Detecting Time-dependent Bias in the Fed's Greenbook Forecasts of Foreign GDP Growth*

Neil R. Ericsson† Emilio J. Fiallos‡ J E. Seymour§

Revised: February 2017

Abstract
Building on Sinclair, Joutz, and Stekler (2010) and Ericsson, Hood, Joutz, Sinclair, and Stekler (2013), this paper examines publicly available Fed Greenbook forecasts of several foreign countries' GDP growth, focusing on potential biases in the forecasts. While standard tests typically fail to detect biases, recently developed indicator saturation techniques detect economically sizable and highly significant time-varying biases. Estimated biases differ not only over time, but by country and across the forecast horizon.

Key Words: Autometrics, bias, Federal Reserve, forecasts, foreign countries, GDP, Greenbook, impulse indicator saturation, Tealbook, United States

1. Introduction

The Fed's monetary policy has attracted considerable attention domestically and abroad; see Bernanke (2012) and Yellen (2012) *inter alia* for recent discussions. Monetary policy decisions at the Fed are based in part on the "Greenbook" forecasts, which are economic forecasts produced by the Fed's staff. The Greenbook forecasts of *U.S.* economic variables have been extensively analyzed, including by Romer and Romer (2008), Sinclair, Joutz, and Stekler (2010), Nunes (2013), and Ericsson, Hood, Joutz, Sinclair, and Stekler (2015). Surprisingly, Greenbook forecasts of *foreign* economic variables have not been examined, even though foreign economic activity is often a topic of discussion in the Fed's deliberations on monetary policy; see Yellen (2015) *inter alia*. This paper thus examines the properties of Greenbook forecasts of foreign GDP growth—a key measure of foreign economic activity.

A central focus in forecast evaluation is forecast bias, especially because forecast bias is systematic, and because ignored forecast biases may have substantive adverse consequences for policy. Building on Sinclair, Joutz, and Stekler (2010) and Ericsson, Hood, Joutz, Sinclair, and Stekler (2013), the current paper analyzes Greenbook forecasts of output growth in several foreign countries over 1998–2008. Standard tests typically fail to detect any important forecast biases. However, a recently developed technique—impulse indicator saturation—detects economically large and highly statistically significant time-varying biases. Biases differ across

*The views in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System. The authors are grateful to Shaghil Ahmed, David Hendry, Jun Ma, Ricardo Nunes, Andrea Raffo, John Rogers, and Herman Stekler for helpful discussions and comments. All numerical results were obtained using PcGive Version 14.0B3, Autometrics Version 1.5e, and Ox Professional Version 7.00 in 64-bit OxMetrics Version 7.00: see Doornik and Hendry (2013) and Doornik(2009).

†Division of International Finance, Board of Governors of the Federal Reserve System, Washington, DC 20551 USA (ericsson@frb.gov), and Research Program on Forecasting, Department of Economics, The George Washington University, Washington, DC 20052 USA (ericsson@gwu.edu)

‡Department of Statistics, Rutgers, The State University of New Jersey, New Brunswick, NJ (emilio_f1@yahoo.com)

§Division of International Finance, Board of Governors of the Federal Reserve System, Washington, DC 20551 USA (jedwardseymour@gmail.com)

the country being forecast, the horizon, and the date of the forecast. For example, forecasts of Chinese real GDP growth are systematically biased, with a statistically significant and economically large bias of approximately two percent per annum. For all countries examined, there is little observed predictability beyond two quarters ahead.

This paper is organized as follows. Section 2 describes the data and the forecasts being analyzed. Section 3 discusses different approaches to testing for potential forecast bias and proposes impulse indicator saturation as a generic test of forecast bias. Section 4 describes indicator saturation techniques, including impulse indicator saturation and several of its extensions. Section 5 presents evidence on forecast bias, using the methods detailed in Sections 3 and 4. Section 6 concludes.

2. The Data and the Forecasts

As input to the decision-making process of the Federal Open Market Committee, the staff of the Federal Reserve Board (the "Fed") produce a document called the Greenbook, which includes forecasts of U.S. and foreign economic activity. This section describes the foreign Greenbook forecasts analyzed in this paper and the data being forecast. See Ericsson, Fiallos, and Seymour (2014) for further details.

The data being forecast are real GDP growth rates for nine countries:

- Brazil (BZ),
- Canada (CA),
- China (CH),
- Germany (GE),
- Japan (JA),
- South Korea (KO),
- Mexico (MX),
- the United Kingdom (UK), and
- (for comparison) the United States (US).

Country abbreviations are in parentheses. The sample period is determined by the presence of the forecasts in the publicly available Greenbooks: 1998Q1–2008Q4 for Canada, Germany, Japan, the United Kingdom, and the United States; and 1999Q4–2008Q4 for Brazil, China, Mexico, and South Korea. The Greenbook forecasts are from the final Greenbook of each quarter so as to allow as much information to be available for the forecasts being made in a given quarter. The forecast horizon h (in quarters) is $h = -1, 0, 1, 2, 3, 4$, where $h = -1$ denotes the one-quarter backcast, $h = 0$ denotes the nowcast, and $h = 1, 2, 3, 4$ denote the one-, two-, three-, and four-quarter-ahead forecasts. Output growth is measured in quarterly rates expressed as percent changes at an annual rate. Measured actual values are the GDP growth rates as reported in the Greenbook with a two-quarter lag.

The Greenbooks are publicly available from the Federal Reserve Bank of Philadelphia:

http://www.phil.frb.org/research-and-data/real-time-center/greenbook-data/pdf-data-set.cfm.

These forecasts are made publicly available approximately five years after the fact. The assumptions underlying the Greenbook forecasts, the complex process involved in generating the forecasts, and the goals and objectives of that process are of considerable interest in their own right and merit detailed examination. However, in the spirit of Stekler (1972), Chong and Hendry (1986), and Fildes and Stekler (2002) *inter alia*, the current paper focuses on the properties of the forecasts themselves.

