Evaluating Government Budget Forecasts*

Neil R. Ericsson† Andrew B. Martinez‡

Abstract
This paper reviews the literature on the evaluation of government budget forecasts, outlines a generic framework for forecast evaluation, and illustrates forecast evaluation with empirical analyses of different U.S. government agencies' forecasts of U.S. federal debt. Techniques for forecast evaluation include comparison of mean squared forecast errors, forecast encompassing, tests of predictive failure, and tests of bias and efficiency. Recent extensions of these techniques utilize machine-learning algorithms to handle more potential regressors than observations, a characteristic common to big data. These techniques are generally applicable, including to forecasts of components of the government budget, to forecasts of budgets from municipal, state, provincial, and national governments, and to other economic and non-economic forecasts. Evaluation of forecasts is fundamental to assessing the forecasts' usefulness; and evaluation can indicate ways in which the forecasts may be improved.

Key Words: bias, big data, budget, debt, efficiency, evaluation, forecast encompassing, forecasts, government forecasts, machine learning, MSFE, projections, RMSE, saturation

1. Introduction

Government budgets have attracted considerable attention, especially with federal debt limits, sequestration, and federal government shut-downs in the United States and with continuing discussions about national debt limits in the euro area. Because future outcomes of government revenues and expenditures are unknown, their forecasts may matter in government policy. It is thus of interest to ascertain how good those forecasts are and how they might be improved. Many tools are available for forecast evaluation, including forecast comparisons, tests of predictive failure, and tests of bias and efficiency. The current paper:

- summarizes the literature on the evaluation of forecasts of the government budget;
- systematically reviews tools for forecast evaluation, empirically illustrating each with different U.S. government agencies' one-year-ahead forecasts of the U.S. gross federal debt over 1984–2018; and

*The views in this paper are solely the responsibility of the authors and should not be interpreted as necessarily representing the views of the Board of Governors of the Federal Reserve System, the US Department of the Treasury, or of any other person associated with the Federal Reserve System, or the US government. We are grateful to Jennifer Castle, Mike Clements, David F. Hendry, Jaime Marquez, and Dan Williams for helpful comments and discussions. All numerical results were obtained using PcGive Version 14.1, Autometrics Version 1.5g, and Ox Professional Version 7.10 in 64-bit OxMetrics Version 7.10: see Doornik and Hendry (2013) and Doornik (2009). An earlier version of this paper appeared as Ericsson and Martinez (2019).

†Division of International Finance, Board of Governors of the Federal Reserve System, Washington, DC 20551 USA (ericsson@frb.gov); H. O. Stekler Research Program on Forecasting, Department of Economics, The George Washington University, Washington, DC 20052 USA; and Paul H. Nitze School of Advanced International Studies (SAIS), Johns Hopkins University, Washington, DC 20036 USA

‡Office of Macroeconomic Analysis, US Department of the Treasury, Washington, DC 20220 USA (andrew.martinez@treasury.gov); and H. O. Stekler Research Program on Forecasting, Department of Economics, The George Washington University, Washington, DC 20052 USA

- develops a generic framework for forecast evaluation, drawing on expositions in Clements and Hendry (1998, 1999), Ericsson and Marquez (1998), Martinez (2015), and Ericsson (2017a) *inter alia*.

This paper is organized as follows. Section 2 briefly reviews the literature on forecasts of the government budget. Section 3 describes the data and forecasts used in the empirical illustrations. Section 4 considers various methods for comparing alternative forecasts. Section 5 discusses different approaches to testing for forecast failure, including subsample tests, tests for bias and efficiency, and generalizations thereof. Drawing on the expositions in Sections 4 and 5, Section 6 proposes a unified approach to forecast evaluation. Section 7 draws out some implications, and Section 8 concludes. The Appendix lists the data and one-year-ahead forecasts analyzed.

Although this paper focuses on forecast evaluation *per se*, it is important to highlight that evaluation also provides a promising basis for forecast improvement. Identifying a forecast's weaknesses is key to its improvement. Typically, those weaknesses are not known *ex ante*, so a panoply of evaluation tools is desirable because different evaluation tools have varying power to detect different shortcomings in the forecasts.

2. Literature Review

A large body of literature evaluates government budget forecasts. The current section focuses on forecasts from U.S. federal budget agencies and includes forecasts of the budget and of other economic variables.

Existing studies can be divided into two types. The first type compares agencies' forecasts by looking at the forecast errors directly or by summarizing the forecasts' properties with statistics such as the root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percent error (MAPE). The second type of study compares different forecasts through regression analysis and regression-based tests. Both types of studies can provide valuable information about the forecasts.

Studies of the first type date back to at least Kamlet, Mowery, and Su (1987), who use measures of bias and MAPE to evaluate forecasts of economic variables from their own ARIMA models and from the Congressional Budget Office (CBO), Office of Management and Budget (OMB), and the ASA/NBER survey. McNees (1995) uses MAEs and RMSEs to assess forecasts of economic variables from the Federal Reserve Board (FRB), the CBO, the Council of Economic Advisors (CEA), and several private forecasters. Frendreis and Tatalovich (2000) examine forecast bias in CBO, OMB, and FRB forecasts of economic variables. The CBO also conducts a semi-annual comparison of the bias, MAE, and RMSE for its own economic forecasts and those of the OMB and Blue Chip Consensus; see (e.g.) CBO (2017a). Additionally, CBO (2015b) conducts a similar evaluation of revenue forecasts by the CBO and the OMB.

Studies of the second type use regressions to evaluate and compare the forecasts. For example, Howard (1987) regresses the OMB's forecast errors on the CBO's forecast errors to understand how the two forecasts are related to one another. Belongia (1988) regresses the actual growth rate of economic variables on the growth rates predicted by the CBO, CEA, and private-sector sources in order to assess which forecast outperforms the others. Cohen and Follette (2003) regress the actual budget deficit on forecasts by the OMB, CBO, and FRB in order to determine which forecast

Table 1: Some studies evaluating U.S. federal budget agency forecasts, as characterized by forecaster, forecast horizon, variable forecast, and forecast period.

Study	Forecaster			Forecast	Variable forecast					Forecast
	CBO	OMB	Other	horizon	B	Y	Δp	U	X	period
CBO (various)	•	•	•	2, 5		•	•		•	various
Kamlet, Mowery, and Su (1987)	•	•	•	1–6		•	•	•		1962–1985
Howard (1987)	•	•		1		•	•	•	•	1976–1985
Plesko (1988)	•	•		1–5	•	•				1974–1988
Belongia (1988)	•		•	1		•	•	•		1976–1987
Miller (1991)	•			1	•	•				1980–1987
Blackley and DeBoer (1993)		•		1	•	•	•	•	•	1963–1989
Auerbach (1995)		•		1	•					1982–1993
Campbell and Ghysels (1995)		•		1	•					1969–1990
McNees (1995)	•		•	1–4		•	•	•		1962–1994
Auerbach (1999)	•	•	•	1–11	•					1986–1999
Frendreis and Tatalovich (2000)	•	•	•	1		•	•	•		1962–1997
Kliesen and Thornton (2001)	•			1, 5	•					1981–2000
Lipford (2001)	•	•		1–5	•	•	•	•		1980–1999
Penner (2001, 2002)	•			1, 5	•					1980–2000
Kitchen (2003)	•	•		1–5	•					1982–2001
Cohen and Follette (2003)	•	•	•	1	•					1977–2003
Krause and Douglas (2005)	•	•	•	1	•	•	•	•		1976–2001
Corder (2005)	•	•	•	1–5		•	•	•	•	1976–2003
Krause and Douglas (2006)	•	•		1	•					1947–2001
Penner (2008)	•			1, 2, 5	•					1983–2005
Huntley and Miller (2009)	•		•	1–5		•	•	•	•	1993–2003
Kliesen and Thornton (2012)	•		•	1, 5	•					1976–2010
Krol (2014)	•	•	•	2, 5		•				1976–2008
CBO (2015b)	•	•		1–6	•	•				1982–2014
Martinez (2011, 2015)	•	•	•	1, 5	•					1984–2013
Tsuchiya (2016)	•	•	•	1–4		•			•	1984–2012
CBO (2017b)	•	•		1, 6	•				•	1985–2016
Ericsson (2017a)	•	•	•	1	•					1984–2012
Croushore and Van Norden (2017)	•		•	1	•	•		•		1967–2010
Croushore and Van Norden (2018)	•		•	1	•			•		1981–2010

Notes. "CBO (various)" denotes CBO (2002), CBO (2004), CBO (2005), CBO (2006), CBO (2007), CBO (2009), CBO (2010), CBO (2013), CBO (2015a), and CBO (2017a), which examine forecast periods from 1976 through (respectively) 2000, 2003, 2004, 2005, 2006, 2008, 2009, 2010, 2012, and 2014. "Other" forecasters include the FRB, CEA, SSA, APB, SPF, Blue Chip Consensus, the ASA/NBER survey, and various private-sector sources. Forecast horizons are in years. Variables forecast are the budget and related items (B), gross domestic product (Y), inflation (Δp), unemployment (U), and other economic variables (X).

contains the most information content. Krause and Douglas (2005) run several forecast-encompassing tests on CBO, OMB, and FRB forecasts of the budget and other economic variables. In a related vein, Corder (2005) uses regression-based tests to evaluate the economic forecasts from the Social Security Administration (SSA), CBO, and OMB; and he examines whether an agency's forecasts could be improved by incorporating information from the other agencies' forecasts.

These various studies provide information on the relative performance of forecasts across different samples, variables, and metrics. Table 1 lists a selection of studies that have evaluated the U.S. federal budget agency forecasts. Additional studies evaluate budget forecasts other than those by U.S. federal agencies. See in particular Williams and Calabrese (2016) for an extensive and systematic review of the literature, Frankel (2011) for cross-country comparisons, and Feenberg, Gentry, Gilroy, and Rosen (1989), Gentry (1989), and Sun (2008) on forecasts of U.S. state budgets.

