Overcoming the Forecast Combination Puzzle: Lessons from the Time-Varying Efficiency of Phillips Curve Forecasts of U.S. Inflation*

Christopher G. Gibbs†

June 2, 2015

Abstract

This paper proposes a new dynamic forecast combination strategy for forecasting inflation. The procedure draws on explanations of why the forecast combination puzzle exists and the stylized fact that Phillips curve forecasts of inflation exhibit significant time-variation in forecast accuracy. The forecast combination puzzle is the empirical observation that a simple average of point forecasts is often the best forecasting strategy. The forecast combination puzzle exists because many dynamic weighting strategies tend to shift weights toward Phillips curve forecasts after they exhibit a significant period of relative forecast improvement, which is often when their forecast accuracy begins to deteriorate. The proposed strategy in this paper weights forecasts according to their expected performance rather than their past performance to anticipate these changes in forecast accuracy. The forward-looking approach is shown to robustly beat equal weights combined and benchmark univariate forecasts of inflation in real-time out-of-sample exercises on U.S. and New Zealand inflation data.

JEL Classifications: E17; E47; C53

Keywords: Forecast combination, inflation, forecast pooling, forecast combination puzzle, Phillips curve

*Previous versions of this paper circulated under the title, "Systematic Inflation Forecast Errors, Forecast Combination, and the Forecast Combination Puzzle."

†School of Economics, UNSW, Sydney. Email: christopher.gibbs@unsw.edu.au. I would like to thank George Evans, Jeremy Piger, Ken Wallis, James Morley, Benjamin Wong, participants of 32^{nd} Annual International Symposium of Forecasters, participants of 2013 Canadian Economic Associations Meetings, participants of the Continuing Education in Macroeconomics Workshop hosted by the University of Tasmania, and the research staff at Reserve Bank of New Zealand for useful comments on earlier drafts of this paper. I would also like to thank the Reserve Bank of New Zealand for allowing use of their real-time dataset. All remaining errors are my own.

1 Introduction

Medium horizon forecasts of inflation are key inputs into the decision making process of monetary policymakers. However, the construction of accurate inflation forecasts is fraught with model uncertainty. The uncertainty is best summarized by Stock and Watson (2007) who note that on the one hand inflation forecasting is relatively simple because standard univariate time series models or even a naïve random walk model can produce efficient - meaning a forecast that reliably minimizes a chosen loss function - forecasts of inflation. While on the other hand, it is exceedingly difficult to construct an efficient forecast that improves upon the naïve forecast, despite economic theory providing a host of alternative model specifications that should predict inflation well. Of course if all multivariate or theoretically motivated inflation forecasts failed to outperform univariate time series models, then forecasters could abandon these strategies all together. But, Stock and Watson (2008) show that this is not the case. Forecasts based on the Phillips curve relationship still provide significant improvements over univariate forecasts episodically. Therefore, inflation forecasters are faced with the choice of a relatively efficient forecast using simple univariate models or uncertain but potential improvements by employing a Phillips curve-type specification.

A common solution to mitigate this type of model uncertainty in economic forecasting is to pool individual forecasts together to create a single combined forecast. The combined forecast hedges against choosing the worst forecasting model in any given period and is typically found to improve overall forecast efficiency relative to the individual models being combined. The effectiveness of this strategy was first shown by Bates and Granger (1969) and has been confirmed by dozens of subsequent studies over the last forty plus years.[1] However, here too an inflation forecaster runs into another specification uncertainty problem. Empirically, the most reliable way to construct a combined forecast is to place equal weight on each considered model and take the mean of their point forecasts. The effectiveness of this simple strategy is found to be robust despite the fact that equal weights is only optimal under very restrictive assumptions about the correlation and covariance of the individual models' forecast errors and despite the fact there often exists significant past differences in forecast efficiency among forecasts that should be exploitable when choosing combination weights.[2] In fact, equal weights is found to be the most efficient forecast strategy so frequently that the result is commonly

[1] Surveys and comments on the literature are found in Clemen (1989), Granger (1989), Diebold and Lopez (1996), Timmermann (2006), and Wallis (2011).

[2] See Timmermann (2006) for a discussion of the assumptions under which equal weights is optimal.

referred to as the forecast combination puzzle.[3]

There are two explanations for why the puzzle exists. The first explains the failure of optimal weighting strategies in stationary environments. Optimal weighting strategies require the estimation of combination weights from the joint distribution of all the considered models' forecast errors. This in practice introduces estimation uncertainty into the forecasts because the number of models considered is typically large relative to the amount of available data. Smith and Wallis (2009) formalize this argument and show using Monte Carlo experiments that in general equal weights provide more efficient forecasts compared to estimated optimal weights because of estimation uncertainty. The literature, therefore, often recommends employing combination strategies that do not require estimation of the weights.

The second explanation is that in general economic forecasting models are often misspecified and economic data is subject to relatively frequent structural breaks. Hendry and Clements (2004) show that equal weights in this case is an effective strategy because it mitigates the different biases that arise in differently specified models. In particular, differently specified models may manifest bias in opposite directions following a change in the data generating process, which is averaged out in the combined forecast. Equal weights also prevent a forecaster from shifting the relative weight placed on different models following increases in forecast efficiency that often quickly reverse. The agnosticism of equal weights is, therefore, the key driver of its efficiency gains.

Combined forecasts of univariate and Phillips curve models of inflation provide an excellent example of the type of time-variation in forecast efficiency that generates the forecast combination puzzle. Stock and Watson (2008, 2010) provide evidence that the forecast efficiency of the Phillips curve is related to the state of the business cycle. In particular, the Phillips curve relationship appears to only have forecasting power relative to univariate models during times of economic contraction. Therefore, any backward-looking combination routines that uses recent past performance will place high weight on Phillips curve forecasts following recessions. The high weight, however, comes precisely as univariate forecasts start to provide more efficient forecasts of inflation during expansions.

The time-varying efficiency of the Phillips curve though also provides a clear example of how the forecast combination puzzle can be overcome. The time-varying efficiency of the Phillips curve is not random. As stated previously, it is related to the state of the business cycle and the current state of the business cycle is to a degree predictable in real

[3]An empirical example of the forecast combination puzzle for inflation is found in Stock and Watson (2003). The first formal reference to the forecast combination puzzle in the literature to my knowledge is Stock and Watson (2004), however, the results is certainly known in the literature at least dating back to Bates and Granger (1969).

time as demonstrated by Chauvet and Piger (2008) and Owyang et al. (2014). Therefore, periods when Phillips curve based forecasts are likely to experiences changes in forecast efficiency should be to a degree predictable i.e. as the economy moves into recession it should indicate a period when Phillips curve models perform well and vice versa when the economy begins to recover.

In this paper I propose a new forecast combination strategy to demonstrate that the time-varying efficiency of the Phillips curve is indeed identifiable and exploitable in real-time out-of-sample forecasting experiments. The strategy uses forward-looking weights to anticipate changes in forecast accuracy. The changes in forecast accuracy are anticipated by predicting the forecast error of a considered model using information on past forecast errors and real activity measures. The real activity measures provide information about the current state of the business cycle and hence information about the future performance of Phillips curve-type forecasts. The combined forecast is then constructed by weighting each individual model by its predicted forecast error relative to the predicted errors of all other models. The predicted weights strategy is shown to result in statistically significant reductions in mean squared forecast error (MSFE) and forecast bias relative to simple time series models, simple backward-looking combination strategies, and an equal weights combined forecasts across a number of real-time out-of-sample forecasting experiments.

1.1 Contribution and Related Works

The effectiveness of the proposed strategy illustrates two points to potentially overcome the forecast combination puzzle in a variety of settings. The first point is that time-varying weighting strategies should be forward-looking. Weights should adjust according to how models are expected to perform in the near future, rather than how they have performed in the recent past.

The second point is that other information, beyond past forecast performance, can often be leveraged to anticipate changes in forecast efficiency. The parsimonious models that are most effective in economic forecasting are often purposefully misspecified to avoid data overfitting. This misspecification means that there exists information that theory predicts should be useful for forecasting, but which in practice may not improve individual point forecasts if included in the original model specification. The extra information, however, may be useful for constructing forward-looking model combination weights.

