Improving Macroeconomic Model Validity and Forecasting Performance with Pooled Country Data using Structural, Reduced Form, and Neural Network Models

Cameron Fen and Samir Undavia

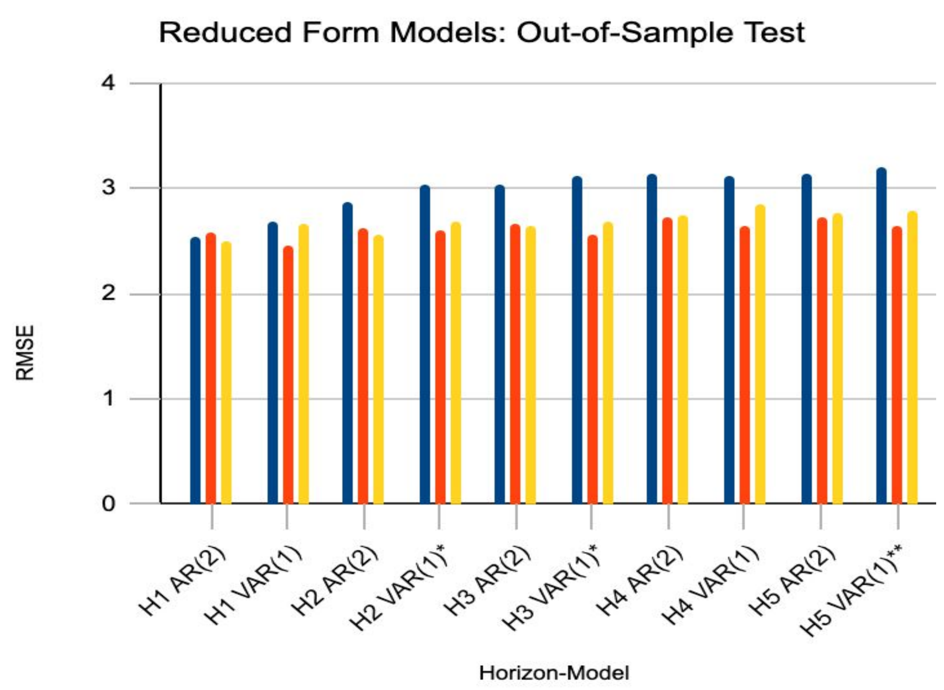
Abstract

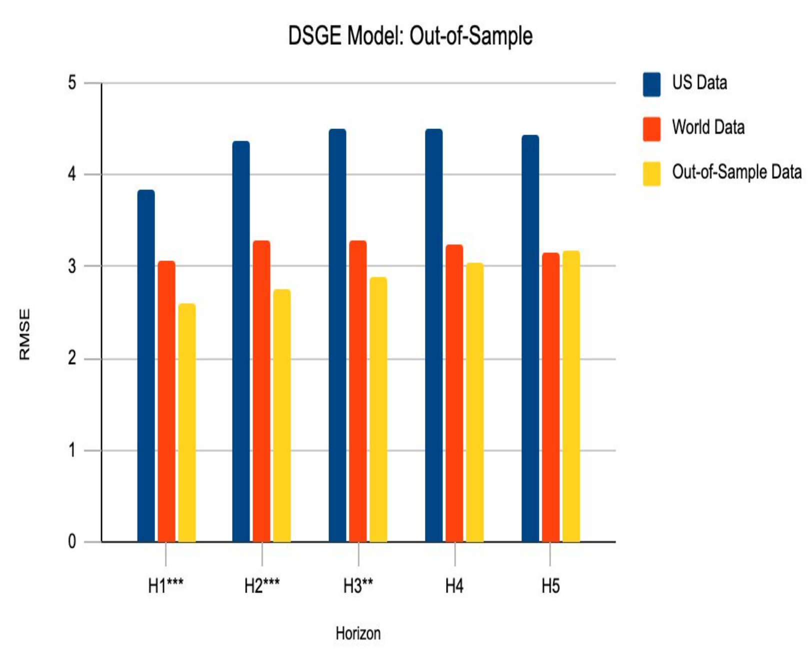
We show that pooling countries' macroeconomic data across a panel dimension produces a statistically significant improvement in the generalizability of structural, reduced form, and machine learning (ML) methods, producing state-of-the-art results. Using GDP forecasts evaluated on an out-of-sample test set, this procedure reduces RMSE anywhere from 12% to 24% depending on the type of model. Forecasting using non-US-pooled-data, we show that reduced-form and structural models are more policy-invariant and outperform a US-data-only baseline. Our deep learning approaches outperform all tested baseline economic models. Robustness checks indicate that our outperformance is reproducible, numerically stable, and generalizable across models.

