

Hybrid VARIMA–Machine Learning Models for Multivariate Macroeconomic and Energy Forecasting

Emmanuel GNANDI

kpanteemmanuel@gmail.com

INSA de Toulouse

Jules SADEFO KAMDEM

Montpellier University

Fredy POKOU

Inria research centre Lille - Nord Europe

Research Article

Keywords: Linear models, machine learning, hybrid models, time series

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9734421/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Hybrid VARIMA–Machine Learning Models for Multivariate Macroeconomic and Energy Forecasting

Emmanuel GNANDI ^{*} ¹, Jules SADEFO KAMDEM ^{†2}, and Fredy POKOU ^{‡3}

¹INSA de Toulouse, Département de Génie Mathématique, Toulouse, France

²MRE UR 209 and Faculty of Economics. Montpellier University, France

³Inria, CNRS, Univ. of Lille, Centrale Lille, UMR 9189 - CRISTAL, F-59000 Lille, France

May 16, 2026

Abstract

Accurate forecasting of macroeconomic and energy time series is complicated by structural instability, strong cross-dependencies, and pervasive nonlinear dynamics. While VARIMA models provide a transparent multivariate framework for linear shock propagation, they often leave economically meaningful structure in forecast residuals. Conversely, machine learning methods offer flexibility but may suffer from instability and limited interpretability in multivariate economic environments. This paper proposes a residual-based hybrid forecasting framework that integrates VARIMA and machine learning in a disciplined and interpretable manner. VARIMA serves as a linear filter capturing common dynamics across macroeconomic, financial, and energy variables, while machine learning models estimate nonlinear corrections from multivariate residuals. Using monthly data from 1999 to 2025, including multiple crisis episodes, we document systematic violations of linear model assumptions through extensive diagnostics. Out-of-sample results show that hybrid models consistently outperform both standalone VARIMA and pure machine learning approaches in terms of accuracy and robustness, particularly during periods of stress. The findings highlight residual-based hybridization as an effective and economically interpretable strategy for multivariate forecasting.

Keywords: Linear models, machine learning, hybrid models, time series.

1 Introduction

Problem statement. Accurate forecasting of macroeconomic, financial, and energy-related time series remains a central challenge for policymakers, financial institutions, and market participants. Forecasts play a crucial role in investment decisions, risk management, and the design of monetary, fiscal, and regulatory policies. However, the empirical behavior of macro–financial systems has become increasingly complex over recent decades, reflecting structural transformations, recurrent crises, and heightened global interconnectedness. Episodes such as financial crashes, geopolitical conflicts, energy supply disruptions, and global health crises have revealed persistent departures from the assumptions underlying standard linear time-series models.

Multivariate econometric models, including ARIMA and VARMA-type specifications, provide a disciplined and interpretable framework for modeling dynamic interactions across economic variables (Box et al., 2015; Lütkepohl, 2013). When correctly specified, these models are able to capture linear propagation mechanisms and contemporaneous cross-dependencies in a coherent statistical setting. Yet, mounting empirical evidence suggests that even well-specified linear multivariate models often

^{*}kpanteemmanuel@gmail.com

[†]jules.sadefo-kamdem@umontpellier.fr

[‡]fredypokou@gmx.fr

fail to adequately describe real-world macro-financial data. In practice, economic time series frequently exhibit heavy-tailed distributions, time-varying volatility, nonlinear dependence structures, and regime-dependent dynamics that remain unaccounted for by linear conditional mean models (Shiller, 1980; Cont, 2001).

Importantly, this limitation should not be interpreted as a mere lack of flexibility or as a failure of estimation, but rather as a form of *structural misspecification*. By construction, VARIMA models are restricted to modeling linear conditional expectations. As a result, even when stability, invertibility, and parsimony conditions are satisfied, these models may leave economically meaningful information embedded in their residuals. Empirically, such residuals often display strong deviations from Gaussianity, conditional heteroskedasticity, and nonlinear dependence patterns, indicating that they cannot be treated as white noise. While extensions such as GARCH-type models or copula-based approaches partially address these issues, they rely on strong parametric assumptions and rapidly become difficult to estimate and interpret in multivariate and high-dimensional settings (Bollerslev, 1986; Tsay, 2013).

In parallel, machine learning (ML) methods have gained increasing attention in economic forecasting due to their ability to approximate complex nonlinear relationships. Algorithms such as tree-based ensembles, support vector machines, and deep neural networks have demonstrated strong empirical performance in a wide range of applications. However, purely data-driven ML approaches also face important limitations in macroeconomic contexts. They typically lack structural interpretability, may suffer from instability in moderate sample sizes, and can exhibit limited robustness when applied directly to multivariate systems characterized by strong cross-correlations and structural breaks (Granger, 1993). These shortcomings are particularly problematic in policy-relevant forecasting, where transparency, robustness, and economic coherence are essential.

These observations raise a fundamental question at the intersection of econometrics and computational economics: how can one construct forecasting models that retain the structural discipline and interpretability of multivariate econometric frameworks, while systematically correcting their structural limitations in the presence of nonlinear and regime-dependent dynamics?

Main contributions. This paper addresses this question by proposing and empirically validating a hybrid multivariate forecasting framework that combines VARIMA models with machine learning techniques in a residual-based architecture. The core premise of the paper is that VARIMA models, when correctly specified, act as disciplined linear filters that capture common propagation mechanisms and cross-dependencies across macroeconomic, financial, and energy variables, but leave a structured residual component that reflects nonlinear dynamics and conditional effects. Machine learning models are then employed as nonlinear correction operators applied to this residual component, rather than as substitutes for econometric modeling.

The contributions of the paper are fourfold.

First, we provide a formal empirical assessment of the structural limitations of multivariate VARIMA models in a macro-energy forecasting context. Using an extensive battery of diagnostic tests, including tests for residual autocorrelation, conditional heteroskedasticity, non-normality, and nonlinear dependence, we document the presence of persistent residual structures that are inconsistent with the white-noise assumption. These results establish that the inadequacy of linear models is not merely a matter of forecast accuracy, but reflects a deeper form of structural misspecification.

Second, we propose a fully multivariate hybrid modeling strategy grounded in econometric principles. By decomposing the data-generating process into a linear component captured by VARIMA and a nonlinear residual component learned by machine learning algorithms, the proposed framework preserves interpretability while reducing dimensional complexity and improving robustness. Unlike approaches that apply machine learning directly to raw data, residual-based learning allows nonlinear models to focus on economically meaningful corrections rather than rediscovering linear dependencies already captured by the econometric structure.

Third, we conduct a large-scale forecasting exercise over multiple decades, covering periods of pronounced macroeconomic and energy-market stress. Forecasting performance is evaluated using rolling-window procedures consistent with the standards of forecast econometrics. In addition to conventional accuracy measures, we employ formal tests of predictive accuracy, including Diebold–Mariano statistics, to assess whether observed performance differences are statistically significant. The results show that hybrid VARIMA–ML models systematically outperform both standalone VARIMA specifications and pure machine learning models across aggregated multivariate performance metrics.

Fourth, the paper contributes to the interpretability of hybrid forecasting models by linking diagnostic evidence, forecasting gains, and economic structure. Using explainable machine learning tools, we provide insights into how nonlinear corrections adjust the linear forecasts, and we show that performance improvements are particularly pronounced during periods of heightened economic uncertainty and structural stress. This analysis clarifies why hybrid models outperform both linear and purely data-driven alternatives, highlighting a transparent bias–variance trade-off and the complementary roles of linear filtering and nonlinear learning.

Related literature. The first strand of literature relevant to this paper concerns linear multivariate time-series models, in particular Vector Autoregressive (VAR) and Vector Autoregressive Moving Average (VARMA) frameworks. Since the seminal contribution of Sims (1980), VAR-type models have played a central role in empirical macroeconomics by providing a flexible yet interpretable representation of dynamic interdependencies among economic variables. Subsequent theoretical developments by Hannan (1979) and comprehensive treatments such as Lütkepohl (2013) established the statistical foundations of VAR and VARMA processes, including issues of identification, estimation, and stability. Despite their widespread use, numerous empirical studies document the limitations of linear multivariate models when applied to macro–financial data characterized by non-Gaussian distributions, volatility clustering, and structural instability (Cont, 2001; Tsay, 2013). These features undermine the adequacy of linear conditional mean models, particularly in forecasting applications during periods of economic stress.

A second strand of the literature addresses these limitations through nonlinear and volatility-based extensions of linear models. ARCH and GARCH-type models, introduced by Engle (1982) and extended by Bollerslev (1986), explicitly model time-varying conditional variances and have become standard tools in financial econometrics. Multivariate extensions, including BEKK, DCC, and copula-based specifications, aim to capture joint volatility dynamics and cross-sectional dependence. However, these models often suffer from parameter proliferation and estimation difficulties in high-dimensional systems, and they primarily target second-moment dynamics, leaving nonlinearities in the conditional mean insufficiently modeled (Engle, 2002; Tsay, 2013).

A third and rapidly growing body of literature explores the use of machine learning methods for economic and financial forecasting. Tree-based algorithms such as Random Forests (Breiman, 2001) and Gradient Boosting (Friedman, 2001), as well as support vector regression methods rooted in statistical learning theory (Vapnik, 2013), have been shown to capture complex nonlinear interactions with strong empirical performance. Deep learning architectures, including recurrent neural networks and LSTM models (Hochreiter and Schmidhuber, 1997; Cho et al., 2014), have also been applied to sequential economic data. Nevertheless, several studies highlight concerns regarding the stability, interpretability, and economic coherence of purely data-driven models, especially in multivariate settings with limited sample sizes and strong cross-dependencies (Makridakis et al., 2018).

The strand most closely related to the present study concerns hybrid models that combine linear econometric structures with nonlinear learning mechanisms. Early contributions by Box and Jenkins (1976) and Zhang (2003) introduced the idea of decomposing time series into linear and nonlinear components. More recent studies document empirical gains from such hybrid approaches, but often focus on univariate settings or provide limited insight into the mechanisms underlying performance improvements. Recent multivariate extensions, such as Pokou et al. (2024) and Pokou

et al. (2026), represent important steps forward, yet frequently lack a systematic diagnostic justification and an explicit economic interpretation of residual-based nonlinear corrections. The present paper contributes to this literature by proposing a fully multivariate hybrid VARIMA–machine learning framework explicitly grounded in econometric diagnostics, thereby providing a transparent explanation of when and why hybridization improves forecasting performance.

Paper organization. The remainder of the paper is organized as follows. Section 2 introduces the linear modeling framework and presents the VARMA and VARIMA specifications used as benchmarks. Section 3 describes the machine learning techniques employed, while Section 4 presents the hybridization strategy combining VARIMA models with machine learning methods. Section 5 describes the data, empirical methodology, and numerical experiments. Section 6 discusses the main empirical results and their economic interpretation. Section 6 concludes and outlines the limitations and potential extensions of the proposed approach.

2 Linear Models

This section sets out the linear multivariate time-series framework underlying our empirical analysis and, crucially, clarifies its role as a *structural linear benchmark*. We focus on the Vector Autoregressive Moving Average (VARMA) family and its integrated counterpart (VARIMA), which provide a principled and parsimonious representation of intertemporal dependence and cross-variable feedback in multivariate systems. VARMA models are theoretically attractive because they jointly accommodate autoregressive propagation and moving-average transmission of shocks; however, their empirical adoption has historically lagged behind VAR specifications, largely due to computational and identification challenges relative to least-squares VAR methods. The computational and practical aspects of VARMA modeling are discussed extensively in Athanasopoulos and Vahid (2008a).

The popularity of VAR models nonetheless comes at a substantive cost. A VAR(p) implicitly approximates a potentially richer Wold representation by truncation. When the underlying data-generating process exhibits non-negligible moving-average dynamics, VAR residuals can remain serially correlated, compromising diagnostic adequacy and potentially distorting inference and causality assessments. Moreover, capturing persistent or intricate dynamics by increasing the lag order rapidly inflates parameter dimensionality, eroding statistical power and aggravating finite-sample instability. These issues are especially salient in multivariate settings, where the number of coefficients grows quadratically in the dimension and linearly in the lag length. In forecasting applications, the omission of MA terms may therefore be detrimental, and VARMA specifications can deliver material accuracy gains relative to high-order VAR approximations (Athanasopoulos and Vahid, 2008b).

Against this background, the VARMA family offers a natural route to parsimony: by explicitly modeling both autoregressive and moving-average components, VARMA(p, q) can replicate rich *linear* dependence structures with fewer parameters than a purely autoregressive approximation of comparable fidelity. The trade-off is methodological: estimation typically requires nonlinear likelihood-based procedures and careful attention to identification, stability, and invertibility constraints. These considerations motivate our emphasis on transparent specification and well-established regularity conditions, drawing on the standard time-series literature (Hannan, 1979; Lütkepohl, 2013; Reinsel, 2003; Brockwell and Davis, 2009; Tsay, 2013; Kilian and Lütkepohl, 2017).

Importantly, even when a VARMA/VARIMA model is correctly specified according to information criteria and satisfies stability and invertibility conditions, it remains constrained by construction: it models the *linear conditional mean* of the multivariate process. Consequently, departures from Gaussianity, conditional heteroskedasticity, nonlinear dependence, or regime-dependent behavior are not mere nuisances; they constitute evidence that the linear model is *structurally misspecified* relative to the data-generating process. In our empirical setting, this perspective is essential: the residual process generated by the best linear model is not treated as a by-product, but as an object

of inference whose structure will motivate the residual-based hybrid correction developed later in the paper.

2.1 Model specification and foundations

Let $\{\mathbf{Y}_t\}_{t \in \mathbb{Z}}$ denote an N -dimensional vector of endogenous variables. A VARMA(p, q) process is defined by

$$\mathbf{Y}_t = \mathbf{c} + \sum_{i=1}^p \Phi_i \mathbf{Y}_{t-i} + \varepsilon_t + \sum_{j=1}^q \Theta_j \varepsilon_{t-j}, \quad t \in \mathbb{Z}, \quad (1)$$

where $\mathbf{c} \in \mathbb{R}^N$ is a vector of intercepts, $\{\Phi_i\}_{i=1}^p$ and $\{\Theta_j\}_{j=1}^q$ are $N \times N$ coefficient matrices, and $\{\varepsilon_t\}$ is an innovation process satisfying $\mathbb{E}(\varepsilon_t) = \mathbf{0}$, $\mathbb{E}(\varepsilon_t \varepsilon_t^\top) = \Sigma_\varepsilon$ with Σ_ε positive definite, and $\mathbb{E}(\varepsilon_t \varepsilon_s^\top) = \mathbf{0}$ for $s \neq t$. Equation (1) nests several canonical models. When $q = 0$, it reduces to the VAR(p) model (Sims, 1980):

$$\mathbf{Y}_t = \mathbf{c} + \sum_{i=1}^p \Phi_i \mathbf{Y}_{t-i} + \varepsilon_t. \quad (2)$$

When $p = 0$, it yields the VMA(q) model:

$$\mathbf{Y}_t = \mathbf{c} + \varepsilon_t + \sum_{j=1}^q \Theta_j \varepsilon_{t-j}. \quad (3)$$

A compact operator form is obtained using the lag operator L , where $L^s \mathbf{Y}_t = \mathbf{Y}_{t-s}$. Define the matrix polynomials

$$\Phi(L) = \mathbf{I}_N - \Phi_1 L - \cdots - \Phi_p L^p, \quad \Theta(L) = \mathbf{I}_N + \Theta_1 L + \cdots + \Theta_q L^q,$$

with $\mathbf{I}_N$ the $N \times N$ identity matrix. Then (1) can be written as

$$\Phi(L) \mathbf{Y}_t = \mathbf{c} + \Theta(L) \varepsilon_t. \quad (4)$$

2.1.1 Stability, invertibility, and fundamental representations

For VARMA(p, q) processes, stability and invertibility are standard regularity conditions ensuring well-defined infinite-order representations and interpretable dynamic propagation (Kilian, 2006; Kilian and Lütkepohl, 2017). Stability requires that all roots $z \in \mathbb{C}$ of the characteristic equation

$$\det(\mathbf{I}_N - \Phi_1 z - \cdots - \Phi_p z^p) = 0 \quad (5)$$

lie outside the unit circle ($|z| > 1$). Under stability, the process admits a convergent VMA(∞) representation, which is central for impulse-response analysis and forecast construction in multivariate systems (Kilian and Lütkepohl, 2017).