3. Data

In Sections 4–6 below, empirical examples illustrate different forecast evaluation methods in order to clarify how those methods are implemented and to highlight their strengths and limitations. Section 3.1 describes the forecasts in those examples, which are all of U.S. gross federal debt. Section 3.2 provides a graphical perspective as a prelude to the numerical illustrations in Sections 4–6.

3.1 Data Description

The variable being forecast in the empirical examples is total U.S. gross federal debt outstanding, in billions of dollars, from 1984 through 2018, as measured for fiscal years ending on September 30. The data on total U.S. gross federal debt ("DEBT") are published by the U.S. Department of the Treasury's Financial Management Service in the December issue of the *Treasury Bulletin* and in its *Monthly Treasury Statement*.

For the most part, the forecasts examined are the one-year-ahead forecasts. Those forecasts are denoted by their sources:

- the Congressional Budget Office (CBO), from its *Budget and Economic Outlook*;
- the Office of Management and Budget (OMB), from its *Budget of the United States Government*; and
- the *Analysis of the President's Budget* (APB).

The Congressional Budget Office and the Office of Management and Budget are different agencies within the U.S. federal government. The *Analysis of the President's Budget* is produced by the Congressional Budget Office, but the policy assumptions embedded in the forecasts from the *Analysis of the President's Budget* differ from those in the forecasts from the CBO's *Budget and Economic Outlook*. Thus, these

two forecasts are referred to as the "APB forecast" and the "CBO forecast" respectively, while noting that both are produced by the Congressional Budget Office. For expositional convenience, the three forecasts listed above are referred to as "agency forecasts", although only two agencies are involved. Also, the empirical illustrations below always use logs (rather than levels) of debt and its forecasts.

The three forecasts are released at the beginning of the calendar year—usually about a month apart in January, February, and March—and are of the level of the U.S. federal debt at the end of the (then) current fiscal year and of future fiscal years. Thus, the forecasts are not precisely one year ahead or an integer number of years ahead, but they will be referred to as such for ease of reference. So, the forecast horizon h is denoted $h = 1$ (denoting the end of the current fiscal year) or $h > 1$ (denoting the end of future fiscal years). See Martinez (2015, Figure 1) for an illustrative timeline, Martinez (2011, Table 2) for specific dates, and Martinez (2015) and Ericsson (2017a) for more detailed descriptions of this measure of debt and of its forecasts.

Importantly, the forecasts are conditioned on different policy assumptions. The CBO forecast assumes that current laws will remain unchanged over the forecast horizon, whereas the OMB and APB forecasts assume that the policy changes proposed in the president's budget will be implemented. From this perspective, the forecasts represent different policy scenarios (or "projections") rather than unconditional forecasts *per se*. That said, it is still of interest to determine how useful these different forecasts are, both relatively and absolutely, and whether any individual forecast subsumes the information in the other forecasts—especially given the prominence that the forecasts play in policy formulation. With that in mind, the agencies' forecasts *are* referred to as "forecasts" below, while recognizing that some of these forecasts may also be usefully viewed as policy scenarios. This broader usage of the term "forecast" is in line with Clements and Hendry (2002, p. 2): "A forecast is any statement about the future". For more information on the forecasts' assumptions, see Martinez (2011).

Discussions of the deficit often overshadow discussions of the federal debt, since the deficit is commonly thought of as equaling the change in debt. Nonetheless, the change in debt differs from the deficit. The latter excludes certain items that are included in the change in debt, such as the Troubled Asset Relief Program (TARP) and changes in cash balances held by the Treasury. The inclusion or exclusion of such items can substantially alter the implied measure of debt; and the CBO and OMB debt forecasts and their relative merits may depend on which measure of debt is used. For example, the difference between the 2009 debt forecasts by the CBO and the OMB was largely due to differences in the agencies' forecasts of the change in financial assets and liabilities in response to the financial crisis. Equally, focusing on the deficit could miss components of debt that are important for policy. Gross federal debt *per se* may be a particularly relevant measure for policy because it is a closer measure of the debt subject to the debt ceiling than is (e.g.) debt held by the public.

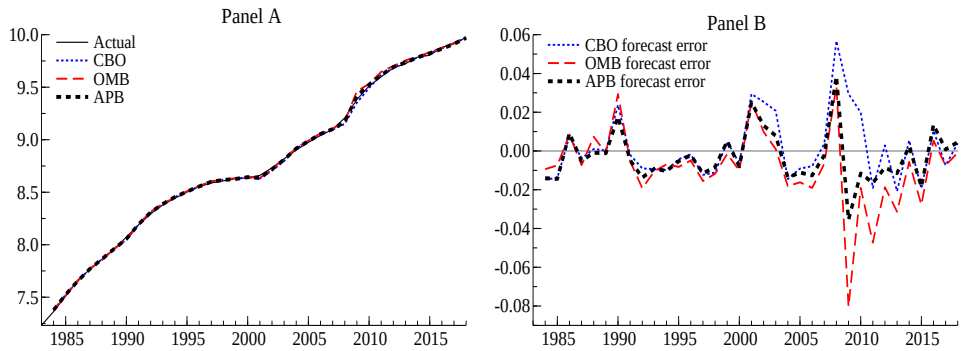

Figure 1: Government agency forecasts and outcomes (in logs) and forecast errors (in percent, expressed as a fraction) of the federal debt.

3.2 Graphical Analysis

Graphs furnish a useful preamble to a more formal statistical comparison and econometric evaluation. To highlight the value of graphical analysis, the current subsection considers the actual debt, its forecasts, and the implied forecast errors in various representations that afford a variety of perspectives on the forecasts themselves. Direct visual comparison of forecasts relative to the outcomes being forecast provides an initial assessment of forecast performance. Figures 1–3 present a smorgasbord of such comparisons.

To start, consider a comparison of forecasts at a *given* horizon with the outcomes being forecast, as in Figure 1 for the one-year-ahead forecasts. Panel A in Figure 1 plots the logs of actual debt and its forecasts, showing just how much debt has grown over the sample period and indicating how the forecasts have performed. Panel B in Figure 1 plots the corresponding forecast errors, where the forecast error is calculated as the log of actual debt minus the log of the forecast. The forecast errors in Panel B are thus in percent of debt, expressed as a fraction. The largest forecast errors were in 1990, 2001, 2002, 2008, 2009, 2011, and 2013. By way of interpretation, in each of these years the United States was entering a recession or expansion (as dated by the National Bureau of Economic Research), or there were major policy changes. For 2008 and 2009, some forecast errors are 4% of debt or more—very large for forecasts of a stock (debt) made within a year of its realization.

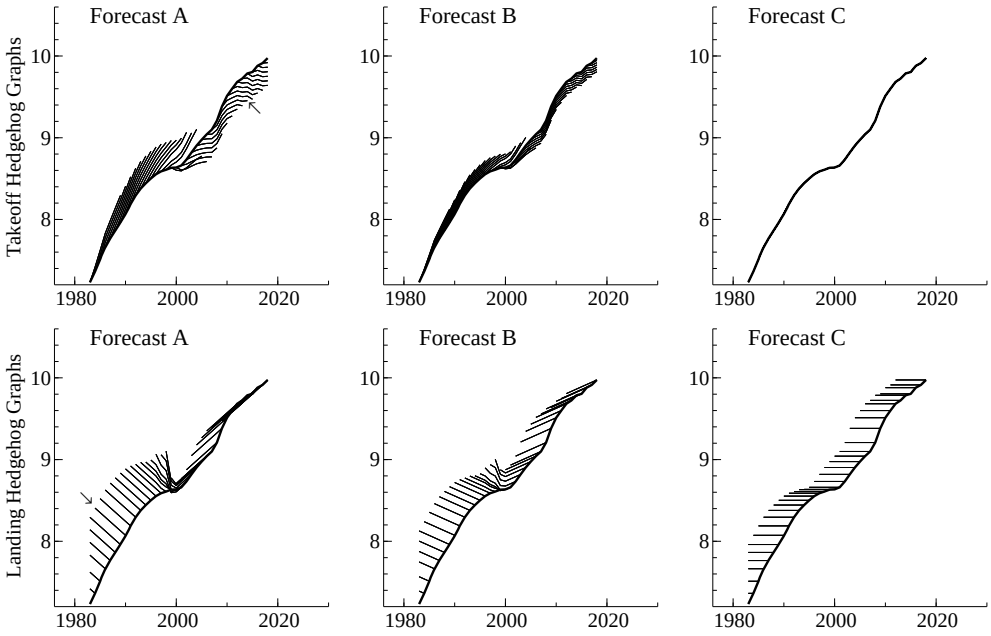

Figure 2: Hedgehog graphs of illustrative forecasts (Forecasts A, B, and C) of the federal debt (in logs).

Another form of graph—the hedgehog graph—helps ascertain systematic features of the forecasts over *multiple* horizons. There are two types of hedgehog graphs: "takeoff" and "landing". Some stylized forecasts help demonstrate what these graphs convey. Then, hedgehog graphs of the CBO, OMB, and APB forecasts are considered.

Figure 2 presents both types of hedgehog graphs for the log of three illustrative (and purely artificial) forecasts—denoted Forecasts A, B, and C—along with the log of actual debt. The top row of panels in Figure 2 are hedgehog graphs of takeoff, one graph for each of Forecasts A, B, and C. Each "spine" on a takeoff graph plots a path of forecasts that were made on a given date, across multiple forecast horizons. For instance, in the takeoff graph for Forecast A, the upward-angled arrow points to the spine of the forecasts that were (hypothetically) made at the beginning of 2009 for debt at the end of fiscal years 2009, 2010, ..., 2014, and 2015. These forecasts substantially under-predict actual debt, and the magnitude of under-prediction increases as the forecast horizon increases. The spines in this takeoff graph illustrate under-prediction in the second half of the sample (spines below actual) and over-prediction in the first half of the sample (spines above actual). On average over the sample, though, the forecast errors for Forecast A are approximately zero (i.e., unbiased), even though the forecasts are systematically biased over each subsample. Section 5 considers forecast bias in greater detail.