Models Used

Var(4) - A 4-period linear autoregressive model using GDP, consumption, and employment as inputs
AR(2) - A 2-period linear autoregressive model using only GDP
Smets-Wouters DSGE - Standard DSGE from the Smets-Wouters paper
Factor - A linear model that takes in 248 data series and uses PCA to reduce dimensionality
Our Custom RNN - A gated-recurrent-unit-based neural network with skip connections and dropout
AutoML - An algorithm that chooses the best machine learning model from a collection of models

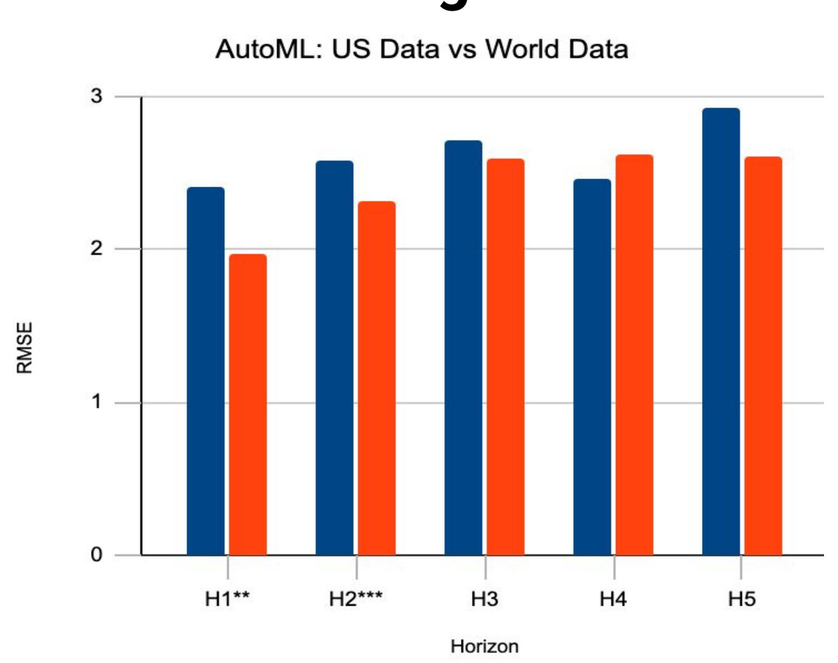
Structural and Reduced Form Models

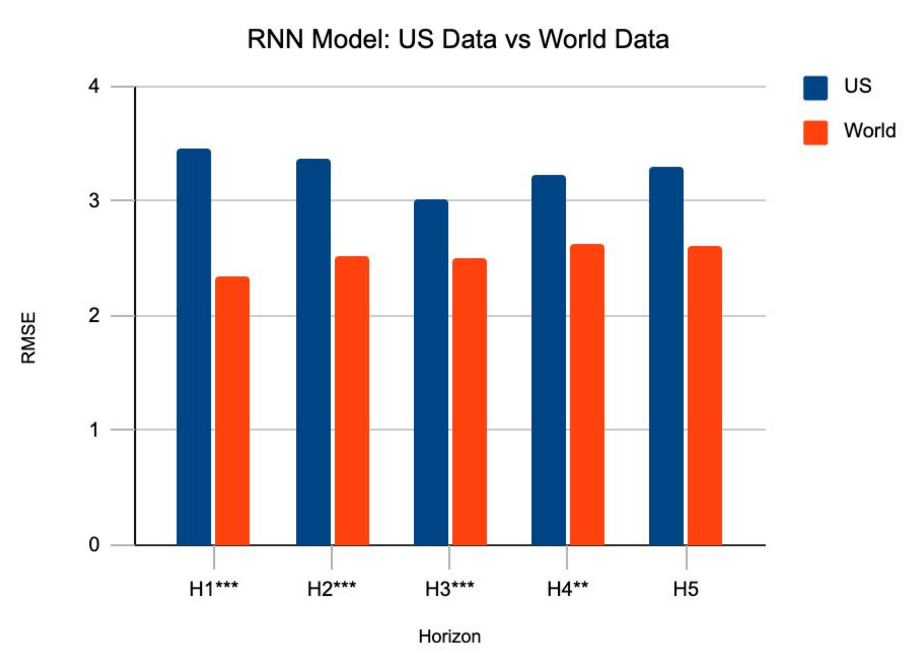




The left chart shows the performance of Var(1) and AR(2) reduced form models. The right chart shows the performance of the Smets-Wouters DSGE estimated with maximum likelihood. For both charts, the first bar in each triplet is the model trained on US-only data. The second bar is the model trained on entire cross-section of data, including US data, and forecasted on only US data. The third bar shows generalizability and policy invariance by training the model on all country data except US data and validated on US-only data.

