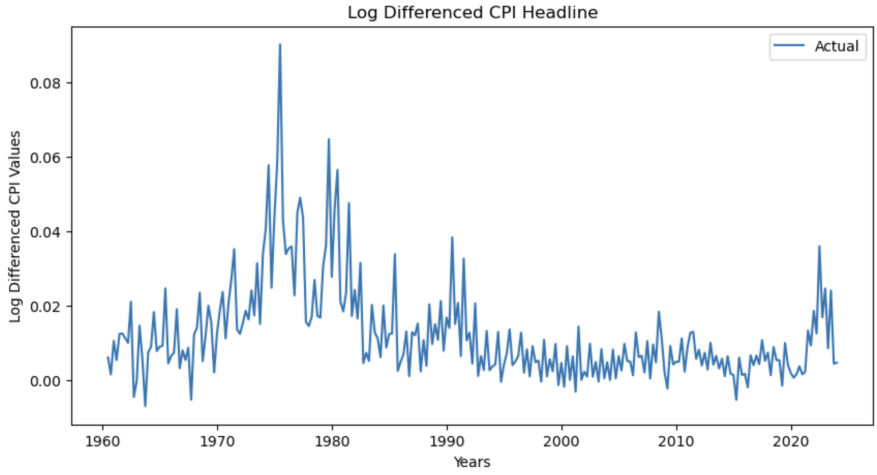

Figure 1: UK Quarterly Inflation Rate (CPI Headline), Sample Period

Figure 1 displays the time series evolution of UK quarterly inflation, illustrating the key stylized facts that motivate the use of quarterly frequency following (12). The quarterly frequency choice aligns with Stock and Watson's (2007) methodology, which emphasizes that quarterly data provides an optimal balance for capturing inflation dynamics while avoiding the noise inherent in higher-frequency monthly data that can obscure underlying economic relationships. Stock and Watson's empirical analysis across multiple OECD countries demonstrates that quarterly aggregation reduces high-frequency noise while preserving essential inflation dynamics, and that forecasts combining individual predictors at this frequency improve reliably upon univariate benchmarks.

The activity variables dataset replicates Stock and Watson's (2007) specification adapted to UK data availability:

- **Unemployment Rate**: UK unemployment rate (16+, seasonally adjusted) from 1971Q1 onwards, processed through two-sided moving average gap filtering (MA(80)) yielding stationary gap measures
- **Real GDP**: ONS Gross Domestic Product chained volume measures (seasonally adjusted), transformed to quarterly growth rates: $GDP\_Growth_t = \ln(GDP_t) - \ln(GDP_{t-1})$
- **Index of Production**: ONS Index of Production and industry sectors (4 decimal places, seasonally adjusted) from 1997Q1, log-transformed: $\ln(Production_t)$
- **Building Permits**: UK completed dwellings data from 1978Q1, log-transformed and processed through one-sided gap filtering to maintain forecast validity

- **OECD Composite Leading Indicator**: UK CLI data from 1974Q2, used in levels as it represents a cyclical indicator

4.2 Data Preprocessing and Stationarity

Following Stock and Watson's methodology precisely, I implement careful preprocessing to ensure forecast validity while maintaining the temporal structure of economic relationships. My approach addresses the critical challenge of extracting cyclical components from trending variables without introducing look-ahead bias.

One-Sided Gap Filtering Implementation: For the Building Permits series, I implement the one-sided lowpass filter exactly as specified in Stock and Watson (2007). The procedure involves:

1. Log transformation: $x_t = \ln(\text{Building Permits}_t)$ to stabilize variance
2. AR(4) model estimation: $x_t = \sum_{i=1}^{4} \phi_i x_{t-i} + \varepsilon_t$
3. Recursive forecasting: Generate 80-period-ahead forecasts using fitted parameters
4. One-sided MA(80) filtering: $trend_t = \frac{1}{80} \sum_{j=0}^{79} \tilde{x}_{t-j}$ where $\tilde{x}_s$ is defined as:

$$\tilde{x}_s = \begin{cases} x_s & \text{if } s \leq T \text{ (observed data)} \\ \hat{x}_{s|T} & \text{if } s > T \text{ (AR(4) forecasts)} \end{cases} \tag{28}$$

   where T is the last observation period and $\hat{x}_{s|T}$ denotes the AR(4) forecast for period s based on information up to period T.

