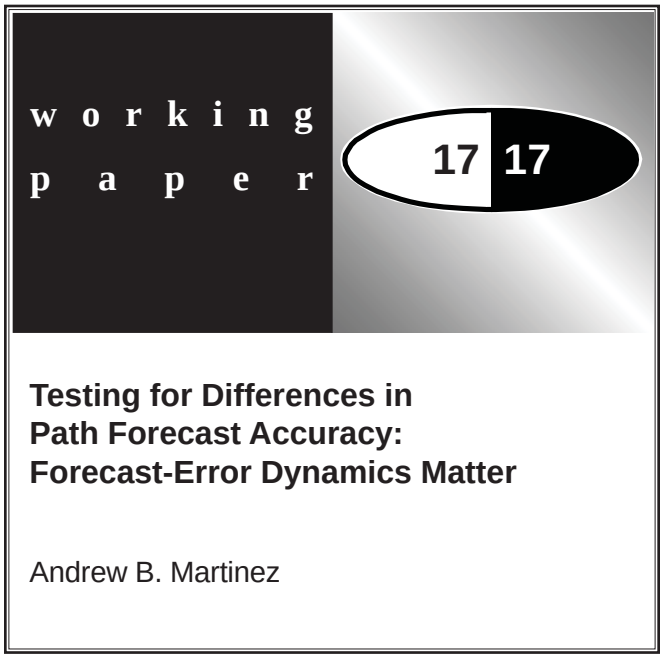


Testing for Differences in Path Forecast Accuracy: Forecast-Error Dynamics Matter

Andrew B. Martinez

FEDERAL RESERVE BANK OF CLEVELAND

ISSN: 2573-7953

Working Paper 17-17 November 2017

Testing for Differences in Path Forecast Accuracy: Forecast-Error Dynamics Matter

Andrew B. Martinez

Although the trajectory and path of future outcomes plays an important role in policy decisions, analyses of forecast accuracy typically focus on individual point forecasts. However, it is important to examine the path forecasts errors since they include the forecast dynamics. We use the link between path forecast evaluation methods and the joint predictive density to propose a test for differences in system path forecast accuracy. We also demonstrate how our test relates to and extends existing joint testing approaches. Simulations highlight both the advantages and disadvantages of path forecast accuracy tests in detecting a broad range of differences in forecast errors. We compare the Federal Reserve's Greenbook point and path forecasts against four DSGE model forecasts. The results show that differences in forecast-error dynamics can play an important role in the assessment of forecast accuracy.

JEL codes: C12, C22, C52, C53.

Keywords: GFESM, log determinant, log score, mean square error.

Suggested citation: Martinez, Andrew B., 2017. "Testing for Differences in Path Forecast Accuracy: Forecast-Error Dynamics Matter," Federal Reserve Bank of Cleveland, Working Paper no. 17-17. https://doi.org/10.26509/frbc-wp-201717.

Andrew Martinez is at the Department of Economics and Institute for New Economic Thinking, Oxford Martin School, University of Oxford, UK (andrew.martinez@economics.ox.ac.uk). Part of this research was conducted while the author was a dissertation intern at the Federal Reserve Bank of Cleveland. This research was also supported in part by a grant from the Robertson Foundation (grant 9907422). The author is grateful for comments and suggestions from Jennifer L. Castle, Todd E. Clark, Michael P. Clements, David F. Hendry, Ryoko Ito, Felix Pretis, participants at the 17th OxMetrics conference, the Oxford Econometrics Lunch Seminar, and the Federal Reserve Bank of Cleveland brownbag seminar. He also thanks Bent Nielsen and Xiyu Jiao for helpful discussions and Maik Wolters for sharing his forecasts. All numerical results and figures were obtained using OxMetrics 7.2 (OSX/U); see Doornik (2013). The manuscript was prepared with LyX 2.2.3.

"Any quibbles I might have with the Greenbook over point estimates for growth are minor because the baseline forecast path in the Greenbook is consistent with our own Atlanta model forecast." Jack Guynn (President of the Federal Reserve Bank of Atlanta) FOMC Meeting Transcript, December 9, 2003

1 Introduction

Path forecasts play an important role in policy decisions. For example, central banks simultaneously care about the trajectory of inflation, output and unemployment. Monetary policy reacts to the expected future path of the economy given lags in policy implementation and the monetary transmission mechanism. Fiscal policy is also in part judged by its expected impact on the trajectory of the debt and deficit; see Martinez (2015). Similarly, disaster management agencies follow the future track and intensity of a tropical cyclone. Decisions about evacuations and emergency responses depend on how the storm is forecast to evolve over time. These simultaneous concerns demonstrate the importance of path forecasts.

A system's forecast path contains additional information beyond the point forecasts at each horizon. It captures the expected dynamics of the individual processes as well as the dynamics of the relationships between variables in the system. Focusing exclusively on the accuracy of the individual point forecasts ignores important aspects of the future trajectory. For example, Schorfheide and Song (2015) find that improvements in nowcasts can lead to improvements in longer-horizon forecasts. This implicitly suggests that there is a high degree of persistence in the underlying forecast. Examining the forecast path allows us to assess this question directly.

The focus on the path of the forecast rather than the point forecasts is a fairly recent development. The evaluation of path forecasts using uncertainty bands was first discussed by Jordà and Marcellino (2010). This spurred a growing literature on the construction of simultaneous confidence bands associated with the path forecasts; see Jordà et al. (2013), Wolf and Wunderli (2015) and Montiel Olea and Plagborg-Møller (2017) among others.

This paper takes a different approach. Rather than computing the uncertainty surrounding the path, we are interested in the path's accuracy. Moreover, we are interested in testing for differences in accuracy between alternative path forecasts. This enables us to understand and detect differences in how well forecasters capture the underlying dynamics of the system.

Assessments of path forecast accuracy cannot rely on traditional point accuracy metrics. Instead, we show that the general forecast-error second-moment (hereafter GFESM), proposed by Clements and Hendry (1993), can be reinterpreted as a metric of path forecast accuracy. The GFESM is effectively a multi-horizon

generalization of the log determinant measure proposed by Doan et al. (1984). The log determinant is widely used to evaluate forecasts of vector autoregressive (VAR) and Dynamic Stochastic General Equilibrium (DSGE) models.[1] While the log determinant captures covariances between variables, the GFESM also captures the covariances across horizons and between variables across horizons.

Despite an extensive literature on testing for differences in point forecast accuracy, there are no tests for differences in path forecast accuracy (see Clark and McCracken 2013b). We use the relationship between the GFESM and the joint density to construct a general likelihood ratio test for differences in path forecast accuracy. We explore how this test statistic differs from existing tests and examine and derive its properties for a special case.

Monte Carlo simulations are used to illustrate the trade-offs associated with our test. We find that although our path forecast accuracy test has lower power to detect differences in biases, it has much higher power to capture differences in variances, covariances and dynamics across forecast models. This is particularly true for differences in forecast-error dynamics, which other tests are unable to capture. Thus, our test captures a new aspect of differences in forecast accuracy. We also demonstrate that the test statistic can be extended to higher dimensional settings.

Finally, we compare the Federal Reserve Board's Greenbook point and path forecasts with four DSGE model forecasts. Although there are important differences in the point and path forecast results, they often complement one another and provide further support to the general findings. We find that the Greenbook dominates in the point and path forecasts of inflation. However, the DSGE models ability to track the long-run mean does well for the GDP growth forecasts over the Great Moderation. Furthermore, while the Greenbook has a solid advantage in point forecasts of interest rates, the DSGE forecasts demonstrate additional value in the path forecasts even after accounting for differences in the nowcasts. We also find mixed effects for changes in the nowcasts on path and point forecast accuracy.

The rest of the paper is structured as follows. The next section introduces measures of path forecast accuracy. Section 3 constructs a likelihood ratio of path forecast accuracy and compares it against existing approaches. Section 4 examines the properties of a special case of the path forecast accuracy test. Section 5 compares the test against alternative loss functions in various simulation settings. Section 6 compares the Federal Reserve Board's Greenbook path forecasts against the path forecasts from four DSGE models. Section 7 concludes.

[1] For example, see Adolfson et al. (2007); Del Negro et al. (2007); Schorfheide and Song (2015); Berg (2016).

2 Measuring path forecast accuracy

We begin by illustrating how the general matrix of the forecast-error second-moment and its determinant, the GFESM, can be interpreted as measures of path forecast accuracy. We also show how measures of path forecast accuracy relate to the mean square forecast error (MSE), the trace statistic, and the log determinant.

Let $\mathbf{y}_t$ be a K-dimensional random vector. We can denote point forecasts of this vector as $\widehat{\mathbf{y}}_t^m(h) = \widehat{E}_t(\mathbf{y}_{t+h} \mid \mathbf{y}_t, \mathbf{y}_{t-1}, \ldots)$ where the sample size used to estimate the parameters required to generate $\widehat{\mathbf{y}}_t^m(h)$ is m and $\{t : m \leq t \leq T\}$. Then the point forecast errors at each horizon are

$$\widetilde{\mathbf{u}}_t^m(h) = \mathbf{y}_{t+h} - \widehat{\mathbf{y}}_t^m(h). \tag{2.1}$$

Point forecasts can be evaluated using any number of loss functions, $L(\mathbf{y}_{t+h}, \widehat{\mathbf{y}}_t^m(h))$. The most common is the quadratic MSE loss function. In multivariate systems, this becomes

$$\widehat{\mathbf{\Sigma}}_{h,N}^m = \frac{1}{N}\sum_{t=m}^{T} \widetilde{\mathbf{u}}_t^m(h)\widetilde{\mathbf{u}}_t^m(h)', \tag{2.2}$$

where $N = T - m + 1$ is the sample of forecast-error observations. A common way to summarize the information within the MSE matrix is to calculate the trace of $\widehat{\mathbf{\Sigma}}_{h,N}^m$, denoted by $Tr\left(\widehat{\mathbf{\Sigma}}_{h,N}^m\right) = \frac{1}{N}\sum_{t=m}^{T} \widetilde{\mathbf{u}}_t^m(h)'\widetilde{\mathbf{u}}_t^m(h)$, which ignores the relationship between errors of different variables. Alternatively, Doan et al. (1984) compute the (log) determinant: $\left|\widehat{\mathbf{\Sigma}}_{h,N}^m\right|$ which captures the covariances between variables. See Komunjer and Owyang (2012) for a multivariate approach that deviates from the MSE loss function.

If $\widehat{\mathbf{Y}}_t^m(H)$ and $\mathbf{Y}_{t,H}$ denote the predicted and observed paths, then the vector of path forecast errors is

$$\widetilde{\mathbf{U}}_{t,H}^m = \mathbf{Y}_{t,H} - \widehat{\mathbf{Y}}_t^m(H) = \begin{bmatrix} \mathbf{y}_{t+1} \\ \vdots \\ \mathbf{y}_{t+H} \end{bmatrix} - \begin{bmatrix} \widehat{\mathbf{y}}_t^m(1) \\ \vdots \\ \widehat{\mathbf{y}}_t^m(H) \end{bmatrix}. \tag{2.3}$$

In order to evaluate the path forecast errors, we can use the general loss function proposed by Clements and Hendry (1993). Their general matrix of the forecast-error second-moment (GMFESM) extends the MSE matrix in (2.2) to multiple horizons and is computed as the mean squared path forecast errors

$$\widehat{\mathbf{\Phi}}_{H,N}^m = \frac{1}{N}\sum_{t=m}^{T} \widetilde{\mathbf{U}}_{t,H}^m \widetilde{\mathbf{U}}_{t,H}^{m\prime}, \tag{2.4}$$

where each K-dimensional block along the main diagonal of (2.4) represents $\widehat{\mathbf{\Sigma}}_{h,N}^m$ for $h = 1, ..., H$. The off diagonals are co-movements between horizons and variables. The determinant $\left|\widehat{\mathbf{\Phi}}_{H,H}^m\right|$ (GFESM) summarizes this information by multiplying the MSEs conditional on the covariances across variables and horizons. Clements and Hendry (1995, 1997) demonstrate that the GFESM captures changes in forecast dynamics, whereas traditional metrics do not.

