

Short-term predictions for country-level energy demand with top-down and bottom-up stacking

Haoyang Li, Erkal Ersoy, Boulis Ibrahim
and Mark E. Schaffer

Edinburgh Business School, Heriot-Watt University, UK

Abstract

Purpose - This paper examines whether stacking (Wolpert, 1992) improves short-term forecasts of global oil consumption and electricity demand during periods of major shocks, including COVID-19, geopolitical conflicts, and trade disruptions. Given diverging trends between OECD and non-OECD economies, we evaluate forecast performance separately across these regions.

Design/methodology/approach - We combine XGBoost, random forest, elastic net, multi-layer perceptron, and a unit fixed-effects model within a stacking framework using a non-negative OLS meta-learner. We implement both a top-down approach, which models aggregate demand directly, and a bottom-up approach, which models disaggregated components and aggregates predictions, to assess the value of decomposition.

Findings - We find that stacking provides a robust benchmark, consistently matching or outperforming the best individual learner across most specifications. In the top-down setting, stacking generally delivers the strongest performance, particularly for global and non-OECD oil demand and global electricity demand. Bottom-up stacking improves accuracy for most oil sub-aggregates and performs comparably to the best base learner for electricity demand by source. Comparing strategies, bottom-up stacking performs better for electricity demand, while both approaches perform similarly for oil demand. Results by demand size suggest that retransformation bias from the log scale (Duan, 1983) and cross-country heterogeneity are particularly important for large economies. Overall, while no single model dominates across all settings, stacking provides a simple and reliable approach to improving forecast accuracy.

Originality/value - Our contribution is threefold. First, we provide a systematic comparison of stacking methods for forecasting global energy demand across OECD and non-OECD economies. Second, we evaluate the relative performance of top-down and bottom-up strategies within a common framework. Third, we examine how forecast performance varies with the scale of energy demand, highlighting the role of retransformation bias (Duan, 1983) and cross-country heterogeneity.

Keywords Stacking, Energy demand, Machine learning, Top-down, Bottom-up, Prediction.

Paper type Research paper

1. Introduction

In recent years, sluggish economic growth associated with the COVID-19 pandemic, geopolitical tensions, and trade disputes, together with a gradual shift away from oil use in transport and power generation, have contributed to slower growth in oil consumption and fossil-fuel-based electricity demand. However, these trends have unfolded differently in advanced and emerging economies. Figure 1 shows the historical demand for oil by fuel category and for electricity by generation source for the 170 economies in our sample. Following the post-pandemic rebound, energy use in OECD countries has largely leveled off, whereas non-OECD countries—driven predominantly by Asian economies—continue to experience robust growth (IEA, 2025). This divergence indicates that future energy demand will increasingly be concentrated in developing economies and will also become more difficult to extrapolate from historical patterns.

JEL Classification — Q47, C53, C52

© Haoyang Li, Erkal Ersoy, Boulis Ibrahim and Mark E. Schaffer. Publish in Asian Journal of Applied Financial Econometrics.

Accurate forecasts of energy demand are crucial for government planning, energy policy design, investment decisions, and trade (Bhattacharyya and Timilsina, 2009). Although there is a long-standing and largely successful record of energy demand forecasting in industrialised nations, these approaches have not transferred effectively to developing countries (Urban et al., 2007). This limitation reflects a broader challenge: no single method delivers consistently accurate predictions across all contexts. Conventional statistical models for energy demand (Hunt and Ninomiya, 2003; Dees et al., 2007; Fatima et al., 2019) remain valuable, but often struggle to capture complex nonlinear relationships and dynamic behaviours. In response, machine learning approaches have increasingly been applied to energy forecasting due to their adaptability and ability to uncover latent patterns (Yu et al., 2019; Aldabbagh et al., 2024; Sharma et al., 2024); however, no individual algorithm performs robustly across all settings.

In addition to the “top-down” approach, which directly models and forecasts aggregate demand, a complementary “bottom-up” strategy forecasts disaggregated sub-series and aggregates them to obtain total demand (Widiarta et al., 2009; Sbrana and Silvestrini, 2013; Spiliotis et al., 2020). The top-down approach generally exhibits greater stability when constituent sub-aggregate series are noisy or strongly correlated, as it avoids the accumulation of forecast errors across components. By contrast, bottom-up methods can outperform top-down approaches when underlying trends and patterns differ substantially across components (Widiarta et al., 2009), or when structural changes occur, such as shifts in the product mix.

In this paper, we combine top-down and bottom-up strategies to forecast oil and electricity demand with greater precision and transparency. To limit the accumulation of small forecast errors, oil demand is disaggregated into just five principal categories: motor gasoline (excluding ethanol), gas/diesel, fuel oil, jet kerosene, and a residual “other” group. Motor gasoline and gas/diesel are primarily associated with road transport and freight, fuel oil is linked to shipping and certain industrial uses, and jet kerosene captures aviation and international tourism. Electricity generation is similarly streamlined into three source-based components: fossil (coal, gas, oil), hydro, and an “other” category. Fossil generation responds mainly to fuel prices, carbon policies, and installed capacity; Hydro output depends on water availability and climate conditions; and the remaining category is dominated by nuclear and renewables such as solar and wind. Figure 1b illustrates a clear and gradual shift away from fossil fuels towards low-carbon energy in OECD countries, alongside a more rapid expansion of green energy deployment in non-OECD economies.

For forecasting we employ the ensembling method “stacked generalisation” (Wolpert, 1992), or “stacking”, which has gained popularity in predictive applications (Khairalla et al., 2018; Ahrens et al., 2022; Reddy et al., 2022; Ersoy et al., 2023; Ahrens et al., 2025). Rather than selecting a single learner based on out-of-sample performance, stacking trains a meta-learner to combine predictions from a diverse set of base learners using different algorithms, with the aim of improving predictive accuracy. Van der Laan et al. (2007) showed theoretically that stacking with cross-validation performs asymptotically as well as the best base learner and achieves a parametric convergence rate when the true model can be approximated by a linear combination of base learners. However, stacking can also be subject to the well-known “forecast combination puzzle,” whereby a simple equal-weight average can outperform optimally weighted combinations in practice. This outcome may arise from limited sample sizes, poor learner selection, or instability in weight estimation due to highly correlated forecasts (Timmermann, 2006; Blanc and Setzer, 2020). To address these issues, we consider five base learners chosen for their methodological diversity and suitability for the data: a fixed effects model (FE), XGBoost, Elastic Net, random forest, and Multi-layer perceptron (MLP) neural network. We show that this design improves the forecasting performance of stacking while reducing the risk of unstable weights. A constrained OLS meta-learner with non-negative coefficients that sum to unity (Breiman, 1996) is used, and a simple equal-weight average serves as a benchmark.

Our contribution is threefold. First, we provide a systematic comparison of stacking methods for forecasting global energy demand across OECD and non-OECD economies. Second, we evaluate the relative performance of top-down and bottom-up strategies within a common framework. Third, we examine how forecast performance varies with the scale of energy demand, highlighting the role of retransformation bias (Duan, 1983) and cross-country heterogeneity.

This paper presents analytical results for both top-down and bottom-up strategies in forecasting global and regional energy demand using stacking. Across most specifications, both stacking estimators deliver robust improvements over, or perform comparably to, the best individual base learner. The remainder of the paper is organised as follows. Section 2 introduces

the data, Section 3 outlines the methodological framework, including the base learners and meta-learner, Section 4 presents the empirical results for the top-down and bottom-up approaches, and Section 5 concludes. The appendices in the supplementary online material report additional results in detail.

2. Data

2.1 Data description

In addition to forecasting global energy demand for the 170 countries in our sample (see the online supplement for details), this paper also examines differences between OECD and non-OECD economies. Oil demand is disaggregated into five primary product categories: motor gasoline, gas/diesel, fuel oil, jet kerosene, and other products. Electricity generation is similarly disaggregated by source into fossil fuels, hydro, and other sources. The annual data are drawn from the 2025 Statistical Review of World Energy, published by the Energy Institute. The sample period spans 1970-2023 for oil consumption and 1985-2023 for electricity demand. The set of predictors includes weather conditions, macro variables, industry shares, and oil prices. Because energy demand is sensitive to temperature (Morris, 2001; Fung et al., 2006; Manderson and Considine, 2025), the number of ice days and summer days per year are used as proxies for heating and cooling demand. Economic activity is a primary driver of energy demand, and real GDP is therefore used to link energy demand to real activity. Population captures the scale of residential demand and transport activity. The industry share of GDP is included to proxy economic structure, as industrial activity is typically more energy-intensive than agriculture or services. Finally, the WTI crude oil price - a global benchmark - is incorporated to capture demand-price elasticities of both oil and electricity generation from fossil fuels. Table 1 summarises the data structure and all variables used in the analysis.

2.2 Data processing

Natural logarithms are applied to total oil consumption, total electricity demand, GDP, population, and oil prices. For sub-aggregate oil and electricity components, the inverse hyperbolic sine transformation is used to accommodate zero values; the inverse of the median is adopted as the scaling parameter to ensure scale invariance. Ice days and summer days are divided by 365 to represent annual shares with extreme cold and heat.

The final 30% of the sample period is reserved as the testing set, on which all models are evaluated. Because some learners are sensitive to variable scale (e.g., Elastic Net), all regressors are normalised using statistics from the training sample, and the resulting scalers are applied to the testing data. Training of the stacking estimator relies on cross-validation (CV); accordingly, a time-series CV scheme is used to prevent future data leakage. Specifically, we use an expanding-window framework with four folds, where each validation fold contains five observations. After each fold is predicted, it is appended to the training set for the next fold. Within each training fold, an additional inner CV procedure tunes hyperparameters for machine-learning models. Predictions from base learners across the full validation sample serve as inputs to the meta-learner. Finally, all base learners are refitted to the entire training sample, and forecasts from each learner, the stacking estimator, and the sample average benchmark are generated for the testing sample.

