International Journal of Forecasting 30 (2014) 12–19

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

# Using forecast evaluation to improve the accuracy of the Greenbook forecast

Natsuki Arai *,[1]

*Department of Economics, Johns Hopkins University, 440 Margenthaler Hall, 3400 N. Charles St., Baltimore, MD 21218, United States*

ARTICLE INFO

*Keywords:*
Evaluating forecasts
Forecast efficiency
Adjusting forecasts
Real-time data
The Greenbook forecast

ABSTRACT

Recently, Patton and Timmermann (2012) proposed a more powerful kind of forecast efficiency regression at multiple horizons, and showed that it provides evidence against the efficiency of the Fed's Greenbook forecasts. I use their forecast efficiency evaluation to propose a method for adjusting the Greenbook forecasts. Using this method in a real-time out-of-sample forecasting exercise, I find that it provides modest improvements in the accuracies of the forecasts for the GDP deflator and CPI, but not for other variables. The improvements are statistically significant in some cases, with magnitudes of up to 18% in root mean square prediction error.


## 1. Introduction

Considerable amounts of research have been concerned with forecast efficiency regressions. Recently, researchers have found evidence against efficiency in forecast efficiency regressions in a multi-horizon system. This paper uses these tests to adjust the original forecasts, and finds modest improvements for the Fed's Greenbook forecast for the GDP deflator and CPI, but not for other variables.

In this paper, I propose a new method which can improve the accuracy of forecasts in real time, using the results of the forecast efficiency test. Based on the evidence against the efficiency of the Greenbook forecast presented by Patton and Timmermann (2012), this paper uses the new method to adjust systematic errors of the Greenbook forecast in real time, building on the suggestion of Croushore (2012).

I find modest, but statistically significant, improvements in the out-of-sample forecast accuracy of the Greenbook forecast for the GDP deflator and CPI. Since Romer and Romer (2000) showed that the Greenbook forecast is more efficient than private forecasts, the better performance of the Greenbook forecast in the 1980s has been documented in the literature. Sims (2002) confirms their result that the Greenbook forecast is more accurate than either private or naïve forecasts. Faust and Wright (2009) show that the Greenbook forecast outperforms the forecasts using reduced-form models, even after giving the Greenbook forecast to these models for several quarters. In more recent periods, the Greenbook forecast has not had this advantage (Edge & Gürkaynak, 2010; Reifschneider & Tulip, 2007), but it is still an interesting benchmark for three reasons. First, the Greenbook forecast is well known in the literature, and researchers have already analyzed its accuracy and efficiency from a range of different perspectives. Second, the Greenbook forecast is the most substantial and detailed judgmental US macroeconomic forecast, being based on an immense range of information. Finally, as is indicated by FOMC minutes and transcripts, the Greenbook forecasts have played an essential role in US monetary policy.[2]

By comparing out-of-sample forecast errors during the Great Moderation, this paper shows that the adjustment based on the forecast efficiency test gives modest improvements for the GDP deflator and CPI, but not for other vari-

* Tel.: +1 410 516 7601; fax: +1 410 516 7600.
*E-mail address:* natsuki.ar@gmail.com.

[1] The online appendix, data, and codes are available on the author's website (https://sites.google.com/site/natsukiarai25/research).

[2] For details, see the NBER working paper version of Faust and Wright (2009).

0169-2070/$ – see front matter 
http://dx.doi.org/10.1016/j.ijforecast.2013.02.002

ables. The magnitude of the improvements in root mean square prediction error can be as high as 18%.

The results that significant improvements can be found in the forecasts for inflation, but not in the forecasts for output growth, are consistent with a finding in the recent literature, namely that forecasts of the output growth are hard to improve given a good estimate of the current state of the economy, and output growth was especially unpredictable during the Great Moderation.

Essentially, the proposed method works by determining whether Greenbook forecasters have over- or under-reacted to incoming news in the past, then proposing a systematic adjustment for their past mistakes. Of course, judgmental forecasters should also be monitoring their own performances at the same time, and making these adjustments. It is possible that my correction, applied to the future Greenbook forecasts, would over-correct and effectively make the adjustment twice. However, the evidence that I show in this paper indicates that a real-time implementation of my proposed adjustment would have given better out-of-sample forecasts than the Greenbook itself. That is, of course, no guarantee of future performance, but indicates that the proposed adjustment might well help.

Two methods of inference for nested forecasts, the bootstrap and the test proposed by Clark and West (2007) (henceforth referred to as the CW test), show that these improvements are statistically significant in some cases. A comparison of these two methods shows that they lead to similar results, but the CW test sometimes rejects the null even when the null model is estimated to be more accurate. This is because the CW test adjusts for the parameter estimation error. It may seem surprising that the CW test can recommend using less accurate forecasts, but this point was made by Clark and West (2007).

Lastly, I provide extensions to the main results. First, I apply the same adjustment scheme to another subjective forecast, the SPF forecast. Second, I use a different sample period for the application to the Greenbook forecast, starting before the Great Moderation. The evidence from these two extensions is mixed, but I still find statistically significant improvements in some cases. Third, I provide an analysis of subperiods to shed some light on how the proposed adjustment improves the forecast accuracy of the Greenbook forecast.

The remainder of the paper is organized as follows: Section 2 describes the forecasts and vintage data I use, and Section 3 explains the methodology, including the adjustment of forecast and inference. Section 4 contains the main results and possible interpretations, and Section 5 provides extensions of the main results. Section 6 concludes.