2.1 Joint densities as general loss functions

The ability to capture forecast dynamics is necessary for a path forecast accuracy metric. Granger (1999) illustrates that the joint predictive distribution is required to evaluate path forecast accuracy. It is possible to show that the GFESM is directly related to a specific form of the joint predictive distribution and so is interpretable as a path forecast accuracy metric.

Following Jordà et al. (2013), we can assume that the distribution of the path forecast errors is iid and elliptically contoured:

$$f_U\left(\widetilde{\mathbf{U}}^m_{t,H}\right) = C\left[\left|\mathbf{\Phi}^m_H\right|\right]^{-1/2} exp\left\{-g\left[\left(\widetilde{\mathbf{U}}^m_{t,H}\right)'\left(\mathbf{\Phi}^m_H\right)^{-1}\left(\widetilde{\mathbf{U}}^m_{t,H}\right)\right]\right\}, \tag{2.5}$$

where C is a normalizing constant, $g(\cdot)$ is a measurable density function and $\mathbf{\Phi}^m_H$ is positive definite. While the assumption of an elliptical distribution is relatively strong, it encompasses a broad class of relevant distributions including the multivariate normal and t distributions. The elliptical distribution also imposes that the loss function is symmetric and allows for a general correlation structure across variables and horizons while accommodating fat-tails.

Given this distributional assumption, we can construct a predictive likelihood following Clements and Hendry (1998).[2] We construct a quasi-profile predictive likelihood by taking the forecasts as being maximized conditional on current and past observations and then concentrating the likelihood further, taking logs and ignoring the constant so that

$$\ell_p\left(\widehat{\mathbf{\Phi}}^m_{H,N}|\widehat{\boldsymbol{Y}}^m_t(H)\right) = -\frac{N}{2}ln\left(\left|\widehat{\mathbf{\Phi}}^m_{H,N}\right|\right) - \sum_{t=m}^{T} g_t\left[\left(\widetilde{\mathbf{U}}^m_{t,H}\right)'\left(\widehat{\mathbf{\Phi}}^m_{H,N}\right)^{-1}\left(\widetilde{\mathbf{U}}^m_{t,H}\right)\right], \tag{2.6}$$

which depends on the scaled log GFESM and an additional term. This additional term drops out for the multivariate normal density since $\widehat{\mathbf{\Phi}}^m_{H,N} = \frac{1}{N}\sum_{t=m}^{T}\widetilde{\mathbf{U}}^m_{t,H}\widetilde{\mathbf{U}}^{m\prime}_{t,H}$. This illustrates that the distributional assumption imposed on the path forecast errors can be viewed more generally as an assumption about the loss function.

Interpreting the joint density as a general loss function allows us to appeal to scoring rules to evaluate path forecast accuracy. Scoring rules are frequently used as loss functions in the evaluation of density forecasts. A common class of scoring rules approximate the normal density and only depend on the first and second moments; see Gneiting and Raftery (2007). These scoring rules are also closely related to a wider class of loss functions; see Cichocki et al. (2015). The next section constructs a likelihood ratio test of differences in path forecast accuracy using the joint density as a loss function.

[2] Bjørnstad (1990) notes that the predictive likelihood was first proposed by Hinkley (1979) and Mathiasen (1979) to extend the likelihood framework to forecasting.

3 A Path Forecast Accuracy Likelihood Ratio Test

This section shows how we can construct a path forecast accuracy test using the joint density as a general loss function. We start by describing the notation and the environment. Next we describe the test and the necessary assumptions. Finally we explore the relationship between this test and alternative approaches.

Consider a stochastic process $\mathbf{Z} \equiv \{\mathbf{z}_t : \Omega \rightarrow \mathbb{R}^s,\ K+S=s \in \mathbb{N}.\ t=1,2,...\}$ defined on a complete probability space (Ω, F, P) and partition the observed vector $\mathbf{z}_t$ as $\mathbf{z}_t \equiv \left(\mathbf{y}_t, \mathbf{x}_t'\right)'$, where $\mathbf{y}_t : \Omega \rightarrow \mathbb{R}^K$ is the vector of variables of interest and $\mathbf{x}_t : \Omega \rightarrow \mathbb{R}^S$ is a vector of predictors. Let $\mathbf{F}_t = \sigma\left(\mathbf{z}_t, \mathbf{z}_{t-1}, \ldots, \mathbf{z}_{t-m+1}; \mathbf{\Pi}_{m,t}\right)$ be the information set at time t and suppose that two competing models (1 and 2) are used to produce a system of path forecasts for the stacked HK vector of variables of interest, $\mathbf{Y}_{t,H}$, using the information in F_t. Denote the path forecasts by $\widehat{\mathbf{Y}}^m_{t,1}(H) \equiv l\left(\mathbf{z}_t, \mathbf{z}_{t-1}, \ldots, \mathbf{z}_{t-m+1}; \hat{\mathbf{\Pi}}_{1,m,t}\right)$ and $\widehat{\mathbf{Y}}^m_{t,2}(H) \equiv v\left(\mathbf{z}_t, \mathbf{z}_{t-1}, \ldots, \mathbf{z}_{t-m+1}; \hat{\mathbf{\Pi}}_{2,m,t}\right)$ where $l(\cdot)$ and $v(\cdot)$ are measurable functions. Subscripts indicate that the time t forecasts are measurable functions of the m most recent observations where m is finite.[3]

The $q \times HK$ matrices $\hat{\mathbf{\Pi}}_{1,m}$ and $\hat{\mathbf{\Pi}}_{2,m}$ collect the parameter estimates from each model respectively. Note that the only requirements that we impose on how the forecasts are produced is that they are measurable functions estimated jointly as a system on a finite estimation window. This allows the forecasts to be produced by a wide range of methods. However, it does require that the same method is used for all variables and horizons for a given model.

Out-of-sample forecast evaluation is performed using a "rolling window" estimation scheme. Let $T+H$ be the total sample size. The forecasts are produced at time m using data indexed $1, \ldots, m$ and the path forecast errors are generated as $\widetilde{\mathbf{U}}^m_{m,H,i} = \mathbf{Y}_{m,H} - \widehat{\mathbf{Y}}^m_{m,i}(H)\ \forall\ i \in \{1,2\}$. The estimation window is rolled forward one observation and the second forecasts are obtained using observations $2, \ldots, m+1$ where the path forecast errors are generated as $\widetilde{\mathbf{U}}^m_{m+1,H,i} = \mathbf{Y}_{m+1,H} - \widehat{\mathbf{Y}}^m_{m+1,i}(H)$. The procedure is thus iterated, so that the last forecasts are obtained using observations $T-m, ..., T$, so that the path forecast errors are generated as $\widetilde{\mathbf{U}}^m_{T,H,i} = \mathbf{Y}_{T,H} - \widehat{\mathbf{Y}}^m_{T,i}(H)$. This yields a sequence of $N = T-m+1$ path forecast-error observations.

This setup is almost identical to Giacomini and White (2006) and Amisano and Giacomini (2007). The main difference is that we are interested in a vector of path forecasts across multiple variables rather than a single variable / horizon at a time. While the framework used here closely follows the unconditional predictive ability test from Giacomini and White (2006), using the log predictive density as a loss function also links our approach with the likelihood ratio test for density forecasts in Amisano and Giacomini (2007).[4]

[3] Note that as in Giacomini and White (2006), m can vary across forecasting systems. In that case we can redefine m as the maximum of the recent observations used across the two forecasting systems.

[4] The unconditional predictive ability test is similar to Diebold and Mariano 1995 and West 1996. However, as discussed in Clark and McCracken (2013a), the former focuses on finite sample results while the latter focus on the population. This difference has

In this sense our framework is also related to Vuong (1989), Mitchell and Hall (2005) and Bao et al. (2007).

We restrict our attention to the log score rule $S(f,U) = ln\{f(U)\}$ which coincides with the log profile predictive likelihood above in (2.6). Unlike in Amisano and Giacomini (2007) where the density, $f(\cdot)$, is chosen by the forecaster, here the density is specified as the loss function and is the same across forecasters. While for density forecasts, the forecaster chooses the density (often based on the data), here it is chosen by the forecast evaluator and is not necessarily directly related to the underlying data generating process but is instead associated with an acceptable loss profile. For two alternative forecasts let

$$LR_{m,t,H} = ln\left\{f_t\left(\widetilde{\mathbf{U}}^m_{t,H,1}\right)\right\} - ln\left\{f_t\left(\widetilde{\mathbf{U}}^m_{t,H,2}\right)\right\}. \tag{3.1}$$

We can also allow for alternative weighting at each horizon. To see this, note that the joint density is equal to the product of the conditional and marginal densities. This conditioning has a unique ordering for path forecast errors such that the joint density can be decomposed as

$$f_t\left(\widetilde{\mathbf{U}}^m_{t,H}\right) = \prod_{h=1}^{H} f_t\left(\tilde{\mathbf{u}}^m_t(h) \mid \tilde{\mathbf{u}}^m_t(1),\ldots,\tilde{\mathbf{u}}^m_t(h-1)\right), \tag{3.2}$$

where the forecast error density at each horizon is conditional on all previous horizons. Given this decomposition, we can construct a weighted test of differences in path forecast accuracy as

$$WLR_{m,t,H} = \sum_{h=1}^{H} w_h\left[ln\left\{f_t\left(\tilde{\mathbf{u}}^m_{t,1}(h) \mid \tilde{\mathbf{u}}^m_{t,1}(1),\ldots,\tilde{\mathbf{u}}^m_{t,1}(h-1)\right)\right\} - ln\left\{f_t\left(\tilde{\mathbf{u}}^m_{t,2}(h) \mid \tilde{\mathbf{u}}^m_{t,2}(1),\ldots,\tilde{\mathbf{u}}^m_{t,2}(h-1)\right)\right\}\right], \tag{3.3}$$

where w_h is the weight assigned to each horizon. In practice, this allows for longer horizons to receive less weight than shorter horizons while accounting for the dependence between them. Re-weighting the conditional and marginal densities differs from Amisano and Giacomini (2007) who re-weight the entire density. A test for equal performance of the weighted path forecast systems can be formulated as

$$H_0 : \mathbb{E}\left[WLR_{m,t,H}\right] = 0, \quad t = 1,2,.... \text{ against} \tag{3.4}$$

$$H_A : \mathbb{E}\left[\overline{WLR}_{m,N,H}\right] \neq 0 \quad \text{for all } N \text{ sufficiently large,} \tag{3.5}$$

where $\overline{WLR}_{m,N,H} = \frac{1}{N}\sum_{t=m}^{T} WLR_{m,t,H}$. Then the test is based on the following statistic

$$t_{m,N,H} = \frac{\sqrt{N} * \overline{WLR}_{m,N,H}}{\hat{\sigma}_N}, \tag{3.6}$$

where $\hat{\sigma}^2_N$ is a heteroskedasticity and autocorrelation consistent (HAC) estimator of asymptotic variance $\sigma^2_N = var\left[\sqrt{N} * \overline{WLR}_{m,N,H}\right]$. See Andrews (1991) and Lazarus et al. (2017).

A level α test rejects the null hypothesis of equal performance of the path forecast systems 1 and 2 whenever $abs\left(t_{m,N,H}\right) > z_{\alpha/2}$, where $z_{\alpha/2}$ is the $(1-\alpha/2)$ quantile of a standard normal distribution. The

implications for the null hypothesis.

following theorem provides asymptotic justification:

Theorem 1 (Likelihood ratio test). For a given estimation window size $m < \infty$, and a weight w_h, $0 \leq \sum_{h=1}^{H} w_h < \infty$, suppose that

1. $\{z_t\}$ is a mixing sequence with ϕ of size $-r/(2r-2)$, $r \geq 2$, or α of size $-r/(r-2)$, $r > 2$;
2. $\mathbb{E}\left[ln\left\{f_t\left(\widetilde{\mathbf{U}}^m_{t,H,i}\right)\right\}\right]^{2r} < \infty$ for all t and $i \in \{1,2\}$;
3. $\sigma^2_N = var\left[\sqrt{N} * \overline{WLR}_{m,N,H}\right] > 0$ for all N sufficiently large.