Predicting forecast failure and exploiting information in forecast errors is of course not a novel idea. For example, Giacomini and Rossi (2009) develop a statistical test for

predicting when forecast accuracy will deteriorate. They show, as in this paper, that the deterioration of forecast accuracy of Phillips curve forecasts is predictable in real time. Wallis and Whitley (1991) and Clements and Hendry (1996) also study the use of forecast errors to improve forecast efficiency through a strategy known as intercept correction. Intercept correction uses the most recent observed forecast errors to correct the bias of a point forecast by adding the errors to the next forecast in order to set the model back on track. Wallis and Whitley (1991) finds that intercept correction produces modest improvements over an uncorrected model for forecasts of UK inflation as well as other macroeconomic variables.

The prediction of forecast errors is also closely related to the idea of inflation gap forecasting explored by Stock and Watson (2007, 2010) and Cogley et al. (2010). The inflation gap is defined as the deviation of inflation from a stochastic trend. An inflation gap forecast is constructed by forecasting the inflation gap and then adding it back to the last observation of the trend. Stock and Watson (2010) and Faust and Wright (2012) both find that inflation gap forecasting offers modest improvements over other parsimonious time series models in pseudo (single vintage of data) and real-time (multiple vintages of data) out-of-sample experiments, respectively. Therefore, the novelty of the combination strategy proposed in this paper is marrying forecast combination to the ideas of identifying changes in forecast efficiency, intercept correction, and forecasting the inflation gap.

The aim of this paper is to provide a proof-of-concept for forward-looking combination weights in a transparent way, rather than to obtain the absolute best possible inflation forecast. To achieve transparency I restrict the analysis to a set of simple model specifications to predict inflation and forecast errors. The success of the concept using simple models, however, suggests that large improvements in forecast efficiency may be possible if more sophisticated forecasting techniques are brought to bear.

I also attempt to address an external validity concern that exists when comparing forecast combination strategies on a fixed set of forecasting models. The individual forecasting specifications chosen by a researcher of course dictate the improvements in forecast accuracy that are possible and these choices often drive the reported results. An example of this is when a number of poor performing forecasts are considered among the set of included forecasts. A sophisticated forecast combination routine may easily detect the poor forecasts and result in a superior combined forecast than equal weights. However, the forecast combination puzzle would re-emerge if the set of models was trimmed to include only the best performing models. This concern is addressed by conducting a forecasting tournament that compares multiple forecast combination strategies on many

different sets of models to illustrate the relative forecast efficiency of the proposed strategy with respect to model choice.

Finally, I test the proposed strategy in real-time forecasting exercise of New Zealand inflation. The U.S. and New Zealand economies clearly differ on a number of dimensions and the confirmation of the forecasting strategies effectiveness on New Zealand data serves to demonstrate the general applicability of the proposed strategy.

The remainder of the paper proceeds as follows. The next section presents the data, forecast models, and the forecast combination strategy. Section 3 presents the results for U.S. data. Section 4 presents the results for New Zealand data. Section 5 concludes.

2 Data and Methods

2.1 Data

The main forecasting experiments are conducted on U.S. data from the Philadelphia Federal Reserve's Real-Time Macroeconomic data set.[4] The measures of inflation I consider are constructed using the Price Index for Personal Consumption Expenditure (PCE) and the GDP Deflator (PGDP). These measures are chosen because real-time data is available dating back to 1965Q4, which allows for the longest possible out-of-sample forecasting period. Quarterly inflation is defined as $\pi_t = ln(\frac{p_t}{p_{t-1}})$ and expressed as an annual rate.

The predictors employed to forecast inflation and to predict model performance are constructed from the real GDP and the civilian unemployment rate measures available in the real-time data set. The real GDP measure is used to create three predictors: 1) GDP growth, constructed as log differenced GDP; 2) output gap, constructed using the standard HP filter; 3) and a growth gap measure, which is constructed as difference of the current GDP growth rate from the maximum growth rate observed over the previous twelve quarters. The unemployment rate is used in levels and as a one sided unemployment gap measure. The unemployment gap measure follows Stock and Watson (2010) and is constructed as the difference in current quarter's unemployment rate from the previous twelve quarter's minimum rate. The growth and unemployment gaps provide one-sided measures of the business cycle to capture the nonlinearity of the Phillips Curve.

The real-time New Zealand data comes from a dataset provided by the Reserve Bank of New Zealand. The dataset does not include real-time measures of PCE or GDP Deflator so quarterly non-tradable CPI inflation is used instead.[5] Inflation is defined and

[4] A detailed description of the data set and an explanation of its usefulness for evaluating forecasting strategies is given by Croushore and Stark (2001).

[5] Since New Zealand is a small open economy, the measure of inflation that has the most similar

Univariate	Phillips Curve	Direct Forecasts
AR(1)	PC Output Gap	DF Output Gap
AR(2)	PC Unemployment Gap	DF Unemployment Gap
AR(4)	PC GDP Growth	DF GDP Growth
ARMA(1, 1)	PC Growth Gap	DF Growth Gap
ARMA(4, 4)	PC Unemployment Rate	DF Unemployment Rate
AO	VAR All	

Table 1: Forecast model specifications.

computed using the same definitions employed for the U.S. data and the same real GDP and unemployment measures are constructed.

2.2 Models

The list of considered models is given in Table 1. The included univariate models are chosen either because they are frequently used as benchmarks in the inflation forecasting literature or to provide variety in the specifications.[6] The AO forecast is based off the naïve model employed in Atkeson and Ohanian (2001) and is the average of the previous four quarters of inflation

$$\hat{\pi}_{t+h}^{AO} = \frac{1}{4}\sum_{i=1}^{4} \pi_{t-1-i}. \tag{1}$$

The AO model serves as the main benchmark for the forecasting experiments.[7]

The Phillips curve (PC) specifications are bi-variate VARs with two lags of inflation and two lags of a real activity measure. VARs are used as the PC-type forecasts to provide a degree of generality to the results. The VARs do not impose a specific theory on the structure of the Phillips curve but incorporate the basic observable information that is used in many different theoretically based PC forecasts. This of course includes nonlinear Phillips curves through the use of one-sided unemployment and output gap measures. The lag length for the VARs is selected for parsimony and held constant throughout the exercise.

The specifications labeled as direct forecasts (DF) are OLS regressions of a given real

relationship to the output gap as PCE and GDP Deflator for U.S. data is non-tradeable CPI inflation.

[6]For example the ARMA(1,1) is the benchmark forecast employed by Ang et al. (2007), who compared dozens of different forecast specifications covering surveys, ARMA models, regressions using real-activity measures, and term structure models or the AR(4), which is the benchmark in Stock and Watson (2010).

[7]The AO forecast performs comparably to the inflation gap forecasting strategy proposed by Stock and Watson (2007) as shown by Faust and Wright (2012).

activity measure on h-quarter ahead inflation

$$\pi_{t+h} = c + \beta x_t + \epsilon_t. \tag{2}$$

where x_t represents a real activity measure and ϵ_t is the error term.[8] The specification is included primarily as a robustness check. The prediction of an individual model's performance is constructed using the same direct forecast specification. The concern is that the prediction step of the combination strategy is picking up a correctable form of model misspecification. The underlying assumption of the proposed combination strategy is that there is model misspecification that can be exploited through the prediction of a model's performance, but which cannot be exploited in the actual forecast model's specification.

The VAR All model is also included for robustness. The VAR All model includes all the information that is found to predict forecast performance well into a single specification (GDP growth, output gap, and the unemployment gap). If the information used to predict a future forecast model's performance is more useful for predicting the level of inflation, then this model should provide relatively efficient forecasts. The lag length of this specification is also two and fixed for all exercises.

2.3 Forecasts and Inference

The forecast of interest in this paper is the four quarter ahead forecast of quarterly inflation expressed as an annual rate. The real-time forecasts are constructed using the latest vintage of data available at each point in time. Due to lags in the release of data to the public, however, the current quarter's observation of inflation is not available at the time a forecast is made. Therefore, the forecast considered is actually the nowcast and the subsequent three quarters. The forecasts are denoted as $E_t^T \pi_{t+4}$, where t is the last observation of data, T is the vintage of data and the time at which the forecast is made (i.e. $T = t + 1$).