Several properties of the data, the Greenbook forecasts, and the corresponding forecast errors are apparent upon graphing. The Chinese forecasts systematically underpredict actual growth, albeit by different amounts, varying over time. Forecasts for other countries' growth likewise under- or over-predict actual growth, and the degree of inaccuracy depends on the country, the horizon, and the date of the forecast. Forecast errors are often persistent, suggestive of systematic biases in the forecasts. See Ericsson, Fiallos, and Seymour (2014) for further details. For some previous analyses of Greenbook and other governmental and institutional forecasts, see Corder (2005), Engstrom and Kernell (1999), Frankel (2011), Joutz and Stekler (2000), Nunes (2013), Sinclair, Joutz, and Stekler (2010), Romer and Romer (2008), Tsuchiya (2013), and Ericsson, Hood, Joutz, Sinclair, and Stekler (2015).

3. Approaches for Detecting Forecast Bias

This section considers different approaches for assessing potential forecast bias, starting with the standard test of (time-invariant) forecast bias by Mincer and Zarnowitz (1969). This section then considers forms of time-dependent forecast bias, with impulse indicator saturation providing a generic test of potentially time-varying forecast bias. This section's exposition draws on Ericsson (2015) and Ericsson, Hood, Joutz, Sinclair, and Stekler (2015).

Mincer and Zarnowitz (1969, pp. 8–11) suggest testing for forecast bias by regressing the forecast error on an intercept and testing whether the intercept is statistically significant. That is, for a variable x_t at time t and its forecast $\hat{x}_t$, estimate the equation:

$$(x_t - \hat{x}_t) \ = \ a \ + \ e_t \qquad t = 1, \ldots, T, \tag{1}$$

where a is the intercept, e_t is the error term at time t, and T is number of observations. A test of $a = 0$ is interpretable as a test that the forecast $\hat{x}_t$ is unbiased for the variable x_t. For current-period and one-step-ahead forecasts, the error e_t may be serially uncorrelated, in which case a t- or F-statistic may be appropriate. For multi-step-ahead forecasts, e_t generally will be serially correlated; hence inference about the intercept a may require some accounting for that autocorrelation.

Holden and Peel (1990) and Stekler (2002) discuss a generalization of equation (1):

$$(x_t - \hat{x}_t) \ = \ b_0 \ + \ b_1' z_t \ + \ e_t \qquad t = 1, \ldots, T, \tag{2}$$

in which the right-hand side variables z_t might be any variables; and they interpret a test of $b_1 = 0$ as a test of efficiency. See Holden and Peel (1990) and Stekler (2002) for expositions on these tests as tests of unbiasedness and efficiency, and Sinclair, Stekler, and Carnow (2012) for a recent discussion.

Many forecast tests are interpretable as being based on equation (2). For example, in Sinclair, Joutz, and Stekler (2010) and Ericsson, Hood, Joutz, Sinclair, and Stekler (2015), the regressor z_t includes a dummy variable that is indicates the business cycle's phase—either contraction or expansion. Another choice of z_t is $\hat{x}_t$, proposed by Mincer and Zarnowitz (1969, p. 11). In Ericsson (2015) and Ericsson, Hood, Joutz, Sinclair, and Stekler (2015), "Mincer–Zarnowitz A" denotes the regression-based test of $a = 0$ in equation (1), whereas "Mincer–Zarnowitz B" denotes the regression-based test of $\{b_0 = 0, b_1 = 0\}$ in equation (2) with $z_t = \hat{x}_t$. Other choices for z_t include an alternative forecast $\tilde{x}_t$ or the differential between the two forecasts $(\tilde{x}_t - \hat{x}_t)$, generating the forecast-encompassing tests in Chong and Hendry (1986). As Ericsson (1992) discusses, a necessary condition for forecast encompassing is having the smallest mean squared forecast error (MSFE); Granger (1989) and Diebold and Mariano (1995) propose tests of whether one model's MSFE is less than another model's MSFE. Also, the "alternative forecast" could be a forecast made in a different time period, in which case $(\tilde{x}_t - \hat{x}_t)$ is the revision of the forecast. Nordhaus (1987) proposes this test based on forecast revisions across multiple horizons as a test of efficiency. Tversky and Kahneman (1974) earlier described "anchoring" as a phenomenon in which $b_1 > 0$ for forecast revisions; see Campbell and Sharpe (2009) for empirical evidence on anchoring.

In equation (2), the term $(b_0 + b_1' z_t)$ is also interpretable as a specific form of time-dependent forecast bias. That time dependence could be completely general, as follows:

$$\begin{aligned}(x_t - \hat{x}_t) &= a_t + e_t \\ &= \textstyle\sum_{i=1}^{T} c_i I_{it} + e_t \qquad t = 1, \ldots, T,\end{aligned} \tag{3}$$

where the impulse indicator I_{it} is a dummy variable that is unity for $t = i$ and zero otherwise, and c_i is the corresponding coefficient for I_{it}. Because the $\{c_i\}$ may have any values whatsoever, the intercept a_t in (3) may vary arbitrarily over time. In this context, a test that all coefficients c_i are equal to zero is a generic test of forecast unbiasedness. Because equation (3) includes T coefficients, equation (3) cannot be estimated unrestrictedly. However, the question being asked can be answered using impulse indicator saturation, as summarized in Section 4.

4. Indicator Saturation Techniques

Impulse indicator saturation (IIS) uses the zero-one dummies $\{I_{it}\}$ to analyze properties of a model. Unrestricted inclusion of all T dummies in the model (thereby "saturating" the sample) is infeasible. However, blocks of dummies *can* be included, and statistically significant dummies can be retained from those blocks. That insight provides the basis for IIS. See Ericsson and Reisman (2012) for an intuitive non-technical exposition of IIS, and Hendry and Doornik (2014) for extensive analysis in the context of automatic model selection. This section's exposition draws on Ericsson (2015) and Ericsson, Hood, Joutz, Sinclair, and Stekler (2015).