Then (a) under H_0 in (3.4), $t_{m,N,H} \xrightarrow{D} N(0,1)$ as $N \to \infty$ and (b) under H_A in (3.5), for some constant $c \in \mathbb{R}$, $P\left[abs\left(t_{m,N,H}\right) > c\right] \to 1$ as $N \to \infty$. The proof follows directly from Giacomini and White (2006, Proof of Theorem 4) where $\mathbf{W}_t \equiv \mathbf{z}_t$ and $\Delta L_{m,t} \equiv WLR_{m,t,H}$.

Choosing $f(\cdot)$ from the class of elliptical densities illustrates the link with the GFESM. Although any number of multivariate densities can be used, the elliptical density is symmetric and coincides closely with the typical MSE loss function. Given this assumption and assuming that $w_h = 1$, then (3.3) becomes

$$\begin{aligned} WLR_{m,t,H} = \frac{1}{2}\left[ln\left(\left|\widehat{\mathbf{\Phi}}^m_{H,N,2}\right|\right) - ln\left(\left|\widehat{\mathbf{\Phi}}^m_{H,N,1}\right|\right)\right] \\ + \left[g_t\left\{\left(\widetilde{\mathbf{U}}^m_{t,H,2}\right)'\left(\widehat{\mathbf{\Phi}}^m_{H,N,2}\right)^{-1}\left(\widetilde{\mathbf{U}}^m_{t,H,2}\right)\right\} - g_t\left\{\left(\widetilde{\mathbf{U}}^m_{t,H,1}\right)'\left(\widehat{\mathbf{\Phi}}^m_{H,N,1}\right)^{-1}\left(\widetilde{\mathbf{U}}^m_{t,H,1}\right)\right\}\right], \end{aligned} \tag{3.7}$$

so that the test statistic in (3.6) can be reformulated as

$$t_{m,N,H} = \frac{\sqrt{N}}{2\hat{\sigma}_N}\left[ln\left(\left|\widehat{\mathbf{\Phi}}^m_{H,N,2}\right|\right) - ln\left(\left|\widehat{\mathbf{\Phi}}^m_{H,N,1}\right|\right) + o_{m,N,H}\right], \tag{3.8}$$

where $o_{m,N,H} = \frac{2}{N}\sum_{t=m}^{T-1}\left[g_t\left\{\left(\widetilde{\mathbf{U}}^m_{t,H,2}\right)'\left(\widehat{\mathbf{\Phi}}^m_{H,N,2}\right)^{-1}\left(\widetilde{\mathbf{U}}^m_{t,H,2}\right)\right\} - g_t\left\{\left(\widetilde{\mathbf{U}}^m_{t,H,1}\right)'\left(\widehat{\mathbf{\Phi}}^m_{H,N,1}\right)^{-1}\left(\widetilde{\mathbf{U}}^m_{t,H,1}\right)\right\}\right]$. Alternatively, if we assume that density is multivariate normal so that $o_{m,N,H} = 0$ and allow for a general weighting function then the test statistic becomes

$$t_{m,N,H} = \frac{\sqrt{N}}{2\hat{\sigma}_N}\sum_{h=1}^{H} w_h\left[ln\left(\left|\widehat{\mathbf{\Sigma}}^m_{2,h|(0,\ldots,h-1)}\right|\right) - ln\left(\left|\widehat{\mathbf{\Sigma}}^m_{1,h|(0,\ldots,h-1)}\right|\right)\right], \tag{3.9}$$

where $ln\left|\widehat{\mathbf{\Sigma}}^m_{h|(0,\ldots,h-1)}\right|$ is the log determinant of the MSE matrix at each horizon, see (2.2), which is orthogonalized by all prior horizons.[5] Then (3.9) is a weighted test for differences in the conditional log determinants of the MSE matrix at each horizon. If $w_h = 1$, (3.9) is interpretable as a test for differences in the log GFESM since $ln\left(\left|\widehat{\mathbf{\Phi}}^m_{H,N}\right|\right) = \sum_{h=1}^{H} ln\left(\left|\widehat{\mathbf{\Sigma}}^m_{h|(0,\ldots,h-1)}\right|\right)$.

3.1 Advantages of the path forecast approach

The difference between the proposed test statistic and existing approaches is seen most clearly by exploring the link with joint tests of predictive accuracy. Setting the density of the score function to be multivariate

[5] For example, the conditional MSE matrix at the second horizon is $\widehat{\mathbf{\Sigma}}_{2|1} = \widehat{\mathbf{\Sigma}}_2 - \widehat{\mathbf{\Sigma}}_{2.1}\widehat{\mathbf{\Sigma}}_1^{-1}\widehat{\mathbf{\Sigma}}_{1.2}$, where $\widehat{\mathbf{\Sigma}}_{1.2} = \widehat{\mathbf{\Sigma}}'_{2.1}$ is the co-movement between the forecast errors at the first and second horizons.

normal, then (3.8) can be expanded as:

$$t_{m,N,H} = \frac{\sqrt{N}}{2\hat{\sigma}_N}\left[ln\left(\left|\widehat{\mathbf{\Phi}}^m_{H,N,2}\right|\right) - ln\left(\left|\widehat{\mathbf{\Phi}}^m_{H,N,1}\right|\right)\right. \tag{3.10}$$
$$+ \frac{1}{N}\sum_{t=m}^{T}\left\{\left(\widetilde{\mathbf{U}}^m_{t,H,2}\right)'\left(\left(\widehat{\mathbf{\Phi}}^m_{H,N,2}\right)^{-1} - \mathbf{I}_{HK}\right)\left(\widetilde{\mathbf{U}}^m_{t,H,2}\right) - \left(\widetilde{\mathbf{U}}^m_{t,H,1}\right)'\left(\left(\widehat{\mathbf{\Phi}}^m_{H,N,1}\right)^{-1} - \mathbf{I}_{HK}\right)\left(\widetilde{\mathbf{U}}^m_{t,H,1}\right)\right\}$$
$$\left. + \sum_{h=1}^{H}\sum_{k=1}^{K}\left(\frac{1}{N}\sum_{t=m}^{T}\left\{\left(\tilde{u}^m_{k,t,2}(h)\right)^2 - \left(\tilde{u}^m_{k,t,1}(h)\right)^2\right\}\right)\right],$$

where $\tilde{u}^m_{k,t,i}(h)$ is the kth element in $\tilde{\mathbf{u}}^m_{t,i}(h)$ as defined in (2.1). The first two lines capture differences in covariances across variables and horizons (i.e. the Gaussian copula). The third line captures differences in each of the marginal densities across variables and horizons. This is identical to the sum of differences in the mean square errors across variables and horizons.[6]

The expansion illustrates that the path forecast accuracy test nests other multi-horizon/variate approaches. If there are no differences in the error covariances then the first two lines drop out and the third line drives the differences. Then, depending on how $\hat{\sigma}_N$ is treated, it is possible to recover the trace test statistic as proposed by Capistrán (2006) or an unweighted version of the average superior predictive ability (aSPA) test discussed in Quaedvlieg (2017). Although it is less obvious, (3.10) also nests differences based on Doan et al. (1984)'s log determinant metric across multiple horizons.

These links demonstrate the differences of the path forecast accuracy test relative to other tests. Equation (3.10) explicitly accounts for differences in covariances between variables and across horizons, whereas other tests do not explicitly capture these differences. This is an important advantage when large differences in the covariances / dynamics between forecasting systems can offset or exacerbate differences between individual variables at each horizon.

3.2 Nested Path System Forecasts

There is an extensive literature examining the link between tests of nested and non-nested forecasts. Clark and McCracken (2001) first demonstrated the link between the Diebold and Mariano (1995) and West (1996) framework for nested models. Hansen and Timmermann (2015) extend this to the Wald statistic when forecast models are nested and $HK = 1$.

While Giacomini and White (2006)'s approach "permits a unified treatment of nested and non-nested models" (p. 1546), their result rests on the assumption that $\sigma^2_N = var\left[\sqrt{N}\bar{LR}_{m,N,H}\right] > 0$. If this is not the case, i.e. when $\sigma^2_N = 0$, then the true forecast-error densities are identical and the test statistic converges to a non-standard distribution; see Clark and McCracken (2013a). Alternatively, Vuong (1989, Theorem

[6] Note that the above discussion implies that the sum of the last two lines of (3.10) is equal to zero.

3.3) illustrates that a nested predictive likelihood ratio test can be formulated.[7] Assuming that the errors are multivariate normally distributed and model 2 strictly nests model 1, then a nested predictive likelihood ratio test can be constructed as

$$-2LR_{m,N,H} = N\left(ln\left(\left|\widehat{\mathbf{\Phi}}^{m}_{H,N,1}\right|\right) - ln\left(\left|\widehat{\mathbf{\Phi}}^{m}_{H,N,2}\right|\right)\right) \xrightarrow{D} \chi^2_r \; as \; N \to \infty, \tag{3.11}$$

where r is the number of restrictions imposed across variables and horizons in the restricted model relative to the unrestricted model. Muirhead (2005) explores the asymptotic properties of the likelihood ratio test statistic when allowing for the broader class of elliptical distributions. Gelper and Croux (2007) propose a similar multivariate likelihood ratio test for testing 1-step-ahead forecasts using a bootstrap of the residuals to get critical values. This provides a more flexible approach without assuming a multivariate normal distribution, but requires direct knowledge of / the ability to estimate the underlying forecast model.

Engle (1984) shows that the nested (predictive) likelihood ratio test is linked to a broader class of test statistics. It is possible to generalize Stewart (1995) for nested models when $HK \geq 1$ by allowing for either a single restriction on a variable for any number of horizons or any number of restrictions on variables for a single horizon. This illustrates that there is a link between the predictive likelihood ratio and the Wald statistic for nested path system forecasts.

4 Path Forecast Accuracy Test: A special case

While the previous section provides general results, this section derives more explicit results for the forecast path test statistic under additional parametric assumptions. This provides further insight into the properties of the general test. We start by laying out the main result from which everything else follows using the following theorem:

Theorem 2 (Asymptotic distribution). Suppose that

1. the forecast errors follow a $MA(p)$ process: $\widetilde{\mathbf{u}}_t(h) = \sum_{i=0}^{p} \mathbf{\Pi}^i \mathbf{v}_{t+h-i}$ for $p \leq h-1$,
2. where $\mathbf{v}_t \overset{\text{iid}}{\sim} N_K[\mathbf{0}, \mathbf{\Omega}]$ and $\mathbf{\Omega}$ is positive definite,
3. and that $H > 0$ and $K < N$ where both H and K are fixed.[8]

Under these assumptions, the estimated log GFESM satisfies

$$\frac{ln\left|\widehat{\mathbf{\Phi}}_{H,N}\right| - ln\left|\mathbf{\Phi}_H\right| - \frac{HK(K+1)}{2N}}{H\sqrt{2K/N}} \xrightarrow{D} N(0,1) \; as \; N \to \infty. \tag{4.1}$$

This extends Cai et al. (2015)'s central limit theorem for log determinants to $MA(p)$ processes.

[7] Vuong (1989) proposes a test of σ^2_N to determine which asymptotic distribution should be used.

[8] This assumption can be relaxed by allowing K to vary with N. See Cai et al. (2015).

