Predicting the Relative Forecasting Performance of the Models: Conditional Predictive Ability Approach*

Eleonora Granziera† and Tatevik Sekhposyan‡

February 18, 2017

Abstract

The relative performance of forecasting models is known to be unstable over time. However, it is not well understood why the forecasting performance of economic models change. We propose to address this question by evaluating the predictive ability of a wide range of economic variables for key U.S. macroeconomic aggregates: output growth and inflation. We take a conditional view on this issue, identifying situations where particular kind of models perform better than simple benchmarks. We, therefore, test whether the relative forecasting performance of models depend on the state of the business cycle, financial conditions, uncertainty or measures of past relative performance. We then investigate whether the conditioning variables help us predict the more accurate forecasting model for a specific future date. In particular, we analyze whether using the conditional performance as a model selection or model averaging criteria can improve the accuracy of the predictions. The proposed strategies deliver sizable improvements especially when the relative performance is predicted using financial variables.

Keywords: Conditional Predictive Ability, Model Averaging, Inflation Forecasts, Output Growth Forecasts

J.E.L. Codes: C22, C52, C53

preliminary and incomplete, do not cite. We are grateful to seminar participants at the Bank of Finland, the participants of 2016 International Symposium on Forecasting and 2015 Society for Nonlinear Dynamics and Econometrics for comments. The views expressed in this paper are those of the authors. No responsibility should be attributed to the Bank of Finland.

†Bank of Finland, Snellmaninaukio Helsinki, Finland; E-mail: eleonora.granziera@gmail.com

‡Texas A&M University, 3060 Allen Building, 4228 TAMU, College Station, TX 77843-4228, USA; E-mail: tsekhposyan@tamu.edu

1 Introduction

The relative forecasting performance of models is known to be unstable over time. For instance, Stock and Watson (2007) document the deterioration of the relative forecasting performance of the economic models over univariate benchmarks of inflation in mid 1980s. Rossi and Sekhposyan (2010) further show that there is a widespread instability in the relative forecasting performance of the models for output growth dating to late 1970s. Despite this evidence (see Rossi, 2013, for an overview), it is still little understood why the forecasting performance change or why they change at different rates across the models.

The literature has approached the understanding of the reversals of models' predictive abilities largely from two perspective. For instance, Rossi and Sekhposyan (2010), Manzan and Zerom (2013), Clark and Doh (2014) document the reversals in the relative predictive abilities of the models. They further seek to identify the economic events that might have occurred at the times of reversals. Alternatively, others have approached the problem by identifying certain recurring periods of economic significance and evaluating the forecasting performance of models in those periods. For instance, Chauvet and Potter (2013) review the relative ability of a wide range of models to forecast output growth in recessions versus expansions. Dotsey, Fujita and Stark (2011) consider the relative forecasting performance of inflation models conditional on the state of the business cycle. Ng and Wright (2013) emphasize the importance of the origin of the business cycle. They suggest that recessions that originate in the financial markets are different from others, and this could explain why some models and economic variables work well at some times and deteriorate in performance in other times.

In this paper we take an economic approach to understanding the relative performance of the models. In addition, we are interested in predicting which model will be more accurate on a specific future date. To this end we employ the test of *conditional* predictive ability proposed by Giacomini and White (2006). Tests of *unconditional* predictive ability ask whether the forecasting models performed equally well on average in the past, and if so, they might have useful recommendations for selecting more accurate models for an unspecified future date. Examples of such tests are Diebold and Mariano (1995), West (1996), Clark and McCracken (2001) and Clark and West (2007), among others. The testable hypothesis is whether the expected loss differences have a zero mean. However, a researcher might be interested in knowing whether relative performance is forecastable. As noted, one could wonder whether the state of the business cycle, i.e. whether the economy is in a recession or expansion, could help us choose a model for a particular future date. It might then be more appropriate to use a test of conditional predictive ability which asks whether there is any information available at the time the forecasts are made, above and beyond past average performance, that can explain the relative performance of the models. Accordingly,

the null hypothesis is whether the expected loss differences have zero mean conditional on some information set, for example, conditional on the economy being in a recession.

The contribution of our paper over the literature is as follows. First, we consider a wide set of univariate and multivariate models of both inflation and output growth. More specifically, we rely on McCracken and Ng (2015) monthly database to consider more than hundred real and nominal variables. Second, we consider a wide set of conditional variables. We not only control for the state of the business cycle, which is commonly done in the literature, but evaluate the importance of measures of financial stress, uncertainty, as well as past relative predictive ability as conditioning variables. Most importantly, we take testing for conditional predictive ability a step further and evaluate its usefulness for model selection and model averaging. To this end, we consider the model selection rule of Giacomini and White (2006) and novel model averaging rules.

In line with previous literature we find that rejections when using the unconditional test are rare, suggesting that the benchmark and the alternative models are equally good on average over the sample. When applying the conditional test, our general finding is that the relative performance of the models can be predicted by a measure of financial stress at short horizons, while at longer horizons the past relative perfomance is a good predictor of future performance. Moreover using the conditioning information in a simple decision rule performs at least as well as the single models or a simple average of them.

The rest of the paper is organized as follows: section 2 presents the econometric framework, section 3 describes the models used to obtain forecasts, section 4 discusses the data and conditional variables, section 5 reports the results and section 6 concludes.

2 Econometric Framework

2.1 Testing for Conditional Predictive Ability

Suppose that $\{y_{s+\tau}, x_s\}_{s=1}^{t}$ are stationary time series variables at each forecast origin $t = R + 1, ..., T - \tau$, where R is the estimation window size and $\tau > 0$ is the forecast horizon. We are interested in forecasting a scalar $y_{t+\tau}$, $\tau \geq 1$, using two alternative models.[1] Denote by $f_{t,R}\left(\hat{\beta}_{0,t}\right) = f\left(y_t, x_t, x_{t-1}, ...; \hat{\beta}_{0,t}\right)$ and $g_{t,R}\left(\hat{\beta}_{1,t}\right) = g\left(y_t, x_t, x_{t-1}, ...; \hat{\beta}_{1,t}\right)$ the $\tau-$period ahead forecasts obtained from the estimated models through either the fixed or rolling window scheme with window size R. In the work that follows we will take $g_{t,R}\left(\hat{\beta}_{1,t}\right)$ to be the benchmark model. The testing framework proposed by Giacomini and White (2006) is valid for general loss functions. In this paper we focus on evaluating point forecasts, and we use the squared error loss as our measure of accuracy, given that this loss function is the most widely used in empirical studies which

[1] This framework allows data to be non-stationary. However, the type of non-stationarity considered rules out unit roots, but allows for changes that could be induced by distributions changing over time.

assess forecast performance of models for inflation and real activity.

Let $\Delta L_{R,t+\tau} = \left(y_{t+\tau} - f_t\left(\hat{\beta}_0, t\right)\right)^2 - \left(y_{t+\tau} - g_t\left(\hat{\beta}_1, t\right)\right)^2$. A positive value for the loss differential, $\Delta L_{R,t+\tau}$, indicates that the alternative model is inferior to the benchmark, while a negative value is indicative of a superior performance of the alternative**.** The null hypothesis is expressed as:

$$H_0 : E\left[\Delta L_{R,t+\tau} | \mathcal{G}_t\right] = 0 \tag{1}$$

The null is formulated in terms of the parameter estimates rather than their population values, therefore, the null is a statement on the forecasting methods: models, size of the estimation window and estimation procedure are all subject to evaluation. Furthermore, this framework allows for comparison of nested as well as non-nested models and of Bayesian as well as classical estimation procedures. When $\mathcal{G}_t = \{\mathcal{F}_t\}$, where $\mathcal{F}_t$ is the time-t information set, the null implies that the forecasting methods are equally accurate given the information available at time t. The unconditional predictive ability test can be considered as a special case of (1), where the conditioning set $\mathcal{G}_t = \{\emptyset, \Omega\}$ is the trivial σ-field. Thus, when testing for unconditional predictive ability, we test $H_0 : E\left[\Delta L_{R,t+\tau}\right] = 0$, i.e. whether the models are equally accurate on average.