IIS provides a general procedure for robust estimation and for model evaluation—in particular, for testing parameter constancy. IIS is a generic test for an unknown number of structural breaks, occurring at unknown times, with unknown duration and magnitude, anywhere in the sample. IIS is a powerful empirical tool for both evaluating and improving existing empirical models. Hendry (1999) proposes IIS as a procedure for testing parameter constancy. Further discussion, recent developments, and applications appear in Hendry, Johansen, and Santos (2008), Doornik (2009), Johansen and Nielsen (2009, 2013, 2015), Hendry and Santos (2010), Ericsson (2011a, 2011b, 2012), Ericsson and Reisman (2012), Bergamelli and Urga (2013), Hendry and Pretis (2013), Hendry and Doornik (2014), Pretis, Mann, and Kaufmann (2015), and Castle, Doornik, Hendry, and Pretis (2015). Ericsson (2015) proposes a new application for IIS—as a generic test for time-varying forecast bias. Section 5 applies IIS to test for potential bias in the Greenbook forecasts.

Many existing procedures can be interpreted as "special cases" of IIS in that they represent particular algorithmic implementations of IIS. Such special cases include recursive estimation, rolling regression, the Chow (1960) predictive failure statistic (including the 1-step, breakpoint, and forecast versions implemented

Table 1: Impulse indicator saturation and super saturation, as characterized by the variables involved.

Name	Description	Variables	Definition of variables
Impulse indicator saturation	Zero-one dummies	$\{I_{it}\}$	$I_{it} = 1$ for $t = i$, zero otherwise
Super saturation	Step functions	$\{I_{it}, S_{it}\}$	$S_{it} = 1$ for $t \geq i$, zero otherwise

in OxMetrics), the Andrews (1993) unknown breakpoint test, the Bai and Perron (1998) multiple breakpoint test, tests of extended constancy in Ericsson, Hendry, and Prestwich (1998, pp. 305ff), tests of nonlinearity, intercept correction (in forecasting), and robust estimation. IIS thus provides a general and generic procedure for analyzing a model's constancy. Algorithmically, IIS also solves the problem of having more potential regressors than observations by testing and selecting over blocks of variables.

Table 1 summarizes IIS and an extension: super saturation. Throughout, T is the sample size, t is the index for time, i and j are the indexes for indicators, k is the index for economic variables (denoted x_{kt}), and K is the total number of potential regressors considered. A few remarks may be helpful for interpreting the entries in Table 1.

Impulse indicator saturation. This is the standard IIS procedure proposed by Hendry (1999), with selection among the T zero-one impulse indicators $\{I_{it}\}$.

Super saturation. Super saturation searches across all possible one-off step functions $\{S_{it}\}$, in addition to $\{I_{it}\}$. Step functions are of economic interest because they may capture permanent or long-lasting changes that are not otherwise incorporated into a specific empirical model. A step function is a partial sum of impulse indicators; equivalently, it is a parsimonious representation of a sequential subset of impulse indicators that have equal coefficients. Castle, Doornik, Hendry, and Pretis (2015) investigate the statistical properties of a closely related saturation estimator—step indicator saturation (SIS)—which searches among only the step indicator variables $\{S_{it}\}$. Autometrics now includes IIS, SIS, super saturation (IIS+SIS), and zero-sum pairwise IIS (mentioned below); see Doornik and Hendry (2013).

Table 1 is by no means an exhaustive list of extensions to IIS. One direct extension is ultra saturation, with searches across $\{I_{it}, S_{it}, T_{it}\}$, where the $\{T_{it}\}$ are broken linear trends. Broken quadratic trends, broken cubic trends, and higher-order broken trends are also feasible. Other extensions include sequential ($j = 1$) and non-sequential ($j > 1$) pairwise impulse indicator saturation for an indicator P_{it}, defined as $I_{it}+I_{i+j,t}$; sequential multiplet indicator saturation for an indicator M_{it}^{j+1}, defined as $I_{it}+\cdots+I_{i+j,t}$ for $j \geq 1$; zero-sum pairwise IIS for an indicator Z_{it}, defined as ΔI_{it}; many many variables for a set of K potential regressors $\{x_{kt},\ k = 1, \ldots, K\}$

for $K > T$; factors; principal components; and multiplicative indicator saturation for the set of $S_{it}x_{kt}$. See Castle, Clements, and Hendry (2013) and Ericsson (2011b, 2012) for details, discussion, and examples in the literature. Also, the IIS-type procedure chosen may itself be a combination of extensions; and that choice may affect the power of the procedure to detect specific alternatives. Notably, dummies for economic expansions and contractions are examples of sequential multiplets.

As a more general observation, different types of indicators are adept at characterizing different sorts of bias: impulse dummies $\{I_{it}\}$ for date-specific anomalies, step dummies $\{S_{it}\}$ for level shifts, and broken trends $\{T_{it}\}$ for evolving developments. Transformations of the variable being forecast also may affect the interpretation of the retained indicators. For instance, an impulse dummy for a growth rate implies a level shift for the (log) level of the variable.

IIS-based tests of forecast bias can serve both as diagnostic tools to detect what is wrong with the forecasts, and as developmental tools to suggest how the forecasts can be improved. Clearly, "rejection of the null doesn't imply the alternative". However, for time series data, the date-specific nature of IIS-type procedures can aid in identifying important sources of forecast error. Use of these tests in forecast development is consistent with a progressive modeling approach; see White (1990).

As equation (3) emphasizes, IIS-based tests generalize the Mincer–Zarnowitz tests to allow for arbitrarily time-varying forecast bias. This observation and the observations above highlight the strength of the Mincer–Zarnowitz tests (that they focus on detecting a constant nonzero forecast bias) and also their weakness (that they assume that the forecast bias *is* constant over time). These characteristics of the Mincer–Zarnowitz tests bear directly on the empirical results in Section 5.

5. Evidence on Biases in the Greenbook Forecasts

This section examines the Greenbook forecasts of output growth for eight foreign countries and for the United States. Standard (Mincer–Zarnowitz) tests of forecast bias typically fail to detect economically and statistically important biases. By contrast, IIS-type tests detect large time-varying biases. Forecast biases differ numerically across the forecast horizon, country being forecast, and the date of the forecast, albeit with some qualitative similarities.