From assumption 1, if the forecast errors are $MA(h-1)$, we can write the path forecast errors as

$$\widetilde{\mathbf{U}}_{H,t} = \mathbf{\Psi}_{\Pi,H}\mathbf{V}_{H,t} = \begin{pmatrix} \mathbf{I}_K & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{\Pi} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{\Pi}^{H-1} & \cdots & \mathbf{\Pi} & \mathbf{I}_K \end{pmatrix} \begin{pmatrix} \mathbf{v}_{t+1} \\ \mathbf{v}_{t+2} \\ \vdots \\ \mathbf{v}_{t+H} \end{pmatrix}, \tag{4.2}$$

where it is possible to construct a lower triangular parameter matrix that drops out when taking the determinant as long as $p \leq h-1$. Then the determinant is identical to the determinant of the covariance of asymmetric Hankel matrices where K governs the degree of asymmetry. From the iid assumption

$$ln\left|\widehat{\mathbf{\Phi}}_{H,N}\right| = ln\left|\frac{1}{N}\sum_{t=1}^{N}\widetilde{\mathbf{U}}_{H,t}\widetilde{\mathbf{U}}'_{H,t}\right| \approx ln\begin{vmatrix} \widehat{\mathbf{\Omega}}_{N+1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \widehat{\mathbf{\Omega}}_{N+H} \end{vmatrix} = \sum_{h=1}^{H} ln\left|\widehat{\mathbf{\Omega}}_{N+h}\right|, \tag{4.3}$$

where we can treat each $\left|\widehat{\mathbf{\Omega}}_{N+h}\right|$ independently. The rest of the proof follows directly from Cai et al. (2015) assuming multivariate normality. This can also be interpreted as the general central limit theorem for the log determinant of the covariance of independent ($H = 1;\ K > 1$) and Hankel ($H > 1;\ K = 1$) matrices.[9]

Given Theorem 2, it is possible construct a test statistic for differences in the GFESM. This test is described in the following corollaries:

Corollary 1 (GFESM Test: zero mean). If the assumptions of Theorem 2 are satisfied for both path forecasting systems 1 and 2, a test statistic can be constructed which satisfies

$$\frac{\sqrt{N}}{2H\sqrt{K}}\left(ln\left(\left|\widehat{\mathbf{\Phi}}_{H,N,2}\right|\right) - ln\left(\left|\widehat{\mathbf{\Phi}}_{H,N,1}\right|\right)\right) \xrightarrow{D} N(0,1)\ as\ N \to \infty. \tag{4.4}$$

This result follows from the properties of a two-sample t-test under the assumption of normality. It is also closely related to Anderson (2003, Theorem 7.5.4). However, the test statistic is smaller than the one proposed by Clements and Hendry (1993) for the simulated GFESM by a factor of $1/\sqrt{H}$.

Corollary 1 coincides with (3.8) when $w_h = 1$ and $\sigma_N^2 = H^2K$. Alternatively, it also coincides with (3.9) when $w_h = 1/H^2$ and $\sigma_N^2 = K$. This illustrates that we can modify corollary 1 to allow for any re-weighting of the horizons. In that case, assuming that $\sum_{h=1}^{H} w_h > 0$, the modified corollary 1 coincides with (3.9) for any w_h where $\sigma_N^2 = \left(H\sum_{h=1}^{H} w_h\right)^2 K$. However, under Theorem 2's assumptions, corollary 1's test statistic is more efficient than the general test since σ_N^2 is not estimated.

Corollary 2 (GFESM Test: non-zero mean). If the assumptions of Theorem 2 are satisfied with the

[9] As far as we know, no other study has determined the distribution of the covariance of Hankel matrices; see Ghodsi et al. (2015).

slight change such that $\mathbf{v}_{t,i} \overset{\text{iid}}{\sim} N_K\left[\boldsymbol{\mu}_i, \boldsymbol{\Omega}_i\right]\ \forall\, i \in \{1,2\}$, a test statistic can be constructed which satisfies

$$\sqrt{N}\frac{ln\left(\left|\widehat{\boldsymbol{\Phi}}_{H,N,2}\right|\right) - ln\left(\left|\widehat{\boldsymbol{\Phi}}_{H,N,1}\right|\right)}{\sqrt{2H\left\{\sum_{i=1}^{2} Tr\left[\left(\mathbf{I}_{HK}+2\boldsymbol{\eta}_i\right)\left(\mathbf{I}_{HK}+\boldsymbol{\eta}_i\right)^{-2}\right]\right\}}} \xrightarrow{D} N(0,1)\ as\ N\to\infty. \tag{4.5}$$

This corollary follows by applying the delta-method to the results from Fujikoshi (1968, Theorem 1) and Muirhead (2005, Chapter 10) for the determinant of a non-central Wishart matrix where $\boldsymbol{\eta}_i = \mathbf{I}_H \otimes \boldsymbol{\mu}_i\boldsymbol{\mu}_i'$ is the non-centrality parameter across horizons. When $\boldsymbol{\eta} = \mathbf{0}$ then the result collapses to (4.4). In general $\boldsymbol{\eta}$ is not known, however it is possible to compute: $\widehat{\boldsymbol{\eta}} = \left(\frac{1}{N}\sum_{t=m}^{T}\widetilde{\mathbf{U}}_{t,H}\right)\left(\frac{1}{N}\sum_{t=m}^{T}\widetilde{\mathbf{U}}_{t,H}\right)'$.

The assumptions for Theorem 2 effectively imply that the conditional MSEs are broadly stable across horizons. However, these are fairly restrictive assumptions. Chong and Hendry (1986) argue that in general, conditional MSEs are "not [...] monotonic in the forecast horizon" (p. 685). This implies that there is either heteroskedasticity across horizons, large outliers, or some form of misspecification. Therefore, Corollaries 1 and 2, which average across horizons, will likely understate the uncertainty so as to over reject the null hypothesis.

There are several ways in which this special case can be extended. First, it is possible to allow for higher order dependence in the forecast errors. For example, if the forecast errors follow a $MA(p)$ process for any p, then we can approximate the log GFESM as

$$ln\left|\widehat{\boldsymbol{\Phi}}_{H,N}\right| \approx \sum_{h=2}^{H} ln\left|\widehat{\boldsymbol{\Omega}}_{N+h}\right| + ln\left|\sum_{j=0}^{\max(p-H+1,0)} \boldsymbol{\Pi}^j\widehat{\boldsymbol{\Omega}}_{N+1-j}\boldsymbol{\Pi}'^j\right| \tag{4.6}$$

If p is close to $H-1$ and $|\Pi| < 1$ then this change has little impact. Otherwise, it is necessary to account for the additional variance. Another extension is to relax the iid and normality assumptions. In order to do so, it is necessary to have an estimator of the GMFESM that is heteroskedasticity consistent (see White, 1980) and robust to outliers or fat-tails (see Rousseeuw, 1985). These extensions are left to future research.

5 Simulations

In this section we report Monte Carlo experiments to compare the properties of the path forecast accuracy tests with existing joint tests. We start by describing the path forecast-error generating process. We then describe the alternative tests and present the simulation results.

5.1 The Path Forecast-Error Generating Process

We generate the path forecast errors as follows:

$$\widetilde{\mathbf{U}}_{t,H,\{b,v,c_k,c_h\}} \equiv \boldsymbol{\theta}_b + \boldsymbol{\Psi}_H\boldsymbol{\Sigma}^{1/2}_{v,c_k,c_h}\mathbf{V}_{t,H}, \tag{5.1}$$

where $\mathbf{V}_{t,H}$ is an asymmetric Hankel matrix where each unique element is $\mathbf{v}_{t,h} \overset{\text{iid}}{\sim} N_K(\boldsymbol{\mu}, \mathbf{I}_K)$. For simplicity we set $\boldsymbol{\mu} = \mathbf{0}$. The forecast errors have a model-specific bias through $\boldsymbol{\theta}_b$, are serially correlated through $\boldsymbol{\Psi}_H$,

and have model-specific variances and correlations across horizons and variables through $\mathbf{\Sigma}_{v,c_k,c_h}$. Thus, the forecast errors follow a $MA(h-1)$ process and exhibit dependence and biases across variables and horizons. Next, we discuss the individual parameters.

First, we define the covariance structure across horizons and variables, at a single point in time. Since forecast errors generally converge to the unconditional mean when the horizon is large, then the correlations should get larger for adjacent horizons when h is large, and smaller for shorter-horizons. We use the correlation matrix C, with elements ρ_{g,h,j,k,c_k,c_h}:

$$\rho_{g,h,j,k,c_k,c_h} = \begin{cases} 1 & \text{if } g = h,\ j = k \\ \exp(-1.2 + 0.025max(g,h) - 0.125abs(h-g)) + c_h & \text{if } g \neq h,\ j = k \\ \exp(-1.8) + c_k & \text{if } g = h,\ j \neq k \\ \exp\left(-1 - \sqrt{abs((k-j)(h-g))}\right) + \frac{c_k + c_h}{2} & \text{if } g \neq h,\ j \neq k \end{cases} \tag{5.2}$$

where c_k governs the differences in how errors are correlated across variables while c_h governs differences in how errors are correlated across horizons. Higher values for either increases the overall correlation. This plays a prominent role in our analysis where under the null, c_k and c_h are set equal to zero while under the alternative they vary across models. We allow the variance to change across horizons so that $\sigma_{v,h,k} = v\left(1 + \frac{\sqrt{h-1}}{2}\right)$ where we set $v = 1$. The variance and correlation are combined so that $\Sigma_{v,c_k,c_h} = diag(\sigma_v) C_{c_k,c_h} diag(\sigma_v)$.

Next, we define the model-specific bias as

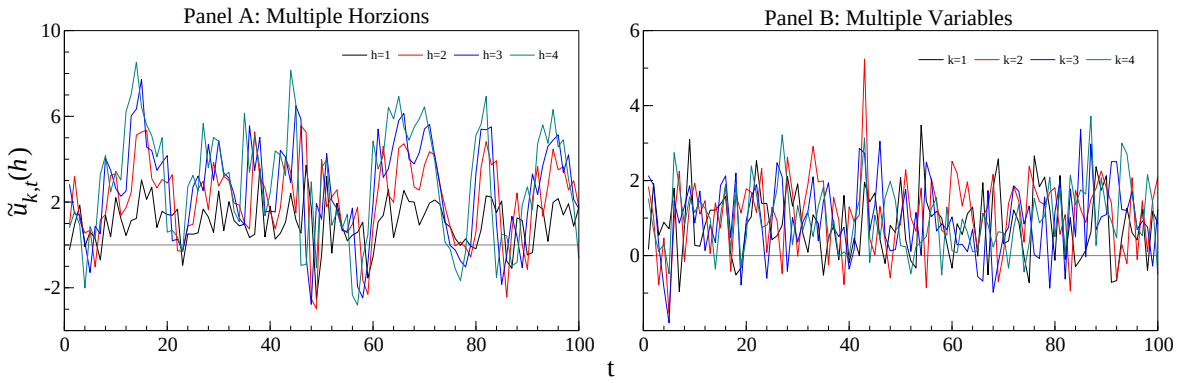
$$\boldsymbol{\theta}_{b,h,k} = b\left(1 + \sqrt{h-1}\right), \tag{5.3}$$

which implies that the bias increases with the horizon but is similar across variables. b governs the degree of bias across models where in the baseline case $b = 1$. Under this formulation, changes in $\boldsymbol{\theta}_b$ effectively shift all horizons or variables up or down based on a fixed spread, whereas changes in $\boldsymbol{\mu}$ increase the spread across horizons.

Finally, we define the serial correlation as a matrix of HK elements where

$$\mathbf{\Psi}_H = \begin{pmatrix} \mathbf{I}_K & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{\Pi} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{\Pi}^{H-1} & \cdots & \mathbf{\Pi} & \mathbf{I}_K \end{pmatrix}, \tag{5.4}$$

so that serial correlation accumulates across horizons just as a $MA(h-1)$ process. We use the following $\pi_{j,k}$



Notes: The figure plots the simulated forecast errors across multiple horizons and variables for a single simulation. The simulations are based on the parameter choices as described above where $b = 1$, $\nu = 1$, $\boldsymbol{\mu} = \mathbf{0}$ and $c_k = c_h = 0$.

