How to Predict Your Next Forecasting Model: Conditional Predictive Ability Approach*

Eleonora Granziera† and Tatevik Sekhposyan‡

This version: December 1, 2017

Abstract

The relative performance of forecasting models is known to be unstable over time. From a practical point of view it is important to know whether the instabilities are predictable, and if so, whether the predictability can be exploited to improve the accuracy of the economic forecasts. We address this question by evaluating the predictive ability of a wide range of economic variables for two key U.S. macroeconomic aggregates, industrial production and inflation, relative to simple benchmarks. We find that the state of the business cycle, financial conditions, uncertainty as well as measures of past relative performance are, on average, useful for explaining the relative forecasting performance of the models. We further construct a pseudo-real-time forecasting exercise where we use the information about the conditional performance for model selection and model averaging. The proposed strategies deliver sizable improvements, particularly when the models are selected or combined consistent with their relative performance predicted by the financial conditions at the forecast origin date and past performance.

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

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

---

*We are grateful to the participants of the 2017 St. Louis FED Central Bank Forecasting Workshop, 2017 Annual Conference of the International Association of Applied Econometrics, 2017 Conference of Computing in Economics and Finance, 2016 International Symposium on Forecasting and 2015 Annual Symposium of the 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. Tatevik Sekhposyan gratefully acknowledges financial support from SAS-IIF Grant that promotes research on forecasting.

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

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

1 Introduction

The relative forecasting performance of time series models is known to be unstable over time. Stock and Watson (2007), Rossi and Sekhposyan (2010), among others document the deterioration of the relative forecasting performance of economic models over univariate benchmarks around the onset of the Great Moderation[1]. There is also evidence that the improvements and deteriorations of the relative forecasting performance could be associated with recurring periods of economic significance. For instance, Chauvet and Potter (2013) review the relative ability of a wide range of models to forecast output growth in recessions versus expansions. In a similar exercise, Dotsey, Fujita and Stark (2011) evaluate relative forecasting performance of inflation models conditional on the state of the business cycle. Ng and Wright (2013) emphasize the importance of the source of business cycle fluctuations. 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.

Figure 1 shows the squared forecast error differentials between an autoregressive model for US output growth and the same model augmented with housing starts (typically considered as a leading indicator) over time. The plot demonstrates the points above: (i) the relative forecasting performance of the models changes over time; (ii) squared forecast error differentials are close to zero and relatively stable since the 1990-s up to late 2000-s, i.e. beginning of the financial crises; (iii) moreover, the reversals of the forecasting performance (in this case improvements of the economic model, measured by positive values of squared loss differential) are associated with NBER recessions (shaded bars) and can be discrete rather than smooth; (iv) further, one can notice some persistence in the forecasting performance of the models.

INSERT FIGURE 1 HERE

In this paper we consider a wide set (more than hundred) of simple forecasting models of autoregressive distributed type. More specifically, we take the best performing autoregressive model (typically a competitive benchmark) and augment it with optimally chosen lags of an economic variable taken from the McCracken and Ng (2016) monthly database. Adding one economic variable at a time will keep the models parsimonious, yet will allow to retain economic interpretability. We further test whether the predictive ability of economic models relative to simple autoregressions can in itself be predicted by some observables, such as indicators of the phase of the business cycle, financial conditions or the level of economic uncertainty. Thus, we identify certain episodes of economic significance in which the economic models statistically dominate the autoregressive ones. We then investigate whether the information provided by the conditioning variables can help us

[1] See Rossi, 2013, for an overview.

predict the relative performance of the models into the future and ultimately produce more accurate forecasts in a pseudo-real-time exercise.

To establish whether relative forecasting performance is predictable 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 give 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, for example, 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. In this contest it might 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. Thus, we use the Giacomini and White (2006) test to study how the relative performance of the models evolve in response to observable, economically relevant/meaningful variables.[2]

A rejection of the conditional predictive ability test indicates that the relative forecasting performance, i.e. the loss difference, depends on some extra information not included in the models. Then, we interpret rejections as evidence of misspecification of the original models and of non-optimality of the forecasts. This suggests that forecasts can be improved. To this end, we suggests two rules that help to either select or average among the existing models. An alternative strategy would be to respecify the models to incorporate the conditioning information. This approach has some clear drawbacks. First, it requires to select a new specification for the time series models; but the conditional test results do not provide guidance on the form of the misspecification. Second, using a richer specification by adding the conditional variables to the model or by using non-linear functional forms might lead to over-fitting. Last, it is costly, as it requires to re-estimate the models and construct the forecasts again. Because of these considerations we advocate the use of our simple decision rule for model selection and model averaging, which can be applied directly on the forecasts of the original, although possibly misspecified, models. We propose this as a low cost, feasible and scalable alternative for improving the forecasting process and its accuracy.

[2] An alternative would be to postulate the relative forecasting performance of the models as a regime-switching model and use the estimated transition matrix for selecting a particular forecasting model. Fossati (2017) and Odendahl, Rossi and Sekhposyan (2017) propose tests of equal predictive ability in a regime-switching framework. Whether model selection/averaging based on a regime-switching setup can outperform our proposed strategy is an empirical question and can be pursued in future research.

The contribution of our paper to the literature is twofold. First, we conduct an extensive examination of conditional predictive ability for US industrial production and inflation over the past fifty years. 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 in explaining the differentials in forecasting performance of the models. Second, 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 a novel model averaging strategy.

There are a few papers in the literature that use conditional performance for model selection and averaging. Aiolfi and Timmermann (2006) exploit the persistence of the past forecast errors in an optimal way for constructing weights for model averaging. Gibbs and Vasnev (2017) consider optimal weighting strategy conditional on *expected future performance* of the models given their past performance. Their approach results in sizable improvements in the accuracy of inflation forecasts. Our weighting strategy is based on heuristics, which though lacks optimality, still allows for a simple way to take into account the expected future performance of the models (conditional on a vast set of variables) for a wide family of loss functions.[3] Kim and Swanson (2016), on the other hand, use a "hybrid" modeling strategy, where they use a threshold controlling for the severity of the business cycle to switch between naive benchmark and sophisticated index driven models for forecasting. They find that this strategy delivers sizable improvements in the accuracy of the GDP growth forecasts.

To summarize the findings, 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. In most cases the economic models outperform the benchmarks during turbulent times, i.e. during recessions, when financial conditional or uncertainty are high, etc. We also find past relative perfomance to be a good predictor of future performance. Moreover, using conditioning information in a simple decision rule in many cases results in gains that are sizable, especially for multiple-step-ahead forecasts.

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.

[3] The theoretical results in Gibbs and Vasnev (2017) hold only for a mean squared forecast error.

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 $t = R, ..., T - \tau$, where R is the estimation window size and $\tau > 0$ is the forecast horizon. Let $P = T - \tau - R + 1$ be the out-of-sample evaluation window size.[4] We are interested in forecasting a scalar $y_{t+\tau}$, $\tau \geq 1$, using two alternative models.[5] Though multi-model comparison might be the best approach if one is ultimately interested in obtaining the most accurate forecast, bi-model comparison helps us in understanding the predictive content of each economic variable in particular periods of time.

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 models estimated with a fixed rolling window of size R. In what follows we take $f_{t,R}\left(\hat{\beta}_{0,t}\right)$ 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 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 superior to the benchmark, while a negative value indicates that the benchmark dominates the alternative in terms of squared forecast errors. The null hypothesis is expressed as:

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

By estimating the models with a fixed rolling window, we ensure that the parameter estimation error is maintained under the null hypothesis and becomes part of the evaluation. The null is a statement on a forecasting method: 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.

[4] 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.

[5] Examples of unconditional equal predictive ability tests for multiple model comparison are given by Clark and McCracken (2012), Hubrich and West (2010) and Granziera, Hubrich and Moon (2014). To the best of our knowledge, tests of conditional predictve ability for multiple model comparison have not been developed.

The test statistic for the unconditional test is the regular t-statistics:

$$t_{R,P,\tau} = \frac{\Delta \bar{L}_{R,P}}{\hat{\sigma}_P/\sqrt{P}}$$

where $\Delta \bar{L}_{R,P} = P^{-1} \sum_{t=R}^{T-\tau} \Delta L_{R,t+\tau}$, i.e. the numerator is just the sample average of the loss difference, and $\hat{\sigma}_P$ is the Heteroskedacticity and Autocorrelation Consistent (HAC) estimator of $\left(P^{-1/2} \sum_{t=R}^{T-\tau} \Delta L_{R,t+\tau} \right)$. Given the Giacomini and White (2006) asymptotics, critical values from the standard normal distribution apply. A statistically significant, negative (positive) value for $t_{R,P,\tau}$ provides evidence of more accurate forecasts from the benchmark (alternative) model on average.