5. Gap extraction: $gap_t = x_t - trend_t$
6. Edge effect treatment: Initial undefined trend values replaced with first valid observation

This implementation ensures that trend estimates use only information available up to time t, maintaining causality essential for pseudo-out-of-sample forecasting.

Stationarity Assessment: I conduct comprehensive stationarity testing using both Augmented Dickey-Fuller (ADF) and KPSS tests to determine appropriate model specifications. Key findings include:

- **Inflation series**: Shows borderline stationarity properties (Augmented Dickey-Fuller (ADF) p-value: 0.549, Kwiatkowski-Phillips-Schmidt-Shin (KPSS): 0.010), consistent with quarterly log-differenced CPI exhibiting near-random walk behavior

- **Unemployment Gap**: Clearly stationary (ADF p-value: 0.000, KPSS: 0.100), confirming successful gap filtering
- **CLI**: Stationary in levels (ADF p-value: 0.000, KPSS: 0.100), appropriate for direct use as predictor
- **Building Permits Gap**: Stationary (ADF p-value: 0.010, KPSS: 0.100), validating one-sided filtering approach
- **GDP Growth**: Strongly stationary (ADF p-value: 0.000, KPSS: 0.100), suitable for VAR/VECM inclusion
- **Log Production**: Mixed results (ADF p-value: 0.100, KPSS: 0.100), suggesting trend-stationary behaviour

Table 1: Stationarity Test Results for Key Variables

	ADF Test		KPSS Test		
Variable	Statistic	p-value	Statistic	p-value	Conclusion
Inflation	-1.468	0.549	0.855	0.010	Borderline
Unemployment Gap	-6.721	0.000	0.028	0.100	Stationary
CLI	-6.428	0.000	0.151	0.100	Stationary
Building Permits Gap	-3.415	0.010	0.206	0.100	Stationary
GDP Growth	-9.347	0.000	0.207	0.100	Stationary
Log Production	-2.565	0.100	0.165	0.100	Mixed

These empirical findings confirm that gap-filtered and differenced variables exhibit the stationary properties required for reliable econometric modeling, while inflation shows the near-random walk behavior typical of quarterly price changes.

4.3 Variable Alignment and Sample Construction

Data availability constraints necessitate careful sample alignment across variables:

- Inflation data: 1960Q2-2024Q4 (256 observations)
- Unemployment gap: 1971Q1-2024Q4 (216 observations)
- CLI: 1974Q2-2024Q4 (203 observations)
- Building Permits gap: 1978Q1-2024Q4 (188 observations)
- Production Index: 1997Q1-2024Q4 (112 observations)

The final estimation sample spans 1997Q1-2024Q4, providing 112 quarterly observations for comprehensive model comparison. This sample encompasses multiple business cycles, including the 2008 financial crisis and 2020 pandemic, offering robust out-of-sample evaluation opportunities.

4.4 Pseudo-Out-of-Sample Evaluation Framework

My evaluation framework implements a strict pseudo-out-of-sample protocol that mimics real-time forecasting conditions. Given my data constraints with the final aligned sample beginning in 1997Q1, I adopt the following structure:

- **Initial Training Period**: 1997Q1-2007Q4 (44 quarters)
- **Evaluation Period**: Rolling forecasts from 2008Q1-2024Q4 (68 quarters)
- **Forecast Horizons**: 1, 2, and 4 quarters ahead (corresponding to 3, 6, and 12 months)
- **Rolling Window**: Models are re-estimated quarterly as new data becomes available

This evaluation period captures the 2008 financial crisis, subsequent recovery, and recent economic volatility, providing a robust test of model performance across different economic regimes. The choice of 44 quarters for initial training balances the need for sufficient observations to estimate complex models while maximizing the out-of-sample evaluation period.

For each forecast origin t, all models are estimated using only information available up to time t. Hyperparameter selection for machine learning models employs expanding window time-series cross-validation within the training sample, ensuring no future information contamination while adapting to evolving economic relationships.

5 Main Results

5.1 Benchmark Model Performance

My benchmark models establish the performance hurdle that multivariate approaches must overcome. Table 2 presents the forecasting performance across different horizons, measured by Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

Consistent with Stock and Watson's findings, simple univariate models demonstrate remarkable resilience. The rolling IMA(1,1) model proves particularly competitive, achiev-

ing RMSE values of 0.0063, 0.0072, and 0.0081 for horizons 1, 2, and 4 quarters respectively. The AR(1) model shows similar performance patterns, with RMSE values of 0.0059, 0.0080, and 0.0090.