The forecasts are evaluated based on root MSFE to measure accuracy and mean forecast error to measure forecast bias. Inference on the observed differences in MSFE are obtained using the Diebold and Mariano (1995) (DM) test for equal within sample forecast accuracy with the Harvey et al. (1997) small sample size and long horizon correction. There is not much guidance in the literature on the correct test statistic to evaluate combined forecasts. This is especially true in the context considered here where it is assumed that the data suffers from frequent structural breaks. Most test statistics are

[8] For an explanation of the merits of direct forecasting see Marcellino et al. (2006).

based on the assumption of asymptotic convergence to stationary distributions for the estimated regressors of the model considered and their forecast errors.[9] None of which can be argued to hold in this case and which of course is part of the reason for considering model combination in the first place. However, the use of Diebold and Mariano test statistic follows recommendations given by Clark and McCracken (2011) for evaluating forecasts on real-time data and Diebold (2014) who notes that the only assumption that must be satisfied in order to use the DM test statistic is that the differences in squared forecast errors are covariance stationary. Inference on the bias results is obtained using a t-test with Newey and West (1987) standard errors.

The target measure of inflation to which the real-time forecasts are compared is a composite series constructed out of the second release quarterly observations of inflation as they appear in the real-time data set. The use of second release data minimizes the influences of large renormalization that occur in the sample due to definitional changes and provides a final measure of inflation that is closer to the actual measure a forecaster would have been attempting to forecast at any given point in time.

2.4 Predicting Forecast Model Performance

The prediction of a forecast model's performance is obtained via a direct forecast of the model's real-time forecast errors. The direct forecast regresses a real activity measure on the four quarter ahead forecast error

$$fe_{i,t+4} = c + \beta x_t + \epsilon_t, \tag{3}$$

where $fe_{i,t} = \pi_t - E_{i,t-4}\pi_t$ for the i^{th} considered model.[10] The forecast error series is constructed using real-time errors obtained from comparing past real-time forecasts to the composite series of second release information. The last forecast error in each period though is compared to first release information as the second release is not available at the time the forecast is made. Note that by construction this procedure introduces new information into the forecasts because the individual forecasts models are estimated on the most recent vintage of data available at the time the forecast is made, while the series of forecast errors contains information from multiple vintages. Therefore, the forecast error series incorporates some information about revision into the final combined

[9] See West (2006) for a review of the literature.

[10] This specification is the same as the specification considered by Stock and Watson (2010) to forecast the inflation gap.

forecast.[11]

In addition to real activity measures, I also consider two specifications that rely solely on past forecast errors for robustness. The first specification uses a rolling average of the last four forecast error observations to predict future errors similar to the AO forecast for inflation. The specification is also similar to simple backward-looking combination strategies that are frequently considered in the literature and partially accounts for the fact that by construction a series of forecast errors follows an MA(h-1) process, where h is the forecast horizon.[12] This specification is denoted AOPW. The second specification attempts to capture trends in the forecasts errors. This specification employs a three year rolling window regression of a constant and a time trend on past forecasts errors.

2.5 Combining Forecasts and Out-of-sample Experiments

The combined forecasts are constructed as a weighted average of point forecasts

$$\hat{\pi}_{t+4} = \sum_{i=1}^{N} \omega_{i,t} E_t \pi_{i,t+4}, \tag{4}$$

where $\omega_{i,t}$ is the weight given to the i^{th} model. The weights are constructed as

$$\omega_{i,t} = \frac{e^{-\beta(E_t fe_{i,t+4})^2}}{Z_t}, \; Z_t = \sum_{i=1}^{17} e^{-\beta(E_t fe_{i,t+4})^2}$$

where β is the shrinkage or intensity of choice parameter.[13] The β parameter governs the relative weight given to each model based on expected differences in squared errors. If β is close to zero, then weights shrink towards equal weights. If β is large, then almost all weight is placed on the best predicted model.

The out-of-sample forecasting exercise requires the data to be separated into three subsets. The required divisions are a training subset to estimate the initial parameters of the candidate forecast models, an in-sample forecasting period to recursively forecast the candidate models to construct an initial series of forecast errors to estimate Equation (3), and an out-of-sample period to conduct out-of-sample forecasts. The three periods are 1947Q2-1965Q4, 1966Q1-1969Q4, and 1970Q1-2014Q1, respectively.

[11] I tested both revised and unrevised forecast error series and the revised series do appear to add a small amount of forecasting power relative to the unrevised.

[12] Predicting forecast errors using an MA(3) specification does not improve upon any of the specifications considered.

[13] The functional form of the weights is a multinomial logit, which is used extensively in the discrete choice econometric literature.

The out-of-sample forecasts are made recursively using the following procedure at each time period T:

1. Each candidate forecast is estimated on vintage T data.
2. Each candidate forecast is used to construct a four quarter ahead forecast.
3. Equation (3) is estimated on the available real-time forecast errors for each of the candidate forecasts.
4. Equation (3) is used to predict the expected forecasts error of each model to construct $E_t fe_{i,t+4}$.
5. The predicted errors are used to construct weights and the weight are used to construct the combined forecast.

3 U.S. Results

The results are separated into five sections. The first section provides the individual forecasting performance of the 17 considered models to establish baselines to compare to the combined results. The second section presents out-of-sample forecasting results using the predicted weights combination strategy. The third section illustrates the source of gains in forecast efficiency for forward-looking weights using an explicit example. The fourth section conducts a tournament that compares the proposed combination strategy to backward-looking combination strategies to illustrate robustness. And finally, the fifth section presents the intercept correction results.

3.1 Individual Model Results

Table 2 reports the real-time out-of-sample forecasting results for the individual model of PCE inflation. Results are reported for the full out-of-sample period (1970Q1-2014Q1) and two sub-sample periods: 1983Q1-2007Q3, which roughly covers the Great Moderation; and the most recent period 2007Q4-2014Q1, which covers the Financial Crisis and the recovery. I only present the results for PCE inflation in this section because I find little difference in the forecast outcomes between the PCE and the PGDP measures of inflation. PGDP results are, however, presented in the forecasting tournament section. All RMSFE results are reported relative to the AO forecast.

The AO forecast clearly dominates all other considered forecasts specifications. The AO forecast results in the lowest RMSFE over the full sample and both subsamples with

Individual Model Results

Model	RMSFE	Bias	RMSFE	Bias	RMSFE	Bias
AO	2.353	0.09†	1.483	0.20†	2.282	0.21†
DF CUR	1.287	0.24†	1.432	1.42	1.349	2.28
DF GDP	1.267	-0.27†	1.189	1.10	1.062	1.51
DF Growth Gap	1.258	-0.08†	1.262	1.27	1.109	1.67
DF Output Gap	1.368	-0.22†	1.461	1.30	1.095	1.45
DF U. Gap	1.263	-0.09†	1.246	1.22	1.153	1.79
VAR ALL	1.073	-0.31†	1.198	0.54	1.090	0.55†
PC CUR	1.064	0.14†	1.169	0.81	1.190	1.33
PC GDP growth	1.041	-0.37†	0.985	0.54	1.062	0.65†
PC Growth Gap	1.002	-0.11†	1.018	0.53	1.082	0.77
PC Output Gap	1.087	-0.31†	1.086	0.71	1.099	0.74
PC U. Gap	1.026	-0.04†	1.055	0.62	1.159	1.10
AR(1)	1.114	-0.22†	1.066	0.84	1.078	1.05
AR(2)	1.021	-0.15†	1.017	0.64	1.073	0.85
AR(4)	1.072	-0.30†	1.019	0.63	1.067	0.82
ARMA(1, 1)	1.003	-0.15†	0.991	0.57	1.057	0.78
ARMA(4, 4)	1.085	-0.22†	1.122	0.70	1.042	0.71
Dates	1970Q1-2014Q1		1983Q1-2007Q3		2007Q4-2014Q1	
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$						

Table 2: This table reports the individual forecast model results. The RMSFE of the AO forecast is reported in the first row. All remaining results are reported relative to the AO forecast. A number less than one represents an improvement in forecast accuracy. Significance for the RMSFE results is only indicated for improvements over the benchmark. The † indicates unbiased forecasts significant at the 10% level.

only two exceptions. The PC GDP growth forecasts and the ARMA(1,1) forecast both result in lower RMSFE during the subsample covering the Great Moderation. However, neither of the improvements are statistically different from the AO forecast. The AO forecast also has the lowest bias of any of the forecast models considered.