For a given choice of a $q \times 1$ vector of conditioning variables h_t, testing for the null is equivalent to testing $E(h_t \Delta L_{R,t+\tau}) = 0$. Let $P = T - \tau - R$. The proposed test statistic is:

$$T^h_{R,P,\tau} = P\left(P^{-1}\sum_{t=R}^{T-\tau} h_t \Delta L_{R,t+\tau}\right)' \hat{V}^{-1}\left(P^{-1}\sum_{t=R}^{T-\tau} h_t \Delta L_{R,t+\tau}\right)$$

where $\hat{V}$ is a Heteroskedacticity and Autocorrelation Consistent (HAC) estimator of the variance of $\left(P^{-1/2}\sum_{t=R}^{T-\tau} h_t \Delta L_{R,t+\tau}\right)$. For our empirical application we use a Newey-West (1987) estimator with a bandwidth of $[0.75T^{1/3}]$.[2] At α level of significance, the test rejects when $T^h_{R,P,\tau} > \chi^2_{q,1-\alpha}$.[3]

Suppose the unconditional predictive ability tests fail to reject the equal predictive performance of the models, yet the conditional predictive ability tests do. As suggested by Giacomini and White (2006), the interpretation of this would be that the two models are the same on average, yet their relative predictive performance could be predicted. On the other hand, if the unconditional test rejects the null hypothesis, the conditional tests should as well.[4]

[2] The choice of the bandwidth parameter is motivated by the recommendation in Stock and Watson (2010, p. 599).

[3] In our empirical implementation we typically consider the conditional variables one by one (in addition to a constant) in order to ease the interpretation of the results. In that context the limiting distribution will always be $\chi_{2,1-\alpha}$.

[4] Giacomini and White (2006) document situations when that might not be the case. Their simulation studies suggest that this could be true because the unconditional tests are slightly oversized given the power properties of the HAC estimators. However, this could also be due to the power of the conditional tests. For instance, if we have situations where the test function h_t includes elements of information set that are at most weakly correlated with the relative performance of the models, then the power of the test will deteriorate. To address some of this econometric issues, in addition to employing a HAC estimator for the variance-covariance matrix of the of

2.2 Forecast Improvements

If the relative forecasting ability of the models can be forecasted, then we could use this information in a constructive way by either selecting the best model for a particular future date or by proposing a model averaging technique that could potentially improve the forecasting performance of the models. We contrast these strategies with a benchmark autoregressive model, as well as with a the simple average, an averaging technique that is a competitive benchmark (see Stock and Watson, 2004).

2.2.1 Model Selection

As a model selection criteria we empirically evaluate Giacomini and White's (2006) model selection rule. More specifically, we divide our out-of-sample period P into two parts: a first sample will be used to "train" the rule and a second sample to evaluate it. Let S be the window size for the implementation of the rule. We then follow the two step rule:

1. Regress the loss differences $\Delta\hat{L}_{S,t+\tau}$ on a single conditioning variable h_t, over the first S observations of the out of sample, $t = R+\tau+1, ..., R+S+\tau$, and denote the regression coefficient as $\hat{\delta}^{\tau}_{S,t}$;

2. Predict $y_{R+S+\tau+1}$ using the forecast of the benchmark model if $\hat{\delta}^{\tau\prime}_{S,t} h_{R+S+\tau+1} > 0$ and use the alternative model otherwise[5];

3. Repeat steps (1) and (2) till $t = T-S-\tau, ..., T-\tau-1$, rolling through the out of sample each time using an estimation sample of size S.

We further consider a modified version of this rule in that we use the information on the statistical significance of $\hat{\delta}^{\tau}_{S,t}$. In other words, in this version of model selection we follow the following steps:

1. Regress the loss differences $\Delta\hat{L}_{S,t+\tau}$ on a single conditioning variable h_t, over the first S observations of the out of sample, $t = R+\tau+1, ..., R+S+\tau$, and denote the regression coefficient as $\hat{\delta}^{\tau}_{S,t}$;

2. Check the statistical significance of $\hat{\delta}^{\tau}_{S,t}$ using a two-sided test. If the coefficient is significant at 5% significance level, then proceed to step 3, otherwise, select the benchmark model and proceed to step 4.

$\left(P^{-1/2}\sum_{t=R}^{T-\tau} h_t \Delta L_{R,t+\tau}\right)$, we use a fixed regressor bootstrap as in Clark and McCracken (2013, p. 1134). Second, given that most conditioning variable we consider could be highly correlated with each other we employ the bootstrap aggregation technique suggested by Inoue and Kilian (2008) designed for regression models with correlated regressors.

[5] Please note that a positive value for a loss differential implied that the benchmark model is better than the alternative and vice versa.

3. Predict $y_{R+S+\tau+1}$ using the forecast of the benchmark model if $\hat{\delta}_{S,t}^{\tau\prime} h_{R+S+\tau+1} > 0$ and use the alternative model otherwise;

4. Repeat steps (1) and (2) till $t = T - S - \tau, ..., T - \tau - 1$, rolling through the out of sample each time using an estimation sample of size S.

The above strategies select only one model, either the benchmark or the alternative, at a given point in time.

2.2.2 Model Averaging

Alternatively, we propose a rule for model averaging where instead of selecting only one model at each origin, we take a weighted average of the benchmark and alternative. The rule can be implemented as follows:

1. Regress the loss differences $\Delta\hat{L}_{S,t+\tau}$ on a single conditioning variable h_t, over the first S observations of the out of sample, $t = R + \tau + 1, ..., R + S + \tau$, and denote the regression coefficient as $\hat{\delta}_{S,t}^{\tau}$;

2. The forecast $\hat{y}_{R+S+\tau+1}$ is constructed as: $\hat{y}_{R+S+\tau+1} = w_{o,j}\ f_{t,R}\left(\hat{\beta}_{0,t}\right) + w_{1,j} g_t\left(\hat{\beta}_1, t\right)$ where the weight assigned to the alternative model is:

$$w_{o,t} = (1/S) \sum_{t=R+\tau+1}^{R+S+\tau} 1\left\{\hat{\delta}_{S,t}^{\tau\prime} h_t < 0\right\}, t = R + 1 + \tau, ..., R + S + \tau$$

   and the weight of the benchmark model is $w_{1,t} = 1 - w_{o,t}$.

3. Repeat steps (1) and (2) till $t = T - S - \tau, ..., T - \tau - 1$, rolling through the out of sample each time using an estimation sample of size S.

We further consider a modified version of this averaging rule in that we look only at the alternative models with $\hat{\delta}_{S,t}^{\tau}$ statistically different than zero at 5% significance level.

3 Forecasting Models

We consider forecasting monthly output growth and inflation τ-periods into the future using autoregressive distributed lag (ADL) models, where we consider lags of one predictor at a time in addition to the lagged dependent variable. The forecasting model is:

$$Y_{t+\tau} = \beta_{k,0} + \beta_{k,1}(L) X_{t,k} + \beta_{k,2}(L) Y_t + u_{k,t+\tau},\ t = 1, ..., T - \tau \tag{2}$$

where the dependent variable is either $Y_{t+\tau} = (1200/\tau)\ln(IP_{t+\tau}/IP_t)$ for output growth or $Y_{t+\tau} = (1200/\tau)\ln(CPI_{t+\tau}/CPI_t) - 1200\ln(CPI_t/CPI_{t-1})$ for inflation; $IP_{t+\tau}$ and $CPI_{t+\tau}$ are the industrial production (IP) index and the consumer price index (CPI), respectively. $X_{t,k}$ denotes the k-th explanatory variable, for $k = 1, ..., K$ and $u_{k,t+\tau}$ is the error term. The total number of individual economic variables considered in our application is $K = 117$.[6] Y_t is either the period t output growth, that is $Y_t = 1200\ln(IP_t/IP_{t-1})]$ or the period t change in inflation, that is $Y_t = 1200\ln(CPI_t/CPI_{t-1}) - 1200\ln(CPI_{t-1}/CPI_{t-2})]$.[7] We consider $\tau = 1, 12$ corresponding to one-month-ahead and one-year-ahead forecast horizons. The regression coefficients are the lag-polynomials $\beta_{k,1}(L) = \sum_{j=0}^{p}\beta_{k,1j}L^j$ and $\beta_{k,2}(L) = \sum_{j=0}^{q}\beta_{k,2j}L^j$, with L being the lag operator. We estimate the number of lags (p and q) recursively by BIC, first selecting the lag length for the autoregressive component, then augmenting with an optimal lag length for the additional predictor. The maximum number of lags considered in each case is 12, which is motivated by the monthly nature of the data.