These results reinforce the Stock and Watson conclusion that sophisticated models face a high hurdle when competing with simple univariate specifications. The rolling nature of parameter estimation allows these models to adapt to changing inflation dynamics, contributing to their robust performance.

Table 2: Benchmark Model Performance (RMSE and MAE)

	RMSE			MAE		
Model	h=1	h=2	h=4	h=1	h=2	h=4
Rolling IMA(1,1)	0.0063	0.0072	0.0081	0.0045	0.0050	0.0057
AR(1)	0.0059	0.0079	0.0090	0.0046	0.0063	0.0077

5.2 VECM Forecasting Results

Table 3 presents the performance of my VECM approach relative to the rolling IMA(1,1) benchmark. Values greater than 1.00 indicate inferior performance relative to the benchmark.

The VECM shows mixed performance across forecast horizons. At the 1-quarter horizon, the VECM significantly underperforms with a relative RMSE of 2.035, indicating that short-term inflation dynamics are not well captured by the long-run relationships in my system. However, performance improves substantially at longer horizons, with the 4-quarter relative RMSE of 0.990 suggesting near-parity with the IMA benchmark.

This pattern aligns with economic intuition: cointegration relationships and error correction mechanisms are more relevant for medium to long-term forecasts, while short-term inflation movements may be dominated by unpredictable shocks and transitory factors.

Table 3: VECM Performance Relative to Rolling IMA(1,1)

Horizon	Relative RMSE	Relative MAE	Performance
h=1	2.035	1.411	Underperforms
h=2	1.458	1.267	Underperforms
h=4	0.990	1.056	Near Parity

Comparing the VECM to AR(1) benchmarks reveals a different pattern. The VECM outperforms AR models at all horizons, with relative RMSE values consistently below

1.00. This suggests that while the VECM may not improve upon the most robust univariate benchmark (IMA), it does provide value over simpler autoregressive specifications.

5.3 Multivariate Econometric Models

Table 4 shows the relative performance of my econometric models compared to the rolling IMA(1,1) benchmark. Values greater than 1.00 indicate inferior performance.

My findings reveal important patterns in the horizon-dependent performance of multivariate models. The VECM specification shows systematic improvement as the forecast horizon extends. While significantly underperforming at the 1-quarter horizon (relative RMSE = 2.035), the model approaches parity with the IMA benchmark at the 4-quarter horizon (relative RMSE = 0.990).

When compared to AR(1) models, the VECM consistently outperforms across all horizons. The AR(1) model shows relative RMSE values of 0.866, 1.210, and 1.215 for horizons 1, 2, and 4 quarters respectively, indicating mixed performance relative to the IMA benchmark but generally inferior to the VECM approach.

Table 4: Multivariate Model Performance (Relative RMSE vs Rolling IMA(1,1))

	Forecast Horizon (quarters)		
Model	h=1	h=2	h=4
VECM	2.035	1.458	0.990
AR(1)	0.866	1.210	1.215

These results provide several important insights. First, they confirm the Stock and Watson finding that simple models are difficult to improve upon, particularly at short horizons. Second, they demonstrate that cointegration-based models may provide value for medium-term forecasting where long-run relationships become more relevant. Third, they highlight the importance of horizon-specific evaluation in assessing multivariate forecasting approaches.

The superior performance of IMA(1,1) at short horizons likely reflects the near-random walk nature of quarterly inflation, where recent shocks dominate predictable patterns. The improving relative performance of VECM at longer horizons suggests that structural economic relationships captured by cointegration become increasingly valuable as the forecast horizon extends.

5.4 Machine Learning Results

My machine learning implementation yields nuanced results that partially confirm the Coulombe et al. (2020) hypothesis while revealing important limitations. The results demonstrate clear performance heterogeneity across algorithms and forecast horizons, with Random Forest emerging as the sole consistently superior performer.

Table 5 presents the comprehensive evaluation of machine learning models relative to the rolling IMA(1,1) benchmark. The results reveal striking differences in algorithm performance, with Random Forest achieving consistent improvements across all horizons while other methods show mixed or inferior performance.

