Predicting Relative Forecasting Performance: an Empirical Investigation*

Eleonora Granziera† and Tatevik Sekhposyan‡

This version: February 12, 2019

Abstract

The relative performance of forecasting models changes over time. This empirical observation raises two questions: is the relative performance itself predictable? If so, can it be exploited to improve forecast accuracy? We address these questions by evaluating the predictive ability of a wide range of economic variables for two key US macroeconomic aggregates, industrial production and inflation, relative to simple benchmarks. We find that business cycle indicators, financial conditions, uncertainty as well as measures of past relative performance are generally useful for explaining the relative forecasting performance of the models. We further conduct a pseudo-real-time forecasting exercise, where we use the information about the conditional performance for model selection and model averaging. The newly proposed strategies deliver sizable improvements over competitive benchmark models and commonly used combination schemes. Gains are larger when model selection and averaging are based on financial conditions as well as past performance measured at the forecast origin date.

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 2018 12th International Conference on Computational and Financial Econometrics, 2018 1st Vienna Workshop on Economic Forecasting, 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. We would also like to thank the editor and three anonymous referees who kindly reviewed the earlier version of this manuscript and provided valuable suggestions and 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 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 (2018) evaluate relative forecasting performance of Phillips curve-type models for inflation 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 1990s up to late 2000s, 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 relative forecasting performance of the models.

INSERT FIGURE 1 HERE

Motivated by these empirical observations, in this paper we answer two questions: (i) whether the relative forecasting performance of different models is predictable; (ii) whether the predictability can be exploited to produce more accurate forecasts.

To answer these questions, we consider a wide set (more than hundred) of simple forecasting models of autoregressive distributed type to predict US industrial production and inflation. 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 test whether the predictive ability of economic models relative to simple autoregressions can in itself be predicted by some

[1] See Rossi (2013) for an overview.

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 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). When comparing forecasting performance, it is common to apply tests of *unconditional* predictive ability, which ask whether the forecasting models performed equally well on average in the past. Then the testable hypothesis is whether the expected loss differences have a zero mean. A rejection of the null gives 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. However, a researcher might be interested in knowing whether a model is more accurate at a specific future date. As noted, one could, for example, wonder whether being in a recession or expansion could help to choose a model for a particular future date. In this context, 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 implies that forecasts can be improved. To this end, we suggest two rules that help to either select among or average across the existing models. An alternative strategy would be to re-specify 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 conditioning 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

[2]An alternative strategy would be to postulate the relative forecasting performance of the models as a regime-switching model and use the estimated state transition probabilities matrix for selecting a particular forecasting model. Fossati (2017) proposes a test 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 models and re-compute the forecasts. 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.

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 an 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 a 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. In contrast, 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 conditions are tight or uncertainty is high, etc. We also find past relative performance to be a good predictor of future performance. Moreover, using conditioning information in a simple decision rule in many cases results in sizable gains, 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 $w_t = \{y_{s+\tau}, x_s\}_{s=1}^{t}$ are time series variables at each forecast origin $t = R, ..., T-\tau$, where T is the total sample size, R is the estimation window size and $\tau \geq 1$ is the integer-valued 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 (the best of) two models. Though a multi-model comparison might be the preferred approach if one is ultimately interested in obtaining the most accurate forecast, bi-model comparisons help us in understanding the predictive content of each economic variable in particular periods of time.[5]

Denote by $f_{t,R}\left(\hat{\beta}_{0,t}^{R}\right) = f\left(w_t, w_{t-1}, ..., w_{t-R+1}; \hat{\beta}_{0,t}^{R}\right)$ and $g_{t,R}\left(\hat{\beta}_{1,t}^{R}\right) = g\left(w_t, w_{t-1}, ..., w_{t-R+1}; \hat{\beta}_{1,t}^{R}\right)$ the τ-period-ahead forecasts obtained from the models estimated with a fixed rolling window of size R, where $\hat{\beta}_{0,t}^{R}$ and $\hat{\beta}_{1,t}^{R}$ are the estimated regression coefficients. In what follows we take $f_{t,R}\left(\hat{\beta}_{0,t}^{R}\right)$ 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 quadratic loss to evaluate the accuracy of the forecasts. This choice is motivated by the wide use of the quadratic loss in empirical studies which assess forecast performance of models for inflation and real activity.

Let $\Delta L_{R,t+\tau} = \left(y_{t+\tau} - f_{t,R}\left(\hat{\beta}_{0,t}^{R}\right)\right)^2 - \left(y_{t+\tau} - g_{t,R}\left(\hat{\beta}_{1,t}^{R}\right)\right)^2$. 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

[4] This framework allows data to be non-stationary. However, the type of non-stationarity considered rules out unit roots, but allows for breaks that could be induced by distributions changing over time. In our empirical application all data have been transformed to eliminate unit roots.

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

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 $H_0 : E[\Delta L_{R,t+\tau}] = 0$, i.e. whether the models are equally accurate on average.

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

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

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 Heteroskedasticity 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 $[0.75P^{1/3}]$.[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 (the regression coefficients denoted by a $\hat{\delta}^{\tau}_{P,t}$ vector.)[8] 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

[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 $\chi^2_{2,1-\alpha}$.

[7] The choice of the bandwidth parameter is motivated by the recommendation in Stock and Watson (2010, p. 599). For robustness we try two alternative approaches to estimating the variances σ_P and V: (i) for one-step-ahead forecasts we use the sample covariance (this is consistent with the recommendation in Giacomini and White, 2006); (ii) for twelve-steps-ahead forecasts, we use Newey-West and Equal-Weighted Cosine (EWC) estimators and conduct inference under fixed-b asymptotics, picking relevant truncation parameters consistent with the recommendations in Lazarus et al. (2018). Our results are robust to the use of the alternative estimators/inferential frameworks and are reported in the appendix.

[8] Note that $\hat{\delta}^{\tau}_{P,t}$ is a $q \times 1$ vector, where q is the row dimension of h_t. In our empirical application h_t includes an intercept term as well as one conditioning variable. Consequently, $\hat{\delta}^{\tau}_{P,t}$ is a 2×1 vector.

is concerned with using the information provided by the conditioning variables to pick a model for a 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), $\hat{\delta}_{P,t}^{\tau\prime} h_t \approx E\left[\Delta L_{R,t+\tau} | \mathcal{F}_t\right]$, with the product $\hat{\delta}_{P,t}^{\tau\prime} h_t$ 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 otherwise. Then, if the alternative model is more accurate than the benchmark during recession episodes both the estimated $\hat{\delta}_{P,t}^{\tau\prime}$ coefficient and the fitted values $\hat{\delta}_{P,t}^{\tau\prime} h_t$ would be strictly positive during recessions, assuming no intercept in the set of conditioning variables h_t. In other words, both the marginal impact of the conditioning variable over the sample as well as the value of the conditioning variable at a particular point in time jointly provide valuable information. Certainly, the current discussion is in terms of full sample analysis, however, as we demonstrate further, it is possible to use the conditioning variables in a pseudo-real-time forecasting exercise as well (hence, the subscript t in $\hat{\delta}_{P,t}$).

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 reject. The interpretation of this would be that the two models are the same on average, yet their relative predictive performance could be predicted. We find this to be the most frequent occurrence in our empirical investigation. 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 occur 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 an information set that are at most weakly correlated with the relative performance of the models, then the power of the conditional predictive ability test will deteriorate.

2.2 Predicting Relative Performance: Decision Rules

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 (combined) models. As discussed in the

introduction, we propose model selection and model averaging as a simple, flexible and scalable way of dealing with model misspecification as opposed to postulating new types of models.

2.2.1 Model Selection

As a model selection criterion, we empirically evaluate Giacomini and White's (2006) model selection rule. More specifically, we divide our out-of-sample period P into two parts: the first subsample will be used to "train" the rule and the second subsample to evaluate it. Let S be the initial training window size for the implementation of the rule. Further, at any given point in time, we use a fixed rolling sample of S observations for conditional testing/"re-training". Thus, we follow a two-step rule:

1. Regress the loss differences, $\{\Delta\hat{L}_{R,j+\tau}\}_{j=t-S+1}^{t}$, on a single conditioning variable and a constant, denoted by $\{h_j\}_{j=t-S+1}^{t}$, in a window of S observations in the out of sample. Let the vector of regression coefficients be $\hat{\delta}_{S,j}^{\tau}$.

2. Predict $y_{j+\tau}$ using the forecast of the benchmark model if $\hat{\delta}_{S,j}^{\tau\prime}h_j \leq 0$, $j = R+S, ..., T-\tau$, and use the alternative model otherwise[9].

Note that h_j is the value taken by the conditioning variable in the last observation included in the training sample and therefore the prediction of $y_{j+\tau}$ is done in (pseudo) real time. We further consider a modified version of this rule in that we use the information on the statistical significance of $\hat{\delta}_{S,j}^{\tau}$. In other words, in this version of model selection, step (1) would be followed by steps (2') and (3') as:

2'. Assess the statistical significance of $\hat{\delta}_{S,j}^{\tau}$ using a two-sided test;

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

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 j state of the conditioning variable, h_j, is expected to improve the forecasting performance of the alternative model over the benchmark for a future date $j+\tau$.

2.2.2 Model Averaging

Forecast combinations have been frequently found to outperform strategies based on a single, best model selection in empirical studies. For this reason, we propose a rule for model averaging where

[9]Note that a positive value for a loss differential implies that the alternative model is better than the benchmark and vice versa.

