Benchmarking Statistical, Machine Learning, Deep Learning, and Hybrid Forecasting Models for Global Renewable Energy Consumption: A Walk-Forward Cross-Validation Study with Structural Break Analysis.

Abstract

Energy independence and resilience have become critical policy priorities as geopolitical tensions, supply disruptions, and price volatility expose the vulnerability of fossil-fuel-dependent energy systems. Accurate forecasting of renewable energy consumption is therefore essential for effective energy transition planning, infrastructure investment, and monitoring progress toward international climate targets such as Sustainable Development Goal 7 (SDG-7). In macro-energy policy practice, quantitative forecasts underpin scenario design, capacity planning, and assessment of alignment with net-zero pathways, yet the annual frequency and short length of globally comparable time series severely constrain the effective application of data-intensive forecasting methods. This study benchmarks 13 forecasting model families—spanning baselines (Naïve, Random Walk with Drift, Linear Trend), classical statistical methods (ETS, Damped ETS, Theta, ARIMA), machine learning (XGBoost), deep learning (GRU, LSTM, N-BEATS), an additive model (Prophet), and a novel ETS–GRU hybrid—against the World Bank EG.FEC.RNEW.ZS indicator (1990–2020). All models are evaluated under a unified 5-window expanding walk-forward cross-validation protocol with a 3-year forecast horizon, nested hyperparameter tuning, multi-seed deep learning robustness checks, Diebold–Mariano tests, Model Confidence Set analysis, and bootstrap inference. A Chow structural break test detects a statistically significant regime shift at 2014 ($F = 32.0$, $p < 0.001$), coinciding with the post-Paris Agreement acceleration in renewable deployment. Results indicate that Holt Linear Exponential Smoothing (ETS) achieves the lowest RMSE (0.543) with a Skill Score of +0.148 against the Naïve baseline, outperforming all deep learning architectures. The study introduces three novel energy transition analytics—a Transition Velocity Index, regional beta-convergence analysis, and SDG-7 gap assessment across 11 World Bank regions—and demonstrates that under a methodologically symmetric evaluation protocol that eliminates the information asymmetries present in prior benchmarking studies, parsimonious statistical models offer superior forecasting performance in small-sample annual energy data regimes.

Keywords: *Renewable energy forecasting; Walk-forward cross-validation; Exponential smoothing; Deep learning; Structural break; SDG-7; Energy transition*

1. Introduction

The global energy transition from fossil fuels to renewable sources represents one of the most consequential structural shifts in the contemporary world economy. Recent geopolitical tensions, supply disruptions, and fossil fuel price volatility have underscored the strategic importance of energy independence and exposed the vulnerability of energy systems that

depend heavily on concentrated hydrocarbon supply chains. Driven by climate commitments under the Paris Agreement (UNFCCC, 2015), declining technology costs, and energy security considerations, the share of renewable energy in total final consumption has accelerated markedly since 2014. Monitoring and forecasting this transition at the global and regional scales is essential for policymakers, international organisations, and energy system planners who must allocate resources, design incentive structures, and assess progress toward internationally agreed targets such as Sustainable Development Goal 7 (SDG-7).

Time series forecasting of energy variables has attracted substantial research attention over the past decade, with a pronounced shift toward machine learning (ML) and deep learning (DL) methods. Recurrent neural networks, particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures, have demonstrated competitive performance on electricity demand, wind power, and solar irradiance forecasting tasks where high-frequency data (hourly or sub-hourly) provide thousands to millions of training observations (Ahmed et al., 2024; Wang et al., 2019). More recently, purpose-built architectures such as N-BEATS (Oreshkin et al., 2020) and Transformer-based models have further expanded the methodological frontier.

However, a significant gap exists between the data regimes in which DL methods have proven effective and the data regimes that characterise internationally comparable energy statistics. The World Bank's flagship renewable energy indicator (EG.FEC.RNEW.ZS) provides annual observations from 1990 onward, yielding approximately 31 data points for any given country or regional aggregate. At this sample size, deep learning models with thousands of trainable parameters face severe underdetermination, and the risk of overfitting to noise rather than capturing genuine temporal structure becomes acute (Makridakis et al., 2018). Despite this, a growing number of studies apply DL methods to annual energy data without adequate acknowledgement of the small-sample constraint or rigorous comparison against parsimonious classical alternatives.

Furthermore, methodological inconsistencies in the existing benchmarking literature undermine the reliability of reported model rankings. A common practice is to evaluate ARIMA with rolling one-step-ahead refitting while assessing DL models on a single frozen fit—an information asymmetry that structurally inflates the reported accuracy of ARIMA (Hyndman and Athanasopoulos, 2021). Similarly, many studies report results from a single train–test split, which is particularly problematic when the underlying data-generating process exhibits structural breaks.

This study addresses these limitations through a comprehensive benchmarking framework that enforces methodological fairness across all model families. Whereas prior work has examined individual model families or used inconsistent evaluation protocols, this study makes five specific contributions that collectively advance the state of the art:

(1) A unified walk-forward cross-validation protocol that evaluates 13 model families under identical conditions—including identical information sets, identical forecast

horizons, and no intermediate retraining—eliminating the information asymmetry present in prior benchmarking studies.

(2) Formal structural break detection via the Chow test, integrated into the evaluation design such that walk-forward windows are deliberately positioned to straddle the identified breakpoint, providing a principled measure of each model's adaptability to regime change.

(3) A rigorous statistical inference framework combining Diebold–Mariano tests, paired bootstrap comparison, and Model Confidence Set analysis, providing set-level inference that accounts for the multiple comparison problem inherent in benchmarking many models simultaneously.

(4) Multi-seed deep learning robustness experiments (10 seeds per architecture) that quantify the sensitivity of DL results to random initialisation on small datasets, directly addressing the reproducibility concerns that pervade small-sample DL applications.

(5) Three novel energy transition analytics—a Transition Velocity Index (TVI), regional beta-convergence analysis, and SDG-7 gap assessment across 11 World Bank regions—that bridge the gap between forecasting methodology and actionable policy intelligence for energy independence planning.

Together, these contributions provide the most comprehensive and methodologically rigorous benchmarking study of forecasting models applied to annual macro-energy indicators to date.

2. Literature Review

2.1. Renewable Energy Forecasting

Renewable energy forecasting encompasses a broad methodological spectrum, ranging from physics-based models for wind and solar output to purely data-driven statistical and machine learning approaches. At the macro level, forecasting aggregate renewable energy consumption as a share of total final energy is critical for national energy planning and international policy assessment (IRENA, 2023). Early work in this domain relied predominantly on exponential smoothing, ARIMA, and regression-based approaches (Hyndman and Athanasopoulos, 2021; Makridakis et al., 2018).

The IEA World Energy Outlook and IRENA's Global Energy Transformation reports provide scenario-based projections that incorporate techno-economic assumptions, but these are not strictly time series forecasts and rely heavily on expert judgement (IEA, 2023). Statistical forecasting of energy indicators using World Bank data has been explored by several authors. Karakurt and Aydin (2023) applied ARIMA and ETS to energy intensity indicators, while Khan et al. (2020) used regression-based approaches for renewable energy share projections. However, few studies have systematically benchmarked multiple model families against the same indicator using rigorous cross-validation protocols.

2.2. Machine Learning and Deep Learning in Energy Forecasting

The application of machine learning to energy forecasting has expanded rapidly since 2015. XGBoost and gradient boosting methods have shown strong performance on electricity price and demand forecasting tasks where engineered features (lag values, calendar variables, weather covariates) provide rich input representations (Chen and Guestrin, 2016; Lago et al., 2021). Deep learning approaches, particularly LSTM networks (Hochreiter and Schmidhuber, 1997), have been widely applied to electricity demand (Kong et al., 2019), wind power (Wang et al., 2019), solar irradiance forecasting (Ahmed et al., 2024), and renewable energy consumption forecasting at the macro level (Biswas, Irshad and Roy, 2026).

GRU networks (Cho et al., 2014) offer a computationally lighter alternative to LSTM with comparable performance on many time series tasks. N-BEATS (Oreshkin et al., 2020), a feed-forward architecture using residual stacking and polynomial basis expansion, achieved state-of-the-art results on the M4 competition dataset. Prophet (Taylor and Letham, 2018) employs an additive decomposition framework with automatic changepoint detection, making it accessible for practitioners without specialist forecasting expertise.

Despite these advances, the performance advantage of DL methods is most pronounced with high-frequency data containing thousands of observations. Makridakis et al. (2018) demonstrated in the M4 competition that simple statistical methods outperformed complex ML and DL models on many time series, particularly those with fewer than 100 observations. Petropoulos et al. (2022) reinforced this finding in a comprehensive review of forecasting methodology, noting that model complexity should be calibrated to data availability. De Oliveira and Cyrino Oliveira (2018) further demonstrated that bagging methods applied to ARIMA and exponential smoothing could outperform standalone complex models on mid-to-long-term energy forecasting tasks, while Deb et al. (2017) provided a systematic review showing that classical methods remain competitive for building energy consumption forecasting when data is limited.

2.3. Hybrid Forecasting Models

Hybrid models that combine statistical and ML/DL components have attracted growing interest as a strategy to capture both linear trend structure and nonlinear residual patterns. Zhang (2003) proposed the seminal ARIMA–ANN hybrid that decomposes a series into linear and nonlinear components. More recent hybrids include ETS–LSTM combinations (Smyl, 2020), which won the M4 competition, and ensemble approaches that average forecasts from complementary model families (Atiya, 2020). The rationale is that statistical models efficiently capture trend and seasonality with minimal parameters, while DL components model residual nonlinearity that escapes the statistical specification.

2.4. Forecast Evaluation Methods

Rigorous forecast evaluation extends beyond point accuracy metrics. The Diebold–Mariano test (Diebold and Mariano, 1995) provides a formal framework for testing whether two forecasting methods produce significantly different prediction errors. The Model Confidence

Set (Hansen et al., 2011) generalises this to the multi-model setting, identifying the subset of models whose predictive ability cannot be statistically distinguished from the best model at a given significance level. Bootstrap methods (Efron and Tibshirani, 1993) provide distribution-free confidence intervals that are particularly valuable when the number of test observations is small and distributional assumptions may be violated.