As a benchmark, we consider the autoregressive model, where we use only the lagged dependent variable to forecast output growth and inflation. In other words, the benchmark model is:

$$Y_{t+\tau} = \beta_0 + \beta_2(L)Y_t + u_{t+\tau},\ t = 1, ..., T - \tau \tag{3}$$

The estimation is conducted based on a fixed rolling window scheme, where at each point in time we use the last 120 observations for estimation. This corresponds to 10 years of data. The choice of the forecasting scheme is due to the theoretical validity of the conditional predictive ability tests. Giacomini and White (2006) framework requires the number of observations used in the estimation to stay finite relative to the overall sample size.

4 Data

The data used for forecasting comes from the monthly macroeconomic database of McCracken and Ng (2015).[8] The dataset covers various categories, namely, it includes measures of (i) output and income; (ii) labor market indicators; (iii) housing; (iv) consumption, orders, and inventories; (v) money and credit; (vi) exchange rate; (v) prices, and (vi) stock prices. For the matters of data construction we refer to McCracken and Ng (2015) for details.

For the purposes of this paper we use all their series with the exception of those that start later than 1959:M1. These include the series new private housing permits (SAAR) and its various geographic counterparts, i.e. the permits covering northeast, midwest, south and west. In

[6] The dataset for output growth includes historical data for inflation, but not output growth (and vice versa), as the lagged dependent variable is automatically included in eq. (3).

[7] Note that, like Stock and Watson's (2003) approach, this relies on the assumption that inflation is I(2).

[8] The data is publicly available at https://research.stlouisfed.org/econ/mccracken/fred-databases/

addition, we also exclude the series new orders for consumer as well as durable goods. The next omitted series is the trade weighted U.S. dollar index against major currencies. We also exclude the consumer sentiment index and VXO.[9] We have a total of 117 series. The sample period ends on 2016M1, yielding a total of 685 observations. The data has been transformed to stationarity using the proposed transformations in McCracken and Ng (2015). The mnemonics for target variables correspond to CPIAUCSL (CPI all items) and INDPRO (IP index). We use the September 2016 vintage of the monthly database.

Moreover, in our empirical application we adjust for outliers. We treat a realization that is 4 standard deviations larger than the mean as an outlier. We substitute the outliers with the mean.[10] Given the sample starting period, the number of observations lost due to data transformations as well as the 10 year rolling window used for estimation, the out-of-sample evaluation period across models starts in 1970:M3.

INSERT TABLE 1 HERE

We further consider several conditioning variables. These conditioning variables are divided into four groups: measures of economic activity, financial condition indices, macroeconomic uncertainty indices and measures of past relative performance. Conditioning variables and their samples are summarized in Table 1. The choice of the conditioning variables is motivated by the availability of the data going back to 1970:M2 to make the out-of-sample conditional predictive ability tests feasible. Moreover, we are looking at variables that the literature has documented to be important for understanding the state and properties of the business cycle. The conditioning variables are discussed in more detail below.

Business cycle indicators

Chauvet and Potter (2013) and Stock and Watson (2010), among others, find that relative forecasting performance differs across phases of the business cycle for output growth and inflation respectively so we consider as conditioning variable a dummy that takes the value one in periods of economic recessions and zero during expansions, as indicated by NBER Business Cycle dating committee. However, NBER recession dates are usually known with a lag, so we also construct alternative measures of the business cycle. More specifically, we consider the recession probability index of Chauvet and Piger (2008). We also construct two binary variables: (i) "ip-rec" which takes the value of one when the industrial production index experiences a negative cumulative growth over six months; (ii) "unemp-rec" when the unemployment rate in the economy is above 5%.

[9] The mnemonics for these series accordingly are PERMIT, PERMITNE, PERMITMW, PERMITS, PERMITW, ACOGNO, ANDENOx, TWEXMMTH, UMCSENTx and VXOCLSx, respectively.

[10] In McCraken and Ng (2015) an outlier is defined as an observation that deviates from the sample median by more than ten interquartile ranges. The outliers are removed and treated as missing values in their case.

Financial conditions/stress indicators

Motivated by Ng and Wright (2013) we consider whether the relative forecasting performance of the models depend on the financial conditions. In the benchmark specification we use the National Financial Condition Index (NFCI) of the Chicago Fed which takes positive (negative) values when conditions are tighter (looser) than average. Due to the possible correlation between economic and financial conditions, we also consider the Adjusted National Financial Condition Index (ANFCI), which extracts a component of financial conditions uncorrelated with economic conditions. Moreover, to disentangle different aspects of financial conditions, we also look at three subindexes of the NFCI index: risk, credit and leverage. The first one captures volatility and funding risk in the financial sector, the second credit conditions, the last debt and equity measures. For robustness we also use alternative measures of financial stress indices produced by the Federal Reserve Banks of St. Louis, Cleveland and Kansas City, "SFSI", "CFSI" and "KFSI", respectively. The samples of these series are unfortunately shorter.

Uncertainty Indices

In light of the importance of the impact of uncertainty into the macroeconomy as discussed in Jurado et al. (2015), we also consider the relevance of uncertainty. We use several measures of uncertainty, namely a longer series of VXO which we construct based on the methodology described in Bloom (2009) by putting together the VXO series from Chicago Board Options Exchange (CBOE) and the actual volatility of the S&P500 index. Further we also use the uncertainty indices of Jurado et al. (2015): the macro uncertainty and finance uncertainty indeces at horizons one, three and twelve months. The uncertainty indices are important also in the light of the fairly vast literature documenting the relationship between uncertainty and levels of macroeconomic variables such as inflation.

Past Relative Performance

Finally, the last conditioning category includes measures of past relative performance: it is the lagged loss difference in the squared forecast errors between the benchmark and the alternative model. We consider the relative forecasting performance of the last period, as well as the smoothed average of the past 5 years.[11]

INSERT TABLE 3 HERE

Table 3 shows the correlation among the conditioning variables. The real time recessionary

[11] We should note there could potentially be a mapping between the uncertainty indices and the variables accounting for the past performance of the models, since, for instance the Jurado et al. (2015) uncertainty measures are forecast error based measures.

dummy based on the growth of industrial production and the one based on the unemployment rate show a very low correlation (even negative for the unemployment dummy) with the other conditioning variables. The other conditioning variables show higher positive correlations, but only in a handful of cases the correlation reaches above eighty percent, suggesting that the information content provided by these variables is not perfectly overlapping.

5 Results

5.1 Predictive Ability Tests

We evaluate the unconditional and conditional predictive ability of the models and methods described in the previous section using the testing framework introduced in section 2. Figure 1 shows the results for the unconditional predictive ability test. The horizonthal axis displays the root mean squared forecast errors relative to the autoregressive model. Ratios greater than one, i.e. to the right of the vertical red line, indicate that the considered models performance is worse than the autoregressive benchmark. The vertical axis indicates the $p-$values from the Giacomini and White unconditional test with the 10 per cent significance level marked by a horizonthal red line. Each dot denotes the result from one of our 117 bimodel comparisons. Then, models which significantly outperform the benchmark will be located in the lower left quadrant of each panel. In line with previous literature we find that unconditional equal predictive ability tests reject only in a handful of cases. Moreover, when the forecasts are statistically significantly different from each other, then usually the economic models are worse than the autoregressive benchmark.

INSERT FIGURE 1 AND TABLE 2 HERE

Table 2, on the other hand, lists the models which are on average statistically better than the benchmark. Comparing Panels A and B, we see that there is a lot less forecastability in inflation than in industrial production growth. In fact, only the model with a real M2 measure delivers statistically significantly different results from the benchmark. Moreover, there is more forecastabiliy in output growth at longer horizon ($h = 12$) relative to the one-month-ahead ($h = 1$). At one-month-ahead forecast horizon measures of real economic activity, i.e. industrial production in the manufacturing sector, help-wanted index, initial unemployment claims as well as the average weekly manufacturing hours are the statistically relevant variables. On the other hand, besides from capacity utilization and real M2 series, the forecastability of industrial production growth at one-year-ahead horizon comes primarily from asset prices.