For a given choice of a $q \times 1$ vector of conditioning variables h_t, testing for the conditional equal predictive ability null is equivalent to testing $E(h_t \Delta L_{R,t+\tau}) = 0$. 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 HAC estimator of the variance of $\left(P^{-1/2} \sum_{t=R}^{T-\tau} h_t \Delta L_{R,t+\tau} \right)$. At α level of significance, the test rejects when $T^h_{R,P,\tau} > \chi^2_{q,1-\alpha}$.[6] Moreover, for our empirical application we use a Newey-West (1987) estimator with a bandwidth of $[0.75 T^{1/3}]$.[7]

It should be noted that both of these tests can be implemented in a regression-based framework, where the loss differentials are regressed either on a constant only or on a constant and (a set of) conditioning variables. In both of these cases we can report the test statistics as well as the marginal impact a particular conditioning variable has on the loss differential (say, denoted by $\hat{\delta}^{\tau}_{P,t}$). Though the marginal impact is important for understanding the average role of the conditioning variables on the loss differential, it might not be the relevant criteria if one is concerned with using the information in the conditioning variables to pick a model for the future date. For that particular question it is useful to think of an approximation to the conditional loss difference proposed by Giacomini and White (2006), $\hat{\delta}^{\tau\prime}_{P,t} h_t \approx E\left[\Delta L_{R,t+\tau} | \mathcal{F}_t\right]$, with the product $\hat{\delta}^{\tau\prime}_{P,t} h_t$ being the fitted value obtained by regressing the loss difference on the conditioning variables. Then, positive (negative) values of the expected conditional difference imply that the alternative (benchmark) should be chosen at time t.

The rationale behind this method can be understood through a simple example. Suppose the conditioning variable is a dummy that takes the value one if the economy is in a recession and zero

[6] 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 $\chi_{2,1-\alpha}$.

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

otherwise. Then, if the alternative model is more accurate than the benchmark during recession episodes both the estimated $\hat{\delta}^{\tau\prime}_{P,t}$ coefficient and the fitted values $\hat{\delta}^{\tau\prime}_{P,t} h_t$ would be strictly positive during recessions. In other words, it is not only the marginal impact of the conditioning variable that is important, but also the particular value of the conditioning variable at a particular point in time. Certainly, the current discussion is in terms of full sample analysis, however, as we show further, it is possible to use the conditioning variables in a pseudo-real-time exercise.

A few comments on the properties of the conditional and unconditional tests are needed. It is possible that the unconditional predictive ability tests fail to reject the equal predictive performance of the models, yet the conditional predictive ability tests do. 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. 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 small-sample 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.

2.2 Picking the Next Forecasting Model

On the one hand, rejecting the null of conditionally equal predictive ability might be interpreted as bad news because it means that models are misspecified and the forecasts made with these models are not optimal. On the other hand, if the relative forecasting ability of the models can be predicted, 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. As discussed in the introduction, we propose model selection and model averaging as a simple, flexible and scalable way of dealing with model mis-specification relative to postulating new types of models.

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 initial window size for the implementation of the rule. Further, at any given point in time, we use a fixed rolling sample of size S for conditional testing. Thus, we follow the two step rule:

1. Regress the loss differences, $\{\Delta \hat{L}_{S,j+\tau}\}_{j=t-S}^{t}$, on a single conditioning variable (plus a con-

stant) $\{h_j\}_{j=t-S}^{t}$ over S observations of the out of sample, $t = R+S, ..., T-\tau$ and denote the regression coefficient as $\hat{\delta}_{S,t}^{\tau}$;

2. Predict $y_{t+\tau}$ using the forecast of the benchmark model if $\hat{\delta}_{S,t}^{\tau\prime} h_t < 0$, $t = R+S, ..., T-\tau$, and use the alternative model otherwise[8].

We further consider a modified version of this rule in that we use the information on the statistical significance of $\hat{\delta}_{S,t}^{\tau}$. In other words, in this version of model selection step (1) would be followed by the following steps (2') and (3').

2'. Check the statistical significance of $\hat{\delta}_{S,t}^{\tau}$ using a two-sided test.

3'. For $t = R+S, ..., T-\tau$, predict $y_{t+\tau}$ using the forecast of the benchmark model if $\hat{\delta}_{S,t}^{\tau\prime} h_t < 0$ and $\hat{\delta}_{S,t}^{\tau}$ is statistically different from zero in the previous step. Use the alternative model otherwise.

The above strategies select only one model, either the benchmark or the alternative, at a given point in time. The default strategy is to use the benchmark, unless, time t state of the conditioning variable, h_t, is expected to improve the forecasting performance of the alternative model over the benchmark for a future date τ.

2.2.2 Model Averaging

Alternatively, we propose a rule for model averaging where instead of selecting only one model (either the benchmark or the alternative) at each forecast origin date, we take a weighted average of the benchmark and the alternative. The rule is implemented as follows:

1. Regress the loss differences, $\{\Delta \hat{L}_{S,j+\tau}\}_{j=t-S}^{t}$, on a single conditioning variable (plus a constant) $\{h_j\}_{j=t-S}^{t}$ over the S observations of the out of sample, $t = R+S, ..., T-\tau$ and denote the regression coefficient as $\hat{\delta}_{S,t}^{\tau}$;

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

$$w_{0,t} = \frac{1}{S} \sum_{j=t-R}^{t} 1\left\{\hat{\delta}_{S,j}^{\tau\prime} h_j < 0\right\}, \quad t = R+S, ..., T-\tau$$

and the weight of the alternative model is $w_{1,t} = 1 - w_{0,t}$. $1\{.\}$ denotes the indicator function.

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

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.

By construction, model selection and model averaging strategies that we propose are pseudo-real-time exercises in that at a particular point in time, from $R+S$ till $T-\tau$, the practitioner considers a bi-model comparison or averaging exercise, where (s)he uses the results of the conditioning test to based on the most recent sample of S observations to (i) select between a benchmark and an alternative model or (ii) to construct weights for the linear combination of τ-period-ahead point forecasts.[9]

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, and we are concerned with annualized average growth rate τ periods ahead. $X_{t,k}$ 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 $K = 117$.[10] 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 $Y_t = 1200\ln(CPI_t/CPI_{t-1}) - 1200\ln(CPI_{t-1}/CPI_{t-2})]$.[11] 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.

[9] The exercise is pseudo-real-time, as we do not consider the real time nature of neither the data that goes into the forecasting models, nor the conditioning variables. Certainly, for the variables that are not subject to revisions the exercise would be real time by construction. However, these variable are a small proportion of the total. It is infeasible to extend this analysis to a real-time as for most conditioning variables real-time data vintages are not available or start much later in the considered sample period.

[10] 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).

[11] Note that, s Rossi and Sekhposyan (2010), this relies on the assumption that inflation is I(2).

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 Description of the Data

We first discuss the data used to construct the autoregressive benchmark and the alternative autoregressive distributed lag models presented in Section 3. We then go into more details on the conditioning variables that are used for the conditional predictive ability tests as well as for the pseudo-out-of-sample real time exercise of model selection and model averaging based on the conditional predictive ability test results.

4.1 Data Used to Forecast

The data used for forecasting comes from the monthly macroeconomic database of McCracken and Ng (2016).[12] 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 purposes of this paper we use all their series with the exception of those that start later than 1959:M1. These include the series on new private housing permits and its various geographic counterparts, i.e. the permits covering northeast, midwest, south and west. In addition, we exclude the series on new orders for consumer as well as durable goods. We also exclude the trade weighted U.S. dollar index against major currencies, consumer sentiment index and VXO.[13] We have a total of 117 series. The sample period ends in 2016M1, yielding a total of 685 monthly observations. The data has been transformed to stationarity using the proposed transformations in McCracken and Ng (2016). 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 for our analysis.

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

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

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.[14] Given the sample starting period, the number of observations lost due to data transformations as well as the 10-year-rolling window used for the estimation, the out-of-sample evaluation period across models starts in 1970:M3.