Section 5.1 reports a standard summary statistic on forecast performance: root mean squared forecast errors. Section 5.2 reports standard Mincer–Zarnowitz tests of forecast bias. Section 5.3 employs IIS-type procedures to test for and estimate time-varying forecast bias.

5.1 Summary Statistics of Forecast Performance

Figure 1 plots the root mean squared forecast errors (RMSEs) for the nine countries as a function of the forecast horizon h. The RMSEs for four developed countries (Germany, Canada, the United Kingdom, and the United States) are considerably smaller at every horizon than the RMSEs for the remaining countries. Notably, the remaining countries include Japan—a developed country—and all of the emerging market economies analyzed (Brazil, Mexico, South Korea, and China). For all

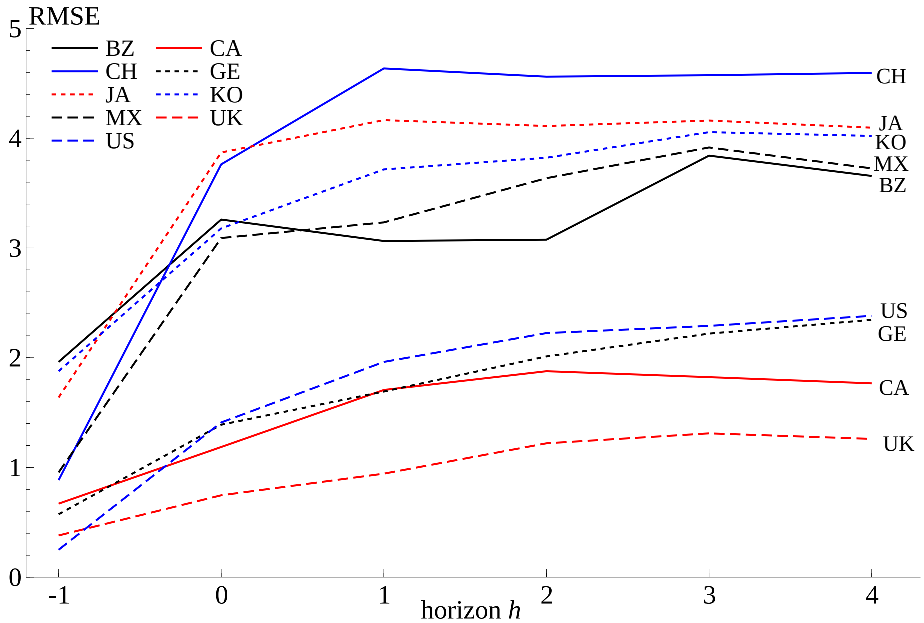

Figure 1: RMSEs of the Greenbook forecasts for GDP growth of nine countries at different forecast horizons h ($h = -1, 0, 1 \ldots, 4$).

countries, the RMSEs generally increase with the forecast horizon; and the RMSEs increase little beyond a horizon of two quarters ahead.

5.2 Standard Tests of Forecast Bias

This subsection examines the Greenbook forecasts for bias using the standard (Mincer–Zarnowitz) test. With the exception of China, the Mincer–Zarnowitz test finds little evidence of economically and statistically important biases.

Table 2 reports estimated intercepts and estimated standard errors for the Mincer–Zarnowitz regression in equation (1). Here and in Table 3, HAC estimated standard errors appear under regression coefficients in square brackets $[\cdot]$. The symbols $^+$, *, and ** respectively denote significance at the 10%, 5%, and 1% levels. Statistically, there is little evidence of forecast bias, except for Chinese output growth, for which the estimated bias is 1.5%–2.0% per annum for the nowcast and the forecasts.

5.3 Estimated Time-varying Bias Using Indicator Saturation

To assess possible time dependence of the forecast biases, this subsection estimates IIS-type equations in the form of equation (3). Time dependence is detected for all countries at some or all forecast horizons.

Table 3 reports estimated intercepts and estimated standard errors for the Mincer–Zarnowitz regression in equation (3); and Table 4 lists the impulse and step dummies retained from super saturation with a target size of 0.5%. Tables 3 and 4 highlight the dependence of bias on the country and on the forecast hori-

Table 2: Estimated intercepts and HAC standard errors for the Mincer–Zarnowitz test of bias in Greenbook forecasts for GDP growth.

Country	Forecast horizon h					
	−1	0	1	2	3	4
Brazil	0.51	0.61	0.63	0.39	0.55	0.44
	[0.27]	[0.57]	[0.56]	[0.60]	[0.77]	[0.77]
Canada	−0.04	0.17	0.01	−0.09	−0.23	−0.35
	[0.10]	[0.22]	[0.32]	[0.37]	[0.38]	[0.36]
China	0.01	1.96**	1.80**	1.73**	1.60**	1.82**
	[0.14]	[0.34]	[0.42]	[0.45]	[0.56]	[0.47]
Germany	0.10	−0.05	−0.16	−0.44	−0.62	−0.75
	[0.09]	[0.17]	[0.26]	[0.37]	[0.44]	[0.48]
Japan	−0.42^{+}	0.47	0.18	−0.07	−0.18	−0.32
	[0.24]	[0.52]	[0.62]	[0.65]	[0.68]	[0.67]
Korea	0.66*	0.70	0.49	0.20	0.11	−0.03
	[0.25]	[0.51]	[0.67]	[0.72]	[0.74]	[0.73]
Mexico	−0.03	−0.30	−0.64	−1.10	−1.21	−1.36^{+}
	[0.16]	[0.48]	[0.56]	[0.67]	[0.77]	[0.75]
United Kingdom	0.10^{+}	0.01	−0.16	−0.19	−0.23	−0.28
	[0.05]	[0.11]	[0.15]	[0.23]	[0.26]	[0.25]
United States	−0.06*	0.56**	0.39	0.14	−0.01	−0.06
	[0.03]	[0.18]	[0.30]	[0.45]	[0.50]	[0.51]

Table 3: IIS-estimated intercepts and HAC standard errors for the Mincer–Zarnowitz test of bias in the Greenbook forecasts for GDP growth.