3. Methodology

3.1 Stacking

There is no universal forecast method that dominates across all real world applications. Unlike model selection approaches, where a single winner model is chosen, forecast combination methods integrate information from multiple sources and have been shown to improve predictive accuracy (Wang et al., 2023). “Stacking” or “stacked generalisation” (Wolpert, 1992) adaptively combines models using a meta-learner trained on the predictions of base learners. In particular, consider a library with M learners $\{f_1, f_2, \dots, f_M\}$, the stacking estimator is a weighted combination

$$\hat{f}_S(X) = \sum_{m=1}^M \hat{w}_m \hat{f}_m(X)$$

where weights $\hat{\mathbf{w}} = (\hat{w}_1, \hat{w}_2, \dots, \hat{w}_M)'$ are estimated using out-of-sample cross-validation to avoid data leakage and overfitting. Van der Laan et al. (2007) subsequently provided a theoretical foundation for the validity of stacking. Let $R(f) = \mathbb{E}[L(Y, f(X))]$ be the expectation of the loss of a model f , then

$$R(\hat{f}_s) \leq \inf_{w \in \mathcal{W}} C_0 R\left(\sum_{m=1}^M w_m \hat{f}_m\right) + o_p(1), \quad (1)$$

for any constant $C_0 > 1$ where $\mathcal{W} \subseteq \mathbb{R}^M$. This inequality implies that stacking performs asymptotically as well as the best convex combination of base learners. Let $\mathcal{F}$ be the model space induced by stacking. If the true model lies in $\mathcal{F}$, the stacking estimator converges to the truth at a parametric rate; otherwise, it converges to the best misspecified approximation within $\mathcal{F}$.

Although there is no restriction on the choice of meta-learner, constrained OLS with non-negative coefficients that sum to unity (Breiman, 1996) is widely used, as it yields a convex combination of base learners without extreme extrapolation and with interpretable weights. We adopt constrained OLS as the meta-learner in this paper. In addition, because stacking can be subject to the forecast combination puzzle, a simple equal-weight average of the base learners is included as a benchmark.

3.2 Base learners

Stacking relies on both the diversity and accuracy of base learners to improve prediction. The role of diversity is well established in ensemble methods (Krogh and Vedelsby, 1994; Opitz and Shavlik, 1995; Kuncheva and Whitaker, 2003; Atiya, 2020). Learners based on different algorithms capture distinct aspects of the data and therefore tend to make less correlated errors, increasing the potential gains from combinations. By contrast, similar learners add little new information and generate highly correlated errors, leading to limited gains from stacking and potentially unstable weights. Recall that stacking performs asymptotically at least as well as the best convex combination of base learners, which is effective only if the convex combination space is sufficiently rich to contain, or closely approximate, the true data-generating process. Although stacking can theoretically down-weight poorly performing learners, in finite samples their inclusion increases uncertainty in weight estimation and may weaken performance. Consequently, many studies select a subset of strong and diverse learners prior to model combination (Zhou et al., 2002; Zhang et al., 2006; Budescu and Chen, 2015; Qian et al., 2015). In this paper, five diverse base learners suited to panel data are employed.

Unit fixed effects model (FE) Consider a panel data with N units observed over T periods. For each dependent variable, the model is

$$Y_{it} = \alpha_i + \rho Y_{i,t-1} + X_{i,t-1}'\beta + u_{it}, \quad i = 1, 2, \dots, N, \quad t = 2, \dots, T, \quad (2)$$

where α_i captures unit-specific time-invariant effects and $\mathbf{X}$ includes regressors listed in Table 1. Time fixed effects absorb unobserved common shocks but are not available for ex-ante forecasting and therefore excluded. Persistent common shocks and dynamics are instead partially captured by the lagged structure of the model. Accordingly, a dynamic specification with unit fixed effects is adopted as the benchmark forecasting model. Moreover, given the large cross-section ($N=170$) and relatively short time dimension ($T=50$), a fully heterogeneous specification with country-specific coefficients β_i would require estimating a large number of additional parameters, increasing variance and potentially degrading out-of-sample performance. The assumption of homogeneous coefficients therefore reflects a standard bias-variance trade-off.

Extreme Gradient Boosting (XGBoost) Boosting constructs an additive model by sequentially adding weak learners, each fitted to the residuals left by the previous learner, to improve predictive accuracy (Friedman, 2001). However, without constraints, boosting methods can overfit the data. XGBoost addresses this issue by introducing an explicit regularisation term (Chen and Guestrin, 2016). Specifically, let $f_j \in \mathcal{F}$ denote the weak learner at iteration j ; the final prediction after J iterations is

$$\hat{f}_{XGB}(X) = \sum_{j=1}^J \hat{f}_j(X). \quad (3)$$

The prediction after iteration $j-1$ is

$$\hat{Y}^{j-1} = \sum_{j=1}^{j-1} \hat{f}_j(X). \quad (4)$$

The j -th learner then solves

$$\min_{f_j \in \mathcal{F}} L(Y, \hat{Y}^{j-1} + f_j(X)) + \Omega(f_j), \quad (5)$$

where $L(\cdot)$ denotes the loss function and $\Omega(f_j)$ is a regularisation term penalising deep and complex trees. This iterative updating allows XGBoost to capture hidden nonlinearities and cross-unit heterogeneity while mitigating overfitting.

Elastic Net While the Lasso (Tibshirani, 1996) assumes a sparse model and shrinks some coefficients exactly to zero, ridge regression (Hoerl and Kennard, 1970) does not require sparsity and can outperform Lasso when true effects are dense and weak. Without prior knowledge of the true model, Zou and Hastie (2005) proposed a data-driven approach that combines the two methods. This resulting “Elastic Net” estimator solves

$$\hat{\beta}_{\text{EN}} = \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^n (Y_i - X_i^\top \beta)^2 + \lambda (\alpha \|\beta\|_1 + (1-\alpha) \|\beta\|_2^2), \quad (6)$$

where $\|\beta\|_1 = \sum |\beta_j|$ is the lasso penalty and $\|\beta\|_2^2 = \sum_{j=1}^p \beta_j^2$ is the ridge penalty, and $0 \leq \alpha \leq 1$ controls the relative weights.

Random forest random forest (Breiman, 2001) is an ensemble of decision trees constructed on bootstrap replications of the original sample, with random subsets of features selected at each split to reduce the dominance of a few predictors across trees and thereby improve out-of-sample generalisation. Specifically, for each tree $b=1,2,\dots,B$, a bootstrap sample is drawn with replacement from the observed data, and a regression tree is grown by randomly selecting a subset of regressors at each node and splitting the sample to minimise a given impurity or loss function. Let $T_b(\cdot)$ denote the prediction function of the b -th tree. For a new regressor vector X , the random forest prediction is

$$\hat{f}_{\text{RF}}(X) = \frac{1}{B} \sum_{b=1}^B T_b(X). \quad (7)$$

Averaging substantially reduces the instability of individual trees decorrelated through random sampling and feature selection. Random forest thus provides a non-parametric approach for approximating nonlinear relationships and high-order interactions through recursively partitioning the covariate space.

Multi-layer Perceptron (MLP) The MLP approach is a feed-forward neural network consisting of an input layer, one or more hidden layers, and an output layer, with full connectivity between consecutive layers. With nonlinear activation functions in the hidden layers, the MLP can learn highly flexible mappings from inputs to outputs (Hornik et al., 1989; Goodfellow, 2016). To illustrate, consider an MLP with a single hidden layer of size H and activation function $\sigma(\cdot)$. The regression function is

$$f_{\text{MLP}}(X, \theta) = \sum_{h=1}^H v_h \sigma(w_h' X + b_h) + c, \quad (8)$$

where the parameter vector $\theta = (\{v_h, w_h, b_h\}_{h=1}^H, c)$ contains all weights and biases. Here, w_h and b_h are the weights and biases of unit h , v_h are output weights, and c is the global bias. Parameters are estimated by minimising

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{2n} \sum_{i=1}^n (Y_i - f_{\text{MLP}}(X_i, \theta))^2 + \lambda R(\theta), \quad (9)$$

where n is the sample size, $R(\theta)$ is a regularisation term on θ , and λ controls its strength. Optimisation is typically performed using stochastic gradient-based methods via backpropagation. Despite its flexibility in capturing nonlinearities and interactions, increasing hidden layer depth and width enlarges the parameter space and raises overfitting risk, particularly in small samples. Accordingly, the number of layers and hidden units must be selected carefully, for example via cross-validation, to balance flexibility and generalisation.

Hyperparameter grids for all machine learning algorithms are summarised in Table 2.

3.3 Assessing performance

We report predictive performance of all learners using root mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). We also report direction accuracy (DA) in % in selected cases.

As a useful diagnostic, we also report Diebold-Mariano (DM) test statistics (Diebold and

Mariano, 2002) for all learners for oil consumption, using a simple average of all base learners as the benchmark. A positive DM statistic indicates that the competing model yields a smaller loss than the benchmark, while a negative value implies a superior benchmark performance. Under the null hypothesis of equal predictive accuracy, the DM statistic is asymptotically standard normal; accordingly, absolute values exceeding 1.96 indicate statistically significant differences in predictive accuracy. Below we discuss the DM test results and report DM statistics graphically; detailed tables with the full set of DM results and discussion are available in the online supplement.