Walk-forward cross-validation (Tashman, 2000) is the standard evaluation protocol for time series, expanding the training window sequentially and producing out-of-sample forecasts at each step. This approach respects the temporal ordering of data and provides multiple evaluation points, yielding more robust performance estimates than a single train–test split.

2.5. Research Gap

Despite the extensive literature on energy forecasting, several gaps remain. First, few studies benchmark a comprehensive set of model families—from simple baselines through statistical methods to deep learning—under strictly identical evaluation conditions on annual macro-energy indicators. Second, the small-sample challenge inherent to annual energy statistics is rarely addressed explicitly; most DL applications to energy data use high-frequency datasets where the data regime is fundamentally different. Third, structural break detection is seldom integrated into the forecasting evaluation framework, despite the known acceleration in renewable energy deployment following the Paris Agreement. Fourth, no prior study has combined rigorous forecasting benchmarking with novel policy-relevant analytics (transition velocity measurement, convergence testing, SDG-7 gap analysis) in a unified framework. This study addresses all four gaps.

3. Data

The dataset used in this study is the World Bank indicator EG.FEC.RNEW.ZS, defined as renewable energy consumption as a percentage of total final energy consumption. The indicator covers 1990–2020, providing 31 annual observations for each of 11 aggregate regional series: East Asia and Pacific, Europe and Central Asia, High income, Latin America and Caribbean, Low income, Lower middle income, North America, South Asia, Sub-Saharan Africa, Upper middle income, and World. Pre-1990 values were excluded from the analysis as they contain backfilled identical values from a single 1990 anchor, carrying no additional temporal signal.

It is important to note that the EG.FEC.RNEW.ZS indicator includes traditional solid biomass, which constitutes a substantial proportion of renewable energy consumption in developing regions. High renewable shares in Sub-Saharan Africa (approximately 70–91%) and Low income regions (approximately 55–75%) primarily reflect reliance on wood fuel and charcoal for cooking and heating rather than modern renewable energy deployment such as wind, solar, or geothermal. This compositional distinction is critical for interpreting cross-regional comparisons and policy implications.

As of the date of this study, 2020 remains the most recent year for which the EG.FEC.RNEW.ZS indicator is available across all World Bank regional aggregates. The underlying data are sourced from the IEA Energy Statistics Data Browser, whose SE4ALL tracking framework updates with a multi-year lag for global coverage. The 1990–2020 series therefore represents the complete available dataset for this indicator.

The primary forecasting target is the World aggregate series, which exhibited a U-shaped trajectory over the study period. From 1990 to approximately 2008, the global renewable share declined gradually from 16.7% to 16.6%, reflecting the more rapid growth of fossil fuel consumption relative to renewables. A turning point emerged around 2010–2013, followed by a marked upward acceleration from 2014 onward, with the World aggregate reaching 19.8% by 2020.

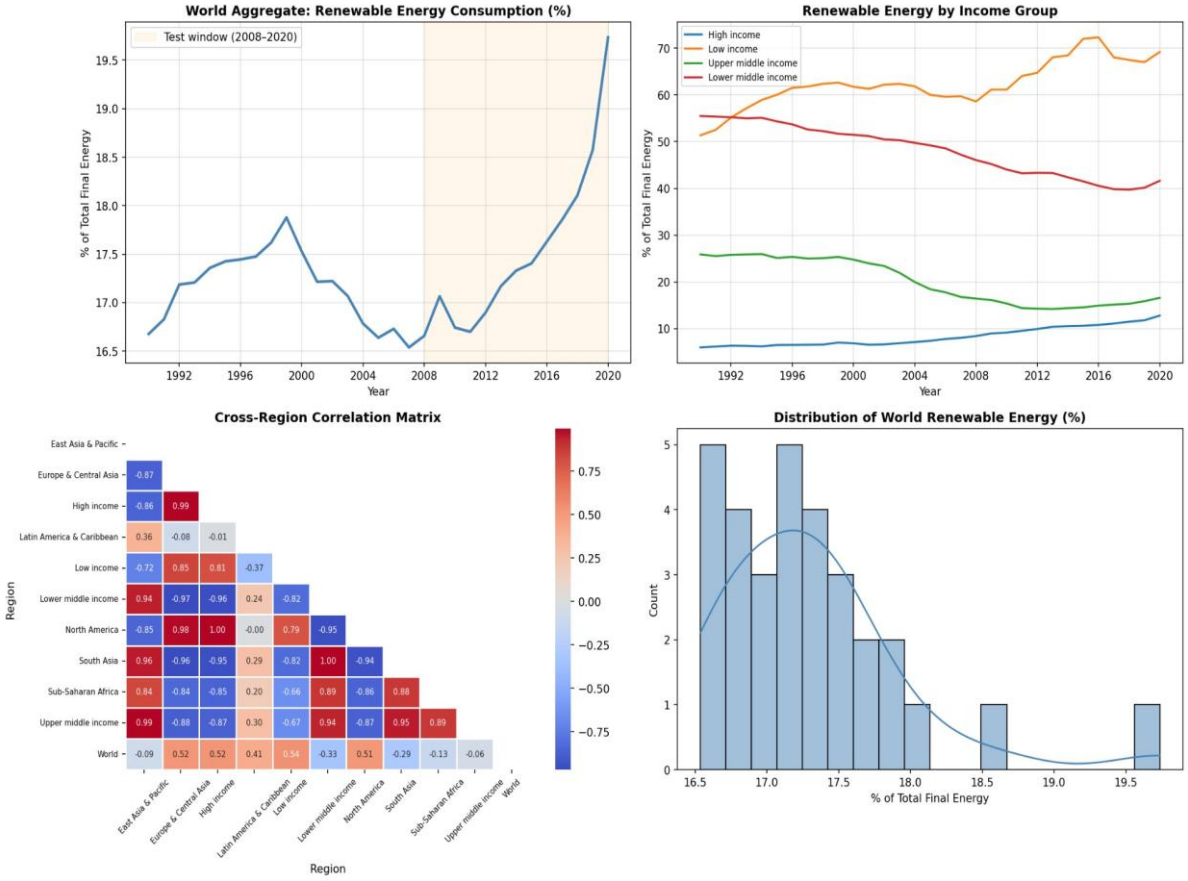

Figure 1. Exploratory Data Analysis: World renewable energy time series, income-group trajectories, cross-region correlation matrix, and distributional summary.

3.1. Structural Break Evidence

A Chow F-test applied iteratively across all candidate breakpoints in the 1990–2020 World series identified 2014 as the year of the most statistically significant structural change ($F = 32.0$, $p < 0.001$). This result was confirmed by CUSUM test diagnostics. Segmented regression analysis revealed two distinct data-generating regimes: Phase 1 (1990–2013) with a near-zero trend of −0.021 percentage points per year, and Phase 2 (2014–2020) with a steep upward trend of +0.359 percentage points per year. This breakpoint coincides with the period

surrounding the Paris Agreement negotiations and the acceleration in global renewable energy investment and policy deployment.

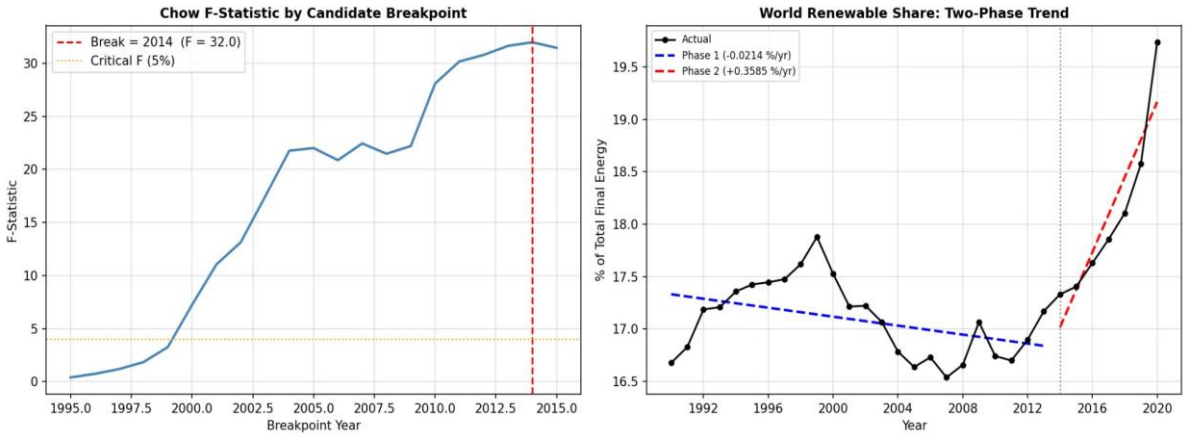

Figure 2. Structural break detection: Chow F-statistic by candidate breakpoint (left) and two-phase segmented regression (right).

4. Methodology

This section describes the 13 forecasting models evaluated in the study, organised by methodological category.

4.1. Baseline Models

Three baseline models establish the minimum performance threshold. The Naïve method carries forward the last observed value for all forecast horizons. The Random Walk with Drift extends this by adding the historical mean period-to-period change as a constant drift term. The Linear Trend model fits an ordinary least squares regression of the series on a time index and extrapolates the fitted line.

4.2. Statistical Models

Four classical statistical methods are evaluated. Holt Linear Exponential Smoothing (ETS) uses two optimised parameters (α for level smoothing and β for trend smoothing) fitted via maximum likelihood, re-optimised on each expanding window. Damped ETS extends Holt's method with a damping parameter φ that gradually attenuates the trend, mitigating over-extrapolation on longer horizons. The Theta method decomposes the series into two theta-lines with different curvatures and combines them, as proposed by Assimakopoulos and Nikolopoulos (2000). ARIMA with automatic order selection (via the AIC criterion) is fitted using the pmdarima library, with orders determined independently for each expanding window.

4.3. Machine Learning Model

XGBoost (Chen and Guestrin, 2016) is evaluated with engineered lag features comprising lags 1 through 5, a 3-period rolling mean, and a 3-period rolling standard deviation. Multi-step forecasting is achieved iteratively: each predicted value is fed back as input for the

subsequent step. Hyperparameters (maximum tree depth and number of estimators) are tuned via the nested cross-validation procedure described in Section 5.