Country	Forecast horizon h					
	−1	0	1	2	3	4
Brazil	0.14	0.61	0.64	0.39	0.15	0.16
	[0.11]	[0.57]	[0.56]	[0.60]	[0.61]	[0.68]
Canada	0.05	0.00	−0.13	−0.34	−0.44	−0.26
	[0.05]	[0.17]	[0.29]	[0.32]	[0.28]	[0.24]
China	0	1.52**	1.81**	1.78**	1.62**	1.86**
	[0]	[0.31]	[0.46]	[0.49]	[0.59]	[0.49]
Germany	0.06	−0.22	−0.16	−0.44	−0.62	−0.75
	[0.05]	[0.18]	[0.26]	[0.37]	[0.44]	[0.48]
Japan	−0.11	0.47	−0.08	−0.33	−0.46	−0.59
	[0.16]	[0.52]	[0.62]	[0.67]	[0.71]	[0.71]
Korea	0.02	0.70	0.49	0.96	0.11	−0.03
	[0.02]	[0.51]	[0.67]	[0.65]	[0.74]	[0.73]
Mexico	0.13	−0.54	−0.89^{+}	−1.36*	−1.53*	−1.64*
	[0.13]	[0.44]	[0.52]	[0.62]	[0.68]	[0.68]
United Kingdom	0.06	0.01	−0.16	−0.10	−0.20	−0.12
	[0.05]	[0.12]	[0.15]	[0.20]	[0.20]	[0.20]
United States	−0.00	0.40*	0.39	0.14	−0.01	−0.06
	[0.03]	[0.18]	[0.30]	[0.45]	[0.50]	[0.51]

Table 4: Dummy variables selected by super saturation at the 0.1% level at at least one forecast horizon, by country.

Country	Indicators selected	
	Impulse	Step
Brazil	1999(4)	2000(3), 2001(3)
Canada	1998(2), 2002(1)	1998(3), 2000(1), 2000(3), 2001(3), 2001(4), 2002(1), 2002(2)
China	2003(2), 2003(3)	2003(1), 2003(2), 2003(3)
Germany	—	2006(2), 2007(4), 2008(1)
Japan	2001(2)	1998(4), 1999(2), 1999(4), 2000(1), 2001(4), 2008(1)
Korea	2000(4)	2000(3), 2000(4), 2001(3), 2002(4), 2003(2)
Mexico	2000(1)	2000(3), 2001(3), 2001(4), 2002(1), 2003(1)
United Kingdom	1999(3) 2000(4)	2002(4), 2003(1), 2003(4), 2007(4), 2008(1), 2008(2)
United States	2003(3)	2000(1), 2000(2), 2001(3), 2001(4), 2002(1), 2003(2), 2003(3)

zon h. The selected dummies in Table 4 indicate the pervasiveness of time-varying bias. Because forecast errors tend to be very small for backcasts ($h = -1$), only non-negative forecast horizons are considered in Table 4.

Graphs directly convey a sense of the magnitude and extent of the biases present. Figures 2–5 thus plot actual values and forecasts and the forecast errors for two of the countries analyzed: China and the United States. Each figure is a panel of 2×3 for forecast horizons h ($h = -1, 0, 1 \ldots, 4$).

Figure 2 plots the actual and forecast values for Chinese growth; and Figure 3 plots the corresponding forecast errors and the bias as estimated from super saturation. Large biases are evident for all horizons except $h = -1$, and the biases appear somewhat different before and after 2003.

Figure 4 plots the actual and forecast values for U.S. growth; and Figure 5 plots the corresponding forecast errors and the bias as estimated from super saturation. The biases are notably time-dependent and persistent at all non-negative horizons. Ericsson, Hood, Joutz, Sinclair, and Stekler (2015) show that those biases depend primarily on the phase of the business cycle.

Forecast biases vary markedly over time, being sometimes positive and othertimes negative. The Mincer–Zarnowitz tests have particular difficulty in detecting such biases because the Mincer–Zarnowitz tests average all biases (both negative and positive) over time, and because the Mincer–Zarnowitz tests assign any time variation in bias to the residual rather than to the bias itself. As an extreme example, the Mincer–Zarnowitz A test has no power to detect a forecast bias that is +10% for the first half of the sample and −10% for the second half of the sample, even though this bias would be obvious from (e.g.) graphing the data. Super saturation often detects time-varying bias, and for historically and economically consequential years. The dates of the retained dummies are important and informative, and those dummies often appear to reflect cyclical movements.

6. Conclusions

Building on Sinclair, Joutz, and Stekler (2010), the current paper analyzes Greenbook forecasts of foreign output growth for potential biases over 1999–2008. Standard tests typically fail to detect bias. However, super saturation detects economically large and highly statistically significant time-dependent biases across all countries being forecast. Biases depend on the country, the forecast horizon, and the date of the forecast. Saturation as a technique defines a generic procedure for examining forecast properties; it explains why standard tests fail to detect bias; and it provides a potential mechanism for improving forecasts. In particular, such biases imply an opportunity to robustify the forecasts, as with intercept correction; see Clements and Hendry (1999, 2002), Hendry (2006), and Castle, Fawcett, and Hendry (2010).