4. Results

Forecast performance on oil consumption and electricity demand across all learners is summarised below. We first discuss the top-down and bottom-up stacking estimators separately, and then compare level predictions. Figure 1 shows that oil and electricity demand in OECD countries plateaued by the 2000s, while the demand in the remaining economies has continued to grow steadily. Given the potential differences in patterns between developed and developing economies, stacking performance is evaluated separately for global, OECD, and non-OECD samples. Moreover, forecasting patterns may differ across economies with varying energy demand scales. Accordingly, all economies are classified into “small”, “medium”, and “large” groups of equal size based on their ranking by average energy demand.

4.1 Top-down stacking

Figures 2 and 3 report Diebold-Mariano test statistics for all learners for oil consumption on oil and electricity demand over time; the benchmark is the average of all base learners, as noted above. The stacking estimator (solid red line) clearly dominates across most regions for both oil and electricity demand, followed by the elastic net and unit fixed effects models. The improvement from stacking is more pronounced for oil demand in global and non-OECD economies, while gains for electricity demand are comparatively weaker. This finding is confirmed by the detailed results for the DM test (see the online supplement): stacking outperforms all base learners and the simple average across all regions.

Table 3 shows the RMSE, MAE and MAPE for all learners for oil consumption including the simple average. For global economies, stacking reduces the RMSE and MAE by 2.42% and 3.86%, respectively, relative to the best base learner. For OECD economies, stacking achieves comparable accuracy to the best base learner, while for non-OECD economies it reduces RMSE and MAE by 1.55% and 1.47%, respectively. Taken together, these results indicate that stacking performs best across all regions, although improvements for OECD economies are marginal.

As shown in Table 4, stacking reduces both RMSE and MAE by 3.4% relative to the best base learner, random forest, for global electricity demand. For the OECD and non-OECD regions, stacking slightly underperforms the best base learner in terms of squared loss, with RMSE higher by 1.3% and 2.6%, respectively, while reducing MAE by 2.5% and 4.4%.

In sum, the stacking estimator improves accuracy in direct total energy demand forecasting, although effects vary across energy types and regions. For oil demand, gains are driven mainly by improvements in non-OECD economies. For electricity demand, there is an overall improvement at the global level, but the effect is more modest within individual sub-regions. Overall, the stacking estimator is robust, consistently ranking as the best or second-best predictor across all samples.

4.2 Bottom-up stacking

This section summarises forecast performance by component for all learners for oil consumption. Predictions are generated for each component, transformed back to levels, and then aggregated. Figures 4 and 5 present DM test statistics for forecasts of the five oil sub-aggregates and three electricity generation sources, together with the corresponding stacking weights on base learners, for global economies. Corresponding figures for the OECD and non-OECD regions are reported in Figures 6-7 and Figures 8-9, respectively.

The stacking estimator clearly delivers the best performance across all components of both oil and electricity demand, with similar patterns observed for OECD and non-OECD economies. Stacking weights on base learners vary modestly across components, with elastic net consistently receiving the largest weight for global and non-OECD economies. By contrast, for OECD economies no single base learner is consistently favoured (see Figures 6 and 8).

Table 5 reports detailed performance metrics for all learners for oil consumption across oil demand sub-aggregates in the global, OECD, and non-OECD samples. For global economies,

relative to the best base learner, stacking reduces RMSE for gas/diesel, motor gasoline, fuel oil, jet kerosene and other oil products by 0.8%, 1.4%, 2.7%, 0.05%, and 1.5%, respectively, and reduces MAE by 2.2%, 1%, 4.5%, 1.4%, and 2.9%. Results for the non-OECD region are similar. By contrast, stacking does not improve upon base learners in the OECD region. In particular, RMSEs are 1.1%, 2.1%, 5.8%, 0.9%, and 2.8% higher than the corresponding best base learner for the five sub-aggregates, while MAEs are higher by 5.2%, 1.6%, 6.3%, and 5% for gas/diesel, motor gasoline, fuel oil, and other oil products. Notably, when stacking does not improve performance, it is almost always outperformed only by the best base learner across components.

Detailed performance comparisons for aggregated electricity demand are reported in Table 6. The elastic net learner is the best performer in predicting global electricity demand from fossil fuels and other sources (mainly renewable and nuclear), closely followed by stacking, with RMSE higher by 0.5% and 0.3% and MAE higher by 1% and 0.3%, respectively. For hydro generation, stacking achieves the lowest squared loss, while random forest yields the lowest mean absolute loss, which is 2.2% lower than stacking. For OECD electricity demand by source, the elastic net learner performs best for fossil fuels, with RMSE and MAE lower than stacking by 0.4% and 0.2%, respectively. Stacking outperforms the best base learner for hydro and other sources, with RMSE reductions of 0.2% and 0.1%. Results for the non-OECD region are similar, with only small quantitative differences between stacking and the best base learner.

4.3 Comparison of two stacking methods

The results above show that stacking is a robust predictor, in the sense that it performs at least as well as the best base learner in almost all cases and occasionally delivers moderate accuracy gains, while no single base learner is universally dominant. In this section, we examine whether gains from forecasting disaggregated components outweigh losses from error accumulation.

Figure 11 presents DM test comparisons for top-down and bottom-up stacking estimators applied to level predictions of energy demand, using the top-down simple average as the benchmark. Both estimators exhibit similar patterns for oil demand across all regions. However, the bottom-up approach shows clearer improvements over the top-down approach for electricity demand in global and non-OECD economies.

Table 7 reports the detailed performance evaluation for both estimators. The bottom-up approach has smaller RMSE for global and OECD region but larger RMSE in non-OECD region than the top-down approach on oil demand prediction. However, the top-down approach shows significantly larger RMSE for global and non-OECD region in terms of electricity demand. This might be caused by the retransformation bias from the natural logarithm back to level for large economies. As evidence, we note that top-down stacking has a significantly smaller MAPE versus the bottom-up approach.

4.4 Subsets by demand size

The different patterns between RMSE and MAPE of level predictions in Table 7 suggest that the heterogeneity in size across countries matters. This is expected since predictions are obtained on the log scale and then exponentiated to recover forecasts in levels, which means that the objective function is approximately minimising squared loss percentage errors. In addition, the naive retransformation from natural logarithm to level leads to biased estimates. Duan (1983) proposed a smearing estimator as a nonparametric correction relying on the estimated residuals. However, with short panels and strong cross-country heteroskedasticity, the correction can result in more noise across countries. Therefore, we proceed with the naive conversion knowing that there exists a retransformation bias. We divide global economies into small, medium, and large size of demand subsets according to ranked country-level average demand for oil and electricity separately to investigate how stacking performs differently for economies of different sizes with a global prediction.

As shown in Figure 10, the top-down and bottom-up stacking estimators are comparable for small- and medium-sized economies in terms of DM test statistics against a simple average over time. But the top-down approach is outperformed by the bottom-up approach for electricity demand in larger economies.

Table 8 provides detailed information on the forecast performance comparisons of all learners on oil demand for subsets of different size. Both stacking estimators have comparable smaller RMSE than all base learners, while the top-down stacking has the smallest MAE and MAPE for small economies. But the improvement from stacking is tiny compared to the best base learner. For

medium-sized economies, the bottom-up stacking estimator achieves the smallest RMSE, MAE, and MAPE across all learners while the bottom-up XGBoost has the highest direction accuracy. The situation is more complicated for large economies where no learner dominates. Specifically, the bottom-up elastic net learner has the smallest RMSE, the bottom-up FE model has the smallest MAE, bottom-up stacking has the smallest MAPE, and the top-down FE model has the highest direction accuracy.

Detailed forecast performance for electricity demand is listed in Table 9. The top-down random forest has the smallest RMSE and second smallest MAE for small economies, which are 0.3% lower and 0.7% higher than the bottom-up stacking estimator respectively. At the same time, the top-down stacking has the smallest MAPE and the elastic net has the highest direction accuracy. For medium economies, the top-down stacking estimator reduces the RMSE, MAE, and MAPE by 6.3%, 2.3%, and 0.5% relative to the best base learner. The bottom-up elastic net learner dominates for large economies in terms of RMSE and MAE, followed by the FE model. The two stacking learners only achieve comparable results in terms of MAPE.

5. Conclusion

This paper applies the ensemble learning method of stacking (Wolpert, 1992) to forecast global oil consumption and electricity demand during a period of structural change. Over this period, energy demand growth has plateaued in OECD economies while shifting increasingly towards emerging countries. We therefore explicitly examine differences between OECD and non-OECD regions. Recognising that no single forecasting model performs well across all regions and energy types, we combine traditional econometric panel models, such as fixed effects specifications, with modern machine-learning algorithms through stacking, which trains a meta-learner to aggregate predictions from multiple base learners. Because different components of oil and electricity demand may respond heterogeneously to shocks and changes in predictors, we consider both a top-down approach that directly models aggregate demand and a bottom-up approach that models a small number of economically meaningful components and then re-aggregates them. The analysis focuses on whether carefully designed stacking can deliver robust gains over individual learners, overcome the forecast combination puzzle, and how these gains vary across energy type, region, and demand scale.

Overall, the results show that stacking is a robust predictor for global energy demand forecasting. In the top-down setting, stacking almost always ranks as the best or second-best model in terms of squared and absolute loss, and it consistently outperforms a simple equal-weight average of base learners. For oil consumption, most improvements originate from non-OECD economies, where demand growth remains strong and more volatile. For electricity demand, stacking delivers clear gains at the global level, although its advantage in squared loss diminishes when OECD and non-OECD regions are considered separately. Across all specifications, the simple average remains a strong benchmark; nevertheless, stacking with a constrained non-negative OLS meta-learner and a small, diverse set of base models achieves moderate and robust improvements overall.