4.4. Deep Learning Models

Three deep learning architectures are evaluated. The GRU (Cho et al., 2014) uses a gated recurrent architecture with a hidden size of 64, employing input and reset gates to control information flow without a separate cell state. The LSTM (Hochreiter and Schmidhuber, 1997) uses the canonical architecture with separate cell state and hidden state, also with a hidden size of 64. Both recurrent models use a sequence length of 3, are trained for 300 epochs with a learning rate of 0.001 and batch size of 8, using the Adam optimiser. Min–max normalisation is applied within each training window to prevent data leakage.

N-BEATS (Oreshkin et al., 2020) is a feed-forward architecture using stacked residual blocks with polynomial basis expansion. The implementation uses 2 stacks with a hidden dimension of 64 and a polynomial degree of 2. All deep learning models use iterative multi-step forecasting, producing one-step-ahead predictions that are fed back as input for subsequent steps.

4.5. Additive Model

Prophet (Taylor and Letham, 2018) is an additive decomposition model with automatic changepoint detection. Configuration includes a changepoint_prior_scale of 0.3 (moderate flexibility), with yearly, weekly, and daily seasonality disabled (annual data). Explicit changepoints at 2001 (dot-com recession) and 2007 (global financial crisis) are provided as prior knowledge.

4.6. Hybrid Model

The ETS–GRU hybrid is an equal-weight ensemble that averages the point forecasts of ETS and GRU at each forecast step. This design is intentionally parsimonious: rather than introducing learned combination weights (which would require additional training data that this small-sample regime cannot support), the equal-weight average follows the theoretical result of Atiya (2020) that simple averaging of complementary forecasters often matches or exceeds optimally weighted combinations when the number of forecast observations is small. The rationale is that ETS efficiently captures the dominant linear trend component while GRU captures potential nonlinear residual structure. Both component forecasts are computed independently under the same walk-forward protocol, and the ensemble is constructed post hoc without additional parameter estimation.

5. Experimental Design

5.1. Walk-Forward Cross-Validation

The evaluation protocol uses a 5-window expanding walk-forward cross-validation design with a fixed horizon of H = 3 years, yielding 15 total test observations. The training window expands from 20 observations (1990–2009) in Window 1 to 28 observations (1990–2017) in

Window 5. Critically, every model—including ARIMA—trains once per window and produces a single H-step-ahead forecast with no access to intermediate test actuals. This eliminates the information asymmetry that arises when ARIMA is evaluated with rolling one-step-ahead refitting while DL models receive a single forecast call.

The window placement is designed to interact with the structural break at 2014. Windows 1 and 2 forecast exclusively within the pre-break regime. Window 3 straddles the breakpoint, forecasting 2014–2016 from training data that is predominantly pre-break. Windows 4 and 5 forecast increasingly into the post-break acceleration regime, providing a direct measure of each model's ability to adapt to a changed data-generating process.

5.2. Nested Hyperparameter Tuning

Deep learning and XGBoost hyperparameters are tuned via nested cross-validation to prevent information leakage between the tuning and evaluation stages. For each outer walk-forward window, an inner loop performs leave-one-out validation on the training portion, evaluating candidate configurations and selecting the one with the lowest inner RMSE. The winning configuration is then used to produce the outer-loop forecast. This ensures that hyperparameter choices do not benefit from exposure to test-period data.

The hyperparameter search grids include hidden dimensions of 32 and 64 with sequence lengths of 3 and 5 for GRU and LSTM; hidden dimensions of 32 and 64 with 2 and 3 stacks for N-BEATS; and maximum tree depths of 2 and 3 with 50 and 100 estimators for XGBoost.

5.3. Multi-Seed Deep Learning Robustness

A fundamental concern with deep learning results on small datasets is sensitivity to random weight initialisation. A single result may be atypically favourable or unfavourable. To quantify this variability, the full walk-forward evaluation is repeated across 10 random seeds for GRU, LSTM, N-BEATS, and XGBoost. The mean RMSE and standard deviation across seeds are reported alongside the primary (seed = 42) results, providing a measure of the reliability of observed performance differences.

6. Forecast Evaluation Framework

Model performance is assessed through a multi-layered evaluation framework comprising point accuracy metrics, statistical significance tests, and prediction interval analysis.

6.1. Point Accuracy Metrics

Three metrics are computed across all 15 test observations. Root Mean Squared Error (RMSE) serves as the primary ranking metric due to its sensitivity to large errors, which is particularly relevant when forecasting through a structural break. Mean Absolute Error (MAE) provides a complementary scale-dependent measure that is less sensitive to outliers. Mean Absolute Percentage Error (MAPE) offers scale-invariant comparison. The Skill Score (SS = 1 − RMSE_model / RMSE_Naïve) measures improvement relative to the Naïve

baseline: positive values indicate the model outperforms the Naïve; negative values indicate it is worse.

6.2. Diebold–Mariano Tests

The Diebold–Mariano (DM) test (Diebold and Mariano, 1995) is applied to all pairwise model comparisons using squared error loss. The test statistic employs a Newey–West heteroskedasticity and autocorrelation consistent (HAC) variance estimator. Statistical significance is assessed at the 5% level. With only 15 test observations, the DM test has limited statistical power, and results should be interpreted cautiously.

6.3. Model Confidence Set

The Model Confidence Set (MCS) procedure (Hansen et al., 2011) is applied at the 10% significance level using 1,000 bootstrap resamples. The MCS identifies the smallest set of models for which the null hypothesis of equal predictive ability cannot be rejected, providing a set-level inference that accounts for the multiple comparison problem inherent in benchmarking many models simultaneously.

6.4. Bootstrap Prediction Intervals

Bootstrap prediction intervals at the 80% and 95% levels are constructed for the top-performing models using empirical residual resampling (500 draws per window). Coverage probability is reported as the fraction of actual test observations falling within the constructed intervals. This provides a calibration check: a well-calibrated 80% interval should contain approximately 80% of realisations.

7. Results

7.1. Benchmark Model Comparison

Table 1 presents the complete benchmark results across all 13 model families, ranked by RMSE. ETS (Holt Linear Exponential Smoothing) achieved the lowest RMSE of 0.543, the lowest MAE of 0.434, and the lowest MAPE of 2.435%, with a Skill Score of +0.148 against the Naïve baseline. The ETS–GRU hybrid ranked second (RMSE = 0.593, SS = +0.070), followed by the Random Walk with Drift (RMSE = 0.593, SS = +0.069), Damped ETS (RMSE = 0.597, SS = +0.063), and Prophet (RMSE = 0.606, SS = +0.048).

Rank	Model	Category	RMSE	MAE	MAPE (%)	Skill Score
1	**ETS**	Statistical	0.543	0.434	2.435	+0.148
2	**ETS–GRU**	Hybrid	0.593	0.453	2.511	+0.070
3	**RW-Drift**	Baseline	0.593	0.450	2.510	+0.069
4	**Damped ETS**	Statistical	0.597	0.459	2.547	+0.063

5	**Prophet**	Additive	0.606	0.470	2.612	+0.048
6	**Naïve**	Baseline	0.637	0.483	2.659	+0.000
7	**GRU**	Deep Learning	0.662	0.527	2.902	−0.039
8	**LSTM**	Deep Learning	0.730	0.584	3.216	−0.145
9	**ARIMA**	Statistical	0.775	0.586	3.226	−0.217
10	**XGBoost**	Machine Learning	0.788	0.584	3.244	−0.237
11	**Theta**	Statistical	0.788	0.588	3.333	−0.237
12	**N-BEATS**	Deep Learning	0.867	0.605	3.216	−0.361
13	**LinTrend**	Baseline	0.979	0.750	4.117	−0.538

Table 1. Benchmark results: 13 model families ranked by RMSE. Walk-forward CV with 5 windows, H = 3, 15 test observations.

A notable finding is that only 5 of the 13 models achieved positive Skill Scores, indicating that the majority of models—including all three deep learning architectures (GRU, LSTM, N-BEATS), XGBoost, Theta, and ARIMA—performed worse than the Naïve baseline. This result underscores the challenging nature of forecasting annual renewable energy consumption with small sample sizes.

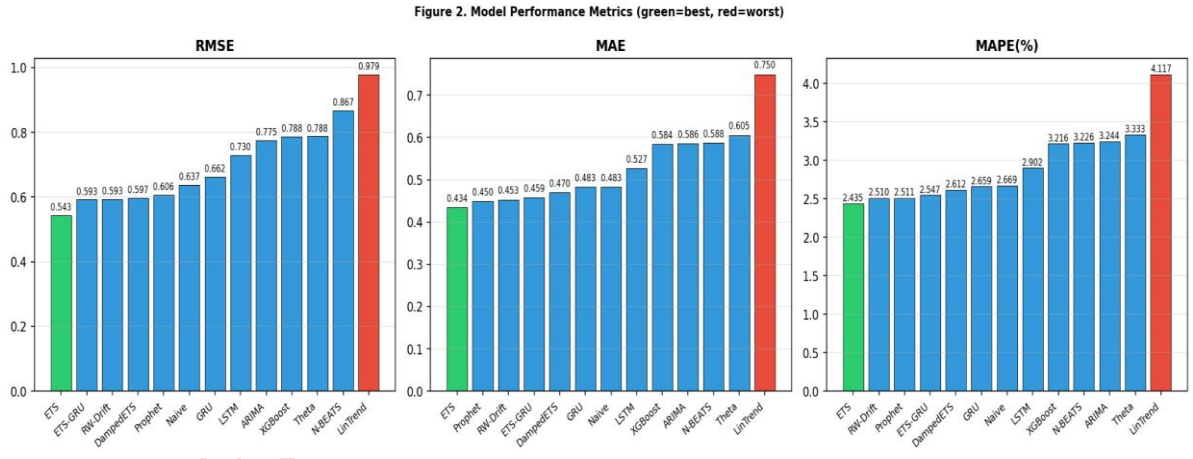

Figure 3. Model performance metrics: RMSE, MAE, and MAPE across all 13 model families.