4.2 Data Used for Model Selection and Averaging

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 frequency (monthly) and availability of the data going back to 1970:M2 in order make the pseudo-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.

INSERT TABLE 1 HERE

Business cycle indicators: Chauvet and Potter (2013) and Stock and Watson (2007), among others, find that relative forecasting performance differs across phases of the business cycle for output growth and inflation, respectively. Therefore, we consider as a 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 look at alternative measures of the business cycle, available in a more timely manner. More specifically, we construct two binary variables: (i) "ip-rec" which takes the value of one when the industrial production index experiences a negative cumulative growth over the past six months prior to the forecast origin date; (ii) "unemp-rec" when the unemployment rate in the economy is above 6% in the period preceding the forecast origin date. Moreover, we include in our analysis the U.S. recession probability index of Chauvet and Piger (2008), which consists of the smoothed probabilities from a dynamic-factor Markov-switching model applied to four indicators of real economic activity.

Financial conditions/stress indicators: Motivated by Ng and Wright (2013) we consider whether the relative forecasting performance of the models depend on 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

[14] In McCraken and Ng (2016) 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.

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 captures credit conditions, the last one proxies debt and equity measures. For robustness we also use alternative measures of financial stress produced by the Federal Reserve Banks of St. Louis and Kansas City, "SFSI" and "KFSI", respectively.

All these indexes are constructed using principal component analysis over a number of financial variables, including interest rates, spreads and stock prices indicators. The Chicago Fed indexes are available since 1973M1, while the indexes from the St. Louis and Kansas City Feds start much later, i.e. in the beginning of 1990-s. For some of these series, such as the NFCI, there is a real time database, but it starts much later: the first vintage dates to 2011M5. Thus, relying on real time vintages for the evaluation will dramatically cut our out-of-sample observations and make the current study infeasible. Further, many of this indices, namely the ones coming from the Federal Reserve Banks of Chicago and St. Louis are available at a weekly frequency. We use its monthly aggregate acknowledging the fact that one could potentially extend the analysis in this paper on model selection and averaging to higher frequencies. However, at this point we leave those consideration to future research.

Uncertainty Indices: Since the onset of the Great Recession aggregate macroeconomic uncertainty has been identified as one of the major drivers of business cycle fluctuations both in microfunded structural and in VAR models (Bloom, 2009, Ludvigson, Ma and Ng, 2015, Jurado, Ludvigson and Ng, 2015). To capture the different definitions of uncertainty suggested by the literature, we use several indicators. First, we consider the realized volatility of stock returns based on the S&P500 index. We then consider VXO, an implied volatility index based on the S&P100 options.[15] Further, we use the macroeconomic and financial uncertainty indexes from Jurado, Ludvigson and Ng (2015) and Ludvigson, Ma, and Ng (2015). These measures are associated with the variance of the unpredictable components in economic variables, and we use the measures associated with one-, three- and twelve-month-ahead horizons. Baker, Bloom and Davis (2016) provide with measures of Economic Policy Uncertainty (EPU) based on newspaper articles. In addition to the EPU, we also consider a direct measure of monetary policy uncertainty provided by Husted, Rogers and Sun (2016). This measure is a refined version of the one provided by Baker, Bloom and Davis (2016).[16]

[15] An alternative would be to consider the VIX, an implied volatility index based on the S&P500 options. However, that data for the VIX start in 1993, thus we opt for the VXO.

[16] We could also use Rossi and Sekhposyan (2015) macroeconomic uncertainty index, yet it comes at a quarterly frequency instead of monthly.

Past Relative Performance: Finally, the last conditioning category includes a measure of past relative performance. As our motivating example in Figure 1 suggests, the relative forecasting performance of the models might exhibit some persistence. We use the lagged mean squared forecast error difference between the benchmark and the alternative models to predict the relative forecasting performance of the models in the future.

INSERT FIGURE 2 HERE

We should note that the various conditioning variables could be correlated with each other. For instance, the macroeconomic uncertainty index of Jurado, Ludvigson and Ng (2015) identifies episodes where macroeconomic variables are unpredictable. These types of episodes happen to be clustered around the recession dates. Figure 2 shows the correlation among the conditioning variables more formally.[17] The real time recessionary dummy based on the past growth rate of the industrial production and the one based on the historic unemployment rate show a very low correlation (even negative for the unemployment dummy) with the other conditioning variables. The remaining 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, though co-moving, is not perfectly overlapping.

5 Results

We first present the full out-of-sample results on unconditional and conditional equal predictive ability in section 5.1. In this section the tests are applied over our full out-of-sample period, i.e. from 1970:M3 to 2016:M1. Then, in section 5.2, we show the results from our pseudo-out-of-sample exercises essentially addressing the issue of whether the conditional test results, obtained based on a fixed rolling sample, are exploitable, i.e. whether they can be used to improve the accuracy of the forecasts in pseudo-real-time.

5.1 Predictive Ability Tests

Figure 3 shows the results for the unconditional predictive ability test for one-month-ahead ($h = 1$) and twelve-month-ahead ($h = 12$) forecast horizons. The horizontal axis displays the root mean squared forecast errors of the alternative ADL models relative to the benchmark autoregressive model. Ratios greater than one, i.e. to the right of the vertical (red) line, indicate that the economic models performance is worse than that of the autoregressive benchmark. The vertical axis indicates the $p-$values from the Giacomini and White (2016) unconditional predictive ability

[17] The figure shows all the cross correlations except that of the past performance. Since past performance is model specific and we have more than hundred models, displaying cross correlations in a legible fashion would be infeasible.

test with the 10 per cent significance level marked by a horizontal (red) line. Each dot represents one of our 117 bi-model comparisons. 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, as most dots below the (red) horizontal line are located in the right quadrant.

INSERT FIGURE 3 AND TABLE 2 HERE

Figure 3 provides a snapshot of how the models behave. Table 2, on the other hand, reports the models which are on average statistically better than the benchmark. The figure has two Panels. Panel A directly compares to Figure 3 and shows the results for the full evaluation sample. As Panel A shows, there is a lot less predictability 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 for inflation at one-year-ahead ($h = 12$) forecast horizon. Interestingly, an inspection of the loss difference for this economic model shows that it is positive in the early portion of the out-of-sample, while it goes to about zero during the Great Moderation, and it alternates between positive and negative values for the last part of the sample, starting with the Great Recession. Moreover, there is more predictability 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 predictability of industrial production growth at one-year-ahead horizon comes primarily from asset prices.

Panel B of Table 2 is provided for robustness: the panel reports the same unconditional equal predictive ability results, but for a different subsample. The evaluation period for this analysis is from 1979M2-2016M1. We provide this result for direct comparability with our pseudo-real-time exercise. When we construct the pseudo-real-time forecasts based on model selection and model averaging, we use the first 10 years of out-of-sample data as an initial window (S) for the conditional predictive ability test. A natural question that one could ask in that context is that, if we find improvements to our model selection and model averaging strategy, would that be dictated by the evaluation sample or not? As Panel B shows, even in the shorter evaluation sample there is not much evidence of predictive ability. Oil prices seem to help forecast one-month-ahead inflation. There are a few more variables, mostly asset prices, that help with one-month-ahead output growth predictability. However, there are a few less variables that help with twelve-month-ahead predictability of the industrial production relative to the benchmark sample.

The test statistic of the unconditional test is proportional to the sample average of the loss difference. This implies that even if there are large differentials, but they switch between positive and negative values, i.e. the relative performance of the model changes over time, the loss differentials could cancel out over the sample, leading to the inability to reject the null. However, positive (or negative) values of the loss differentials might be clustered around periods of economic significance, summarized by observable time series. To investigate whether this is the case, we apply the Giacomini and White (2006) conditional predictive ability test to the models under consideration using the conditioning variables discussed previously.