Results from the bottom-up approach confirm that modelling sub-aggregates can be beneficial when components exhibit heterogeneous dynamics. Stacking dominates or closely tracks the best base learner across most oil products and electricity-generation sources, particularly for global and non-OECD economies. For aggregated level forecasts, bottom-up stacking outperforms all base learners in terms of the share of countries where the simple average yields higher squared loss.

Comparing the two strategies, top-down and bottom-up stacking perform similarly for oil demand in levels across regions, whereas bottom-up stacking clearly dominates in electricity-demand forecasting for global and non-OECD economies. Country-level Diebold–Mariano tests show that the bottom-up approach outperforms the simple average in a larger share of economies than the top-down approach for electricity demand. However, the divergence between RMSE and MAPE across the two approaches suggests that top-down stacking performs better for small economies but worse for large ones. This pattern is consistent with retransformation bias from the log scale (Duan, 1983) and strong cross-country heterogeneity in demand levels: naïve log-to-level conversion disproportionately penalises large values in terms of RMSE, even when percentage errors remain comparable.

We further classify global economies into three groups based on average energy demand. For small and medium-sized economies, both stacking approaches perform comparably and deliver small but systematic gains over the best base learner for both oil and electricity demand. For large economies, no single learner dominates across all metrics, with the bottom-up elastic

net performing most competitively, followed by the fixed-effects model. This highlights the greater influence of retransformation bias and heterogeneity for large economies, which should be considered when applying stacking.

From a methodological perspective, the results indicate that stacking is a useful and robust estimator for improving energy-demand forecasts. A diverse set of base learners expands the effective model space without introducing excessive uncertainty in weight estimation, while non-negativity and sum-to-unity constraints ensure that stacking remains a stable linear combination of base models. Empirically, the findings underscore clear differences between developed and developing economies, suggesting that historical experience from OECD countries cannot be directly extrapolated to emerging-economy demand growth.

Several limitations remain. All forecasts are point predictions evaluated using standard loss functions; extending the framework to probabilistic forecasting and risk measures would enhance its relevance for investment and supply-chain planning. Retransformation bias could also be addressed using smearing estimators, particularly for large economies with sufficient data. Moreover, the trade-off between gains from disaggregation and losses from error accumulation warrants further investigation using alternative component structures. Finally, beyond bottom-up aggregation, time–frequency decomposition methods such as wavelet transforms or empirical mode decomposition could be explored, although such approaches require higher-frequency data.

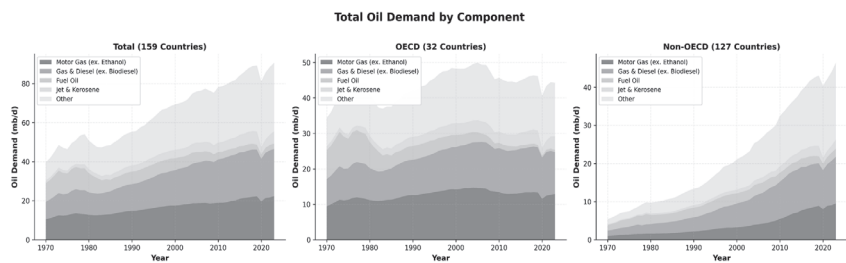
In summary, this paper shows that while no single base learner is universally optimal, stacking provides a simple and robust approach that consistently matches or improves upon the best individual model across energy types, regions, and country sizes. As global energy demand increasingly shifts towards developing economies and becomes harder to infer from past patterns, such robustness represents a valuable benchmark for applied energy forecasting.

References

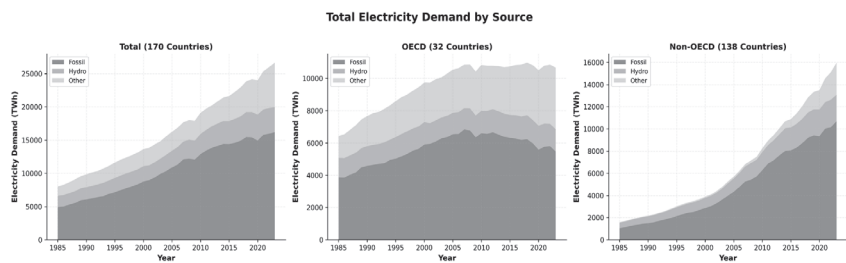
- Ahrens, A., E. Ersoy, V. Iakovlev, H. Li, and M. E. Schaffer (2022). An introduction to stacking regression for economists. In *International Conference of the Thailand Econometrics Society*, pp. 7–29. Springer.
- Ahrens, A., C. B. Hansen, M. E. Schaffer, and T. Wiemann (2025). Model averaging and double machine learning. *Journal of Applied Econometrics* 40(3), 249–269.
- Aldabbagh, A. M., A. Economou, and C. Chariton (2024). Forecasting global oil demand: Application of machine learning techniques. *Technical report*, OIES Paper.
- Atiya, A. F. (2020). Why does forecast combination work so well? *International Journal of Forecasting* 36(1), 197–200.
- Bhattacharyya, S. C. and G. R. Timilsina (2009). Energy demand models for policy formulation: a comparative study of energy demand models. *World Bank Policy Research Working Paper* (4866).
- Blanc, S. M. and T. Setzer (2020). Bias–variance trade-off and shrinkage of weights in forecast combination. *Management science* 66(12), 5720–5737.
- Breiman, L. (1996). Stacked regressions. *Machine learning* 24(1), 49–64.
- Breiman, L. (2001). Random forests. *Machine learning* 45(1), 5–32.
- Budescu, D. V. and E. Chen (2015). Identifying expertise to extract the wisdom of crowds. *Management science* 61(2), 267–280.
- Chen, T. and C. Guestrin (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.
- Dees, S., P. Karadeloglou, R. K. Kaufmann, and M. Sanchez (2007). Modelling the world oil market: Assessment of a quarterly econometric model. *Energy policy* 35(1), 178–191.
- Diebold, F. X. and R. S. Mariano (2002). Comparing predictive accuracy. *Journal of Business & economic statistics* 20(1), 134–144.
- Duan, N. (1983). Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association* 78(383), 605–610.
- Ersoy, E., H. Li, M. E. Schaffer, and T. Szendrei (2023). Stacking regression for time-series, with an application to forecasting quarterly us gdp growth. In *Optimal Transport Statistics for Economics and Related Topics*, pp. 131–149. Springer.
- Fatima, T., E. Xia, and M. Ahad (2019). Oil demand forecasting for china: a fresh evidence from structural time series analysis. *Environment, Development and Sustainability* 21(3), 1205–1224.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

- Fung, W., K. S. Lam, W. Hung, S. Pang, and Y. Lee (2006). Impact of urban temperature on energy consumption of hong kong. *Energy* 31(14), 2623–2637.
- Goodfellow, I. (2016). *Deep learning*.
- Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics* 12(1), 69–82.
- Hornik, K., M. Stinchcombe, and H. White (1989). Multilayer feedforward networks are universal approximators. *Neural networks* 2(5), 359–366.
- Hunt, L. C. and Y. Ninomiya (2003). Unravelling trends and seasonality: a structural time series analysis of transport oil demand in the uk and japan. *The Energy Journal* 24(3), 63–96.
- IEA (2025). Oil 2025. Paris. <https://www.iea.org/reports/oil-2025>. Licence: CC BY 4.0.
- Khairalla, M. A., X. Ning, N. T. Al-Jallad, and M. O. El-Faroug (2018). Short-term forecasting for energy consumption through stacking heterogeneous ensemble learning model. *Energies* 11(6), 1605.
- Krogh, A. and J. Vedelsby (1994). Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems* 7.
- Kuncheva, L. I. and C. J. Whitaker (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning* 51(2), 181–207.
- Manderson, E. and T. Considine (2025). The effect of temperature on energy demand and the role of adaptation. *Journal of the Association of Environmental and Resource Economists* 12(3), 739–774.
- Morris, M. (2001). *The impact of temperature trends on short-term energy demand*.
- Opitz, D. and J. Shavlik (1995). Generating accurate and diverse members of a neural-network ensemble. *Advances in neural information processing systems* 8.
- Qian, C., Y. Yu, and Z.-H. Zhou (2015). Pareto ensemble pruning. In *Proceedings of the AAAI conference on artificial intelligence*, Volume 29.
- Reddy, S., S. Akashdeep, R. Harshvardhan, and S. Kamath (2022). Stacking deep learning and machine learning models for short-term energy consumption forecasting. *Advanced Engineering Informatics* 52, 101542.
- Sbrana, G. and A. Silvestrini (2013). Forecasting aggregate demand: analytical comparison of top-down and bottom-up approaches in a multivariate exponential smoothing framework. *International Journal of Production Economics* 146(1), 185–198.
- Sharma, K., Y. K. Dwivedi, and B. Metri (2024). Incorporating causality in energy consumption forecasting using deep neural networks. *Annals of Operations Research* 339(1), 537–572.
- Spiliotis, E., F. Petropoulos, N. Kourentzes, and V. Assimakopoulos (2020). Cross-temporal aggregation: Improving the forecast accuracy of hierarchical electricity consumption. *Applied Energy* 261, 114339.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58(1), 267–288.
- Timmermann, A. (2006). Forecast combinations. *Handbook of economic forecasting* 1, 135–196.
- Urban, F., R. Benders, and H. Moll (2007). Modelling energy systems for developing countries. *Energy policy* 35(6), 3473–3482.
- Van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007). Super learner. *Statistical applications in genetics and molecular biology* 6(1).
- Wang, X., R. J. Hyndman, F. Li, and Y. Kang (2023). Forecast combinations: An over 50-year review. *International Journal of Forecasting* 39(4), 1518–1547.
- Widiarta, H., S. Viswanathan, and R. Piplani (2009). Forecasting aggregate demand: An analytical evaluation of top-down versus bottom-up forecasting in a production planning framework. *International Journal of Production Economics* 118(1), 87–94.
- Wolpert, D. H. (1992). Stacked generalization. *Neural networks* 5(2), 241–259.
- Yu, L., Y. Zhao, L. Tang, and Z. Yang (2019). Online big data-driven oil consumption forecasting with google trends. *International Journal of Forecasting* 35(1), 213–223.
- Zhang, Y., S. Burer, W. Nick Street, K. P. Bennett, and E. Parrado-Hernández (2006). Ensemble pruning via semi-definite programming. *Journal of machine learning research* 7(7).
- Zhou, Z.-H., J. Wu, and W. Tang (2002). Ensembling neural networks: many could be better than all. *Artificial intelligence* 137(1-2), 239–263.
- Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67(2), 301–320.