## 2. Data

This paper focuses primarily on the Greenbook forecasts for inflation and output growth: the GDP deflator, CPI inflation, Core CPI inflation and GDP growth. The data for the Greenbook forecast are obtained from the Philadelphia Fed's website. Since the Greenbook forecast is the forecast prepared before the FOMC meeting, which is usually held eight times a year, I pick the forecast that is closest to the middle of each quarter when constructing quarterly forecasts. Although the Greenbook forecast has been being released since 1964, its forecast horizons have varied, especially in the early periods. For four-quarter forecast horizons, the GDP deflator and GDP growth forecasts are available from the second quarter of 1974, the CPI forecast is available from the fourth quarter of 1979, and the Core CPI forecast is available from the first quarter of 1986.

In order to ensure comparability between different series and to focus on the forecasts made during the Great Moderation, I use the data from the first quarter of 1984 to the fourth quarter of 2005 as the benchmark. Given Tulip's (2009) observation that the forecast errors made by the Greenbook forecast were the largest early in the sample period, Faust and Wright (2009) set the sample beginning in 1984 as their baseline case. I also follow this convention, in order to prevent the volatility of the data before the Great Moderation from affecting the whole analysis. Since the Greenbook forecast becomes available to the public with a lag of five years, the fourth quarter of 2005 is the most recent data available.

All of the forecasts and variables are quarterly, and all vintages are recorded quarterly. The vintage data are obtained from the Philadelphia Fed's website. Inflation and output growth rates are computed as annualized percentage changes, $100 * \left(\left(\frac{x_t}{x_{t-1}}\right)^4 - 1\right)$, where $x_t$ is a price or output level at time $t$. The results using the continuously compounding annual rate of change, $400 * \log\left(\frac{x_t}{x_{t-1}}\right)$, are listed in the online appendix, but the differences between these results are very small. For CPI and Core CPI, I use the data recorded in December 2010 and treat the data up until time $t-1$ as a vintage of time $t$, ignoring the issues associated with the lack of real-time data. This is because the availability of vintage data for CPI and Core CPI is limited and the revisions to these two measures are trivial. Quarterly price levels are computed by averaging the monthly values for the three months in the quarter.

When using real-time data, it is important to adopt a definition of "the realized value". In this paper, I follow Faust and Wright (2009) and treat the data released two quarters after the forecasted date as the realized value.[3] For example, I treat the output growth from the second quarter of 1994, recorded in the fourth quarter of 1994, as the realized value for the computation of forecast errors.

## 3. Method

### 3.1. Multi-horizon forecast efficiency evaluation

First, I apply the forecast efficiency evaluation across multiple horizons, proposed by Patton and Timmermann (2012), to the Greenbook forecast. Let $y_{t+1}$ be a variable to be forecasted at time $t+1$, and $\hat{y}_{t+1|t}$ be a forecast of

[3] For a more detailed discussion, see Faust and Wright (2009) and Tulip (2009). On the way in which data revisions affect the qualitative implications of forecasting, see Croushore (2006).

$y_{t+1}$ at time $t$. The standard Mincer–Zarnowitz regression for testing forecast efficiency is given by the following equation:

$$y_{t+1} = \alpha + \beta \hat{y}_{t+1|t} + \varepsilon_{t+1}. \quad (1)$$

Since standard forecast efficiency implies that forecasts are the conditional mean of forecasted variables, the null hypothesis is $[\alpha, \beta] = [0, 1]$.

Now define a revision of the forecast for $t$ between $t-i$ and $t-j$, for $0 < i < j$, as $d_{t|i,j} \equiv \hat{y}_{t|t-i} - \hat{y}_{t|t-j}$. By definition, a recent forecast is described as the sum of the forecast at a longer horizon and subsequent forecast revisions: $\hat{y}_{t|t-i} = \hat{y}_{t|t-j} + \sum_{k=i}^{j-1} d_{t|k,k+1}$. By replacing the nowcast in Eq. (1), $\hat{y}_{t+1|t}$, with the sum of an old forecast and subsequent revisions, $\hat{y}_{t+1|t-j} + \sum_{k=1}^{j} d_{t+1|k,k+1}$, Eq. (1) can be rewritten as follows:

$$y_{t+1} = \alpha + \beta\, \hat{y}_{t+1|t-j} + \sum_{k=1}^{j} \gamma_k\, d_{t+1|k,k+1} + \varepsilon_{t+1}, \quad (2)$$

with the null hypothesis of $[\alpha, \beta, \gamma_1, \ldots, \gamma_j] = [0, 1, 1, \ldots, 1]$, where $j$ denotes the number of forecast revisions included in the regression. This regression tests the implication of forecast efficiency that forecasts are the conditional mean and the subsequent revisions are orthogonal to the past forecasts. The F-test is used for this regression to test the null hypothesis jointly. Using a Monte Carlo simulation, Patton and Timmermann (2012) show that this multi-horizon forecast efficiency evaluation has a greater power to detect forecast inefficiency in finite samples.

In addition, Patton and Timmermann (2012) apply this method to the Greenbook forecast and reject its efficiency. Although Clements, Joutz, and Stekler (2007) also reject the efficiency of the Greenbook forecast by pooling forecast errors at different horizons, this multi-horizon approach is more straightforward. This paper also focuses primarily on the Greenbook forecast in the following sections.