Figure 5.1: Illustration of Forecast-Error Data Generating Process

elements for each $K \times K$ matrix $\boldsymbol{\Pi}$:

$$\pi_{j,k} = \begin{cases} 0.4 + min\left(\frac{k}{10}, 0.5\right) & j = k \\ 0.2 & j \neq k \end{cases}. \tag{5.5}$$

Note that $\boldsymbol{\Pi}$ only plays a role when $\boldsymbol{\theta}_b \neq \mathbf{0}$ since $|\boldsymbol{\Psi}_H| = 1$. We visualize the choice of the DGP by plotting the forecast errors across 100 observations in Figure 5.1. The pattern of the forecast errors across horizons is similar to that of the Congressional Budget Office's forecast errors for the gross federal debt. Similar results were also obtained using alternative parameterizations so that the simulated forecast errors follow patterns similar to the Federal Reserve's Greenbook forecasts errors for GDP growth and inflation.

We compare the performance of four test statistics: (1) the equally weighted univariate aSPA (i.e. the trace statistic across horizons and variables), (2) the equally weighted multivariate aSPA (i.e. the single horizon version of (3.8)), (3) the unweighted general likelihood ratio test statistic from equation (3.8), and (4) the GFESM test of Corollary 2 from (4.5). The first test represents existing approaches in the literature as seen in Capistrán (2006) and Quaedvlieg (2017) that do not explicitly target differences in covariances or dynamics. The second test is effectively the log determinant measure from Doan et al. (1984) applied to the test described in Quaedvlieg (2017) which accounts for covariances but not dynamics. Standard errors are computed using the HAC estimator in Andrews (1991) for the first three test statistics.[10] Standard errors for the fourth test statistic are computed using estimates of the bias.

Comparisons across tests allow for an assessment of the strengths and weaknesses of alternative loss functions. The univariate aSPA test does not explicitly capture differences in covariances or dynamics. The multivariate aSPA explicitly captures differences in covariances but not differences in dynamics. Finally the general likelihood approach and Corollary 2 both capture differences in covariances and dynamics. This

[10] The standard errors could also be computed using bootstrap procedures as in Hansen et al. (2011) or Quaedvlieg (2017). However, simulations indicate that these procedures only perform well when there is an abundance of forecast-error observations (i.e. > 200).

allows us to illustrate the advantages and disadvantages of alternative approaches in different scenarios.

Performance is evaluated by examining rejection frequencies under the null and various alternatives for a given critical value. Section 5.2 examines the null rejection frequency across alternative sample sizes and dimensions. Section 5.3 considers the non-null rejection frequency across various differences between the forecasts. This allows for the determination of how well different tests or loss functions detect alternative forms of misspecification. Finally, section 5.4 analyzes the null rejection frequency when the number of path forecast-error observations is small relative to the number of variables or the length of the path.

5.2 Null rejection frequency

We first consider the case where forecast errors from the two forecasting systems are essentially the same. In this case, since we are conducting our tests using a nominal size of 5%, we expect that the null rejection frequency for each test statistic should be approximately 5%. We assess how this changes with the number of forecast-error observations as well as the dimension of the system and the length of the forecast horizon. Table 5.1 presents the results.

	Univariate aSPA				**Multivariate aSPA**				**General LR**				**Corollary 2**			
N↓ ‖ H,K→	1,1	1,2	4,1	4,2	1,1	1,2	4,1	4,2	1,1	1,2	4,1	4,2	1,1	1,2	4,1	4,2
32	3.07	3.23	0.10	0.25	4.21	4.57	0.82	1.13	4.21	4.57	0.59	0.83	5.39	5.80	5.60	5.81
64	4.26	4.54	1.05	1.50	4.89	5.42	3.04	3.36	4.89	5.42	2.43	2.51	5.16	5.93	6.12	5.09
128	4.54	4.51	2.73	3.39	4.54	5.11	4.02	4.57	4.54	5.11	3.82	4.27	4.71	5.54	5.56	5.26
256	4.80	4.94	4.00	4.16	4.80	4.87	4.85	5.00	4.80	4.87	4.54	5.04	4.80	5.31	5.28	5.15
512	4.89	4.94	4.58	4.49	4.89	4.84	4.97	4.97	4.89	4.84	5.09	4.91	5.11	5.07	6.05	4.98
1000	4.72	4.84	4.68	4.81	4.72	4.99	5.29	4.99	4.72	4.99	4.96	5.11	4.79	5.34	5.77	5.34

Notes: The nominal size is 5%. Univariate aSPA is identical to the average trace statistic across all variables and horizons. Multivariate aSPA averages across horizons. General LR test refers to the general likelihood ratio test statistic. 10,000 Replications.

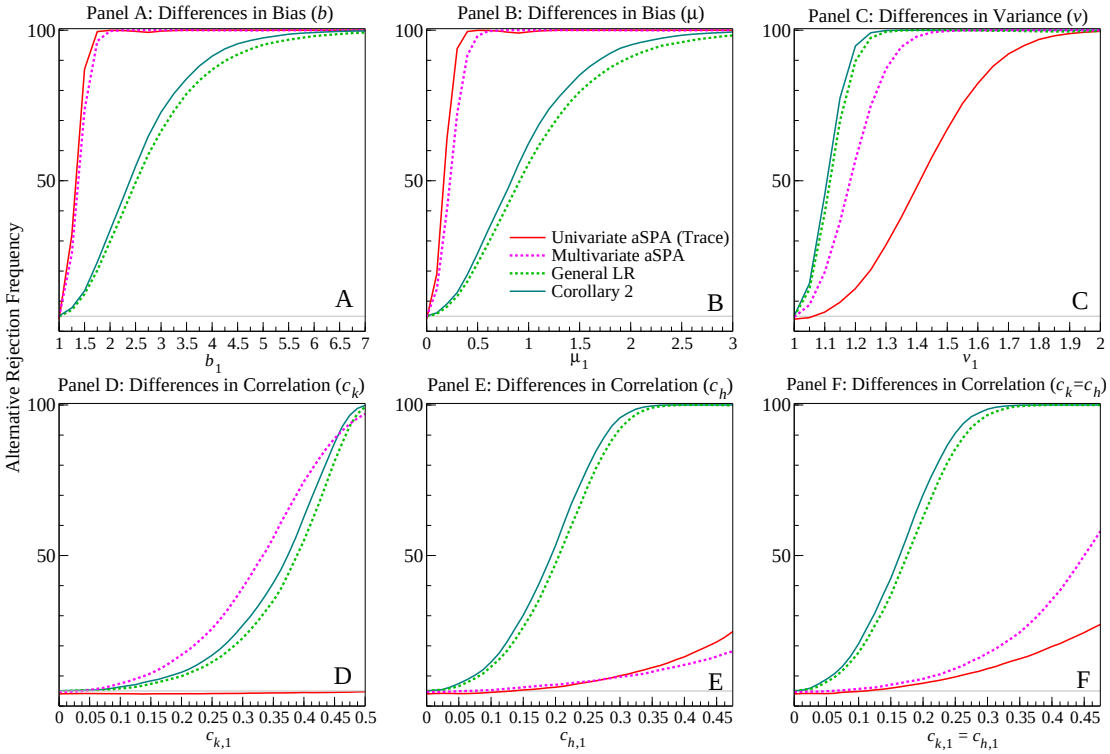
Table 5.1: Null Rejection Frequencies

Table 5.1 illustrates that while most test statistics start off undersized when $N = 32$, they all generally converge to the correct nominal size as the number of forecast-error observations increases. This dynamic is even more pronounced for larger systems and longer horizons. Note that when $H = 1$ the multivariate aSPA and general LR tests are identical, which stems from the fact that they both rely on the multivariate normal density as the loss function. If there were no correlations between variables, then the univariate aSPA would also be identical.

The test statistic from Corollary 2 also performs well. Its null rejection frequency is somewhat above 5% particularly when $H > 4$, which stems from the fact that the bias in the forecast errors does not follow the assumed structure: $\boldsymbol{\eta}_i \neq \mathbf{I}_H \otimes \boldsymbol{\mu}_i \boldsymbol{\mu}_i'$. However, despite this, the null rejection frequency remains close to the nominal size, especially when $K > 1$.

5.3 Non-null rejection frequency

Next we assess the tests ability to capture differences across forecasting systems. While only differences in biases are typically considered, we examine differences in biases, variances, and correlations across variables and horizons. This allows us to assess how well alternative tests capture various differences in the forecast errors. For simplicity, we focus on the case where $K = 2$, $H = 4$ and $N = 200$. Figure 5.2 presents the results across model differences.



Notes: The figure plots the rejection frequencies when the null is false for different statistics across different degrees of model differences for fixed $N = 200$ when $K = 2$ and $H = 4$. The nominal size is 5%. General LR refers to the general likelihood ratio test statistic in equation (3.8). 10,000 replications.

Figure 5.2: Non-null rejection frequency by type of difference

The results in Figure 5.2 illustrate an important trade-off between path forecast accuracy tests and alternative tests. As panels A and B indicate, the path forecast tests are less able to distinguish differences in bias than joint accuracy tests. This is due to the fact that increasing bias across multiple variables and horizons also increases the forecast-error covariance, which partially offsets any increase in the bias. Therefore, path forecast accuracy tests, which account for differences in covariances require larger differences in biases. This is not true for differences in the error variance in panel C when the path forecast accuracy tests have higher power.

The flip side of the bias is presented in panels D, E and F of Figure 5.2 where the path forecast tests are much better at detecting differences in error covariances across forecasts. It is particularly true for differ-

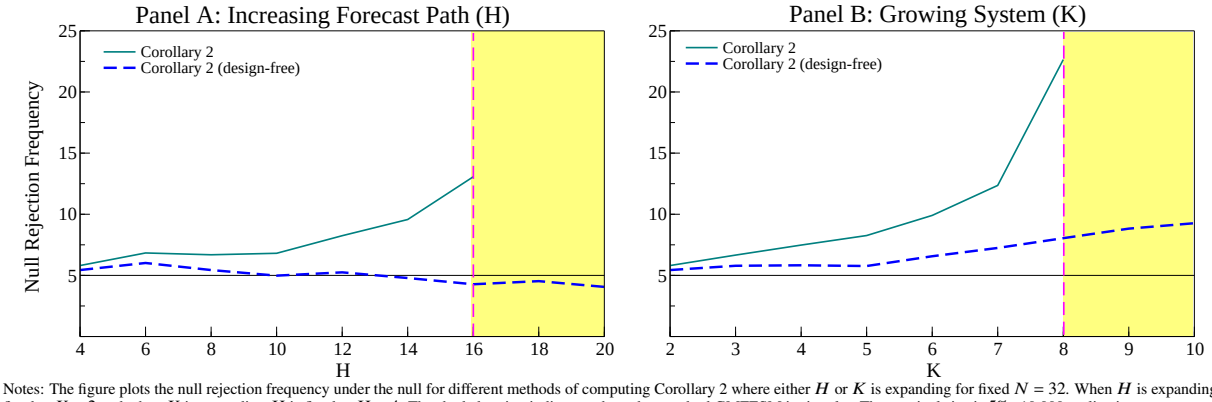
ences in dynamics / correlations (panel E) across horizons where even the multivariate aSPA test is unable to capture differences in dynamics. This illustrates the advantages of the path forecast accuracy metric and the associated tests in being able to capture differences in forecast-error dynamics. These advantages become even greater when there is a higher degree of error covariance across variables and horizons.

5.4 Extensions to higher dimensions

The test statistics can also be extended to allow for higher dimensional systems and longer forecast horizons with a finite number of forecast-error observations. Using standard methods, the sample estimates of the general matrix of the forecast-error second-moment, i.e. the GMFESM in equation (2.4), are singular when the dimension of the system times the length of the forecast path is larger than the number forecast-error of observations ($HK > N$). In practice, this is quite limiting as the reliability of the GFESM deteriorates rapidly. It also has important implications for test statistics that use the GFESM.