INSERT FIGURE 4 HERE

Results for the conditional predictive ability tests are provided in Figure 4. On the horizontal axis we show 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 the conditioning variable. Recall that $\hat{\delta}^{\tau\prime}_{P,t} h_t \approx E\left[\Delta L_{P,t+\tau} | \mathcal{F}_t\right]$; in practice we compute the statistic $I_{GW} = \frac{1}{P}\sum_{t=R}^{T-\tau} 1\left\{\hat{\delta}^{\tau\prime}_{P,t} h_t < 0\right\}$. We calculated the loss differential as that of the squared loss of the benchmark and economic models: a negative value of $\hat{\delta}^{\tau\prime}_{P,t} h_t$ indicates a better performance for the benchmark model. We mark the significance of the marginal effects of the coefficients based on the Giacomini and White (2006) conditional preditive ability test at 10% significance level. Then, consistent with the figure showing the unconditional test results, models that perform significantly better than the benchmark will be located on the lower left quadrant. For both target variables (i.e. industrial production and inflation) and for both forecasting horizons we only report results for the conditioning variables that give us the highest number of rejections of equal predictive ability across the wide set of ADL models considered[18]. For both inflation and output growth, at one-year-ahead forecast horizon, lagged performance of the models seems to have the most predictive power. In other words, this is the conditioning variable that gives the highest number of improvements of the alternative model over the benchmark in the full out-of-sample. The evidence is stronger for inflation than for output growth. For the one-month-ahead forecast horizon, however, it appears that the financial indices matter more. For the growth rate of industrial production, the Adjusted (of macroeconomic factors) National Financial Conditions Index (ANFCI) appears to be more important, while for inflation the improtant conditioning variable comes out to be the Kansas City Fed Financial Stress Index (KCFCI). The detailed description of the models for which, conditional on the chosen conditioning variable (i.e. ANFCI, KCFCI and lagged performance), we obtain statistically significant results based on the conditional predicitive ability test are provided in Table 3.

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

In general, the unreported results for the full set of conditioning variables shows that the conditional test rejects more frequently than the unconditional test at both one-step-ahead and one-year-ahead forecast 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 output growth.

Table 4 lists the models which perform significantly better than the benchmark in Figure 4, i.e. the models in the low left quadrant of the figure. The relative performance column shows the statistic:

$$M_{GW} = \frac{\sum_{t=R}^{T-\tau} \mid \hat{\delta}_{P,t}^{\tau\prime} h_t \mid 1 \left\{ \hat{\delta}_{P,t}^{\tau\prime} h_t < 0 \right\}}{\sum_{t=R}^{T-\tau} \mid \hat{\delta}_{P,t}^{\tau\prime} h_t \mid},$$

which is bounded between zero and one. This gives us an idea of the magnitude of the improvement of the benchmark model over the alternative, given the average improvement induced by the conditioning variable. Since $\hat{\delta}_{P,t}^{\tau\prime} h_t \approx E\left[\Delta L_{P,t+\tau} | \mathcal{F}_t\right] = E\left[\varepsilon_{0,t+\tau}^2 - \varepsilon_{1,t+\tau}^2 | \mathcal{F}_t\right]$, a value close to zero indicates that the alternative model has a much better performance than the benchmark. The lower this number, the better the performance of the alternative. Then, this paper further contributes to the literature by suggesting this new statistic to summarize the conditional, relative performance of the models.

INSERT TABLE 3 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 it is picked up only from the conditional test. For inflation, oil prices are important at one-step-ahead forecast horizon. At twelve-steps-ahead, on top of money measures, real activity measures and, in particular, measures of employment prove to be useful. In addition, the conditional predictive ability test finds evidence of an empirical relationship between inflation and money measures and between inflation and unemployment.

The indicator variable suggested by Giacomini and White (2006) provides with a summary statistic of the conditional relative performance of the models. However, it does not help to identify which model is useful under which circumstances. We suggest, as a complementary analysis to understand the reason for the rejection, to compare the conditioning variable to the time series $1\left\{\hat{\delta}_{P,t}^{\tau\prime} h_t < 0\right\}$ over the out-of-sample. This essentially would recover the value of $\hat{\delta}_{P,t}^{\tau\prime}$. It is instructive to focus on some specific interesting examples, namely on the predictability over the business cycle and based on financial conditions.

5.1.1 Predictability and Business Cycle Phases

First, we analyze conditional predictability in recessions versus expansions. There is a vast literature documenting that the behaviour of macroeconomic variables differs in these two phases of the business cycle. As long as this translates into changes in the interdependencies among economic variables, it could also affect the relative forecasting performance of time series models. The Giacomini and White (2006) test that uses the NBER recession dates as conditioning variables do reject the null in a number of bi-model comparisons. Figure 5 plots our target variables, industrial production and inflation, as well as the NBER recession dates (shaded bars) and lists the models for which the conditional test rejects the null. We consider only models for which the unconditional test is unable to reject or it rejects but points to a superior performance of the benchmark. We find that most of the economic models are useful during recessions. For example, housing starts, total non-revolving credit, new orders of durable goods and labor market condition indicators such as the help-wanted index and civilian employment, help predicting industrial production at one-step ahead during recessions. This provides statistically supported evidence to the findings in Chauvet and Potter (2013), which argue that during expansions simple univariate autoregressive models for GDP are as good as more complex models, while during downturns there can be large gains in forecast accuracy using additional variables or larger models. Interestingly, for inflation, Phillips-curve type models which include measures of employment such as the civilian unemployment rate or retail trade employees, are useful in prediction during recessions. This confirms the results in Dotsey, Fujita and Stark (2011).

INSERT FIGURE 5 HERE

5.1.2 Predictability and Financial Conditions

A number of recent studies document non-linear dynamics between macro variables depending on the financial conditions of the economy. Galvao and Owyang (2017) find that macro variables dynamics change during period of high financial stress. Adrian, Boyarchenko and Giannone (2016) studies the evolution of the distribution of output over time and documents that only the left tail varies with financial conditions. Del Negro, Hasegawa and Schorfheide (2016) find that models with financial frictions produce superior forecasts in periods of financial distress relative to models without financial frictions. Our results confirm these non-linearities, as illustrated in Figure 6 by two examples for industrial production at one-step ahead when the conditioning variable is the ANFCI index. The solid line represents the index, while the shaded areas are the periods of times in which the indicator $\hat{\delta}_{P,t}^{\tau\prime} h_t$ is positive, i.e. the test selects the alternative model. The ADL model that includes new orders for durable goods (AMDMNOx), as most of the other ADL models for which we obtain a rejection, is more useful when financial conditions are tight. Interestingly, interest rates and spreads, such as the 1-year Treasury spread, are useful when financial conditions are loose. Note that for interest rates and spreads we were obtaining that they were significantly more accurate than the benchmark even unconditionally, while this was not the case for the AMDMNOx model.

INSERT FIGURE 6 HERE

As a general finding, we report that ADL models are more accurate than simple benchmarks during turbulent times. This helps us understand why, in general, rejections from the unconditional test point to better performance of the benchmark model: if the alternative model is more accurate, i.e. the loss differential is positive, only during turbulent times and these times are less frequent and shorter lived than tranquil times, then, on average, the loss differential will be negative. Failing to reject the unconditional test might lead to dismiss the alternative model. However, our results from the conditional tests show that we can redeem many economic models: while simple benchmarks might be enough when we navigate tranquil waters, economic models are most valuable when economic conditions are deteriorating.

5.2 Decision Rule

We interpret rejections of the null of conditional equal predictive ability as indication of misspecification of the models. In other words, the conditioning variable represents information available at the time 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 selection and/or averaging rules proposed in this paper. The first strategy requires a formulation of a new forecasting model as well as an ability to estimate it and produce forecasts.[19] The second strategy, on the other hand, is based on the forecasts of the benchmark and alternative models, which are already available.

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

[19] Moreover, some of the misspecifications suggest regime dependence, which might require an estimation of a non-linear model. This could be computationally cumbersome.

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 (autoregressive) model. It should be noted that the conditioning test is conducted in real time, i.e. in each of the 465 forecast origin dates we have an updated result on the conditional predictive performance test.

INSERT FIGURE 7 HERE

Figure 7 shows, for each conditioning variable, the relative RMSE of the model selection rule versus the benchmark. The figure plots only the models for which the model selection rule provides a lower RMSE than the benchmark. Gains are larger at twelve-steps-ahead than at one-step-ahead, and more so for industrial production than for inflation. Reductions in RMSFE can reach 12%, which is a large number compared to the literature. Though lagged performance seems to be the most robust conditioning variable across the various targets and horizons, uncertainty and financial stress indices help to improve the accuracy of the one-year-ahead industrial production growth forecasts, while uncertainty indices are useful for improving the twelve-month-ahead inflation forecasts. These gains are sizable and similar to the improvements recorded in the literature. For instance, McCracken and Ng (2016) find similar relative RMSFE for US industrial production using as alternative model with a single factor. Note that their exercise is also a pseudo-real-time one since the factor is obtained using the whole sample data. Diebold and Sin (2017) also obtain improvements of a similar magnitude by applying their two step model selection and subsequent model averaging exercise to the European Survey of Professional Forecasters.