An equivalent characterization uses the companion form of the autoregressive polynomial. Let B denote the $Np \times Np$ companion matrix,

$$B = \begin{bmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_{p-1} & \Phi_p \\ \mathbf{I}_N & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{I}_N & \mathbf{0} & \mathbf{0} \end{bmatrix}. \quad (6)$$

Then (5) holds if and only if all eigenvalues of B have modulus strictly less than one (Hyndman et al., 2008; Kilian and Lütkepohl, 2017). This equivalence is convenient for empirical diagnostics and for imposing stability in estimation algorithms.

Invertibility analogously requires that all roots $z \in \mathbb{C}$ of

$$\det(\mathbf{I}_N + \Theta_1 z + \cdots + \Theta_q z^q) = 0 \quad (7)$$

lie outside the unit circle ($|z| > 1$). Invertibility ensures that the innovation process can be recovered from present and past observations and that the model admits a convergent VAR(∞) representation. Specifically, under invertibility,

$$\Psi(L)Y_t = \eta + \varepsilon_t, \quad (8)$$

where $\Psi(L) = \Theta(L)^{-1}\Phi(L) = I_N - \sum_{i=1}^{\infty} \Psi_i L^i$ and $\eta = \Theta(L)^{-1}c$. The matrices $\{\Psi_i\}$ can be obtained recursively from $\{\Phi_i\}$ and $\{\Theta_j\}$ via

$$\Psi_i = \Phi_i + \Theta_i - \sum_{j=1}^{i-1} \Theta_{i-j} \Psi_j, \quad i = 1, 2, \dots, \quad (9)$$

with $\Theta_0 = I_N$, $\Theta_i = \mathbf{0}$ for $i > q$, and $\Phi_i = \mathbf{0}$ for $i > p$.

2.1.2 Identification and canonical parameterizations

A central methodological issue in VARMA modeling is identification. While the composite operator $\Psi(L) = \Theta(L)^{-1}\Phi(L)$ is pinned down by the second-order properties (autocovariances) of $\{Y_t\}$, the decomposition into separate autoregressive and moving-average polynomials is generally not unique, so distinct pairs $(\Phi(L), \Theta(L))$ may imply the same stochastic structure for the observed series (Tsay, 2013). This non-uniqueness necessitates additional restrictions to obtain a well-defined parameterization and enable consistent estimation.

The identification problem and its remedies are classical, with influential contributions including Hannan (1969) and Akaike (1976). In practice, researchers typically impose canonical forms or structured restrictions that deliver uniqueness while preserving flexibility. A widely used approach is the block-identifiability condition of Dunsmuir and Hannan (1976), which provides a tractable framework for constructing identified VARMA specifications and plays a key role in likelihood-based estimation procedures. In our empirical strategy, these identification considerations are not only technical requirements: they ensure that the linear benchmark is economically interpretable and statistically well-defined, thereby making any remaining residual structure meaningful rather than an artifact of ill-posed parametrization.

2.1.3 Linear benchmarks and diagnostic adequacy: why residuals matter

In forecast econometrics, a linear model is not validated solely by achieving a low in-sample fit or by being selected by information criteria; it must also satisfy *diagnostic adequacy*. For VARMA/VARIMA models, diagnostic adequacy means that the fitted innovations behave as an approximately uncorrelated, homoskedastic, and weakly Gaussian white-noise process. When these conditions fail, the residual process carries systematic information that the linear model is structurally unable to encode.

This point is central to the logic of our hybrid approach. A VARIMA model is designed to capture linear conditional mean dynamics and linear cross-variable propagation. If the fitted residuals exhibit (i) significant serial correlation, (ii) conditional heteroskedasticity (ARCH effects), (iii) strong deviations from multivariate normality, or (iv) detectable nonlinear dependence, then the interpretation is not merely that the model “could be improved” but that the linear conditional mean specification is incomplete. In macro-financial and energy contexts, such residual patterns are economically interpretable: they may reflect regime switches, threshold effects, asymmetric responses to shocks, or state-dependent transmission mechanisms that are activated during stress episodes.

In the empirical sections of the paper, we therefore treat residual diagnostics as *evidence of structural misspecification*. Specifically, the joint presence of residual autocorrelation, multivariate non-normality, and conditional heteroskedasticity indicates that the VARIMA benchmark fails to deliver innovations that can be regarded as pure shocks. This motivates a residual-based correction strategy: rather than replacing the linear model, we preserve it as an interpretable first-stage filter and

explicitly model the remaining residual structure using flexible nonlinear learners. This decomposition is econometrically coherent because it isolates linear propagation mechanisms while letting the nonlinear component target only what is demonstrably left unexplained.

2.2 Vector Autoregressive Integrated Moving Average (VARIMA)

Macroeconomic, financial, and energy series frequently exhibit stochastic trends, structural persistence, or unit roots, making integration an essential feature of the empirical time-series environment. VARIMA models extend VARMA by allowing the observed vector to be integrated of finite order while preserving a stationary linear dynamic structure for the appropriately differenced process (Hyndman et al., 2008; Kilian and Lütkepohl, 2017).

Formally, an N -dimensional VARIMA(p, d, q) process $\{\mathbf{Y}_t\}$ with $d \in \mathbb{Z}_{\geq 0}$ satisfies: (i) $\mathbf{Y}_t \sim I(d)$, and (ii) the d -th difference $\nabla^d \mathbf{Y}_t$ follows a stationary VARMA(p, q) model. Using the difference operator $\nabla \mathbf{Y}_t = \mathbf{Y}_t - \mathbf{Y}_{t-1}$ and its d -fold composition $\nabla^d = \underbrace{\nabla \circ \dots \circ \nabla}_{d \text{ times}}$, the model is

$$\nabla^d \mathbf{Y}_t = \mathbf{c} + \sum_{i=1}^p \Phi_i \nabla^d \mathbf{Y}_{t-i} + \varepsilon_t + \sum_{j=1}^q \Theta_j \varepsilon_{t-j}, \quad (10)$$

where $\{\varepsilon_t\}$ is a multivariate white-noise innovation with $\mathbb{E}(\varepsilon_t) = \mathbf{0}$ and $\mathbb{E}(\varepsilon_t \varepsilon_s^\top) = \Sigma_\varepsilon \mathbf{1}_{\{t=s\}}$.

Under the stability and invertibility conditions stated above (applied to the VARMA dynamics of $\nabla^d \mathbf{Y}_t$), VARIMA admits the standard infinite-order representations and inherits the usual interpretability properties of stationary multivariate linear systems (Hyndman et al., 2008; Kilian and Lütkepohl, 2017). At the same time, because VARIMA is restricted to linear conditional mean dynamics, it provides a well-defined benchmark against which residual diagnostics can reveal the presence of nonlinear or state-dependent mechanisms. This benchmark role is essential for the hybrid methodology developed subsequently: the linear component captures systematic linear propagation, while the residual component becomes the target of a nonlinear correction that is justified empirically by diagnostic evidence.

3 Machine Learning approach in time series modeling

Linear multivariate models such as VARMA/VARIMA provide an analytically transparent benchmark for dynamic dependence, impulse propagation, and multi-step forecasting. Yet, in many economic and energy-market applications, their practical adequacy is challenged by empirical features that systematically violate the Gaussian linear paradigm: conditional heteroskedasticity, fat tails, asymmetric responses to shocks, regime changes, and state-dependent interactions. While the econometric literature offers sophisticated linear-nonlinear extensions including VARMA-GARCH and DCC-type specifications, copula-based dependence structures, and time-varying parameter SVARs their empirical gains in predictive accuracy are often limited, highly model-dependent, and obtained at the cost of substantial parametric complexity and fragile estimation in moderate samples. Recent evidence in energy markets emphasizes this trade-off between interpretability and forecasting performance, and documents that even copula-enhanced TVP-SVAR frameworks may remain outperformed by flexible nonlinear predictors in real-time forecasting exercises (Pokou et al., 2026).

Accordingly, we adopt machine learning (ML) methods not as a substitute for economic structure, but as a disciplined complement designed to capture nonlinearities and interactions that persist after controlling for linear dynamics. In our setting, the VARIMA baseline delivers a parsimonious linear approximation; ML models are then used either as standalone nonlinear forecasters or as residual learners in a hybrid specification. This section is intentionally pedagogical: it provides a transparent description of the learning problems, the role of regularization, and the hyperparameters that control the bias-variance trade-off in time-series forecasting.

Two considerations motivate this choice. First, conditional variance and nonlinear dependence can be modeled econometrically, but the associated estimation (e.g., high-dimensional VARMA-GARCH/DCC or copula-based multivariate dynamics) is computationally demanding and sensitive to specification choices; the literature indicates that such models do not uniformly dominate in predictive accuracy in energy applications (Pokou et al., 2026). Second, modern ML methods provide flexible function classes with built-in regularization, allowing systematic evaluation under rolling-window forecasting protocols, thereby prioritizing predictive reliability while maintaining methodological transparency through fully documented training and validation design.

3.1 Forecasting as regularized risk minimization under temporal dependence

Let $\{(y_{t+1}, \mathbf{x}_t)\}_{t=1}^{T-1}$ denote the supervised learning representation of a one-step-ahead forecasting problem, where y_{t+1} is the target (a component of $\mathbf{Y}_{t+1}$) and $\mathbf{x}_t$ is an information vector constructed from lagged values, transformations, and (in the hybrid case) VARIMA residual features. We consider predictors f_θ from a model class $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ and estimate them by *regularized empirical risk minimization*:

$$\hat{\theta} \in \arg \min_{\theta \in \Theta} \frac{1}{|\mathcal{T}_{\text{train}}|} \sum_{t \in \mathcal{T}_{\text{train}}} \ell(y_{t+1}, f_\theta(\mathbf{x}_t)) + \lambda \mathcal{R}(\theta), \quad (11)$$

where $\ell(\cdot, \cdot)$ is a forecasting loss (typically squared or absolute error), $\mathcal{R}(\theta)$ is a regularization term, and $\lambda > 0$ controls model complexity. Equation (11) highlights the central forecasting principle: better accuracy does not come from flexibility alone, but from selecting a complexity level that generalizes to future periods (bias-variance trade-off).

Because time-series observations are dependent and ordered, validation and tuning must respect chronology. We adopt rolling (or expanding) windows and restrict all hyperparameter choices to past information. Concretely, for a sequence of forecast origins $t \in \mathcal{T}_{\text{test}}$, a model is (re-)trained on $\{1, \dots, t\}$ and evaluated on $t + 1$, producing a forecast path. Hyperparameters are selected by minimizing validation loss computed on a set of past-origin forecasts. This prevents look-ahead bias and targets temporal generalization.

3.2 Tree-based learners

Tree-based methods are natural candidates for time-series prediction when nonlinear threshold effects, interaction terms, and piecewise regimes are plausible. They are also attractive in applied macro-finance and energy contexts due to numerical stability and competitive performance under relatively mild tuning.

3.2.1 Classification and Regression Trees (CART)

The CART methodology of Breiman et al. (2017) approximates an unknown regression function through recursive partitioning. A regression tree defines a partition $\{S_j\}_{j=1}^J$ of the feature space and predicts a constant on each region:

$$f(\mathbf{x}) = \sum_{j=1}^J c_j \mathbf{1}_{\{\mathbf{x} \in S_j\}}, \quad c_j = \mathbb{E}[Y \mid \mathbf{X} \in S_j]. \quad (12)$$

The hyperparameters controlling complexity are *max depth*, *minimum samples per leaf*, and the *minimum impurity decrease* required to split. For time-series forecasting, these parameters are crucial: deeper trees can fit idiosyncratic patterns (low bias) but may overfit regime-specific noise (high variance), especially in moderate samples.

3.2.2 Random Forests

Random Forests (RF), introduced by [Breiman \(2001\)](#), reduce variance by averaging many randomized trees. A Breiman-type RF predictor can be written as

$$\hat{f}_{\text{RF}}(\mathbf{x}) = \frac{1}{D} \sum_{d=1}^D T(\mathbf{x}; \xi_d, Z_d^*), \quad (13)$$

where $T(\cdot)$ is a tree trained on a bootstrap sample Z_d^* and randomized by ξ_d (e.g., random feature subsampling at each split). Although classical RF theory assumes i.i.d. data, recent extensions address dependent settings relevant for time-series prediction ([Goehry, 2020](#); [Goehry et al., 2023](#)).

Key hyperparameters (practical interpretation).

- *Number of trees D* : larger D reduces Monte Carlo variance and stabilizes forecasts, at the cost of computation.
- *Max depth / min leaf size*: controls individual-tree variance; shallower trees generalize better under structural breaks.
- *Feature subsampling* (mtry): smaller values increase tree diversity (variance reduction) but may increase bias if too small.
- *Bootstrap vs subsampling*: affects robustness; subsampling may be preferable in dependent data.

In macro-energy forecasting, RFs often perform well because they capture nonlinear interactions while remaining strongly regularized by averaging.

3.2.3 Gradient Boosting Decision Trees

Gradient Boosting constructs an additive predictor by sequentially fitting weak learners to the negative gradient of a chosen loss ([Friedman, 2001](#)). Let f_m denote the model at iteration m , with initialization f_0 . At step m , pseudo-residuals are

$$r_t^{(m)} = - \left. \frac{\partial \ell(y_{t+1}, f(\mathbf{x}_t))}{\partial f(\mathbf{x}_t)} \right|_{f=f_{m-1}}, \quad (14)$$

and a regression tree h_m is fit to $\{(\mathbf{x}_t, r_t^{(m)})\}$. The boosted predictor is updated as

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \nu h_m(\mathbf{x}), \quad (15)$$

where $\nu \in (0, 1)$ is the learning rate (shrinkage).

Key hyperparameters (why they matter).

- *Number of boosting iterations M* : higher M increases flexibility but can overfit; controlled via early stopping.
- *Learning rate ν* : smaller ν improves generalization but requires larger M (more computation).
- *Tree depth* (or number of leaves): depth 1-3 captures low-order interactions and is often robust in moderate samples.
- *Subsampling* (row/column): adds randomness and reduces overfitting; particularly useful under regime instability.

This explicit set of regularization levers explains why boosting can outperform deeper neural networks in macroeconomic time series: it provides strong performance while keeping variance under control.

3.3 Kernel methods

3.3.1 Support Vector Regression

Support Vector Regression (SVR) implements structural risk minimization through a convex optimization problem with a unique optimum (Vapnik, 2013). Given $(\mathbf{x}_t, y_{t+1})_{t \in \mathcal{T}_{\text{train}}}$, SVR estimates

$$f(\mathbf{x}) = \langle w, \phi(\mathbf{x}) \rangle + b, \quad (16)$$

where $\phi(\cdot)$ is a feature map induced by a kernel. Using the ε -insensitive loss, the primal problem is

$$\min_{w, b, \xi, \xi^*} \frac{1}{2} \|w\|^2 + C \sum_{t \in \mathcal{T}_{\text{train}}} (\xi_t + \xi_t^*) \quad \text{s.t.} \quad \begin{cases} y_{t+1} - \langle w, \phi(\mathbf{x}_t) \rangle - b \leq \varepsilon + \xi_t, \\ \langle w, \phi(\mathbf{x}_t) \rangle + b - y_{t+1} \leq \varepsilon + \xi_t^*, \\ \xi_t, \xi_t^* \geq 0. \end{cases} \quad (17)$$

This formulation highlights why SVR is stable: regularization is explicit, and training is convex.

Key hyperparameters.

- C : penalty for errors beyond ε ; large C fits training data closely (risk of overfitting), small C increases bias.
- ε : width of the insensitive tube; larger ε yields sparser solutions and can improve robustness to noise.
- Kernel choice and bandwidth (e.g., RBF γ): controls smoothness of the fitted function; overly large γ overfits local noise.

SVR is therefore a strong nonlinear benchmark when sample sizes are limited and stable optimization is desirable (Smola and Schölkopf, 2004; Welling, 2004; Mukherjee et al., 1997; Müller et al., 1997).

3.4 Neural architectures for multivariate time series

Neural-network models can represent complex nonlinear relationships through layered compositions. In forecasting, their practical success depends critically on regularization and validation design, particularly in moderate samples with structural breaks.