Results for the conditional ability tests are provided in Figure 2. For the conditional test the figure shows on the horizonthal axis the proportion of times over the out-of-sample in which the

decision rule chooses the benchmark model, i.e. the proportion of times the benchmark model is better than the alternative model given the value taken by conditioning variable. Recall that $\hat{\delta}'_{P,t}h_t \approx E\left[\Delta L_{P,t+\tau}|\mathcal{G}_t\right]$; in practice we compute the statistic $I_{GW} = \frac{1}{P}\sum_{t=R+\tau+1}^{T} I\left\{\hat{\delta}'_{P,t}h_t > 0\right\}$. Since we calculated the loss differential as that of the squared loss of the economic and benchmark models, the smallest the statistic, the better the performance of the alternative model. We mark the significance of the marginal effects of the coefficients based on the Giacomini and White (2006) conditional preditive ability tests at 10% significance level. Therefore models that perform significantly better than the benchmark will be located on the left lower quadrant of the figures. For both target variables (i.e. industrial production and inflation) and for both forecasting horizons we only report results for the conditioning variable that give us the highest number of rejections[12].

In general the conditional test rejects more frequently than the unconditional test at both horizons, especially for industrial production at twelve steps ahead. As observed for the unconditional predictive ability test, there is a lot less forecastability in inflation than in industrial production growth. At one step ahead forecasting horizon conditioning on current financial conditions provides with further information regarding the future relative predictive ability of the models. At twelve steps ahead, for both variables rejections are more frequent when conditioning on past performance.

Table 4 lists the models which perform significantly better than the benchmark, i.e. the models for which the statistic I_{GW} is lower than 0.5 and the p-value from the conditional predictive ability test is lower than 0.10. The relative performance column shows the statistic:

$$M_{GW} = \frac{\sum_{t=R+\tau+1}^{T} \mid \hat{\delta}'_{P,t}h_t \mid I\left\{\hat{\delta}'_{P,t}h_t > 0\right\}}{\sum_{t=R+\tau+1}^{T} \mid \hat{\delta}'_{P,t}h_t \mid}$$

which is bounded between zero and one. This gives us an idea of the magnitude of the improvement of the alternative model over the benchmark model. Because $\hat{\delta}'_{P,t}h_t \approx E\left[\Delta L_{P,t+\tau}|\mathcal{G}_t\right] = E\left[\varepsilon^2_{0,t+\tau} - \varepsilon^2_{1,t+\tau}|\mathcal{G}_t\right]$ a value close to zero indicates that the alternative model has a much better performance than the benchmark. Then, this paper further contributes to the literature by suggesting this new statistic to summarize the conditional, relative performance of the models.

INSERT TABLE 4 HERE

While for the twelve step ahead forecasting horizon the usefulness of asset prices in predicting industrial production emerged also in the unconditional evaluation, for the one step ahead is picked up only from the conditional test. For inflation oil prices are important at one step ahead. At twelve step ahead on top of money measures, real activity measures and in particular measures of unemployment prove useful. Then conditioning test finds evidence of an empirical relashionship between inflation and money measures and between inflation and unemployment.

[12] Additional results are available from the authors upon request.

We interpret rejection of the null of conditional equal predictive ability as indication of misspecification of the models, as the conditioning variable represents information available at the time the forecasts are made that is able to explain the relative performance of the models. Following a rejection then, a researcher aiming at improving the accuracy of the forecasts can adopt two strategies: (i) modify the original models to incorporate the information provided by the conditioning variable or (ii) adopt the simple model averaging rule described in the next subsession. The first strategy requires to formulate a specification of a new forecasting model as well as to estimate the new model, and produce the forecasts, while the second strategy is based on the forecasts of the benchmark and alternative models which are already available.

5.2 Decision Rule

We evaluate the usefulness of the information contained in the conditioning variables by implementing the model selection and the model averaging strategies outlined in Section 2.2. The goal of this exercise is to assess whether we can ultimately produce more accurate forecasts, either by selecting or averaging across models, given that the relative performance of the forecasting models can be predicted by the conditioning variables. To apply these strategies we first need to split the overall forecast sample into two subsamples: one for the training of the rule and one for its evaluation. We choose the window size for the implementation of the rule to be ten years, S=120. Given the size of the out-of sample, $P = 685$, this leaves us with 465 observations for the evaluation of the decision rule. Then, for each conditioning variable $n = 1, ..., N$, and for each forecasting model $m = 1, .., M$ we produce forecasts of the target variables at one-step ahead and twelve-step ahead following the steps detailed above. We then compute the RMSE associated with the forecasts produced with those rules and compared them to the RMSFE of the benchmark model. We report only the results for the averaging rule as they are very similar to the ones for the selection rule.

INSERT FIGURE 3 HERE

Figure 3 shows for each conditioning variable, the relative RMSE of the decision rule versus the benchmark. The figure plots only the models for which the decision rule provides a lower RMSE than the benchmark. Gains are larger at twelve steps ahead than at one step ahead, for industrial production than inflation. Reductions in RMSFE can reach 12% which is a large number compared to the literature. Figure 4 shows for selected conditioning variables, the models that perform better than the benchmark. For industrial production, models including measures of real activity and asset prices are successful both at one and twelve steps ahead. For inflation on top of real activity measures and asset prices, also commodity prices and money measures are helpful predictors.