Hendry and Martinez (2017) describe a procedure which circumvents these limitations and produces reliable estimates of the GFESM even when $HK > N$. Their procedure extends Abadir et al. (2014)'s "design-free" approach to the GFESM by using sub-samples of the eigenvectors to improve estimates of the eigenvalues and thereby ensure that the GMFESM remains positive definite. They show that their approach does well in a variety of simulations and produces reliable estimates of the GFESM when $HK \geq N$.

Here we show how the test statistics introduced above perform when using these "design-free" estimates.[11] For simplicity we focus on the properties of the test statistic associated with Corollary 2. In theory these results can also be extended to the general test statistic. However, there are important complications when using the HAC estimator. Figure 5.3 presents the null rejection frequencies for longer forecast horizons and larger systems.



Notes: The figure plots the null rejection frequency under the null for different methods of computing Corollary 2 where either H or K is expanding for fixed $N = 32$. When H is expanding K is fixed at $K = 2$ and when K is expanding H is fixed at $H = 4$. The shaded region indicates where the standard GMFESM is singular. The nominal size is 5%. 10,000 replications.

Figure 5.3: Null rejection frequencies for relatively small samples

[11] This is closely related to the literature on testing for differences in high dimensional matrices. See Li and Chen (2012), Cai et al. (2013) and Cho and Phillips (2017).

Figure 5.3 shows that increases in the horizon or the size of the system relative to the number of forecast-error observations result in Corollary 2 being substantially oversized. The GMFESM is singular when $H > 16$ in Panel A or $K > 8$ in Panel B and so cannot be computed in these cases. However, the "design-free" methods overcome these limitations by enabling computation of the GFESM even when $HK > N$. Panel A demonstrates the benefits of these methods where even when extending the forecast paths out to twenty-steps-ahead, the modified Corollary 2 test remains close to the correct nominal value. The benefits are less pronounced in Panel B, however there is still a substantial reduction in the null rejection frequency relative to the original test statistic. This suggests that there are substantial gains to using these methods to extend the path forecast accuracy tests to higher dimensions.

6 Evaluating the FRB Greenbook path forecasts

This section compares the accuracy of the Federal Reserve Board's Greenbook path forecasts of real GDP growth, inflation and interest rates against the path forecasts from four DSGE models. We start by describing the data and the forecasts. Next we perform an initial analysis using both point and path forecast accuracy metrics. Finally, we test for differences in the path forecasts using the general path forecast accuracy test described in section 3.

6.1 Data and Forecasts

The data and forecasts are of real GNP/GDP growth, the GDP deflator and the federal funds rate spanning 1985Q4-2000Q4.[12] The DSGE model forecasts are described in detail by Wolters (2015). While Wolters (2015) utilizes every forecast release per quarter to compute and evaluate point (and density) forecast-error metrics, we focus on a single forecast release per quarter. This allows us to have a consistent set of path forecasts to evaluate each quarter. Note that earlier in the sample, several quarters do not contain forecast releases. In these cases, the forecast made in the previous quarter which corresponds to the missing quarter's forecast is used instead.[13] The 'actual' values against which we compare the forecasts are those published in the Greenbook two quarters after the quarter to which data refer. This is similar to the common practice of using the third (final) data estimates.

We compare the Greenbook (GB) forecasts against four DSGE model forecasts. The models are: Del Negro and Schorfheide (2004, "DS"), Edge et al. (2008, "EDO"), Fuhrer (1997, "FM"), and Smets and Wouters (2007, "SW"). They range in size from small (DS and FM) and medium (SW) to large (EDO) and were chosen because of their academic / policy relevance as well as their wide-spread use before the Great Recession. The forecasts were generated using the real-time dataset from Faust and Wright (2009).

[12] Special thanks to Maik Wolters for sharing his data and forecasts.

[13] For example, the previous quarter's 1-step-ahead forecast becomes the current quarter's nowcast and so on.

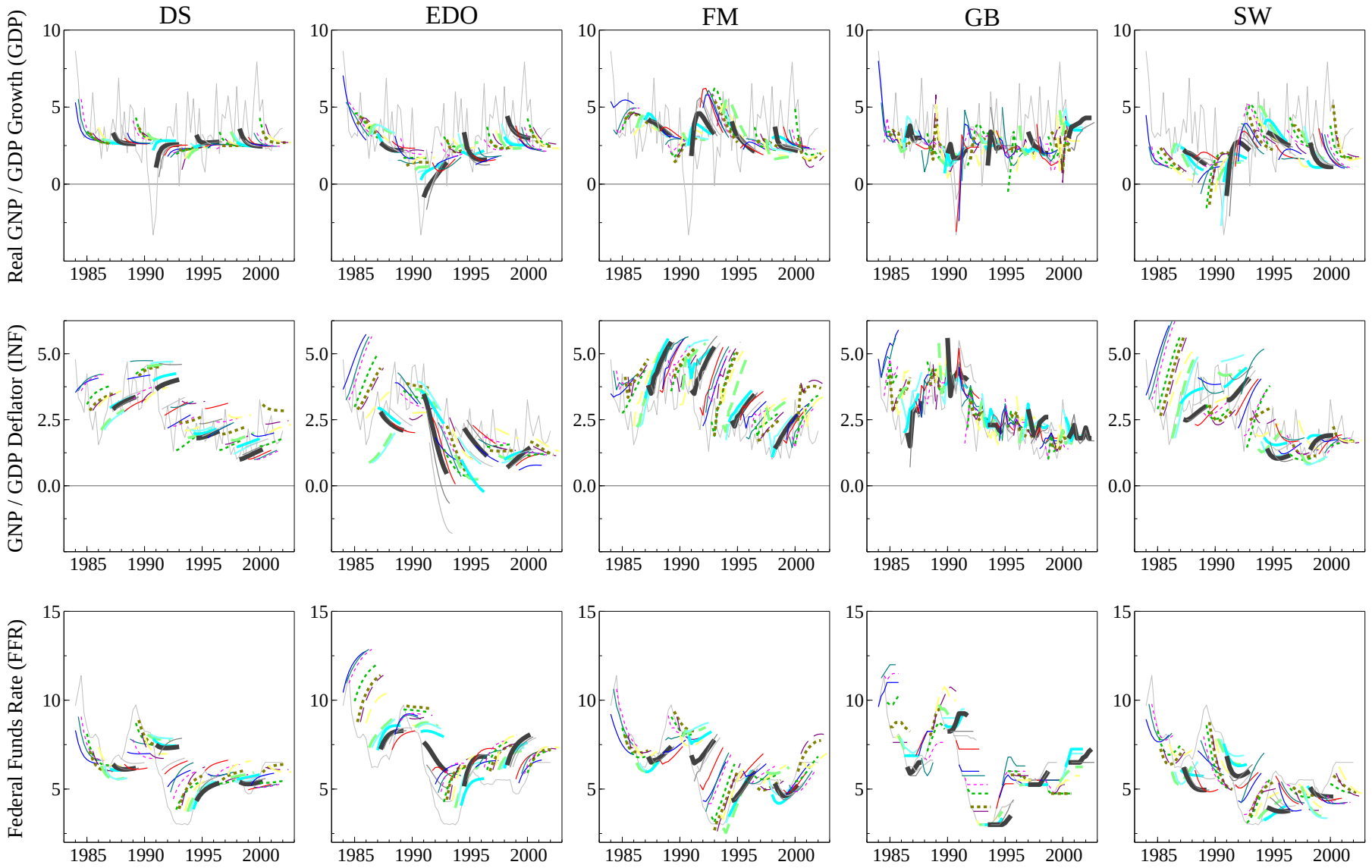


Figure 6.1: 'Path' forecasts and actuals, 1984-2002

Figure 6.1 plots the actual values and the path of the forecasts made in every quarter for each model and variable. Each row of figures represents forecasts of a different variable, while each column presents the forecasts from a separate model. This type of plot is commonly referred to as a 'hedgehog plot' since the forecasts often point away from the actual series like the quills on a hedgehog. This illustrates that the forecasts consistently fail to capture the correct dynamics of the series.

There are both similarities and differences in the path forecasts across models. Looking first at the path forecasts for real GDP, it is clear that the DSGE models generally do not capture the dynamics of the series and instead adhere closely to the long-run mean of the series. This is particularly true for the DS model but also to some extent for the other models. On the other hand, the Greenbook path forecasts do capture some of the dynamics.

This is even more pronounced for the inflation forecasts. In this case the DSGE models tend to have linear path forecast trajectories, especially when compared with the Greenbook. That being said, the properties of the path forecasts are very different across alternative models. The DS model inflation forecasts are generally fairly constant across horizons. On the other hand, the FM model inflation forecasts generally trend upwards far above the actual values. The EDO and SW inflation forecasts tend to be somewhat more reactive to the business cycle despite having periods of substantial over / under prediction.

Forecasts of the federal funds rate appear to be the most common across models. This is particularly interesting since Greenbook 'forecasts' of the federal funds rate are not forecasts in the traditional sense but rather represent the interest rate path upon which the Greenbook forecasts are conditioned. In fact, the Greenbook appears to have used a simple no change rule for at least part of the period. Despite this, the DS model forecasts tend to vary the least and adheres to the mean of the series, while the EDO and FM forecasts tend to consistently over predict the FFR. The SW model appears to do the best at trying to capture the dynamics of the data despite under predicting the FFR for significant periods.

Although the path forecasts often extend out through eight-quarters-ahead, we only evaluate them up through four-quarters-ahead due to data limitations. We also exclude the nowcasts to allow for more consistent comparisons (although this does not change the results). Thus, four forecast horizons are evaluated along with a maximum of three variables so that $HK = 12$. Given that there are approximately 70 forecast-error observations, there is no concern of relatively few forecast-error observations. In addition to these forecasts, two simple forecast combinations are constructed: AV1 is an average of the DSGE model forecasts and AV2 is an average of all the forecasts.

	GB		DS		EDO		FM		SW		AV1	
	b	**s**	**b**	**s**	**b**	**s**	**b**	**s**	**b**	**s**	**b**	**s**
GDP	0.72	1.46	0.44	2.06	0.70	1.89	-0.25	2.18	1.06	2.19	0.48	1.86
INF	-0.21	0.63	-0.02	0.79	0.40	0.76	-0.32	0.88	0.21	0.76	0.07	0.72
FFR	0.01	0.09	-0.03	0.61	-0.66	0.47	0.08	0.40	0.34	0.46	-0.07	0.41

Notes: 'b' denotes the nowcast error bias while 's' denotes the standard deviation of the nowcast errors. See text for a description of the models. GDP refers to real GNP/GDP growth. INF refers to the GNP/GDP deflator. FFR refers to the federal funds rate.

Table 6.1: Nowcast Error Properties

6.2 Assessing point and path forecast accuracy

We evaluate point forecast accuracy using the root mean squared forecast errors (RMSE) for real GDP growth, the GDP deflator and the federal funds rate. We also compute point forecast accuracy for the system of three variables using the root standardized determinant of the MSE matrix. This is similar to the root geometric mean of the sequentially conditioned mean square forecast errors. Path forecast accuracy is evaluated using the root standardized GFESM for each variable and for the system of three variables. These measures provide a comprehensive view of point and path forecast accuracy and allow us to assess different drivers of forecast accuracy.

The DSGE forecasts face an important disadvantage relative to the Greenbook forecasts. They are constructed such that the nowcasts only rely on data from the previous quarter. This does not provide an accurate depiction of how they are used in practice since nowcasts are augmented with higher frequency data. As a result, Wolters (2015) generated two different DSGE forecasts. The first (i.e. Jumping off -1) uses data from the previous quarter to generate the nowcast and forecasts. The second (i.e. Jumping off 0) takes the Greenbook nowcasts as given to generate the forecasts.