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

1. Construct a binary variable $\{D_{S,j+\tau}\}_{j=t-S+1}^{t}$, which takes the value one if the loss differences $\Delta\hat{L}_{R,t+\tau}$ is positive (alternative model is better) and zero otherwise;

2. Regress the binary variable, $\{D_{S,j+\tau}\}_{j=t-S+1}^{t}$, on a single conditioning variable and a constant, denoted by $\{h_j\}_{j=t-S+1}^{t}$, in a window of S observations in the out of sample. Let the vector of regression coefficients be $\hat{\delta}_{S,j}^{\tau}$.

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

$$w_{1,j} = \hat{\delta}_{S,j}^{\tau\prime} h_j, \quad j = R+S, ..., T-\tau.$$

   Consequently, the implied weight of the benchmark model is $w_{0,j} = 1 - w_{1,j}$.

We use a linear probability model to compute the forecast combination weights. The weight assigned to the alternative model is just an estimate of the probability that the τ-step-ahead squared forecast error from the alternative model is smaller than the squared forecast error from the benchmark model, conditional on h_t, i.e.

$$w_{1,j} = P\left(e_{1,j+\tau}^{2} < e_{0,j+\tau}^{2} \mid h_t\right).$$

Note that a probit model can be used to estimate the weights instead of a linear probability model. However, we found the estimation of a probit model (multiple times) to be computationally more expensive, thus, we resorted to a linear probability model. Moreover, even though in the linear probability model we leave the estimation unrestricted, we obtain weight estimates bounded between zero and one.

By construction, the 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 and follows up with a model selection or an averaging exercise, where (s)he uses the results of the conditioning test 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.[10]

[10]The exercise is pseudo-real-time as we do not consider the real time nature of either the data that goes into the forecasting models or the conditioning variables. Certainly, for the variables that are not subject to revisions the exercise is real-time by construction. However, these variables are a small proportion of the total. It is infeasible to extend the analysis in the paper to real-time as for most conditioning variables real-time data vintages are not available or start much later in the considered sample period.

3 Forecasting Models

We consider forecasting monthly industrial production 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 lags of a 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 growth of industrial production 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 rates τ-periods into the future. $x_{t,k}$ 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 $K = 117$.[11] Similarly, y_t is either the period t growth rate of industrial production, that is $y_t = 1200\ln(IP_t/IP_{t-1})$ or the period t change in inflation, that is $y_t = 1200\ln(CPI_t/CPI_{t-1}) - 1200\ln(CPI_{t-1}/CPI_{t-2})$.[12] 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, also chosen with the BIC criterion. 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 an autoregressive model, where we use only the lagged dependent variable to forecast growth in industrial production 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's (2006) framework requires the number of observations used in the estimation to stay finite relative to the overall sample size.

[11] The dataset for industrial production includes historical data for inflation, but not industrial production (and vice versa) as the lagged dependent variable is automatically included in eq. (2).

[12] Note that, as in Rossi and Sekhposyan (2010), this relies on the assumption that CPI is I(2).

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).[13] 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 1959M1. 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 US dollar index against major currencies, consumer sentiment index and VXO.[14] 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 eliminate unit roots, 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.

Moreover, in our empirical application we adjust for outliers. We treat realizations that are 4 standard deviations larger than the mean as outliers and we substitute them with the mean.[15] 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, divided into four groups: measures of economic activity, financial condition indices, macroeconomic uncertainty indices and measures of past rel-

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

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

[15] 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.

ative 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 have been 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 on average over the past six months prior to the forecast origin date; (ii) "unemp-rec,"which takes the value of one when the unemployment rate in the economy is above 6% in the month preceding the forecast origin date. Moreover, we include in our analysis the US recession probability index of Chauvet and Piger (2008), which consists of the smoothed probabilities from a Markov-switching dynamic-factor 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 depends on financial conditions. In the benchmark specification we use the National Financial Conditions 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 Conditions Index (ANFCI), which extracts a component of financial conditions uncorrelated with economic conditions. Moreover, to disentangle the 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, "STLFSI" and "KFSI,"respectively.

All these indexes are constructed using principal component analysis over a number of financial variables, including interest rates, spreads and stock price 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 1990s. 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 would dramatically cut the number of our out-of-sample observations and make the current study infeasible. Further, many of these indices, namely the ones coming from the Federal Reserve Banks of Chicago and St. Louis, are available at a weekly frequency. We use their monthly aggregate, acknowledging the fact that one could potentially extend the analysis in this paper to model selection and averaging at higher frequencies. However, at this point we leave those considerations 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 structural and time-series models (Bloom, 2009, Ludvigson, Ma and Ng, 2015, Jurado, Ludvigson and Ng, 2015). To capture the different definitions of uncertainty suggested in 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.[16] 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 unpredictable components of economic variables, and we use the uncertainty 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).[17]

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 squared forecast error differential between the benchmark and alternative models to predict the relative forecasting performance of the models in the future. We consider both the latest value of the loss differential as well as the average over the past 12 months as indicators of past performance.

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 episodes happen to be clustered

[16] 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.

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

around the recessions. Figure 2 shows the contemporaneous correlation between the conditioning variables.[18] The real time recession dummy variables that are based on past output growth and level of unemployment show a very low correlation (even negative for the unemployment-based 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.

Multiple Conditioning Variables: While in the paper we condition on each variable one at a time (in addition to a constant), one could ask whether using the information contained in multiple variables simultaneously could be useful. To answer this question we consider the first principal component of the conditioning variables as a parsimonious way to summarize the information content of multiple variables. Specifically, we take the first principal component of: (i) all conditioning variables, excluding past relative performance; (ii) business cycle indicators; (iii) financial conditions/stress indicators and (iv) uncertainty indices. While it might be interesting to design a procedure to optimally choose the variables to be included in the vector h_t, we leave this for future research.

5 Results

We first present the results for 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 1970M3 to 2016M1. 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 are exploitable and can be used to improve the accuracy of the forecasts in pseudo-real-time.

5.1 Equal Predictive Ability Tests

Figure 3 shows the results for the unconditional predictive ability test for one-month-ahead ($\tau = 1$) and twelve-month-ahead ($\tau = 12$) forecast horizons. The horizontal axis displays relative root mean squared error expressed as the ratio between the RMSE of the alternative ADL models over the RMSE of the benchmark autoregressive model.[19] Ratios greater than one, i.e. to the right of the vertical (red) line, indicate that the economic models perform worse than the autoregressive benchmark. The vertical axis indicates the $p-$values from the Giacomini and White (2016) unconditional predictive ability 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

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

[19] Note that the horizontal axis is different across variables and forecasting horizons in order to accommodate outliers in a visually pleasant manner.

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.[20] 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 table has two Panels. Panel A directly compares to Figure 3 and shows the results for the full evaluation sample.[21] 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 period, and it alternates between positive and negative values for the last part of the sample, starting with the Great Recession. Another general finding is that there is more predictability in industrial production at longer horizon ($\tau = 12$) relative to the one-month ahead ($\tau = 1$). At this shorter 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, aside 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 test results, but for a different subsample which spans 1979M2-2016M1. We provide this result for direct comparability with our pseudo-real-time exercise. When constructing 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 is whether improvements in accuracy found by applying our

[20] The large amount of alternative models considered, might raise concerns about multiple testing bias. Therefore, we check the robustness of the results by running another experiment where the p-values are adjusted as in Benjamini and Hochberg (1995) or using the Bonferroni approach. In this case rejections are substantially less frequent. However, we still obtain many more rejections in the conditional equal predictability framework relative to the unconditional one. Results are reported in Table A.1. and Figure A.1 of the appendix.

[21] Note that RMSEs associated with twelve-steps-ahead forecasts are smaller than those associated with one-step-ahead forecasts for both inflation and output growth. This is due to the definition of the target variable: the twelve-steps-ahead forecast is for the *average* cumulative growth in the next twelve periods. If you were to change the definition of the target to be the twelve-steps-ahead monthly growth rates, the longer-horizon forecasts would have higher RMSEs relative to one-step-ahead forecasts.

model selection and model averaging strategy could be driven by the choice of the evaluation sample. As Panel B shows, even in the shorter evaluation sample there is not much evidence of improved relative predictability for the economic models. 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 industrial production predictability. However, there are a few less variables that help with twelve-month-ahead predictability of the industrial production relative to the longer sample.

What drives these results? The test statistic of the unconditional predictive ability test is proportional to the sample average of the loss differential. This implies that even if there are large differences between the losses, 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. This is illustrated through our example in Figure 1, where the p-value of the unconditional equal predictive ability test is 0.27. 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 FIGURES 4 AND 5 HERE

Results for the conditional predictive ability tests are provided in Figure 4 for industrial production and in Figure 5 for inflation. In both figures the horizontal axis shows 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 alternative model is worse than the benchmark given the value of the conditioning variable. Recall that $\hat{\delta}_{P,t}^{\tau\prime} h_t$, i.e. the fitted values in the regression of loss differentials on the conditioning variables, is an approximation to the expected conditional loss differential, $E\left[\Delta L_{P,t+\tau} | \mathcal{F}_t\right]$. Given that, we compute statistic $I_{GW} = \frac{1}{P} \sum_{t=R}^{T-\tau} 1\left\{\hat{\delta}_{P,t}^{\tau\prime} h_t \leq 0\right\}$ as suggested in Giacomini and White (2006). Since our definition of loss differential is the squared loss of the benchmark model minus the squared loss of the alternative, a positive value of $\hat{\delta}_{P,t}^{\tau\prime} h_t$ indicates a better performance for the alternative model. In addition to presenting the I_{GW} indicator, we also mark the $p-$ values associated the marginal effects of the conditioning variables. In the figures, models that perform significantly better than the benchmark for most of the sample will be located in the lower left quadrant. Note that an indicator value larger than 0.5 does not necessarily mean that the alternative model should be deemed worse than the benchmark. In fact, it is consistent with a scenario in which the alternative model performs better than the benchmark but only in episodes (e.g. recessions) which are not very frequent over the sample.[22]

[22]We also control for multiple testing bias by applying the procedures by Benjamini and Hochberg (1995) and Bonferroni. Similar to the unconditional tests, with these adjustments we obtained fewer rejections. Results are displayed in Figures A.2 and A.3 of the appendix.