### 3.2. Adjustment of forecast

Given the forecast evaluation process described in the previous section, I propose a simple method which is able to improve the accuracy of forecasts. Essentially, I treat the predicted value from the forecast efficiency regression as a new forecast, which adjusts the systematic errors of the original forecast.

Suppose that I evaluate the forecast efficiency using the regression in Eq. (2) every period. Specifically, I observe the following series in period $t$: the vintage of the forecasted variable in period $t$, $\{y_{j+1|t}, \ldots, y_{t|t}\}$; the Greenbook forecast $j+1$ periods before, $\{\hat{y}_{j+1|0}, \ldots, \hat{y}_{t|t-j-1}\}$; and subsequent revisions of the Greenbook forecast from $j+1$ periods before to 1 period before, $\left\{\{d_{j+1|k,k+1}\}_{k=1}^{j}, \ldots, \{d_{t|k,k+1}\}_{k=1}^{j}\right\}$. Having evaluated the forecast efficiency by using these series up until time $t$, I then plug the forecast for period $t+1$ made $j+1$ periods before and subsequent revisions, $\hat{y}_{t+1|t-j}$ and $\{d_{t+1|k,k+1}\}_{k=1}^{j}$, into the estimated equation, and treat its prediction as another forecast. In order to simplify the algorithm and to use all available information, I treat the vintage at each period as the realized value. This definition is given by the following equation:

$$\tilde{y}_{t+1|t,j} \equiv \hat{\alpha}_t + \hat{\beta}_t\, \hat{y}_{t+1|t-j} + \sum_{k=1}^{j} \hat{\gamma}_{t,k}\, d_{t+1|k,k+1}, \quad (3)$$

where $\hat{\alpha}_t$, $\hat{\beta}_t$ and $\{\hat{\gamma}_{t,k}\}_{k=1}^{j}$ are the estimated coefficients from Eq. (2) in period $t$. By repeating this procedure every period to obtain the predictions from the forecast efficiency regression, I form the adjusted forecast in real time.

The intuition behind this adjustment is that the new forecast adjusts the systematic errors which the original forecast made in the past. If the null hypothesis in Eq. (2) is rejected, this means that previous forecast revisions over- or under-reacted to incoming news, or were systematically too optimistic or pessimistic, leading to inefficiencies in the original forecast. As a result, correcting these systematic errors as in Eq. (3) gives researchers an opportunity to improve the accuracy of the original forecast. By construction, the adjusted forecasts will not contain these systematic errors, and so the adjustment cannot then be applied again. In his discussion of the work of Patton and Timmermann (2012), Croushore (2012) suggested the idea of using the test to create improved forecasts, which I am implementing in this paper.

### 3.3. Extension to longer-horizon forecasts

The extension of the baseline method to longer-horizon forecasts is straightforward: extend the forecast horizon to $h$ and focus on $y_{t+h}$ instead of $y_{t+1}$. Then, the extended forecast evaluation becomes

$$y_{t+h} = \alpha_h + \beta_h\, \hat{y}_{t+h|t-j} + \sum_{k=h}^{h+j-1} \gamma_{h,k}\, d_{t+h|k,k+1} + \varepsilon_{t+h}, \quad (4)$$

with the null hypothesis of $[\alpha_h, \beta_h, \gamma_{h,h}, \ldots, \gamma_{h,h+j-1}] = [0, 1, 1, \ldots, 1]$. The only difference between Eqs. (2) and (4) is that Eq. (4) does not contain recent forecast revisions, $\{d_{t+h|1,2}, \ldots, d_{t+h|h-1,h}\}$, since they are not available at time $t$. In other words, I replace recent forecast revisions with old forecast revisions, in order to evaluate the forecast efficiency in real time. Then, I define the adjusted forecast in exactly the same way:

$$\tilde{y}_{t+h|t,j} \equiv \hat{\alpha}_{h|t} + \hat{\beta}_{h|t}\, \hat{y}_{t+h|t-j} + \sum_{k=h}^{h+j-1} \hat{\gamma}_{h|t,k}\, d_{t+h|k,k+1}, \quad (5)$$

where $\hat{\alpha}_{h|t}$, $\hat{\beta}_{h|t}$ and $\{\hat{\gamma}_{h|t,k}\}_{k=h}^{h+j-1}$ are the estimated coefficients in Eq. (4) using the series up until $t$. Trivially, it is a generalization of the forecast in Eq. (3). Note that I can only extend this model for a few periods, since the number of horizons in the Greenbook forecast is limited, being only four quarters ahead in my dataset.

### 3.4. Test statistic and inference

The natural metric for comparing the forecast accuracies of the Greenbook forecast and the adjusted forecast is the out-of-sample Relative Root Mean Square Prediction Error (RRMSPE). The RRMSPE is defined as the ratio of the

RMSPE of the adjusted forecast for $h$ periods ahead, with the adjustment using $j$ forecast revisions, to the RMSPE of the Greenbook forecast for $h$ periods ahead:

$$\text{RRMSPE}_{h|j} \equiv \sqrt{\frac{\sum_{t=1}^{T}(\tilde{y}_{t+h|t,j} - y_{t+h})^2}{\sum_{t=1}^{T}(\hat{y}_{t+h|t} - y_{t+h})^2}}, \quad (6)$$

where $\tilde{y}_{t+h|t,j}$ is the adjusted forecast for $t+h$ made at $t$ using $j$ forecast revisions, $\hat{y}_{t+h|t}$ is the Greenbook forecast for $t+h$ made at $t$, and $T$ is the number of predictions. Since the RMSPE of the Greenbook forecast is in the denominator, a value of the RRMSPE which is larger than unity indicates that the Greenbook forecast outperforms the adjusted forecast. Under the null hypothesis that the Greenbook forecast is efficient, the RRMSPE has an expected value greater than unity, because the in-sample over-fitting worsens the out-of-sample predictive accuracy in small samples.