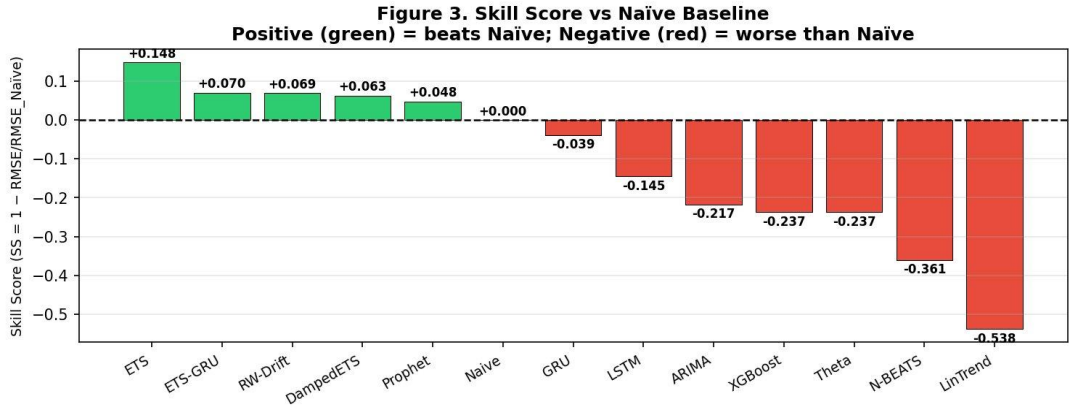

Figure 4. Skill Score relative to Naïve baseline. Positive (green) = outperforms Naïve; negative (red) = worse.

7.2. Walk-Forward Forecast Trajectories

Figure 5 presents the walk-forward forecast trajectories for all models across the five evaluation windows. In the pre-break windows (W1, W2), most models produce reasonably accurate forecasts, and performance differences are modest. The critical divergence occurs in Windows 4 and 5, where the post-2014 acceleration causes the actual series to rise steeply. Models trained predominantly on the stagnant Phase 1 data systematically underpredict the post-break acceleration, with errors escalating sharply in 2019–2020.

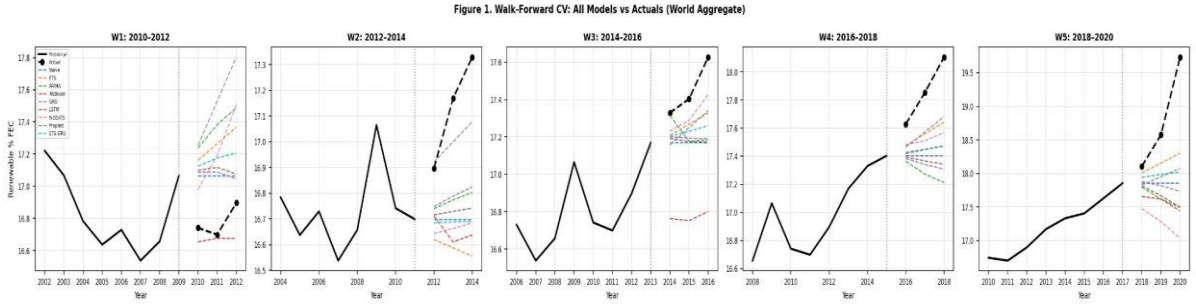

Figure 5. Walk-forward CV: All models vs actuals across 5 evaluation windows (World aggregate).

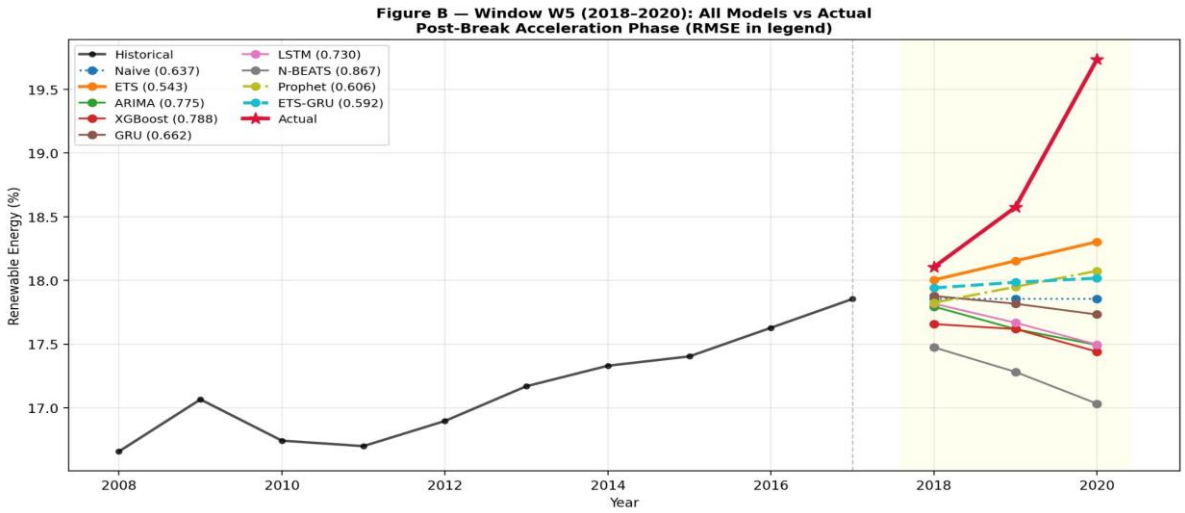

Figure 6. Window W5 (2018–2020): All models vs actual in the post-break acceleration phase. RMSE values shown in legend.

7.3. Per-Window RMSE and Structural Break Impact

The per-window RMSE decomposition reveals the impact of the structural break on model performance. All models exhibit substantially higher RMSE in Window 5 (2018–2020) compared to earlier windows. ETS demonstrates the most graceful degradation, with W5 RMSE of 0.866 compared to W1 RMSE of 0.493. In contrast, Prophet exhibits dramatic degradation (W5 RMSE = 1.776 vs. W1 RMSE = 0.768), and N-BEATS degrades similarly (W5 RMSE = 1.039 vs. W1 RMSE of 0.474). The consistent pattern of W5 degradation across all model families confirms that the post-break acceleration represents a genuinely novel data-generating regime that challenges all forecasting approaches.

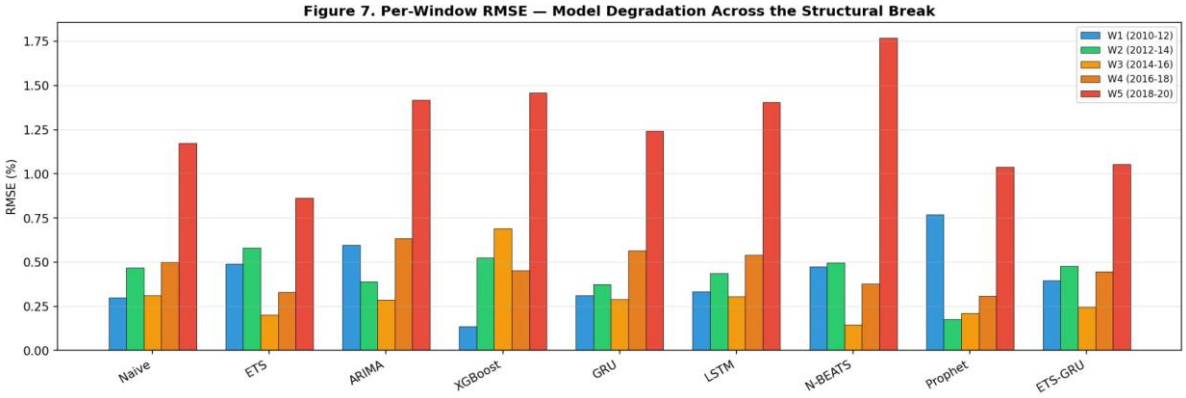

Figure 7. Per-window RMSE: Model degradation across the structural break.

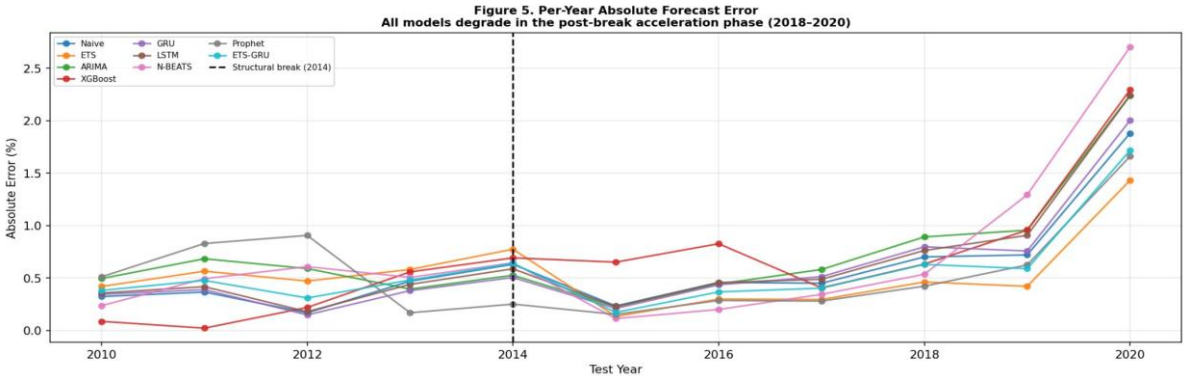

Figure 8. Per-year absolute forecast error. All models degrade sharply in the post-break phase (2018–2020).

7.4. Statistical Significance

The Diebold–Mariano significance matrix (Figure 9) reveals that, despite clear differences in point RMSE, few pairwise comparisons achieve formal statistical significance at the 5% level. The only significant comparison is ARIMA vs. GRU (DM statistic = 1.85, $p < 0.05$), where ARIMA is significantly less accurate than GRU. The limited statistical power is expected given only 15 test observations and the relatively modest RMSE differences among the top-ranked models. The Model Confidence Set at the 10% significance level retains ETS, ETS–GRU, RW-Drift, Damped ETS, Prophet, and Naïve, confirming that these six models cannot be statistically distinguished from the best performer.