(The appendix follows overleaf)

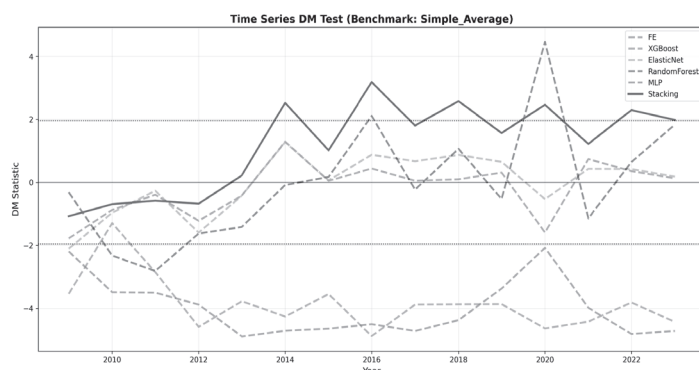


(a) Oil demand (1970 - 2023)

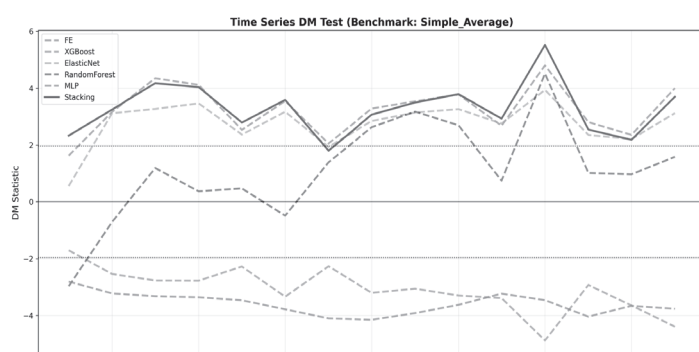


(b) Electricity demand (1985 - 2023)

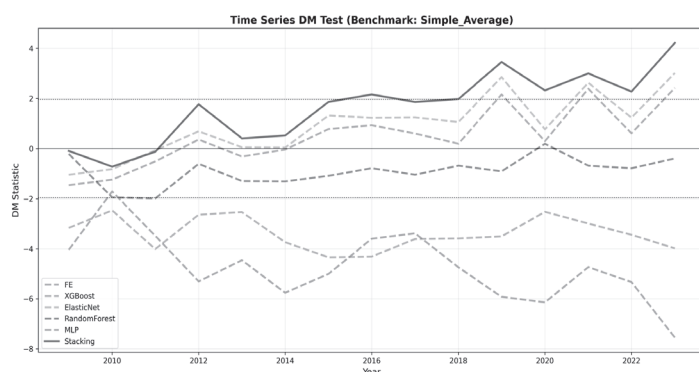
Source(s): Figure by the authors
Figure 1. Energy demand by components/sources and OECD membership



(a) Global

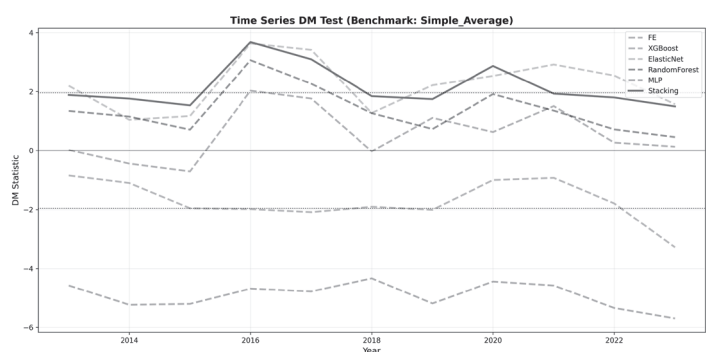


(b) OECD

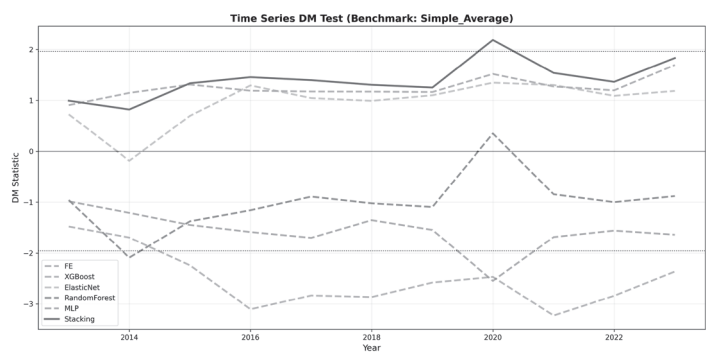


(c) Non-OECD

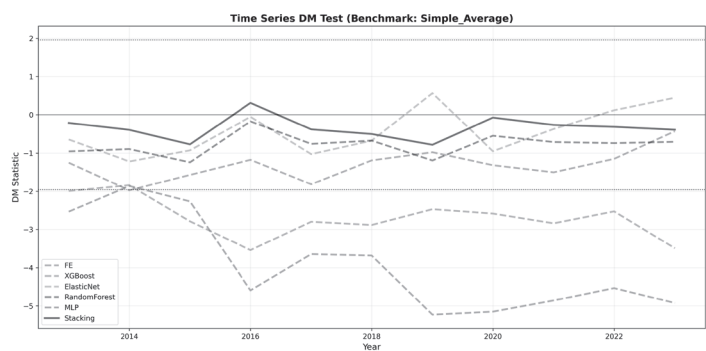
Source(s): Figure by the authors
Figure 2. Diebold–Mariano test statistics on oil demand across models and regions with simple average as benchmark — Top-down



(a) Global

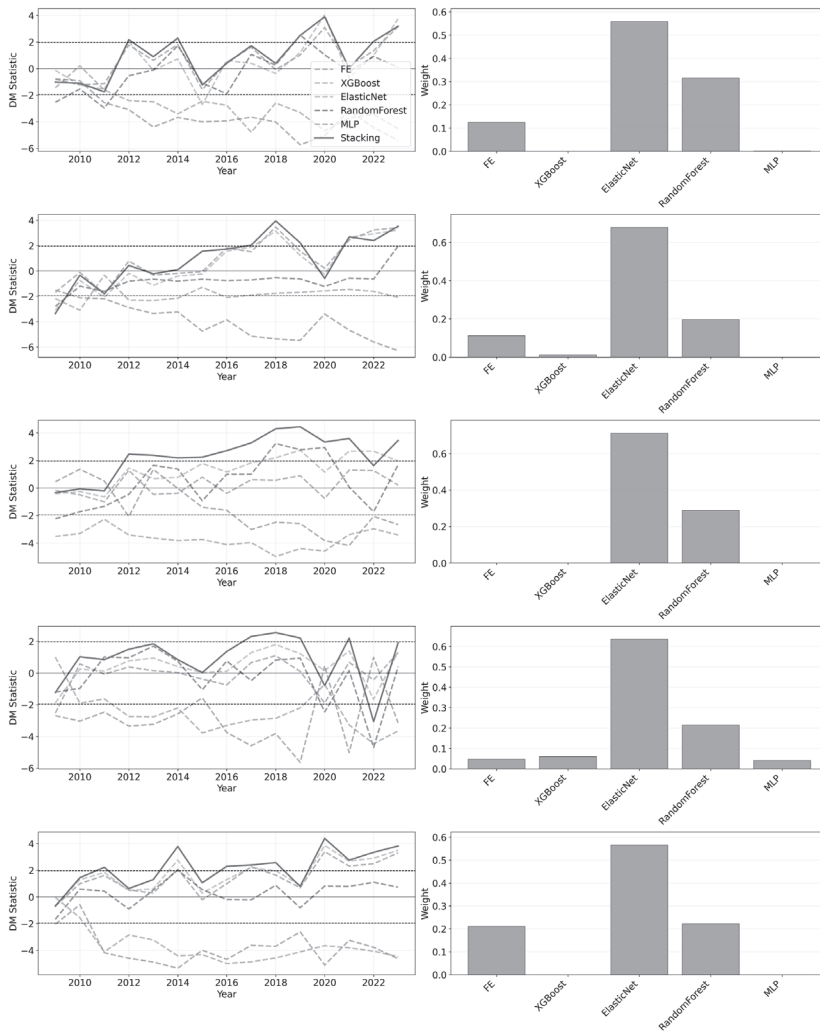


(b) OECD



(c) Non-OECD

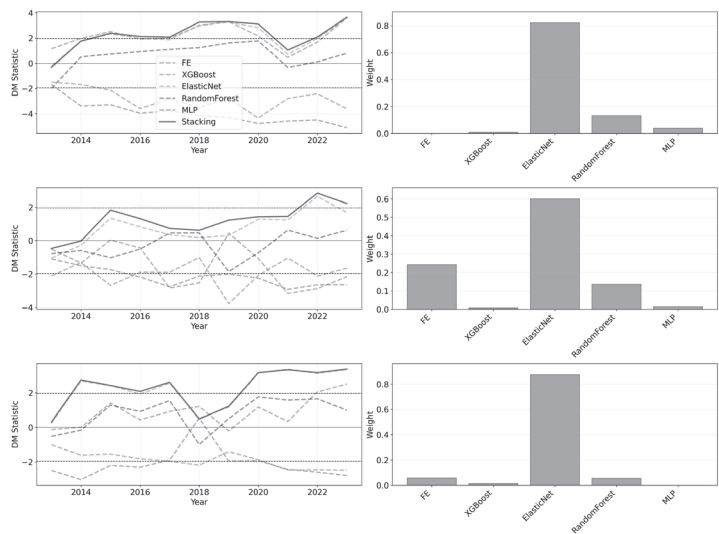
Source(s): Figure by the authors
Figure 3. Diebold–Mariano test statistics on electricity demand across models and regions with simple average as benchmark — Top-down