In theory, augmenting the DSGE forecasts with the Greenbook nowcasts should improve the DSGE forecasts. Recent work by Faust and Wright (2013) and Schorfheide and Song (2015) shows that improvements in the nowcasts can lead to improvements in the forecasts. However, the DSGE and Greenbook nowcasts have very different properties which may not result in forecast improvements.

Table 6.1 presents the nowcast error properties across models and variables. The Greenbook nowcast errors have a higher bias than the DSGE models, especially for GDP growth. The Greenbook only has a lower nowcast bias for the federal funds rate. While the Greenbook more than makes up for its higher bias with much lower standard deviations, this suggests that using the Greenbook nowcasts may not improve the DSGE model forecasts at longer horizons.

Table 6.2 presents the forecast accuracy results. The bold values indicate the models that have the lowest loss for a given variable / horizon / jumping off point. The results show that Greenbook point forecasts of

	Del Negro and Schorfheide (DS)										Edge et. al. (EDO)									
	Jump off -1					Jump off 0					Jump off -1					Jump off 0				
	RMSE				GFESM	RMSE				GFESM	RMSE				GFESM	RMSE				GFESM
K↓ || H→	1	2	3	4	1 - 4	1	2	3	4	1 - 4	1	2	3	4	1 - 4	1	2	3	4	1 - 4
GDP	2.13	2.20	2.21	2.21	**0.33**	2.08	2.15	2.21	2.21	**0.52**	2.08	2.24	2.30	2.27	0.65	2.10	2.13	2.27	2.29	0.85
INF	0.90	0.86	0.85	1.00	0.64	0.77	0.86	0.83	0.91	0.67	1.00	1.03	1.17	1.25	0.77	0.85	0.97	1.00	1.23	0.79
FFR	0.98	1.27	1.49	1.63	0.54	0.62	1.00	1.27	1.48	0.47	1.31	1.67	1.97	2.23	0.70	0.82	1.34	1.72	2.03	0.66
SYS	1.22	1.32	1.37	1.48	**0.35**	0.99	1.22	1.31	1.40	**0.41**	1.35	1.51	1.69	1.79	0.52	1.12	1.37	1.52	1.74	0.61
	Fuhrer and Moore (FM)										**Greenbook (GB)**									
	Jump off -1					Jump off 0					Jump off -1					Jump off 0				
	RMSE				GFESM	RMSE				GFESM	RMSE				GFESM	RMSE				GFESM
K↓ || H→	1	2	3	4	1 - 4	1	2	3	4	1 - 4	1	2	3	4	1 - 4	1	2	3	4	1 - 4
GDP	2.44	2.43	2.38	2.29	0.70	2.30	2.53	2.51	2.41	0.71	2.13	2.08	2.24	2.23	0.96	2.13	2.08	2.24	2.23	0.96
INF	1.08	1.08	1.24	1.50	0.65	0.93	1.02	1.02	1.26	0.59	**0.71**	**0.72**	**0.78**	**0.82**	**0.47**	**0.71**	**0.72**	0.78	**0.82**	**0.47**
FFR	0.66	0.97	1.23	1.44	0.56	0.43	0.71	1.00	1.25	0.51	**0.32**	**0.65**	**0.92**	**1.21**	0.51	**0.32**	**0.65**	**0.92**	**1.21**	0.51
SYS	1.17	1.32	1.48	1.64	0.50	0.95	1.20	1.33	1.52	0.52	**0.77**	**0.92**	**1.11**	**1.22**	0.57	**0.77**	**0.92**	**1.11**	**1.22**	0.57
	Smets and Wouters (SW)										**DSGE Model Average (AV1)**									
	Jump off -1					Jump off 0					Jump off -1					Jump off 0				
	RMSE				GFESM	RMSE				GFESM	RMSE				GFESM	RMSE				GFESM
K↓ || H→	1	2	3	4	1 - 4	1	2	3	4	1 - 4	1	2	3	4	1 - 4	1	2	3	4	1 - 4
GDP	2.31	2.31	2.42	2.51	0.93	2.59	2.31	2.41	2.46	1.17	**2.01**	**2.08**	**2.15**	**2.17**	0.47	**1.98**	**2.01**	**2.12**	**2.15**	0.62
INF	0.95	0.95	1.05	1.24	0.69	0.79	0.96	0.99	1.12	0.70	0.86	0.77	0.82	0.95	0.55	0.72	0.83	**0.74**	0.87	0.57
FFR	1.01	1.37	1.64	1.81	0.66	0.58	1.06	1.41	1.68	0.54	0.73	1.01	1.24	1.41	**0.47**	0.41	0.74	1.02	1.24	**0.38**
SYS	1.19	1.32	1.47	1.63	0.58	0.96	1.19	1.36	1.50	0.63	1.06	1.15	1.24	1.35	0.37	0.83	1.06	1.14	1.27	0.42

Notes: Bolded values represent the lowest value for a given metric, jumping off point, variable and horizon. The Greenbook is identical for both jumping off points. The DSGE Models use the Greenbook nowcasts when jumping off from 0. The GFESM is computed as the root of the standardized derterminant where the determinant is raised to the power of $1/2HK$*. GDP refers to real GNP/GDP growth. INF refers to the GNP/GDP deflator. FFR refers to the federal funds rate. SYS refers to the root of the standardized determinant of the three variables in the system.*

Table 6.2: Point and Path Forecast Accuracy Metrics

inflation and interest rates have lower RMSEs at all horizons and jumping off points. However, the DS model and DSGE model average forecasts of GDP growth have lower RMSEs than the Greenbook. This implies that GDP growth forecasts which capture the long-run mean of the series over this period do well. However, since the sample coincides with the Great Moderation, it is unclear whether this finding applies to more volatile periods; see Del Negro and Schorfheide (2013), Hendry and Mizon (2014) and Hendry and Muellbauer (2017).

Next we look at the system point forecasts. Although the DSGE model forecasts benefit from the Greenbook nowcasts and do well for the GDP forecasts at longer horizons, it is not enough to overcome their poor performance in the inflation and interest rate forecasts. Thus, the Greenbook system point forecasts perform best at each horizon.

Path forecast accuracy metrics convey a slightly different story. The Greenbook path forecasts of inflation do best even when excluding the nowcasts and accounting for different nowcasts. The DSGE model forecast average does slightly better for interest rates, which is not surprising since its RMSEs are not substantially different than the Greenbook's for Jumping off 0. Although the DSGE model forecast average consistently has lower RMSEs for GDP growth, the DS model has a lower GFESM. This suggests that although the model average has a lower bias than the DS model at each horizon, it comes at a cost of not capturing all of the forecast-error dynamics. It indicates that the advantages of forecast averaging that are prevalent for point forecasts may not extend completely to path forecasts.

Finally we can look at the system path forecasts. Although the Greenbook did best for the system point forecasts, this is no longer true for the system path forecasts. While the system point forecast performance was driven by the Greenbook's dominance in inflation point forecasts, the system point forecast performance is driven by the DS model's dominance in the GDP path forecasts. This suggests that the persistence in the GDP forecast errors is considerably larger than any other aspect of overall forecast accuracy across variables or horizons. In fact, it has such a large impact that only the EDO and SW forecasts do as poorly as or worse than the Greenbook according to this metric.

The path forecast accuracy metric indicates few improvements when using the Greenbook nowcasts. In fact, path forecast accuracy across all of the DSGE models generally decreases when going from jumping off point -1 to jumping off point 0. For GDP growth and inflation using the Greenbook nowcasts generally worsens the path forecast accuracy of the DSGE forecasts. The exception to this are the federal funds rate forecasts, which do see improvements in path forecast accuracy when using the Greenbook nowcasts. These results contradict the point forecast accuracy results which suggest that using the Greenbook nowcasts generally lowers RMSEs across horizons. However, it is consistent with the nowcast results in Table 6.1,

where the Greenbook nowcasts tend to have a higher bias for all variables except the federal funds rate.

6.3 Testing for differences in path forecasts

The last subsection demonstrated important differences in path forecast accuracy between the Greenbook and the DSGE models. However, it is not clear whether these differences were statistically significant. In this subsection we re-examine and test for differences in path forecast accuracy using the general test statistic from section 3. For simplicity, rather than testing each of the forecasts against one another we take the Greenbook as the baseline against which we compare all other forecasts.

Table 6.3 presents the test results. We focus specifically on the statistical significance of the test. Larger differences between the metrics do not necessarily always translate into greater statistical significance. This is because the standard errors may be more or less impacted by different correlations between the forecasts. It is more relevant to focus on the test statistic instead.

The test statistics and their associated p-values in Table 6.3 provide fairly consistent results. Overall, the Greenbook path forecasts of GDP perform significantly worse than the DSGE forecasts with the exception of the EDO and SW models (especially after controlling for starting point differences). On the other hand, the Greenbook path forecasts of inflation perform significantly better than the DSGE forecasts and the averages. The differences in path forecasts of the interest rates are generally not highly significant, especially after controlling for differences in the nowcasts.

		Jump off -1							Jump off 0						
Var.		GB	DS	EDO	FM	SW	AV1	AV2	GB	DS	EDO	FM	SW	AV1	AV2
GDP	GFESM:	0.96	**0.33**	0.65	0.70	0.93	0.47	0.53	0.96	**0.52**	0.85	0.71	1.17	0.62	0.64
	test stat:		1.75*	2.38**	2.02*	0.29	2.11*	2.18*		2.15*	1.41	2.23*	1.66*	2.40**	2.44**
	p-value:		[0.040]	[0.009]	[0.022]	[0.388]	[0.017]	[0.015]		[0.016]	[0.080]	[0.013]	[0.048]	[0.008]	[0.007]
INF	GFESM:	**0.47**	0.64	0.77	0.65	0.69	0.55	0.50	**0.47**	0.67	0.79	0.59	0.70	0.57	0.55
	test stat:		1.64	2.00*	1.60	1.80*	1.89*	1.06		1.87*	1.94*	1.87*	1.92*	1.97*	2.40**
	p-value:		[0.051]	[0.023]	[0.055]	[0.036]	[0.029]	[0.145]		[0.031]	[0.026]	[0.031]	[0.027]	[0.024]	[0.008]
FFR	GFESM:	0.51	0.54	0.70	0.56	0.66	0.47	**0.46**	0.51	0.47	0.66	0.51	0.54	**0.38**	0.39
	test stat:		0.34	1.29	0.466	1.74*	0.67	0.99		0.56	1.17	0.08	0.52	1.44	2.03*
	p-value:		[0.366]	[0.099]	[0.321]	[0.041]	[0.253]	[0.162]		[0.287]	[0.121]	[0.468]	[0.302]	[0.075]	[0.021]
SYS	GFESM:	0.57	**0.35**	0.52	0.50	0.58	0.37	0.42	0.57	**0.41**	0.61	0.52	0.63	0.42	0.45
	test stat:		1.67*	1.37	1.12	0.24	1.81*	2.27*		1.79*	1.17	1.19	1.54	1.87*	2.28*
	p-value:		[0.048]	[0.086]	[0.132]	[0.407]	[0.035]	[0.012]		[0.036]	[0.122]	[0.118]	[0.062]	[0.030]	[0.011]

Notes: In all cases, only horizons 1-4 are considered. GDP refers to real GNP/GDP growth, INF refers to GNP/GDP deflator and FFR refers to the Federal Funds Rate. The three entries within a given block of numbers are: the respective metric, the test statistic for testing the null hypothesis of equal predictive accuracy (relative to GB), and the tail probability associated with that value of the test statistic (in square brackets). The GFESM is technically computed as the root standardized GFESM where the GFESM is raised to $1/2HK$. It is akin to root of the geometric mean of the sequentially conditioned MSEs across variables and horizons. The test statistic associated with testing no differences in the GFESM is the general LR test as described above and is compared against the standard normal distribution. * Denotes rejection of the null hypothesis at the 5% critical value. ** Denotes rejection at the 1% critical value. *** Denotes rejection at the 0.1% critical value. Bold values represent the lowest value of a respective metric at a given horizon. The forecasts are defined as DS: Del Negro and Schorfheide (2004), EDO: Edge et al. (2008), FM: Fuhrer (1997), GB: Federal Reserve Board Greenbook and SW: Smets and Wouters (2007). AV1 is an average of the DSGE model forecasts and AV2 is an average of all the forecasts. The test statistic is the same as (3.8).