The best performing PC specification in terms of RMSFE is the specification that utilizes the one-sided growth gap measure. The growth gap specification is statistically no different from AO forecast over the full sample and the Great Moderation subsample. The worst performing PC specification in terms of RMSFE is the output gap specification. It is statistically significantly worse than the AO forecast on the full sample and on the two subsamples.

One perhaps surprising result of this exercise is that all models provide unbiased forecasts on the full sample. Although this result is largely driven by the fact that all models, except the AO forecast, have a negative bias in the period leading up to the

Predicted Weights Combined Results

Predictor	Rel. RMSFE	Bias	Rel. RMSFE	Bias	Rel. RMSFE	Bias
Equal Weights	1.054	-0.14†	1.056	0.80	1.046	1.07
Output Gap	0.922***	-0.19†	0.949	0.23†	0.998	0.74
U. Gap	0.979	0.22†	1.028	0.53	1.044	1.06
GDP Growth	0.931***	0.15†	0.974	0.43	1.058	1.11
Growth Gap	1.009	0.26†	1.007	0.46	1.076	1.21
CUR	1.003	0.29†	1.028	0.48	1.047	1.06
Trend	1.155	0.02†	1.060	0.62	1.042	0.88
AOPW	1.075	0.20†	1.054	0.56	1.014	0.76
Dates	1970Q1-2014Q1		1983Q1-2007Q3		2007Q4-2014Q1	

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 3: Forecasts are constructed with $\beta = 5$. The RMSFE are shown relative to the AO forecast. Significance for the RMSFE results is only indicated for improvements over the benchmark. The † indicates unbiased forecasts significant at the 10% level.

Great Moderation and a positive bias thereafter.

3.2 Forward-Looking Combined Forecasts

Table 3 reports the results for the predicted weights combined forecasts incorporating all seventeen models. The RMSFE results are normalized relative to the AO forecast. The *Predictor* column denotes the real activity measure used to predict forecast errors to construct the weights. The shrinkage parameter for all exercises is set to $\beta = 5$. The β parameter is chosen by searching over whole number values of β to minimize the MSFE in the pre-sample 1967Q1-1969Q4.

The largest improvements in forecast efficiency relative to equal weights and the AO forecasts are obtained on the full sample. In particular, the output gap and real GDP growth predictors generate statistically significant reduction in forecast accuracy relative to both benchmarks. The combined forecasts constructed using only the past information available in the real-time forecast errors (Trend and AOPW), however, do not show any notable improvements in forecast accuracy. These forecasts are about as efficient as the equal weights forecasts across the different samples. The lack of any significant or qualitative improvements in RMSFE in these two cases illustrates that the correlation of the forecast errors with real activity is the relationship exploited to improve forecast efficiency.

The absolute performance of the predicted weight forecasts is attenuated in the two subsamples. Although, the combined forecasts do retain a statically significant advantage

Best Predicted Model in Each Period

Predictor	Rel. RMSFE	Bias	Rel. RMSFE	Bias	Rel. RMSFE	Bias
Equal Weights	1.054	-0.14†	1.056	0.80	1.046	1.07
Output Gap	0.941**	0.28†	1.002	0.22†	0.987	0.52†
U. Gap	1.011	0.33†	1.117	0.57	1.124	1.28
GDP Growth	0.962	0.09†	0.973	0.26†	1.167	1.04
Growth Gap	1.029	0.14†	1.005	0.25†	1.220	1.68
CUR	1.019	0.26†	1.070	0.41†	1.093	1.09
Trend	1.201	-0.02†	1.123	0.59	1.080	0.82
AORW	1.106	0.14†	1.228	0.47	1.156	0.44†
Dates	1970Q1-2014Q1		1983Q1-2007Q3		2007Q4-2014Q1	
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$						

Table 4: Forecasts are constructed with $\beta \to \infty$. The RMSFE are shown relative to the AO forecast. Significance for the RMSFE results is only indicated for improvements over the benchmark. The † indicates unbiased forecasts significant at the 10% level.

over equal weights in most cases and a qualitative advantage over the AO forecasts for at least one predictors in each subsample. The predicted weights forecasts also exhibit a reduction in bias in almost every case compared to the underlying individual models and a comparable or improved bias relative to equal weights across all samples.

Table 4 reports the forecasting results when the β parameter is sent to infinity. This case is equivalent to placing a weight of one on the best predicted forecast in each time period, which removes the hedging component of combining forecasts. The case is of particular interest because, as noted in Timmermann (2006), choosing a single model in every period typically results in very poor out-of-sample forecasting efficiency. Therefore, the results are quite surprising. Choosing the expected best model in each period actually leads to reductions in relative RMSFE compared to equal weights in a majority of out-of-sample forecast experiments and even results in lower RMSFE compared to the AO benchmark for some predictors in every sample period considered. It also leads in most cases to a reduction in bias compared to the combined forecasts. The results illustrates that the gains in forecast efficiency observed in Table 3 are partly due to correct predictions of the actual best performing forecast model in each period.

The best overall predictor for the weights observed in these exercises is the output gap measure. This is somewhat surprising because Orphanides and Van Norden (2005) show that the HP filtered output gap has very little predictive power over inflation in real-time and because the output gap is the worst predictor of inflation on average among the PC and DF forecasts reported in Table 2. One explanation for the finding is that the HP filter provides an estimate of the output gap that only captures large business cycle

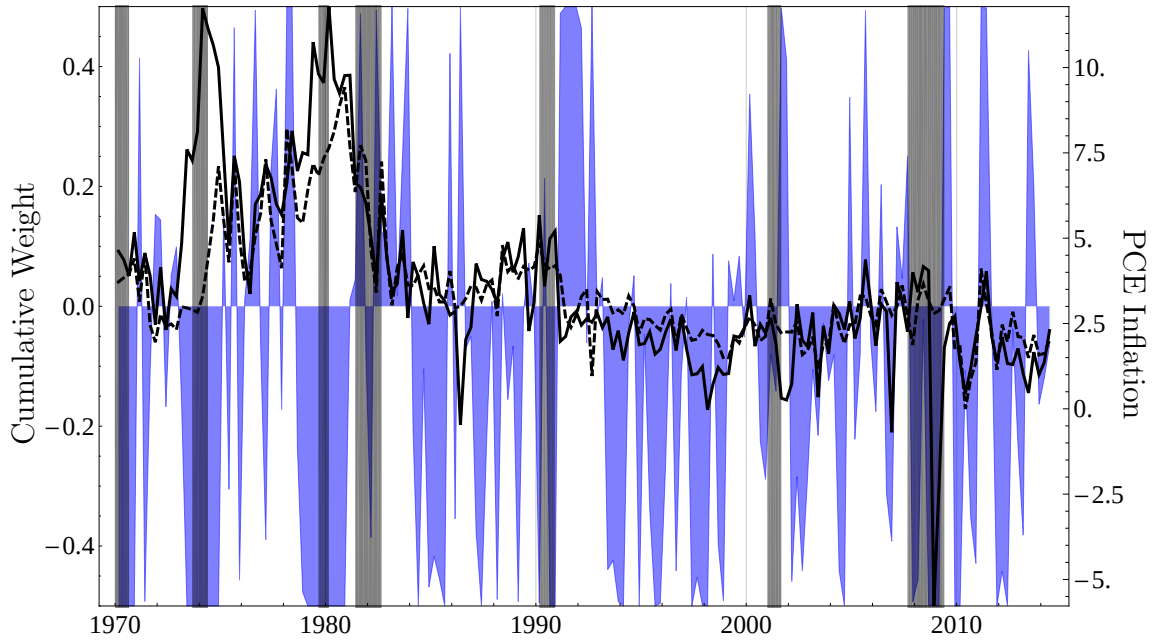

Figure 1: The figure depicts the PCE measure of U.S. inflation from 1970Q1 to 2014Q1 (solid), the ex post best possible forward-looking combined forecast (dashed), and the cumulative weight placed on the Phillips curve forecasts relative to equal weights constructed using ex post forecast errors (shaded blue). The dark bars indicate the NBER recession dates.

fluctuation, which is when a change in relative efficiency between the Phillips curve and univariate models is most likely.