3.4.1 Feedforward networks (MLP)

A multilayer perceptron implements nonlinear regression by composing affine maps and nonlinear activations (Rosenblatt, 1957; Azoff, 1994). For an input $\mathbf{x}$, an L -layer MLP (Fig. 1) can be written compactly as

$$h^{(0)} = \mathbf{x}, \quad h^{(\ell)} = \varphi\left(W^{(\ell)}h^{(\ell-1)} + b^{(\ell)}\right) \quad (\ell = 1, \dots, L), \quad \hat{y} = W^{(L+1)}h^{(L)} + b^{(L+1)}. \quad (18)$$

Training minimizes a loss (e.g., MSE) via stochastic optimization, with generalization controlled by regularization.

Key hyperparameters.

- *Width/depth*: more layers/neurons increase flexibility but can overfit in moderate samples.
- *Weight decay* (L2 penalty) and *dropout*: reduce variance and improve robustness.
- *Learning rate* and *batch size*: control optimization stability.
- *Early stopping*: stops training when validation loss no longer improves, crucial for time-series generalization.

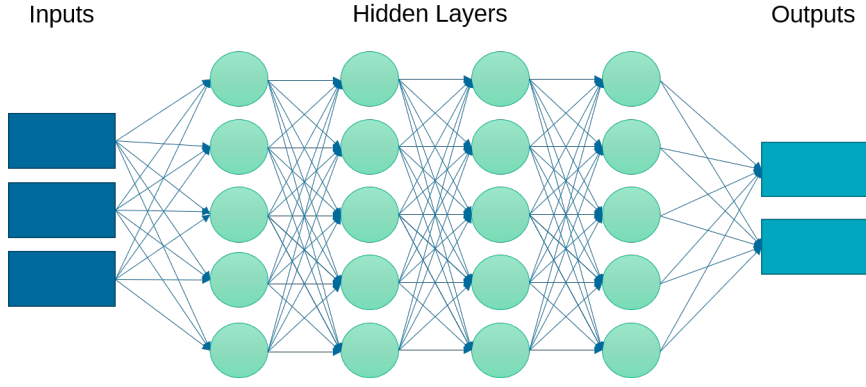


Figure 1: MLP (Rahman et al., 2024)

3.4.2 Recurrent neural networks (RNN)

Recurrent neural networks (Fig. 2) maintain a hidden state to model sequential dependence. A standard RNN updates

$$h_t = \psi(W_{xh}\mathbf{x}_t + W_{hh}h_{t-1} + b_h), \quad \hat{y}_{t+1} = W_{hy}h_t + b_y, \quad (19)$$

where $\psi(\cdot)$ is an activation function. Basic RNNs can suffer from vanishing or exploding gradients in long sequences, motivating gated architectures.

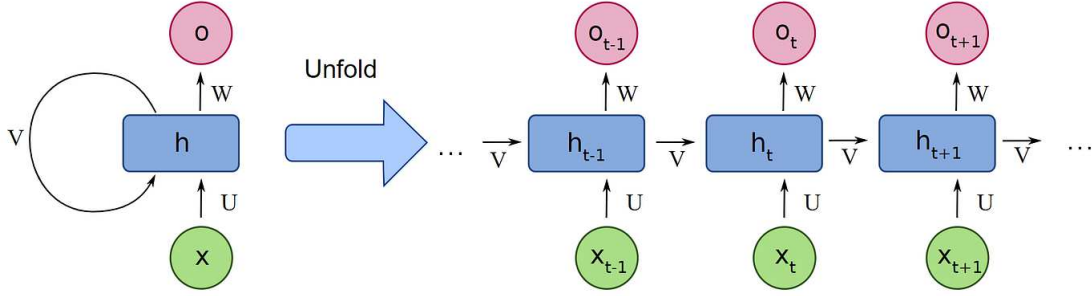


Figure 2: Recurrent neural network architecture (Lin, 2023)

3.4.3 Long short-term memory (LSTM) and gated recurrent units (GRU)

LSTM networks (Fig. 3) introduce gating mechanisms to regulate information flow and stabilize training (Hochreiter and Schmidhuber, 1997). Using $\sigma(\cdot)$ and $\otimes$ for element-wise multiplication, an LSTM cell can be summarized by

$$\begin{aligned} f_t &= \sigma(W_f[\mathbf{x}_t, h_{t-1}] + b_f), & i_t &= \sigma(W_i[\mathbf{x}_t, h_{t-1}] + b_i), \\ \tilde{c}_t &= \tanh(W_c[\mathbf{x}_t, h_{t-1}] + b_c), & c_t &= f_t \otimes c_{t-1} + i_t \otimes \tilde{c}_t, \\ o_t &= \sigma(W_o[\mathbf{x}_t, h_{t-1}] + b_o), & h_t &= o_t \otimes \tanh(c_t). \end{aligned} \quad (20)$$

GRU networks provide a streamlined alternative with fewer parameters (Cho et al., 2014).

Key hyperparameters.

- *Sequence length* (lookback window): too short misses dependence; too long increases variance and training instability.
- *Hidden size and number of layers*: control capacity; must be conservative in moderate samples.
- *Dropout/recurrent dropout*: essential to prevent overfitting.
- *Learning rate schedule and early stopping*: stabilize optimization under non-stationarity.

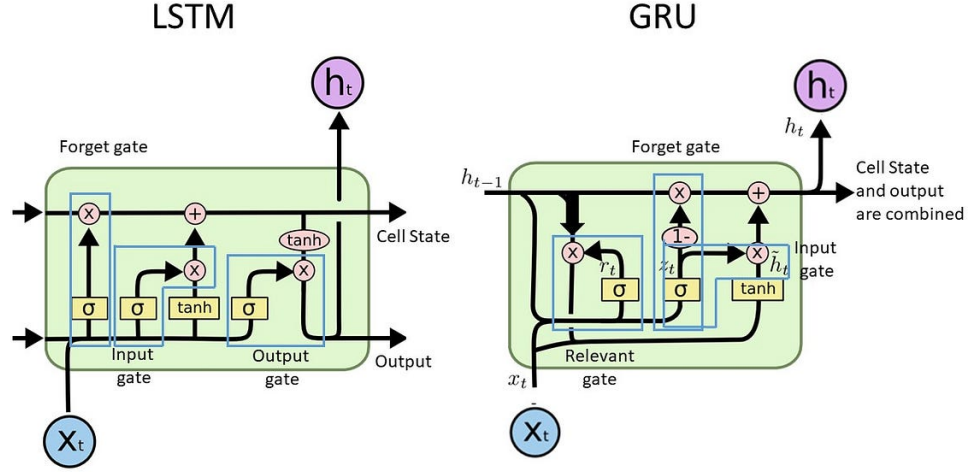


Figure 3: LSTM architecture and GRU architecture (Goswami, 2020)

3.4.4 Convolutional models and temporal convolutional networks (TCN)

Convolutional neural networks (CNNs) (Fig. 4) and temporal convolutional networks (TCN) (Fig. 5) extract local and multiscale temporal patterns through causal and dilated convolutions, often with improved training stability relative to recurrent models (LeCun et al., 2002; Ding et al., 2015; Yu et al., 2017; Navon and Keller, 2017). TCNs combine causal convolutions, dilation, and residual connections to enlarge receptive fields while preserving temporal order (Waibel et al., 2013; Yu and Koltun, 2015; He et al., 2016; Bai, 2018).

Key hyperparameters.

- *Kernel size and dilation schedule*: determine receptive field (effective memory).
- *Number of filters and depth*: control capacity and variance.
- *Dropout and weight decay*: prevent overfitting under regime shifts.

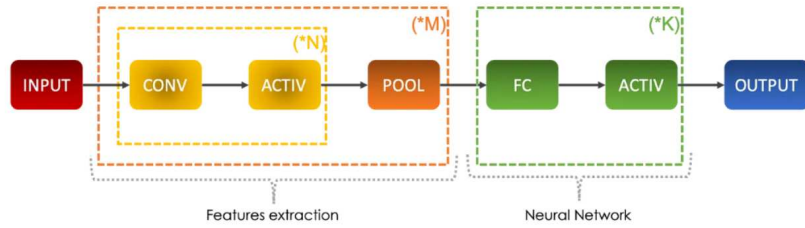


Figure 4: CNN architecture (Pokou et al., 2024)

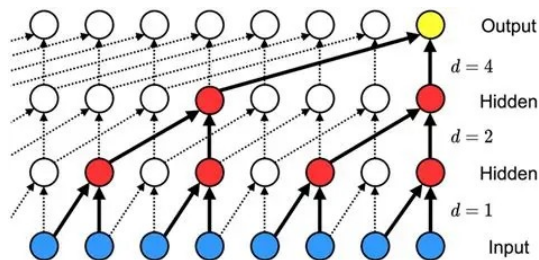


Figure 5: TCN architecture (Lee et al., 2021)

3.5 Summary: regularization knobs and their role in our empirical design

Across all ML families, the crucial point for time-series forecasting is that *hyperparameters are not technical details*: they are the mechanism through which each model implements a bias-variance compromise under temporal dependence. Tree-based models regularize through depth, subsampling, and shrinkage; SVR regularizes through margin parameters and kernel smoothness; neural networks regularize through architectural conservatism, weight decay, dropout, and early stopping. This common structure motivates our empirical protocol: we tune hyperparameters using rolling validation windows and report the selected values explicitly to ensure reproducibility and to make the comparison between pure ML and hybrid residual learning economically interpretable.

4 Hybridization of VARIMA and ML

Empirically, macro-financial and energy time series often display a layered structure: (i) a component consistent with approximately linear propagation mechanisms (distributed lag adjustment, cross variable transmission of shocks), and (ii) nonlinear and state-dependent effects that become salient during periods of stress (threshold dynamics, regime switches, asymmetric responses, and interactions activated by volatility bursts). In such environments, a well-specified linear multivariate model can provide a disciplined and interpretable benchmark, yet still leave forecast errors with systematic structure that is economically meaningful and statistically detectable.

While nonlinear econometric multivariate models (e.g., VARMA-GARCH/DCC specifications, copula-based dependence, or time-varying parameter systems) represent natural candidates, the applied forecasting literature including energy-market evidence suggests that their gains in real-time predictive accuracy are not uniform and may come at the cost of high parametric complexity, fragile estimation, and limited scalability in moderate samples. This motivates a hybrid strategy: preserve the interpretability and statistical discipline of VARIMA as a first-stage linear filter, while introducing a flexible nonlinear learner to capture structured residual dynamics that remain after linear propagation has been accounted for.

Our approach follows the residual-based hybrid principle introduced in [Zhang \(2003\)](#) and developed in subsequent applications ([Pai and Lin, 2005](#); [Askari Moghadam and Keshmirpour, 2011](#); [Ordóñez et al., 2019](#); [Xu et al., 2022](#); [Pokou et al., 2024](#)). The key methodological point, which we emphasize throughout, is that the nonlinear learner is not trained on the raw series as an unrestricted black box; it is trained to approximate a correction term whose existence is empirically justified by residual diagnostics. This positioning makes the hybrid architecture both forecast-oriented and econometrically interpretable.

4.1 A structural decomposition: linear propagation plus nonlinear correction

Let $\{\mathbf{Y}_t\}_{t \in \mathbb{Z}}$ be an N -dimensional process (macro, financial, and energy variables). We postulate an additive decomposition

$$\mathbf{Y}_t = \mathbf{L}_t + \mathbf{N}_t, \quad (21)$$

where $\mathbf{L}_t$ captures *linear conditional-mean propagation* and cross-variable transmission of shocks, while $\mathbf{N}_t$ captures residual nonlinear and state-dependent dynamics. The decomposition in (21) is not an arbitrary modeling convenience: it corresponds to a common econometric viewpoint in which linear propagation mechanisms can coexist with nonlinear corrections triggered by changes in regimes, uncertainty, or volatility states.

Stage 1 (VARIMA as linear filter). We approximate $\mathbf{L}_t$ by a VARIMA model fitted on the training window. Denote the VARIMA one-step-ahead linear forecast by

$$\hat{\mathbf{L}}_{t+1|t} := \mathbb{E}[\mathbf{Y}_{t+1} \mid \mathcal{F}_t]_{\text{VARIMA}}, \quad (22)$$

where $\mathcal{F}_t$ is the information set generated by past observations. The associated innovation (forecast error) is

$$\varepsilon_{t+1} := Y_{t+1} - \widehat{L}_{t+1|t}. \quad (23)$$

Under a correctly specified linear Gaussian conditional-mean model, $\{\varepsilon_t\}$ should behave approximately as a white-noise innovation sequence (up to mild deviations). However, in macro-financial and energy data, residual diagnostics typically reveal persistent serial dependence, conditional heteroskedasticity, non-normality, and nonlinear dependence patterns. In this case, ε_t is not merely “noise”: it contains systematic information that VARIMA cannot encode by construction because it is restricted to linear conditional-mean dynamics.

Stage 2 (ML as nonlinear residual corrector). We model the nonlinear component as the conditional expectation of the residual given a feature map:

$$N_{t+1} \approx \mathbb{E}[\varepsilon_{t+1} \mid \mathbf{Z}_t] = g(\mathbf{Z}_t), \quad (24)$$

where $\mathbf{Z}_t$ collects predictors designed to capture structured residual dynamics. In a purely residual-based hybrid specification, we take

$$\mathbf{Z}_t = (\varepsilon_t, \varepsilon_{t-1}, \dots, \varepsilon_{t-\ell+1}), \quad (25)$$

with $\ell \geq 1$ denoting the residual lag length. In multivariate systems, $\varepsilon_t \in \mathbb{R}^N$, so $\mathbf{Z}_t$ includes lagged residual vectors and therefore retains cross-variable nonlinear interactions in a parsimonious manner.

More generally, one may augment $\mathbf{Z}_t$ to improve robustness to regime dependence by including lagged levels/returns and simple state proxies (e.g., realized volatility or absolute residuals), i.e.

$$\mathbf{Z}_t = (\varepsilon_{t:t-\ell+1}, \mathbf{Y}_{t:t-k+1}, \mathbf{s}_t), \quad (26)$$

where $\mathbf{s}_t$ is a low-dimensional state indicator capturing stress or volatility conditions. In our baseline implementation, we emphasize residual-based features to preserve the division of labor: linear propagation is handled by VARIMA, while ML targets the remaining structure.

Let $\widehat{g}$ denote the ML estimator of $g(\cdot)$ trained on the rolling training window using the protocol described in Section 3. The nonlinear correction forecast is then

$$\widehat{N}_{t+1|t} := \widehat{g}(\mathbf{Z}_t). \quad (27)$$

The final one-step-ahead hybrid forecast is

$$\widehat{Y}_{t+1|t} = \widehat{L}_{t+1|t} + \widehat{N}_{t+1|t}. \quad (28)$$

4.2 Interpretation

Equation (28) implements a clear and testable division of labor.

Statistical interpretation. VARIMA provides an estimate of the linear conditional mean $\mathbb{E}[Y_{t+1} \mid \mathcal{F}_t]$ under a parsimonious multivariate specification. If residual diagnostics indicate that ε_{t+1} is not white noise, then $\mathbb{E}[\varepsilon_{t+1} \mid \mathbf{Z}_t] \neq \mathbf{0}$ for a suitable feature set $\mathbf{Z}_t$, and the correction term $\widehat{N}_{t+1|t}$ in (27) is not ad hoc: it estimates a genuine conditional structure left unexplained by the linear benchmark. This makes the architecture falsifiable: if residuals were truly innovation-like, the ML correction would collapse to (approximately) zero and add no systematic predictive value.

Economic interpretation. The linear VARIMA component can be read as a model of *propagation of shocks* and *cross-variable transmission* (e.g., macro shocks affecting energy prices with delays, feedback loops across financial conditions). The nonlinear correction captures *state-dependent adjustments* for example, asymmetric responses during stress, threshold effects linked to uncertainty, or nonlinear amplification mechanisms activated in high-volatility regimes. In this sense, the hybrid model is not a black-box competitor to econometrics: it is an econometric benchmark augmented by an adaptive mechanism that becomes particularly relevant when the economy moves away from “normal” regimes.

4.3 One-step-ahead versus multi-step forecasting

The hybrid forecast in (28) is naturally formulated for one-step-ahead prediction, which is the building block for robust rolling evaluation. Multi-step-ahead forecasting can be handled in two standard ways.

Iterated (recursive) strategy. One can iterate the one-step-ahead mapping by feeding predicted values back into the feature construction, producing $\hat{\mathbf{Y}}_{t+h|t}$ recursively for $h > 1$. This approach preserves model parsimony but may accumulate errors.