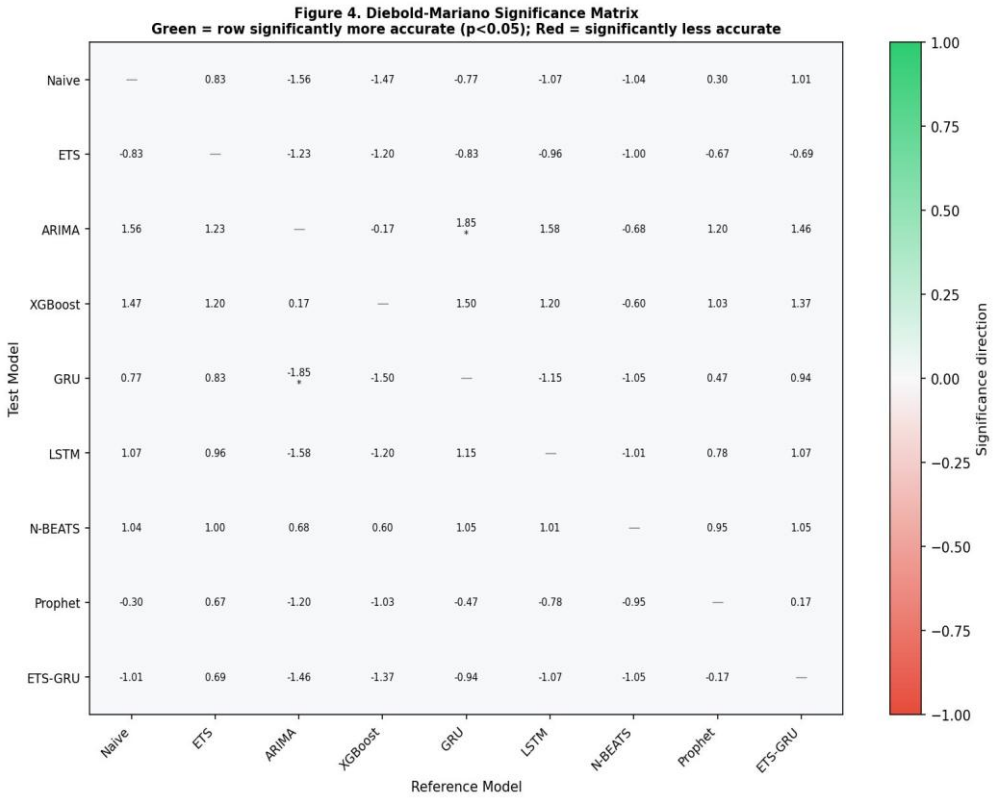

Figure 9. Diebold–Mariano significance matrix. Green = row model significantly more accurate ($p < 0.05$); Red = significantly less accurate.

7.5. Model Ranking Stability

Ranking stability analysis across the five walk-forward windows reveals considerable variation. Prophet achieves the best mean rank (approximately 1.8) but with high standard deviation (4.6), indicating inconsistent performance across windows. ETS shows a more moderate mean rank with lower variance, suggesting greater reliability. The ranking stability map (Figure 10) plots mean rank against rank standard deviation; models in the bottom-left quadrant (low mean rank, low variance) represent the most reliably good performers.

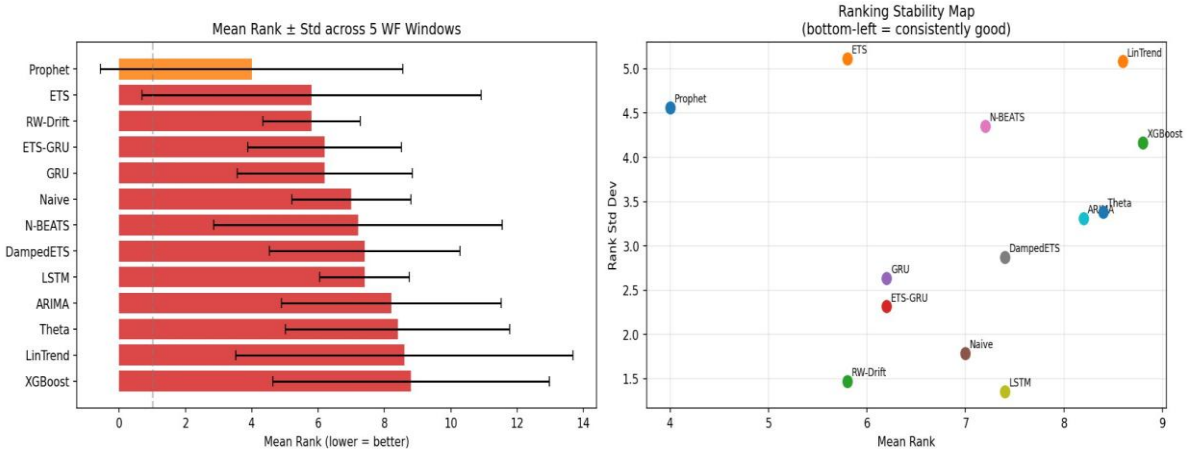

Figure 10. Model ranking stability: Mean rank ± standard deviation across 5 WF windows (left); Ranking stability map (right).

7.6. Residual Diagnostics and Prediction Intervals

Residual diagnostics indicate that all models exhibit negative forecast bias—systematically underpredicting the renewable energy share—consistent with the challenge of forecasting through a positive structural break. ETS shows the smallest bias (−0.240), while XGBoost exhibits the largest (−0.584). Ljung–Box tests suggest no significant residual autocorrelation for any model, though several models show departures from normality (Shapiro–Wilk $p < 0.05$), supporting the use of bootstrap rather than parametric inference.

Bootstrap prediction intervals for the top four models (ETS, ETS–GRU, Prophet, Naïve) show moderate calibration. ETS achieves 80% coverage of 53% and 95% coverage of 53%, indicating that the prediction intervals are narrower than ideal—a consequence of the structural break causing actual values to fall outside the range anticipated by residual-based bootstrapping.

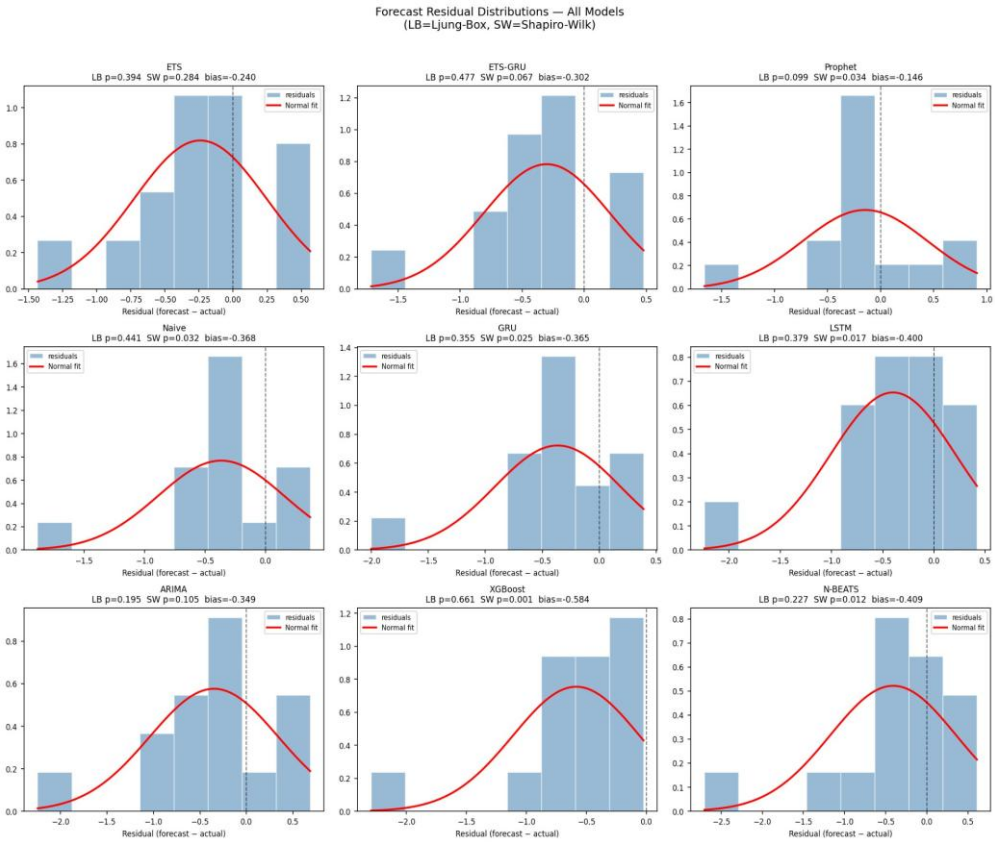

Figure 11. Forecast residual distributions for all models. LB = Ljung–Box p-value; SW = Shapiro–Wilk p-value.

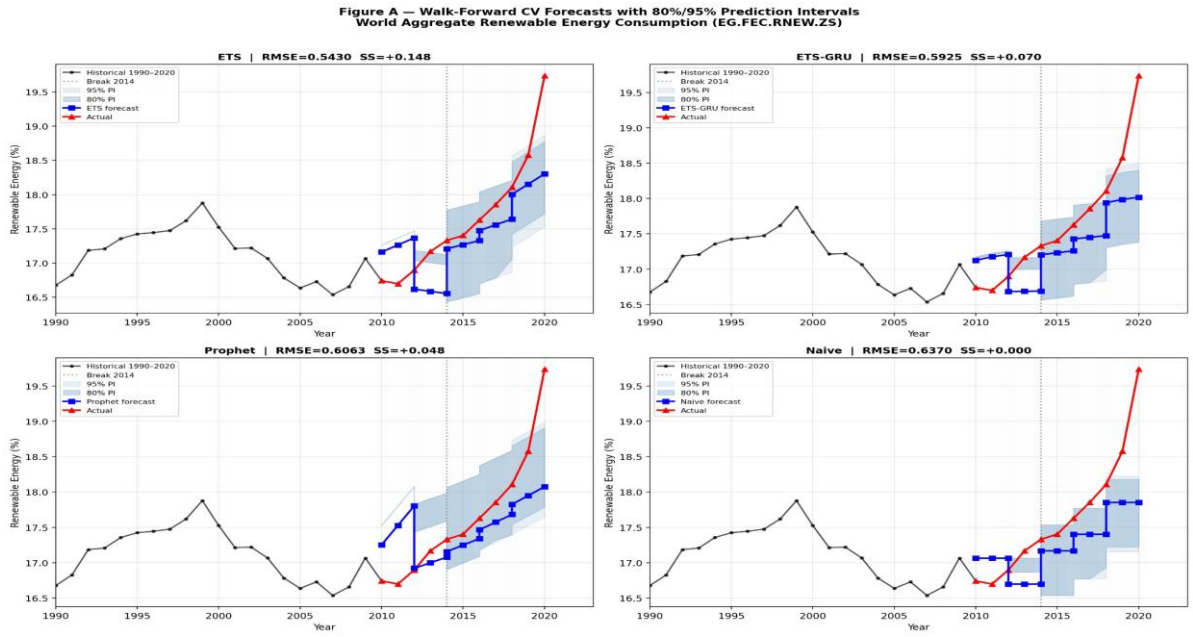

Figure 12. Walk-forward forecasts with 80% and 95% bootstrap prediction intervals for top 4 models.