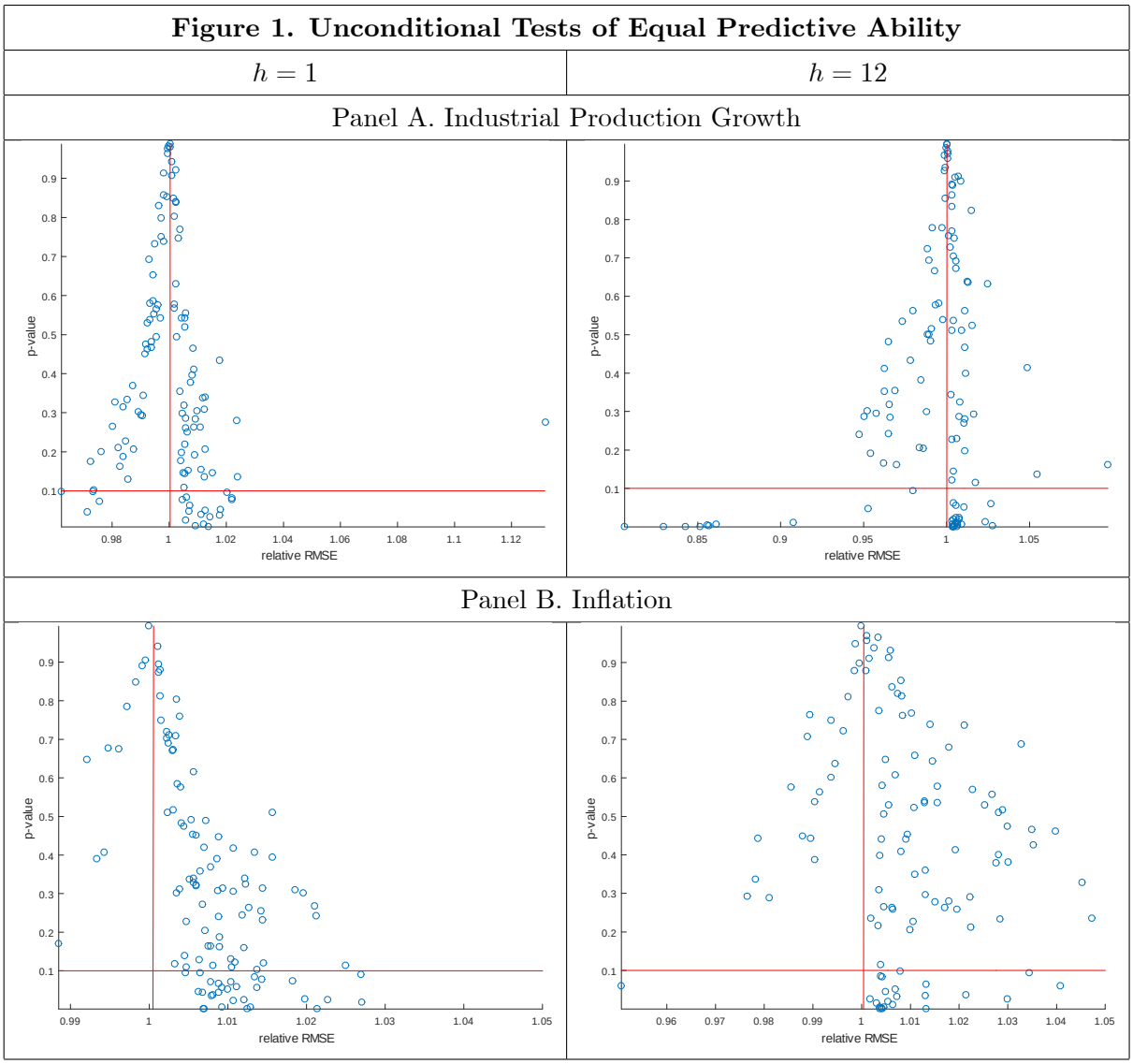
INSERT FIGURE 4 HERE

6 Conclusions

In this paper we conducted a systematic evaluation of the conditional predictive ability of various economic variables that represented asset prices, measures of real economic activity, wages and prices, as well money. We consider a wide range of autoregressive distributed lag models for forecasting. We ask whether the relative performance of the models depends on the state of the economy, financial conditions, macroeconomic uncertainty or whether it can be predictive based on past out-of sample relative accuracy. We find that for both variables at longer horizons the past relative performance is a good indicator of future relative performance, while for the short run financial conditions have predictive content for next period's relative performance. Our results suggest using the conditional test in an informative way. In particular, we document that using the conditioning information as a criteria for model selection in fact performs better than the single forecasting models.

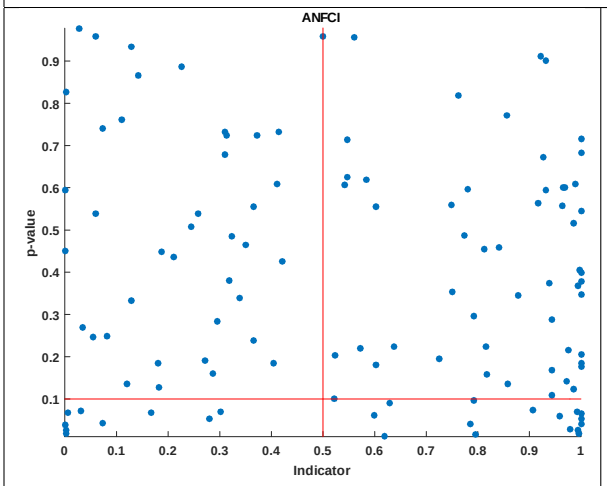
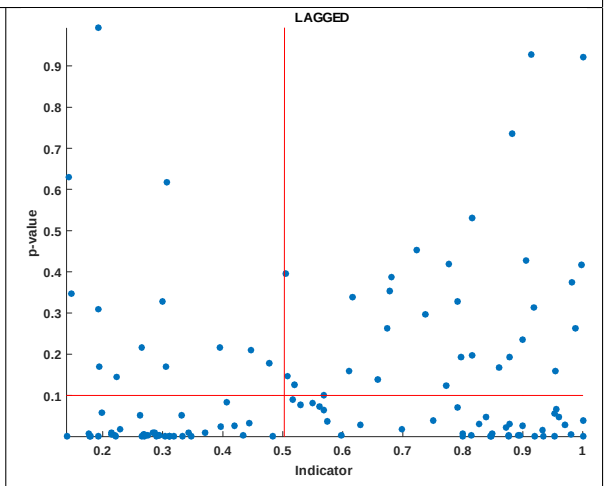
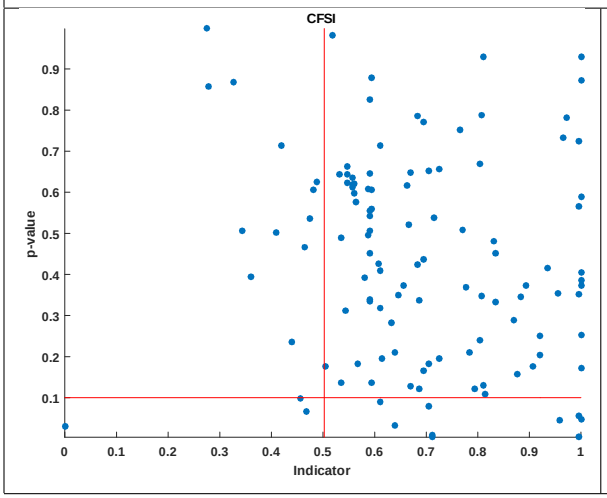
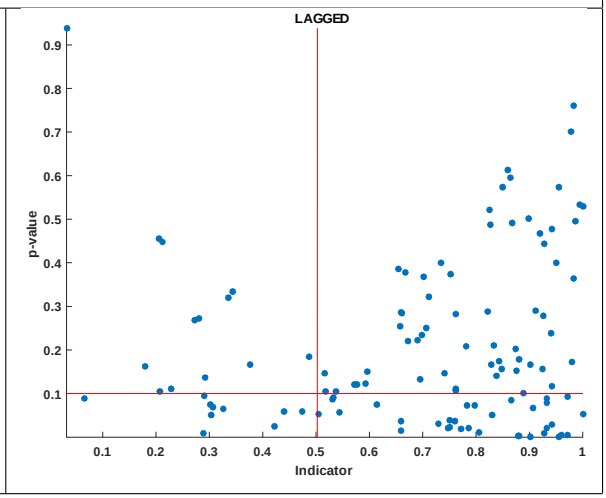
7 Figures and Tables