Direct strategy. Alternatively, one can train horizon-specific predictors $g_h(\cdot)$ to correct the h -step-ahead VARIMA forecast, i.e.

$$\hat{\mathbf{Y}}_{t+h|t} = \hat{\mathbf{L}}_{t+h|t} + \hat{g}_h(\mathbf{Z}_t), \quad h = 1, 2, \dots, H. \quad (29)$$

The direct strategy is computationally heavier but often more stable when nonlinearities are horizon-dependent. In the empirical section, we focus on one-step-ahead forecasts to avoid compounding errors and to align with standard forecast evaluation practices.

4.4 Why strongly regularized nonlinear learners can dominate deep architectures

A recurrent empirical finding in macroeconomic and energy forecasting is that relatively simple nonlinear learners (e.g., gradient boosting or random forests) may outperform deep neural architectures, especially when sample sizes are moderate and regimes are unstable. The hybrid framework provides a principled explanation.

First, the VARIMA filter absorbs a substantial fraction of predictable linear variation. The residual learning task is therefore a *correction problem* rather than a full representation-learning problem. In such settings, models with strong built-in regularization and low variance such as boosting with shrinkage and early stopping can be particularly effective.

Second, deep architectures typically require large and stable datasets to reliably estimate high-dimensional representations. In macro-energy time series, training samples are limited, dependence is strong, and structural breaks are plausible; this increases estimation variance and sensitivity to hyperparameters. By contrast, tree-based ensembles implement conservative complexity control (depth, subsampling, shrinkage) that is well aligned with rolling-window evaluation and tends to yield robust out-of-sample performance.

Third, the hybrid decomposition itself reduces the risk of spurious nonlinear fit: because the nonlinear layer is trained on residual structure whose existence is documented by diagnostics, the ML model is less likely to “hallucinate” patterns in raw data and more likely to target economically meaningful deviations from linear propagation. This is precisely the sense in which hybridization improves not only forecast accuracy but also methodological credibility.

4.5 Evaluation metrics

Model comparison is conducted using standard loss functions that quantify average forecast errors and penalize large deviations. Let $\{y_i\}_{i=1}^N$ denote realized values and $\{\hat{y}_i\}_{i=1}^N$ the corresponding forecasts. We report:

1. **Mean Absolute Error (MAE):** $\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$.
2. **Mean Squared Error (MSE):** $\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$.
3. **Root Mean Squared Error (RMSE):** $\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$.

MAE is scale-consistent and robust to occasional large errors, whereas MSE and RMSE assign greater weight to large deviations and are therefore informative when tail risks and volatility bursts are economically relevant, as in energy markets.

5 Numerical Experiments

This section reports the core empirical evidence of the paper and is organized to make the econometric logic explicit and falsifiable. We proceed in five steps. First, we justify the economic relevance of the multivariate macro–financial–energy system and the out-of-sample design. Second, we document the distributional and stochastic properties of the raw data, motivating the transformation strategy. Third, we select a disciplined VARIMA benchmark and interpret its linear propagation content. Fourth, we conduct a comprehensive diagnostic analysis of the VARIMA residuals and interpret these diagnostics as evidence of structural misspecification of the linear conditional-mean benchmark. Finally, we evaluate pure ML and hybrid VARIMA–ML forecasts under a time-series protocol consistent with forecast econometrics and interpret the observed performance ranking through the lens of a bias–variance trade-off and regime instability.

Symbol	Variable name	Description	Source
DGS10	10-Year Treasury Constant Maturity Rate	U.S. long-term interest rate, proxy for financing and monetary stance	FRED
UMCSENT	University of Michigan Consumer Sentiment Index	Proxy for household expectations and consumption demand	FRED
GEPUCURRENT	Global Economic Policy Uncertainty Index	Global economic policy uncertainty, proxy for uncertainty shocks	FRED
INDPRO_US	Industrial Production Index (U.S.)	Proxy for real economic activity and industrial energy demand	FRED
USD/EUR	U.S. Dollar to Euro Exchange Rate	Measures relative USD/EUR value, proxy for global oil affordability	FRED
APE	Average Price: Electricity per kWh (U.S.)	Final energy price reflecting household and industrial electricity costs	FRED
BRENT	Brent Crude Oil Price	Global crude oil price benchmark, dependent variable of interest	FRED

Table 1: Variables, Definitions, and Data Sources

5.1 Economic relevance of the dataset and temporal design

The empirical analysis relies on a multivariate system of seven monthly macroeconomic, financial, and energy-related variables: DGS10, UMCSENT, GEPUCURRENT, INDPRO_US, USD/EUR, APE, and BRENT. These variables are drawn from authoritative sources, primarily FRED, and are summarized in Table 1. Together, they capture complementary dimensions of the global macro-energy environment: monetary and financial conditions (DGS10), household expectations (UMCSENT), global policy uncertainty (GEPUCURRENT), real economic activity (INDPRO_US), international transmission channels (USD/EUR), final energy costs faced by consumers and firms (APE), and global energy market dynamics (BRENT).

From an economic standpoint, this system is designed to capture both (i) *propagation channels* (how shocks transmit across macro, financial, and energy blocks) and (ii) *state dependence* (how responses change under stress). Monetary conditions, uncertainty shocks, and real activity jointly shape energy demand and pricing pressures, while exchange rates and electricity prices transmit international and sectoral shocks across markets. These interactions are contemporaneous and dynamic, which makes a fully multivariate modeling framework economically coherent: univariate specifications would necessarily discard cross-variable feedback effects and potentially misattribute predictive content.

The sample spans January 1999 to August 2025 and is divided into an in-sample period (January 1999–February 2023) and an out-of-sample evaluation period (March 2023–August 2025). This split is motivated by the forecasting objective: models must be trained on a long historical record encompassing multiple regimes (financial crises, COVID-19 disruption, geopolitical tensions) and then evaluated on a turbulent and policy-sensitive macro-energy environment. The out-of-sample window therefore constitutes a stringent test of real-time predictive reliability and, importantly, of the ability of hybrid models to adapt when the economy departs from “*normal*” regimes.

5.2 Statistical properties of the raw data

Descriptive statistics for the level series are reported in Table 2. Several salient features emerge. First, the distributions of the raw variables exhibit pronounced deviations from Gaussianity, as reflected in non-zero skewness, excess kurtosis, and large Jarque-Bera statistics. These features are consistent with well-documented stylized facts in macro-financial and energy data (Cont, 2001). Second, Augmented Dickey-Fuller (ADF) tests applied to the level series fail to reject the null hypothesis of a unit root for all variables, providing strong evidence of non-stationarity in levels.

The econometric implication is immediate: estimating multivariate linear models on non-stationary series may induce spurious dynamics and unstable forecasts. Since our objective is out-of-sample forecasting, we adopt a transformation strategy that enforces stationarity while preserving economically meaningful co-movements.

Series	Mean	Std	Skewness	Kurtosis	Min	Max	ADF	Jarque-Bera
DGS10	3.3173	1.3869	0.2291	-0.8033	0.6236	6.661	-1.8185	10.3349
UMCSENT	84.8517	13.5011	-0.3562	-0.5708	50	112	-2.2980	10.0691
GEPUCURRENT	138.6402	70.4967	1.1802	1.1734	47.8623	420.8612	-0.5092	83.9540
INDPRO_US	96.5080	5.0799	-0.4031	-1.0318	84.6746	104.1038	-2.1549	20.7183
USD/EUR	1.1921	0.1580	-0.0618	-0.4141	0.8525	1.5759	-1.9350	2.2564
APE	0.1211	0.0210	-0.3254	-0.8780	0.0840	0.1680	-0.1149	14.4344
BRENT	63.5525	30.4069	0.3179	-0.9280	10.2716	132.7182	-2.6801	15.2896

Table 2: Descriptive statistics of raw data.

5.3 Returns, cointegration, and transformed series

To address non-stationarity, all variables are transformed into logarithmic returns defined as

$$r_t = \log(p_t) - \log(p_{t-1}),$$

where p_t denotes the level of the corresponding series at time t . This transformation removes stochastic trends, reduces scale heterogeneity, and yields stationary series suitable for multivariate conditional-mean modeling.

A potential concern is whether differencing removes economically meaningful long-run equilibrium relations. We therefore test for cointegration. Johansen trace tests reported in Table 3 fail to reject the null hypothesis of no cointegration at the 5% level. This supports the validity of a differenced/return-based representation for forecasting.

r_0	r_1	Test statistic	Critical value
0	7	103.7	125.6

Table 3: Johansen cointegration test using trace test statistic at 5% significance.

Series	Mean	Std	Skewness	Kurtosis	Min	Max	ADF	Jarque-Bera
DGS10	-0.000876	0.08339	-1.17290	7.74906	-0.54753	0.25559	-10.6870	792.071
UMCSENT	-0.00178	0.05420	-0.53500	1.21751	-0.21588	0.12762	-14.9993	31.746
GEPUCURRENT	0.004203	0.18628	0.61043	1.12455	-0.48229	0.75070	-6.49842	33.291
INDPRO_US	0.000567	0.01193	-5.74180	72.27143	-0.14205	0.06377	-12.9562	64706.64
USD/EUR	-0.000272	0.02227	-0.00333	0.45558	-0.07794	0.06186	-10.8550	2.508
APE	0.002349	0.02039	0.61208	2.14104	-0.05237	0.07490	-3.89402	73.498
BRENT	0.006738	0.10679	-1.05915	5.75767	-0.55491	0.46910	-11.9524	454.792

Table 4: Descriptive statistics of log-returns.

After transformation, the descriptive statistics of the return series (Table 4) confirm stationarity according to ADF tests for all variables. Importantly, strong non-normality persists, with significant skewness, excess kurtosis, and Jarque-Bera statistics. This persistence is not incidental: it suggests that even after eliminating unit roots, higher-order features (tails, asymmetry, state dependence) remain, precisely the type of structure that can induce systematic residual patterns under linear modeling.

5.4 VARIMA specification and economic interpretation

The VARIMA framework provides a disciplined linear benchmark for modeling the joint dynamics of the stationary return series. Model identification follows a transparent and reproducible procedure. Based on unit-root results, the integration order is fixed at $d = 1$. We then explore a systematic grid of autoregressive and moving-average orders, as reported in Table 5, allowing $p, q \in \{1, \dots, 5\}$ and considering specifications with and without deterministic terms.

Model	Parameter	Values
VARIMA	p	1, 2, 3, 4, 5
	d	1
	q	1, 2, 3, 4, 5
	tr	'n' (no constant), 'c' (with constant)

Table 5: VARIMA model specification and parameter values.

Model selection is conducted using standard information criteria: AIC, BIC, and HQIC. The results in Table 6 unanimously select a VARIMA(1, 1, 1) specification without a constant term. This choice reflects a balance between parsimony and explanatory power. Economically, it implies that the monthly dynamics of returns are dominated by short-memory propagation and transitory shock transmission, without deterministic drift. Importantly, the selected VARIMA benchmark can be interpreted as a *linear propagation operator*: it captures how shocks in one block (e.g., monetary conditions or uncertainty) transmit to other blocks (including energy prices) through lagged cross-dependencies and MA shock channels.

The selected model is given by

$$\nabla \mathbf{Y}_t = \Phi_1 \nabla \mathbf{Y}_{t-1} + \Theta_1 \varepsilon_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \Sigma),$$

where $\mathbf{Y}_t = [DGS10_t, UMCSSENT_t, GEPU_t, INDPRO_{US,t}, USD/EUR_t, APE_t, BRENT_t]^\top$. Stability of the estimated system is confirmed by the eigenvalue analysis reported in Figure 6, ensuring the validity of the linear representation and the interpretability of impulse propagation.

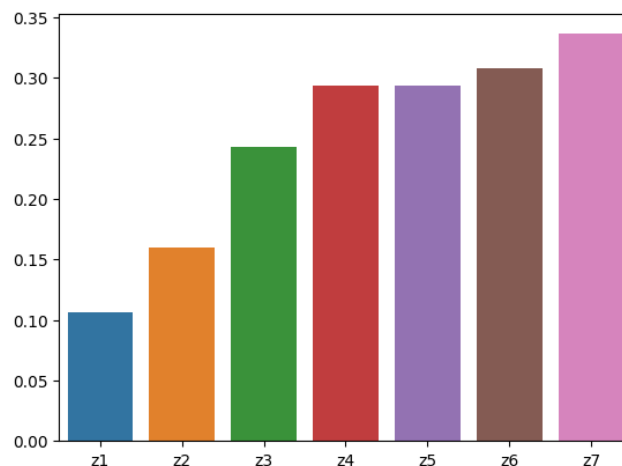


Figure 6: Absolute values of the eigenvalues λ_k of the companion matrix, all satisfying $|\lambda_k| < 1$.

	p	q	tr	AIC	BIC	HQIC
0	1	1	n	-6852.957577	-6390.987814	-6667.848783
1	1	1	c	-6844.624188	-6356.989438	-6649.231572
2	1	2	n	-6827.042518	-6185.417848	-6569.946971
10	2	1	n	-6821.352563	-6179.727892	-6564.257016
3	1	2	c	-6817.000600	-6149.710943	-6549.621232
11	2	1	c	-6811.917225	-6144.627568	-6544.537856
4	1	3	n	-6807.807042	-5986.527464	-6478.724742
20	3	1	n	-6807.361134	-5986.081556	-6478.278834
30	4	1	n	-6798.044798	-5797.110312	-6396.975745
21	3	1	c	-6797.863722	-5950.919157	-6458.497600
5	1	3	c	-6796.513898	-5949.569333	-6457.147777
31	4	1	c	-6791.168703	-5764.569231	-6379.815828
6	1	4	n	-6787.318161	-5786.383675	-6386.249107
7	1	4	c	-6783.624184	-5757.024711	-6372.271309
40	5	1	n	-6752.362015	-5571.772622	-6279.306209
12	2	2	n	-6748.567534	-5927.287956	-6419.485234
41	5	1	c	-6746.264844	-5540.010463	-6262.925216
13	2	2	c	-6738.958043	-5892.013478	-6399.591921
22	3	2	n	-6721.326332	-5720.391846	-6320.257279
8	1	5	n	-6718.675269	-5538.085876	-6245.619463

Table 6: Information criteria for VARMA models

5.5 Residual diagnostics and limits of linear modeling

A rigorous residual analysis is essential for establishing whether the VARIMA benchmark is *diagnostically adequate*. In forecast econometrics, adequacy requires that fitted residuals behave approximately as innovations (uncorrelated, homoskedastic, and weakly Gaussian). When this fails, the residual process carries systematic information that a linear conditional-mean model cannot encode, implying structural misspecification.

The Modified Portmanteau Test of Hosking (Table 7) strongly rejects the null hypothesis of no residual autocorrelation, showing that the VARIMA benchmark does not fully exhaust temporal dependence. Distributional assumptions are further challenged by the multivariate normality tests reported in Table 8: both Mardia skewness and kurtosis statistics reject Gaussianity, revealing asymmetry and heavy tails consistent with crisis-driven macro-energy data.

Volatility dynamics provide additional evidence of inadequacy. Univariate ARCH-LM tests (Table 9) reveal significant conditional heteroskedasticity in several series, while multivariate ARCH tests (Table 10) confirm cross-volatility dependence. Together, these results imply time-varying and interconnected risk conditions—economically plausible in systems exposed to uncertainty shocks and energy-market disruptions—that are not captured by a VARIMA conditional-mean specification.

Finally, nonlinear dependence is formally documented using the Teräsvirta neural network test (Table 11), the Tsay test (Table 12), and linearity tests against SETAR alternatives (Table 13). The convergence of these diagnostics is critical: it establishes that the residuals are structured in ways consistent with state dependence and regime switching. In other words, the “noise” left by VARIMA contains predictable components that are economically interpretable and statistically measurable. This provides the empirical justification for residual-based nonlinear correction.