Forecast B has smaller forecast errors than Forecast A for all horizons and dates, with the spines for Forecast B's takeoff graph all lying closer to the line for actual federal debt than do the spines for Forecast A. Forecast C represents a set

of "perfect" forecasts, in that the forecasts equal the actual values being forecast. In the takeoff graph for Forecast C, the spines lie on top of the actual values being forecast.

The bottom row of panels in Figure 2 are hedgehog graphs of landing. Each spine on a landing graph plots a sequence of forecasts from the longest horizon to the shortest horizon, where the outcome being forecast occurs on a *given* date. For instance, in the landing graph for Forecast A, the downward-angled arrow points to the spine of the forecasts that were (again, hypothetically) made in 1984, 1985, ..., 1989, and 1990 for debt at the end of fiscal year 1990. As the negative slope of the spine implies, the forecasts were revised downward year by year. The corresponding forecast errors were negative, declining in magnitude as the forecast horizon decreased. All spines prior to 2000 share those features—a negative slope, and implied negative forecast errors that decline in magnitude as the forecast horizon decreased. After 2000, the positive slope of the spines implies that the forecasts were revised upward year by year. Their corresponding forecast errors were positive, declining in magnitude as the forecast horizon decreased.

Forecast B has smaller forecast errors than Forecast A for all horizons and dates, with the spines for Forecast B's landing graph all lying flatter than those for Forecast A. Forecast C represents a set of "perfect" forecasts, in that the forecasts equal the actual values being forecast. In the landing graph for Forecast C, the spines are horizontal, lying at the actual values being forecast. These hedgehog graphs of stylized forecasts provide context for interpreting hedgehog graphs of the CBO, OMB, and APB forecasts.

Figure 3 presents both types of hedgehog graphs for the log of each agency's forecast, along with the log of actual debt.

The top row of panels in Figure 3 are hedgehog graphs of takeoff. Each spine on a takeoff graph plots a path of forecasts that were made on a given date, across multiple forecast horizons. For instance, in the CBO takeoff graph, the horizontal arrow points to the spine of CBO forecasts that were made in January 2009 for debt at the end of fiscal years 2009, 2010, ..., 2018, and 2019. These forecasts substantially under-predict actual debt, and the magnitude of under-prediction increases as the forecast horizon increases. This spine and others in the takeoff graphs illustrate that longer-horizon forecasts perform particularly poorly around turning points. Forecasts made in 2001 and 2008–2009 (both beginnings of recessions) tended to under-predict future debt. That said, debt forecasts in the late 1990s (an expansionary period) tended to over-predict somewhat: the economy grew faster than expected, with tax receipts bringing in more revenue than anticipated. More generally, takeoff graphs portray how "optimistic" or "pessimistic" the forecasts were, relative to outcomes, and how that optimism or pessimism evolved across horizons and over time.

The bottom row of panels in Figure 3 are hedgehog graphs of landing. Each spine on a landing graph plots a sequence of forecasts from the longest horizon to the shortest horizon, where the outcome being forecast occurs on a given date. For instance, in the CBO landing graph, the vertical arrow points to the spine of CBO forecasts that were made in 1984, 1985, ..., 1989, and 1990 for debt at the end

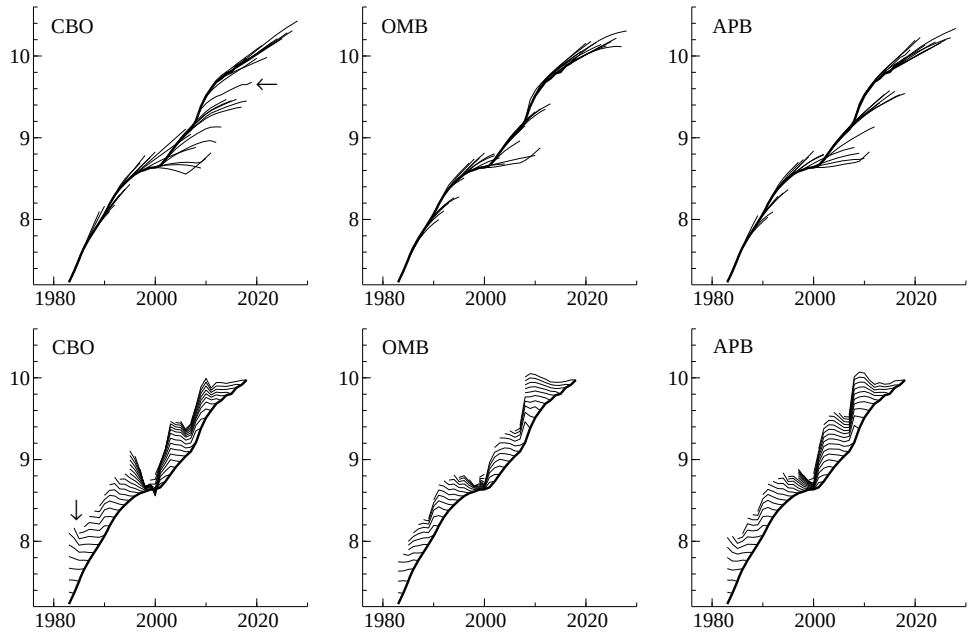

Figure 3: Hedgehog graphs of U.S. government agency forecasts of the federal debt (in logs).

of fiscal year 1990. As the relative flatness of that spine implies, these forecasts changed little as they were updated year by year. This spine and others in the landing graphs show that forecast revisions are typically small, with many forecast paths being remarkably flat. Occasionally, however, forecasts have large upward or downward revisions, often corresponding to significant changes in macroeconomic conditions, government policies, or both. More generally, the landing graphs show how forecasts of a particular outcome are revised over time, as might arise from new information obtained about the economy and about policy.

4. Comparison of Alternative Forecasts

This section considers various methods for comparing alternative forecasts. These methods include mean squared forecast errors (MSFEs, in Section 4.1), forecast encompassing (in Section 4.2), and (closely related) the pooling and combination of forecasts (in Section 4.3).

4.1 Comparisons of RMSEs

In many frameworks, good forecasts produce small expected losses, while bad forecasts produce large expected losses. One very common loss function is squared error loss, also known as quadratic mean squared error (MSE) loss. The MSE satisfies requirements laid out by Granger (1999) that a loss function (i) has a minimum of zero for a zero forecast error, (ii) is greater than zero for nonzero forecast errors, and (iii) is non-decreasing as the magnitude of the error increases. Additionally, the

Table 2: A comparison of root mean squared forecast errors.

Statistic	CBO	OMB	APB	DDD
RMSE	1.68%	2.17%	1.39%	2.78%
Relative RMSE	0.60	0.78	0.50	1
Diebold–Mariano t-statistic	−2.88** [0.007]	−2.06* [0.048]	−3.51** [0.001]	
Diebold–Mariano t-statistic (HAC)	−2.70* [0.011]	−1.68 [0.101]	−2.99** [0.005]	

Notes. Asterisks * and ** denote statistical significance at the 5% and 1% levels respectively, and p-values are in square brackets.

MSE is symmetric, and its quadratic nature penalizes larger forecast errors more than proportionately.

In this vein, a forecast can be empirically evaluated by estimating its expected loss with the sample average of the loss. For the MSE, the sample average of the squared forecast errors is:

$$\text{MSE} \;=\; \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2\,, \tag{1}$$

where y_t is the outcome at time t (i.e., the variable being forecast), $\hat{y}_t$ is a forecast of y_t, and the forecasts are made for T observations ($y_t;\ t = 1, ..., T$).

In practice, the square root of the MSE in equation (1) is typically reported, rather than the MSE itself, noting that the root mean squared error (RMSE) is the out-of-sample equivalent to the in-sample residual standard error. From the properties of the RMSE, a smaller value indicates a better forecast performance. Hence, it is common to compare forecast performance by comparing the RMSEs across forecasts, with smaller RMSEs indicating better performance. Granger (1989b, pp. 186–187) proposes how to test for statistically significant differences between RMSEs. Diebold and Mariano (1995), and subsequently Giacomini and White (2006), extend and generalize that approach to testing. See Clements and Hendry (1993) on limitations of the RMSE and Diebold (2015) on the use and misuse of the Diebold–Mariano test statistic.

Illustration. Table 2 reports the RMSEs for each of the three agencies' one-year-ahead debt forecasts. It also reports the RMSE for forecasts from a simple double-differenced device (DDD), which is a "robust" naive forecast device that is calculated as the previous year's debt plus the change in the previous year's debt; see Hendry (2006). The agency forecasts have smaller RMSEs than the naive forecast. The APB forecast has the smallest (1.39%), followed by the CBO forecast (1.68%) and the OMB forecast (2.17%). Table 2 also reports the RMSEs, *relative* to the RMSE of the naive forecast (the DDD forecast). Relative RMSEs are a common way of numerically comparing forecasts to a benchmark forecast.

The penultimate row in Table 2 reports Diebold–Mariano test statistics that compare each agency's forecast with the DDD forecast. Associated p-values are in

square brackets. The RMSE of each agency's forecast is statistically significantly smaller than the RMSE for the DDD forecast at the 95% level. The final row in Table 2 reports the same Diebold–Mariano test statistics but with Andrews's (1991) heteroscedasticity- and autocorrelation-consistent (HAC) correction. The results are similar but at somewhat reduced significance levels.

4.2 Forecast Encompassing

Chong and Hendry (1986) develop the concept of forecast encompassing as an approach for comparing alternative forecasts and determining whether one of them is "sufficient" in a very specific statistical sense. Importantly, having the smallest RMSE is necessary but not sufficient for a given forecast to forecast-encompass other forecasts; see Ericsson (1992). This subsection motivates forecast encompassing through the regression used to test for it. Transformations of and restrictions on that regression provide additional insight on the nature of forecast encompassing; and they also link directly to subsequent sections.