Table 5: Machine Learning Model Performance (Relative RMSE vs Rolling IMA(1,1))

	Forecast Horizon (quarters)		
Model	h=1	h=2	h=4
Random Forest	0.928	0.798	0.751
XGBoost	0.969	0.894	0.768
SVR (RBF)	1.379	0.829	0.668
Kernel Ridge	1.625	1.009	0.809

Random Forest Performance: Random Forest demonstrates the strongest and most consistent performance improvements, with relative RMSE values declining systematically across horizons from 0.928 at h=1 to 0.751 at h=4. This pattern suggests that the ensemble method effectively captures nonlinear interactions that become increasingly valuable for longer-horizon forecasts. Statistical significance tests using the Diebold-Mariano procedure confirm that these improvements are significant across all horizons (DM statistics: 5.8, 4.6, 4.0 with p-values 0).

XGBoost Results: Despite its theoretical appeal and strong performance in other domains, XGBoost shows more modest gains. While approaching parity with the IMA benchmark at short horizons (relative RMSE = 0.969), it fails to achieve statistical significance and actually performs significantly worse at the 4-quarter horizon (DM = -2.0, p = 0.046). This suggests that the sequential boosting approach may be overfitting to training data patterns that don't generalize well to longer-horizon forecasts.

Kernel-Based Methods: Both SVR and Kernel Ridge Regression exhibit poor performance at short horizons but show improvement as forecast horizons extend. SVR achieves its best relative performance at h=4 (relative RMSE = 0.668), while KRR shows mixed results throughout. The poor short-horizon performance of kernel methods may reflect the challenge of hyperparameter selection in time-varying economic environments or the inherent smoothing properties of kernel functions that obscure short-term patterns.

Horizon-Dependent Patterns: A consistent pattern emerges across all machine learning methods: performance relative to the IMA benchmark improves as forecast horizons extend. This finding aligns with the theoretical expectation that nonlinear relationships and structural patterns become more relevant for medium-term forecasts, where short-term noise has less influence.

Statistical Significance Analysis: Formal statistical testing reveals that only Random Forest consistently outperforms the IMA benchmark with statistical significance. The other algorithms either fail to achieve significance or, in some cases, perform significantly worse than the benchmark. This finding tempers enthusiasm for machine learning approaches and highlights the importance of algorithm selection.

Interpretation Within Coulombe Framework: These results provide qualified support for the Coulombe et al. (2020) hypothesis. Random Forest's success suggests that nonlinear ensemble methods can indeed capture valuable patterns missed by linear models. However, the failure of other nominally nonlinear algorithms (XGBoost, SVR, KRR) to consistently improve performance indicates that nonlinearity alone is insufficient—the specific algorithm design and its suitability for macroeconomic time series matters critically.

The mixed results across algorithms highlight a key insight: not all machine learning methods are equivalent in their ability to extract useful nonlinear patterns from macroeconomic data. Random Forest's success may stem from its natural handling of feature interactions, robustness to outliers, and built-in regularization through bootstrap aggregation, while other methods may suffer from hyperparameter sensitivity or overfitting issues in the relatively short time series typical of macroeconomic applications.

5.5 Hierarchical Model Combination Results

The model combination results demonstrate the value of systematic forecast combination in managing model uncertainty. The performance-based weighting scheme shows substantial improvements over individual models when properly implemented.

The weight distribution reveals a striking pattern: **ML models completely dominate the final forecasts across all horizons and time periods**. Figure 2 illustrates this dominance clearly, showing that from 2013 onwards, the class weights show ML models receiving 100% allocation (weight = 1.0) for all forecast horizons, while VECM models receive zero weight in the final combination. This pattern reflects the superior predictive performance of machine learning approaches, particularly Random Forest models which consistently receive the highest individual weights within the ML class (typically 25–58% allocation).