Statistic	Value
Data	Residuals of VARIMA
Test	The Modified Portmanteau Test of Hosking
Test statistic χ^2	217.8149
Degrees of freedom (df)	98
p -value	4.3058×10^{-11}

Table 7: The Modified Portmanteau Test of Hosking applied to the residuals of VARIMA

Test	Statistic	p-value	Method	MVN
Mardia Skewness	652.050	< 0.001	Asymptotic	Not normal
Mardia Kurtosis	37.331	< 0.001	Asymptotic	Not normal

Table 8: Multivariate Mardia Normality Test for VARIMA Residuals

Statistic	Values						
	DGS10	UMCSENT	GEPUCURRENT	INDPRO_US	USD/EUR	APE	BRENT
Data	Residual	Residual	Residual	Residual	Residual	Residual	Residual
LM Statistic (χ^2)	1.2888	18.5626	2.5521	44.1283	15.9395	17.2071	86.1425
p-value	0.8633	0.0010	0.6353	0.0000	0.0031	0.0018	0.0000
ARCH Effects	No	Yes	No	Yes	Yes	Yes	Yes

Table 9: Univariate ARCH-LM test on the residuals of the VARIMA

Test	Statistic	p-value	H ₀	Conclusion (5% level)
Q(m) of squared (LM)	Inf	0	No ARCH effects	Reject H ₀
Rank-based	42.26207	6.74×10^{-6}	No ARCH effects	Reject H ₀

Table 10: Multivariate ARCH test results for VARIMA residuals

	DGS10	UMCSENT	GEPUCURRENT	INDPRO_US	USD.EUR	APE	BRENT
DGS10	0.8213	0.4885	0.7874	0.4438	0.0018	0.5333	0.0001
UMCSENT	0.4524	0.6874	0.5162	0.0656	0.8523	0.4142	0.0013
GEPUCURRENT	0.2470	0.0883	0.0948	0.9796	0.1388	0.7248	0.1012
INDPRO_US	0.1035	0.0000	0.0761	0.8948	0.1693	0.2490	0.0000
USD.EUR	0.6051	0.3427	0.6465	0.3661	0.8148	0.2679	0.2439
APE	0.4504	0.2859	0.8111	0.4105	0.6015	0.8633	0.4924
BRENT	0.0016	0.0006	0.3547	0.0361	0.1024	0.3929	0.0000

Table 11: P-value Matrix of the Teräsvirta Neural Network Test on VARIMA Residuals
H₀: Linearity $Y \sim X$ (rows = Y, columns = X)

Series	Test Statistic	p-value
DGS10	1.532	0.04752
UMCSENT	1.817	0.009194
GEPUCURRENT	1.197	0.2537
INDPRO_US	12.09	4.599×10^{-12}
USD.EUR	1.596	0.03344
APE	6.602	3.751×10^{-9}
BRENT	2.516	0.001697

Table 12: Univariate Tsay Nonlinearity Test Results for VARIMA Residual Series

Series	1 vs 2 Test	1 vs 2 P-value	1 vs 3 Test	1 vs 3 P-value	Interpretation
APE	25.60	0.003	36.07	0.005	Nonlinear, SETAR recommended
BRENT	10.42	0.015	23.18	0.025	Nonlinear, SETAR recommended
USD.EUR	22.80	0.645	51.86	0.520	Linear, sufficient
INDPRO_US	59.99	0.015	79.49	0.025	Nonlinear, SETAR recommended
GEPUCURRENT	19.11	0.675	46.43	0.630	Linear, sufficient
UMCSENT	32.33	0.001	62.64	0.002	Nonlinear, SETAR recommended
DGS10	24.04	0.001	17.09	0.001	Nonlinear, SETAR recommended

Table 13: Test of linearity against SETAR (2 and 3 regimes) for different VARIMA residual series.

5.6 Feature design, forecasting horizon, and time-series protocol

This subsection makes explicit the forecasting horizon, the exact inputs to each model class, and the safeguards against information leakage. We focus on one-step-ahead forecasting, which is the fundamental unit of rolling evaluation and avoids compounding errors in turbulent regimes. All models are trained and evaluated under a strict chronological protocol with rolling windows.

Inputs (pure ML). For pure ML models, the feature vector at time t is constructed from lagged multivariate returns. Consistent with the short-memory structure suggested by the VARIMA selection and to preserve comparability across model classes, we use lag-one predictors: $\mathbf{X}_t = \mathbf{Y}_{t-1}$. This design choice is conservative: it prevents ML models from benefiting from a richer lag structure than the linear benchmark, and therefore makes performance comparisons more interpretable.

Inputs (hybrid residual learning). For hybrid models, the nonlinear learner is trained on the VARIMA residuals. Specifically, letting $\boldsymbol{\varepsilon}_t$ denote the VARIMA residual vector, we use $\mathbf{Z}_t = \boldsymbol{\varepsilon}_{t-1}$ as inputs to the ML correction. This operationalizes the division of labor emphasized in Section 4: VARIMA captures linear propagation, and ML captures the predictable nonlinear component embedded in residual dynamics. Importantly, residual features are orthogonal to the linear fit by construction, which reduces the risk that ML simply re-learns linear dependencies and improves robustness in finite samples.

Normalization. All variables are standardized by series using training-window statistics and then applied to validation/test windows. This prevents scale dominance and avoids leakage.

Hyperparameter selection. Hyperparameters are selected via rolling cross-validation on the in-sample period, and the selected configurations are reported in Table 15. This is essential in time-series forecasting: the aim is not to minimize in-sample error but to control complexity so that models generalize under regime instability.

5.7 Forecasting performance and interpretation of hybrid gains

All quantitative forecasting results are reported in Table 14, where models are evaluated in a fully multivariate setting and performance metrics (MAE, MSE, RMSE) are aggregated across all series. The ranking reveals three robust patterns.

(i) Linear benchmark inadequacy. VARIMA is dominated by most nonlinear competitors (Table 14). This is consistent with the diagnostic evidence: residual autocorrelation (Table 7), heavy tails (Table 8), ARCH effects (Tables 9-10), and nonlinear dependence (Tables 11-13) imply that the linear conditional-mean benchmark leaves exploitable structure in forecast errors.

(ii) Pure ML improves responsiveness but may suffer from variance. Pure ML models partially alleviate VARIMA’s lag and under-reaction during turbulence, but they do not dominate overall. This is visible in the figures referenced below and is consistent with the fact that purely data-driven learning on raw returns can be variance-prone under moderate samples and regime instability: models may overreact to idiosyncratic shocks, producing noisy forecast paths and degraded aggregate accuracy.

(iii) Hybrid residual learning yields the best bias-variance compromise. Hybrid VARIMA-ML models systematically dominate both VARIMA and their pure ML counterparts (Table 14), with the notable exception of VARIMA-LSTM. This dominance is structural: the VARIMA component removes a large share of predictable linear dynamics (reducing bias relative to naive models), while the residual learner targets nonlinear corrections evidenced by diagnostics (reducing the remaining bias) without paying the full variance cost of learning on raw data.

Model	MAE	MSE	RMSE
VARIMA-GB	7.418	649.308	25.481
LSTM	7.504	688.057	26.230
VARIMA-SVR	7.576	670.207	25.888
VARIMA-RF	7.585	676.467	26.008
VARIMA-MLP	7.590	677.678	26.032
VARIMA-GRU	7.603	696.635	26.393
VARIMA-TCN	7.608	659.368	25.678
VARIMA-Tree	7.646	650.964	25.514
VARIMA-LSTM	7.750	682.605	26.126
GRU	8.413	890.932	29.848
TCN	8.463	838.582	28.958
MLP	9.153	673.078	25.943
Tree	9.382	1334.864	36.535
GB	9.620	1076.030	32.802
VARIMA	9.957	1178.000	34.322
SVR	10.515	1866.528	43.203
RF	10.887	1626.592	40.331

Table 14: Model performance comparison.

To facilitate interpretation beyond aggregate numbers, Figures [A1](#), [A2](#), [A3](#), [A4](#), [A5](#), [A6](#), [A7](#), and [A8](#) complement Table 14 by illustrating realized and predicted trajectories generated by the VARIMA benchmark, pure ML models, and hybrid VARIMA-ML specifications. A robust visual regularity emerges: VARIMA forecasts tend to under-react and lag during turning points in turbulent periods, while hybrid forecasts improve responsiveness without generating the high-frequency oscillations sometimes produced by pure ML. This is precisely the empirical signature of a bias-variance improvement through residual learning.

Why boosting dominates deep hybrids. The top-ranked model in Table 14 is VARIMA-GB. This ranking is economically and statistically coherent. After VARIMA filtering, the residual learning task is a *correction problem* dominated by short-run nonlinearities and threshold effects (as suggested by Teräsvirta/Tsay/SETAR results). Gradient boosting is particularly effective in this regime because it implements strong regularization through shrinkage, shallow trees, subsampling, and early stopping, yielding a low-variance nonlinear approximator. By contrast, recurrent deep architectures (LSTM/GRU) are parameter-intensive and typically require large effective sample sizes and stable regimes to outperform simpler learners. When the remaining signal is short-memory and regime-dependent rather than long-memory, deep recurrence adds variance without commensurate bias reduction, explaining the weaker VARIMA-LSTM results.

Overall, the combined evidence from Tables 7-13, Table 14, and Figures [A1-A8](#) supports a coherent structural conclusion: VARIMA captures linear propagation and cross-variable transmission channels, while the ML correction captures adaptive nonlinear adjustments that become particularly relevant under stress. The gains are not driven by isolated series; they emerge at the system level under a turbulent out-of-sample period, which is precisely the setting in which robustness and interpretability are most valuable.

Model	Hyperparameters	Value	Description
GB	learning rate, max depth, n estimators, subsample	0.05, 3, 500, 0.8	Gradient boosting complexity control (shrinkage + shallow trees)
RF	n estimators, max features, min samples leaf	800, $\sqrt{p}$, 2	Ensemble size and randomness (stability + de-correlation)
SVR	C , ε , kernel (γ)	8, 0.05, RBF ($\gamma = \text{scale}$)	Margin and kernel flexibility (regularized nonlinear regression)
MLP	hidden layers, neurons, activation, optimizer, batch size	[64, 32], ReLU, Adam, 32	Network capacity with standard training protocol
GRU	hidden units, dropout, sequence length, epochs	32, 0.2, 12, 100	Sequential memory depth (moderate capacity)
LSTM	hidden units, dropout, sequence length, epochs	32, 0.2, 12, 100	Sequential memory depth (comparable capacity)
TCN	filters, kernel size, dilation levels, dropout	32, 3, [1, 2, 4, 8], 0.2	Temporal receptive field (dilated causal convolutions)
Hybrid (VARIMA-GB)	learning rate, max depth, n estimators, residual features	0.05, 3, 500, 0.8, ε_{t-1}	VARIMA residual learning with boosting
Hybrid (VARIMA-SVR)	C , ε , kernel (γ), residual features	8, 0.05, RBF ($\gamma = \text{scale}$), ε_{t-1}	VARIMA residual learning with SVR
Hybrid (VARIMA-MLP)	hidden layers, neurons, activation, residual features	[64, 32], ReLU, ε_{t-1}	VARIMA residual learning with MLP
Hybrid (VARIMA-RF)	n estimators, max features, min samples leaf, residual features	800, $\sqrt{p}$, 2, ε_{t-1}	VARIMA residual learning with Random Forest
Hybrid (VARIMA-LSTM)	hidden units, dropout, sequence length, residual features	32, 0.2, 12, ε_{t-1}	VARIMA residual learning with LSTM
Hybrid (VARIMA-GRU)	hidden units, dropout, sequence length, residual features	32, 0.2, 12, ε_{t-1}	VARIMA residual learning with GRU
Hybrid (VARIMA-TCN)	filters, kernel size, dilation levels, dropout, residual features	32, 3, [1, 2, 4, 8], 0.2, ε_{t-1}	VARIMA residual learning with TCN

Table 15: Machine learning and hybrid model hyperparameter protocol. Numerical values are selected via rolling cross-validation.

6 Discussions and Conclusion

6.1 Discussion

This section goes beyond average performance and explains the mechanisms behind the superiority of hybrid VARIMA–machine learning models. We organize the discussion around three questions explicitly raised by the reviewers: *when* hybridization delivers the largest gains, *why* it does so (statistical and economic mechanisms), and *which variables and channels* drive the improvement. The overarching message is that the hybrid architecture is not merely a performance trick: it implements an economically meaningful division of labor in which VARIMA captures linear shock propagation while ML provides adaptive nonlinear corrections that become particularly valuable during stress regimes.

(i) When does the hybrid model win? Stress regimes and tail-risk episodes. A robust empirical regularity emerging from the out-of-sample evaluation is that hybrid models deliver the largest improvements precisely in periods characterized by abrupt regime changes and volatility bursts. This is visible in the distributional evidence reported in Figure 7: relative to pure ML, hybrid models display both lower central error and markedly thinner tails across MAE, MSE, and RMSE. In forecasting applications, this distinction is central: outliers and tail losses dominate economic costs in risk management and in policy-sensitive energy settings.

The diagnostic evidence provided earlier in the paper directly rationalizes this finding. Residual autocorrelation (Table 7), strong departures from multivariate normality (Table 8), and pervasive conditional heteroskedasticity (Tables 9–10) imply that the linear benchmark leaves structured residual components that are amplified during turbulent regimes. Moreover, formal nonlinearity tests (Tables 11, 12, and 13) show that the residual dynamics are inconsistent with a purely linear conditional-mean specification. In economic terms, these residual nonlinearities are consistent with state-dependent adjustment, asymmetries, and threshold effects that become binding precisely when uncertainty rises and energy prices respond nonlinearly to macro-financial conditions.

Importantly, the out-of-sample window was chosen to be economically demanding: it is characterized by persistent uncertainty and macro-energy dislocations. In this environment, the performance advantage of the hybrid model is not an artifact of isolated episodes: it is the systematic consequence of a model design that explicitly accommodates nonlinear corrections when regimes shift.

(ii) Why does the hybrid model win? A structural bias–variance mechanism. From a statistical standpoint, the hybrid gains are best interpreted through a bias–variance lens. Pure VARIMA models are disciplined but biased when the conditional mean is state-dependent or non-linear. Pure ML models are flexible but can become high-variance learners in moderate samples with strong cross-dependencies and structural breaks. The hybrid model mitigates both failures through residual learning.

The first-stage VARIMA acts as a multivariate *linear propagation filter* that absorbs the dominant cross-lag and cross-variable dependence in the conditional mean. This reduces the complexity of the remaining learning problem. The second-stage learner is then trained on residual structure which (i) is empirically documented to be non-white-noise and (ii) concentrates precisely the nonlinear corrections that VARIMA cannot encode by construction. This reduces estimation variance relative to training ML directly on raw returns, because the nonlinear learner no longer needs to rediscover linear dynamics and cross-correlations.

The distributional evidence in Figure 7 makes this mechanism tangible: hybridization reduces dispersion and outliers, indicating that it improves not only average accuracy but also robustness. The Welch two-sample tests in Table 16 corroborate that these gains are statistically significant across MAE, MSE, and RMSE, even under unequal variances, which is appropriate given the heteroskedastic nature of forecast errors.