3.3 An Illustration of Forward-Looking Weights and the Forecast Combination Puzzle

This section provides an explicit example of the forecast combination puzzle and how forward-looking weights can improve forecast efficiency. For the example, I take the six univariate and six PC forecast specifications given in the first two columns of Table 1 and compare backward-looking and forward-looking weighting strategies to the ex post best weights a forecaster could obtain using the forecasting strategy proposed in this paper.

Figure 1 illustrates the ex post best weights and their implied combined forecast for PCE inflation. The weights are constructed using the actual ex post observed squared forecast errors of each model. In particular, the weights are constructed as

$$\omega_{i,t} = \frac{e^{-\beta fe^2_{i,t+4}}}{Z_t}, \; Z_t = \sum_{i}^{12} e^{-\beta fe^2_{i,t+4}} \tag{5}$$

where $fe_{i,t+4}$ is the actual forecast error and $\beta = 5$. The weights represent those that

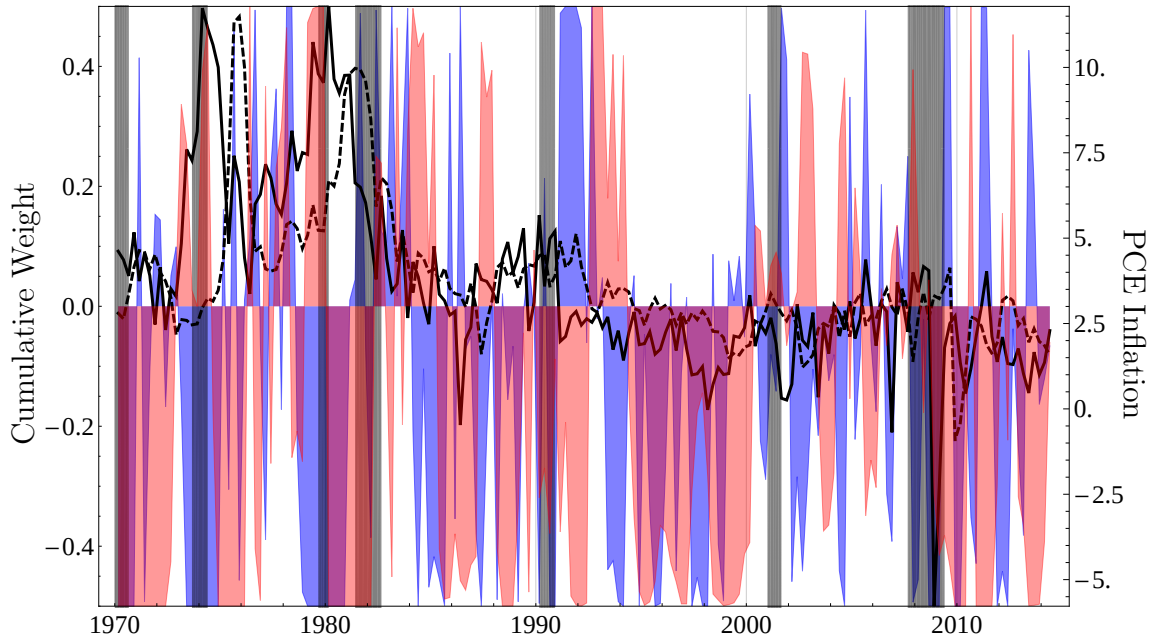

Figure 2: The figure depicts the PCE measure of U.S. inflation from 1970Q1 to 2014Q1 (solid), the backward-looking weights combined forecast (dashed), the cumulative benchmark weights relative to equal weights (shaded blue), and the cumulative backward-looking weights relative to equal weights (shaded red). The cumulative weight placed on the Phillips curve forecasts from the backward-looking weighting strategy are constructed using the four-quarter rolling average of past MSFE. The dark bars indicate the NBER recession dates.

would be obtained if it were possible to perfectly forecast the error of each model in real time. The benchmark weights produce an unbiased combined forecast that is a 26% improvement over both the AO and equal weights combined forecast in RMSFE.[14] The cumulative weights illustrated in the graph are constructed by summing the weights placed on the PC forecasts in each quarter and subtracting it from 0.5 ($\sum_{PC} \omega_{i,t} - 0.5$). The graph, therefore, provides an approximate description of how the cumulative weight placed on PC forecasts shifts relative to equal weights over time. Points above zero indicate that greater than half of all weight is on the PC forecast specifications. Points below zero represent that greater than half of all weight is on the univariate forecast specifications.

Figure 2 illustrates the weights and combined forecasts that results from a simple backward-looking weighting strategy similar to one proposed by Bates and Granger (1969). Each model is weighted relative to the four-quarter rolling average of past observed forecast errors. Specifically, the weights are constructed using Equation (5), where

[14]The AO and equal weights combined forecasts have a relative RMSFE of 1.0004. The benchmark combined weights also result in a slight increase in RMSFE compared to actually forecasting with the ex post best model in each period.

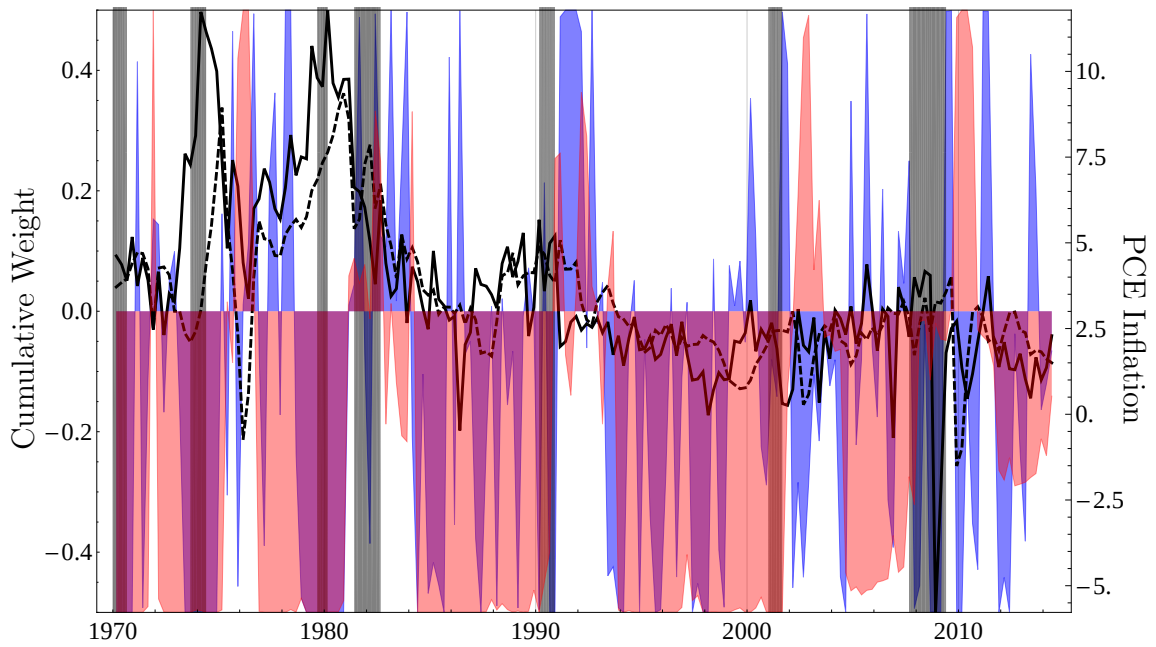

Figure 3: The figure shows the PCE measure of U.S. inflation from 1970Q1 to 2014Q1 (solid), the forward-looking predicted weights combined forecast (dashed), the cumulative benchmark weights relative to equal weights (shaded blue), and the cumulative predicted weights relative to equal weights (shaded red). The cumulative predicted weights are constructed using Equation (3) with the output gap real activity measure. The dark bars indicate the NBER recession dates.

$fe_{i,t+4}$ is replaced with $\frac{1}{4}\sum_{i=1}^{4} fe_{i,t-i}^2$. This strategy results in a modest but statistically significant 5% loss in relative RMSFE compared to an equal weights combined forecast and the AO forecast.

The reason this strategy fails to improve upon equal weights is clearly visible in the figure. The backward-looking weights are negatively correlated with the benchmark weights (corr = -0.1382). The strategy shifts weight to the PC forecasts after periods where the PC forecasts performs well, which is of course precisely when the strategy is about to lose forecast efficiency relative to the univariate forecasts.