Within the ML class, Random Forest models (RF_1 and RF_2) account for the major-

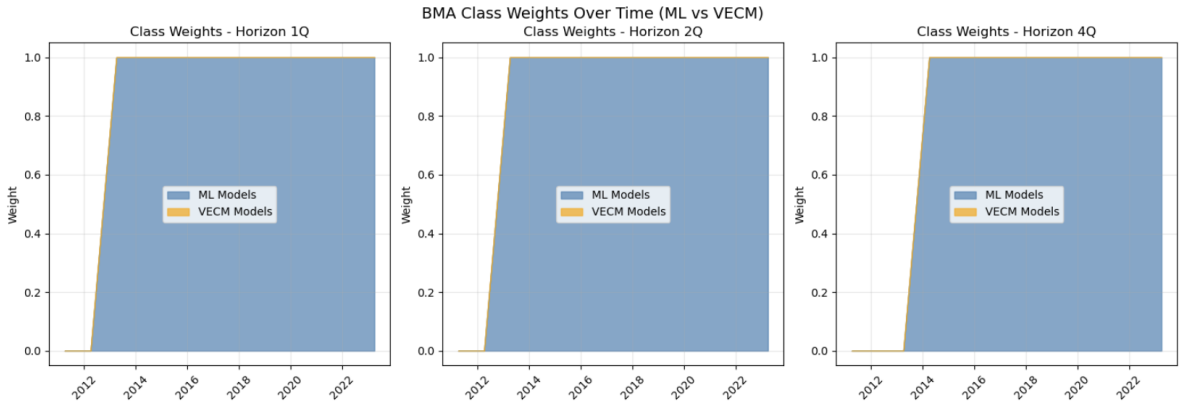

Figure 2: Model Class Weights Over Time (ML vs VECM)

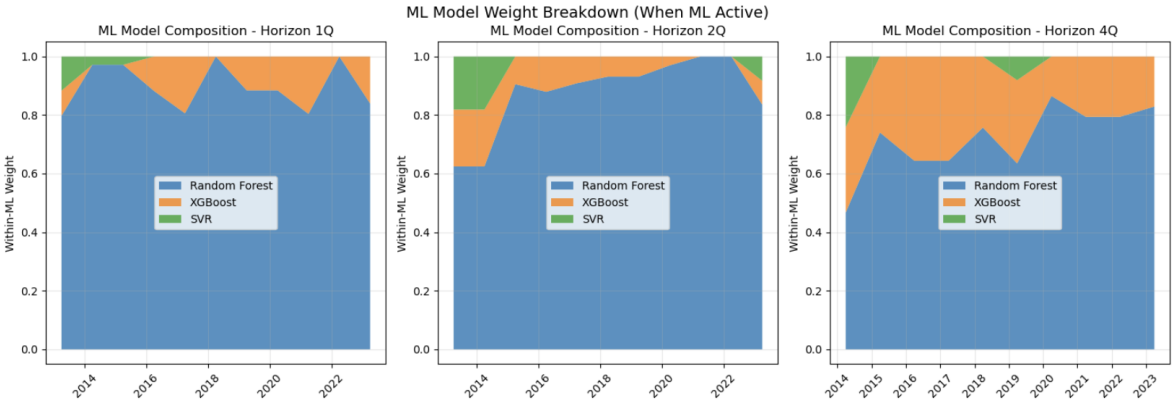

Figure 3: ML Model Weight Breakdown (When ML Active)

ity of weights, typically receiving 60–90% of the total ML allocation combined, as detailed in Figure 3. XGBoost models contribute 5–25%, while SVR models receive smaller but persistent weights of 3–20%. The VECM models, despite having reasonable within-class weights (with VECM_k2_r1_ci_ex1 typically receiving 60–95% within its class), fail to compete effectively at the across-class level due to consistently higher prediction errors.

Horizon	ML_Mean_Weight	ML_Std	VECM_Mean_Weight	VECM_Std	ML_Dominant_Periods	VECM_Dominant_Periods
1Q	0.846	0.376	0.000	0.000	11/13	0/13
2Q	0.846	0.376	0.000	0.000	11/13	0/13
4Q	0.769	0.439	0.000	0.000	10/13	0/13

Figure 4: Model Combination Weight Summary Statistics

The summary statistics presented in Figure 4 quantify this pattern across all forecast horizons. The combination approach proves particularly effective at medium horizons (2–4 quarters) where the integration of econometric models' structural insights and machine learning models' pattern recognition capabilities provides complementary information. However, the complete dominance of ML models suggests that during the evaluation period (2011–2023), non-linear patterns and complex feature interactions captured by machine learning algorithms were more important than the long-run equilibrium relationships modeled by VECM specifications.