For both target variables, i.e. industrial production and inflation, and for both forecast horizons we only report results for the two conditioning variables that give us the highest number of rejections of the null of equal predictive ability across the wide set of ADL models considered.[23] For both inflation and industrial production, at one-year-ahead forecast horizon, the unemployment rate-based recession indicator as well as the lagged performance of the models are the conditioning variables that offer the highest frequency of incidents where the alternative model improves over the benchmark in the full out-of-sample.[24] For the one-month-ahead forecast horizon, however, it appears that financial indices matter more. For the growth rate of industrial production, two indices developed by the Chicago Fed, the Adjusted National Financial Conditions Index (ANFCI) and the National Financial Conditions Index-Leverage (NFCILeverage) appear to be more useful, while for inflation the important conditioning variables are the financial indices constructed by the Kansas City and St. Louis Feds (KCFCI and STLFSI). Table 3 provides a list of the models for which, given a conditioning variable (e.g. ANFCI, KCFCI, and lagged performance), we reject the null of equal conditional predictive ability at the 10 percent significance level and the indicator variable, I_{GW}, takes a value lower than 0.5.

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. The results are particularly stronger for industrial production at a one-year-ahead horizon. As observed for the unconditional predictive ability test, there is a lot less forecastability in inflation than in output growth.

Table 3 lists the models which perform significantly better than the benchmark in Figures 4 and 5, i.e. the models in the low left quadrant of the figure. The relative performance column shows the statistic:

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

which is bounded between zero and one. This gives us an idea about the *magnitude* of the expected improvement of the alternative over the benchmark induced by the conditioning variable. Given our definition of the loss differentials (squared losses of the benchmark minus that of the alternative), 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

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

[24] Because the unemployment rate-based recession indicator is a dummy variable, it can only take four possible values: zero if the alternative is always better than the benchmark, 0.46 if the alternative is better during recessions (as their frequency over the sample is 46%), 0.54 if the alternative is better during expansions (as expansions occur over 54% of the sample), one if the alternative is always worse than the benchmark.

While for the twelve-steps-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. Specifically, the conditional predictive ability test finds evidence of an empirical relationship between inflation and narrow money (M1) and between inflation and unemployment (civilians unemployed for 15-26 weeks).

The indicator variable suggested by Giacomini and White (2006), I_{GW}, provides with a summary statistic of the conditional relative performance of the models. However, it does not tell us the reason for the rejection. In particular, it does not help to identify under which circumstances one model is more accurate than the other. Then, as a complementary analysis, we suggest two exercises: (i) to check the sign of the regression coefficient $\hat{\delta}'_{P,t}$; (ii) to plot the conditioning variable against the time series $\left\{\hat{\delta}'_{P,t}h_t\right\}_{t=R}^{T-\tau}$.

The results of the exercise (i) are displayed in Figure 6 for industrial production and Figure 7 for inflation.[25] A positive (negative) value for the coefficient indicates that the higher the conditioning variable, the higher (lower) the conditional loss differential, i.e. the better (worse) the performance of the alternative model. All the conditioning variables in our sample are defined such that they take higher values during turbulent times. For example, the dummy indicators associated with the recessions take the value of one during recessions and zero otherwise, the financial condition indices are high when financial conditions are tight, while the uncertainty indices increase as uncertainty rises. Then, a positive coefficient means that the alternative is more useful when economic conditions are deteriorating. In the case of lagged performance, a higher coefficient indicates higher persistence in the relative performance of the models. For industrial production the coefficients are mostly positive, while for inflation results are more mixed: they are generally positive for lagged performance but mostly negative when conditioning on the STLFSI or KCFSI indices.

INSERT FIGURES 6 AND 7 HERE

Regarding the exercise (ii), it is instructive to focus on some specific interesting examples, namely on predictability based on the phases of the business cycle and on financial conditions, which we do below.

5.1.1 Predictability and Business Cycle Phases

First, we analyze conditional predictability in recessions. There is a vast literature documenting that the behavior of macroeconomic variables differs in various phases of the business cycle. As

[25] We should note that in the reported conditional predictive ability test results we only show the regression coefficient associated with the economic variables – the coefficient associated with the constant is not reported.

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 a conditioning variable does reject the null of equal predictive ability in a number of bi-model comparisons. Figure 8 plots our target variables, industrial production and inflation, as well as the NBER recession dates (shaded bars) and lists some of 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 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 in regular times 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 labor market such as the civilian unemployment rate or employment in the retail trade sector, are useful in prediction during recessions. Personal consumption expenditures on durable goods turn out to be a useful indicator in contractions as well. These results are in line with those reported in Dotsey, Fujita and Stark (2018).

INSERT FIGURE 8 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 (2018) find that the dynamics of macro variables change during period of high financial stress. Adrian, Boyarchenko and Giannone (2016) study the evolution of the distribution of output over time and document that only the left tail varies with changing 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 9 by two examples for industrial production at one-step-ahead horizon 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. On the contrary, 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 9 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. Rejecting the unconditional test in favor of the benchmark or 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 and financial conditions are deteriorating.

5.2 Decision Rule

We interpret rejections of the null of equal conditional predictive ability as indication of misspecification of the models, such as, for example, non-linearities in the relation among macroeconomic variables. 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 the ability to estimate it and to produce new forecasts.[26] Moreover, a rejection of the conditional test does not point to the source of the misspecification thus is not helpful with the specification of new models. The second strategy, on the other hand, is based on the forecasts of the benchmark and alternative models, which are already available. Therefore, it is quite inexpensive and convenient.

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

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

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 $k = 1, .., K$ we produce forecasts of the target variables at one-step ahead and twelve-step ahead following the procedures detailed above. We then compute the RMSE associated with the forecasts produced with the model selection and model averaging rules and compare them to the RMSE of the benchmark (autoregressive) model. It should be noted that the conditioning test is conducted in pseudo-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 10 HERE

Figure 10 shows, for each conditioning variable, the relative RMSE of the model selection rule versus the benchmark, where the rule is based on the point estimate of the regression coefficient regardless of whether it is statistically significant or not. The figure plots only the models for which the model selection rule delivers a lower RMSE than the benchmark. To provide some additional insight, we mark the models that use economic variables in a similar category with the same symbols and colors. The gains from the model selection are larger at twelve-steps-ahead relative to one-step-ahead, and more so for industrial production than for inflation. Reductions in RMSE can reach 21%, which is a large number compared to the literature. Improved accuracy over the benchmark is achieved for most models in the case of industrial production (on average about 59 models for one-step-ahead and 66 for twelve-steps-ahead) and for a sizable number of models for inflation (about 29 for one-step-ahead and 40 for twelve-steps-ahead). At twelve-steps-ahead, lagged performance seems to be the most robust conditioning variable across target variables. Uncertainty and financial stress indices also help to improve the accuracy of the one-year-ahead industrial production growth forecasts, while uncertainty indices are useful for improving inflation forecasts.

The average gains are sizable and similar to the improvements recorded in the literature. For instance, McCracken and Ng (2016) find similar relative RMSE for US industrial production using an alternative model with a single factor. Note that their exercise is also a pseudo-real-time one since the factor is obtained using the full sample of observations. Diebold and Shin (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.

In terms of specific models, the ones which lead to higher gains are: for industrial production – measures of real economic activity, such as industrial production of durable materials, initial claims, wholesale trade employees at one-step-ahead and also interest rate spreads at twelve-steps-ahead; interestingly, for inflation – oil prices and industrial production at one-step-ahead, and housing

starts, real money supply as well as measures of unemployment at twelve-steps-ahead. Finally, accuracy improvements obtained conditioning on principal components are as good, and in many cases even better, than those obtained from considering a single or a small set of conditioning variables. Therefore, conditioning on principal components might be a viable strategy in the absence of a statistical procedure or economic rationale on how to select one or more conditioning variables.

Figure 11, on the other hand, shows the results of the averaging strategy relative to the autoregressive benchmark in the pseudo-real-time exercise under consideration. The results are grouped by conditioning variables and are similar to that of model selection. In fact, it appears that the model averaging exercise delivers marginally better results than the model selection, as measured by the number of models that improve in forecast accuracy for each conditioning variables and by the magnitude of the improvements at one-month-ahead horizon. On the contrary, the magnitude of the improvements is larger for the model selection rule when looking at forecast accuracy at twelve-steps-ahead forecast horizon.

INSERT FIGURE 11 HERE

While we make general statements about the performance of the large number of models and conditioning variables, we note that specific combination of the conditioning variable and economic model could be of relevance. For instance, in both model selection and model averaging contexts, regardless of the conditioning variable, the model with oil prices provides improved predictions for one-step-ahead inflation. On the other hand, the models with housing starts appear to improve over the benchmark when conditioning on the past forecasting performance, and employment in mining and logging when conditioning on the St. Louis Fed Financial Stress Index for twelve-steps-ahead inflation forecasts.