Figures 1 and 2, however, illustrate the relationship between Phillips curve forecast efficiency and economic downturns. In particular, the benchmark weights consistently shift towards the Phillips curve forecasts in the periods surrounding the NBER recession dates. A forward-looking strategy can take advantage of this regularity by shifting weight towards PC forecasts when real activity is weak and by shifting weights towards the univariate models when real activity is strong.

Figure 3 demonstrates the forward-looking strategy. The expected error of each considered model is predicted in real-time using Equation (3) with the output gap real activity measure. The figure shows that predicted weights often deviate far from the benchmark but are positively correlated with it over time (corr = 0.1780). The posi-

tive correlation translates into a statistically significant 7% improvement in RMSFE over both equal weights and the AO forecasts for the set of considered models.

3.4 Forecast Tournament

Comparisons of combined forecast techniques face an external validity problem because the results are sensitive to the set of forecast models considered. This concern is especially relevant when comparing a combination strategy to an equal weights forecast. Equal weights of course has no mechanism to filter out obviously poor forecasts. Therefore, it is easy to construct a straw man equal weights forecast by considering poor performing forecasts that a sophisticated model combination strategy can easily detect and which an actual forecasters would not consider. To overcome this issue and provide a relatively fair comparison of equal weights to the proposed predicted weights strategy, I conduct a forecast tournament that varies the set of combined models. The tournament is conducted by selecting the twelve most efficient models from the 2007Q4-2014Q1 subsample found in Table 2 and then considering every combination of the twelve distinct forecast taken n at time, where $n = 2, 3, ..., 12$. This provides 4,083 different sets of models to combine.[15]

The tournament compares four different combination strategies: 1) equal weights, 2) AOPW weights, 3) weights based of the past observed MSFE of each model, and 4) predicted weights that use the output gap as the predictor.[16] The MSFE weights follow Stock and Watson (2004). The weight are constructed using relative past mean squared forecast error

$$\omega_{i,t} = \frac{(1/MSFE_{i,t})^k}{\sum_{i=1}^{n}(1/MSFE_{i,t})^k}, \tag{6}$$

where $MSFE_{i,t} = (1/m)\sum_{\tau=t-m}^{t} fe_{i,\tau-4}^2$, m is the sample size, and k is a shrinkage parameter.[17] The MSFE weights are one of the simple combination procedures that, like equal weights, consistently improves upon more sophisticated weighting procedures.[18]

Figure 4 gives a summary of the results for the real-time out-of-sample forecasting exercises conducted on the 1970Q1-2014Q1 PCE and PGDP measures of inflation. The figure shows the median, minimum, and maximum RMSFE observed for each subset

[15] The number of distinct combination of the twelve models for each n is as follows: $n = 2 \rightarrow 66$ sets, $n = 3 \rightarrow 220$ sets, $n = 4 \rightarrow 495$ sets, $n = 5 \rightarrow 792$ sets, $n = 6 \rightarrow 924$ sets, $n = 7 \rightarrow 792$ sets, $n = 8 \rightarrow 495$ sets, $n = 9 \rightarrow 220$ sets, $n = 10 \rightarrow 66$ sets, $n = 11 \rightarrow 21$ sets, $n = 12 \rightarrow 1$ set.

[16] The predicted weights assume $\beta = 5$.

[17] For the forecast tournament I choose $k = 5$.

[18] I also considered the optimal weights implied by regressing all past forecasts on the actual realization of inflation proposed by Granger and Ramanathan (1984). However, the weights perform so poorly compared to the other four methods considered that it did not provide a useful comparison.

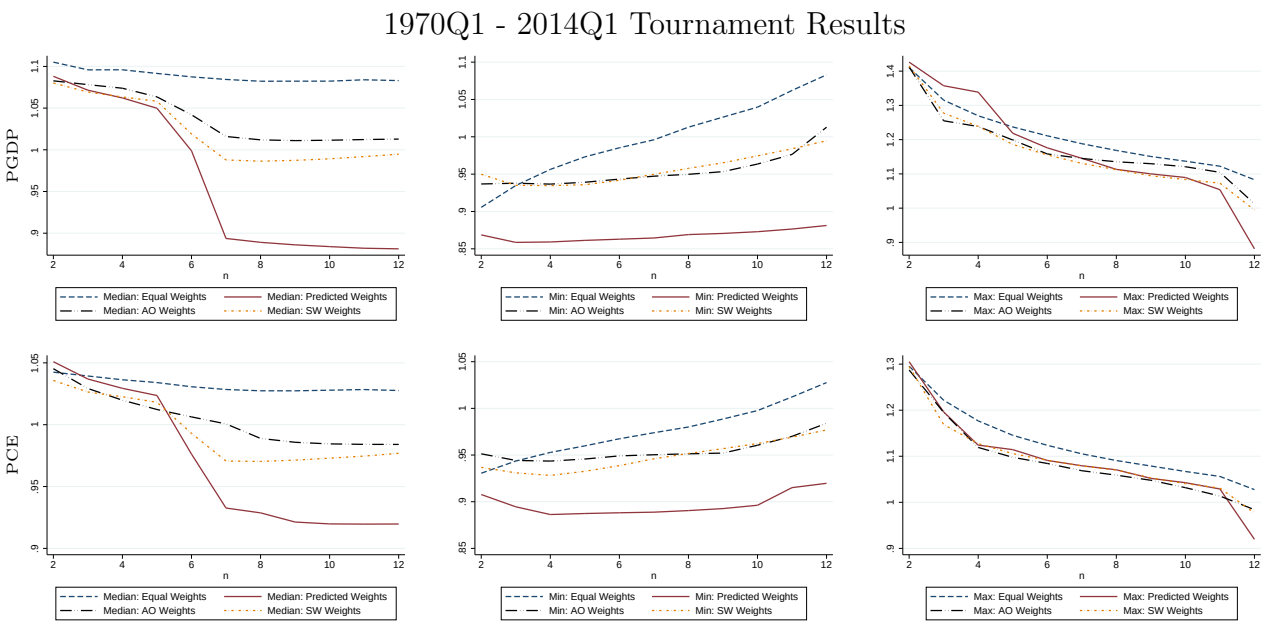

Figure 4: Median, minimum, and maximum relative RMSFE results for combination of n different models. The results are relative to the RMSFE of the AO forecast.

of models of size n relative to the AO forecast RMSFE. The maximum RMSFE plot shows the worst case scenario for each of the combination methods for combining n different forecasts, the minimum RMSFE plot shows the best case scenario for combining n different forecasts, and the median RMSFE provides a measure of the distribution of RMSFE observed for combining n different forecasts.

The maximum RMSFE results shows that each combination method exhibits roughly equal risk over the full sample. The maximum RMSFE or worst forecasting outcomes of all four strategies are comparable to each other across all sets of size n. Efficiency though is increasing in the number of forecast considered in all cases.

The minimum RMSFE results show a clear advantage for the predicted weights strategy. The predicted weights strategy consistently results in the lowest observed RMSFE among the four different forecast combination strategies. The PGDP results are especially impressive with consistent improvements nearing 15% relative to the AO forecast for the minimum RMSFE observations for all n.

The median RMSFE results in Figure 4 also show an advantage for the predicted weights forecasts. The median improvements in efficiency are as high as 10% relative to the AO forecast for combinations of $n > 5$ models. The increase in efficiency at $n > 5$ also provides some evidence of the exploitable time-varying trade-off between PC and univariate forecasts. For $n > 5$, almost all experiments include at least one PC and one univariate forecast model.