Consider two alternative forecasts $\hat{y}_t$ and $\tilde{y}_t$ of the variable y_t. Chong and Hendry (1986, equation (7)) propose running the following "unrestricted" regression with coefficients $\{b_1, b_2\}$ and residual e_t:

$$y_t \;=\; b_1 \hat{y}_t + b_2 \tilde{y}_t + e_t \,, \tag{2}$$

and testing $\{b_1 = 1, b_2 = 0\}$. This hypothesis holds when the first forecast $\hat{y}_t$ is an "adequate" forecast for y_t (hence $b_1 = 1$) and, given that first forecast $\hat{y}_t$, the second forecast $\tilde{y}_t$ is redundant (hence $b_2 = 0$).

In that light, equation (2) has a useful second representation. Subtracting $\hat{y}_t$ from both sides, equation (2) can be rewritten as:

$$(y_t - \hat{y}_t) \;=\; c_1 \hat{y}_t + b_2 \tilde{y}_t + e_t \,, \tag{3}$$

where the dependent variable is the first forecast's forecast error $(y_t - \hat{y}_t)$, and $c_1 = b_1 - 1$. Under the same null hypothesis as before, then $\{c_1 = 0, b_2 = 0\}$: that is, the two forecasts $\hat{y}_t$ and $\tilde{y}_t$ are uninformative in explaining the first forecast's forecast error $(y_t - \hat{y}_t)$.

Chong and Hendry (1986, equation (8)) also consider a restricted version of equation (3) in which $c_1 = 0$ is imposed:

$$(y_t - \hat{y}_t) \;=\; b_2 \tilde{y}_t + e_t \,. \tag{4}$$

Equation (4) is used to test whether the second forecast $\tilde{y}_t$ is informative about the first forecast's forecast error $(y_t - \hat{y}_t)$. As Chong and Hendry note, this expresses the regression in a "residual diagnostics" form, with the "residual" being the first forecast's forecast error $(y_t - \hat{y}_t)$.

Equation (2) has a third representation that provides yet additional insight. To obtain that representation, add $+b_2\hat{y}_t - b_2\hat{y}_t$ to the right-hand side of equation (3). Then, re-arrange terms in that equation to obtain $\hat{y}_t$ (the first forecast) and $(\tilde{y}_t - \hat{y}_t)$ (the differential of the two forecasts) as regressors:

$$(y_t - \hat{y}_t) \;=\; c_2 \hat{y}_t + b_2 (\tilde{y}_t - \hat{y}_t) + e_t \,, \tag{5}$$

where $c_2 = c_1 + b_2 = b_1 + b_2 - 1$. Under the null hypothesis discussed above, then $\{c_2 = 0, b_2 = 0\}$: that is, neither the second forecast nor the differential between the two forecasts is informative in explaining the first forecast's forecast error.

In practice, unit homogeneity of the two forecasts with respect to the outcome is sometimes imposed on equation (5), as would occur if each forecast is cointegrated $(+1 : -1)$ with the outcome; see Ericsson (1993). In equation (5), that homogeneity restriction corresponds to $c_2 = 0$, resulting in:

$$(y_t - \hat{y}_t) \;=\; b_2(\tilde{y}_t - \hat{y}_t) + e_t \,. \tag{6}$$

Intuitively, equation (6) examines whether the *additional* information in the second forecast—as captured by the forecast differential $(\tilde{y}_t - \hat{y}_t)$—can help explain the first forecast's forecast error $(y_t - \hat{y}_t)$. Put somewhat differently, the forecast differential $(\tilde{y}_t - \hat{y}_t)$ measures the relevance of information in the second forecast that is not contained in the first forecast. Equivalently, equation (6) imposes the unit homogeneity restriction $b_1 + b_2 = 1$ on equation (2). Equation (6) also solves a "balance" problem in equation (4) for integrated-cointegrated forecasts and outcomes; see Ericsson (1992). Equations (3) and (5) do not impose that homogeneity restriction: they are directly equivalent to equation (2) but are written in different representations.

To summarize, in each of equations (2)–(6), the basic question is whether additional information can help explain the first forecast's forecast error or, in essence, help improve the first forecast. These equations can be easily extended in useful directions to include an intercept, to reverse the forecasts' roles, and to compare more than two forecasts; see Ericsson and Marquez (1993), Marquez and Ericsson (1993), and Martinez (2015) *inter alia*. The empirical illustrations below employ these extensions, with three (rather than two) forecasts.

Illustration. To start, consider the "unrestricted" regression (2) as applied to actual debt and the three debt forecasts:

$$\begin{array}{rcccccccc} debt_t & = & -0.064 & - & 0.15\, cbo_t & - & 1.02\, omb_t & + & 2.17\, apb_t \,, \\ & & (0.032) & & (0.20) & & (0.31) & & (0.45) \end{array} \tag{7}$$

where estimated coefficients are reported for the intercept and $\{b_{cbo}, b_{omb}, b_{apb}\}$, which generalize $\{b_1, b_2\}$. Lowercase variables denote the logs of uppercase variables, estimated standard errors are in parentheses, and the sample period is 1984–2018.

Building on the discussion of equation (2), the forecast-encompassing hypothesis $\{b_{cbo} = 1,\ b_{omb} = 0,\ b_{apb} = 0\}$ examines whether the CBO forecast is an adequate forecast for actual debt, with the OMB and APB forecasts being redundant, given the CBO forecast. This hypothesis is strongly rejected, with an F-statistic of 11.5 and a p-value of less than 0.1%. Similar tests can be calculated for the OMB and APB forecasts. As reported in the row for "unrestricted" regressions in Table 3, each agency's forecasts could benefit from the other agencies' forecasts: all three statistics for the unrestricted equation reject at standard significance levels.

Next, consider the "residual diagnostic" formulation in equation (4), with the CBO's forecast error as the dependent variable and the OMB and APB forecasts as regressors:

$$\begin{array}{rcccccc} (debt_t - cbo_t) & = & -0.018 & - & 0.03\, omb_t & + & 0.03\, apb_t \,. \\ & & (0.044) & & (0.38) & & (0.38) \end{array} \tag{8}$$

Table 3: Forecast-encompassing test statistics.

Regression type	CBO	OMB	APB
Unrestricted [equation (2)]	11.5** [0.000] $F(3,31)$	20.7** [0.000] $F(3,31)$	4.01* [0.016] $F(3,31)$
Residual diagnostic [equation (4)]	0.13 [0.880] $F(2,32)$	4.60* [0.018] $F(2,32)$	2.27 [0.119] $F(2,32)$
Forecast differential [equation (6)]	13.8** [0.000] $F(2,32)$	26.5** [0.000] $F(2,32)$	3.61* [0.039] $F(2,32)$

Notes. The three entries within a given block of numbers are the value of the F-statistic for testing the null hypothesis of forecast encompassing by the forecasting agency listed at the top of the column, the tail probability associated with that value of the test statistic (in square brackets), and the distribution under the null hypothesis, with degrees of freedom in parentheses. Asterisks * and ** denote statistical significance at the 5% and 1% levels respectively.

In equation (8), the coefficients on *omb* and *apb* are jointly insignificant, with an F-statistic of 0.13. However, the implicit unit restriction on *cbo* is strongly rejected, with an F-statistic of 34.0, as is apparent from the coefficient on *cbo* in equation (7). Table 3 reports F-statistics for the residual diagnostic form for all three forecasts.

Finally, consider the "forecast differential" regression in equation (6), which imposes $b_{cbo} + b_{omb} + b_{apb} = 1$ on the unrestricted formulation (7). With the CBO forecast error as the dependent variable, that forecast-differential regression is:

$$(debt_t - cbo_t) \;=\; \underset{(0.002)}{-\,0.000} \;-\; \underset{(0.25)}{0.61}\,(omb_t - cbo_t) \;+\; \underset{(0.40)}{1.66}\,(apb_t - cbo_t)\;. \qquad (9)$$

Jointly, the coefficients on both of the forecast differentials $(omb_t - cbo_t)$ and $(apb_t - cbo_t)$ in equation (9) are statistically highly significant, with an F-statistic of 13.8. The final row in Table 3 reports the forecast-differential form of the forecast-encompassing statistic for all three forecasts. For each forecast, its forecast error can be explained in part by the forecast differentials relative to the two other forecasts.

In summary, each agency forecast could be improved by using information in the other two forecasts, as the unrestricted form and forecast-differential form of the forecast-encompassing test statistic indicate. The residual diagnostic form of the forecast-encompassing test appears less informative here, probably because that test imposes an empirically rejectable implicit unit restriction.

4.3 Pooling and Combining Forecasts

Bates and Granger (1969) propose combining or "pooling" forecasts to improve forecast accuracy. In essence, forecast combination implies choosing nonzero values for both b_1 and b_2 in equation (2), which could be advantageous if neither forecast forecast-encompasses the other. Many options have been considered for selecting the weights b_1 and b_2 on the forecasts, including equal weights, regression-based

Table 4: RMSEs of some individual and pooled forecasts.

	CBO	OMB	APB	DDD	Average	Reg-Un	Reg-FD
RMSE	1.68%	2.17%	1.39%	2.78%	1.51%	1.15%	1.23%

weights, and Bayesian weights. Granger (1989a), Clemen (1989), and Timmermann (2006) review the literature on forecast combinations; Diebold (1989) discusses links and differences between forecast encompassing and forecast combination; Hendry and Clements (2004) consider the possible benefits to pooling the forecasts of differentially mis-specified models; and Hansen (2007) examines estimated weights. Forecast combination has potential benefits, and also important caveats, as Hendry and Doornik (2014) discuss.

> ... A combination of forecasts can outperform, on some measures, all the individual forecasts when there are offsetting biases, offsetting breaks, or diversification across relatively uncorrelated forecasts which reduces the variance of the average. Conversely, averaging without any selection for the set of forecasts involved has obvious drawbacks: by way of analogy, with 10 glasses of pure drinking water and one of a virulent poison, it does not seem wise to mix all of these before drinking, rather than select out the glass of poison. (p. 286)

Illustration. The forecast-encompassing equations in Section 4.2 can be interpreted as motivation for forecast combination. To illustrate, Table 4 augments the RMSEs in Table 2 with RMSEs from three pooled forecasts:

- an equally weighted average of the CBO, OMB, and APB forecasts ("Average");
- the regression-based combination of the forecasts from the unrestricted forecast-encompassing regression in equation (7) (denoted "Reg-Un"); and
- the regression-based combination of the forecasts from the forecast-differential forecast-encompassing regression in equation (9) (denoted "Reg-FD").