The out-of-sample RRMSPEs are calculated after forty quarters from the starting point of the sample of each series. Adjusted forecasts are computed in two ways: recursive, where the entire sample is used to estimate the model; and rolling, where the samples of the forty most recent quarters are used.

If I set $[\hat{\alpha}_{h|t}, \hat{\beta}_{h|t}, \hat{\gamma}_{h|t,h}, \ldots, \hat{\gamma}_{h|t,h+j-1}] = [0, 1, 1, \ldots, 1]$ for all $t$ in Eq. (5), as is consistent with the null hypothesis, the adjusted forecast becomes identical to the Greenbook forecast. In other words, the adjusted forecast nests the Greenbook forecast. When forecasting models are nested, the distribution of the test statistic presented by Diebold and Mariano (1995) is not asymptotically normal. The literature on forecast evaluation shows that testing the null hypothesis of equal MSPEs for nested models with normal critical values results in severe size distortions and poor power in practice.[4] Similarly, the RRMSPE for nested models has a nonstandard distribution, and assessing its statistical significance raises a number of econometric issues. In order to avoid these issues, this paper uses two different methods of inference: the bootstrap and the CW test.

#### 3.4.1. Bootstrap

The bootstrap $p$-values are constructed by using the null hypothesis that the Greenbook forecast is efficient, and therefore it is a conditional mean of the realized series. By resampling from the residuals of the AR(4) model, I first make an artificial realized series. Then I treat the conditional mean of the artificial series as artificial Greenbook forecasts and construct adjusted forecasts in exactly the same way. By computing the RRMSPEs of these artificial forecasts and repeating this procedure an arbitrarily large number of times, I can form the distribution of the bootstrap RRMSPE and report $p$-values of the realized RRMSPEs. The specific algorithm is described in detail in the Appendix.

[4] For details, see Faust and Wright (2012) and West (2006).

#### 3.4.2. The CW test

This paper uses the CW test as an alternative method of inference for nested forecasts. The CW test first adjusts the noise in the MSPE which is due to the estimation of additional parameters in an alternative forecasting model, which nests a parsimonious null forecasting model. It then tests the hypothesis that the null forecasting model is correctly specified and the prediction errors of these two models are the same in the population, by using normal critical values.[5]

The specific procedure is summarized succinctly in Section 2 of Clark and West (2007). In the context of this paper, I first compute the following statistic:

$$f_{t+h|t,j} \equiv (\hat{y}_{t+h|t} - y_{t+h})^2 - [(\tilde{y}_{t+h|t,j} - y_{t+h})^2 - (\hat{y}_{t+h|t} - \tilde{y}_{t+h|t,j})^2], \quad (7)$$

where $\hat{y}_{t+h|t}$ is the Greenbook forecast for $t+h$ made at $t$, $\tilde{y}_{t+h|t,j}$ is the adjusted forecast for $t+h$ made at $t$ using $j$ revisions, and $y_{t+h}$ is the realized value at $t+h$. Then I then regress $f_{t+h|t,j}$ on a constant and derive the $t$-statistic. If this $t$-statistic is larger than 1.282 (1.645), I reject the null hypothesis at the 10% (5%) significance level, respectively. Based on the evidence from the Monte Carlo exercise in finite samples, Clark and West (2007) argue that this $t$-statistic is approximated well by a normal distribution, even though it has a nonstandard asymptotic distribution.

## 4. Results

Recursive and rolling RRMSPEs of the adjusted forecasts are reported in Table 1. Adjusted forecasts are computed for nowcasts through four-quarter-ahead forecasts ($h = 1, \ldots, 5$), using as many revisions as possible; for example, the adjusted nowcasts are based on four subsequent revisions, and the one-period-ahead adjusted forecasts are based on three subsequent revisions. The results using all the possible numbers of forecast revisions are listed in the online appendix.

### 4.1. Inflation and output growth

The results on the forecast accuracy of adjusted forecasts for inflation are mixed. I find significant improvements for the CPI and GDP deflator forecasts both for nowcasts and for forecasts at longer horizons, whereas significant improvements for the Core CPI forecasts are only found in nowcasts.

For the CPI forecast, the RRMSPEs are smaller than one at almost all horizons, for both the recursive and rolling regressions. In addition, most of the improvements are statistically significant, with magnitudes ranging from 4.1% to 13.0%, where the significance level varies from 1% to 10%. This implies that the Greenbook forecast made systematic errors in its CPI forecasts, the adjustment of which leads to significant gains in forecast accuracy in

[5] On the other hand, Clark and McCracken (2009, 2011) discuss the use of inference to test the null hypothesis that two models have equal RMSPEs in *finite sample*.