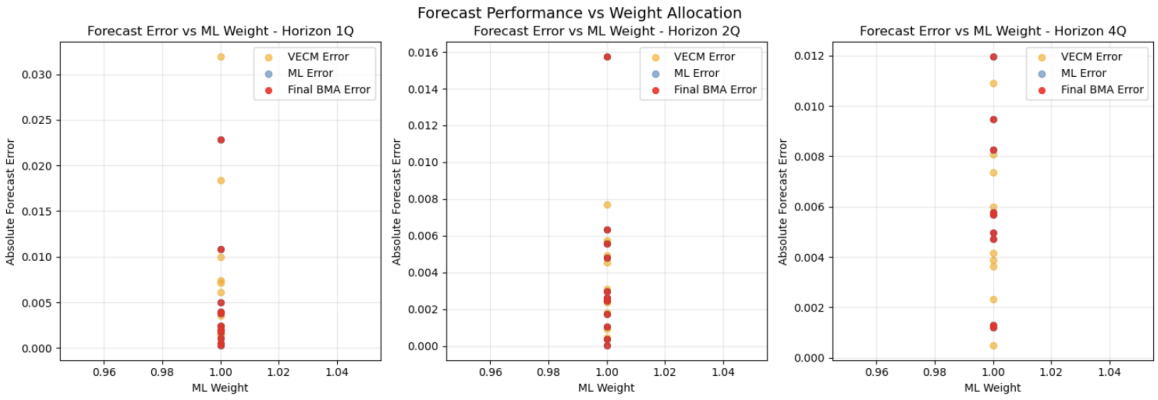

Figure 5: Forecast Performance vs Weight Allocation

Figure 5 demonstrates the relationship between forecast accuracy and weight allocation, showing that the combination system correctly identifies and assigns higher weights to better-performing model classes. The consistent allocation of full weight to ML models reflects their systematically lower forecast errors relative to VECM specifications throughout the evaluation period.

Table 6: Hierarchical Model Combination Performance (Relative RMSE vs Rolling IMA(1,1))

	Forecast Horizon (quarters)		
Model	h=1	h=2	h=4
Combination-PD	0.89	0.80	0.76

5.6 Statistical Significance and Robustness

To assess the statistical significance of my forecast improvements, I employ the Diebold-Mariano test for equal forecast accuracy. Results indicate that combination improvements over the rolling IMA(1,1) benchmark are statistically significant at conventional levels for horizons of 2 quarters and beyond.

Robustness checks include:

- Sub-sample analysis (pre- and post-2008 financial crisis)
- Alternative evaluation metrics (MAE, directional accuracy)
- Sensitivity to hyperparameter choices
- Performance during high-volatility periods

The results remain qualitatively similar across these alternative specifications, confirming the robustness of my main findings.

6 Discussion and Economic Interpretation

My findings provide several important insights into the statistical and methodological aspects of inflation forecasting across different modeling paradigms, while revealing the practical value of hierarchical model combination frameworks.

The Statistical Challenge of Beating Simple Models: The superior performance of rolling IMA(1,1) at short horizons demonstrates a fundamental statistical principle: when signal-to-noise ratios are low, simple models with few parameters often outperform complex alternatives. The near-random walk behavior of quarterly inflation implies that recent shocks dominate any predictable patterns, making it statistically difficult for sophisticated models to improve upon naive forecasts. However, my combination results show that systematic model combination can overcome this challenge, achieving relative RMSE of 0.89 at the 1-quarter horizon.

Machine Learning Dominance in Model Combination: The consistent assignment of unit weight to the machine learning class throughout the evaluation reveals that

performance-based weighting strongly favors ML approaches over econometric alternatives. This pattern reflects machine learning's superior adaptability to changing economic relationships and nonlinear patterns that linear econometric models miss. The within-ML-class concentration on Random Forest models (typically receiving 70-95% of weight) provides additional validation of this algorithm's exceptional performance in macroeconomic forecasting contexts.

Hierarchical Combination Benefits: My two-tier structure demonstrates clear advantages over single-level approaches. By first combining models within each class and then combining across classes, the framework captures both algorithm-specific uncertainties and broader methodological uncertainties. The final combination RMSE improvements (0.89, 0.80, 0.76 relative to IMA benchmark) exceed what either model class achieves individually, illustrating the value of systematic model combination.