Both of the regression-based forecast combinations have smaller RMSEs than the APB forecast, which itself has the smallest RMSE among the individual agency forecasts. Thus, the APB forecast appears to lack some relevant information that is available from the other agencies' forecasts. The forecast-encompassing statistics in the final column of Table 3 indicate that the CBO and OMB forecasts could improve the APB forecast. The RMSEs in Table 4 for the regression-based forecast combinations indicate what that improvement might be. That said, the RMSE for the APB forecast is still smaller than the RMSE for the equally weighted average of the three individual forecasts.

5. Forecast Failure

This section discusses different approaches to testing for forecast failure, including subsample tests (Section 5.1), tests for bias and efficiency (Section 5.2), and generalizations thereof (Section 5.3). Section 6 then develops a unified framework that includes these tests and the ones in Section 4, which compare alternative forecasts.

5.1 Comparisons across Subsamples

This subsection examines how forecasts may be evaluated across subsamples, and in particular how such evaluation can help detect predictive failure. In this vein, Chow (1960) proposes comparing the in-sample performance of a given *model* with that same model's out-of-sample performance, utilizing the prediction interval of the (out-of-sample) forecasts. Numerically, Chow's test statistic compares the in-sample estimated residual variance with the out-of-sample mean squared forecast error. Chow's statistic is thus designed to detect a worsening performance of a given model in the out-of-sample period, i.e., predictive failure.

Chow distinguishes his statistic from the Fisher (1922) covariance test statistic, which compares the coefficient estimates from one subsample with the coefficient estimates from another subsample. Andrews's (1993) unknown breakpoint test and Bai and Perron's (1998) multiple breakpoint test generalize Fisher's test; see also Section 5.3.

As illustrated below, Chow's statistic can also be used for comparing *forecasts* across different subsamples, and not just for comparing model-based in-sample and out-of-sample results. That is, the Chow statistic can compare the performance of a *given* forecast across different subsamples. That contrasts with the Diebold–Mariano and forecast-encompassing statistics, which compare *different* forecasts across the same sample. The Chow statistic thus provides information about the forecasts that is distinct from the information in the Diebold–Mariano and forecast-encompassing statistics; see Ericsson (1992) for further discussion.

Illustration. Table 5 reports the RMSEs for the three agencies' forecasts over the subsamples 1984–2000 and 2001–2018, with the RMSEs for the full sample (1984–2018) given as reference. For all agencies, the RMSEs increase markedly from the first subsample to the second. The last row in Table 5 reports the Chow statistics for that split of the sample: the increases in RMSEs are statistically highly significant for all three agencies' forecasts.

These Chow statistics quantify what is apparent visually in Panel B of Figure 1: forecast performance worsens substantially after 2000. That worsening could have resulted from any of many potential causes—for instance, greater challenges in forecasting over 2001–2018, which included two major recessions. Chow's (1960) statistic is specific to the particular sample split chosen: Sections 5.3 and 6 discuss how that restriction can be relaxed.

5.2 Tests of Bias and Efficiency

An additional approach for assessing forecast performance is through tests of forecast bias and efficiency. Mincer and Zarnowitz (1969, pp. 8–11) propose testing for

Table 5: Subsample RMSEs and corresponding Chow statistics.

Statistic	CBO	OMB	APB
RMSE (1984–2018)	1.68%	2.17%	1.39%
RMSE (1984–2000)	0.99%	1.17%	0.95%
RMSE (2001–2018)	2.13%	2.80%	1.71%
Chow statistic	4.63** [0.001] $F(18, 17)$	5.77** [0.000] $F(18, 17)$	3.20* [0.010] $F(18, 17)$

Notes. The three entries within a given block of numbers for the Chow statistic are the value of the statistic itself, the tail probability associated with that value of the statistic (in square brackets), and the statistic's distribution under the null hypothesis, with degrees of freedom in parentheses. Asterisks * and ** denote statistical significance at the 5% and 1% levels respectively.

forecast bias by regressing the forecast error on an intercept and testing whether the intercept is statistically significant. Continuing in the notation of Section 4.2, that regression is:

$$(y_t - \hat{y}_t) \;=\; b_0 + e_t \,, \tag{10}$$

where b_0 is the intercept. A test of $b_0 = 0$ is interpretable as a test that the forecast $\hat{y}_t$ is unbiased for the variable y_t. That is, the forecast error is zero on average. For one-step-ahead forecasts, the error e_t may be serially uncorrelated, in which case a standard t- or F-statistic for $b_0 = 0$ may be appropriate. For multi-step-ahead forecasts, e_t generally will be serially correlated; hence, inference about the intercept may require accounting for that autocorrelation.

Mincer and Zarnowitz (1969, p. 11) also propose how to assess a forecast's efficiency. Their efficiency test uses a slightly more general version of equation (10) in which the coefficient on the forecast itself is estimated, rather than imposed to be unity:

$$y_t \;=\; b_0 + b_1 \hat{y}_t + e_t \,, \tag{11}$$

where b_1 is the coefficient on $\hat{y}_t$, and $b_1 = 1$ in equation (10). Mincer and Zarnowitz (1969) interpret a test that $b_1 = 1$ as a test of the efficiency of the forecast $\hat{y}_t$ for the outcome y_t. The joint hypothesis $\{b_0 = 0,\ b_1 = 1\}$ of unbiasedness and efficiency is also of interest.

Equation (11) has a useful alternative representation. Subtracting $\hat{y}_t$ from both sides, equation (11) can be rewritten in a residual diagnostic form:

$$(y_t - \hat{y}_t) \;=\; b_0 + c_1 \hat{y}_t + e_t \,, \tag{12}$$

where $c_1 = b_1 - 1$, as in equation (3). The hypothesis $\{b_0 = 0,\ c_1 = 0\}$ in equation (12) is equivalent to $\{b_0 = 0,\ b_1 = 1\}$ in equation (11). A large literature on forecast efficiency further develops tests of such hypotheses. For example, Patton and Timmermann (2012) extend tests of forecast efficiency to multi-horizon forecasts by examining the forecast revisions across horizons; see also Nordhaus (1987) and Coibion and Gorodnichenko (2015).

Table 6: Mincer–Zarnowitz t- and F-statistics for testing unbiasedness and efficiency.

Null hypothesis	CBO	OMB	APB
Unbiasedness	0.50	−2.47*	−1.38
[equation (10): $b_0 = 0$]	[0.624]	[0.018]	[0.178]
	$t(34)$	$t(34)$	$t(34)$
Efficiency	0.45	−2.06*	−0.20
[equation (12): $c_1 = 0$]	[0.655]	[0.047]	[0.845]
	$t(33)$	$t(33)$	$t(33)$
Unbiasedness and efficiency	0.22	5.49**	0.94
[equation (12): $b_0 = c_1 = 0$]	[0.803]	[0.009]	[0.401]
	$F(2, 33)$	$F(2, 33)$	$F(2, 33)$

Notes. The three entries within a given block of numbers are the value of the test statistic (either t or F) for testing the null hypothesis, the tail probability associated with that value of the test statistic (in square brackets), and the distribution under the null hypothesis, with degrees of freedom in parentheses. Asterisks * and ** denote statistical significance at the 5% and 1% levels respectively.

Illustration. Using the CBO forecast to illustrate, the regression in equation (10) is:

$$(debt_t - cbo_t) \;=\; \underset{(0.0029)}{+0.0014}\;, \tag{13}$$

and the regression in equation (12) is:

$$(debt_t - cbo_t) \;=\; \underset{(0.035)}{-0.014} \;+\; \underset{(0.0040)}{0.0018}\; cbo_t\,. \tag{14}$$

From equation (13), the CBO forecast has a numerically small bias of +0.14%; and that bias is statistically insignificant, with a p-value of 62.4%. From equation (14), the estimates of the intercept and the slope coefficient are individually numerically small and statistically insignificant. Jointly, they also appear statistically insignificant, with an F-statistic of 0.22 and a p-value of 80.3%. Table 6 reports tests of unbiasedness and efficiency for all three agencies' forecasts. The CBO and APB forecasts appear unbiased and efficient in Mincer and Zarnowitz's sense, whereas the OMB forecasts appear both biased and inefficient.

5.3 General Tests of Forecast Bias

Mincer and Zarnowitz's (1969) test for forecast bias implicitly assumes that the bias b_0 is time-invariant; see equation (10). In practice, however, the forecast bias may vary over time. If it does, other tests may be more effective than the Mincer–Zarnowitz test at detecting that bias. Moreover, the Mincer–Zarnowitz test may lack power to detect certain forms of time-varying forecast bias, as when a positive forecast bias over part of the sample offsets a negative forecast bias elsewhere in the sample.

By allowing the intercept b_0 in equation (10) to vary freely over time, a com-

pletely general model of time-varying forecast bias may be formulated, as follows:

$$(y_t - \hat{y}_t) \;=\; \sum_{i=1}^{T} d_i I_{i,t} + e_t \,, \tag{15}$$

where $I_{i,t}$ is an impulse dummy equal to unity for $t = i$ and zero otherwise, and d_i is the coefficient on $I_{i,t}$. That is, d_i is the forecast bias in period i. Equation (15) can also be written with the intercept b_0 explicit:

$$(y_t - \hat{y}_t) \;=\; b_0 + \sum_{i=1}^{T} a_i I_{i,t} + e_t \,, \tag{16}$$

in which case a_i captures the deviation of the forecast bias in observation i from the average forecast bias b_0. When $a_1 = a_2 = \ldots = a_T = 0$ is imposed, equation (16) simplifies to equation (10) for Mincer and Zarnowitz's test.