8. Discussion

The central finding of this study—that Holt Linear Exponential Smoothing outperforms all deep learning architectures on this dataset—is not a general claim about forecasting methodology. It is a data-regime-specific finding arising from four quantifiable properties of the World Bank annual renewable energy series.

First, the observations-to-parameters ratio is severely unfavourable for deep learning. With 20–28 training observations per window and DL models carrying 3,393 (GRU) to 4,513 (LSTM) trainable parameters, the ratio is approximately 1:150. At this extreme, gradient-based optimisation cannot meaningfully constrain the parameter space, and the models effectively memorise the training data without learning generalisable temporal structure. By contrast, ETS has only 2 free parameters (α, β), yielding a ratio of approximately 10:1 to 14:1.

Second, the signal-to-noise ratio is dominated by a single linear trend component that accounts for most of the explainable variance in the post-2014 regime. ETS is precisely specified for this structure: a two-parameter linear trend model applied to a series with a dominant linear trend is near-optimal in the bias–variance sense. DL models, with their vastly greater representational capacity, incur variance costs without commensurate bias reduction.

Third, the structural break at 2014 compounds the small-sample problem. In Windows 3–5, models trained on 24–28 observations must forecast into a regime that differs qualitatively from the majority of their training data. ETS adapts because its exponential weighting naturally up-weights recent observations where the new trend is most evident. ARIMA, by contrast, selects its order based on the full training sample, which is dominated by the stagnant Phase 1, and consequently anchors its forecasts to the old regime.

Fourth, the annual frequency eliminates the high-frequency patterns (seasonality, intra-day cycles, weather effects) that provide the feature-rich environment in which DL methods excel. Without these patterns, the additional representational capacity of neural networks provides no advantage and merely introduces noise sensitivity.

These findings align with the broader M4 competition results (Makridakis et al., 2018) and the forecast methodology review by Petropoulos et al. (2022), both of which emphasise that model complexity should be calibrated to the information content of the data. The practical implication is that practitioners and policymakers working with annual energy statistics should not default to complex ML or DL approaches without first establishing that simpler methods have been outperformed under rigorous evaluation conditions.

9. Policy Implications

The forecasting results are complemented by three novel energy transition analytics that provide direct policy relevance.

9.1. Twenty-Year Regional Forecasts

Using the champion ETS model retrained on the full 1990–2020 series, 20-year point forecasts (2021–2040) were generated for all 11 World Bank regions. These forecasts extrapolate the exponentially smoothed trend observed in each region's historical data. Latin America and Caribbean, South Asia, and Sub-Saharan Africa show the steepest projected growth trajectories, while North America and Europe and Central Asia show more moderate increases from higher base levels.

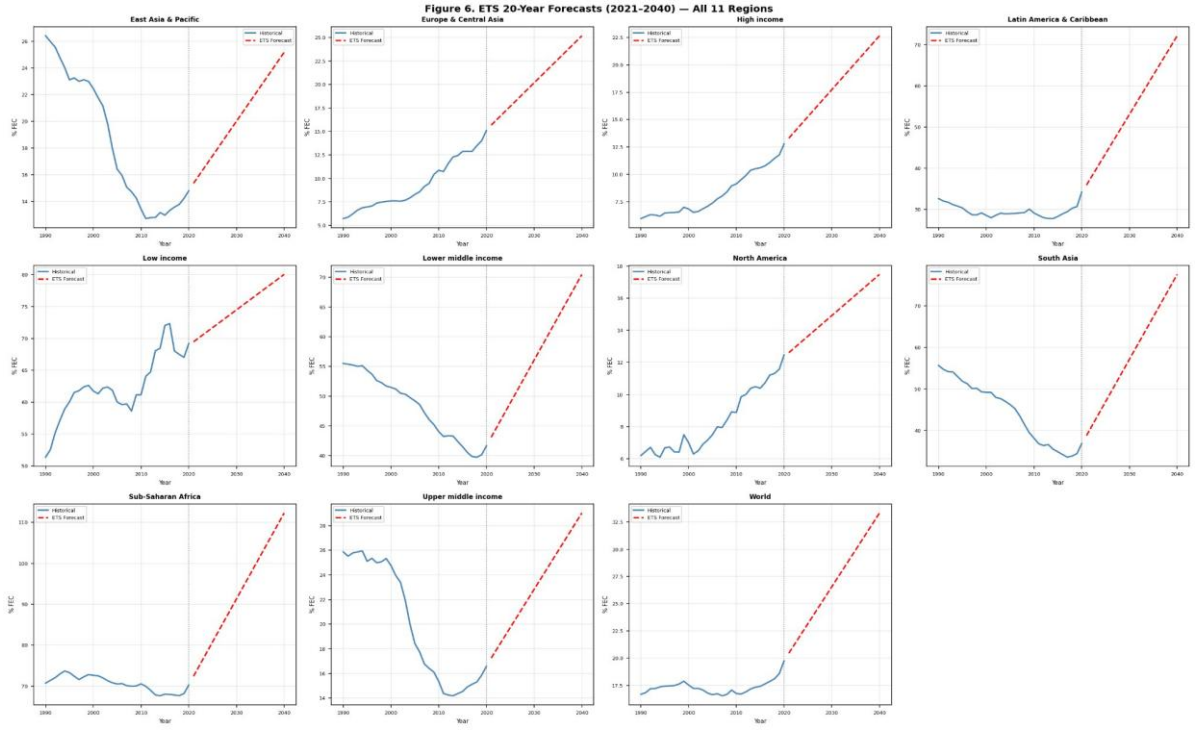

Figure 13. ETS 20-year forecasts (2021–2040) for all 11 World Bank regions.

9.2. Transition Velocity Index

The Transition Velocity Index (TVI) measures the rate of renewable energy transition relative to a fixed 2013 baseline (the last pre-break year), enabling direct comparison across regions with different absolute renewable shares. Historically (2013–2020), Latin America and Caribbean, High income, and Europe and Central Asia exhibited the highest transition velocities (approximately +23% relative change). Notably, Lower middle income was the only grouping with a negative historical TVI (−3.8%), indicating regression. Forecast TVI values (2013–2040, ETS-based) project all regions to achieve positive long-term transition velocities, with Latin America and Caribbean leading at +160%.

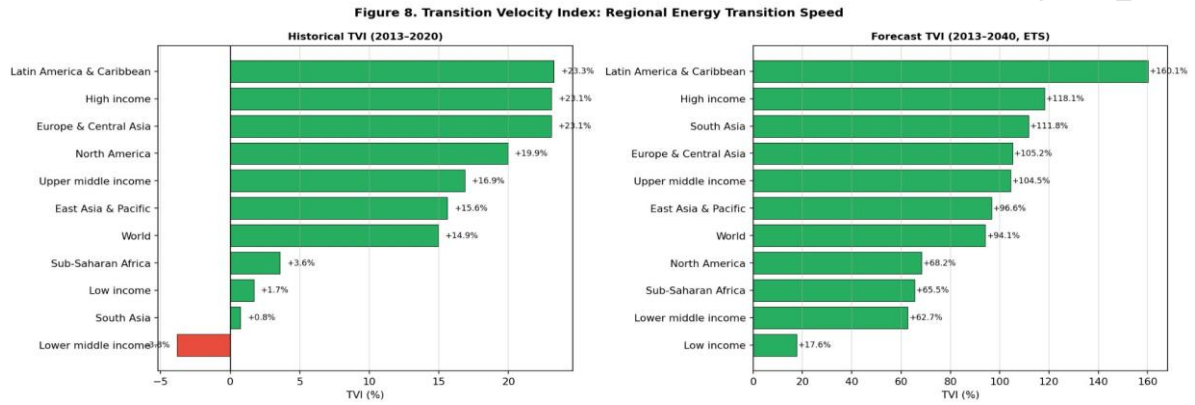

Figure 14. Transition Velocity Index: Historical (2013–2020) and forecast (2013–2040, ETS) regional energy transition speed.

9.3. SDG-7 Gap Analysis

The SDG-7 gap analysis compares ETS 2030 forecasts for each region against the IEA Net Zero by 2050 intermediate milestone of 30% renewable final energy consumption. Six regions—Sub-Saharan Africa (91.3%), Low income (74.5%), South Asia (57.1%), Lower middle income (56.0%), Latin America and Caribbean (53.1%), and the World aggregate (26.6%)—are projected to meet or exceed the 30% target by 2030. However, the high forecasted shares for Sub-Saharan Africa and Low income regions are driven predominantly by traditional biomass rather than modern renewables, highlighting the compositional limitation of the EG.FEC.RNEW.ZS indicator. North America (14.9%), High income (17.7%), East Asia and Pacific (20.0%), and Europe and Central Asia (20.2%) are projected to fall short of the 30% target, indicating substantial policy gaps in regions that are major global energy consumers.