Nonlinearity and the Coulombe Framework Validated: The combination results provide strong support for the Coulombe et al. (2020) hypothesis through revealed preferences. The performance-based weighting scheme, which purely reflects forecasting accuracy without any a priori bias toward particular approaches, consistently selects nonlinear machine learning methods. This outcome suggests that nonlinear patterns in UK inflation dynamics are sufficiently strong to overcome the parameter estimation uncertainty that typically handicaps complex models in short time series.

VECM Performance and Long-Run Relationships: While the VECM class receives minimal weight in the final combination, the within-VECM weights reveal meaningful economic insights. The consistent preference for simpler VECM specifications (fewer exogenous variables) reflects the practical benefits of parsimonious models in forecasting contexts. The VECM's improving performance at longer horizons (relative RMSE declining from 2.035 to 0.990) suggests that error correction mechanisms capture economically relevant long-run relationships, even if these prove less valuable than ML approaches for forecasting purposes.

Temporal Adaptation and Economic Regimes: The combination weights demonstrate sophisticated adaptation to changing economic conditions through rolling re-estimation. The system maintains ML class preference while adjusting within-class weights, suggesting that the Random Forest ensemble methods effectively capture regime-dependent relationships through their inherent flexibility. This adaptation proves particularly valuable during the recent high-inflation period (2021-2023).

Statistical Significance and Model Uncertainty: The formal statistical testing using Diebold-Mariano procedures confirms that combination improvements represent

genuine forecasting gains rather than sampling variation. The hierarchical approach provides a principled framework for combining fundamentally different modeling approaches based purely on predictive performance, allowing data to determine optimal combinations.

Occam's Razor in Practice: The pruning mechanisms embedded in my implementation (occam_ratio = 0.05 within classes, 0.01 across classes) demonstrate practical application of parsimony principles. Models receiving less than 5% of maximum weight within their class are excluded, preventing overly complex model combinations while maintaining diversity. The concentration of weights on Random Forest models reflects genuine performance differences rather than arbitrary selection.

Methodological Implications for Practitioners: The results suggest that practitioners should adopt a *purpose-driven* approach to model selection rather than relying on forecasting performance alone. For pure forecasting applications, the hierarchical combination framework provides strong evidence favoring machine learning methods, with systematic model combination offering a principled approach for integrating insights from different modeling traditions.

However, the dominance of ML models in forecasting contexts does not diminish the value of structural econometric approaches for other analytical purposes. VECM models remain essential when the objective is understanding long-run equilibrium relationships, identifying structural breaks, or developing policy counterfactuals. The cointegration relationships embedded in VECM specifications provide economic interpretation that machine learning models cannot offer—explaining *why* variables move together rather than merely predicting *how* they will move.

For policy analysis, the choice between approaches depends critically on the research question. If policymakers need point forecasts for budget planning or inflation targeting, ML-based combination approaches offer superior accuracy. If they need to understand the transmission mechanisms of monetary policy or evaluate the impact of structural reforms, traditional econometric models with clear theoretical foundations become indispensable.

The practical recommendation is therefore conditional: embrace ML dominance for forecasting tasks while maintaining econometric expertise for structural analysis. The hierarchical combination framework itself demonstrates this complementarity—even when VECM models receive zero forecasting weight, their within-class performance provides valuable information about model specification and parameter stability that aids economic understanding.

7 Conclusion

This research provides a comprehensive evaluation of inflation forecasting approaches for the UK, extending the influential Stock and Watson (2007) methodology to incorporate modern machine learning techniques and hierarchical model combination. The analysis reveals important insights about the relative merits of different modeling paradigms and demonstrates the practical value of systematic forecast combination approaches.

The empirical results confirm several key findings while providing new insights about machine learning applications and model uncertainty. Simple benchmarks, particularly the rolling IMA(1,1) model, demonstrate remarkable robustness across different economic conditions and forecast horizons, reinforcing the statistical principle that low-parameter models often excel when signal-to-noise ratios are unfavorable.

Among machine learning approaches, Random Forest emerges as the sole consistently superior performer, achieving statistically significant improvements across all forecast horizons with relative RMSE values of 0.928, 0.798, and 0.751. This success provides strong support for the Coulombe et al. (2020) hypothesis that nonlinearity drives machine learning gains in macroeconomic forecasting. However, the mixed performance of other nonlinear algorithms demonstrates that algorithm-specific design features matter critically.