For unrestricted a_i, equation (16) is not directly implementable in regression because it has T dummy coefficients for T observations. However, blocks of dummies can be included in regression, and that insight provides the basis for a technique known as impulse indicator saturation (IIS). IIS proceeds in two phases. In the first phase, equation (16) is estimated for subsets of impulse dummies and, for each subset, significant dummies are retained. In the second phase, equation (16) is re-estimated with the retained dummies from those subsets, followed by re-selection across those retained dummies. These two phases may be iterated as well. IIS has well-defined statistical properties, including (in the current context) high power to detect time-varying forecast bias.

Hendry (1999) originally proposed IIS as a procedure for testing parameter constancy. As such, IIS is a generic test for an unknown number of breaks, occurring at unknown times anywhere in the sample, with unknown duration, magnitude, and functional form. IIS is a powerful empirical tool for both evaluating and improving existing empirical models. Furthermore, many existing procedures can be interpreted as special cases of IIS in that they represent particular algorithmic implementations of IIS. Special cases include recursive estimation, rolling regression, Chow's (1960) predictive failure statistic, the unknown breakpoint tests by Andrews (1993) and Bai and Perron (1998), tests of extended constancy in Ericsson, Hendry, and Prestwich (1998), tests of nonlinearity, intercept correction (in forecasting), tests of aggregation, and robust estimation.

By testing and selecting over blocks of variables, IIS implements a machine-learning algorithm that solves the problem of having more potential regressors than observations. Notably, that is a problem common to the analysis of big data. Ericsson (2017a, Section 4) formalizes how IIS can also be used to test for time-varying forecast bias, as in equation (16). See also Johansen and Nielsen (2009, 2016), Doornik (2009), Hendry and Doornik (2014), and Ericsson (2017a) *inter alia* for theoretical developments and empirical applications of saturation techniques.

Illustration. Again, using the CBO forecast, IIS applied to equation (16) obtains:

$$\begin{array}{rcccc} (debt_t - cbo_t) & = & -\,0.0002 & + & 0.0571\; I_{2008,t} \,, \\ & & (0.0024) & & (0.0144) \end{array} \tag{17}$$

Table 7: IIS-based estimates of time-varying bias.

Estimates	CBO	OMB	APB
Estimated coefficients of retained impulse indicators			
$I_{1990,t}$		+3.80	
$I_{2001,t}$		+3.41	+2.95
$I_{2008,t}$	+5.71	+4.21	+4.29
$I_{2009,t}$		−7.18	−3.11
$I_{2011,t}$		−3.87	
IIS estimate of the average bias b_0 [equation (16)]	−0.02 (0.24)	−0.86** (0.18)	−0.44* (0.17)
Mincer–Zarnowitz estimate of the average bias b_0 [equation (10)]	0.14 (0.29)	−0.85* (0.34)	−0.32 (0.23)

Notes. Estimated biases are reported as percentages. The retained impulse indicators are detected at a 1% target size. Estimated standard errors for the impulse indicators are 1.4, 1.0, and 1.0 for the CBO, OMB, and APB forecasts respectively. Asterisks * and ** on estimated average biases denote statistical significance at the 5% and 1% levels respectively. Estimated standard errors are in parentheses.

where $I_{2008,t}$ (the impulse indicator for 2008) is retained at a tight (1%) target size or "gauge". Thus, from equation (17), the CBO forecast appears to have a time-varying bias, with a numerically large and statistically highly significant bias of over 5% in 2008 and a near-zero and statistically insignificant bias for all other years.

Accounting for such time variation can also affect inferences about the average forecast bias, as Table 7 highlights. In particular, the average bias for the APB forecast is statistically significant at close to the 1% level when using IIS, but it is statistically insignificant when estimated without IIS in the Mincer–Zarnowitz framework. IIS allows detection of time-varying forecast bias; and, it permits more robust and efficient estimation of time-*invariant* bias that may be present.

6. A Unified Approach

Impulse indicator saturation is not only a valuable tool for forecast evaluation: it also underpins a unified framework for all of the forecast evaluation procedures discussed above. This section sketches that framework.

As a preface, it is useful to note that the saturation approach discussed above applies to linear transformations of the impulse indicators, and not just to the impulse indicators themselves. Examples of such transformations include step functions, broken trends, economic variables, principal components and factors, time-dependent changes in variables' slope coefficients ("multiplicative indicator saturation"), and "designer" breaks. Ericsson (2011) proposes a systematic structure for discussing and developing such extensions.

With this in mind, consider the following equation:

$$(y_t - \hat{y}_t) \;=\; b_0 + \sum_{k=1}^{K} a_k x_{kt} + e_t \,, \tag{18}$$

Table 8: A summary of tools for forecast evaluation.

Type of evaluation	Statistical basis	Reference
Alternative forecasts	Graphical analysis	—
	RMSEs	Granger (1989b), Diebold and Mariano (1995)
	Forecast encompassing	Chong and Hendry (1986)
Forecast failure	Graphical analysis	—
	Known subsamples	Fisher (1922)
	Unknown subsamples	Andrews (1993), Bai and Perron (1998)
	Predictive failure	Chow (1960)
	Bias	Mincer and Zarnowitz (1969)
	Efficiency	Mincer and Zarnowitz (1969)
Generic	IIS	Hendry (1999), Johansen and Nielsen (2009)
	Saturation techniques	Ericsson (2011)

which includes an intercept b_0 and K potential regressors x_{kt} with slope coefficients a_k; and K may be greater than the number of observations T. For suitable choices of b_0, a_k, x_{kt}, and K, equation (18) can be re-expressed as each of the equations above, which motivate the different forecast evaluation tools. The regressions for the forecast-encompassing, Diebold–Mariano, and efficiency test statistics can be written as equation (18) because those regressions are all based on the forecasts $\hat{y}_t$ and $\tilde{y}_t$, which can be written as $\sum_{i=1}^{T} \hat{y}_t I_{i,t}$ and $\sum_{i=1}^{T} \tilde{y}_t I_{i,t}$, i.e., as linear combinations of impulse indicators. The Chow predictive failure statistic includes impulse indicators for only the out-of-sample period, simply testing their joint significance and not selecting among them; see Salkever (1976). For Mincer and Zarnowitz's test of forecast bias, the regression intercept can be written as $\sum_{i=1}^{T} 1 \cdot I_{i,t}$, which is the sum of the impulse indicators. In this way, the saturation framework provides a basis for interpreting these and many other tests for forecast evaluation.

Table 8 summarizes techniques for forecast evaluation, as categorized by the type of evaluation. The first category evaluates forecasts by comparing one forecast with other forecast(s): through graphical analysis, RMSEs, and forecast encompassing. The second category evaluates forecasts by their properties: across subsamples, bias, and efficiency. The third category, as represented by equation (18), includes generic procedures for evaluating forecasts and subsumes the first two categories.

Equation (18) emphasizes that these tools for forecast evaluation are in the spirit of Lagrange-multiplier residual-diagnostic tests; see Engle (1982, 1984). Moreover, regressors from different evaluation procedures can be included together in equation (18), allowing joint hypotheses to be tested. Also, some regressors may be "forced" to enter equation (18), as when those regressors are of central importance to the hypotheses being examined; see Martinez (2011), Hendry and Johansen (2015), and Ericsson (2017a) for examples.

Illustration. As Section 5.3 found, turning points in the business cycle may give rise to large errors in forecasts of government debt—and unsurprisingly so because actual outcomes of both expenditures and revenues are liable to be affected

when the economy moves from an expansion to a recession or from a recession to an expansion. Following Hendry and Johansen (2015), a natural extension of IIS in this context is to force NBER-based turning-point dummies to enter equation (16) (and hence equation (18)), with IIS applied to all remaining observations so as to capture any other important events that might bias the forecasts. That is, equation (16) becomes:

$$(y_t - \hat{y}_t) \;=\; b_0 \;+\; \sum_{i \,\in\, NBER} a_i I_{i,t} \;+\; \sum_{i \,\notin\, NBER} a_i I_{i,t} \;+\; e_t \,, \tag{19}$$

where $NBER$ denotes the set of turning-point observations, and selection of impulse indicators is across only the second summation, i.e., for $i \notin NBER$. This variation of IIS is thus "focused saturation" in that it focuses attention on certain key regressors (here, the intercept and the turning-point dummies) while still saturating the sample with impulse indicator dummies. Because the focus variables themselves are impulse indicator dummies, saturation does not need to include those particular dummies.

Applying focused saturation at a 1% target size to equation (19) with the CBO forecast obtains:

$$\begin{array}{lcccccccc} (debt_t - cbo_t) & = & -\,0.0057 & + & 0.0264\, I_{2003,t} + & 0.0256\, I_{2010,t} & + & \{NBER\}\,, \\ & & (0.0017) & & (0.0089) & (0.0089) & & \end{array} \tag{20}$$

where $\{NBER\}$ denotes the inclusion of impulse indicators for 1990, 1991, 2001, 2002, 2008, and 2009, i.e., the NBER-dated turning points in this sample. In equation (20), positive biases of about +2.6% are detected for both 2003 and 2010, and a small statistically significant bias of about −0.6% is present for the sample as a whole.

Table 9 summarizes the estimated forecast biases for the three agencies. Turning points typically have numerically large and statistically significant biases, with 2008 and 2009 dominating. Additional time-dependent biases are detected for the CBO and OMB forecasts, but not for the APB forecasts. The IIS estimate of the average bias b_0 indicates relatively small, negative, but highly statistically significant time-invariant biases for all agencies. At a more general level, equation (18) and the example in Table 9 illustrate the flexibility of the saturation approach—how it can incorporate into the model's structure the economic, institutional, and political insights of the researcher, while allowing for detection of additional phenomena.

7. Remarks

This section summarizes some implications of forecast evaluation, focusing on policy, predictability, diagnostics, interpretability, and extensions.

First, because forecasts of government budgets play important roles in policy, it is valuable to ascertain how good those forecasts are, and how they might be improved. The procedures discussed above provide a host of tools for evaluating those forecasts and for seeking ways in which to improve them. For the illustration with U.S. gross federal debt, agency forecasts are relatively good during quiescent periods, but they do sometimes deviate significantly from outcomes, particularly at turning points in the business cycle. So, budget forecasts might benefit from improving forecasts of the business cycle itself.