Table 6.3: Testing for differences in path forecasts (1985 Q4 - 2000 Q4)

Given these variable by variable results, it is not surprising that the differences in the overall system are not highly significant. However, the results do indicate that the DS model and the model averages have significantly lower system path forecast errors than the Greenbook regardless of the starting point. This is consistent with the earlier results where the overall system metric is driven by the persistence of the GDP forecast errors rather than other aspects of the system.

These results present a more nuanced and comprehensive view of differences in forecast accuracy between the Greenbook and DSGE point and path forecasts. The Greenbook inflation forecasts dominate all other forecasts considered here. This is true both across the point and path forecasts. However, the DSGE models ability to track the long-run mean does well for the GDP growth forecasts over this period. Furthermore, while the Greenbook has a solid advantage in point forecasts of interest rates (despite them not being actual forecasts), the DSGE forecasts demonstrate some value in the path forecasts even after accounting for differences in the nowcasts. Comparing across DSGE model forecasts, we find that both point and path forecast accuracy metrics (and tests) indicate that smaller DSGE models do better (particularly DS) than medium and large-scale DSGE models (i.e. SW and EDO). While this is an interesting finding, it is beyond the scope of this analysis to speculate as to why this may be the case.

The differences in the results between the point and path forecast metrics indicate that forecast-error dynamics play an important role. It is true both when assessing the relative accuracy of the forecast systems and also when assessing whether changes in the nowcasts affect the overall path forecast. These findings indicate that an important aspect of forecast accuracy has been overlooked up until now.

7 Conclusions

This paper provides a new approach to assessing forecast accuracy. We argue that the focus on point forecasts ignores forecast dynamics which are crucial to understanding the trajectory of the forecast. We propose instead that additional focus should be placed on the accuracy of the path forecast. We show that the GFESM, originally proposed by Clements and Hendry (1993), can be used as a metric of path forecast accuracy and is closely related to the joint density of the stacked forecast errors.

We then construct a general test for differences in path forecast accuracy using the link between the joint density and the loss function. We also demonstrate that when using a multivariate normal density, this test statistic nests other joint test statistics. Furthermore, we illustrate that it explicitly captures differences in error covariances and dynamics. We also examine and derive the properties of this test for an important special case.

Monte Carlo simulations are used to illustrate the trade-offs associated with our test. We find that although our path forecast accuracy test has lower power to detect differences in biases, it has much higher

power to capture differences in variances, covariances and dynamics across forecast models. This is particularly true for differences in forecast-error dynamics, which other tests are unable to capture. Thus, our test captures a new aspect of differences in forecast accuracy. We also demonstrate that the test statistic can be extended to higher dimensional settings.

Finally we compare the Federal Reserve Board's Greenbook point and path forecasts with four DSGE model forecasts. Although there are important differences in the point and path forecast results, they often complement one another and provide further support to the general findings. We find that the Greenbook dominates with its point and path inflation forecasts. However, the DSGE models ability to track the long-run mean does well for the GDP growth forecasts over the Great Moderation. Furthermore, while the Greenbook has a solid advantage in point forecast accuracy of the federal funds rate, the DSGE forecasts demonstrate additional value in the path forecasts even after accounting for differences in the nowcasts. We also find mixed effects for changes in the nowcasts on path and point forecast accuracy.

Our results indicate that forecast-error dynamics can play an important role in forecast accuracy. This is true both when assessing the relative accuracy of forecast systems and also when assessing whether improvements in the nowcast will improve the overall path forecast. These findings coincide with the results from the simulation exercise and indicate that a potentially important aspect of forecast accuracy has been overlooked up until now. This demonstrates the value of using path forecast accuracy measures and tests to assess differences in multi-horizon forecasting systems.

Bibliography

Abadir, K. M., Distaso, W., and Žikeš, F. (2014). Design-free estimation of variance matrices. *Journal of Econometrics*, 181(2):165 – 180.

Adolfson, M., Lindé, J., and Villani, M. (2007). Forecasting performance of an open economy DSGE model. *Econometric Reviews*, 26(2-4):289–328.

Amisano, G. and Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics*, 25(2):177–190.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. Wiley series in probability and statistics. Wiley, third edition.

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, pages 817–858.

Bao, Y., Lee, T.-H., and Saltoğlu, B. (2007). Comparing density forecast models. *Journal of Forecasting*, 26(3):203–225.

Berg, T. O. (2016). Multivariate forecasting with BVARs and DSGE models. *Journal of Forecasting*, 35(8):718–740.

Bjørnstad, J. F. (1990). Predictive likelihood: a review. *Statistical Science*, pages 242–254.

Cai, T. T., Liang, T., and Zhou, H. H. (2015). Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional gaussian distributions. *Journal of Multivariate Analysis*, 137:161–172.

Cai, T. T., Liu, W., and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association*, 108(501):265–277.

Capistrán, C. (2006). On comparing multi-horizon forecasts. *Economics Letters*, 93(2):176–181.

Cho, J. S. and Phillips, P. C. B. (2017). Pythagorean generalization of testing the equality of two symmetric positive definite matrices.

Chong, Y. Y. and Hendry, D. F. (1986). Econometric evaluation of linear macro-economic models. *The Review of Economic Studies*, 53(4):671–690.

Cichocki, A., Cruces, S., and Amari, S.-i. (2015). Log-determinant divergences revisited: Alpha-beta and gamma log-det divergences. *Entropy*, 17(5):2988–3034.

Clark, T. E. and McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105(1):85–110.

Clark, T. E. and McCracken, M. W. (2013a). Advances in forecast evaluation. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2B, chapter 20, pages 1107–1201. Elsevier.

Clark, T. E. and McCracken, M. W. (2013b). Evaluating the accuracy of forecasts from vector autoregressions. In Fomby, T. B., Kilian, L., and Murphy, A., editors, *VAR Models in Macroeconomics-New Developments and Applications: Essays in Honor of Christopher A. Sims*, volume 32, chapter 4, pages 117–168. Emerald Group.

Clements, M. P. and Hendry, D. F. (1993). On the limitations of comparing mean square forecast errors. *Journal of Forecasting*, 12(8):617–637.

Clements, M. P. and Hendry, D. F. (1995). Forecasting in cointegrated systems. *Journal of Applied Econometrics*, 10(2):127–146.

Clements, M. P. and Hendry, D. F. (1997). An empirical study of seasonal unit roots in forecasting. *International Journal of Forecasting*, 13(3):341 – 355.

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

Del Negro, M. and Schorfheide, F. (2004). Priors from general equilibrium models for VARs. *International Economic Review*, 45(2):643–673.

Del Negro, M. and Schorfheide, F. (2013). DSGE model-based forecasting. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2A, chapter 2, pages 57–140. Elsevier.

Del Negro, M., Schorfheide, F., Smets, F., and Wouters, R. (2007). On the fit of New Keynesian models. *Journal of Business & Economic Statistics*, 25(2):123–143.

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

Doan, T., Litterman, R., and Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric reviews*, 3(1):1–100.

Doornik, J. A. (2013). *Ox 7: An Object-Oriented Matrix Programming Language*. Timberlake Consultants Ltd.

Edge, R. M., Kiley, M. T., and Laforte, J.-P. (2008). Natural rate measures in an estimated DSGE model of the US economy. *Journal of Economic Dynamics and control*, 32(8):2512–2535.

Engle, R. F. (1984). Wald, likelihood ratio, and lagrange multiplier tests in econometrics. *Handbook of econometrics*, 2:775–826.

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

Faust, J. and Wright, J. H. (2013). Forecasting inflation. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2, chapter 1, pages 3–56. Elsevier.

Fuhrer, J. C. (1997). Inflation/output variance trade-offs and optimal monetary policy. *Journal of Money, Credit, and Banking*, pages 214–234.

Fujikoshi, Y. (1968). Asymptotic expansion of the distribution of the generalized variance in the non-central case. *Journal of Science of the Hiroshima University, Series AI (Mathematics)*, 32(2):293–299.

Gelper, S. and Croux, C. (2007). Multivariate out-of-sample tests for granger causality. *Computational statistics & data analysis*, 51(7):3319–3329.

Ghodsi, M., Alharbi, N., and Hassani, H. (2015). The empirical distribution of the singular values of a random hankel matrix. *Fluctuation and Noise Letters*, 14(03):1550027.

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Granger, C. W. (1999). Outline of forecast theory using generalized cost functions. *Spanish Economic Review*, 1(2):161–173.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Hansen, P. R. and Timmermann, A. (2015). Equivalence between out-of-sample forecast comparisons and wald statistics. *Econometrica*, 83(6):2485–2505.

Hendry, D. F. and Martinez, A. B. (2017). Evaluating multi-step system forecasts with relatively few forecast-error observations. *International Journal of Forecasting*, 33(2):359 – 372.

Hendry, D. F. and Mizon, G. E. (2014). Unpredictability in economic analysis, econometric modeling and forecasting. *Journal of Econometrics*, 182(1):186–195.

Hendry, D. F. and Muellbauer, J. (2017). The future of macroeconomics: Macro theory and models at the bank of england. *Oxford Review of Economic Policy, forthcoming*.

Hinkley, D. (1979). Predictive likelihood. *The Annals of Statistics*, pages 718–728.

Jordà, Ò., Knüppel, M., and Marcellino, M. (2013). Empirical simultaneous prediction regions for path-forecasts. *International journal of forecasting*, 29(3):456–468.

Jordà, Ò. and Marcellino, M. (2010). Path forecast evaluation. *Journal of Applied Econometrics*, 25(4):635–662.

Komunjer, I. and Owyang, M. T. (2012). Multivariate forecast evaluation and rationality testing. *Review of Economics and Statistics*, 94(4):1066–1080.

Lazarus, E., Lewis, D. J., and Stock, J. H. (2017). The size-power tradeoff in HAR inference.

Li, J. and Chen, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics*, 40(2):908–940.

Martinez, A. B. (2015). How good are us government forecasts of the federal debt? *International Journal of Forecasting*, 31(2):312–324.

Mathiasen, P. E. (1979). Prediction functions. *Scandinavian Journal of Statistics*, pages 1–21.

Mitchell, J. and Hall, S. G. (2005). Evaluating, comparing and combining density forecasts using the KLIC with an application to the bank of england and NIESR 'fan' charts of inflation. *Oxford Bulletin of Economics and Statistics*, 67(s1):995–1033.

Montiel Olea, J. L. and Plagborg-Møller, M. (2017). Simultaneous confidence bands: Theoretical comparisons and suggestions for practice.

Muirhead, R. J. (2005). *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons.

Quaedvlieg, R. (2017). Multi-horizon forecast comparison. Available at SSRN: https://ssrn.com/abstract=2979352.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8:283–297.

Schorfheide, F. and Song, D. (2015). Real-time forecasting with a mixed-frequency VAR. *Journal of Business & Economic Statistics*, 33(3):366–380.

Smets, F. and Wouters, R. (2007). Shocks and frictions in US business cycles: A Bayesian DSGE approach. *The American Economic Review*, 97(3):586–606.

Stewart, K. G. (1995). The functional equivalence of the W, LR, and LM statistics. *Economics Letters*, 49(2):109–112.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, pages 307–333.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, pages 1067–1084.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, pages 817–838.

Wolf, M. and Wunderli, D. (2015). Bootstrap joint prediction regions. *Journal of Time Series Analysis*, 36(3):352–376.

Wolters, M. H. (2015). Evaluating point and density forecasts of DSGE models. *Journal of Applied Econometrics*, 30(1):74–96.