5.3 Robustness

We conduct a series of experiments to check the robustness of our results. These results are delegated to the Appendix, but we provide a short summary here.

First, we repeat the evaluation of the decision rules for various initial window sizes S used to quantify the importance of the conditioning variable. We consider setting S either equal to 60 or 240. The trade-offs are quite clear: a smaller (larger) window size implies more (less) observations in the forecast evaluation sample (625 for S =60 and 445 for S =240), but leaves fewer (more) observations for estimating the relationship between the loss differential and the conditioning variable. The results, reported in the appendix, are analogous to the ones reported in the paper, obtained using $S = 120$ for both the model selection and averaging exercises.

We also report the results based on the significance of the tests, applying the selection rule that chooses the alternative model only if it is statistically better than the benchmark. This strategy delivers slightly worse results than the ones shown in Figure 10, due to the fact that using this strategy results into substantially fewer instances in which the alternative model is chosen over the benchmark.

While the autoregressive model is a competitive benchmark, as shown, for instance, in Figure 3, it is instructive to compare the performance of our model averaging rule to other forecast combination strategies commonly used in the literature. We first consider the simple average, which is typically found to dominate more complex combination schemes in empirical applications. In general, our averaging rule can deliver lower RMSEs also relative to the equal weighting. However, the gains for industrial production are smaller when compared to the equal weighted average than when compared to an autoregressive model. For instance, we get a maximum of 4% gain instead of 12% at one-month-ahead horizon and 11% instead of 16% at twelve-months-ahead horizon. On average, the gains over the simple average are smaller for inflation as well. Nevertheless, in some instances improvements are substantially larger than when comparing to an autoregressive model, e.g. maximum of 7% vs 3.5% at one-step-ahead horizon and 10.5% vs 8.5% at twelve-steps-ahead. Given the success of the simple average documented in the literature, these results confirm that our averaging scheme is quite competitive.

We further evaluate our averaging rule relative to the optimal weighting scheme suggested by Bates and Granger (1969), which is obtained by minimizing the RMSE of the combined forecast. In this framework the weight associated with the benchmark model is given by

$$\hat{w}_{0,t} = \frac{\hat{\sigma}^2_{1,t} - \hat{\sigma}_{01,t}}{\hat{\sigma}^2_{1,t} + \hat{\sigma}^2_{0,t} - 2\hat{\sigma}_{01,t}},$$

where $\hat{\sigma}^2_{i,t}$ $(i = 0, 1)$ is the sample variance of the forecast errors from model i, and $\hat{\sigma}_{01,t}$ is the sample covariance between the forecast errors of the alternative and benchmark models. We document that this optimal forecast combination scheme does not perform well compared to our averaging rule, consistent with other studies which point to the poor empirical performance of the optimal weighting. The explanation for the high RSMFE of the optimal weights is that it is a costly strategy as it involves the estimation of a covariance. Motivated by this finding, as an alternative combination scheme we try a sub-optimal weighting, where the weight assigned to the benchmark model only depends on the sample variances: $\hat{w}_{0,t} = \hat{\sigma}^2_{1,t} / \left(\hat{\sigma}^2_{1,t} + \hat{\sigma}^2_{0,t}\right)$. In general, the gains from our proposed combination scheme are less pronounced in this case relative to when we compare our results to the optimal combination, except for inflation at twelve-steps-ahead. However, our rule still provides a lower RMSE in many instances.

Finally, we conduct a model averaging exercise with a simple modification to our benchmark

rule: the weight assigned to the alternative model still determined with a linear probability model, but in this case the explanatory variable includes only a constant – there is no additional conditioning variable. The modified rule seldom outperforms the benchmark rule which takes into account the conditioning variable.

The robustness exercises conducted had three main conclusions. First, they confirm that the simple autoregressive model is a difficult benchmark to beat. Second, while forecast pooling can provide some improvements with respect to the autoregressive model, the choice of an averaging scheme matters. In line with existing studies, we find that in many cases simple combination schemes, such as equal weighting, can deliver a lower RMSE than the autoregressive model. Finally, they show that the conditioning variable can help improve the forecast accuracy above and beyond other averaging schemes which neglect the information content of the conditioning variables.

6 Conclusions

In this paper we consider the forecasting performance of a wide range of economic models (in the autoregressive distributed lag family) that use asset prices, measures of real economic activity, wages and prices, as well as money for US industrial production and inflation. We compare the forecasting performance of the economic models to an autoregressive benchmark and study 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 relative performance. We document predictability in the relative forecasting performance based on certain economic fundamentals such as past relative performance and financial conditions indicators.

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 twenty percent improvements in the root mean squared forecast error relative to a competitive autoregressive benchmark in a pseudo-real-time forecasting exercise. Motivated by our results, it might be interesting to design a procedure to optimally choose the variables to include in the conditioning set. In principle, one can use strategies such as bagging outlined in Inoue and Kilian (2008) in our context of tests for conditional forecast evaluation, which will be an interesting avenue to explore in future research.

References

[1] Adrian, T., Boyarchenko, N., and D. Giannone (2016). "Vulnerable Growth," *American Economic Review,* forthcoming.

[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] Bates J.M. and C.W.T. Granger (1969). "The Combination of Forecasts," *Operational Research Quarterly* 20(4), 451-468.

[5] Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rates: a Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1), 289-300.

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

[7] 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.

[8] 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.

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

[10] 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.

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

[12] 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.

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

[14] Diebold, F.X. and M. Shin (2017). "Machine Learning for Regularized Survey Forecast Combination: Partially-Egalitarian Lasso and Its Derivatives," *International Journal of Forecasting,* forthcoming.

[15] Dotsey, M., Fujita, S. and T. Stark (2018). "Do Phillips Curves Conditionally Help to Forecast Inflation?," *International Journal of Central Banking* 14(4) 43-92.

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

[17] Galvao, A.B. and M.T. Owyang (2018). "Financial Stress Regimes and the Macroeconomy," *Journal of Money, Credit and Banking* 50(7), 1479-1505.

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

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

[20] 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.

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

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

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

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

[25] Lazarus, E., Lewis, D. J., Stock J. H. and M. W. Watson (2018). "HAR Inference: Recommendations for Practice," *Journal of Business and Economic Statistics* 36(4), 541-559.

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

[27] 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.

[28] Newey W. K. and K. D. West (1987). "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55(3), 703-708.

[29] 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.

[30] 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.

[31] 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.

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

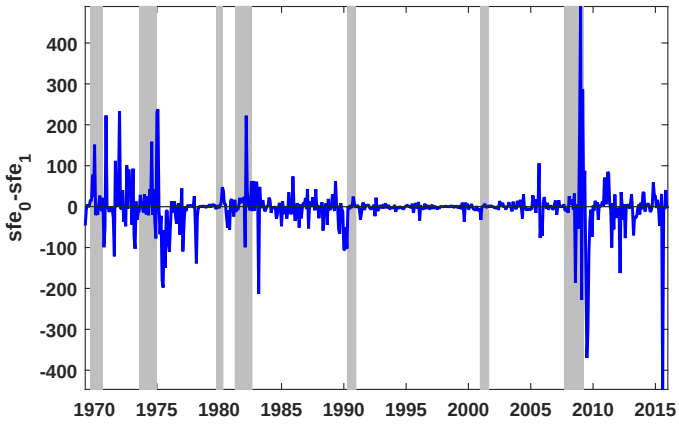
[33] Stock, J.H. and M.W. Watson (2010). *Introduction to Econometrics*, 3rd ed., Addison-Wesley.

[34] 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.

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

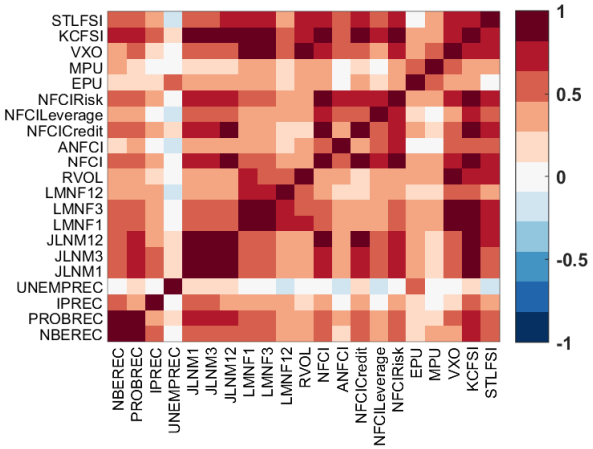
7 Figures and Tables

Figure 1. Illustrative Example



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

Figure 2. Cross-correlation of Conditioning Variables

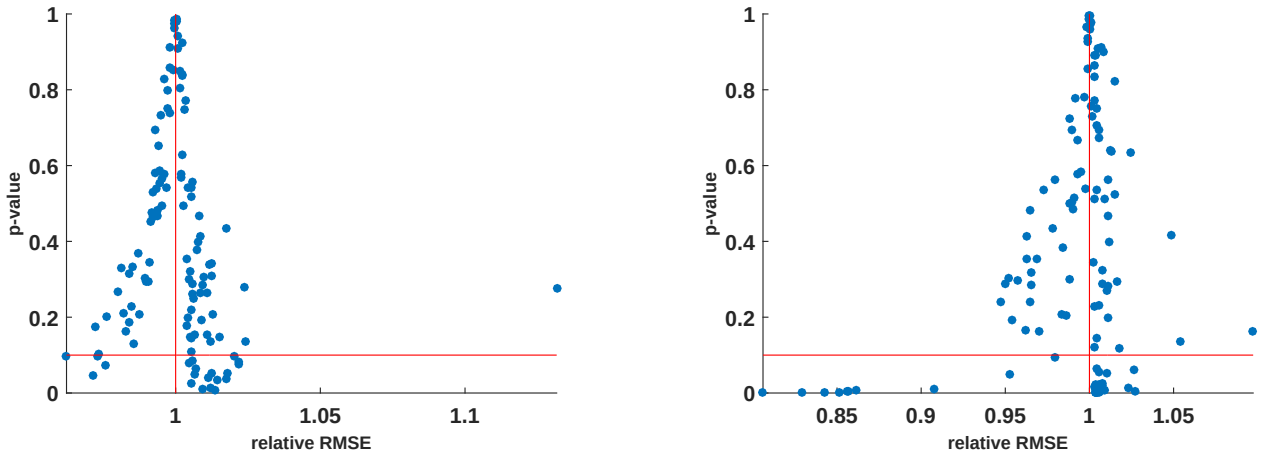


Note: Contemporaneous cross-correlations of conditioning variables. The labels of conditioning variables, marked on the horizontal axis, are consistent with the definitions in Table 1.