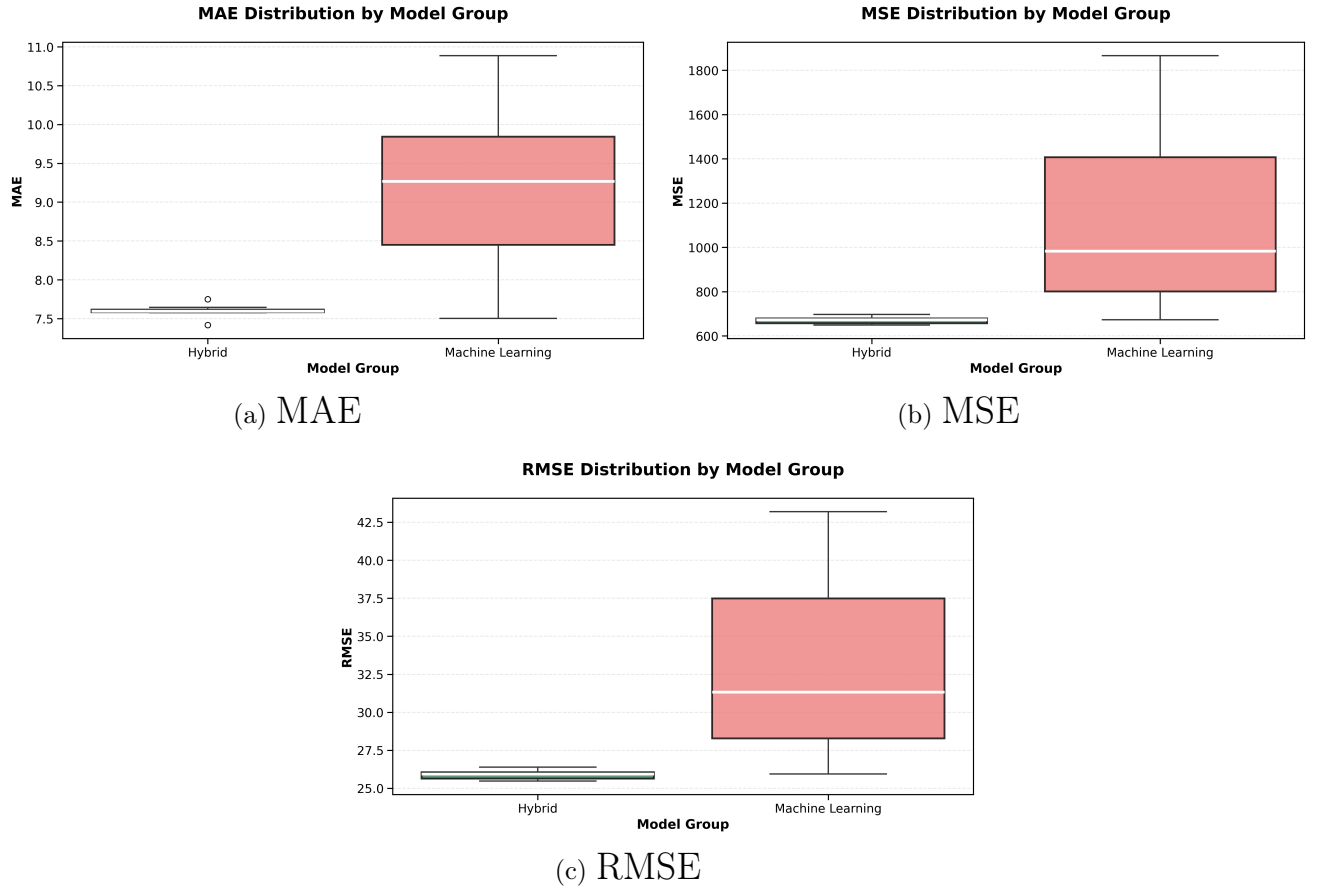


Figure 7: Boxplots of forecasting error distributions for the hybrid VARIMA–machine learning model and the machine learning model

Metric	t-statistic	p-value
MAE	-4.13	0.0043
MSE	-2.89	0.0233
RMSE	-3.10	0.0172

Table 16: Welch two-sample t-test comparing hybrid and machine learning models

This reduction in dispersion is economically meaningful. In macroeconomic and energy-related environments characterized by recurrent shocks, regime changes, and abrupt volatility bursts, forecast robustness is at least as important as average accuracy. The boxplots show that pure machine learning models are more prone to extreme forecast errors, as reflected in wider tails and more frequent outliers. By contrast, hybrid models effectively dampen these extreme deviations, suggesting a superior ability to generalize under stress conditions. This property is particularly relevant in the out-of-sample period, which encompasses episodes of heightened uncertainty and market instability.

The visual evidence from Figure 7 is formally corroborated by the Welch two-sample t -tests reported in Table 16. For all three error measures, the null hypothesis of equal predictive accuracy between hybrid and pure machine learning models is rejected at the 5% significance level. The negative t -statistics indicate that hybrid models deliver significantly lower forecasting errors on average. Importantly, the use of Welch’s test accounts for potential heteroskedasticity and unequal variances between error distributions, ensuring that the statistical inference remains valid despite the non-Gaussian features documented earlier in the paper.

From an econometric standpoint, these results confirm that the observed performance gains are not driven by isolated episodes or favorable realizations but reflect a systematic improvement in predictive accuracy. The simultaneous reduction in MAE, MSE, and RMSE implies that hybrid models outperform their competitors both in terms of average forecast precision and in their ability to control large deviations. This dual improvement is rarely achieved by purely data-driven models, which often face a trade-off between flexibility and stability.

The economic interpretation of these findings is closely linked to the structure of the hybrid framework. The VARIMA component captures linear cross-dependencies, common macro-financial dynamics, and short-run propagation effects across the multivariate system. This linear filtering step removes a substantial portion of predictable variation and isolates residual components that primarily reflect nonlinear adjustments, regime-dependent behavior, and higher-order interactions. Machine learning models are then applied to these residuals, where they can focus exclusively on nonlinear structures without being confounded by strong linear dependence.

This division of labor explains why hybrid models consistently dominate pure machine learning approaches. When machine learning models are trained directly on raw or weakly filtered data, they must simultaneously learn linear dynamics, cross-correlations, and nonlinear patterns, which increases estimation noise and amplifies overfitting risks. In contrast, residual-based learning in the hybrid framework reduces dimensional complexity and improves signal-to-noise ratios, leading to more stable out-of-sample performance.

6.2 Concluding remarks

The results of this study carry several important implications for the literature on multivariate macro-energy forecasting. First, they demonstrate that the limitations of linear econometric models are not confined to volatility misspecification or distributional assumptions but extend to systematic nonlinear dependence structures in the conditional mean. Second, they show that purely machine learning approaches, while flexible, may lack robustness when applied directly to complex multivariate economic systems characterized by strong cross-dependencies and structural breaks.

The main scientific contribution of this paper lies in providing a transparent and empirically validated framework that reconciles these two modeling traditions. By combining VARIMA models with machine learning algorithms in a residual-based hybrid architecture, the proposed approach achieves superior predictive accuracy while preserving economic interpretability. The linear VARIMA component retains its role as a structural benchmark that reflects economically meaningful interactions, while the machine learning component enhances performance by capturing nonlinear corrections that are empirically present but difficult to model parametrically.

Beyond forecasting accuracy, the hybrid framework offers a form of interpretability that is particularly valuable in multivariate settings. Because nonlinear learning is applied to residuals rather than raw data, the contribution of machine learning can be interpreted as an adjustment to a well-defined linear baseline. This structure facilitates economic reasoning about when and why nonlinear effects become relevant, for instance during periods of heightened uncertainty or abrupt regime shifts.

Several limitations should nevertheless be acknowledged. First, the hybrid framework remains computationally more demanding than purely linear models, particularly when multiple machine learning architectures are considered. Second, while residual-based learning improves robustness, it does not fully eliminate the sensitivity of machine learning models to hyperparameter choices. Third, the analysis focuses on one-step-ahead forecasts; extending the framework to multi-horizon forecasting may require additional structural constraints to avoid error accumulation.

These limitations naturally suggest directions for future research. Potential extensions include the integration of time-varying parameter VARIMA filters, the incorporation of distributional forecasting objectives, and the use of explainable machine learning tools to further dissect the nonlinear

components learned by hybrid models. Exploring alternative residual representations and regime-dependent hybridizations may also provide additional gains in highly unstable macroeconomic environments.

In conclusion, this paper provides strong statistical and economic evidence that hybrid VARIMA-machine learning models constitute a powerful and interpretable tool for multivariate macro-energy forecasting. By systematically combining structural econometric modeling with adaptive nonlinear learning, the proposed framework delivers robust improvements in predictive accuracy and stability, addressing key challenges at the intersection of econometrics and machine learning. These results support the broader view that hybrid approaches represent a promising and methodologically sound direction for future research in applied macro-finance and energy economics, particularly in environments characterized by regime shifts, heavy tails, and time-varying uncertainty. Future work can naturally extend this framework toward multi-horizon and distributional forecasting (densities and quantiles), explicit structural break detection, and explainable AI-based decomposition of nonlinear correction channels to further strengthen economic attribution and policy relevance.

References

- H. Akaike. Canonical correlation analysis of time series and the use of an information criterion. In *Mathematics in science and engineering*, volume 126, pages 27–96. Elsevier, 1976.
- R. Askari Moghadam and M. Keshmirpour. Hybrid arima and neural network model for measurement estimation in energy-efficient wireless sensor networks. In *International conference on informatics engineering and information science*, pages 35–48. Springer, 2011.
- G. Athanasopoulos and F. Vahid. A complete varma modelling methodology based on scalar components. *Journal of Time Series Analysis*, 29(3):533–554, 2008a.
- G. Athanasopoulos and F. Vahid. Varma versus var for macroeconomic forecasting. *Journal of Business & Economic Statistics*, 26(2):237–252, 2008b.
- E. M. Azoff. *Neural network time series forecasting of financial markets*. John Wiley & Sons, Inc., 1994.
- S. Bai. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.
- T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327, 1986.
- G. Box and G. Jenkins. *Analysis: Forecasting and control*. San francisco, 1976.
- G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.
- L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Chapman and Hall/CRC, 2017.
- P. J. Brockwell and R. A. Davis. *Time series: theory and methods*. Springer science & business media, 2009.
- K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative finance*, 1(2):223, 2001.

-
- X. Ding, Y. Zhang, T. Liu, and J. Duan. Deep learning for event-driven stock prediction. In *Ijcai*, volume 15, pages 2327–2333, 2015.
- W. Dunsmuir and E. J. Hannan. Vector linear time series models. *Advances in Applied Probability*, 8(2):339–364, 1976.
- R. Engle. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of business & economic statistics*, 20(3):339–350, 2002.
- R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the econometric society*, pages 987–1007, 1982.
- J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- B. Goehry. Random forests for time-dependent processes. *ESAIM: Probability and Statistics*, 24: 801–826, 2020.
- B. Goehry, H. Yan, Y. Goude, P. Massart, and J.-M. Poggi. Random forests for time series. *REVSTAT-Statistical Journal*, 21(2):283–302, 2023.
- A. Goswami. Understanding Transformers: A straightforward breakdown of “Attention is All You Need”. <https://medium.com/data-science/understanding-transformers-3344d16c8c36>, 2020. Consulté le 6 février 2026. Blog Medium.
- C. W. Granger. Strategies for modelling nonlinear time-series relationships. *Economic Record*, 69 (3):233–238, 1993.
- E. Hannan. The identification of vector mixed autoregressive-moving average system. *Biometrika*, 56(1):223–225, 1969.
- E. J. Hannan. The statistical theory of linear systems. In *Developments in Statistics*, volume 2, pages 83–121. Elsevier, 1979.
- K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- R. Hyndman, A. Koehler, K. Ord, and R. Snyder. *Forecasting with exponential smoothing: the state space approach*. Springer, 2008.
- L. Kilian. New introduction to multiple time series analysis, by helmut lütkepohl, springer, 2005. *Econometric theory*, 22(5):961–967, 2006.
- L. Kilian and H. Lütkepohl. *Structural vector autoregressive analysis*. Cambridge University Press, 2017.
- Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.
- K. Lee, J. Ray, and C. Safta. The predictive skill of convolutional neural networks models for disease forecasting. *Plos one*, 16(7):e0254319, 2021.
- Q. Lin. Price prediction of bitcoin using lstm neural network. *BCP Business & Management*, 38: 2738–2744, 2023.
- H. Lütkepohl. *Introduction to multiple time series analysis*. Springer Science & Business Media, 2013.
-

-
- S. Makridakis, E. Spiliotis, and V. Assimakopoulos. Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one*, 13(3):e0194889, 2018.
- S. Mukherjee, E. Osuna, and F. Girosi. Nonlinear prediction of chaotic time series using support vector machines. In *Neural Networks for Signal Processing VII. Proceedings of the 1997 IEEE Signal Processing Society Workshop*, pages 511–520. IEEE, 1997.
- K.-R. Müller, A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik. Predicting time series with support vector machines. In *International conference on artificial neural networks*, pages 999–1004. Springer, 1997.
- A. Navon and Y. Keller. Financial time series prediction using deep learning. *arXiv preprint arXiv:1711.04174*, 2017.
- C. Ordóñez, F. S. Lasheras, J. Roca-Pardinas, and F. J. de Cos Juez. A hybrid arima–svm model for the study of the remaining useful life of aircraft engines. *Journal of Computational and Applied Mathematics*, 346:184–191, 2019.
- P.-F. Pai and C.-S. Lin. A hybrid arima and support vector machines model in stock price forecasting. *Omega*, 33(6):497–505, 2005.
- F. Pokou, J. Sadefo Kamdem, and F. Benhmad. Hybridization of arima with learning models for forecasting of stock market time series. *Computational Economics*, 63(4):1349–1399, 2024.
- F. Pokou, J. S. Kamdem, and K. E. Gnandi. Predictive accuracy versus interpretability in energy markets: A copula-enhanced tvp-svar analysis. *arXiv preprint arXiv:2601.19321*, 2026.
- J. Rahman, A. Raihan, T. Tanchangya, and M. Ridwan. Optimizing the digital marketing landscape: A comprehensive exploration of artificial intelligence (ai) technologies, applications, advantages, and challenges. *Frontiers of finance*, 2(2):6549, 2024.
- G. C. Reinsel. *Elements of multivariate time series analysis*. Springer Science & Business Media, 2003.
- F. Rosenblatt. *The perceptron, a perceiving and recognizing automaton:(Project Para)*. Cornell Aeronautical Laboratory, 1957.
- R. J. Shiller. The use of volatility measures in assessing market efficiency, 1980.
- C. Sims. *Macroeconomics and reality*’econometrica, vol. 48. 1980.
- A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.
- R. S. Tsay. *Multivariate time series analysis: with R and financial applications*. John Wiley & Sons, 2013.
- V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.
- A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang. Phoneme recognition using time-delay neural networks. In *Backpropagation*, pages 35–61. Psychology Press, 2013.
- M. Welling. Support vector regression. *Department of Computer Science, University of Toronto, Toronto (Canada)*, 2004.
- D. Xu, Q. Zhang, Y. Ding, and D. Zhang. Application of a hybrid arima-lstm model based on the spei for drought forecasting. *Environmental Science and Pollution Research*, 29(3):4128–4144, 2022.
- B. Yu, H. Yin, and Z. Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.
-

-
- F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.
- G. P. Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.

A Appendix 1: Forecasting Plots

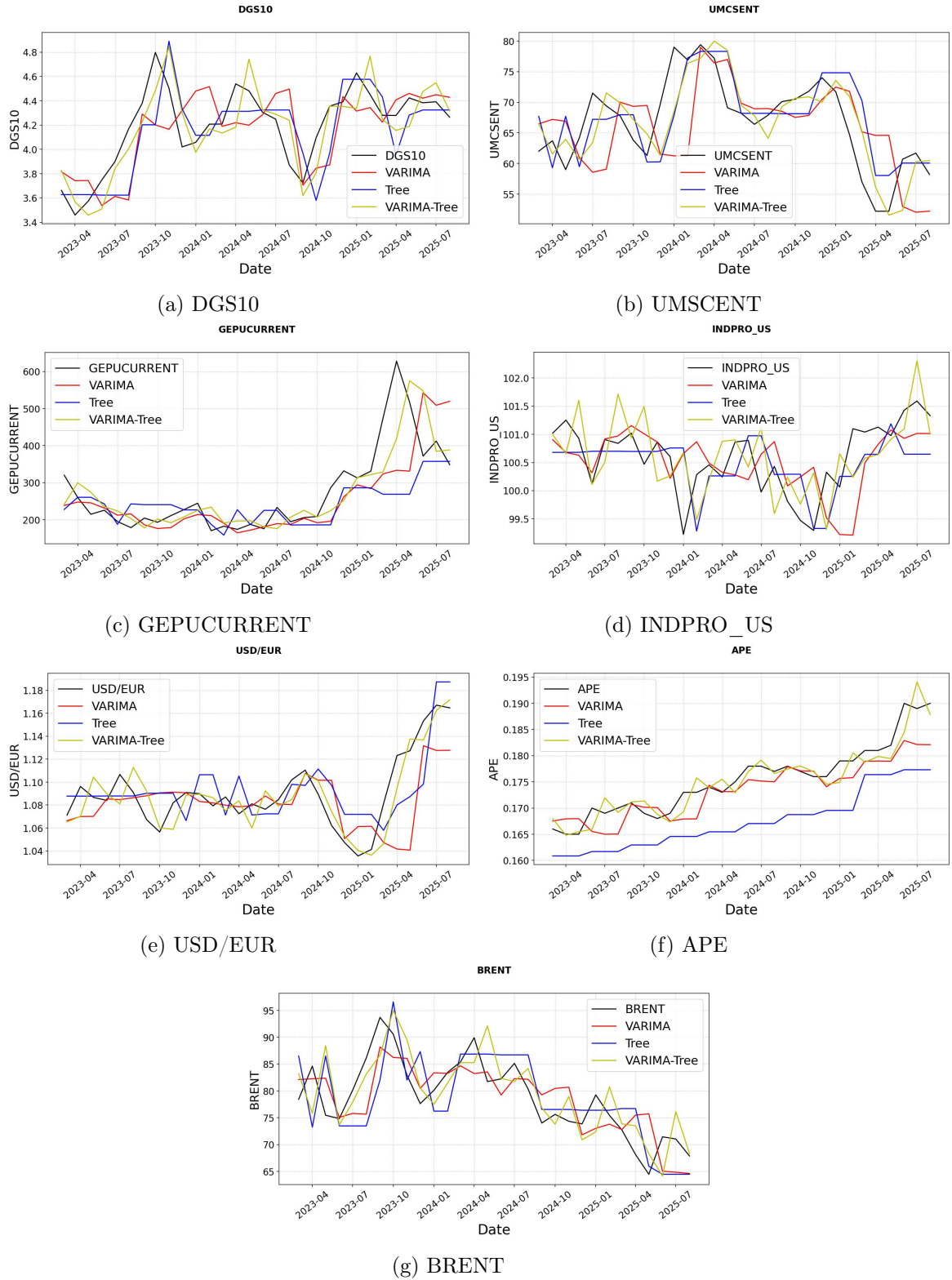
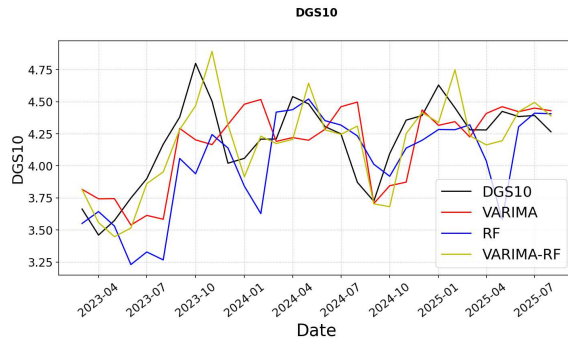
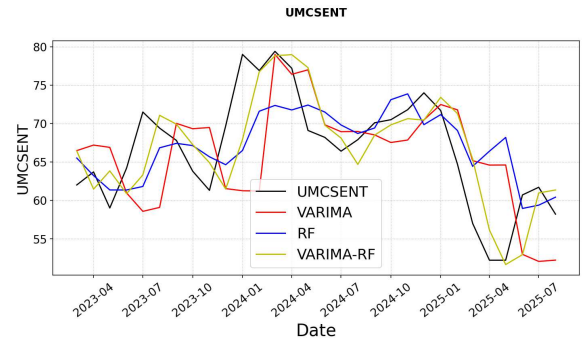