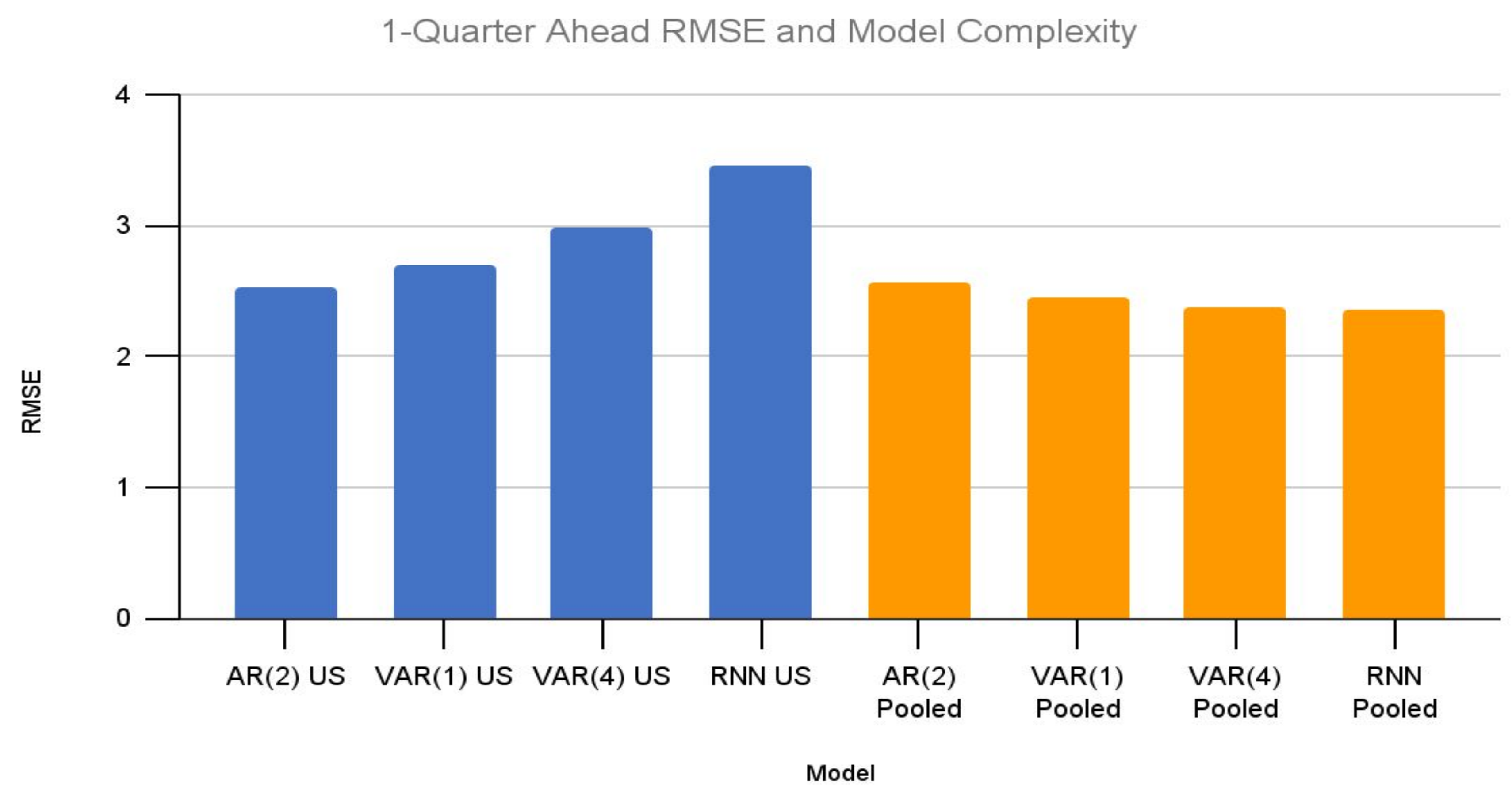
Machine Learning Models





The left chart shows the performance of AutoML, and the right chart shows the performance of our custom GRU (RNN) model. For both charts, the first bar in each double is the model trained on US-only data. The second bar is the model trained on entire cross-section of data, including US data, and forecasted on only US data. We did not include results from our world out-of-sample (third bar in the reduced form and DSGE charts) as there was not enough time-series data for the deep learning models. For AutoML, the bars within each double are not directly comparable since, for example, a model trained on US data only might use XGBoost and the world data test might have best results with a deep learning model. We provide the RNN model to show effective comparison for a flexible machine learning model.

Model Complexity



This chart shows model complexity increases from AR(2) to RNN in two different situations: the blue bars using only US data and the orange bars using the larger pooled dataset. The blue bars decrease in performance significantly as model complexity increases, while the orange bars increase in performance as model