Figure 8, on the other hand, shows, for selected conditioning variables, the result of the averaging strategy relative to the benchmark in a pseudo-real-time exercise under consideration. The results are similar to that of model selection. In fact, it appears that the model averaging exercise delivers marginally better results than the model selection one measured by the number of models that improve for each of the conditioning variables. However, the improvements are similar in magnitude. We should note that the results based on significance testing, i.e. model selection and averaging strategies that rely on the alternative only if the alternative is statistically better than the benchmark, deliver comparable results. There is not much of a gain relative to the results presented in Figures 7 and 8. Thus, we omitt those results to save space.

INSERT FIGURE 8 HERE

6 Conclusions

In this paper we conduct a systematic evaluation of the conditional predictive ability of various economic variables that represent asset prices, measures of real economic activity, wages and prices, as well as money. We consider a wide range of autoregressive distributed lag models for forecasting and compare its forecasting performance to an autoregressive benchmark. 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 predicted based on past out-of sample relative accuracy. We find that all these variable are, to some extent, useful for predicting a better performing model in the future. Selecting models based on their past performance improves the predictive ability for both inflation and output growth. Both financial indices, as well as uncertainty indices seem to be useful for improving both output growth and inflation forecasts at a one-month-ahead forecast horizon. Our results suggest that using the conditional equal predictive ability tests in an informative way could indeed be useful for model selection and model averaging strategies. In particular, we document that using the conditioning information as a criteria for model selection and averaging in fact can result in up to ten percent improvements in the root mean squared forecast error relative to a competitive autoregressive benchmark.

References

[1] Adrian, T., Boyarchenko, N., and D. Giannone (2016). "Vulnerable Growth," New York Fed *Staff Report* 794.

[2] Aiolfi, M. and A. Timmermann (2006). "Persistence in Forecasting Performance and Conditional Combination Strategies," *Journal of Econometrics* 135 (1-2), 31-53.

[3] Baker, S.R., N. Bloom and S.J. Davis (2016), "Measuring Economic Policy Uncertainty," *Quarterly Journal of Economics*, 131(4), 1593-1636.

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

[5] 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.

[6] 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.

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

[8] Clark, T.E. and M.W. McCracken (2012), "Reality Checks and Comparisons of Nested Predictive Models," *Journal of Business and Economic Statistics* 30, 53-66.

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

[10] Del Negro, M., Hasegawa, R.B. and F. Schorfheide (2016), "Dynamic Prediction Pools: An Investigation of Financial Frictions and Forecasting Performance," *Journal of Econometrics* 192(2), 391-405.

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

[12] Diebold, F.X. and M. Shin (2017), "Beating the Simple Average: Egalitarian LASSO for Combining Economic Forecasts," *mimeo.*

[13] Dotsey, M., Fujita, S. and T. Stark (2011), "Do Phillips Curves Conditionally Help to Forecast Inflation," Philadelphia FED *Working Paper* 17-26.

[14] Fossati, S. (2017), "Testing for State-Dependent Predictive Ability," University of Alberta *Working Paper* 2017-09.

[15] Galvao, A.B. and M.T. Owyang (2017). "Financial Stress Regimes and the Macroeconomy," St. Louis Fed *Working Paper* 2014-020C.

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

[17] Gibbs, C. and A.L. Vasnev (2017), "Conditionally Optimal Weights and Forward-Looking Approaches to Combining Forecasts," *mimeo.*

[18] Granziera E., Hubrich, K. and H.R. Moon (2014), "A Predictability Test for a Small Number of Nested Models," *Journal of Econometrics* 182, 174-185.

[19] Hubrich, K. and K.D. West (2010), "Forecast Evaluation of Small Nested Model Sets," *Journal of Applied Econometrics* 25, 574-594.

[20] Husted, L., Rogers, J. and B. Sun (2016) "Measuring Monetary Policy Uncertainty: The Federal Reserve, January 1985-January 2016," *IFDP Notes*, Board of Governors.

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

[22] Kim, K. and N.R. Swanson (2016), "Mixing Mixed Frequency and Diffusion Indices in Good Times and In Bad," *mimeo.*

[23] Ludvigson, S., Ma, S. and S. Ng (2015), "Uncertainty and Business Cycles: Exogenous Impulse or Endogenous Response?" *mimeo.*

[24] 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.

[25] 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.

[26] Odendahl, F., Rossi, B. and T. Sekhposyan (2017), "Comparing Model Forecast Performances under Markov Switching," *mimeo.*

[27] 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.

[28] 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.

[29] Rossi, B. and T. Sekhposyan (2015), "Macroeconomic Uncertainty Indices Based on Nowcast and Forecast Error Distributions," *American Economic Review* 105(5), 650-655.

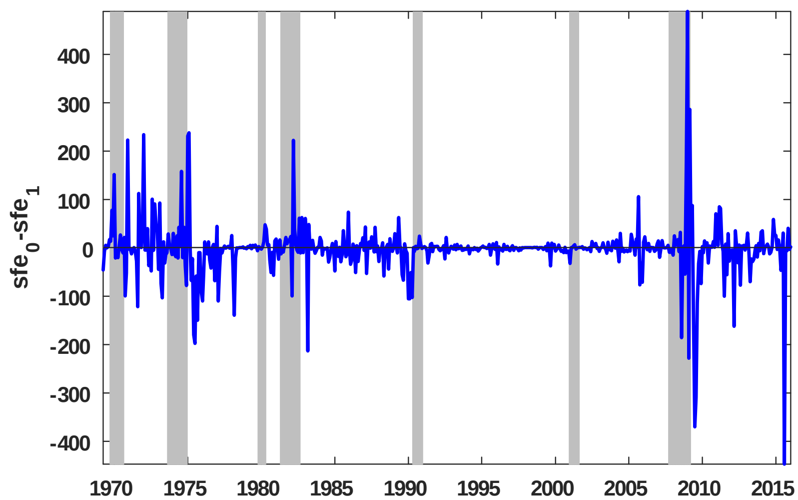
[30] Stock, J.H. and M.W. Watson (2010), *Introduction to Econometrics*, 3rd ed., Addison-Wesley.

[31] 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.

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

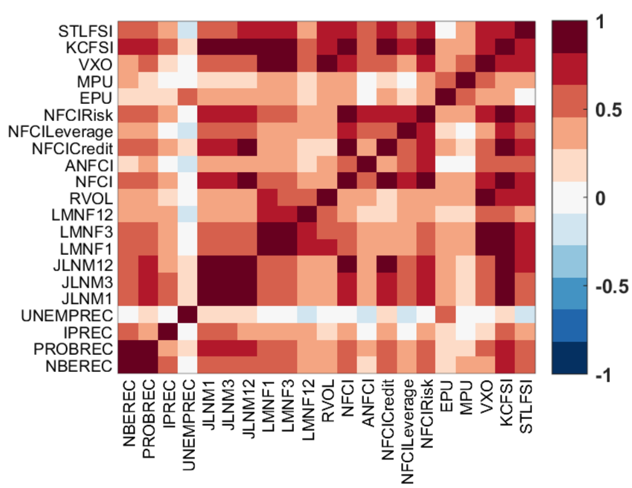
7 Figures and Tables

Figure 1. Illustrative Example



Note: The figure shows the squared forecast error differential between the benchmark AR(2) model (sfe_0) for US IP and an alternative (sfe_1) ADL model with housing starts. Shaded areas represent NBER recession dates.