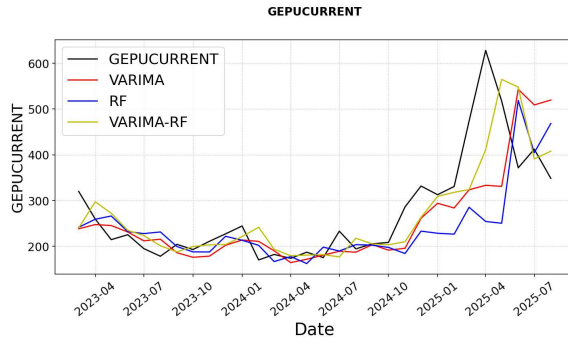
Figure A1: Multivariate time series forecasting via VARIMA, pure TREE method, and hybrid strategy (VARIMA-TREE).



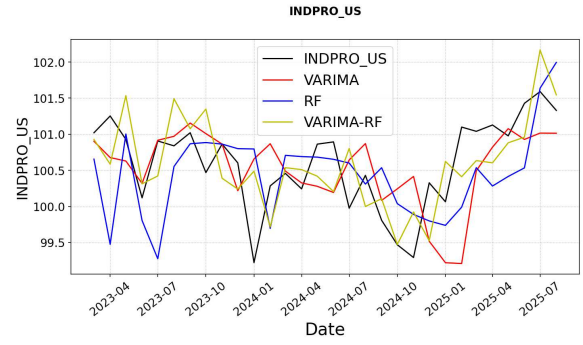
(a) DGS10



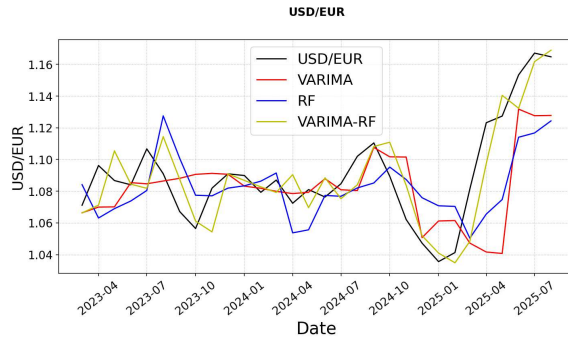
(b) UMCSENT



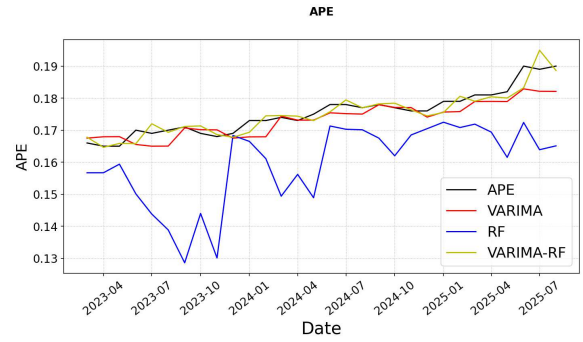
(c) GEPUCURRENT



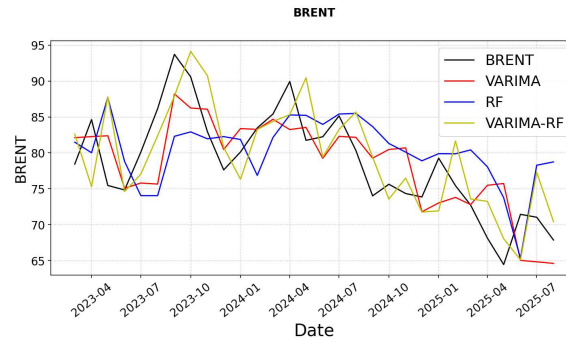
(d) INDPRO_US



(e) USD/EUR

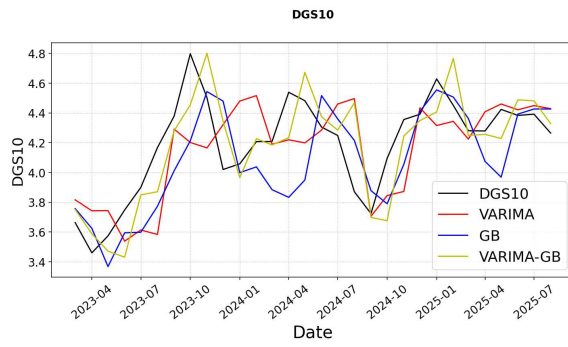


(f) APE

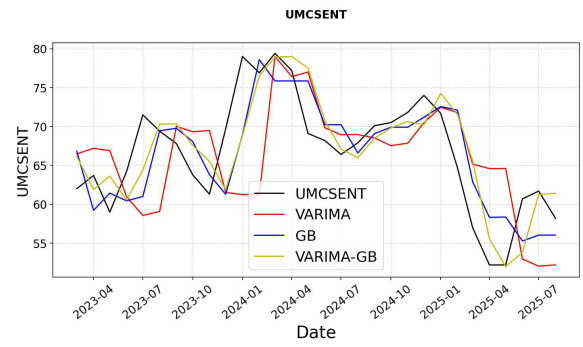


(g) BRENT

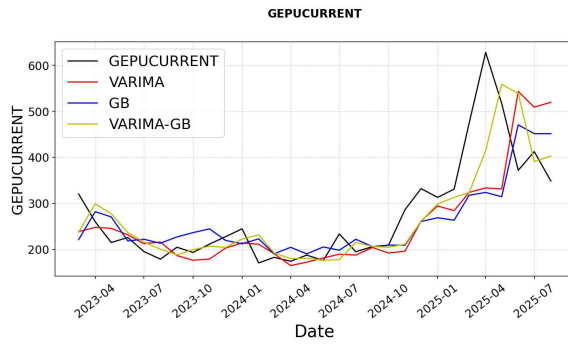
Figure A2: Multivariate time series forecasting via VARIMA , Pure Random Forest method and Hybrid strategie(VARIMA-RF)



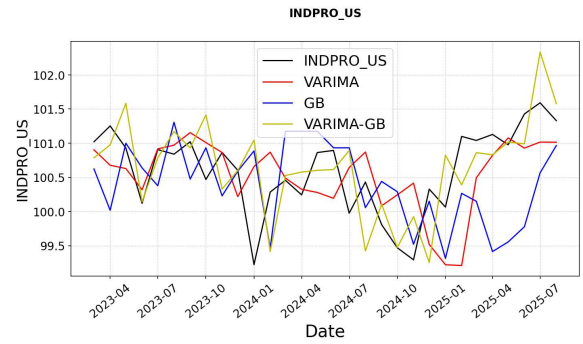
(a) DGS10



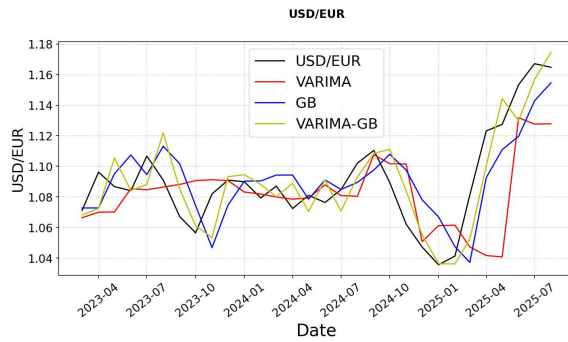
(b) UMSCENT



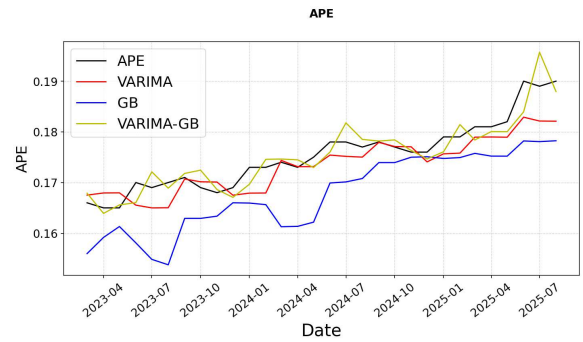
(c) GEPUCURRENT



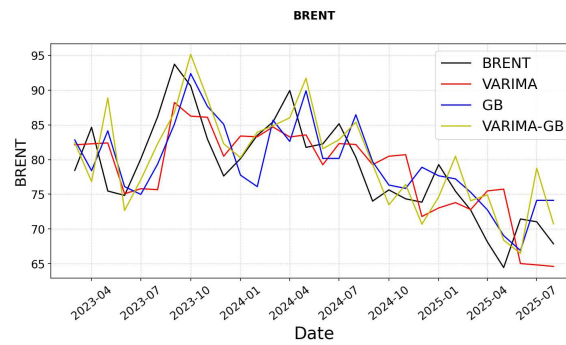
(d) INDPRO_US



(e) USD/EUR



(f) APE



(g) BRENT

Figure A3: Multivariate time series forecasting via VARIMA , Pure Gradient Boosting method and Hybrid strategie(VARIMA-GB)

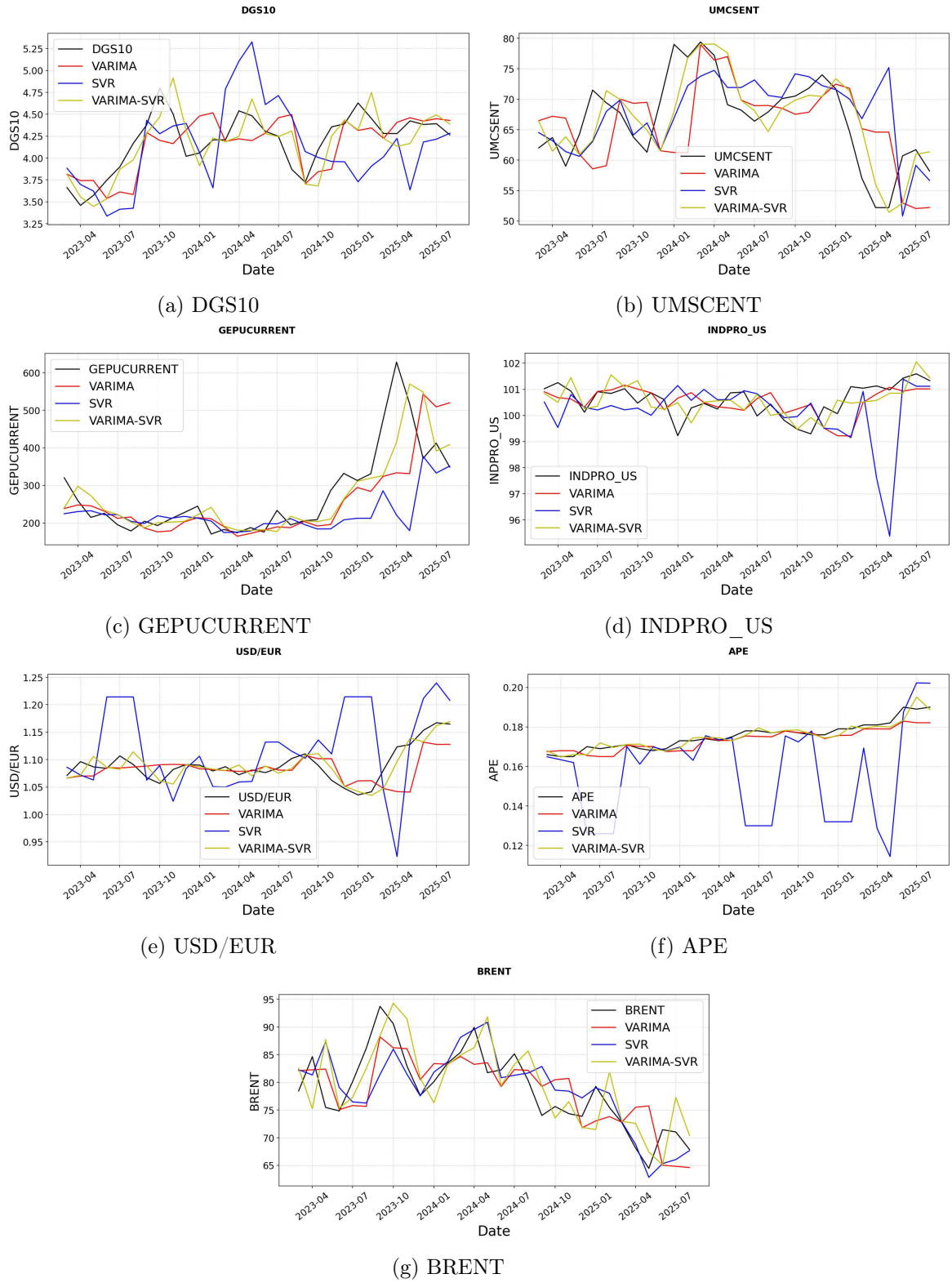
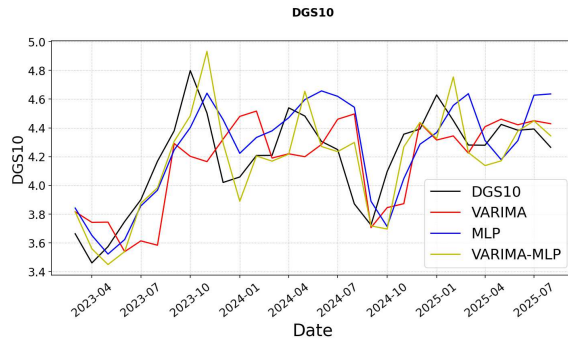
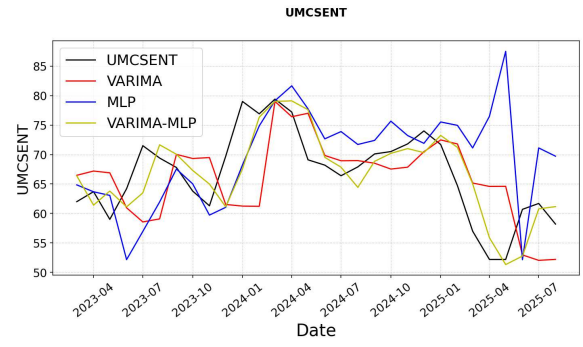