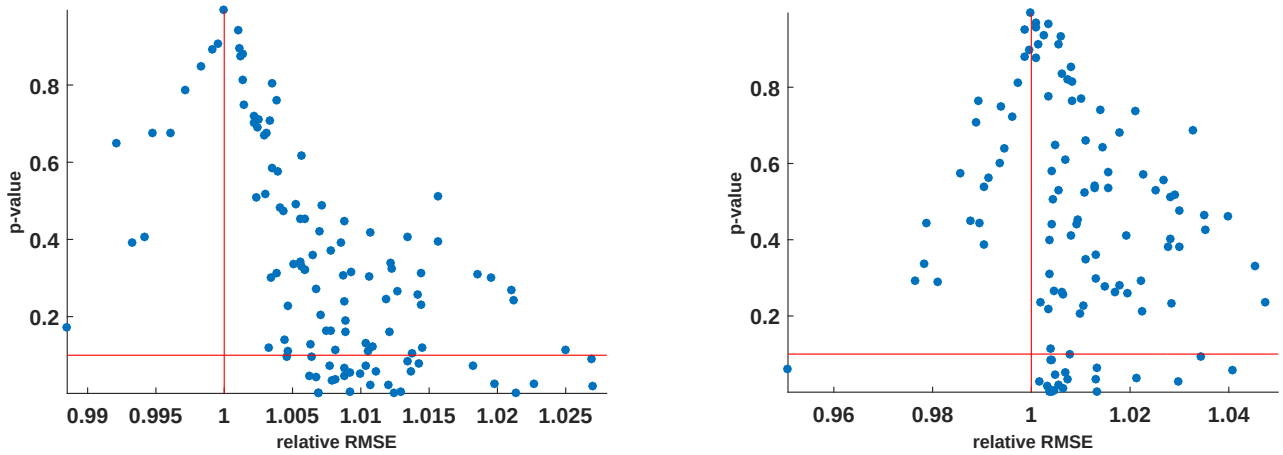
Figure 3. Unconditional Tests of Equal Predictive Ability

$\tau = 1$ $\tau = 12$

Panel A. Industrial Production Growth

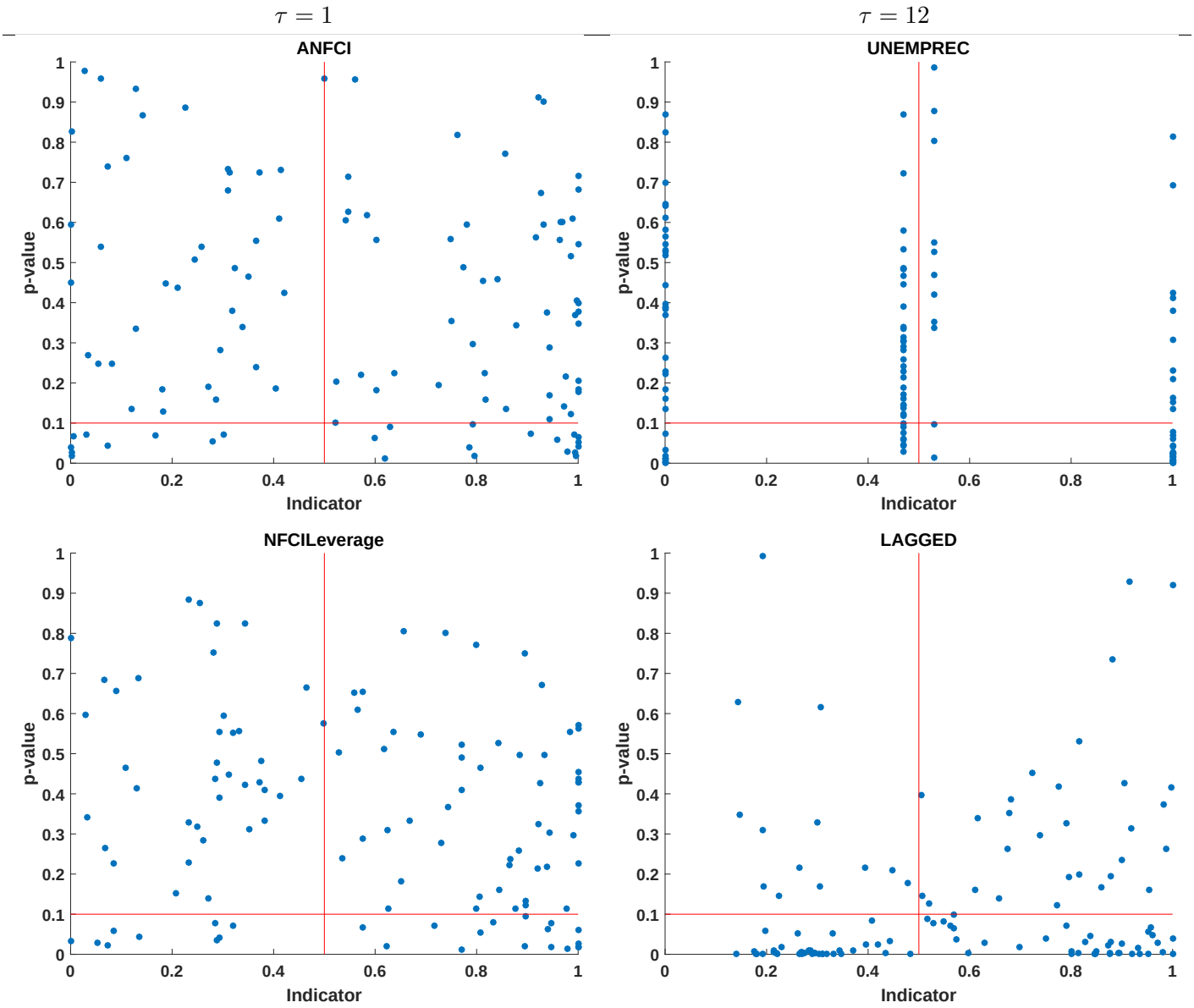


Panel B. Inflation



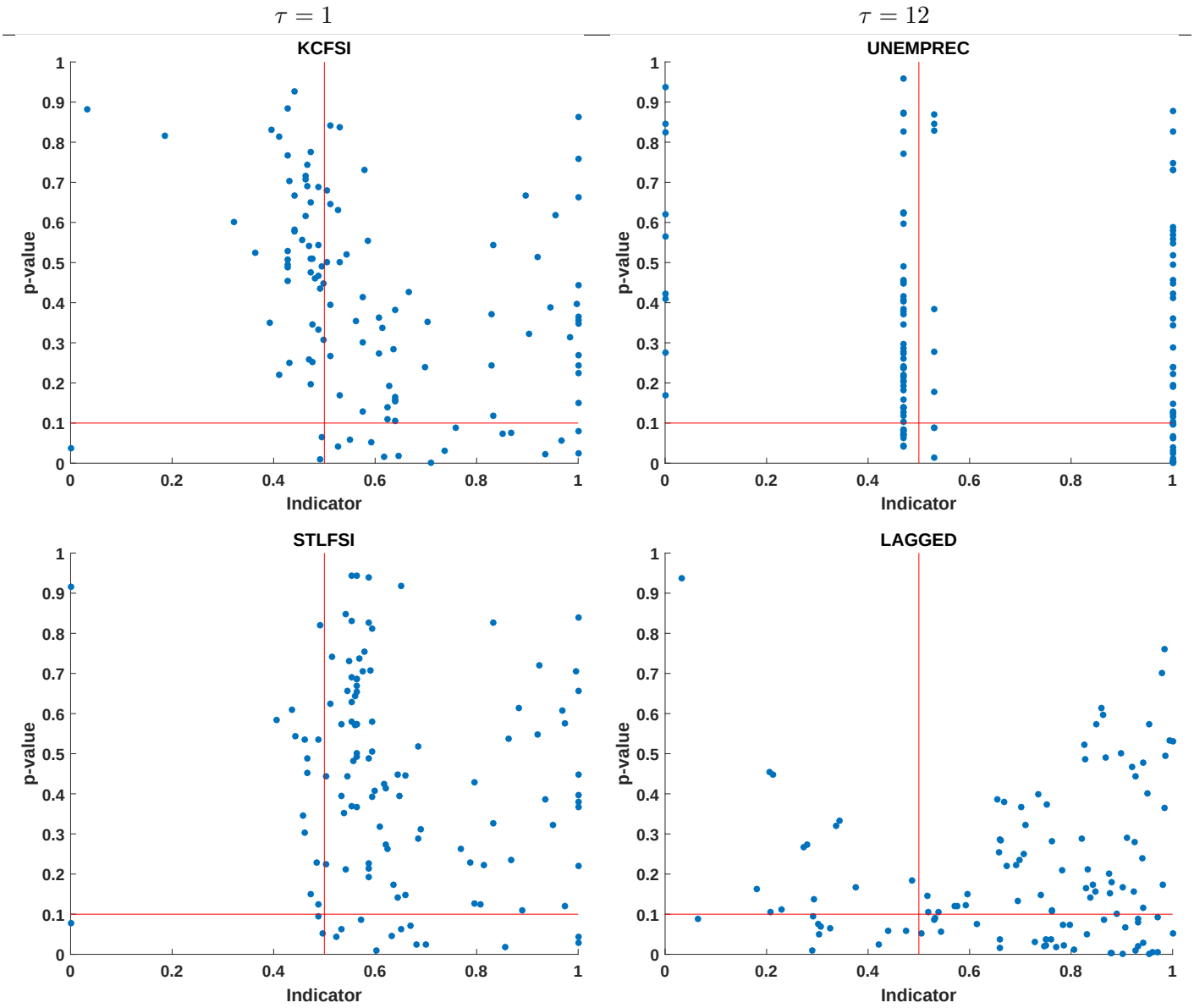
Notes: The figure shows the unconditional equal predictive ability test results for the models of inflation and industrial production growth. The benchmark is an AR, while the alternatives are ADLs, where we consider each economic variable one at a time. The relative RMSE is defined as the ratio of the RMSE of the alternative model over the RMSE of the benchmark. Values greater than one (to the right of the vertical, red, line) favor the benchmark model. Benchmark RMSE for industrial production is 7.91 (one-step-ahead) and 4.08 (twelve-steps-ahead), while for inflation it is 3.05 (one-step-ahead) and 2.51 (twelve-steps-ahead). The horizontal (red) line marks the significance level, which is 0.1 in this case.