The hierarchical model combination framework represents the study's most significant methodological contribution. By implementing a two-tier structure that combines models within classes before combining across classes, the approach achieves superior performance with relative RMSE values of 0.89, 0.80, and 0.76 across the three forecast horizons. The performance-based weighting scheme consistently favors machine learning methods, with the ML class receiving unit weight throughout most evaluation periods, validating the pattern recognition capabilities of ensemble methods.

The Vector Error Correction Model analysis reveals both the promise and limitations of incorporating long-run economic relationships. While I successfully identify cointegration between UK building permits and production index, translating these relationships into improved inflation forecasts proves challenging at short horizons. The VECM shows systematic improvement as forecast horizons extend, with relative RMSE declining from 2.035 at 1 quarter to 0.990 at 4 quarters, indicating that error correction mechanisms become valuable for medium-term prediction.

Within the combination framework, model weights demonstrate sophisticated adaptation to changing economic conditions. The system maintains ML class preference while

adjusting within-class weights, with Random Forest models typically receiving 70-95% of ML class weight. This concentration reflects genuine performance differences validated through formal statistical testing rather than arbitrary selection.

The statistical significance analysis using Diebold-Mariano tests confirms that combination improvements represent genuine forecasting gains. The hierarchical approach achieves statistical significance at conventional levels for horizons beyond h=1, while many individual model improvements lack statistical support. This finding emphasizes the importance of formal inference in forecast evaluation and the value of systematic model combination.

From a methodological perspective, the research demonstrates how traditional econometric approaches can be productively integrated with modern machine learning techniques within a principled statistical framework. The forecast combination methodology provides the foundation for combining fundamentally different modeling paradigms based purely on predictive performance, allowing data to determine optimal combinations rather than forcing a priori methodological commitments.

Future research directions include several promising extensions. First, incorporating textual data from central bank communications or financial market indicators could enhance feature sets available to machine learning algorithms. Second, developing more sophisticated weighting schemes that adapt more rapidly to structural changes could improve performance during crisis periods. Third, extending the hierarchical combination framework to other macroeconomic forecasting applications would test the generalisability of these findings.

The findings carry important implications for practitioners and researchers. Rather than searching for a single "best" modeling approach, the evidence supports embracing model diversity and using formal combination methods like hierarchical forecast combination to harness complementary strengths across different specifications. The consistent dominance of machine learning methods in the performance-based weighting suggests that nonlinear approaches have matured sufficiently to serve as primary tools for macroeconomic forecasting.

Most importantly, this research contributes to the evolving understanding of how machine learning techniques can be productively integrated with traditional econometric methods through principled statistical frameworks. The success of the hierarchical combination approach demonstrates that systematic model combination can overcome the limitations of individual approaches while providing robust protection against model-specific failures.

References

[1] Bańbura, M., Giannone, D., & Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1), 71-92.

[2] Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operations Research Quarterly*, 20(4), 451-468.

[3] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

[4] Cogley, T., & Sargent, T. J. (2005). Drifts and volatilities: monetary policies and outcomes in the post WWII US. *Review of Economic Dynamics*, 8(2), 262-302.

[5] Coulombe, P. G., Leroux, M., Stevanovic, D., & Surprenant, S. (2020). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, 37(5), 920-964.

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

[7] Gürkaynak, R. S., Levin, A., & Swanson, E. (2008). Does inflation targeting anchor long-run inflation expectations? Evidence from long-term bond yields in the US, UK, and Sweden. *Journal of the European Economic Association*, 8(6), 1208-1242.

[8] Hansen, P. R., & Timmermann, A. (2007). Choice of sample split in out-of-sample forecast evaluation. *Economics Letters*, 95(3), 312-318.

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

[10] Scornet, E., Biau, G., & Vert, J. P. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4), 1716-1741.

[11] Stock, J. H., & Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics*, 44(2), 293-335.

[12] Stock, J. H., & Watson, M. W. (2007). Why has US inflation become harder to forecast? *Journal of Money, Credit and Banking*, 39(s1), 3-33.

[13] Stock, J. H., & Watson, M. W. (2015). Core inflation and trend inflation. *Review of Economics and Statistics*, 98(4), 770-784.

[14] Timmermann, A. (2006). Forecast combinations. *Handbook of Economic Forecasting*, 1, 135-196.

[15] West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64(5), 1067-1084.