Figure 2. Cross-correlation of Conditioning Variables

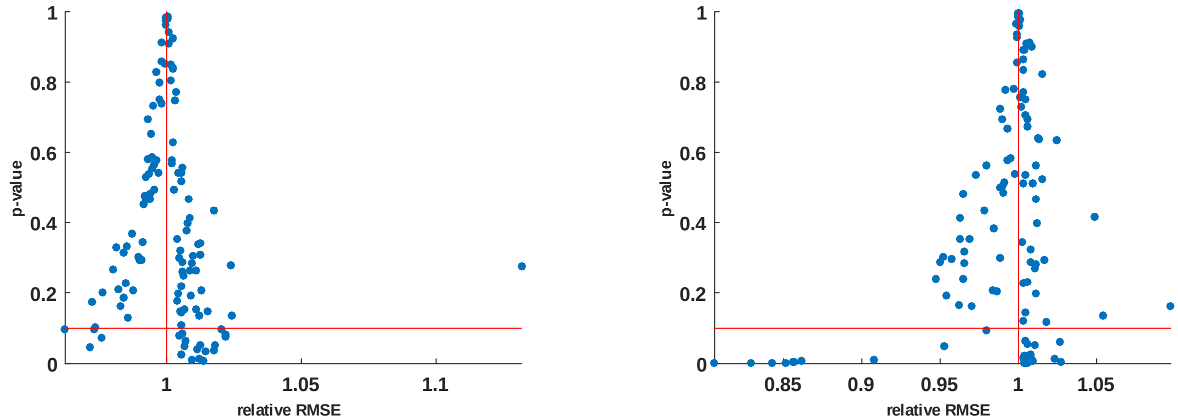


Note: Cross-correlation of conditioning variables. The labels are consistent with those in Table 1.

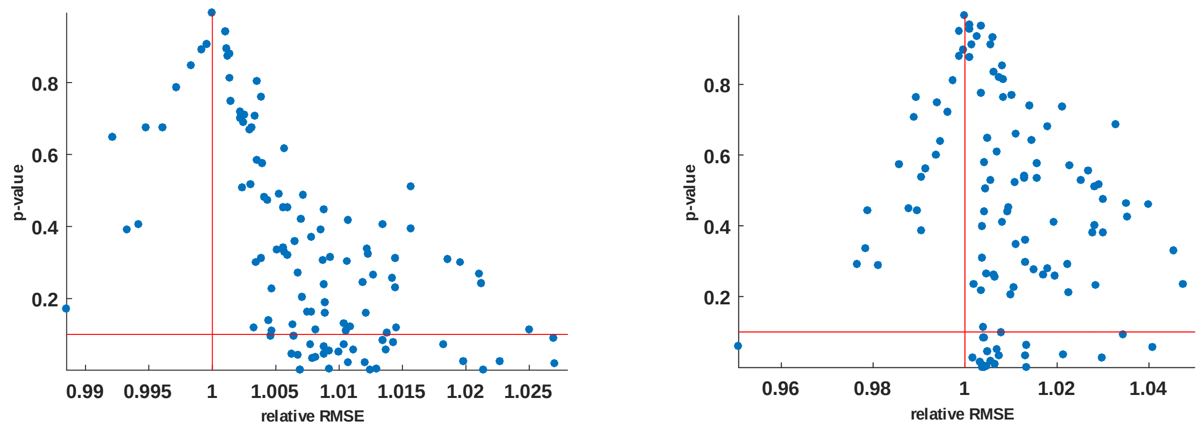
Figure 3. Unconditional Tests of Equal Predictive Ability

$h = 1$ | $h = 12$

Panel A. Industrial Production Growth



Panel B. Inflation

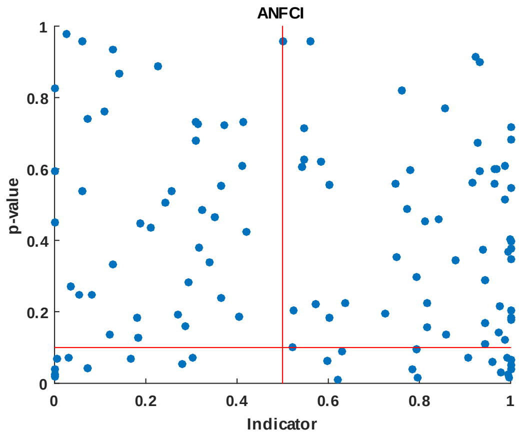


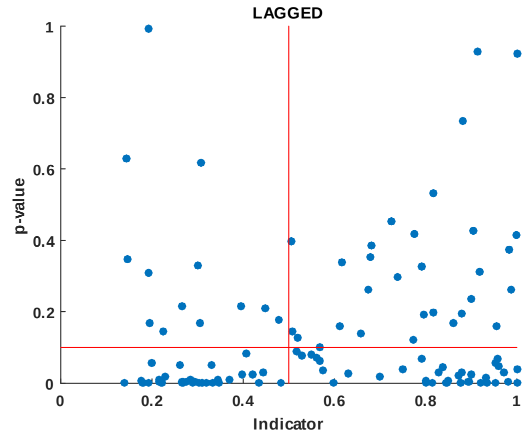
Notes: The figure shows the unconditional test results for the models of inflation and industrial production growth. The benchmark is an AR, while the alternatives are ADLs, where we consider each economic variable one at a time. Relative RMSE values greater than one favor the benchmark model.

Figure 4. Conditional Tests of Equal Predictive Ability

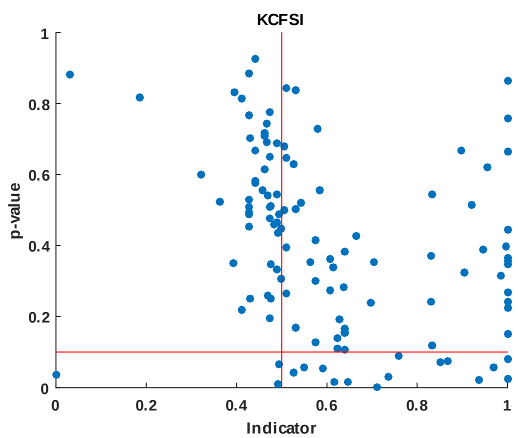
$h = 1$ | $h = 12$

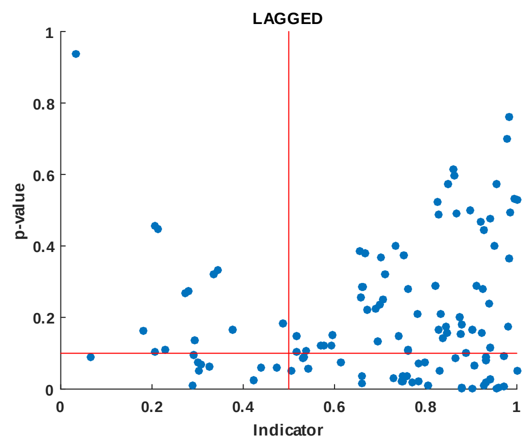
Panel A. Industrial Production Growth





Panel B. Inflation

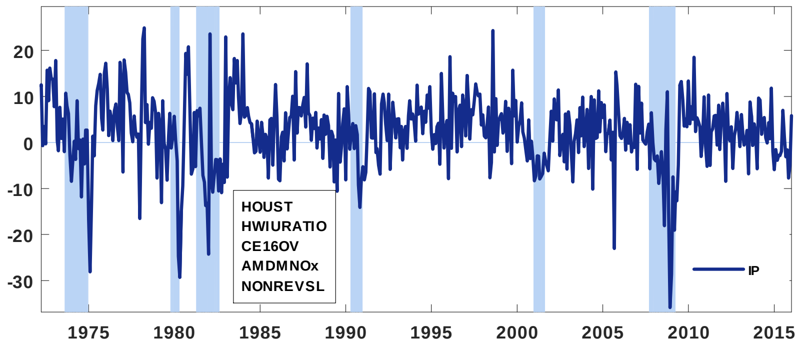




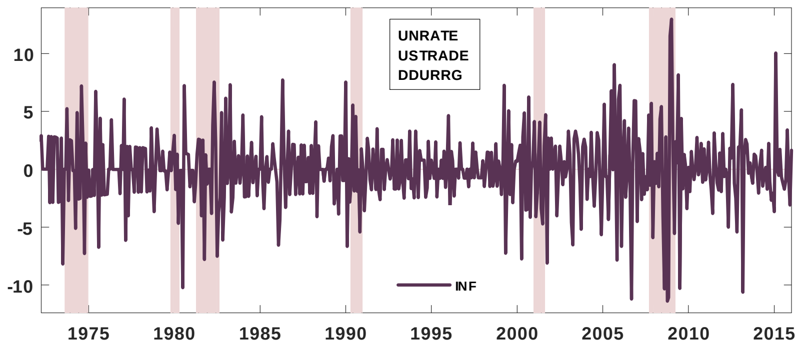
Notes: The figure shows the conditional predictive ability test results for the models of inflation and output growth. We report the results for the conditioning variables resulting in the largest number of rejections across the various models. Horizontal axis captures the proportion of the time the benchmark model is better than the alternative, while the vertical axis shows the p-values associated with the Giacomini and White (2006) conditional predictive ability test.