**Table 1**
RRMSPE of the adjusted forecast relative to the Greenbook forecast (the sample during the Great Moderation).

| Series | GDP deflator | CPI | Core CPI | GDP growth |
|---|---|---|---|---|
| Panel A: recursive | | | | |
| Nowcasts | 1.033 | $0.959^{***}_{‡}$ | $0.952^{**}_{‡}$ | 1.026 |
| 1Q ahead | $0.958^{*}_{‡}$ | 1.009 | 1.036 | 1.073 |
| 2Q ahead | 1.002 | $0.952_{†}$ | 1.105 | 1.040 |
| 3Q ahead | 1.005 | $0.906^{**}_{‡}$ | 1.050 | 1.031 |
| 4Q ahead | 0.997 | $0.954^{*}_{†}$ | 1.018 | 1.021 |
| Panel B: rolling with a forty-quarter window | | | | |
| Nowcasts | $0.877^{***}_{‡}$ | $0.942^{***}_{‡}$ | $1.001^{**}_{‡}$ | 1.045 |
| 1Q ahead | $0.820^{***}_{‡}$ | $0.906^{***}_{‡}$ | 1.080 | 1.081 |
| 2Q ahead | $0.920^{**}_{‡}$ | $0.899^{**}_{†}$ | 1.133 | 1.065 |
| 3Q ahead | 0.984 | $0.870^{***}_{‡}$ | 1.137 | $1.065_{†}$ |
| 4Q ahead | 0.965 | $0.912^{**}_{†}$ | 1.085 | $1.060_{‡}$ |

a. This table shows the RRMSPEs of the adjusted forecast to the Greenbook forecast from 40 quarters after the first period. The Core CPI series start from 1986Q1, and all other series start from 1984Q1. All series end in 2005Q4.
b. The superscripts *, ** and *** denote significance at the 10%, 5% and 1% levels, respectively, based on the bootstrap with 10,000 replications. For the construction of bootstrap *p*-values, see Section 3.4.1 and the Appendix.
c. The subscripts † and ‡ denote significance at the 10% and 5% levels, respectively, based on the CW test. Newey–West standard errors with a lag truncation of four are used.

real time. In addition, there are also many cases where I find significant improvements for the GDP deflator forecast from the adjustment, especially in rolling regressions. The improvements range in magnitude from 4.2% to 18.0%. However, for the GDP deflator forecast, the recursive results show few significant improvements, unlike the rolling results. Unlike the cases of the CPI and GDP deflator forecasts, significant improvements for the Core CPI forecast are found only with nowcasts.

Faust and Wright (2009) show that the Greenbook forecast is such a good forecast of inflation that it outperforms reduced-form forecasts, even after giving the Greenbook's nowcasts and longer-horizon forecasts to reduced-form forecasting models for several quarters. Sims (2002) also suggests that the superiority of the Greenbook forecast arises from its advantage in the timing of information. However, the results presented here suggest that the Greenbook forecasts (which incorporate economic judgment that may make it perform better than reduced-form models) still made systematic errors, meaning that there is still room to improve upon the Greenbook forecast.

On the other hand, the performance of the adjusted forecast for output growth is quite different to that of the inflation forecasts. I find no improvement in either recursive or rolling regressions. These results are consistent with the findings of Faust and Wright (2009) and Tulip (2009) that output growth during the Great Moderation is largely unpredictable, especially at longer horizons. Tulip (2009) argues that the predictable volatility of output growth vanishes during the Great Moderation.

### 4.2. Bootstrap and the CW test

One objective of this paper is to compare the two different methods of inference, the bootstrap and the CW test, in the context of an important practical application. As is evident from the results, the bootstrap and the CW test generally lead to the similar results. However, there are some cases where the CW test rejects the null even though the RRMSPE is larger than unity. This is because the CW test adjusts for parameter estimation error. If the restricted model is correct, then it should give substantially *more* accurate forecasts in small samples, because of parameter uncertainty. If the improvement is small enough, then we could still conclude that the restricted model is to be rejected. It may seem surprising that the CW test could recommend using the less accurate forecasts, but this is due to the effects of parameter estimation error, as was pointed out by Clark and West (2007).[6]

A Monte Carlo simulation reported in the online appendix confirms these results. It shows that the bootstrap inference is generally more conservative than the CW test. The CW test is modestly oversized and has a higher power, whereas the bootstrap inference is modestly undersized and has a lower power. Which test gives a higher size-adjusted power depends on the simulations.

## 5. Extensions

### 5.1. Comparison with the SPF forecast

In order to see whether the adjusted forecasts improve upon the original forecasts in the case of other judgmental forecasts, I apply the same adjustment to the Survey of Professional Forecasters (SPF) forecasts of the GDP deflator, CPI inflation and GDP growth. The data are obtained from the Philadelphia Fed's website.[7] The recursive and rolling results are reported in Table 2. More detailed results are reported in the online appendix.

The results using the SPF forecast are mixed. Unlike the case of the Greenbook forecast, I find few significant improvements for either inflation or output growth. Even though the overall accuracy of the SPF forecast is not necessarily better than that of the Greenbook forecast, its errors are not as systematic as the Greenbook, and the adjustments using forecast evaluation can improve the original forecast in a few cases, but not all.

### 5.2. Different subsamples

Generally, the accuracy of the forecasts is very sensitive to the choice of the sample period. For example, Edge and Gürkaynak (2010) show that none of their forecasts, including the Greenbook forecast and forecasts produced using DSGE models, perform better than a constant forecast when looking at forecasts from 1992 to 2006, but this disagrees with the results from earlier samples.