Source(s): Figure by the authors

Note: Five sub-aggregates: Gas/Diesel (ex. biodiesel), Mogas (ex. ethanol), Fuel oil, Jet/Kerosene, and other.

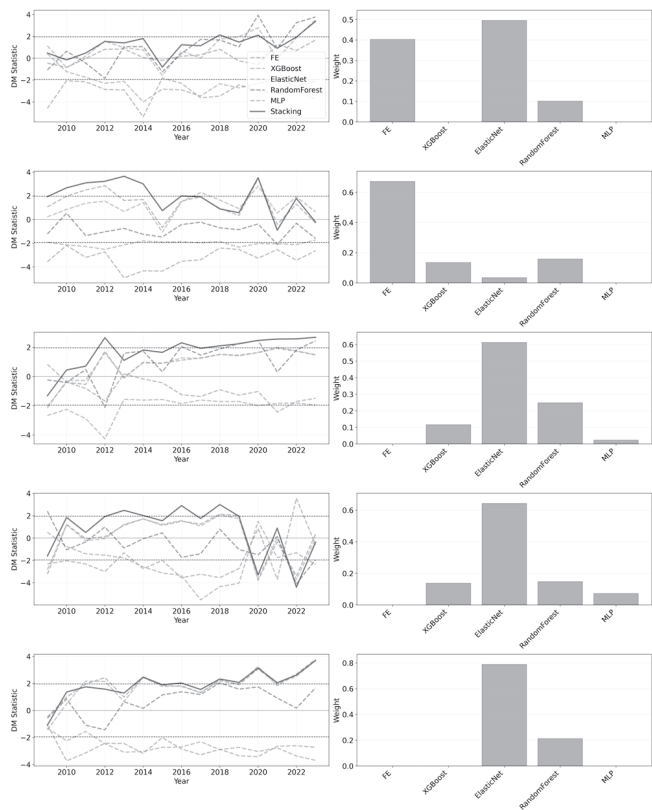
Figure 4. Diebold–Mariano test statistics on oil demand sub-aggregates for global economies across models with simple average as benchmark and stacking weights — Bottom-up



Source(s): Figure by the authors

Note: Three sources: Fossil fuels, Hydro and other (renewable and nuclear).

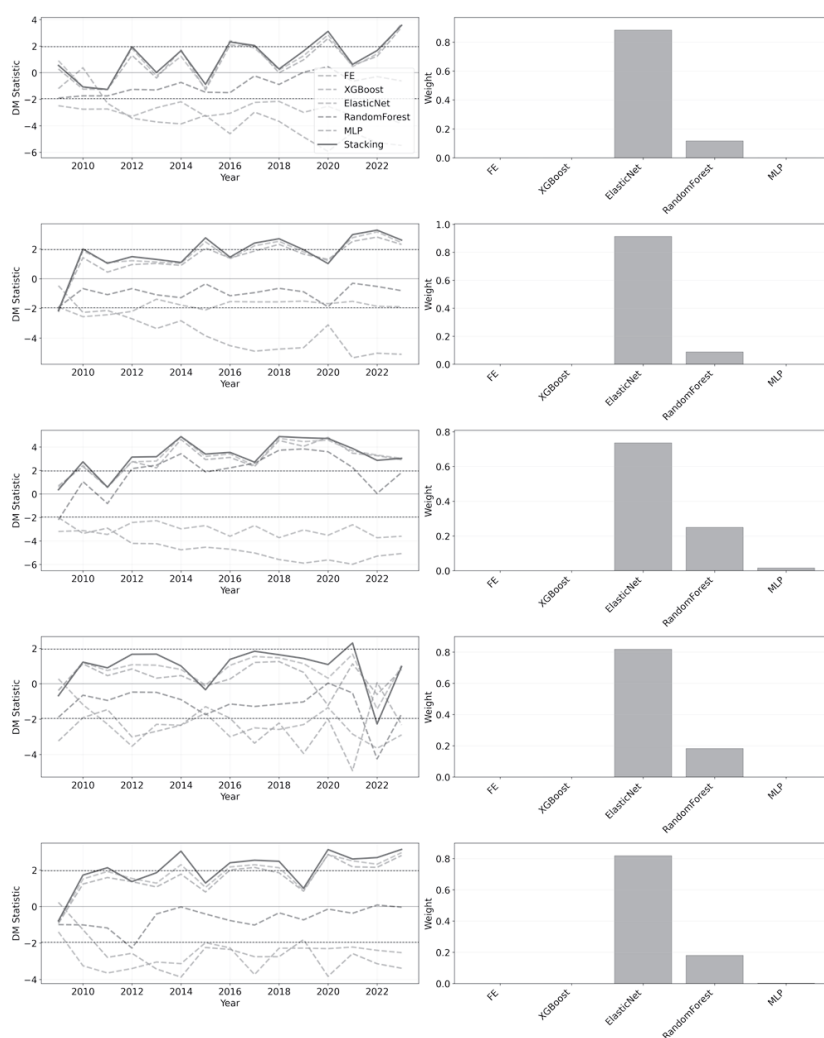
Figure 5. Diebold–Mariano test statistics on electricity demand by source for global economies across models with simple average as benchmark and stacking weights — Bottom-up



Source(s): Figure by the authors

Note: Five sub-aggregates: Gas/Diesel (ex. biodiesel), Mogas (ex. ethanol), Fuel oil, Jet/Kerosene, and other.

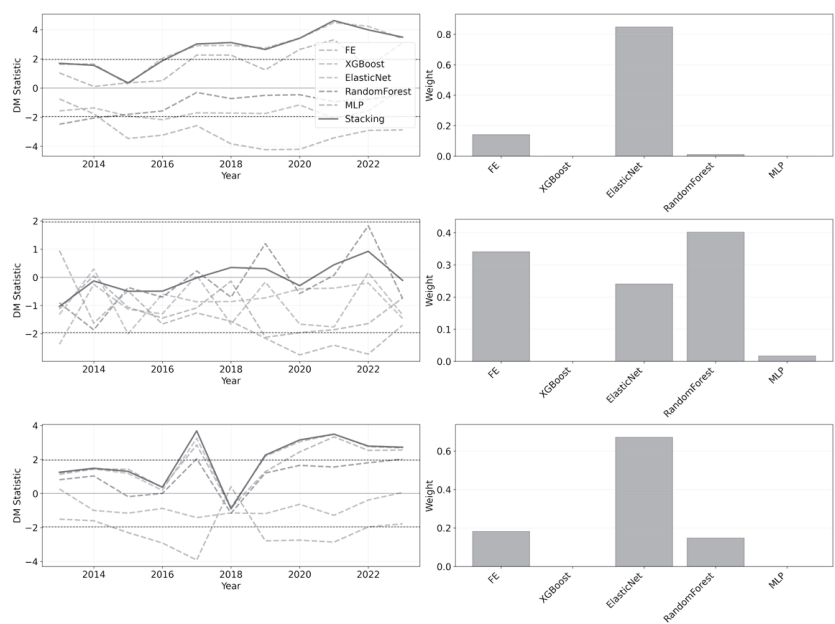
Figure 6. Diebold–Mariano test statistics on oil demand sub-aggregates for OECD economies across models with simple average as benchmark and stacking weights — Bottom-up



Source(s): Figure by the authors

Note: Five sub-aggregates: Gas/Diesel (ex. biodiesel), Mogas (ex. ethanol), Fuel oil, Jet/Kerosene, and other.