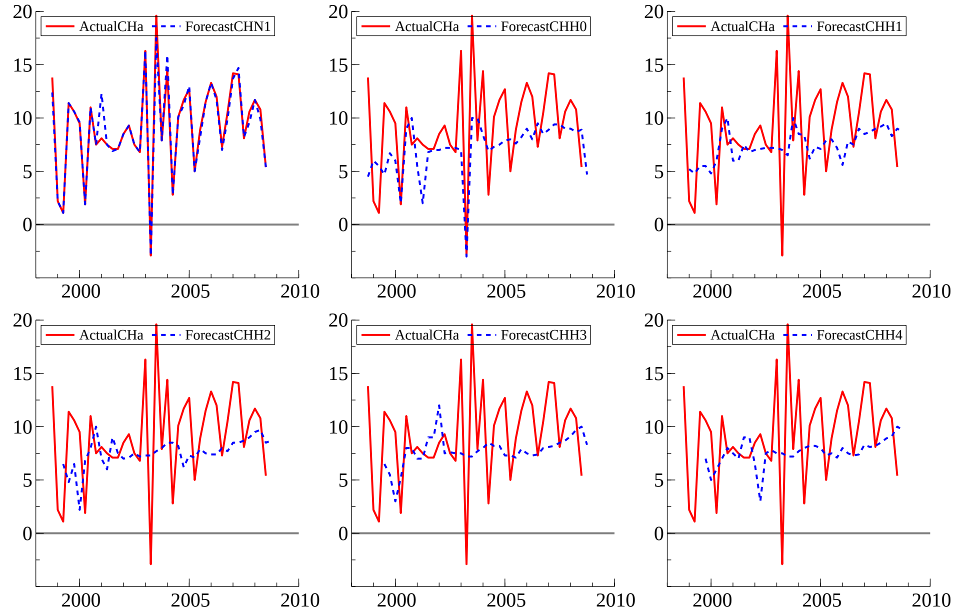

Figure 2: Chinese GDP growth and its Greenbook forecasts at different forecast horizons h $(h = -1, 0, 1 \ldots, 4)$.

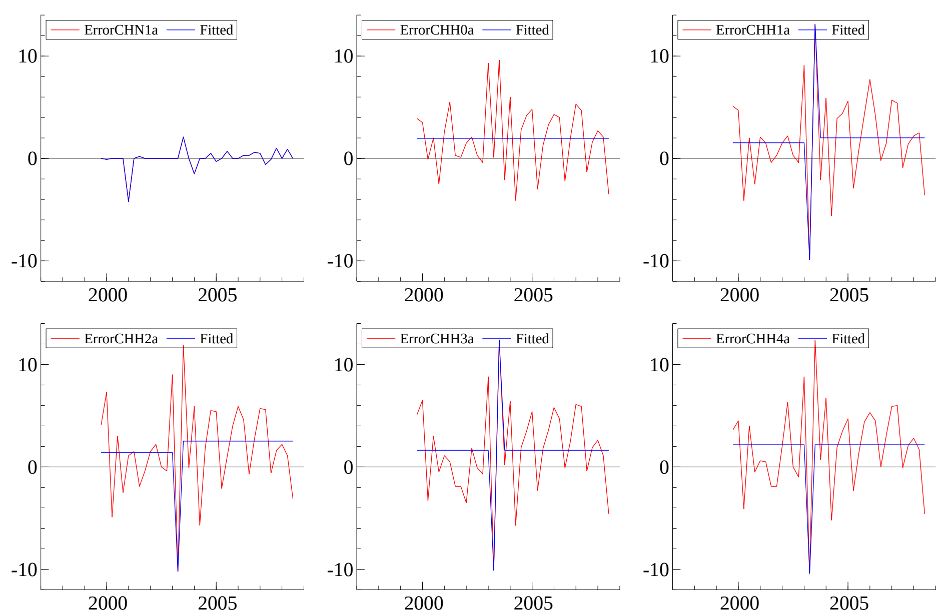

Figure 3: Greenbook forecast errors for Chinese GDP growth at different forecast horizons h $(h = -1, 0, 1 \ldots, 4)$, and estimated forecast biases as calculated using super saturation.

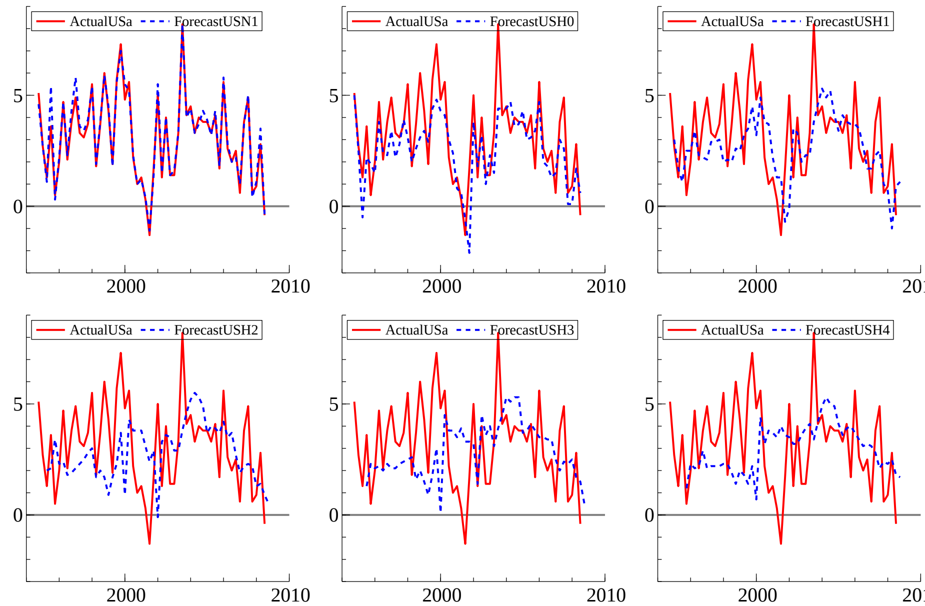

Figure 4: US GDP growth and its Greenbook forecasts at different forecast horizons h ($h = -1, 0, 1 \ldots, 4$).