6 For example, see Clark and West (2007, p. 309).
7 Core CPI is not included because the sample period is too short.

**Table 2**
RRMSPE of the adjusted forecast relative to the SPF forecasts (the sample during the Great Moderation).

| Series | GDP deflator | | CPI | | GDP growth | |
|---|---|---|---|---|---|---|
| | Mean | Median | Mean | Median | Mean | Median |
| Panel A: recursive | | | | | | |
| Nowcasts | 1.053 | 1.043 | $0.987^{**}_{\ddagger}$ | 1.017 | $1.001^{*}$ | $0.996^{*}$ |
| 1Q ahead | $1.059_{\dagger}$ | 1.031 | 1.005 | 1.010 | 1.042 | 1.031 |
| 2Q ahead | $1.017_{\dagger}$ | 1.010 | 1.011 | 1.025 | 1.044 | 1.025 |
| 3Q ahead | 1.015 | 1.002 | 1.014 | 1.021 | 1.061 | 1.042 |
| 4Q ahead | $0.984_{\dagger}$ | 1.012 | 1.012 | 1.016 | 1.000 | 1.022 |
| Panel B: rolling with a forty-quarter window | | | | | | |
| Nowcasts | $0.983^{**}_{\ddagger}$ | 1.062 | $1.016^{**}_{\ddagger}$ | 1.053 | 1.207 | 1.169 |
| 1Q ahead | $1.047_{\ddagger}$ | 1.029 | 1.027 | 1.063 | 1.179 | 1.145 |
| 2Q ahead | $0.984_{\ddagger}$ | $1.017_{\ddagger}$ | 1.051 | 1.071 | 1.118 | 1.100 |
| 3Q ahead | $1.009_{\ddagger}$ | $1.044_{\dagger}$ | 1.058 | 1.073 | 1.124 | 1.128 |
| 4Q ahead | $1.021_{\dagger}$ | 1.033 | 1.035 | 1.050 | 1.084 | 1.110 |

a. This table shows the RRMSPEs of the adjusted forecast to the SPF forecast from 40 quarters after the first period. All series start from 1984Q1 and end in 2005Q4.
b. Same as Table 1.
c. Same as Table 1.

In order to see the effect of the choice of sample period on the results, I conduct the same analysis using the entire available sample: from the second quarter of 1974 for GDP deflator and output growth, and from the fourth quarter of 1979 for CPI inflation. The recursive and rolling results are reported in Table 3. Even though there are some differences, the results of the samples from 1974 and 1979 are to some extent similar to the results of the subsample during the Great Moderation. For inflation, there are significant improvements from the adjustments in both nowcasts and longer-horizon forecasts in rolling regressions. On the other hand, for output growth, I see no significant improvements from using either recursive or rolling regressions.

The RRMSPEs of the adjusted SPF forecasts for the entire available samples are listed in the online appendix in order to conserve space here, but the results are generally similar to the subsample from 1984. There are fewer improvements than in the case of the Greenbook forecast.

### 5.3. Analysis in subperiods

To shed some light on the effect of the proposed adjustment in improving the forecast accuracy, I provide a simple analysis of forecast revisions and a breakdown of the rolling RRMSPEs in three subperiods: 1994–1997, 1998–2001, and 2002–2005.

First, Table 4 shows the sample average first-order autocorrelation and the average bias of forecast revisions for each subperiod. Since forecast revisions for the fixed target period are unpredictable under forecast efficiency, both the first-order autocorrelation and the bias of forecast revisions should be zero. However, the sample average first-order autocorrelations are negative for all subperiods and series, and the average biases are sometimes notably different from zero for all series. These statistics suggest that there is inefficiency in several subperiods. For example, the average autocorrelations for CPI are −0.250, −0.293 and −0.089 in the three subperiods.

**Table 3**
RRMSPE of the adjusted forecast relative to the Greenbook forecast (the entire available sample).

| Series | GDP deflator | CPI | GDP growth |
|---|---|---|---|
| Panel A: recursive | | | |
| Nowcasts | 1.103 | 1.706 | 1.062 |
| 1Q ahead | 1.047 | $1.101_{\dagger}$ | $1.023_{\ddagger}$ |
| 2Q ahead | 1.040 | 1.102 | 1.053 |
| 3Q ahead | 1.055 | 1.048 | 1.018 |
| 4Q ahead | 1.067 | 1.002 | 1.031 |
| Panel B: rolling with a forty-quarter window | | | |
| Nowcasts | $1.024^{***}_{\ddagger}$ | $1.063_{\dagger}$ | 1.091 |
| 1Q ahead | $0.945^{***}_{\ddagger}$ | $0.973^{**}_{\ddagger}$ | $1.037_{\ddagger}$ |
| 2Q ahead | $0.993^{*}_{\dagger}$ | $0.951_{\dagger}$ | $1.058_{\ddagger}$ |
| 3Q ahead | 1.036 | $0.856^{***}_{\ddagger}$ | $1.069_{\dagger}$ |
| 4Q ahead | 1.054 | $0.931^{**}_{\dagger}$ | $1.076_{\dagger}$ |

a. This table shows the RRMSPEs of the adjusted forecast to the Greenbook forecast from 40 quarters after the first period. The GDP deflator and growth series start from 1974Q2, and the CPI series start from 1979Q4. All of the series end in 2005Q4.
b. Same as Table 1.
c. Same as Table 1.