Figure 1. Unconditional Tests of Equal Predictive Ability



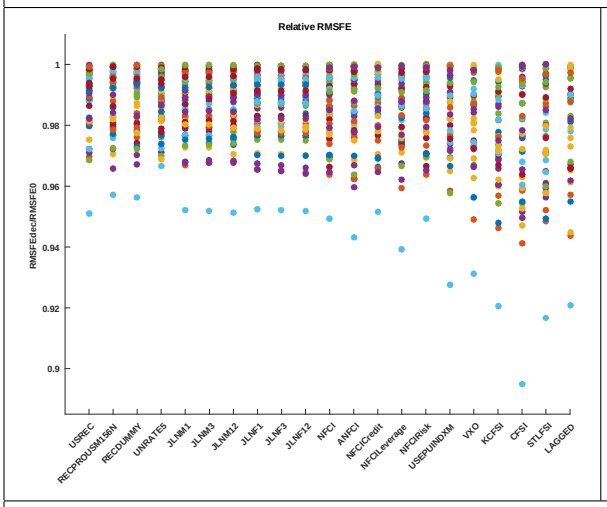
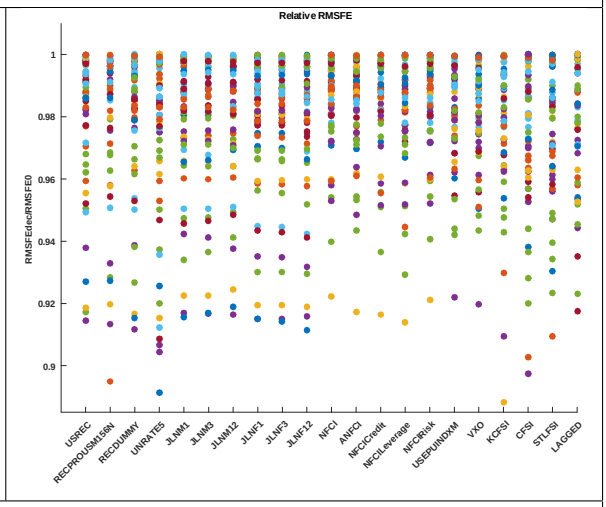
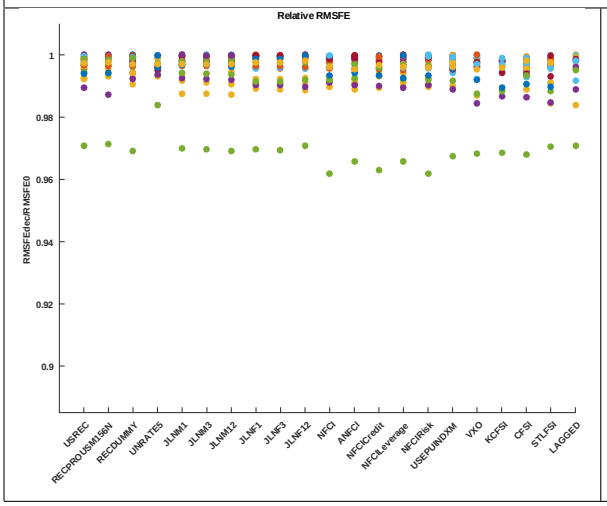
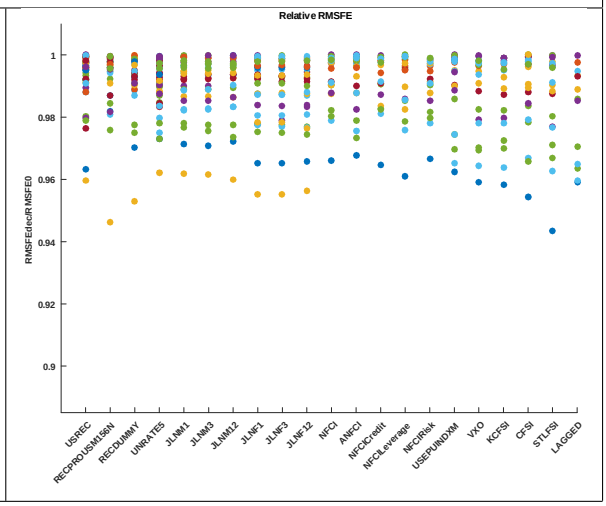
Notes: The figure shows the unconditional test results for the wide range of models for inflation and industrial production growth. The benchmark is the autoregressive model, while the alternatives are autoregressive distributed lag models, where we consider each economic variable one at a time. For the relative RMSE, values greater than one indicate favor the benchmark model.

Figure 2. Conditional Tests of Equal Predictive Ability

$h = 1$	$h = 12$
Panel A. Industrial Production Growth	
	
Panel B. Inflation	
	

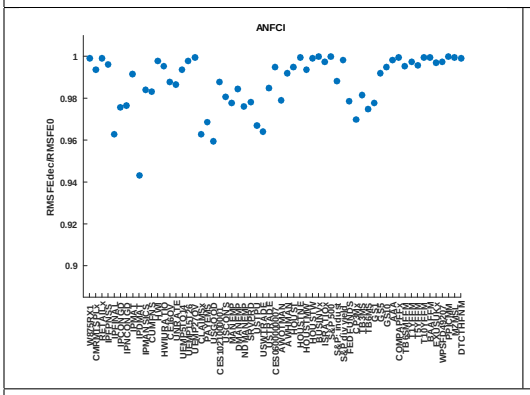
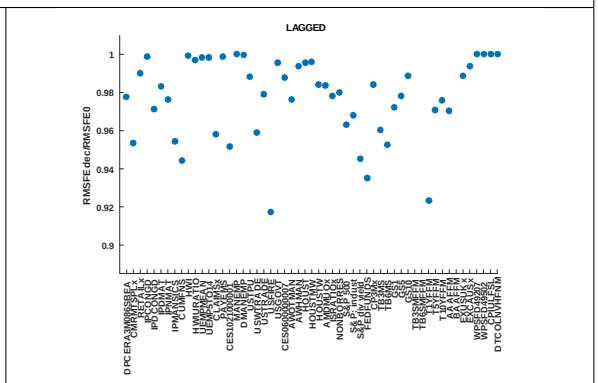
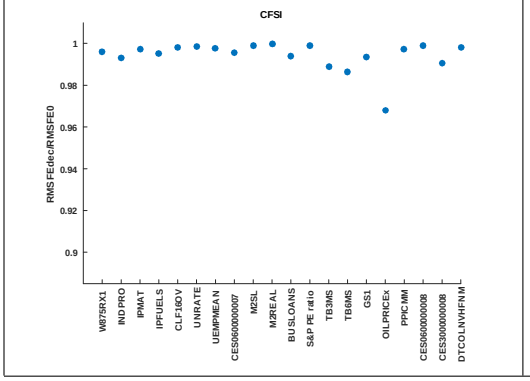
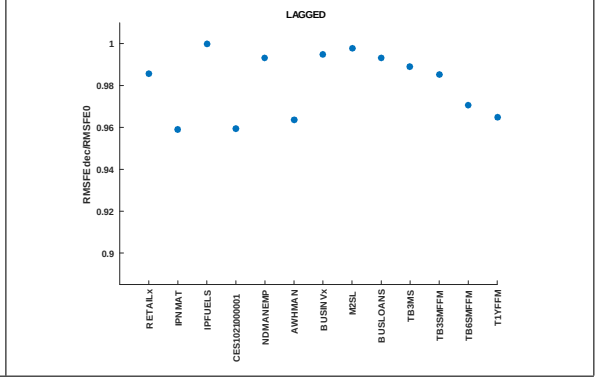
Notes: The figure shows the unconditional test results for the wide range of models for inflation and industrial production growth. The benchmark is the autoregressive model, while the alternatives are autoregressive distributed lag models, where we consider each economic variable one at a time. For the relative RMSE, values greater than one indicate favor the benchmark model.

Figure 3. Conditional Tests of Equal Predictive Ability: Relative RMSE

$h = 1$	$h = 12$
Panel A. Industrial Production Growth	
	
Panel B. Inflation	
	

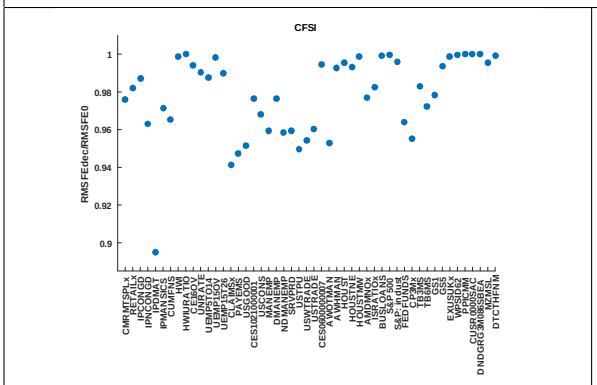
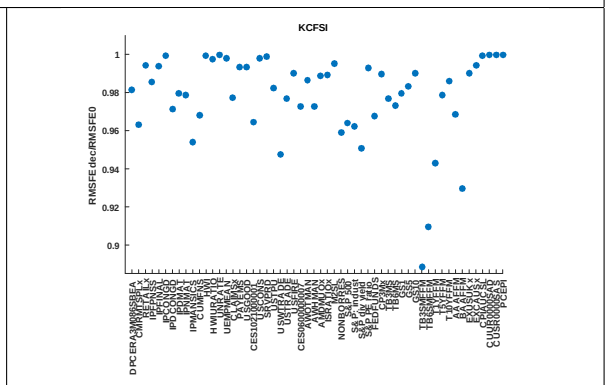
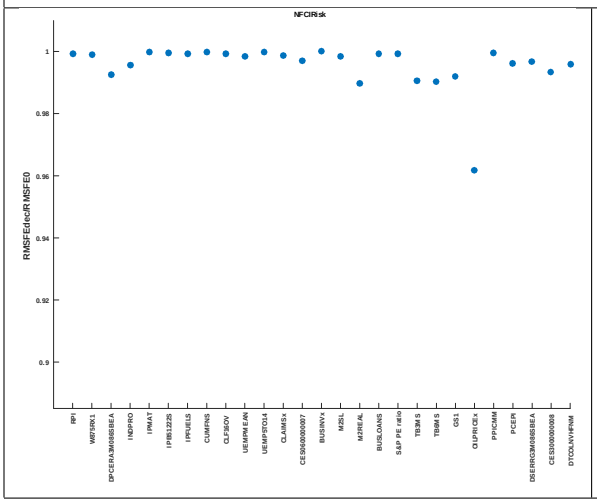
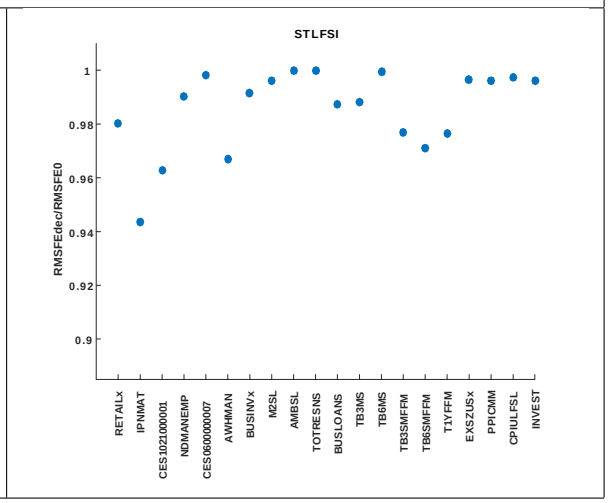
Notes: The figure shows the unconditional test results for the wide range of models for inflation and industrial production growth. The benchmark is the autoregressive model, while the alternatives are autoregressive distributed lag models, where we consider each economic variable one at a time. For the relative RMSE, values greater than one indicate favor the benchmark model.