Table 9: Estimates of time-varying bias from focused impulse indicator saturation with NBER-based turning-point dummies.

Estimates	CBO	OMB	APB
Estimated coefficients of NBER-based turning-point dummies			
$I_{1990,t}$	+2.95	+3.93	+2.34
$I_{1991,t}$	+0.37	+0.46	+0.10
$I_{2001,t}$	+3.51	+3.54	+3.09
$I_{2002,t}$	+3.10	+1.98	+1.89
$I_{2008,t}$	+6.25	+4.34	+4.43
$I_{2009,t}$	+3.52	−7.05	−2.98
Estimated coefficients of retained impulse indicators			
$I_{1986,t}$		+1.83	
$I_{1988,t}$		+1.72	
$I_{2003,t}$	+2.64		
$I_{2010,t}$	+2.56		
$I_{2011,t}$		−3.74	
$I_{2013,t}$		−2.15	
Focused IIS estimate of the average bias b_0 [equation (19)]	−0.57** (0.17)	−0.99** (0.16)	−0.57** (0.16)

Notes. Estimated biases are reported as percentages. Estimated standard errors for impulse indicators are 0.9, 0.8, and 0.8 for the CBO, OMB, and APB forecasts respectively. The retained impulse indicators are detected at a 1% target size. Asterisks * and ** on estimated average biases denote statistical significance at the 5% and 1% levels respectively. Estimated standard errors are in parentheses.

Second, forecast evaluation with equation (18) emphasizes that evaluation focuses on the possible predictability of forecast errors—specifically, on whether or not the forecast errors have a systematic component. In essence, forecast evaluation with equation (18) examines whether the forecasts fully utilize the information in the regressors $\{x_{kt}\}$. If the forecasts don't, then improvement in the forecasts may be possible by better utilizing that information. That information may reflect information in another agency's forecasts (as with the Diebold–Mariano and forecast-encompassing statistics) or information specific to subsamples (as with the Chow statistic). Systematic forecast errors need not be persistent, as Granger (1983) highlights in his paper "Forecasting White Noise".

Third, certain challenges arise when interpreting rejection by any diagnostic statistic in forecast evaluation: the diagnostic statistic may have power to detect features other than the ones that it was designed for. Saturation-based tests in particular can detect not only time-varying forecast bias but also other forms of mis-specification, such as outliers due to heteroscedasticity and thick tails. Two items can help resolve this interpretational challenge. The structure of the retained dummies may have implications for their interpretation, as with the pattern of their estimated coefficients over time. And, outside information—such as from economic, institutional, and historical knowledge—can assist in interpreting the results, as

with the dates of business-cycle turning points in the empirical illustrations above. While "rejection of the null doesn't imply the alternative", the date-specific nature of saturation procedures can aid in identifying and potentially adjusting for important sources of forecast error. See Ericsson (2017b) for further discussion.

Fourth, from a more constructive perspective, different indicators are adept at characterizing different types of bias: impulse dummies for date-specific anomalies, step dummies for level shifts, and broken trends for evolving developments. Conversely, multiple tools are needed for forecast evaluation because the nature of the forecast errors is not known *ex ante*. Transformations of the variable being forecast may also affect the interpretation of the retained indicators. For instance, an impulse dummy for a growth rate implies a level shift in the (log) level of the variable.

Finally, many extensions are of interest. For instance, Clements and Hendry (1993) analyze system-based multivariate forecasts over multiple horizons; and Hendry and Martinez (2017) further develop that framework. In a policy context, it is often valuable to evaluate the discrepancies between the paths of forecasts and to relate policy decisions to the underlying forecasts; see Martinez (2017) and Castle, Hendry, and Martinez (2017) respectively.

8. Conclusions

This paper describes a spectrum of interrelated new and old techniques for evaluating forecasts in general, and forecasts of the government budget in particular. These tools permit rigorous assessment of forecasts and offer directions for their potential improvement. In so doing, these tools help glean the implications of different forecast errors over time and across forecasting techniques, and they provide a basis for understanding the sources of forecast errors.

Appendix. The Data and the Forecasts

This paper analyzes data on U.S. gross federal debt and the one-year-ahead CBO, OMB, and APB forecasts of that debt, as compiled by Martinez (2015) and extended herein. Table A lists those data. See Martinez (2015) and Section 3 above for details, including sources and definitions. Multi-year forecasts (for the hedgehog graphs) appear in the original sources.

Table A: U.S. gross federal debt and the one-year-ahead CBO, OMB, and APB forecasts of that debt.

Year	DEBT	CBO	OMB	APB
1983	1381.886	–	–	–
1984	1576.748	1600.	1591.573	1599.
1985	1827.47	1853.	1841.077	1854.
1986	2129.964	2114.	2112.	2110.6
1987	2355.206	2364.	2372.4	2367.2
1988	2600.679	2598.	2581.6	2603.
1989	2865.664	2865.	2868.8	2869.
1990	3206.26	3131.	3113.3	3150.
1991	3598.919	3606.	3617.837	3616.
1992	4002.815	4039.	4080.3	4058.
1993	4351.149	4392.	4396.7	4391.
1994	4643.996	4690.	4676.	4692.
1995	4920.95	4942.	4961.5	4947.
1996	5181.923	5191.	5207.3	5193.
1997	5369.7	5436.	5453.7	5432.
1998	5478.717	5540.	5543.6	5524.
1999	5606.486	5579.	5614.9	5578.
2000	5629.009	5665.	5686.	5674.
2001	5770.249	5603.	5625.	5627.
2002	6198.129	6043.	6137.1	6117.
2003	6758.722	6620.	6752.	6706.
2004	7352.017	7459.	7486.4	7453.
2005	7902.8	7975.	8031.4	7991.
2006	8448.991	8515.	8611.5	8556.
2007	8948.534	8915.	9007.8	8968.
2008	9983.694	9432.	9654.4	9606.
2009	11873.812	11529.	12867.5	12303.
2010	13526.633	13260.	13786.6	13684.
2011	14762.223	15047.	15476.2	15006.
2012	16048.111	16002.	16350.9	16187.
2013	16716.791	17068.	17249.2	16909.
2014	17792.023	17694.	17892.6	17750.
2015	18117.866	18472.	18627.6	18455.
2016	19537.417	19332.	19433.3	19274.
2017	20203.891	20355.	20354.4	20188.
2018	21460.692	21375.	21478.2	21363.

REFERENCES

Andrews, D. W. K. (1991) "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estmation", *Econometrica*, 59, 3, 817–858.

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

Auerbach, A. J. (1995) "Tax Projections and the Budget: Lessons from the 1980's", *American Economic Review: Papers and Proceedings*, 85, 2, 165–169.

Auerbach, A. J. (1999) "On the Performance and Use of Government Revenue Forecasts", *National Tax Journal*, 52, 4, 767–782.

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

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

Belongia, M. T. (1988) "Are Economic Forecasts by Government Agencies Biased?", *Federal Reserve Bank of St. Louis Review*, 70, 6, 15–23.

Blackley, P. R., and L. DeBoer (1993) "Bias in OMB's Economic Forecasts and Budget Proposals", *Public Choice*, 76, 3, 215–232.

Campbell, B., and E. Ghysels (1995) "Federal Budget Projections: A Nonparametric Assessment of Bias and Efficiency", *Review of Economics and Statistics*, 77, 1, 17–31.

Castle, J. L., D. F. Hendry, and A. B. Martinez (2017) "Evaluating Forecasts, Narratives and Policy Using a Test of Invariance", *Econometrics*, 5, 3 (39), 1–27.

CBO (2002) "CBO's Economic Forecasting Record: A Supplement to The Budget and Economic Oulook: Fiscal Years 2003-2012", Technical Report, Congressional Budget Office, Washington, D.C., February.

CBO (2004) "CBO's Economic Forecasting Record: An Evaluation of Economic Forecasts CBO made from January 1976 through January 2002", Technical Report, Congressional Budget Office, Washington, D.C., September.

CBO (2005) "CBO's Economic Forecasting Record: An Evaluation of Economic Forecasts CBO made from January 1976 through January 2003", Technical Report, Congressional Budget Office, Washington, D.C., October.

CBO (2006) "CBO's Economic Forecasting Record: An Evaluation of Economic Forecasts CBO made from January 1976 through January 2004", Technical Report, Congressional Budget Office, Washington, D.C., November.

CBO (2007) "CBO's Economic Forecasting Record: 2007 Update", Publication No. 3042, Congressional Budget Office, Washington, D.C., November.

CBO (2009) "CBO's Economic Forecasting Record: 2009 Update", Publication No. 3255, Congressional Budget Office, Washington, D.C., July.

CBO (2010) "CBO's Economic Forecasting Record: 2010 Update", Publication No. 4138, Congressional Budget Office, Washington, D.C., July.

CBO (2013) "CBO's Economic Forecasting Record: 2013 Update", Publication No. 4431, Congressional Budget Office, Washington, D.C., January.

CBO (2015a) "CBO's Economic Forecasting Record: 2015 Update", Publication No. 49891, Congressional Budget Office, Washington, D.C., February.

CBO (2015b) "CBO's Revenue Forecasting Record", Publication No. 50831, Congressional Budget Office, Washington, D.C., November.

CBO (2017a) "CBO's Economic Forecasting Record: 2017 Update", Publication No. 53090, Congressional Budget Office, Washington, D.C., October.

CBO (2017b) "An Evaluation of CBO's Past Outlay Projections", Publication No. 53328, Congressional Budget Office, Washington, D.C., November.

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.

Clemen, R. T. (1989) "Combining Forecasts: A Review and Annotated Bibliography", *International Journal of Forecasting*, 5, 4, 559–583.

Clements, M. P., and D. F. Hendry (1993) "On the Limitations of Comparing Mean Square Forecast Errors", *Journal of Forecasting*, 12, 8, 617–637 (with discussion).

Clements, M. P., and D. F. Hendry (1998) *Forecasting Economic Time Series*, Cambridge University Press, Cambridge.