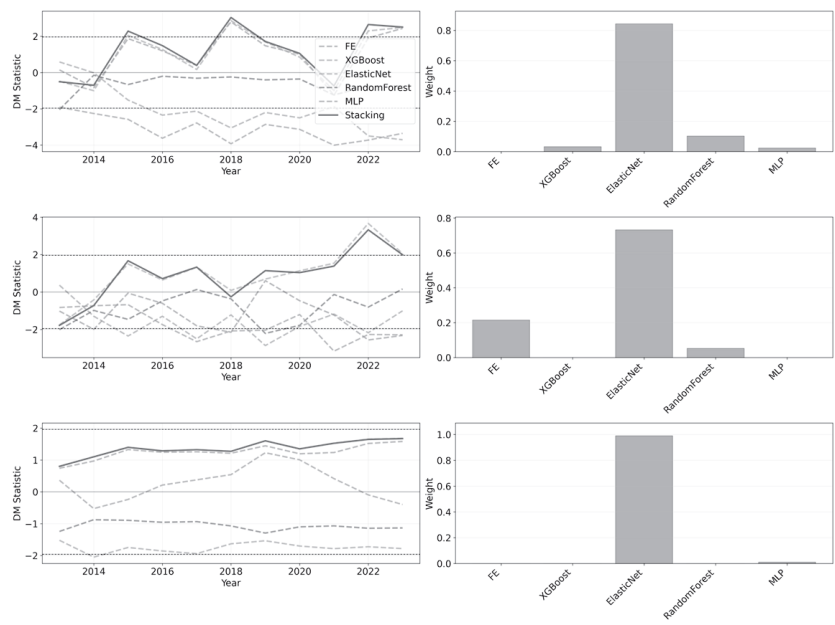
Figure 7. Diebold–Mariano test statistics on oil demand sub-aggregates for Non-OECD economies across models with simple average as benchmark and stacking weights — Bottom-up



Source(s): Figure by the authors

Note: Three sources: Fossil fuels, Hydro and other (renewable and nuclear).

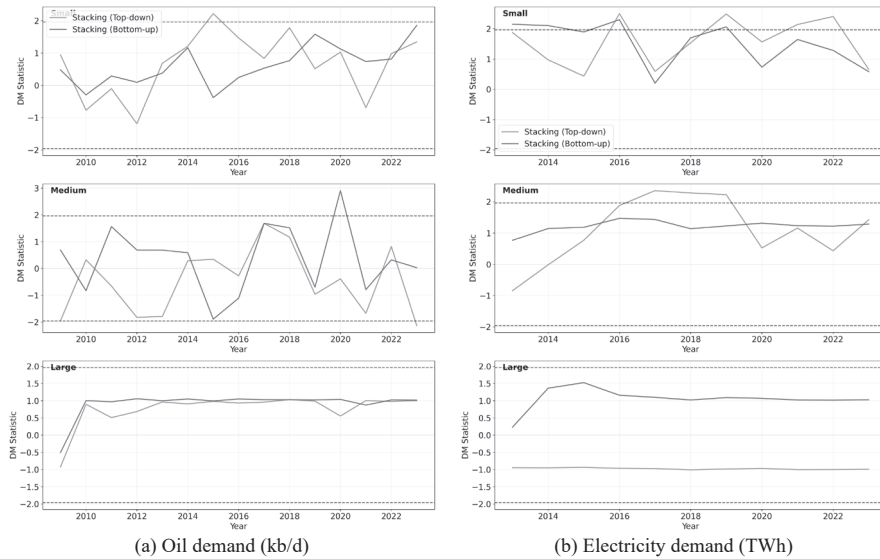
Figure 8. Diebold–Mariano test statistics on electricity demand by source for OECD economies across models with simple average as benchmark and stacking weights — Bottom-up



Source(s): Figure by the authors

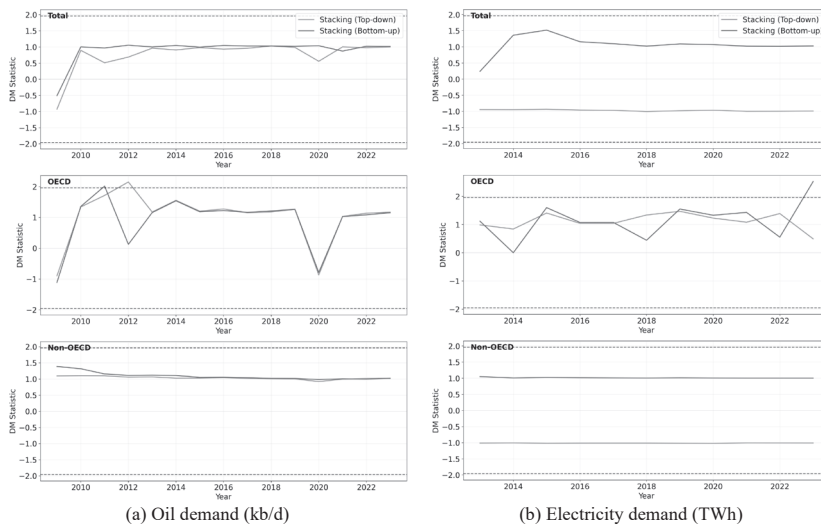
Note: Three sources: Fossil fuels, Hydro and other (renewable and nuclear).

Figure 9. Diebold–Mariano test statistics on electricity demand by source for Non-OECD economies across models with simple average as benchmark and stacking weights — Bottom-up



Source(s): Figure by the authors

Figure 10. Diebold–Mariano test statistics on energy demand in level for Top-down and Bottom-up stacking estimators across subsets by country size with simple average as benchmark



Source(s): Figure by the authors

Figure 11. Diebold–Mariano test statistics on energy demand in level for Top-down and Bottom-up stacking estimators across regions with simple average as benchmark

Table 1. Summary of data and variables

Description	
Economies	
OECD	32 OECD members. Czech Republic, Estonia, Latvia, Lithuania, Slovak Republic, and Slovenia are omitted from the OECD group in this dataset due to data availability.
Non-OECD	All 138 remaining economies (see online supplement).
Oil products (kb/d)	
Motor gasoline	Motor gasoline consumption excluding ethanol blending.
Gas/Diesel	Exclude biodiesel and bunkers consumption.
Fuel oil	Exclude bunkers consumption.
Jet kerosene	Jet kerosene consumption.
Other products	aviation gasoline (avgas), bitumen, other kerosene, LPG/ethane, lubricants, other oil products category, international bunkers for fuel oil and gas/diesel, plus refinery fuel use and refinery losses.
Electricity generation (TWh)	
Fossil	Electricity generated from fossil fuels, including coal, natural gas, and oil-fired generation.
Hydro	Electricity generated from hydropower.
Other	Biomass, geothermal, nuclear, solar, wind, and other renewables.
Regressors	
Ice Days (ID)	Number of ice days (daily maximum temperature < 0°C).
Summer Days (SD)	Number of summer days (daily maximum temperature > 25°C).
GDP	Real GDP at constant 2015 prices in US Dollars.
Population (POP)	Total population.
Industry % of GDP (IND)	Industry value added as a share of GDP. “Industry” combines: Mining, Manufacturing, Utilities (ISIC C–E), Manufacturing (ISIC D), and Construction (ISIC F).
Oil price (PRICE)	annually averaged WTI price.

Source(s): Table by the authors

Table 2. Hyperparameter grids for tuning machine learning base learners

Learner	Hyperparameter	Grid values
Elastic Net	Penalisation λ	{0.001, 0.01, 0.1, 1.0, 10.0}
	Mixing ratio α	{0, 0.3, 0.5, 0.7, 1.0}
Random Forest	Number of trees	{300}
	Maximum depth	{5, 10, no limit}
	Fraction of features per split	{ $\sqrt{p}$, 0.5, 1.0}
	Minimum split size	{5}
XGBoost	Number of trees	{300}
	Maximum depth	{2, 4, 6}
	Learning rate	{0.03, 0.05, 0.1}
	Subsample fraction	{0.8}
	Column subsample per tree	{0.8}
Multi-layer Perceptron	Hidden layer size	{{16}, {32, 16}, {64, 32}}
	Weight decay	{0.0001, 0.001, 0.01, 0.1}
	Initial learning rate	{0.001, 0.01, 0.1}
	Activation function	{ReLU, tanh}

Source(s): Table by the authors

Table 3. Forecast performance for oil demand across models and regions — Top-down

	FE	XGBoost	Elastic Net	Random Forest	MLP	Simple Average	Stacking
Global							
RMSE	0.1215	0.1571	0.1207	0.1200	0.1537	0.1207	0.1171
MAE	0.0826	0.1125	0.0828	0.0829	0.1128	0.0828	0.0796
MAPE	12.95%	14.34%	12.86%	14.26%	11.94%	11.98%	12.96%
OECD							
RMSE	0.0692	0.1467	0.0747	0.0831	0.1959	0.0942	0.0692
MAE	0.0463	0.1108	0.0520	0.0598	0.1541	0.0718	0.0469
MAPE	0.85%	1.89%	0.94%	1.05%	2.57%	1.25%	0.86%
Non-OECD							
RMSE	0.1305	0.1711	0.1291	0.1448	0.1729	0.1327	0.1271
MAE	0.0895	0.1237	0.0885	0.0962	0.1305	0.0930	0.0872
MAPE	15.77%	16.22%	15.64%	17.05%	16.69%	15.11%	15.73%

Source(s): Table by the authors**Table 4.** Forecast performance for electricity demand across models and regions — Top-down

	FE	XGBoost	Elastic Net	Random Forest	MLP	Simple Average	Stacking
Global							
RMSE	0.1405	0.1768	0.1308	0.1279	0.3190	0.1441	0.1235
MAE	0.0870	0.1110	0.0799	0.0742	0.2255	0.0941	0.0717
MAPE	37.16%	40.58%	28.61%	31.41%	85.27%	32.79%	28.29%
OECD							
RMSE	0.0608	0.1308	0.0623	0.0926	0.0904	0.0753	0.0616
MAE	0.0443	0.0913	0.0462	0.0554	0.0614	0.0504	0.0432
MAPE	1.41%	3.30%	1.45%	2.44%	2.33%	1.96%	1.48%
Non-OECD							
RMSE	0.1525	0.1880	0.1421	0.1531	0.1894	0.1390	0.1426
MAE	0.0952	0.1232	0.0864	0.0858	0.1334	0.0870	0.0820
MAPE	46.20%	53.06%	35.82%	36.80%	53.60%	33.26%	35.44%