Figure 4.a Conditional Tests of Equal Predictive Ability: Relative RMSE

$h = 1$	$h = 12$
Panel A. Industrial Production Growth	
	
Panel B. Inflation	
	

Notes: The figure shows the unconditional test results for the wide range of models for inflation and industrial production growth. The benchmark is the autoregressive model, while the alternatives are autoregressive distributed lag models, where we consider each economic variable one at a time. For the relative RMSE, values greater than one indicate favor the benchmark model.

Figure 4.b Conditional Tests of Equal Predictive Ability: Relative RMSE

$h = 1$	$h = 12$
Panel A. Industrial Production Growth	
	
Panel B. Inflation	
	

Notes: The figure shows the unconditional test results for the wide range of models for inflation and industrial production growth. The benchmark is the autoregressive model, while the alternatives are autoregressive distributed lag models, where we consider each economic variable one at a time. For the relative RMSE, values greater than one indicate favor the benchmark model.

Table 1. Description of Conditioning Variables

Label	Starting Date	Description	Source
		Business Cycle Indicators	
NBER-rec	1959M1	NBER recession dates: from Peak to Trough	F
prob-rec	1967M6	Smoothed US recession probabilities, percent	F
ip-rec	1959M1	Six months of negative industrial production growth	F*
unemp-rec	1959M1	Unemployment rate above 5%	F*
		Financial Conditions/Stress Indicators	
NFCI	1973M1	National Financial Conditions Index	F
ANFCI	1973M1	Adjusted National Financial Conditions Index	F
NFCI-Credit	1973M1	National Financial Conditions: Credit Subindex	F
NFCI-Leverage	1973M1	National Financial Conditions: Leverage Subindex	F
NFCI-Risks	1973M1	National Financial Conditions: Risk Subindex	F
STLFSI	1994M1	St. Louis Fed Financial Stress Index	F
CFSI	1991M10	Cleveland Fed Financial Stress Index	F
KFSI	1990M2	Kansas City Fed Financial Stress Index	F
		Uncertainty Indices	
VXO	1986M1	CBOE Implied Volatility Index based on S&P100 options	C
BBD	1985M1	Economic Policy Uncertainty Index	BBD
JLN-Macro	1960M7	Macroeconomic Uncertainty Index, 1-3 and 12 periods ahead	JLN
JLN-Financial	1960M7	Financial Uncertainty Index, 1-3 and 12 periods ahead	JLN
		Measures of Past Relative Performance	
rMSFE-1		relative MSFE for the past period	O
rMSFE-5		relative MSFE for the past 5 years	O

Notes: Sources are abbreviated as follows: "F"- Federal Reserve Economic Data (FRED), "BBD"- Baker et al. (2016), "JLN"- Jurado et al. (2015), "O "- calculations from the paper, "* "- indicates additional calculations on the source data .

Table 2. Unconditional Tests of Equal Predictive Ability

	h=1			h=12	
Model	relative RMSE	p-value	Model	relative RMSE	p-value
Panel A. Industrial Production					
IP: Durable Materials	0.96	0.10	Capacity Utilization: Manufacturing	0.98	0.10
Help-Wanted Index	0.97	0.10	Real M2 Money Stock	0.86	0.01
Initial Claims	0.97	0.05	S&P's Common Stock Price Index: Industrials	0.95	0.05
Avg Weekly Hours: Manufacturing	0.98	0.07	3-Month Treasury C Minus FEDFUNDS	0.85	0.00
			6-Month Treasury C Minus FEDFUNDS	0.85	0.00
			1-Year Treasury C Minus FEDFUNDS	0.91	0.01
			5-Year Treasury C Minus FEDFUNDS	0.86	0.00
			10-Year Treasury C Minus FEDFUNDS	0.86	0.00
			Moody's Aaa Corporate Bond Minus FEDFUNDS	0.83	0.00
			Moody's Baa Corporate Bond Minus FEDFUNDS	0.81	0.00
Panel B. Inflation					
			Real M2 Money Stock	0.95	0.06

Notes: The table shows the unconditional test for models which are statistically different than the benchmark at 10% significance level.

Table 3. Conditioning Variables Cross-Correlation

	REC	REC-P	IP-REC	UNE	JLNM1	JLNM3	JLNM12	JLNF1	JLNF3	JLNF12	NFCI	ANFCI	NFCI-C	NFCI-L	NFCI-R	EPUN	VXO	KCFSI	CFSI	STLFSI
REC	1	0.88	0.46	0.06	0.62	0.62	0.59	0.48	0.47	0.33	0.59	0.24	0.58	0.44	0.56	0.27	0.40	0.65	0.45	0.55
REC-P		1	0.45	0.09	0.69	0.68	0.64	0.50	0.49	0.33	0.61	0.27	0.62	0.44	0.58	0.27	0.46	0.80	0.50	0.63
IP-REC			1	0.10	0.49	0.48	0.43	0.35	0.35	0.40	0.31	0.07	0.32	0.06	0.30	0.14	0.26	0.46	0.35	0.32
UNE				1	0.07	0.06	0.09	-0.17	-0.18	-0.18	0.09	-0.19	0.15	-0.07	0.07	0.39	-0.11	-0.09	-0.01	-0.18
JLNM1					1	0.99	0.96	0.58	0.58	0.41	0.76	0.30	0.76	0.53	0.72	0.35	0.55	0.85	0.61	0.56
JLNM3						1	0.98	0.60	0.60	0.43	0.80	0.34	0.78	0.56	0.76	0.34	0.59	0.87	0.64	0.59
JLNM12							1	0.54	0.54	0.39	0.84	0.38	0.84	0.58	0.82	0.30	0.61	0.89	0.65	0.66
JLNF1								1	0.99	0.72	0.52	0.34	0.35	0.44	0.54	0.37	0.84	0.84	0.64	0.65
JLNF3									1	0.72	0.52	0.34	0.35	0.44	0.54	0.37	0.83	0.83	0.64	0.64
JLNF12										1	0.34	0.25	0.24	0.28	0.34	0.24	0.55	0.54	0.55	0.33
NFCI											1	0.62	0.84	0.66	0.99	0.36	0.70	0.94	0.68	0.77
ANFCI												1	0.36	0.57	0.65	0.01	0.50	0.55	0.37	0.61
NFCI-C													1	0.48	0.78	0.40	0.57	0.89	0.68	0.68
NFCI-L														1	0.66	0.22	0.40	0.66	0.54	0.49
NFCI-R															1	0.30	0.72	0.94	0.66	0.81
EPUN																1	0.0	0.40	0.52	0.02
VXO																	1	0.80	0.64	0.71
KCFSI																		1	0.71	0.77
CFSI																			1	0.37
STLFSI																				1