Figure 4. Conditional Tests of Equal Predictive Ability: IP



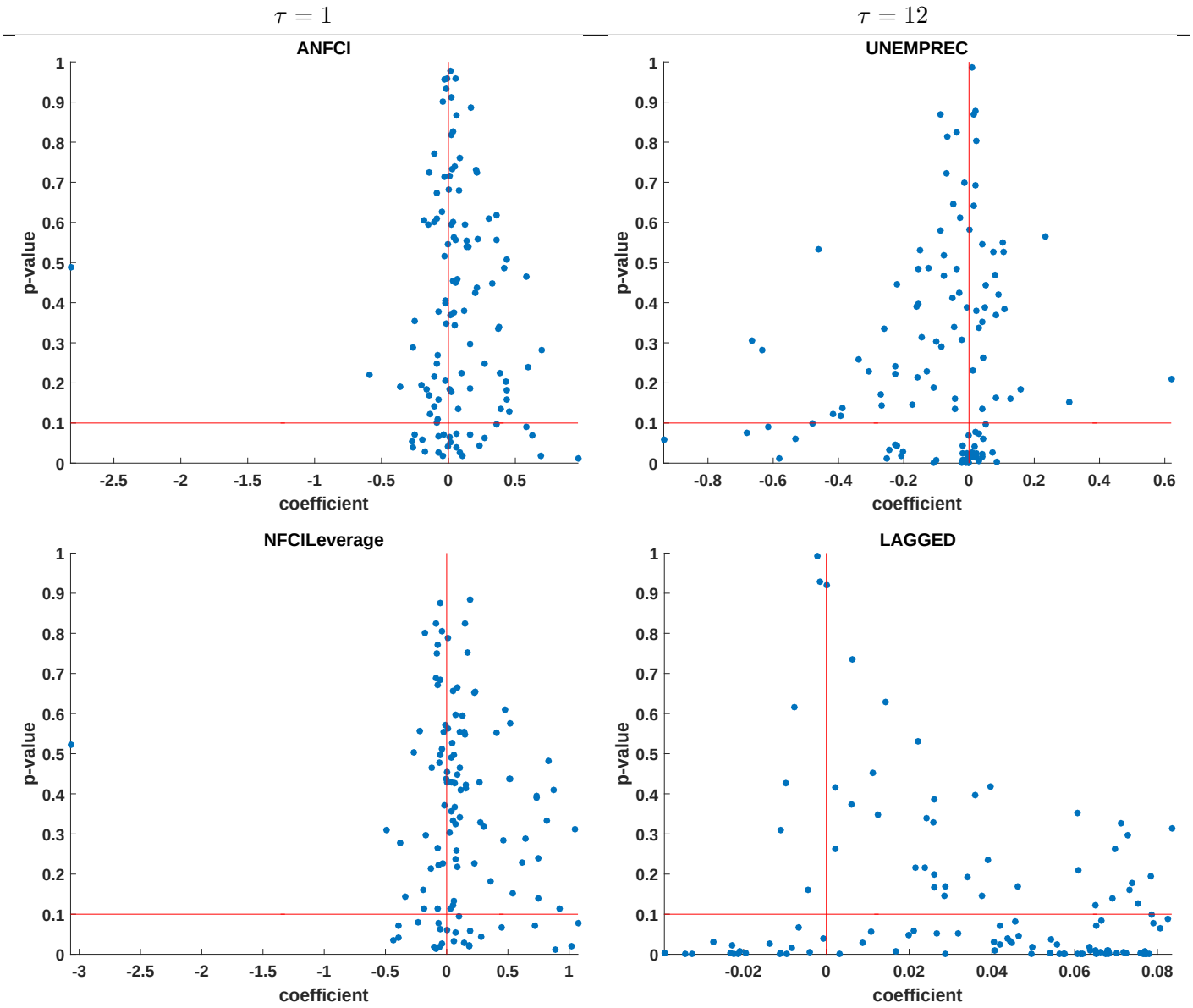
Notes: The figure shows the conditional equal predictive ability test results for industrial production. We report the results for the conditioning variables resulting in the largest number of rejections across the various models. Horizontal axis captures the proportion of the time the benchmark model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

Figure 5. Conditional Tests of Equal Predictive Ability: Inflation



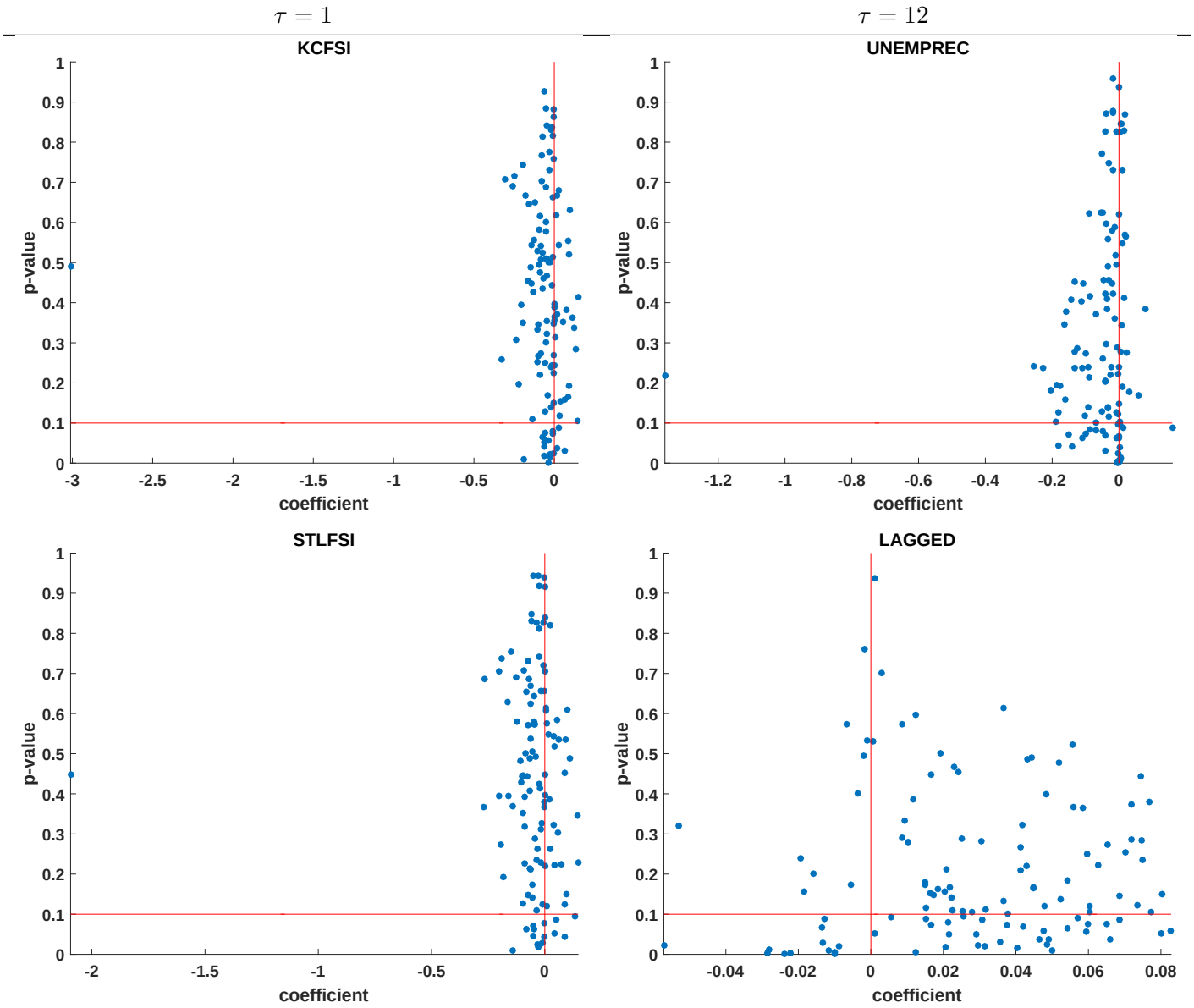
Notes: The figure shows the conditional equal predictive ability test results for inflation. We report the results for the conditioning variables resulting in the largest number of rejections across the various models. Horizontal axis captures the proportion of the time the benchmark model is better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

Figure 6. Conditional Tests of Equal Predictive Ability: Coefficients for IP



Notes: The figure shows the conditional equal predictive ability test results for industrial production. We report the results for the conditioning variables resulting in the largest number of rejections across the various models. Horizontal axis shows the estimated coefficient on the conditioning variable from the regression of the loss differential on the conditioning variable and a constant. A positive coefficient implies that the conditioning variable, on average, improves the relative performance of the alternative model.