Source(s): Table by the authors

Metric: RMSE

Region	FE	XGBoost	Elastic Net	Random Forest	MLP	Simple Average	Stacking
Metric: RMSE							
<i>Gas/diesel (ex. biodiesel)</i>							
Global	0.1022	0.1246	0.1030	0.1042	0.1341	0.1038	0.1014
OECD	0.0580	0.0786	0.0556	0.0548	0.0760	0.0592	0.0554
Non-OECD	0.1248	0.1580	0.1245	0.1379	0.1740	0.1278	0.1239
<i>Mogas (ex. ethanol)</i>							
Global	0.0929	0.1192	0.0940	0.1074	0.1371	0.0961	0.0927
OECD	0.0528	0.0937	0.0531	0.0656	0.1002	0.0593	0.0539
Non-OECD	0.1148	0.1621	0.1131	0.1441	0.1949	0.1262	0.1126
<i>Fuel oil</i>							
Global	0.2189	0.3300	0.2106	0.2166	0.2386	0.2208	0.2049
OECD	0.1493	0.4254	0.1477	0.1411	0.2116	0.1784	0.1493
Non-OECD	0.2444	0.3417	0.2417	0.2505	0.5032	0.2791	0.2417
<i>Jet/kerosene</i>							
Global	0.2260	0.2461	0.2237	0.2310	0.2526	0.2250	0.2236
OECD	0.1981	0.2017	0.1975	0.2026	0.2002	0.1934	0.1951
Non-OECD	0.2438	0.2846	0.2406	0.2661	0.2760	0.2464	0.2406
<i>Other oil</i>							
Global	0.1307	0.1928	0.1289	0.1353	0.2008	0.1381	0.1270
OECD	0.0604	0.1519	0.0625	0.0719	0.1293	0.0818	0.0621
Non-OECD	0.1682	0.2857	0.1669	0.1901	0.2278	0.1817	0.1651
Metric: MAE							
<i>Gas/diesel (ex. biodiesel)</i>							
Global	0.0650	0.0826	0.0658	0.0650	0.0933	0.0667	0.0636
OECD	0.0425	0.0578	0.0409	0.0384	0.0603	0.0442	0.0404
Non-OECD	0.0801	0.1050	0.0789	0.0850	0.1216	0.0831	0.0782
<i>Mogas (ex. ethanol)</i>							
Global	0.0590	0.0745	0.0599	0.0615	0.0918	0.0615	0.0584
OECD	0.0366	0.0639	0.0378	0.0453	0.0784	0.0427	0.0372
Non-OECD	0.0730	0.0976	0.0714	0.0803	0.1319	0.0806	0.0707
<i>Fuel oil</i>							
Global	0.1388	0.2088	0.1311	0.1332	0.1561	0.1390	0.1252
OECD	0.1049	0.2669	0.1037	0.0935	0.1512	0.1205	0.0994
Non-OECD	0.1430	0.2230	0.1407	0.1500	0.3260	0.1742	0.1404
<i>Jet/kerosene</i>							
Global	0.1279	0.1475	0.1254	0.1267	0.1620	0.1279	0.1237
OECD	0.1066	0.1231	0.1065	0.1116	0.1247	0.1060	0.1053
Non-OECD	0.1446	0.1741	0.1404	0.1542	0.1798	0.1457	0.1400
<i>Other oil</i>							
Global	0.0738	0.1180	0.0724	0.0737	0.1296	0.0811	0.0703
OECD	0.0441	0.1050	0.0461	0.0536	0.0944	0.0601	0.0463
Non-OECD	0.0970	0.1638	0.0958	0.1030	0.1466	0.1077	0.0938
Source(s): Table by the authors							

Table 6. Forecasting performance for electricity demand by source by region — Bottom-up

Region	FE	XGBoost	Elastic Net	Random Forest	MLP	Simple Average	Stacking
Metric: RMSE							
<i>Fossil fuels</i>							
Global	0.1333	0.1820	0.1328	0.1401	0.2332	0.1451	0.1335
OECD	0.0875	0.1279	0.0832	0.1129	0.1720	0.1011	0.0835
Non-OECD	0.1507	0.1906	0.1512	0.1653	0.2234	0.1575	0.1508
<i>Hydro</i>							
Global	0.1831	0.2011	0.1672	0.1752	0.2321	0.1731	0.1668
OECD	0.1187	0.1193	0.1174	0.1092	0.1472	0.1078	0.1076
Non-OECD	0.2073	0.2310	0.1876	0.2062	0.2405	0.1956	0.1889
<i>Other sources</i>							
Global	0.1876	0.2530	0.1772	0.1849	0.2150	0.1917	0.1777
OECD	0.1959	0.2487	0.1918	0.2066	0.4466	0.2183	0.1917
Metric: MAE							
<i>Fossil fuels</i>							
Global	0.0761	0.1073	0.0722	0.0786	0.1413	0.0830	0.0729
OECD	0.0631	0.0929	0.0597	0.0744	0.1266	0.0753	0.0598
Non-OECD	0.0813	0.1099	0.0786	0.0886	0.1305	0.0853	0.0783
<i>Hydro</i>							
Global	0.1150	0.1193	0.0947	0.0945	0.1320	0.1015	0.0966
OECD	0.0826	0.0797	0.0735	0.0679	0.0977	0.0709	0.0693
Non-OECD	0.1287	0.1327	0.1056	0.1073	0.1338	0.1122	0.1082
<i>Other sources</i>							
Global	0.0456	0.0756	0.0390	0.0416	0.0698	0.0497	0.0391
OECD	0.0777	0.1371	0.0729	0.0797	0.2605	0.1150	0.0719
Source(s): Table by the authors							

Table 7. Prediction comparisons of Top-down and Bottom-up stacking estimators for energy demand across regions.

Region	Stacking	RMSE	MAE	MAPE	DA (%)
Oil demand (kb/d)					
Global	Top-down	112.2934	26.7425	8.0368	58.7248
Global	Bottom-up	96.5777	22.4039	9.7139	58.3893
OECD	Top-down	153.3818	48.6300	4.8428	59.0814
OECD	Bottom-up	149.1932	49.6137	5.4418	59.7077
Non-OECD	Top-down	114.0007	21.0925	8.8149	58.3508
Non-OECD	Bottom-up	137.9735	22.1147	9.9698	57.8782
Electricity demand (TWh)					
Global	Top-down	199.1173	17.3058	7.2100	65.6501
Global	Bottom-up	39.7697	5.4725	10.4715	62.2793
OECD	Top-down	28.5666	10.6599	4.4686	59.2593
OECD	Bottom-up	19.4549	8.7760	7.4401	57.5499
Non-OECD	Top-down	316.0038	29.1838	8.2080	66.7765
Non-OECD	Bottom-up	25.1040	3.2955	9.5405	66.1173

Source(s): Table by the authors

Table 9. Forecast performance for electricity demand by model, aggregation approach and economy size

Model	Small economies				Medium economies				Large economies			
	RMSE	MAE	MAPE	DA (%)	RMSE	MAE	MAPE	DA (%)	RMSE	MAE	MAPE	DA (%)
<i>Top-down</i>												
Simple Average	0.1093	0.0536	12.2962	60.3252	1.5326	0.8978	9.5007	67.5039	199.6365	38.2671	6.6417	63.4146
FE	0.1239	0.0567	11.8577	63.0894	1.4599	0.8224	8.8462	69.7017	47.1186	16.7182	5.6324	63.7398
XGBoost	0.1185	0.0592	14.6955	58.8618	1.7922	1.0734	11.8575	63.1083	438.7212	76.1950	7.7568	61.4634
Elastic Net	0.1120	0.0505	10.5190	64.2276	1.4226	0.7889	8.5046	68.9168	70.4361	20.5209	5.3165	64.3902
Random Forest	0.0863	0.0409	9.4524	59.6748	1.3652	0.7463	8.1992	67.9749	414.6910	61.1293	4.7536	63.0894
MLP	0.2137	0.1163	28.2201	59.3496	3.6996	2.1690	21.1109	63.1083	179.0574	64.0578	17.4298	62.4390
Stacking	0.0884	0.0415	9.3713	62.6016	1.2788	0.6968	7.6795	68.9168	346.9257	51.7723	4.5624	64.7154
<i>Bottom-up</i>												
FE	0.0873	0.0474	30.7592	60.9756	1.6751	0.9444	10.7641	65.7771	40.0810	11.9141	4.2079	63.5772
XGBoost	0.1311	0.0648	18.8156	58.5366	1.5556	0.8936	11.0681	63.1083	351.1590	61.0398	7.3279	58.6992
Elastic Net	0.0879	0.0419	20.3328	57.2358	1.3691	0.7132	8.8872	67.0330	28.3178	10.3770	3.9791	62.9268
Random Forest	0.0935	0.0441	12.5498	59.8374	1.4123	0.7639	9.4338	63.8932	382.7682	60.1532	5.3378	61.1382
MLP	0.1767	0.0959	44.0529	54.4715	2.1706	1.3119	18.4571	63.4223	96.8017	27.4573	7.2301	64.0650
Stacking	0.0866	0.0406	18.1336	57.2358	1.4018	0.7521	9.2709	66.5620	69.2772	15.7936	4.0529	62.4390
Source(s): Table by the authors												

Corresponding author

Mark E. Schaffer can be contacted at: m.e.schaffer@hw.ac.uk