Figure 5. Predictive Ability Over the Business Cycle

Panel A. Industrial Production Growth

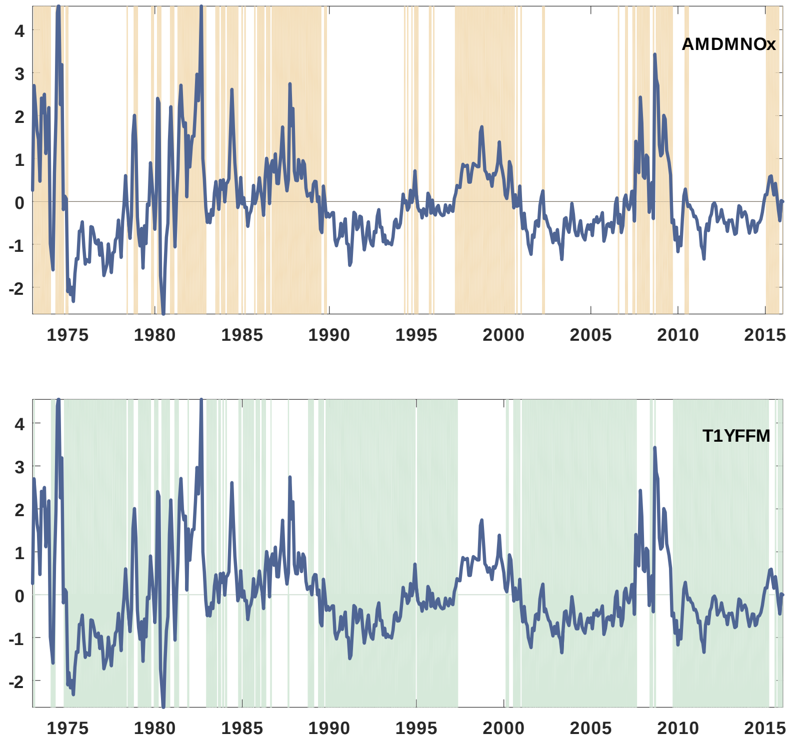


Panel B. Inflation



Notes: The figure displays the conditioning variable, in this case NBER recession dates, as well as the total effect of the recessions on the relative forecasting performance of the models. It also lists the models selected during recessions. The results are for one-step-ahead prediction.

Figure 6. Predictive Ability in Output Growth and Financial Conditions

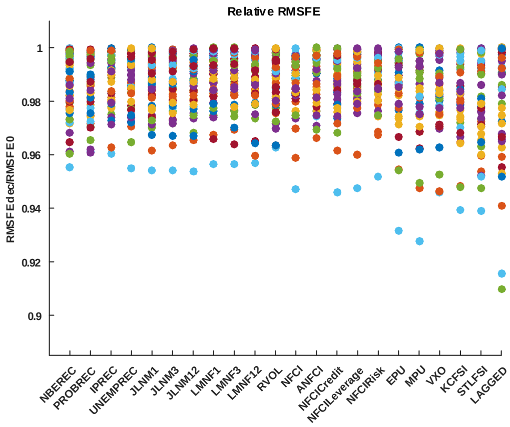


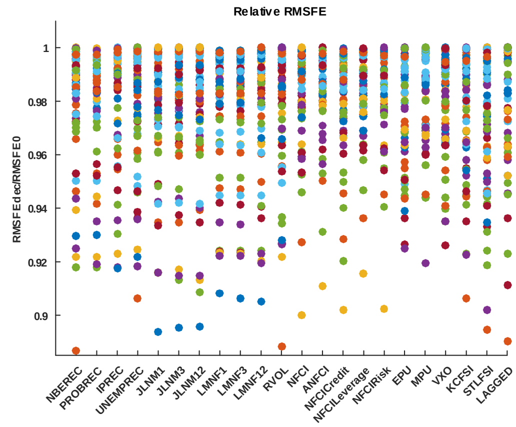
Notes: Conditioning variable is the ANFCI index (blue solid line in both panels). Positive (negative) values of the ANFCI indicate financial conditions that are tighter (looser) than average. Upper panel uses the model with AMDMNOx: new orders for durable goods, while the lower panel uses the model with T1YFFM: 1-year Treasury minus FFR. Target variable is industrial production at one-step-ahead horizon. Shaded areas are the periods of times in which the indicator $\delta^{\tau}_{P,t} h_t$ favors the alternative model.

Figure 7. How to Pick the Next Forecasting Model: Model Selection Approach

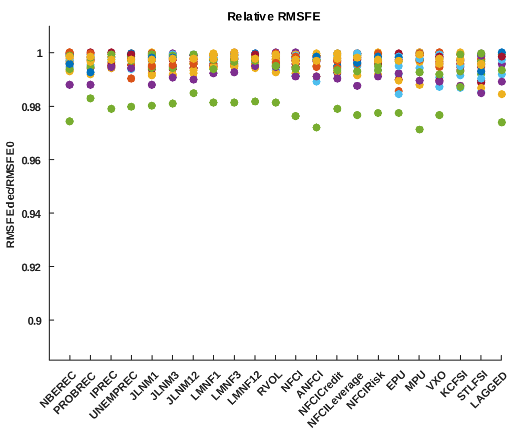
$h = 1$ $h = 12$

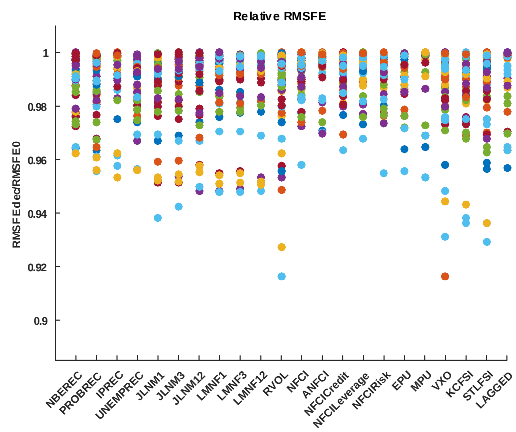
Panel A. Industrial Production Growth





Panel B. Inflation



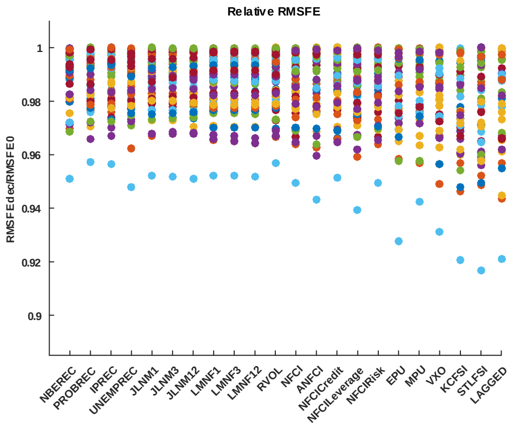


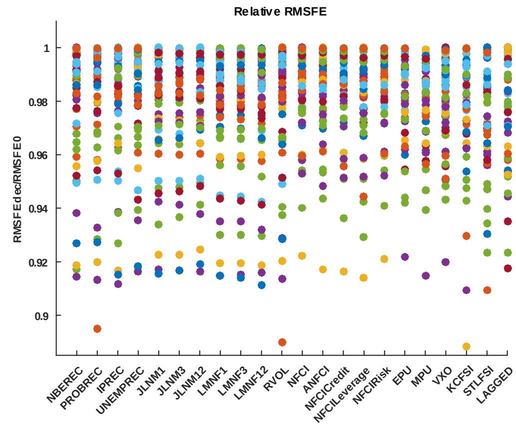
Notes: The figure shows the root mean squared forecast error of the model selection rule relative to the benchmark (benchmark is marked with 'RMSFE0'). Values greater than one are not depicted on the figure since they would indicate that our model selection criteria deteriorates the accuracy of the forecasts relative to the benchmark. Results are group by conditioning variables, labeled consistently with the labels in Table 1.