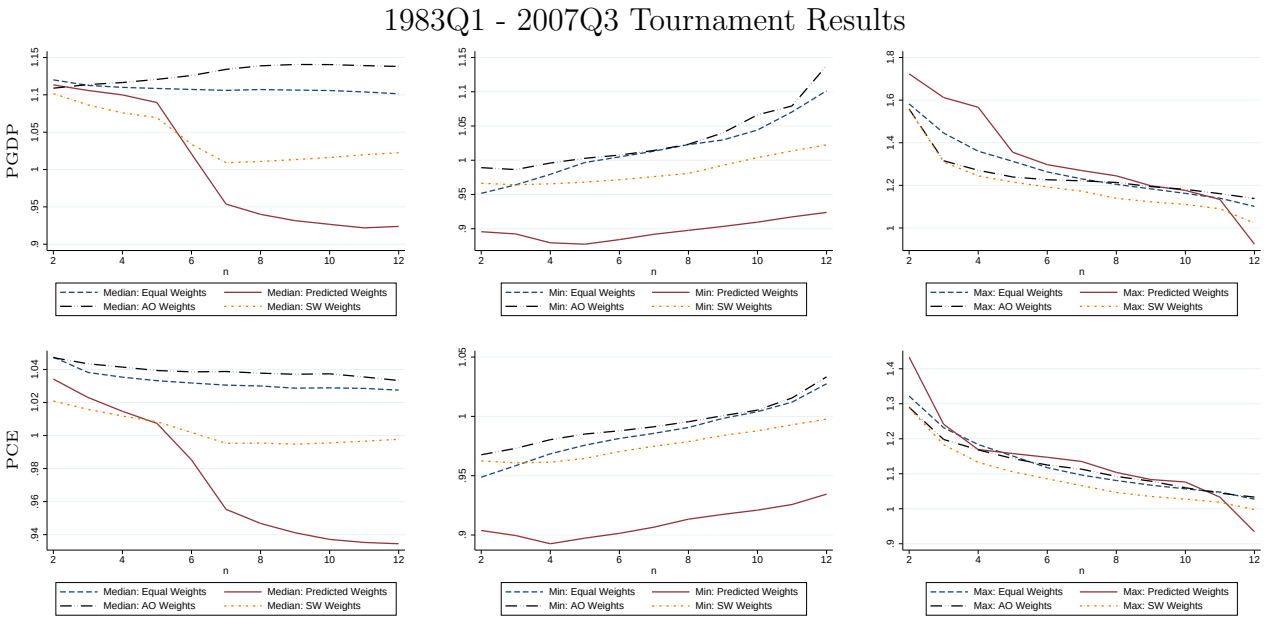

Figure 5: Median, minimum, and maximum relative RMSFE results for combination of n different models. The results are relative to the RMSFE of the AO forecast.

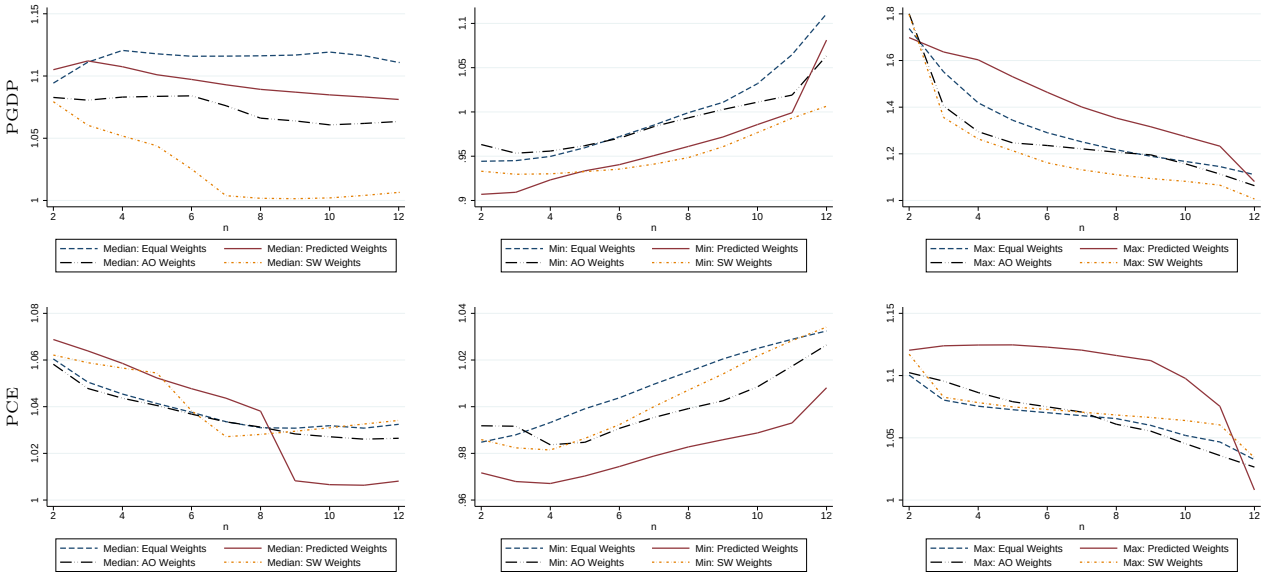

Figure 6: Median, minimum, and maximum relative RMSFE results for combination of n different models. The results are relative to the RMSFE of the AO forecast.

Figure 5 and 6 show the results for the two subsamples. The 1983Q1-2007Q4 subsample is consistent with the full sample results. The 2007Q4-2014Q1 subsample results, however, are attenuated compared to the RMSFEs obtained on the other samples. The attenuation is most pronounced for the GDP Deflator measure of inflation. Here there are no consistent improvements over equal weights. Though subsequent exploration revealed that increasing the β parameter can significantly improve performance in this sample.

The results shown in Figures 4, 5, and 6 can roughly be replicated using either the GDP growth or the unemployment gap predictors to construct the predicted weights. The GDP growth measure in particular produces forecasts that are comparable to the output gap forecasts across all considered samples. The results for the one-sided growth gap and unemployment rate measure, however, are less impressive. These predictors perform well compared to the equal weights combined forecasts, but often fail to outperform the AOPW and SW Weights.

3.5 Intercept Correction

Table 5 presents the intercept correction results. The intercept correction results use the output gap predictions of the forecast errors to correct the points forecast of each model such that

$$E_t\pi^{IC}_{i,t+4} = E_t\pi_{i,t+4} + E_t fe_{i,t+4}. \tag{7}$$

The table shows that the point forecasts of the forecast errors are not very accurate. The intercept corrected forecasts are less efficient than the uncorrected forecasts in almost all cases. The one exception is the AO forecast. Although, the reported improvement in forecast efficiency is not statistically different from the uncorrected AO forecast. The results are similar for any of the five real activity measures considered to predict forecast errors.

The explanation for the disparity in effectiveness between predicted weights forecasts and the intercept corrected forecasts is that the two strategies use the predictions of the forecast errors in different ways. Predicted weights exploits the relative ranking of the forecasts implied by the predicted forecast errors, while intercept correction relies on the accuracy of the actual point forecast of the error. The differences suggest that the predictions of forecast errors contain information about relative performance, but little information about the absolute performance of the considered models.

Intercept Correction Results

Model	RMSFE	Bias	RMSFE	Bias	RMSFE	Bias
AO	0.937	0.32†	1.262	0.49†	1.130	-0.09†
DF CUR	1.562	1.81	2.703	3.06	1.493	2.17
DF GDP	1.513	1.62	2.553	3.20	1.258	1.89
DF Growth Gap	1.512	1.74	2.611	3.25	1.271	1.88
DF Output Gap	1.667	1.61	2.772	3.36	1.278	1.87
DF U. Gap	1.509	1.70	2.579	3.20	1.294	2.03
VAR ALL	1.296	1.09	2.073	1.99	1.229	0.91†
PC CUR	1.232	1.27	1.988	1.87	1.414	1.21†
PC GDP	1.197	1.13	1.895	2.11	1.217	1.03
PC Growth Gap	1.137	1.10	1.771	1.70	1.243	0.87†
PC Output Gap	1.309	1.17	2.036	2.27	1.231	1.13
PC U. Gap	1.179	1.11	1.843	1.82	1.352	1.14†
AR(1)	1.297	1.24	2.141	2.48	1.227	1.28
AR(2)	1.139	1.05	1.806	1.93	1.196	1.00
AR(4)	1.208	1.22	1.917	2.20	1.217	1.15
ARMA(1, 1)	1.100	1.00	1.708	1.79	1.177	0.91†
ARMA(4, 4)	1.252	1.22	1.900	2.09	1.171	0.91†
Dates	1970Q1-2014Q1		1983Q1-2007Q3		2007Q4-2014Q1	
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$						

Table 5: Intercept correction results. The RMSFE results are presented relative to the uncorrected AO forecasts. Significance for the RMSFE results is only indicated for improvements over the benchmark. The † indicates unbiased forecasts significant at the 10% level.

4 New Zealand Inflation

The justification of the proposed strategy in this paper is motivated completely by the stylized facts of U.S. inflation. Therefore, to determine whether the insights from those observed U.S. relationships are more broadly applicable, this section applies forward-looking weights to real-time data for New Zealand.