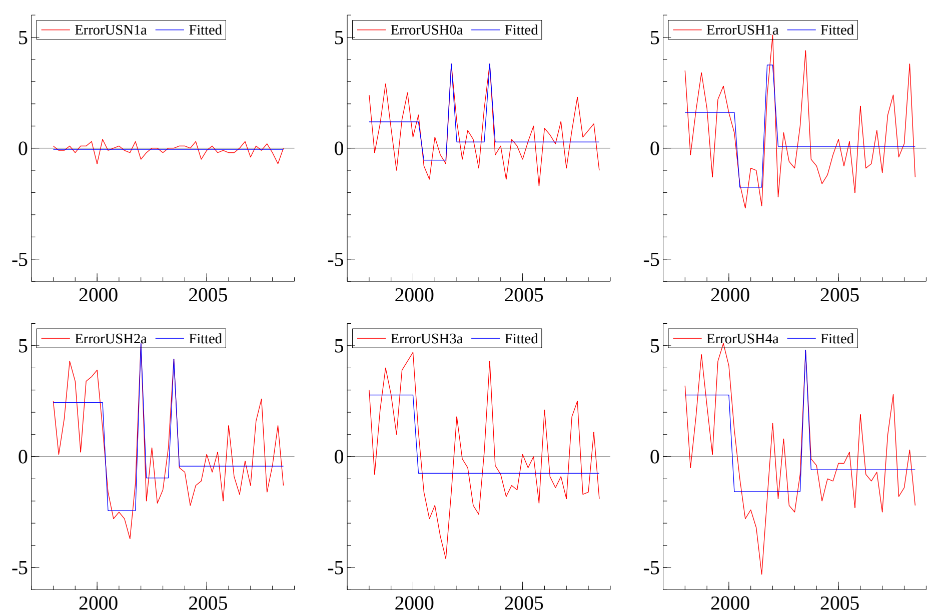

Figure 5: Greenbook forecast errors for US GDP growth at different forecast horizons h ($h = -1, 0, 1 \ldots, 4$), and estimated forecast biases as calculated using super saturation.

REFERENCES

Andrews, D. W. K. (1993) “Tests for Parameter Instability and Structural Change with Unknown Change Point”, *Econometrica*, 61, 4, 821–856.

Bai, J., and P. Perron (1998) “Estimating and Testing Linear Models with Multiple Structural Changes”, *Econometrica*, 66, 1, 47–78.

Bergamelli, M., and G. Urga (2013) “Detecting Multiple Structural Breaks: A Monte Carlo Study and an Application to the Fisher Equation for the US”, draft, Cass Business School, London, March.

Bernanke, B. S. (2012) “U.S. Monetary Policy and International Implications”, remarks at the seminar “Challenges of the Global Financial System: Risks and Governance under Evolving Globalization”, Bank of Japan, Tokyo, Japan, October 14.

Campbell, S. D., and S. A. Sharpe (2009) “Anchoring Bias in Consensus Forecasts and Its Effect on Market Prices”, *Journal of Financial and Quantitative Analysis*, 44, 2, 369–390.

Castle, J. L., M. P. Clements, and D. F. Hendry (2013) “Forecasting by Factors, by Variables, by Both or Neither?”, *Journal of Econometrics*, 177, 2, 305–319.

Castle, J. L., J. A. Doornik, D. F. Hendry, and F. Pretis (2015) “Detecting Location Shifts During Model Selection by Step-indicator Saturation”, *Econometrics*, 3, 2, 240–264.

Castle, J. L., N. W. P. Fawcett, and D. F. Hendry (2010) “Forecasting with Equilibrium-correction Models During Structural Breaks”, *Journal of Econometrics*, 158, 1, 25–36.

Chong, Y. Y., and D. F. Hendry (1986) “Econometric Evaluation of Linear Macro-economic Models”, *Review of Economic Studies*, 53, 4, 671–690.

Chow, G. C. (1960) “Tests of Equality Between Sets of Coefficients in Two Linear Regressions”, *Econometrica*, 28, 3, 591–605.

Clements, M. P., and D. F. Hendry (1999) *Forecasting Non-stationary Economic Time Series*, MIT Press, Cambridge.

Clements, M. P., and D. F. Hendry (2002) “Explaining Forecast Failure in Macroeconomics”, Chapter 23 in M. P. Clements and D. F. Hendry (eds.) *A Companion to Economic Forecasting*, Blackwell Publishers, Oxford, 539–571.

Corder, J. K. (2005) “Managing Uncertainty: The Bias and Efficiency of Federal Macroeconomic Forecasts”, *Journal of Public Administration Research and Theory*, 15, 1, 55–70.

Diebold, F. X., and R. S. Mariano (1995) “Comparing Predictive Accuracy”, *Journal of Business and Economic Statistics*, 13, 3, 253–263.

Doornik, J. A. (2009) “Autometrics”, Chapter 4 in J. L. Castle and N. Shephard (eds.) *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, Oxford University Press, Oxford, 88–121.

Doornik, J. A., and D. F. Hendry (2013) *PcGive 14*, Timberlake Consultants Press, London (3 volumes).

Engstrom, E. J., and S. Kernell (1999) “Serving Competing Principals: The Budget Estimates of OMB and CBO in an Era of Divided Government”, *Presidential Studies Quarterly*, 29, 4, 820–829.

Ericsson, N. R. (1992) “Parameter Constancy, Mean Square Forecast Errors, and Measuring Forecast Performance: An Exposition, Extensions, and Illustration”, *Journal of Policy Modeling*, 14, 4, 465–495.

Ericsson, N. R. (2011a) “Improving Global Vector Autoregressions”, draft, Board of Governors of the Federal Reserve System, Washington, D.C., June.

Ericsson, N. R. (2011b) “Justifying Empirical Macro-econometric Evidence in Practice”, invited presentation, online conference *Communications with Economists: Current and Future Trends* commemorating the 25th anniversary of the *Journal of Economic Surveys*, November.

Ericsson, N. R. (2012) “Detecting Crises, Jumps, and Changes in Regime”, draft, Board of Governors of the Federal Reserve System, Washington, D.C., November.

Ericsson, N. R. (2015) “How Biased Are U.S. Government Forecasts of the Federal Debt?”, *International Journal of Forecasting*, forthcoming.

Ericsson, N. R., E. J. Fiallos, and J. E. Seymour (2014) "Assessing Greenbook Forecasts of Foreign GDP Growth", draft, Board of Governors of the Federal Reserve System, Washington, D.C., August.

Ericsson, N. R., D. F. Hendry, and K. M. Prestwich (1998) "The Demand for Broad Money in the United Kingdom, 1878–1993", *Scandinavian Journal of Economics*, 100, 1, 289–324 (with discussion).

Ericsson, N. R., S. B. Hood, F. Joutz, T. M. Sinclair, and H. O. Stekler (2013) "Greenbook Forecasts and the Business Cycle", draft, Board of Governors of the Federal Reserve System, Washington, D.C., December.