Figure 8. How to Predict the Next Forecasting Model: Model Averaging Approach

$h = 1$	$h = 12$

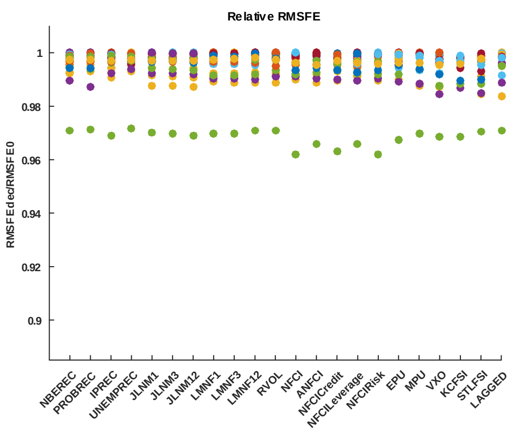
Panel A. Industrial Production Growth

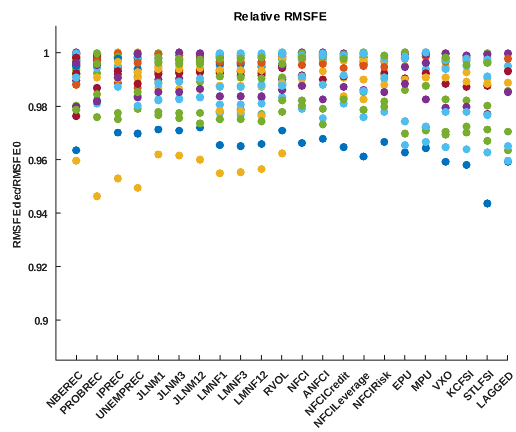




2

Panel B. Inflation





Notes: The figure shows the root mean squared forecast error of the model averaging rule relative to the benchmark (benchmark is marked with 'RMSFE0'). Values greater than one are not depicted on the figure since they would indicate that our model averaging criteria deteriorates the accuracy of the forecasts relative to the benchmark. Results are group by conditioning variables, labeled consistently with the labels in Table 1.

Table 1. Description of Conditioning Variables

Label	Starting Date	Description	Source
		Business Cycle Indicators	
NBERREC	1959M1	NBER recession dates: from Peak to Trough	F
PROBREC	1967M6	Smoothed US recession probabilities, percent	F
IPREC	1959M1	Six months of negative industrial production growth	F*
UNEMPREC	1959M1	Unemployment rate above 6%	F*
		Financial Conditions/Stress Indicators	
NFCI	1973M1	National Financial Conditions Index	F
ANFCI	1973M1	Adjusted National Financial Conditions Index	F
NFCICredit	1973M1	National Financial Conditions: Credit Subindex	F
NFCILeverage	1973M1	National Financial Conditions: Leverage Subindex	F
NFCIRisk	1973M1	National Financial Conditions: Risk Subindex	F
STLFSI	1994M1	St. Louis Fed Financial Stress Index	F
KCFSI	1990M2	Kansas City Fed Financial Stress Index	F
		Uncertainty Indices	
VXO	1986M1	CBOE Implied Volatility Index based on S&P100 options	C
RVOL	1967M7	Realized Volatility	F*
MPU	1985M1	Monetary Policy Uncertainty Index	HRS
EPU	1985M1	Economic Policy Uncertainty Index	BBD
JLNM1, JLNM3, JLNM12	1960M7	Macroeconomic Uncertainty Index, 1-3 and 12 periods ahead	JLN
LMNF1, LMNF2, LMNF12	1960M7	Financial Uncertainty Index, 1-3 and 12 periods ahead	LMN
		Measures of Past Relative Performance	
rMSFE	1970:M3	relative MSFE for the past period	O

Notes: Sources are abbreviated as follows: "F"- Federal Reserve Economic Data (FRED), "BBD"- Baker, Bloom and Davis (2016), "HRS"- Husted, Rogers and Sun (2016), "JLN"- Jurado, Ludvigson and Ng (2015), "LMN"- Ludvigson, Ma and Ng (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	rel. RMSE	p-value	model	rel. RMSE	p-value
Panel A. Full Evaluation Sample (1970M3-2016M1)					
Industrial Production					
IP: Durable Materials	0.96	0.10	Capacity Utilization: Man	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 Common Stock Price Index: Ind.	0.95	0.05
Avg Weekly Hours: Man.	0.98	0.07	3-Month Treasury Minus FFR	0.84	0.00
			6-Month Treasury Minus FFR	0.85	0.00
			1-Year Treasury Minus FFR	0.91	0.01
			5-Year Treasury Minus FFR	0.86	0.00
			10-Year Treasury Minus FFR	0.86	0.01
			Moody's Aaa Corp. Bond Minus FFR	0.83	0.00
			Moody's Baa Corp. Bond Minus FFR	0.81	0.00
Inflation					
			Real M2 Money Stock	0.95	0.06
Panel B: Partial Evaluation Sample (1979M2-2016M1)					
Industrial Production					
IP: Nondurable Consumer Goods	0.98	0.05	All Employees: Mining and Logging: Mining	0.93	0.10
IP: Materials	0.95	0.08	3-Month Treasury Minus FFR	0.92	0.04
Initial Claims	0.96	0.03	6-Month Treasury Minus FFR	0.91	0.02
All Employees: Durable goods	0.97	0.10	Moody's Baa Corp. Bond Minus FFR	0.88	0.05
Effective FFR	0.97	0.05			
3-Month AA Fin. Comm. Paper Rate	0.97	0.06			
3-Month Treasury Minus FFR	0.98	0.03			
6-Month Treasury Minus FFR	0.97	0.02			
1-Year Treasury Minus FFR	0.98	0.02			
Inflation					
Crude Oil, spliced WTI and Cushing	0.97	0.04			

Notes: The table shows the unconditional test for equal predictive ability for the models which are statistically different than the benchmark at 10% significance level. FFR stands for Federal Funds Rate.

Table 3. Conditional Tests of Equal Predictive Ability

Model	h=1 relative perform	h=1 p-value	Model	h=12 relative perform	h=12 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: Man.	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 Fin. Comm. Paper Rate	0.03	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, Transp. & Utilities	0.49	0.03
1-Year Treasury Rate	0.00	0.03	All Employees: Wholesale Trade	0.46	0.03
6-Month Treasury Minus FFR	0.27	0.05	All Employees: Retail Trade	0.40	0.01
1-Year Treasury Minus FFR	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 cons. credit/Pers. 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 FFR	0.38	0.00
			3-Month AA Fin. Comm. 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 Corp. Bond Yield	0.38	0.00
			Moody's Seasoned Baa Corp. Bond Yield	0.38	0.01
			3-Month Commercial Paper Minus FFR	0.36	0.00
			3-Month Treasury Minus FFR	0.17	0.00
			6-Month Treasury Minus FFFR	0.14	0.00
			1-Year Treasury Minus FFR	0.25	0.00
			5-Year Treasury Minus FFR	0.24	0.00
			10-Year Treasury Minus FFR	0.25	0.00
			Moody's Aaa Corp. Bond Minus FFR	0.22	0.00
			Moodyís Baa Corp. Bond Minus FFR	0.20	0.00
			CPI : Medical Care	0.31	0.02
			CPI : All Items Less Food	0.36	0.01
Panel B. Inflation: (KCFSI/STLSFI)			(LAGGED)		
KCFSI:IP: Final Products	0.73	0.01	Real Personal Income	0.45	0.06
All Employees: Trade, Transp. & 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
STLSFI:S&P Price-Earnings Ratio	0.54	0.09	Civilians Unemployed for 15-26 Weeks	0.49	0.06
Canada / U.S. Foreign Exchange Rate	0.61	0.05	All Employees: Service-Providing Industries	0.39	0.07
Crude Oil, spliced WTI and Cushing	0.00	0.08	Housing Starts: Total New Privately Owned	0.38	0.07
			New Orders for Durable Goods	0.51	0.06
			M1 Money Stock	0.43	0.06
			St. Louis Adjusted Monetary Base	0.08	0.09
			6-Month Treasury Minus FFR	0.36	0.10

Notes: The table shows the relative performance of the models measured by M_{GW} statistic, which captures the magnitude of the improvement of a benchmark model over the average improvement induced by the conditioning variable. The smaller this number, the greater the gains for the alternative model. We further show the p-values of the Giacomini and White (2006) conditional predictive ability test for models that based on the selection rule would be preferred over the benchmark at least half of the time in the out of sample. FFR stands for Federal Funds Rate.