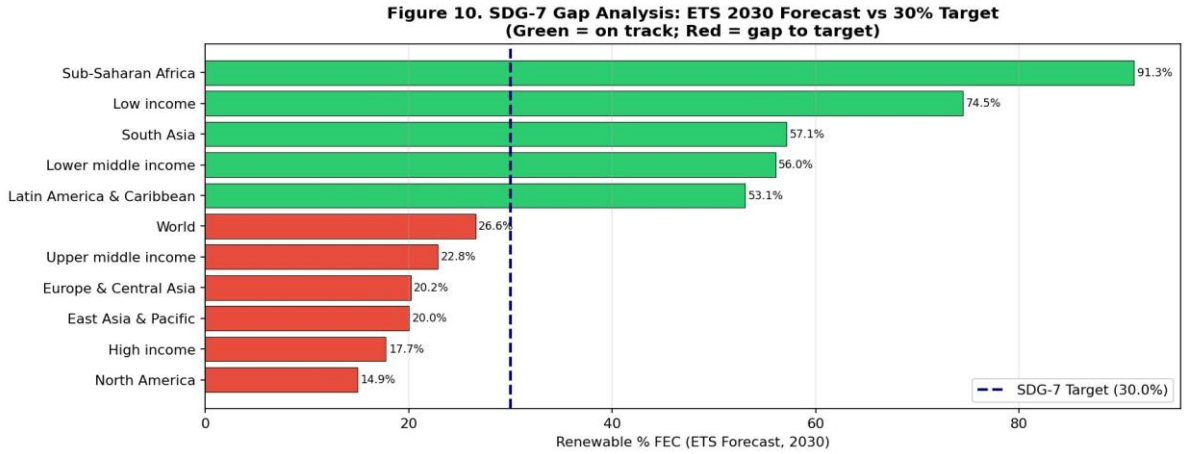

Figure 15. SDG-7 gap analysis: ETS 2030 forecast vs 30% renewable target. Green = on track; Red = gap to target.

9.4. Beta-Convergence

Regional beta-convergence analysis, adapted from economic growth theory (Barro and Sala-i-Martin, 1992), tests whether regions with lower initial renewable shares grow proportionally faster. Regressing annualised log-growth rates on log initial renewable shares across the 11 regions yields a significantly negative β coefficient for both the full 1990–2020 period ($\beta = -0.0140$, $R^2 = 0.578$, $p = 0.011$) and the post-Paris 2014–2020 period ($\beta = -0.0152$, $R^2 = 0.640$, $p = 0.005$). This confirms convergence: regions starting with lower renewable shares are growing proportionally faster, consistent with technology diffusion and late-mover advantage in renewable energy deployment.

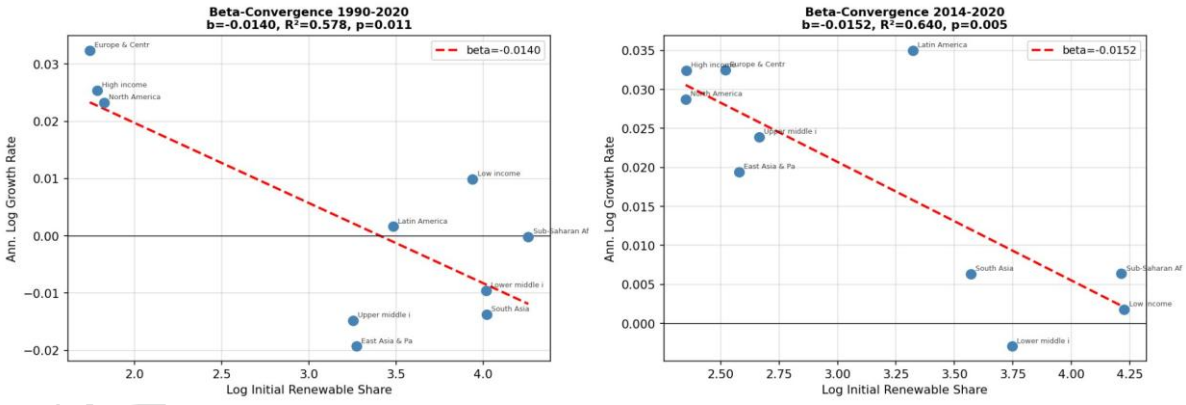

Figure 16. Beta-convergence: 1990–2020 (left) and 2014–2020 (right). Negative β confirms convergence.

9.5. Multi-Region Robustness

The full walk-forward cross-validation was replicated across all 11 World Bank regions for 5 core models (Naïve, ETS, ARIMA, GRU, LSTM). The multi-region heatmap (Figure 17) reveals that no single model dominates across all regions. Naïve achieves the best RMSE in East Asia and Pacific, South Asia, and Sub-Saharan Africa. ETS leads in High income, North America, and the World aggregate. ARIMA performs well in Europe and Central Asia and Lower middle income. This heterogeneity reinforces the conclusion that model selection

should be context-specific and that blanket adoption of any single methodology is inadvisable.

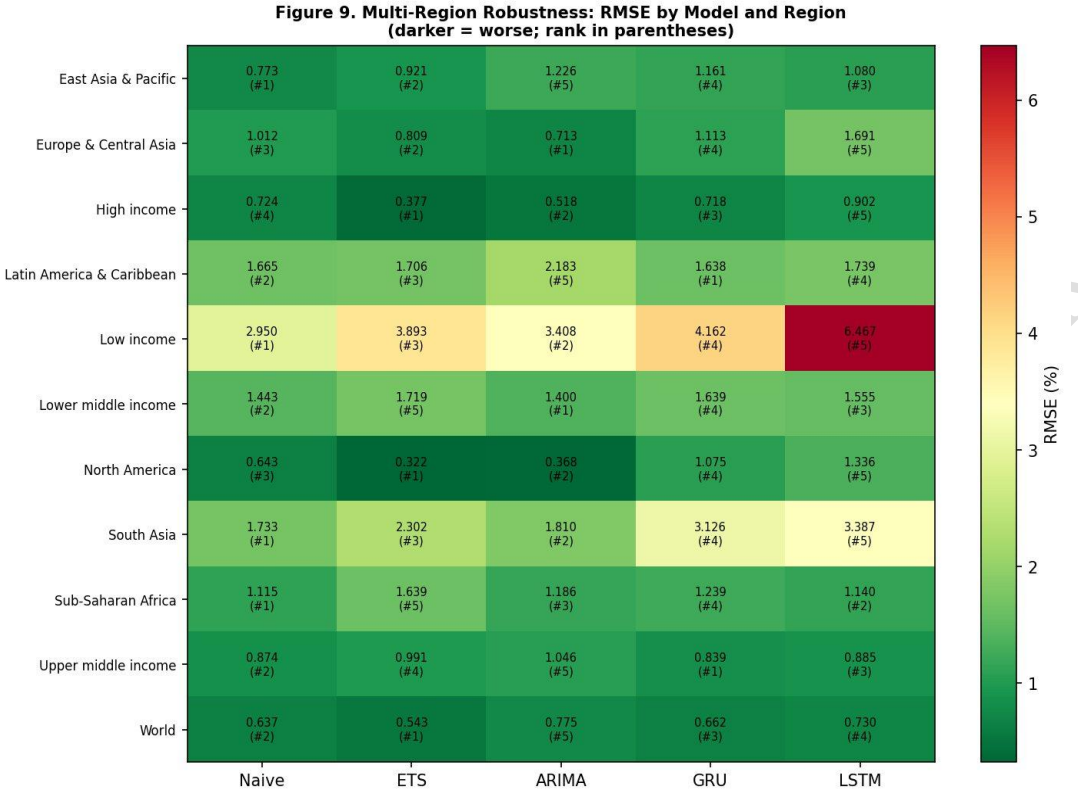

Figure 17. Multi-region robustness: RMSE by model and region with rank in parentheses.

10. Limitations

Several limitations of this study should be acknowledged. First, the dataset contains only 31 annual observations, which constrains the number of walk-forward windows and yields only 15 total test observations. This small test set limits the statistical power of the Diebold–Mariano tests and contributes to the wide confidence intervals observed in the bootstrap analysis.

Second, the study uses a univariate forecasting framework. Renewable energy consumption is influenced by numerous exogenous factors including oil prices, technology costs, policy interventions, economic growth, and carbon pricing mechanisms. Incorporating these as covariates in a multivariate framework could potentially improve forecast accuracy, though at the cost of additional data requirements and model complexity.

Third, the annual frequency of the data is both a constraint and a defining feature of the study. While higher-frequency data (monthly or quarterly) would provide more observations for model training, such data are not consistently available across all World Bank regional aggregates. The annual frequency is representative of the data regime that policymakers and international organisations actually work with for macro-level energy transition monitoring.

Fourth, the EG.FEC.RNEW.ZS indicator includes traditional biomass, complicating cross-regional interpretation. Regions with high renewable shares driven by biomass dependence (Sub-Saharan Africa, Low income) are not directly comparable to regions where renewable growth is driven by modern technologies (wind, solar, geothermal).

Fifth, the 20-year forecasts (2021–2040) are purely trend extrapolations and do not incorporate anticipated policy changes, technological breakthroughs, or macroeconomic shocks. They should be interpreted as conditional projections under the assumption that historical trends continue, not as predictions of likely outcomes.

11. Future Research

Several directions for future research emerge from this study. First, incorporating exogenous variables (oil prices, GDP growth, carbon prices, renewable energy investment flows) in a multivariate framework such as Vector Autoregression (VAR) or multivariate deep learning models could improve forecast accuracy while providing insight into causal mechanisms driving the energy transition.

Second, applying the benchmarking framework to higher-frequency datasets—monthly electricity generation from renewables, for example—would test whether the DL performance disadvantage persists with larger sample sizes. This would help delineate the critical threshold of data availability at which complex models begin to outperform parsimonious alternatives.

Third, Transformer-based architectures (Vaswani et al., 2017), including recent time series variants such as the Temporal Fusion Transformer (Lim et al., 2021) and PatchTST (Nie et al., 2023), were not evaluated in this study. While these models typically require even more data than RNNs, their attention mechanisms may offer advantages in capturing regime changes.

Fourth, regime-switching models that explicitly model the structural break—such as Markov-switching models or threshold autoregressive specifications—could provide a principled framework for handling the identified 2014 breakpoint within the forecasting model itself, rather than treating it as an external evaluation consideration.