**Table 4**
Average first-order autocorrelation and average bias of the revisions of the Greenbook forecast in subperiods.

| Series | GDP deflator | CPI | Core CPI | GDP growth |
|---|---|---|---|---|
| Panel A: average autocorrelation | | | | |
| 1994Q1–1997Q4 | −0.210 | −0.250 | −0.206 | −0.171 |
| 1998Q1–2001Q4 | −0.271 | −0.293 | −0.205 | −0.072 |
| 2002Q1–2005Q4 | −0.110 | −0.089 | −0.134 | −0.260 |
| Panel B: average bias | | | | |
| 1994Q1–1997Q4 | 0.011 | −0.019 | −0.027 | 0.086 |
| 1998Q1–2001Q4 | −0.039 | −0.022 | −0.055 | −0.077 |
| 2002Q1–2005Q4 | 0.030 | 0.191 | 0.019 | −0.177 |

a. This table shows the sample average first-order autocorrelation and average bias of the revisions of the Greenbook forecast, during the subperiods 1994Q1–1997Q4, 1998Q1–2001Q4 and 2002Q1–2005Q4.

Second, Table 5 shows the breakdown of the rolling RRMSPEs of the adjusted forecast relative to the Greenbook forecast for the same subperiods. The improvements vary across the series and subperiods. The improvements of the adjusted forecasts for the GDP deflator and CPI inflation are due mainly to the improvements in the first and the last subperiods (1994–1997 and 2002–2005). To save space, I have listed the breakdown of the recursive RRMSPEs in the online appendix.

## 6. Conclusion

This paper addresses a question in relation to the Fed's Greenbook forecast: Given the evidence against the forecast efficiency of the Greenbook forecast, which has been provided recently by multi-horizon forecast efficiency regressions, can researchers improve its forecast accuracy in real time? I propose a new method that uses this evidence against efficiency to adjust the Greenbook forecast. Using this method in a real-time out-of-sample forecasting exercise, I find that it leads to modest

**Table 5**
Rolling RRMSPE of the adjusted forecasts relative to the Greenbook forecasts in subperiods, with a forty-quarter window.

| Series | GDP deflator | CPI | Core CPI | GDP growth |
|---|---|---|---|---|
| Panel A: nowcasts | | | | |
| 1994Q1–1997Q4 | 0.826 | 1.048 | 1.263 | 0.997 |
| 1998Q1–2001Q4 | 1.208 | 1.204 | 0.805 | 0.907 |
| 2002Q1–2005Q4 | 0.742 | 0.865 | 1.050 | 1.277 |
| Panel B: one-quarter-ahead forecasts | | | | |
| 1994Q1–1997Q4 | 0.783 | 1.099 | 1.043 | 1.067 |
| 1998Q1–2001Q4 | 1.113 | 1.272 | 1.177 | 1.034 |
| 2002Q1–2005Q4 | 0.689 | 0.731 | 1.078 | 1.204 |
| Panel C: two-quarter-ahead forecasts | | | | |
| 1994Q1–1997Q4 | 0.959 | 1.052 | 1.086 | 1.045 |
| 1998Q1–2001Q4 | 1.185 | 1.312 | 0.991 | 0.973 |
| 2002Q1–2005Q4 | 0.792 | 0.735 | 1.171 | 1.315 |
| Panel D: three-quarter-ahead forecasts | | | | |
| 1994Q1–1997Q4 | 1.101 | 0.925 | 1.057 | 1.065 |
| 1998Q1–2001Q4 | 1.173 | 1.209 | 0.967 | 0.947 |
| 2002Q1–2005Q4 | 0.833 | 0.726 | 1.190 | 1.351 |
| Panel E: four-quarter-ahead forecasts | | | | |
| 1994Q1–1997Q4 | 1.023 | 0.922 | 0.964 | 1.029 |
| 1998Q1–2001Q4 | 1.193 | 1.214 | 1.113 | 0.885 |
| 2002Q1–2005Q4 | 0.820 | 0.814 | 1.139 | 1.530 |

a. This table shows the rolling RRMSPEs of the adjusted forecasts to the Greenbook forecast during the subperiods 1994Q1–1997Q4, 1998Q1–2001Q4 and 2002Q1–2005Q4. For Core CPI, the RRMSPEs from 1996Q1 to 1997Q4 are computed for the first subperiod.

improvements in the forecast accuracy of the Greenbook forecasts for inflation. These improvements are statistically significant in some cases.

Specifically, I construct another forecast that adjusts the systematic errors of the Greenbook forecasts in real time, by collecting the predictions of a multi-horizon forecast efficiency regression every period. Then, I compare out-of-sample performances of the adjusted forecast and the Greenbook forecast. By focusing on the Great Moderation, I find modest improvements from the adjustment for the GDP deflator and CPI forecasts, but not for the forecasts for other variables. The magnitude of the improvement in root mean square prediction error can be up to 18%.

Given the results in this paper, one might be tempted to take a more general approach to Patton and Timmermann's forecast evaluation regression, in which the coefficients shrink toward one with a Bayesian algorithm. If the prior is very dogmatic, then that would impose forecast efficiency. On the other hand, a very diffuse prior would correspond to the approach in this paper. There would be some possibilities between these two extreme approaches, but I leave this generalization as a future exercise.

## Acknowledgments

I am grateful to Jon Faust, Jonathan Wright and two anonymous referees for their advice and helpful comments. I thank Blair Chapman, Chris Martin, Rodrigo Sekkel, and David Vera for their valuable comments on earlier drafts, and participants in the seminars at JHU and the the University of Washington in Saint Louis. All errors are the sole responsibility of the author.