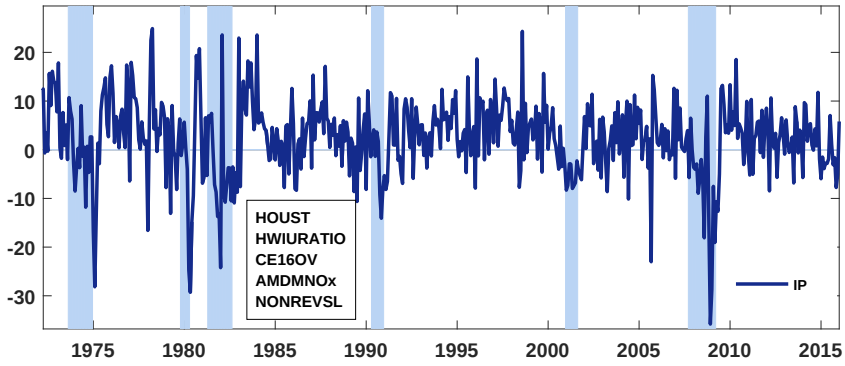
Figure 7. Conditional Tests of Equal Predictive Ability: Coefficients for Inflation



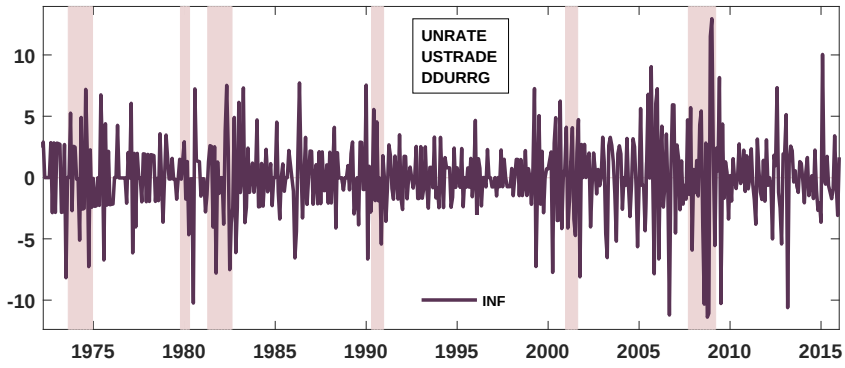
Notes: The figure shows the conditional equal predictive ability test results for inflation. We report the results for the conditioning variables resulting in the largest number of rejections across the various models. Horizontal axis shows the estimated coefficient on the conditioning variable from the regression of the loss differential on the conditioning variable and a constant. A positive coefficient implies that the conditioning variable, on average, improves the relative performance of the alternative model.

Figure 8. Predictive Ability Over the Business Cycle

Panel A. Industrial Production Growth

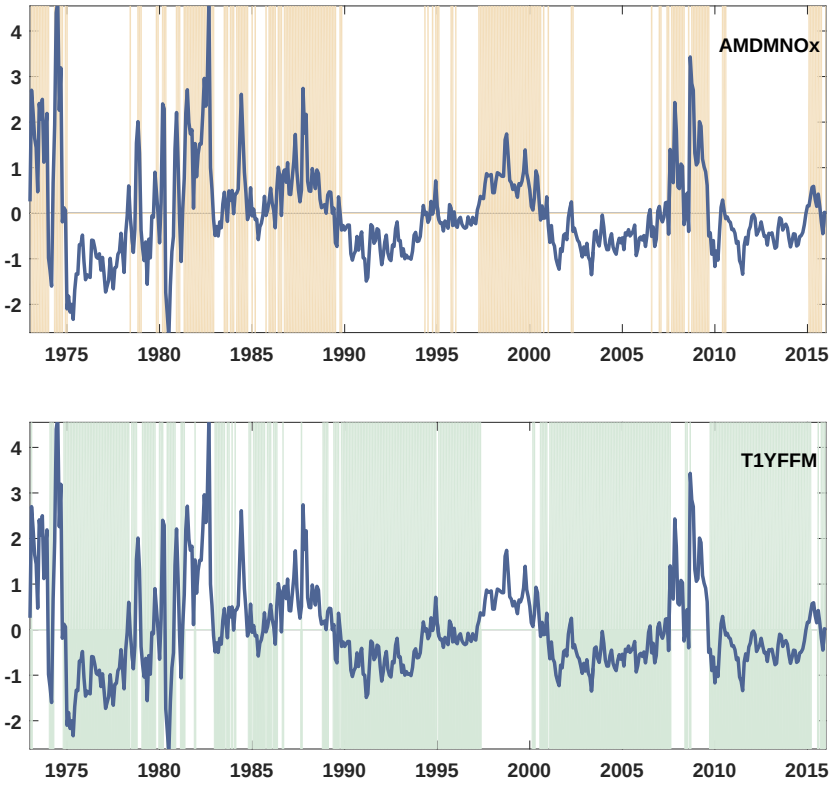


Panel B. Inflation



Notes: The figure displays the conditioning variable, in this case NBER recession dates (shaded areas) and the realizations of target variables (solid lines). It also lists the models selected during recessions. The results are for one-step-ahead prediction."HOUST" stands for Housing Starts: Total New Privately Owned; "HWIURATIO" stands for Help Wanted/No. Unemployed; "CE16OV" stands for Civilian Employment; "AMDMNOx" stands for New Orders for Durable Goods; "NONREVSL" stands for Total Nonrevolving Credit; "UNRATE" stands for Civilian Unemployment Rate; "USTRADE" stands for All Employees: Retail Trade; "DDURRG" stands for Personal Cons. Exp: Durable goods from the McCracken and Ng (2016) database.

Figure 9. Predictive Ability in Industrial Production and Financial Conditions

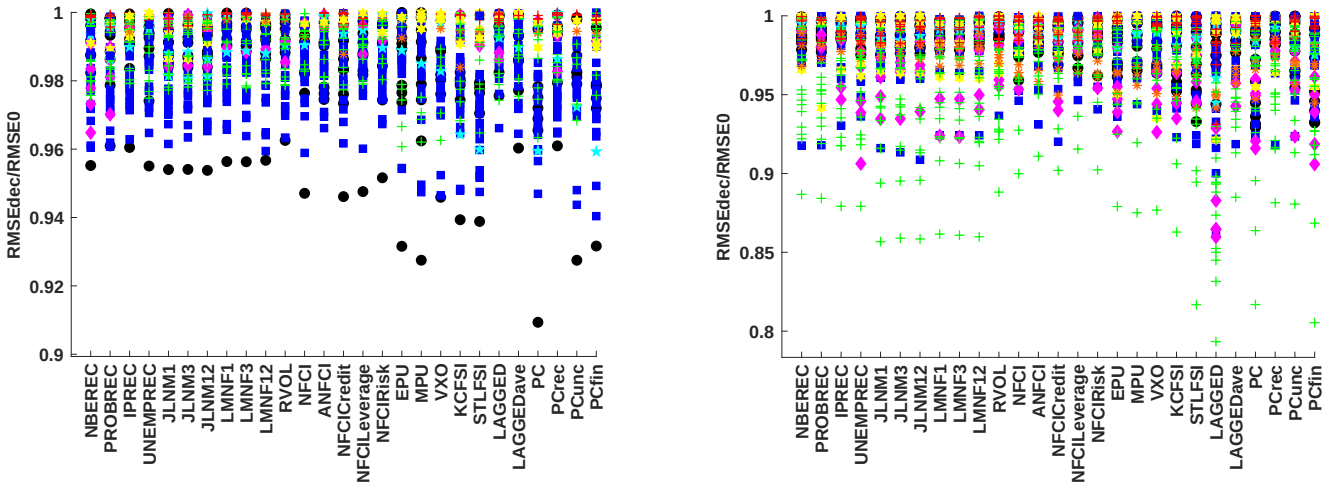


Notes: Conditioning variable is the ANFCI index (blue solid line in both panels). Positive (negative) values of the ANFCI indicate financial conditions that are tighter (looser) than average. Upper panel uses the model with AMDMNOx: new orders for durable goods, while the lower panel uses the model with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