Fifth, the Transition Velocity Index and beta-convergence analysis could be extended to country-level data, disaggregated by renewable technology type (wind, solar, hydro, biomass), enabling more granular policy insights.

12. Conclusion

This study provides a comprehensive benchmarking of 13 forecasting model families for global renewable energy consumption using the World Bank EG.FEC.RNEW.ZS indicator (1990–2020). Under a unified walk-forward cross-validation protocol that eliminates methodological asymmetries present in prior studies, Holt Linear Exponential Smoothing (ETS) emerges as the champion model with an RMSE of 0.543 and a Skill Score of +0.148

against the Naïve baseline. All three deep learning architectures (GRU, LSTM, N-BEATS), XGBoost, and ARIMA perform worse than the Naïve baseline on this dataset.

This outcome is explained by four quantifiable data-regime properties: a severely unfavourable observations-to-parameters ratio for DL models, a signal dominated by a single linear trend, a structural break at 2014 that compounds the small-sample challenge, and annual frequency that eliminates the high-frequency patterns in which DL methods excel. These findings do not imply that DL methods are inferior in general; rather, they demonstrate that model complexity must be calibrated to the information content of the available data.

The study also contributes novel energy transition analytics. The Transition Velocity Index reveals heterogeneous regional transition speeds, with Latin America and Caribbean showing the highest momentum and Lower middle income regions showing the weakest historical progress. Beta-convergence analysis confirms that regions with lower initial renewable shares are growing proportionally faster, consistent with technology diffusion theory. The SDG-7 gap analysis identifies substantial policy gaps in North America, Europe, and High income regions, where ETS 2030 forecasts fall well short of the 30% renewable target.

These findings carry practical implications for both the forecasting and energy policy communities. For forecasting practitioners, the results reinforce that rigorous evaluation protocols—walk-forward cross-validation, nested hyperparameter tuning, multi-seed robustness checks, and formal statistical tests—are essential for credible model comparison, especially with small datasets. For policymakers, the analysis suggests that current trends, if continued, will leave major economies short of SDG-7 targets, underscoring the need for accelerated policy intervention.

CRediT Author Statement

Shaon Biswas: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Paramita Roy:** Validation, Writing – review & editing, Domain expertise (energy systems).

Data Availability Statement

The dataset used in this study is publicly available from the World Bank Open Data repository (indicator EG.FEC.RNEW.ZS) at https://data.worldbank.org/indicator/EG.FEC.RNEW.ZS. The complete analysis code, including all model implementations, evaluation protocols, and figure generation scripts, is available at the corresponding author's GitHub repository (https://github.com/ShaonINT) upon publication.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge the World Bank for making the EG.FEC.RNEW.ZS indicator freely available under open data licence. Computational resources were provided by the Data Analytics and Intelligent Machines (DAIM) Research Centre at the University of Hull.

References

Ahmed, R., Sreeram, V., Mishra, Y., Arif, M.D., 2024. A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. Renewable and Sustainable Energy Reviews 124, 109792.

Assimakopoulos, V., Nikolopoulos, K., 2000. The theta model: a decomposition approach to forecasting. International Journal of Forecasting 16(4), 521–530.

Atiya, A.F., 2020. Why does forecast combination work so well? International Journal of Forecasting 36(1), 197–200.

Barro, R.J., Sala-i-Martin, X., 1992. Convergence. Journal of Political Economy 100(2), 223–251.

Biswas, S., Irshad, A., Roy, P., 2026. Global renewable energy consumption forecasting: A comparative benchmarking study of statistical, machine learning, and deep learning models. Computer Engineering and Intelligent Systems 17(1), 44–57. DOI: 10.7176/CEIS/17-1-05.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734.

De Oliveira, E.M., Cyrino Oliveira, F.L., 2018. Forecasting mid-long term electric energy consumption through bagging ARIMA and exponential smoothing methods. Energy 144, 776–788.

Deb, C., Zhang, F., Yang, J., Lee, S.E., Shah, K.W., 2017. A review on time series forecasting techniques for building energy consumption. Renewable and Sustainable Energy Reviews 74, 902–924.

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

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall/CRC.

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

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Computation 9(8), 1735–1780.

Hyndman, R.J., Athanasopoulos, G., 2021. Forecasting: Principles and Practice, 3rd edition. OTexts, Melbourne, Australia.

IEA, 2023. World Energy Outlook 2023. International Energy Agency, Paris.

IRENA, 2023. World Energy Transitions Outlook 2023. International Renewable Energy Agency, Abu Dhabi.

Karakurt, I., Aydin, G., 2023. Forecasting of energy-related CO2 emissions and energy demand using ARIMA and ETS models. Energy Sources, Part B: Economics, Planning, and Policy 18(1), 2175462.

Khan, I., Hou, F., Irfan, M., Zakari, A., Le, H.P., 2020. Does energy trilemma a driver of economic growth? The roles of energy use, population growth, and financial development. Renewable and Sustainable Energy Reviews 146, 111157.

Kong, W., Dong, Z.Y., Jia, Y., Hill, D.J., Xu, Y., Zhang, Y., 2019. Short-term residential load forecasting based on LSTM recurrent neural network. IEEE Transactions on Smart Grid 10(1), 841–851.

Lago, J., Marcjasz, G., De Schutter, B., Weron, R., 2021. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. Applied Energy 293, 116983.

Lim, B., Arık, S.Ö., Loeff, N., Pfister, T., 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting 37(4), 1748–1764.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2018. Statistical and Machine Learning forecasting methods: Concerns and ways forward. PLoS ONE 13(3), e0194889.

Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J., 2023. A time series is worth 64 words: Long-term forecasting with transformers. In: International Conference on Learning Representations (ICLR).

Oreshkin, B.N., Carpov, D., Chapados, N., Bengio, Y., 2020. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In: International Conference on Learning Representations (ICLR).

Petropoulos, F., Apiletti, D., Assimakopoulos, V., et al., 2022. Forecasting: theory and practice. International Journal of Forecasting 38(3), 845–1130.

REN21, 2023. Renewables 2023 Global Status Report. REN21 Secretariat, Paris.

Shahbaz, M., Raghutla, C., Chittedi, K.R., Jiao, Z., Vo, X.V., 2020. The effect of renewable energy consumption on economic growth: Evidence from the renewable energy country attractive index. Energy 207, 118162.

Smyl, S., 2020. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. International Journal of Forecasting 36(1), 75–85.

Tashman, L.J., 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. International Journal of Forecasting 16(4), 437–450.

Taylor, S.J., Letham, B., 2018. Forecasting at scale. The American Statistician 72(1), 37–45.

UNFCCC, 2015. Paris Agreement. United Nations Framework Convention on Climate Change.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NeurIPS), pp. 5998–6008.

Wang, H., Lei, Z., Zhang, X., Zhou, B., Peng, J., 2019. A review of deep learning for renewable energy forecasting. Energy Conversion and Management 198, 111799.

World Bank, 2024. World Development Indicators: Renewable energy consumption (% of total final energy consumption). World Bank Open Data. Available at: https://data.worldbank.org/indicator/EG.FEC.RNEW.ZS.

Zhang, G.P., 2003. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50, 159–175.

Ziel, F., Weron, R., 2018. Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks. Energy Economics 70, 396–420.

Appendix

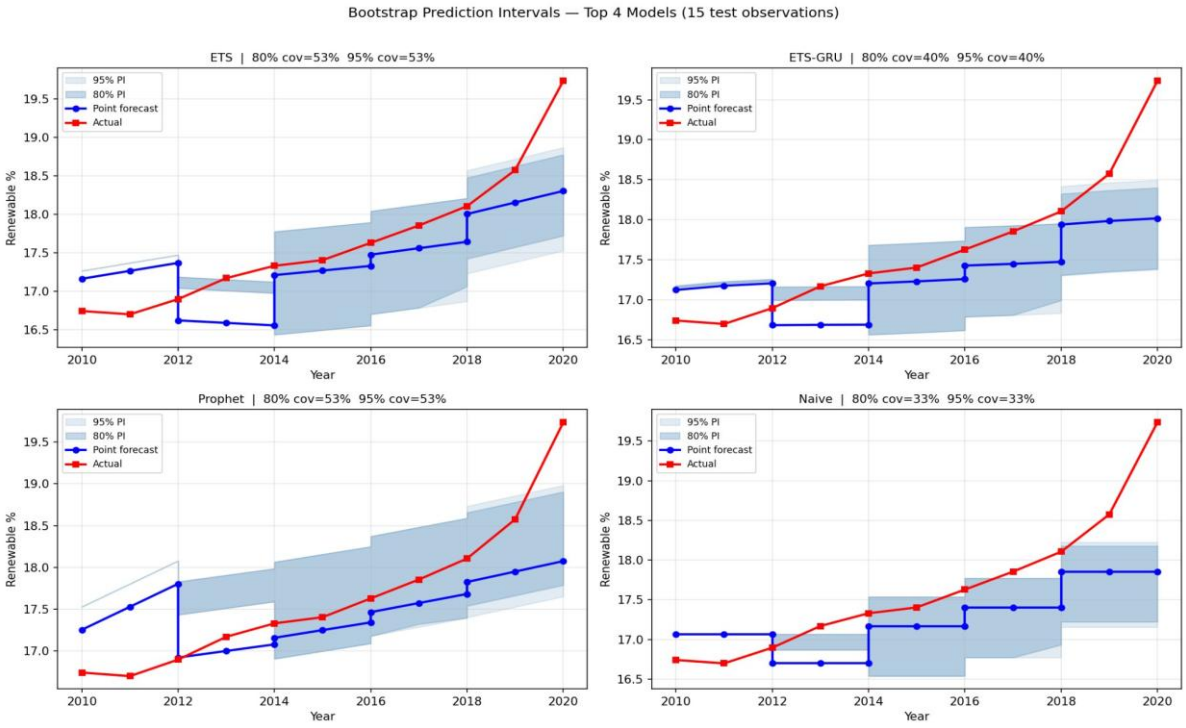

Figure A1. Bootstrap prediction intervals (80% and 95%) for top 4 models across 15 test observations.