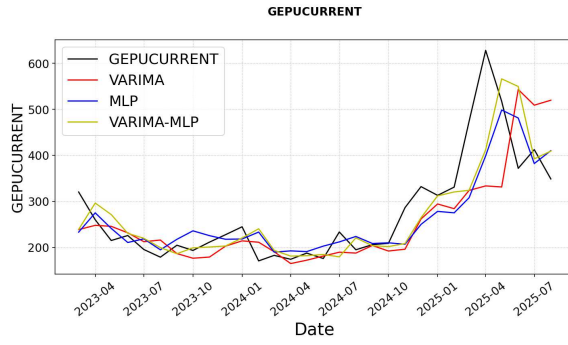
Figure A4: Multivariate time series forecasting via VARIMA , Pure Support Vector Regression method and Hybrid strategie(VARIMA-SVR)



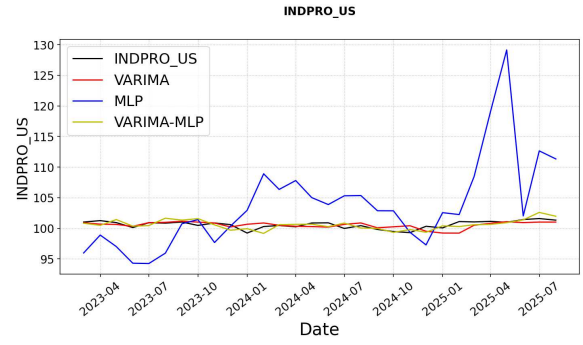
(a) DGS10



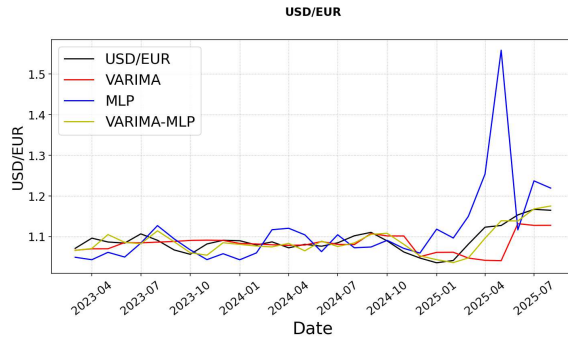
(b) UMCSSENT



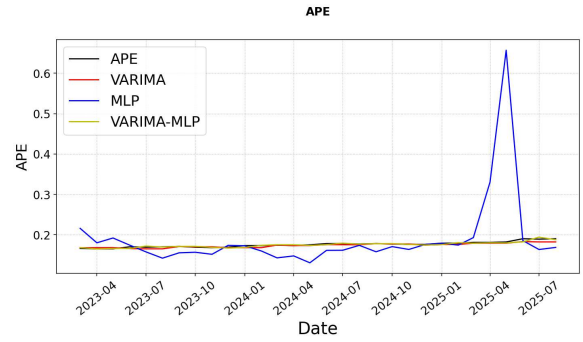
(c) GEPUCURRENT



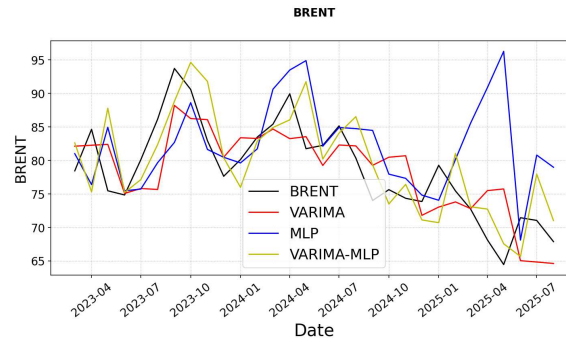
(d) INDPRO_US



(e) USD/EUR

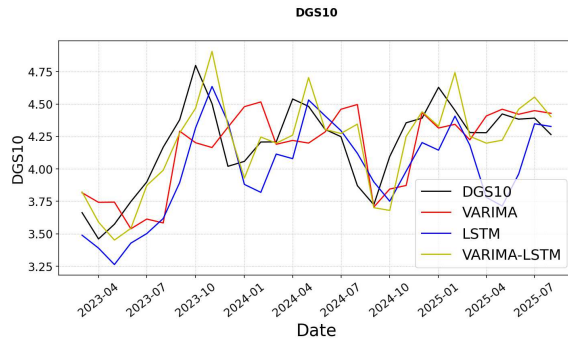


(f) APE

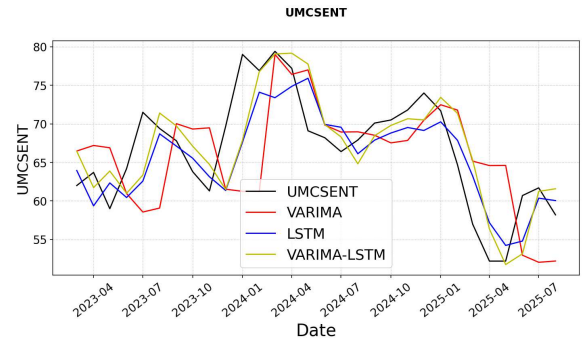


(g) BRENT

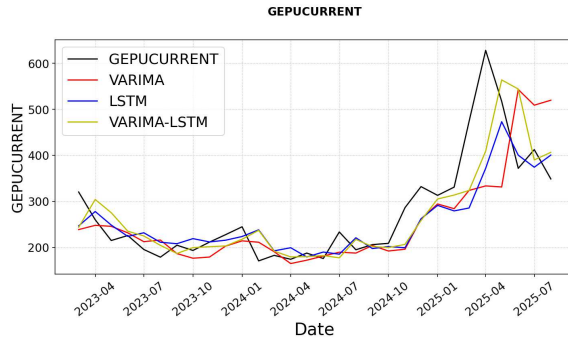
Figure A5: Multivariate time series forecasting via VARIMA , Pure Multi-Layer perceptron method and Hybrid strategy(VARIMA-MLP)



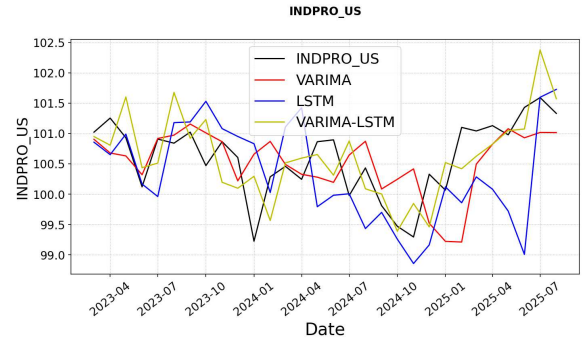
(a) DGS10



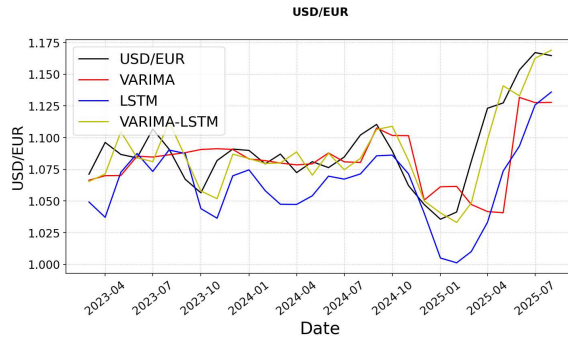
(b) UMSCENT



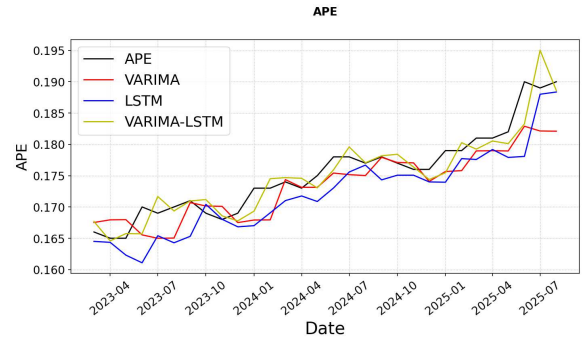
(c) GEPUCURRENT



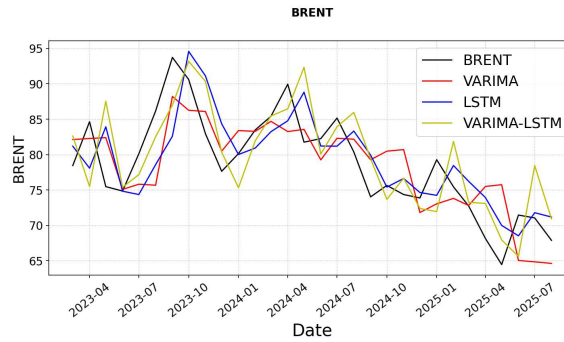
(d) INDPRO_US



(e) USD/EUR



(f) APE



(g) BRENT

Figure A6: Multivariate time series forecasting via VARIMA , Pure LSTM method and Hybrid strategy(VARIMA-LSTM)

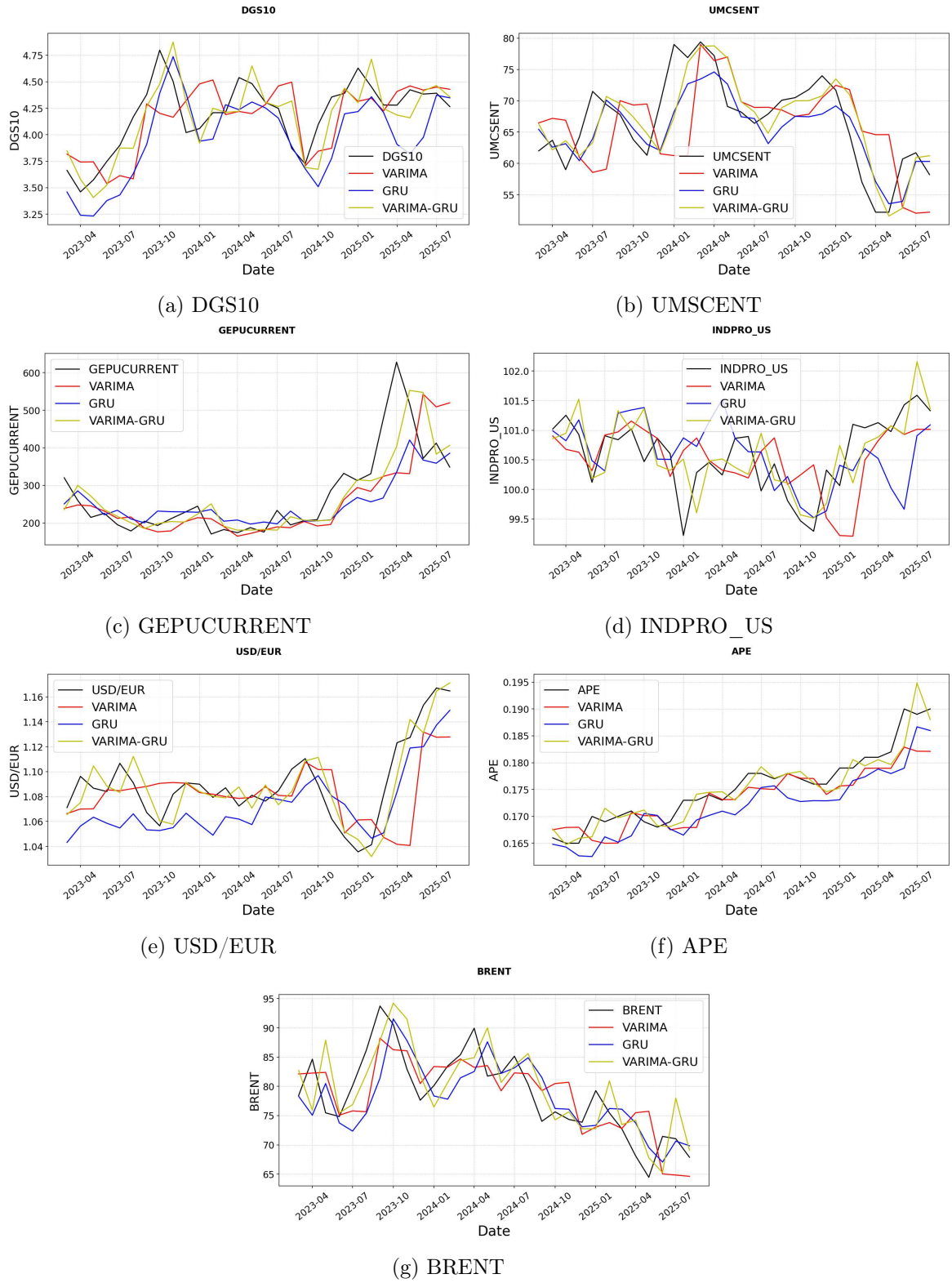


Figure A7: Multivariate time series forecasting via VARIMA , Pure GRU method and Hybrid strategie(VARIMA-GRU)

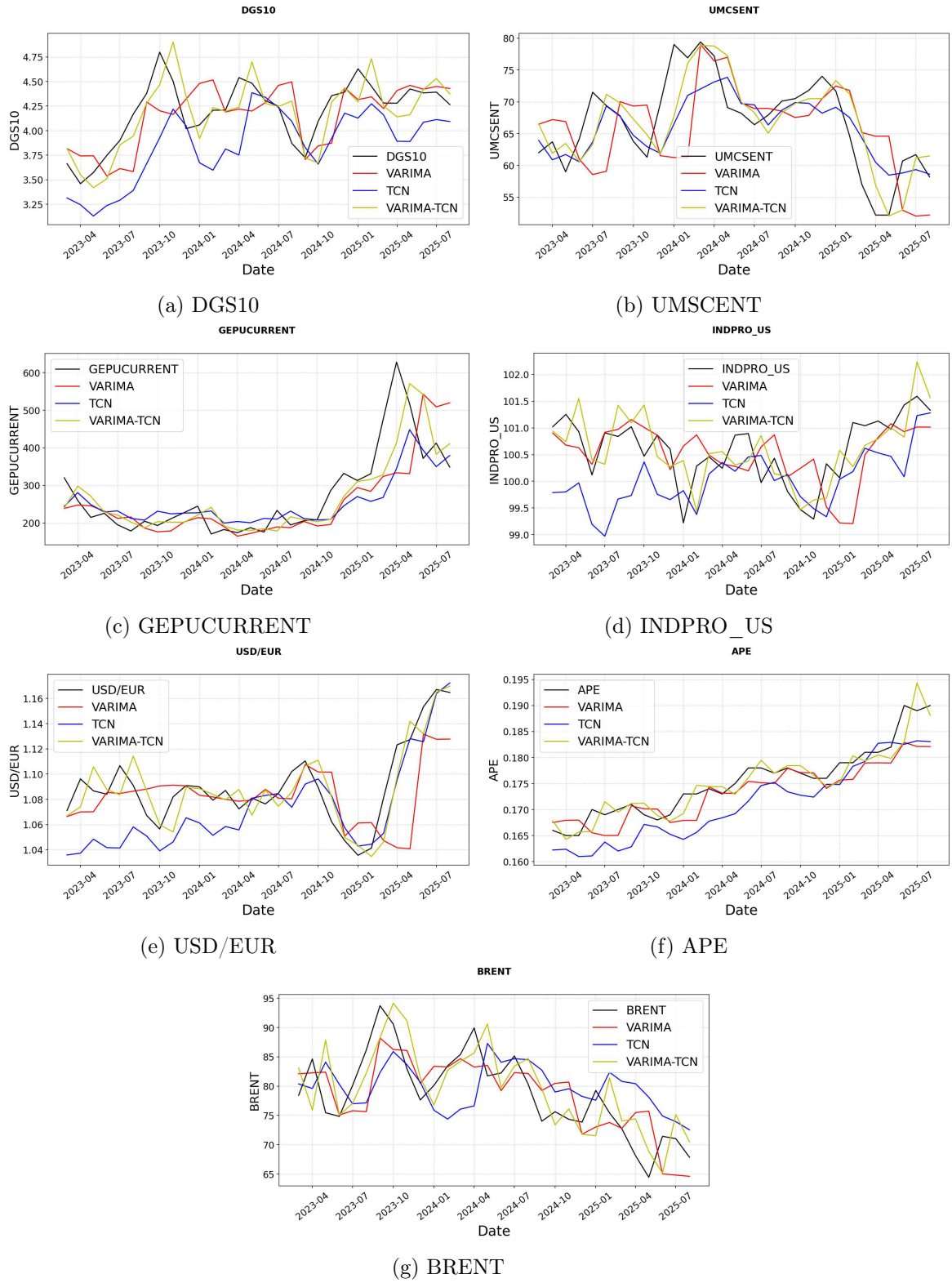


Figure A8: Multivariate time series forecasting via VARIMA , Pure TCN method and Hybrid strategie(VARIMA-TCN)