Ericsson, N. R., S. B. Hood, F. Joutz, T. M. Sinclair, and H. O. Stekler (2015) "Time-dependent Bias in the Fed's Greenbook Forecasts", in *JSM Proceedings,* Business and Economic Statistics Section, American Statistical Association, Alexandria, VA, forthcoming.

Ericsson, N. R., and E. L. Reisman (2012) "Evaluating a Global Vector Autoregression for Forecasting", *International Advances in Economic Research*, 18, 3, 247–258.

Fildes, R., and H. O. Stekler (2002) "The State of Macroeconomic Forecasting", *Journal of Macroeconomics*, 24, 4, 435–468.

Frankel, J. (2011) "Over-optimism in Forecasts by Official Budget Agencies and Its Implications", *Oxford Review of Economic Policy*, 27, 4, 536–562.

Granger, C. W. J. (1989) *Forecasting in Business and Economics*, Academic Press, Boston, Massachusetts, Second Edition.

Hendry, D. F. (1999) "An Econometric Analysis of US Food Expenditure, 1931–1989", Chapter 17 in J. R. Magnus and M. S. Morgan (eds.) *Methodology and Tacit Knowledge: Two Experiments in Econometrics*, John Wiley and Sons, Chichester, 341–361.

Hendry, D. F. (2006) "Robustifying Forecasts from Equilibrium-correction Systems", *Journal of Econometrics*, 135, 1–2, 399–426.

Hendry, D. F., and J. A. Doornik (2014) *Empirical Model Discovery and Theory Evaluation: Automatic Selection Methods in Econometrics*, MIT Press, Cambridge, Massachusetts.

Hendry, D. F., S. Johansen, and C. Santos (2008) "Automatic Selection of Indicators in a Fully Saturated Regression", *Computational Statistics*, 23, 2, 317–335, 337–339.

Hendry, D. F., and F. Pretis (2013) "Anthropogenic Influences on Atmospheric CO_2", Chapter 12 in R. Fouquet (ed.) *Handbook on Energy and Climate Change*, Edward Elgar, Cheltenham, 287–326.

Hendry, D. F., and C. Santos (2010) "An Automatic Test of Super Exogeneity", Chapter 12 in T. Bollerslev, J. R. Russell, and M. W. Watson (eds.) *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*, Oxford University Press, Oxford, 164–193.

Holden, K., and D. A. Peel (1990) "On Testing for Unbiasedness and Efficiency of Forecasts", *The Manchester School*, 58, 2, 120–127.

Johansen, S., and B. Nielsen (2009) "An Analysis of the Indicator Saturation Estimator as a Robust Regression Estimator", Chapter 1 in J. L. Castle and N. Shephard (eds.) *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, Oxford University Press, Oxford, 1–36.

Johansen, S., and B. Nielsen (2013) "Outlier Detection in Regression Using an Iterated One-step Approximation to the Huber-skip Estimator", *Econometrics*, 1, 1, 53–70.

Johansen, S., and B. Nielsen (2015) "Asymptotic Theory of Outlier Detection Algorithms for Linear Time Series Regression Models", *Scandinavian Journal of Statistics*, in press.

Joutz, F., and H. O. Stekler (2000) "An Evaluation of the Predictions of the Federal Reserve", *International Journal of Forecasting*, 16, 1, 17–38.

Mincer, J., and V. Zarnowitz (1969) "The Evaluation of Economic Forecasts", Chapter 1 in J. Mincer (ed.) *Economic Forecasts and Expectations: Analyes of Forecasting Behavior and Performance*, National Bureau of Economic Research, New York, 3–46.

Nordhaus, W. D. (1987) "Forecasting Efficiency: Concepts and Applications", *Review of Economics and Statistics*, 69, 4, 667–674.

Nunes, R. (2013) "Do Central Banks' Forecasts Take Into Account Public Opinion and Views?", International Finance Discussion Paper No. 1080, Board of Governors of the Federal Reserve System, Washington, D.C., May.

Pretis, F., M. L. Mann, and R. K. Kaufmann (2015) "Testing Competing Models of the Temperature Hiatus: Assessing the Effects of Conditioning Variables and Temporal Uncertainties Through Sample-wide Break Detection", *Climatic Change*, 131, 4, 705–718.

Romer, C. D., and D. H. Romer (2008) "The FOMC versus the Staff: Where Can Monetary Policymakers Add Value?", *American Economic Review*, 98, 2, 230–235.

Sinclair, T. M., F. Joutz, and H. O. Stekler (2010) "Can the Fed Predict the State of the Economy?", *Economics Letters*, 108, 1, 28–32.

Sinclair, T. M., H. O. Stekler, and W. Carnow (2012) "A New Approach for Evaluating Economic Forecasts", *Economics Bulletin*, 32, 3, 2332–2342.

Stekler, H. O. (1972) "An Analysis of Turning Point Forecasts", *American Economic Review*, 62, 4, 724–729.

Stekler, H. O. (2002) "The Rationality and Efficiency of Individuals' Forecasts", Chapter 10 in M. P. Clements and D. F. Hendry (eds.) *A Companion to Economic Forecasting*, Blackwell Publishers, Oxford, 222–240.

Tsuchiya, Y. (2013) "Are Government and IMF Forecasts Useful? An Application of a New Market-timing Test", *Economics Letters*, 118, 1, 118–120.

Tversky, A., and D. Kahneman (1974) "Judgment under Uncertainty: Heuristics and Biases", *Science*, 185, 4157, 1124–1131.

White, H. (1990) "A Consistent Model Selection Procedure Based on m-testing", Chapter 16 in C. W. J. Granger (ed.) *Modelling Economic Series: Readings in Econometric Methodology*, Oxford University Press, Oxford, 369–383.

Yellen, J. L. (2012) "Perspectives on Monetary Policy", remarks at the Boston Economic Club Dinner, Federal Reserve Bank of Boston, Boston, Massachusetts, June 6.

Yellen, J. L. (2015) "FOMC Press Conference", transcript, Federal Reserve Board, Washington, D.C., Septeber 17.