## Appendix. Construction of bootstrap *p*-values

The algorithm for constructing bootstrap $p$-values is as follows:

1. Fit an AR(4) model to the realized series, $\{y_t\}$;

$$y_t = \alpha + \sum_{k=1}^{4} \phi_k\, y_{t-k} + \varepsilon_t. \tag{8}$$

2. Randomly resample from the residuals and create an artificial sample, $\{y_t^b\}$;

$$y_t^b = \hat{\alpha} + \sum_{k=1}^{4} \hat{\phi}_k\, y_{t-k}^b + e_t^b, \tag{9}$$

where $\hat{\alpha}$ and $\{\hat{\phi}_k\}_{k=1}^4$ are estimated coefficients of the AR(4) model in Eq. (8), and $e_t^b$ is a randomly resampled residual. I randomly pick a block of four observations to set as the initial observations of an artificial sample.

3. Calculate the conditional mean of an artificial sample at all horizons and take it as an artificial Greenbook forecast. Specifically, the conditional mean is computed in the following way:

$$\hat{y}_{t+h|t}^b = \begin{cases} \hat{\alpha} + \sum_{k=1}^{4} \hat{\phi}_k\, y_{t-k}^b & \text{if } h = 1 \\ \hat{\alpha} + \sum_{k=1}^{h-1} \hat{\phi}_k\, \hat{y}_{t+h-k|t}^b + \sum_{k=h}^{4} \hat{\phi}_k\, y_{t+h-k}^b & \text{if } 2 \le h \le 4 \\ \hat{\alpha} + \sum_{k=1}^{4} \hat{\phi}_k\, \hat{y}_{t+h-k|t}^b & \text{if } h = 5. \end{cases}$$

For the first four observations, I take the unconditional mean to be the forecasts at all horizons.

4. Given an artificial sample and an artificial Greenbook forecast, construct an adjusted forecast in exactly the same way as in Section 3 for the $h$-period-ahead forecast:

$$\tilde{y}_{t+h|t,j}^b \equiv \hat{\alpha}_{h|t}^b + \hat{\beta}_{h|t}^b\, \hat{y}_{t+h|t-j}^b + \sum_{k=h}^{h+j-1} \hat{\gamma}_{h|t,k}^b\, d_{t+h|k,k+1}^b, \tag{10}$$

where $d_{t|i,j}^b \equiv \hat{y}_{t|t-i}^b - \hat{y}_{t|t-j}^b$ for $0 < i < j$ and $\hat{\alpha}_{h|t}^b$, $\hat{\beta}_{h|t}^b$ and $\{\hat{\gamma}_{h|t,k}^b\}_{k=h}^{h+j-1}$ are the coefficients from the forecast efficiency regression using artificial data up until $t$. I plug the forecast for period $t+h$ made $h+j$ periods before, plus subsequent revisions, $\hat{y}_{t+h|t-j}^b$ and $\{d_{t+h|k,k+1}^b\}_{k=h}^{h+j-1}$, into the estimated equation and treat its prediction as another forecast.

5. Repeat this procedure every period to obtain predictions. I take these predictions as the artificial adjusted forecast.

6. Compute the RRMSPE of the artificial adjusted forecast relative to the artificial Greenbook forecast. Unlike the

procedure in Section 3, I assume that the realized series are observable and never revised.

7. Repeat steps 2–6 to form the distribution of the bootstrap RRMSPE. I report $p$-values of the realized RRMSPE according to this distribution.

## References


Clark, T. E., & McCracken, M. W. (2009). *Nested forecast model comparisons: a new approach to testing equal accuracy*. Working paper.

Clark, T. E., & McCracken, M. W. (2011). *Advances in forecast evaluation*. Working paper.

Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, *138*(1), 291–311.

Clements, M. P., Joutz, F., & Stekler, H. O. (2007). An evaluation of the forecasts of the Federal Reserve: a pooled approach. *Journal of Applied Econometrics*, *22*(1), 121–136.

Croushore, D. (2006). Forecasting with real-time macroeconomic data, In G. Elliott, C. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting, vol. 1* (Chapter 17, pp. 961–982). Elsevier.

Croushore, D. (2012). Comment on "Forecast rationality tests based on multi-horizon bounds". *Journal of Business and Economic Statistics*, *30*(1), 17–20.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*(3), 253–263.

Edge, R. M., & Gürkaynak, R. S. (2010). How useful are estimated DSGE model forecasts for central bankers? *Brookings Papers on Economic Activity*, *2*, Fall.

Faust, J., & Wright, J. H. (2009). Comparing Greenbook and reduced form forecasts using a large realtime dataset. *Journal of Business and Economic Statistics*, *27*(4), 468–479.

Faust, J., & Wright, J. H. (2012). *Forecasting inflation*. Working Paper. June.

Patton, A. J., & Timmermann, A. (2012). Forecast rationality tests based on multi-horizon bounds. *Journal of Business and Economic Statistics*, *30*(1), 1–17.

Reifschneider, D., & Tulip, P. (2007). *Gauging the uncertainty of the economic outlook from historical forecasting errors*. Working paper. November.

Romer, C. D., & Romer, D. H. (2000). Federal Reserve information and the behavior of interest rates. *American Economic Review*, *90*(3), 429–457.

Sims, C. A. (2002). The role of models and probabilities in the monetary policy process. *Brookings Papers on Economic Activity*, *33*(2002-2), 1–62.

Tulip, P. (2009). Has the economy become more predictable? Changes in Greenbook forecast accuracy. *Journal of Money, Credit and Banking*, *41*(6), 1217–1231.

West, K. D. (2006). Forecast evaluation. In G. Elliott, C. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting, vol. 1* (Chapter 3, pp. 99–134). Elsevier.