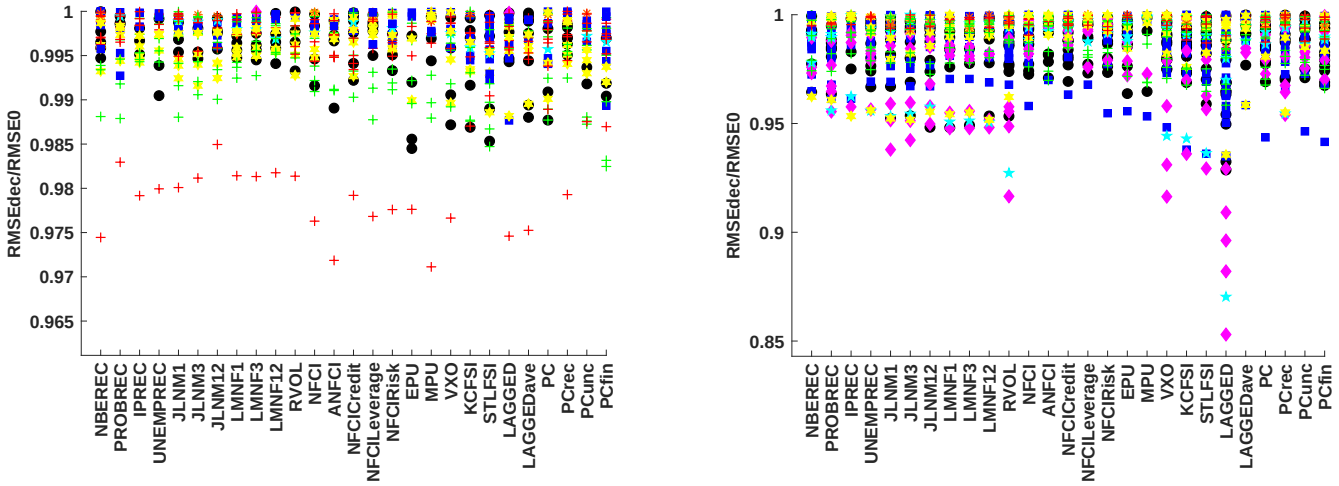
Figure 10. Decision Rule: Model Selection Approach

$\tau = 1$ $\tau = 12$

Panel A. Industrial Production Growth



Panel B. Inflation

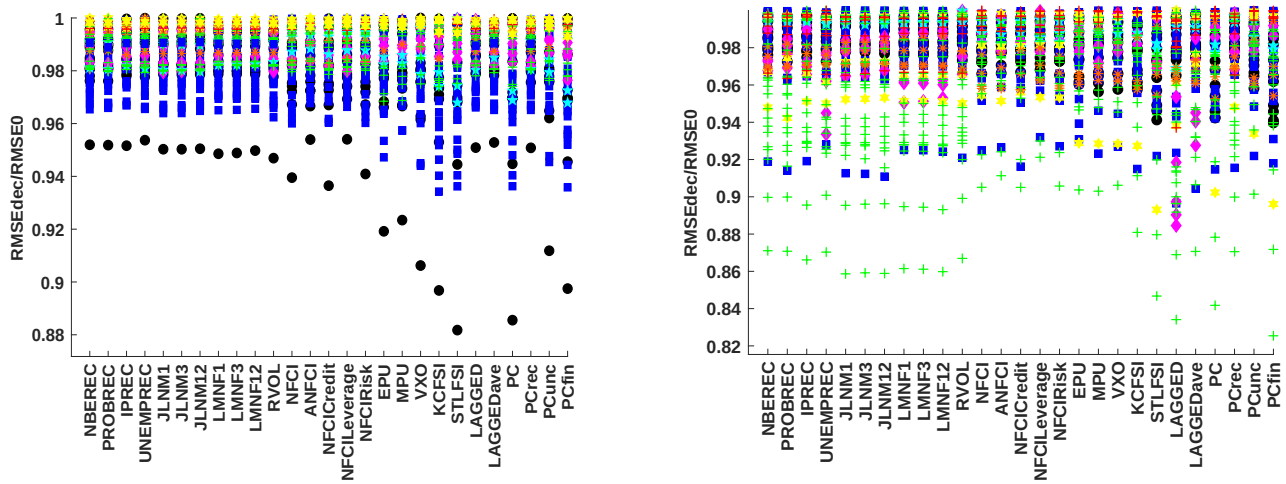


Notes: The figure shows the root mean squared forecast error (RMSE) of the model selection rule relative to an autoregressive benchmark (benchmark is marked with "RMSE0"). Values greater than one are not depicted on the figure since they would indicate that our model selection criteria deteriorates the accuracy of the forecasts relative to the benchmark. Results are grouped by conditioning variables, labeled consistently with the labels in Table 1. Benchmark RMSE for industrial production is 7.91 (one-step ahead) and 4.08 (twelve-steps-ahead), while for inflation it is 3.05 (one-step-ahead) and 2.51 (twelve-steps-ahead). Different colors and symbols refer to alternative models in which the additional variable represents: output and income (circles, black), labor market (squares, blue), housing market (diamonds, magenta), orders and inventories (pentagram, cyan), money and credit (hexagram, yellow), stock market (star, orange), interest and exchange rate (cross, green), prices (cross, red).

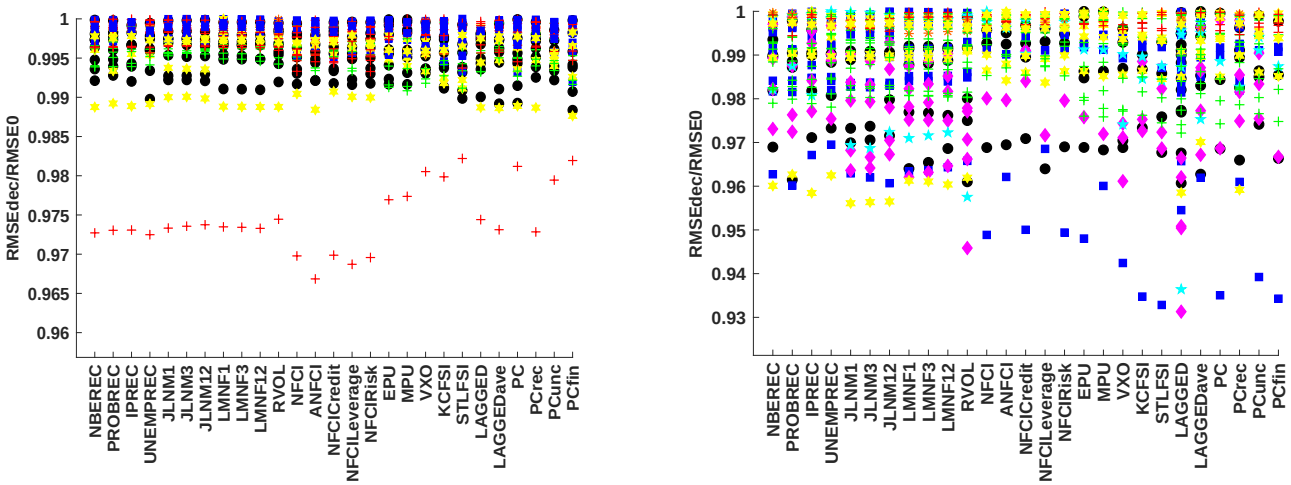
Figure 11. Decision Rule: Model Averaging Approach

$\tau = 1$ $\tau = 12$

Panel A. Industrial Production Growth



Panel B. Inflation



Notes: The figure shows the root mean squared forecast error (RMSE) of the model selection rule relative to an autoregressive benchmark (benchmark is marked with "RMSE0"). 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. Benchmark RMSE for industrial production is 7.91 (one-step ahead) and 4.08 (twelve-steps ahead), while for inflation is 3.05 (one-step ahead) and 2.51 (twelve-steps ahead). Different colors and symbols refers to alternative models in which the additional variable represents: output and income (circles, black), labor market (squares, blue), housing market (diamonds, magenta), orders and inventories (pentagram, cyan), money and credit (hexagram, yellow), stock market (star, orange), interest and exchange rate (cross, green), prices (cross, red).

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 months-ahead	JLN
LMNF1, LMNF3, LMNF12	1960M7	Financial Uncertainty Index, 1-3 and 12 periods ahead	LMN
		Measures of Past Relative Performance	
LAGGED	1970M3	previous month loss differential	O
LAGGEDave	1971M3	average loss differential over past 12 months	O
PC	1994M1	Principal component of all Indicators	O
PCrec	1967M6	Principal component of Business Cycle Indicators	O
PCunc	1986M1	Principal component of Uncertainty Indicators	O
PCfin	1994M1	Principal component of Financial Condition Indicators	O

Notes: Sources are abbreviated as follows: "F"- Federal Reserve Economic Data (FRED), "C"- CBOE, "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

$\tau = 1$			$\tau = 12$		
model	rel. RMSE	p-value	model	rel. RMSE	p-value
Panel A. Full Evaluation Sample (1970M3-2016M1)					
			Industrial Production		
benchmark RMSE	7.91		benchmark RMSE	4.08	
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		
benchmark RMSE	3.05		benchmark RMSE	2.51	
			Real M2 Money Stock	0.95	0.06
Panel B: Partial Evaluation Sample (1979M2-2016M1)					
			Industrial Production		
benchmark RMSE	7.56		benchmark RMSE	3.85	
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		
benchmark RMSE	3.14		benchmark RMSE	2.40	
Crude Oil, spliced WTI and Cushing	0.97	0.04			

Notes: The table shows the results of the unconditional test of 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	τ=1 relative perform	p-value	Model	τ=12 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: 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 / US 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 expected improvement of the alternative over the benchmark 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 more than half of the time in the out of sample. FFR stands for Federal Funds Rate.