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) "An Overview of Economic Forecasting", Chapter 1 in M. P. Clements and D. F. Hendry (eds.) *A Companion to Economic Forecasting*, Blackwell Publishers, Oxford, 1–18.

Cohen, D., and G. Follette (2003) "Forecasting Exogenous Fiscal Variables in the United States", FEDS Working Paper No. 2003–59, Board of Governors of the Federal Reserve System, Washington, D.C., November.

Coibion, O., and Y. Gorodnichenko (2015) "Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts", *American Economic Review*, 105, 8, 2644–2678.

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.

Croushore, D., and S. Van Norden (2017) "Fiscal Surprises at the FOMC", CIRANO Working Paper No. 2017S–09, CIRANO, Montréal, Canada, April.

Croushore, D., and S. Van Norden (2018) "Fiscal Forecasts at the FOMC: Evidence from the Greenbooks", *Review of Economics and Statistics*, 100, 5, 933–945.

Diebold, F. X. (1989) "Forecast Combination and Encompassing: Reconciling Two Divergent Literatures", *International Journal of Forecasting*, 5, 4, 589–592.

Diebold, F. X. (2015) "Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests", *Journal of Business and Economic Statistics*, 33, 1, 1–9.

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).

Engle, R. F. (1982) "A General Approach to Lagrange Multiplier Model Diagnostics", *Journal of Econometrics*, 20, 1, 83–104.

Engle, R. F. (1984) "Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics", Chapter 13 in Z. Griliches and M. D. Intriligator (eds.) *Handbook of Econometrics*, Volume 2, North-Holland, Amsterdam, 775–826.

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. (1993) "On the Limitations of Comparing Mean Square Forecast Errors: Clarifications and Extensions", *Journal of Forecasting*, 12, 8, 644–651.

Ericsson, N. R. (2011) "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. (2017a) "How Biased Are U.S. Government Forecasts of the Federal Debt?", *International Journal of Forecasting*, 33, 2, 543–559.

Ericsson, N. R. (2017b) "Interpreting Estimates of Forecast Bias", *International Journal of Forecasting*, 33, 2, 563–568.

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., and J. Marquez (1993) "Encompassing the Forecasts of U.S. Trade Balance Models", *Review of Economics and Statistics*, 75, 1, 19–31.

Ericsson, N. R., and J. Marquez (1998) "A Framework for Economic Forecasting", *Econometrics Journal*, 1, 1, C228–C266.

Ericsson, N. R., and A. B. Martinez (2019) "Evaluating Government Budget Forecasts", Chapter 3 in D. W. Williams and T. D. Calabrese (eds.) *The Palgrave Handbook of Government Budget Forecasting*, Palgrave Macmillan, Cham, Switzerland, 37–69.

Feenberg, D. R., W. Gentry, D. Gilroy, and H. S. Rosen (1989) "Testing the Rationality of State Revenue Forecasts", *Review of Economics and Statistics*, 71, 2, 300–308.

Fisher, R. A. (1922) "The Goodness of Fit of Regression Formulae, and the Distribution of Regression Coefficients", *Journal of the Royal Statistical Society*, 85, 4, 597–612.

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

Frendreis, J., and R. Tatalovich (2000) "Accuracy and Bias in Macroeconomic Forecasting by the Administration, the CBO, and the Federal Reserve Board", *Polity*, 32, 4, 623–632.

Gentry, W. M. (1989) "Do State Revenue Forecasters Utilize Available Information?", *National Tax Journal*, 42, 4, 429–439.

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

Granger, C. W. J. (1983) "Forecasting White Noise", in A. Zellner (ed.) *Applied Time Series Analysis of Economic Data*, Bureau of the Census, Washington, D.C., 308–314.

Granger, C. W. J. (1989a) "Combining Forecasts—Twenty Years Later", *Journal of Forecasting*, 8, 167–173.

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

Granger, C. W. J. (1999) "Outline of Forecast Theory using Generalized Cost Functions", *Spanish Economic Review*, 1, 2, 161–173.

Hansen, B. E. (2007) "Least Squares Model Averaging", *Econometrica*, 75, 4, 1175–1189.

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 M. P. Clements (2004) "Pooling of Forecasts", *Econometrics Journal*, 7, 1, 1–31.

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., and S. Johansen (2015) "Model Discovery and Trygve Haavelmo's Legacy", *Econometric Theory*, 31, 1, 93–114.

Hendry, D. F., and A. B. Martinez (2017) "Evaluating Multi-step System Forecasts with Relatively Few Forecast-error Observations", *International Journal of Forecasting*, 33, 2, 359–372.

Howard, J. A. (1987) "Government Economic Projections: A Comparison Between CBO and OMB Forecasts", *Public Budgeting and Finance*, 7, 3, 14–25.

Huntley, J., and E. Miller (2009) "An Evaluation of CBO Forecasts", CBO Working Paper Series No. 2009–02, Congressional Budget Office, Washington, D.C., August.

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 (2016) "Asymptotic Theory of Outlier Detection Algorithms for Linear Time Series Regression Models", *Scandinavian Journal of Statistics*, 43, 2, 321–381 (with discussion and rejoinder).

Kamlet, M. S., D. C. Mowery, and T.-T. Su (1987) "Whom Do You Trust? An Analysis of Executive and Congressional Economic Forecasts", *Journal of Policy Analysis and Management*, 6, 3, 365–384.

Kitchen, J. (2003) "Observed Relationships Between Economic and Technical Receipts Revisions in Federal Budget Projections", *National Tax Journal*, 56, 2, 337–353.

Kliesen, K. L., and D. L. Thornton (2001) "The Expected Federal Budget Surplus: How Much Confidence Should the Public and Policymakers Place in the Projections?", *Federal Reserve Bank of St. Louis Review*, 83, 2, 11–24.

Kliesen, K. L., and D. L. Thornton (2012) "How Good Are the Government's Deficit and Debt Projections and Should We Care?", *Federal Reserve Bank of St. Louis Review*, 94, 1, 21–39.

Krause, G. A., and J. W. Douglas (2005) "Institutional Design Versus Reputational Effects on Bureaucratic Performance: Evidence from US Government Macroeconomic and Fiscal Projections", *Journal of Public Administration Research and Theory*, 15, 2, 281–306.

Krause, G. A., and J. W. Douglas (2006) "Does Agency Competition Improve the Quality of Policy Analysis? Evidence from OMB and CBO Fiscal Projections", *Journal of Policy Analysis and Management*, 25, 1, 53–74.

Krol, R. (2014) "Forecast Bias of Government Agencies", *Cato Journal*, 34, 1, 99–112.

Lipford, J. W. (2001) "How Transparent is the US Budget?", *The Independent Review*, 5, 4, 575–591.

Marquez, J., and N. R. Ericsson (1993) "Evaluating Forecasts of the U.S. Trade Balance", Chapter 14 in R. C. Bryant, P. Hooper, and C. L. Mann (eds.) *Evaluating Policy Regimes: New Research in Empirical Macroeconomics*, Brookings Institution, Washington, D.C., 671–732.

Martinez, A. B. (2011) "Comparing Government Forecasts of the United States' Gross Federal Debt", RPF Working Paper No. 2011–002, Research Program on Forecasting, Center of Economic Research, Department of Economics, The George Washington University, Washington, D.C., February.

Martinez, A. B. (2015) "How Good Are US Government Forecasts of the Federal Debt?", *International Journal of Forecasting*, 31, 2, 312–324.

Martinez, A. B. (2017) "Testing for Differences in Path Forecast Accuracy: Forecast-Error Dynamics Matter", Working Paper No. 17–17, Federal Reserve Bank of Cleveland, Cleveland, Ohio, November.

McNees, S. K. (1995) "An Assessment of the 'Official' Economic Forecasts", *New England Economic Review*, 1995, July/August, 13–24.

Miller, S. M. (1991) "Forecasting Federal Budget Deficits: How Reliable Are US Congressional Budget Office Projections?", *Applied Economics*, 23, 12, 1789–1799.

Mincer, J., and V. Zarnowitz (1969) "The Evaluation of Economic Forecasts", Chapter 1 in J. Mincer (ed.) *Economic Forecasts and Expectations: Analyses 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.

Patton, A. J., and A. Timmermann (2012) "Forecast Rationality Tests Based on Multi-Horizon Bounds", *Journal of Business and Economic Statistics*, 30, 1, 1–17.

Penner, R. G. (2001) *Errors in Budget Forecasting*, Urban Institute, Washington, D.C., April.

Penner, R. G. (2002) "Dealing with Uncertain Budget Forecasts", *Public Budgeting and Finance*, 22, 1, 1–18.

Penner, R. G. (2008) "Federal Revenue Forecasting", Chapter 2 in J. Sun and T. D. Lynch (eds.) *Government Budget Forecasting: Theory and Practice*, CRC Press, Boca Raton, Florida, 31–46.

Plesko, G. A. (1988) "The Accuracy of Government Forecasts and Budget Projections", *National Tax Journal*, 41, 4, 483–501.

Salkever, D. S. (1976) "The Use of Dummy Variables to Compute Predictions, Prediction Errors, and Confidence Intervals", *Journal of Econometrics*, 4, 4, 393–397.

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.

Sun, J. (2008) "Forecast Evaluation: A Case Study", Chapter 10 in J. Sun and T. D. Lynch (eds.) *Government Budget Forecasting: Theory and Practice*, CRC Press, Boca Raton, Florida, 223–240.

Timmermann, A. (2006) "Forecast Combinations", Chapter 4 in G. Elliott, C. W. J. Granger, and A. Timmermann (eds.) *Handbook of Economic Forecasting*, Volume 1, Elsevier, 135–196.

Tsuchiya, Y. (2016) "Directional Analysis of Fiscal Sustainability: Revisiting Domar's Debt Sustainability Condition", *International Review of Economics and Finance*, 41, January, 189–201.

Williams, D. W., and T. D. Calabrese (2016) "The Status of Budget Forecasting", *Journal of Public and Nonprofit Affairs*, 2, 2, 127–160.