complexity increases. While the differences in RMSE for the orange bars look slight, it is the difference between a standard model and a state-of-the-art model. Furthermore, the large jump between the blue AR(2) and RNN bars show an extreme difference in performance, where the RNN is not a useable model.

Comparison of Model Performance

Time (Q's Ahead)	1Q	2Q	3Q	4Q	5Q
VAR(4)					
US Data	2.99	3.03	3.10	3.08	3.08
World Data	2.37	2.52	2.56	2.63	2.63
AR(2)					
US Data	2.53	2.88	3.03	3.14	3.13
World Data	2.57	2.62	2.67	2.72	2.72
Smets-Wouters DSGE Bayesian					
US Data	2.79	2.95	2.89	2.80	2.71
Factor					
US Data	2.24	2.48	2.50	2.67	2.86
RNN (Ours)					
US Data	3.46	3.37	3.01	3.23	3.30
World Data	2.35	2.52	**2.50**	**2.62**	**2.60**
AutoML (Ours)					
US Data	2.41	2.58	2.71	**2.45**	2.92
World Data	**1.97**	**2.32**	2.59	**2.62**	**2.61**
SPF Median	1.86	2.11	2.36	2.46	2.65

This table shows the results of our tested models. We compared the performance of all of the above models mentioned in this poster with our RNN and AutoML models. Our RNN with pooled data had the lowest RMSE (highest performance) for 3Q, 4Q, and 5Q ahead. The AutoML model trained with pooled data had the lowest RMSE for 1Q and 2Q ahead.