Table 6 reports the baseline results for the seventeen models used in the previous forecasting experiments. The out-of-sample forecast period for New Zealand is 1997Q1-2014Q1. The table shows the results are very similar to those obtained on U.S. data. Predicted weights results in significant increases in forecast efficiency relative to the AO forecast and an equal weights forecast.

Figure 7 shows the results for the forecast tournament on New Zealand real-time data. The results here as well are nearly identical to those observed for U.S. data.

New Zealand Results

	$\beta = 5$			$\beta \to \infty$		
Predictor	RMSFE	Rel. RMSFE	Bias	RMSFE	Rel. RMSFE	Bias
Equal Weights	1.649	0.971	0.55	-	-	-
Ouput Gap	1.714	1.009	0.38†	1.728	1.017	0.35†
U. Gap	1.681	0.989	0.23†	1.762	1.037	0.05†
GDP Growth	1.583	0.932*	0.25†	1.603	0.943	0.19†
CUR	1.626	0.957	0.44	1.694	0.997	0.39
Trend	1.667	0.981	-0.36†	1.755	1.033	0.62†
AOPW	1.655	0.974	-0.58†	1.673	0.985	-0.39†
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$						

Table 6: The RMSFE are shown relative to the AO forecast. Significance for the RMSFE results is only indicated for improvements over the benchmark. The † indicates unbiased forecasts significant at the 10% level.

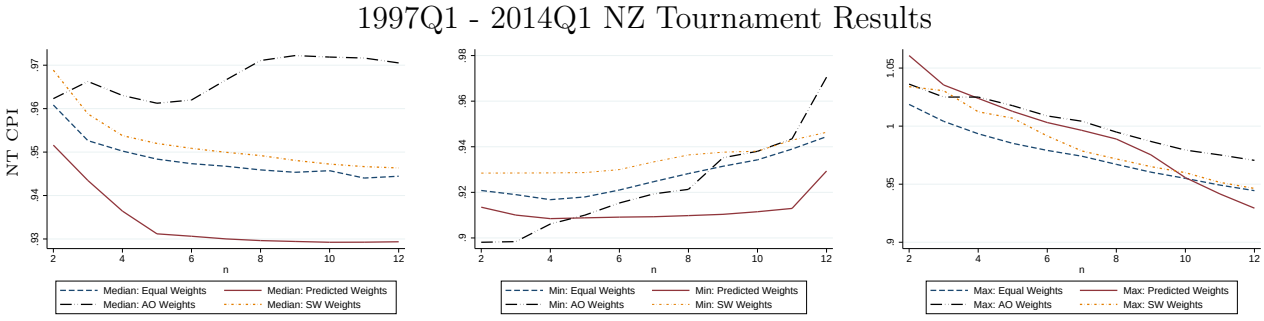

Figure 7: Median, minimum, and maximum relative RMSFE results for combination of n different models with $\beta = 5$ for New Zealand non-tradable inflation. The results are relative to the RMSFE of the AO forecast.

5 Conclusion

The time-varying efficiency of the Phillips curve illustrates a flaw in the assumptions underpinning many forecast combination strategies. Forecast combination strategies almost always assume that recent past performance is positively correlated with performance in the near future. This paper demonstrates that if forecast are weighted by their expected performance, rather than their past performance, that robust improvements in forecast efficiency for inflation are obtained.

The positive result represents a proof-of-concept for forward-looking weights. Time-varying forecast efficiency is not confined only to inflation forecasting. Similar time-variation in forecast efficiency likely exists in many forecasting settings and it may be exploitable by similar forward-looking forecast combination strategies.

References

Ang, A., G. Bekaert, and M. Wei, "Do macro variables, asset markets, or surveys forecast inflation better?," *Journal of Monetary Economics*, 2007, *54* (4), 1163–1212.

Atkeson, Andrew and Lee E Ohanian, "Are Phillips curves useful for forecasting inflation?," *Federal Reserve bank of Minneapolis quarterly review*, 2001, *25* (1), 2–11.

Bates, J.M. and C.W.J. Granger, "The combination of forecasts," *Operations Research*, 1969, pp. 451–468.

Chauvet, Marcelle and Jeremy Piger, "A comparison of the real-time performance of business cycle dating methods," *Journal of Business & Economic Statistics*, 2008, *26* (1), 42–49.

Clark, Todd E and Michael W McCracken, "Advances in forecast evaluation," *working paper*, 2011.

Clemen, R.T., "Combining forecasts: A review and annotated bibliography," *International Journal of Forecasting*, 1989, *5* (4), 559–583.

Clements, M.P. and D.F. Hendry, "Intercept Corrections and Structural Change.," *Journal of Applied Econometrics*, 1996, *11* (5), 475–494.

Cogley, Timothy, Giorgio Primiceri, and Thomas Sargent, "Inflation-gap persistence in the US," *American Economic Journal: Macroeconomics*, 2010, *2*, 43–69.

Croushore, D. and T. Stark, "A real-time data set for macroeconomists," *Journal of Econometrics*, 2001, *105* (1), 111–130.

Diebold, Francis X, "Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold-Mariano tests," *Journal of Business & Economic Statistics*, 2014.

_ **and Robert S Mariano**, "Comparing predictive accuracy," *Journal of Business & economic statistics*, 1995, *13* (3).

Diebold, F.X. and J.A. Lopez, "Forecast evaluation and combination," *Handbook of Statistics*, 1996, *14*, 241–68.

Faust, Jon and Jonathan H Wright, "Forecasting Inflation," *working paper*, 2012.

Giacomini, Raffaella and Barbara Rossi, "Detecting and predicting forecast breakdowns," *The Review of Economic Studies*, 2009, *76* (2), 669–705.

Granger, Clive WJ and Ramu Ramanathan, "Improved methods of combining forecasts," *Journal of Forecasting*, 1984, *3* (2), 197–204.

Granger, C.W.J., "Invited review combining forecasts—twenty years later," *Journal of Forecasting*, 1989, *8* (3), 167–173.

Harvey, David, Stephen Leybourne, and Paul Newbold, "Testing the equality of prediction mean squared errors," *International Journal of forecasting*, 1997, *13* (2), 281–291.

Hendry, D.F. and M.P. Clements, "Pooling of forecasts," *The Econometrics Journal*, 2004, *7* (1), 1–31.

Marcellino, M., J.H. Stock, and M.W. Watson, "A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series," *Journal of Econometrics*, 2006, *135* (1-2), 499–526.

Newey, W.K. and K.D. West, "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix," *Econometrica*, 1987, *55* (3), 703–708.

Orphanides, Athanasios and Simon Van Norden, "The reliability of inflation forecasts based on output gap estimates in real time," *Journal of Money, Credit and Banking*, 2005, pp. 583–601.

Owyang, M.T., J.M. Piger, and H.J. Wall, "Forecasting National Recessions Using State Level Data," *The Journal of Money, Credit and Banking*, 2014, *forthcoming.*

Smith, J. and K.F. Wallis, "A Simple Explanation of the Forecast Combination Puzzle," *Oxford Bulletin of Economics and Statistics*, 2009, *71* (3), 331–355.

Stock, J.H. and M.W. Watson, "Forecasting output and inflation: The role of asset prices," *Journal of Economic Literature*, 2003, *41* (3), 788–829.

_ **and** _ , "Combination forecasts of output growth in a seven-country data set," *Journal of Forecasting*, 2004, *23* (6), 405–430.

_ **and** _ , "Why has US inflation become harder to forecast?," *Journal of Money, Credit and banking*, 2007, *39* (s1), 3–33.

_ **and** _ , "Phillips curve inflation forecasts," Technical Report, National Bureau of Economic Research 2008.

_ **and** _ , "Modeling inflation after the crisis," Technical Report, National Bureau of Economic Research 2010.

Timmermann, A., "Forecast combinations," *Handbook of economic forecasting*, 2006, *1*, 135–196.

Wallis, K.F., "Combining forecasts–forty years later," *Applied Financial Economics*, 2011, *21* (1-2), 33–41.

_ **and J.D. Whitley**, "Sources of error in forecasts and expectations: UK economic models, 1984-88," *Journal of Forecasting*, 1991, *10* (3), 231–253.

West, Kenneth D, "Forecast evaluation," *Handbook of economic forecasting*, 2006, *1*, 99–134.