Notes: Cross-correlation across conditioning variables. "REC": NBER recession probability; "REC-P": Smoothed US recession probabilities "UNE": Dummy for unemployment rate above 5%

Table 4. Conditional Tests of Equal Predictive Ability

	h=1			h=12	
Model	relative perform	p-value	Model	relative perform	p-value
Panel A. Industrial Production (ANFCI)			**(LAGGED)**		
IP: Durable Materials	0.07	0.07	Real personal consumption expenditures	0.40	0.00
Avg Weekly Overtime Hours: Manufacturing	0.00	0.07	IP: Nondurable Materials	0.38	0.05
Effective Federal Funds Rate	0.01	0.07	Help-Wanted Index for United States	0.37	0.05
3-Month AA Financial Commercial Paper Rate	0.02	0.04	Ratio of Help Wanted/No. Unemployed	0.25	0.06
3-Month Treasury Bill	0.00	0.04	All Employees: Mining and Logging: Mining	0.32	0.01
6-Month Treasury Bill	0.00	0.02	All Employees: Trade, Transportation & Utilities	0.49	0.03
1-Year Treasury Rate	0.00	0.03	All Employees: Wholesale Trade	0.46	0.03
6-Month Treasury C Minus FEDFUNDS	0.27	0.05	All Employees: Retail Trade	0.40	0.01
1-Year Treasury C Minus FEDFUNDS	0.32	0.07	All Employees: Goods-Producing Industries	0.44	0.08
			All Employees: Mining and Logging: Mining	0.46	0.00
			Avg Weekly Hours: Manufacturing	0.52	0.00
			Total Business Inventories	0.36	0.00
			Real M2 Money Stock	0.21	0.00
			Nonrevolving consumer credit to Personal Income	0.46	0.03
			S&P 500	0.36	0.00
			S&P: industrials	0.30	0.00
			S&P dividend yield	0.37	0.01
			S&P Price-Earnings Ratio	0.39	0.01
			Effective Federal Funds Rate	0.38	0.00
			3-Month AA Financial Commercial Paper Rate	0.45	0.00
			3-Month Treasury Bill	0.41	0.00
			6-Month Treasury Bill	0.38	0.00
			1-Year Treasury Rate	0.37	0.00
			5-Year Treasury Rate	0.41	0.00
			10-Year Treasury Rate	0.42	0.00
			Moody's Seasoned Aaa Corporate Bond Yield	0.38	0.00
			Moody's Seasoned Baa Corporate Bond Yield	0.38	0.01
			3-Month Commercial Paper Minus FEDFUNDS	0.36	0.00
			3-Month Treasury C Minus FEDFUNDS	0.17	0.00
			6-Month Treasury C Minus FEDFUNDS	0.14	0.00
			1-Year Treasury C Minus FEDFUNDS	0.25	0.00
			5-Year Treasury C Minus FEDFUNDS	0.24	0.00
			10-Year Treasury C Minus FEDFUNDS	0.25	0.00
			Moody's Aaa Corporate Bond Minus FEDFUNDS	0.22	0.00
			Moodyís Baa Corporate Bond Minus FEDFUNDS	0.20	0.00
			CPI : Medical Care	0.31	0.02
			CPI : All Items Less Food	0.36	0.01
Panel B. Inflation: (KCFSI/CSFI/STLSFI)			**(LAGGED)**		
KCFSI: IP: Final Products	0.73	0.01	Real Personal Income	0.45	0.06
All Employees: Trade, Transportation & Utilities	0.75	0.07	Ratio of Help Wanted/No. Unemployed	0.36	0.01
Crude Oil, spliced WTI and Cushing	0.00	0.04	Civilian Labor Force	0.45	0.02
CSFI: IP: Residential Utilities	0.56	0.10	Civilians Unemployed for 15-26 Weeks	0.49	0.06
Ratio of Help Wanted/No. Unemployed	0.58	0.07	All Employees: Service-Providing Industries	0.39	0.07
Crude Oil, spliced WTI and Cushing	0.00	0.03	Housing Starts: Total New Privately Owned	0.38	0.07
STLSFI: S&P Price-Earnings Ratio	0.54	0.09	New Orders for Durable Goods	0.51	0.06
Canada / U.S. Foreign Exchange Rate	0.61	0.05	M1 Money Stock	0.43	0.06
Crude Oil, spliced WTI and Cushing	0.00	0.08	St. Louis Adjusted Monetary Base	0.08	0.09
			6-Month Treasury C Minus FEDFUNDS	0.36	0.10

Notes: The table shows the unconditional test for models which are statistically different than the benchmark at 10% significance level.

References

[1] Baker, S.R., N. Bloom and S.J. Davis (2016), "Measuring Economic Policy Uncertainty," *Quarterly Journal of Economics*, forthcoming.

[2] Bloom, N. (2009). "The Impact of Uncertainty Shocks," *Econometrica* 77(3), 623-685.

[3] Chauvet, M. and S. Potter (2013), "Forecasting Output," in G. Elliott and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Volume 2A, Elsevier-North Holland Publications.

[4] Clark T.E. and T. Doh (2014), "A Bayesian Evaluation of Alternative Models of Trend Inflation," *International Journal of Forecasting* 30(3), 426-448.

[5] Clark T.E. and M.W. McCracken (2001), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics* 105, 85-110.

[6] Clark T.E. and M.W. McCracken (2003), "Advances in Forecast Evaluation,"in G. Elliott and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Volume 2B, Elsevier-North Holland Publications.

[7] Clark T.E. and K.D. West (2007), "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models," *Journal of Econometrics* 138, 291-311.

[8] Chauvet, M. and J. Piger (2008), "A Comparison of the Real-Time Performance of Business Cycle Dating Methods," *Journal of Business and Economic Statistics* 26, 42-49.

[9] Diebold, F. X. and R. S. Mariano (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13(3), 253-263.

[10] Dotsey, M., Fujita, S. and T. Stark (2011), "Do Phillips Curves Conditionally Help to Forecast Inflation," Philadelphia FED wp. 11-40.

[11] Giacomini, R. and H. White (2006), "Tests of Conditional Predictive Ability," *Econometrica* 74(6), 1545-1578.

[12] Gilchrist, S. and E. Zakrajšek (2012), "Credit Spreads and Business Cycle Fluctuations," *American Economic Review*, 102 (4), 1692-1720.

[13] Inoue, A. and L. Kilian (2008), "How Useful is Bagging in Forecasting Economic Time Series? A Case Study of U.S. Consumer Price Inflation," *Journal of the American Statistical Association*, 103 (482), 511-522.

[14] Jurado, K., S. Ludvigson and S. Ng (2015), "Measuring Uncertainty," *American Economic Review* 105 (3), 1177-1216.

[15] McCracken, M. W. and S. Ng (2016), "FRED-MD: A Monthly Database For Macroeconomic Research," *Journal of Business and Economic Statistics*, 34((4), 574-589.

[16] Ng, S. and J. Wright (2013), "Facts and Challenges from the Great Recession for Forecasting and Macroeconomic Modeling," *Journal of Economic Literature* 51(4), 1120-1154.

[17] Rossi, B. (2013), "Advances in Forecasting Under Instabilities,"in G. Elliott and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Volume 2B, Elsevier-North Holland Publications.

[18] Rossi, B. and T. Sekhposyan (2010), "Have Economic Models' Forecasting Performance for US Output Growth and Inflation Changed over Time, and When?" *International Journal of Forecasting,* 26(4), 808-835.

[19] Stock, J.H. and M.W. Watson (2004), "Combination Forecasts of Output Growth in a Seven Country Data Set,"*Journal of Forecasting* 23(6), 405-430.

[20] Stock, J.H. and M.W. Watson (2007), "Why Has U.S. Inflation Become Harder to Forecast,"*Journal of Money, Credit and Banking* 39(s1), 3-33.

[21] Stock, J.H. and M.W. Watson (2010), *Introduction to Econometrics*, 3rd ed., Addison-Wesley.

[22] West, K.D. (1996), "Asymptotic Inference about Predictive Ability," *Econometrica* 64, 1067-1084.