Testing for Differences in Path Forecast Accuracy

Andrew B. Martinez*

November 23, 2019

Abstract

Path forecasts help guide many important policy decisions. However, analyses of forecast accuracy often focus on one horizon at a time and ignore the dynamics along the path. I propose two tests of path forecast accuracy which specify the joint density as a general loss function to capture differences in forecast dynamics. Simulations highlight the benefits of not relying on heteroskedasticity and autocorrelation consistent estimators at long horizons as well as the trade-offs associated with using joint tests. I apply the path tests to evaluate macroeconomic path forecasts, and the accuracy of Hurricane Irma's predicted path in 2017. The results illustrate that differences in forecast dynamics play an important role in path forecast accuracy.

Keywords: joint density, log determinant, loss function, mean square error

JEL classifications: C12, C22, C52, C53

*Office of Macroeconomic Analysis, U.S. Department of the Treasury; Andrew.Martinez@treasury.gov. The views in this paper are mine and do not represent those of the Treasury Department or the U.S. Government. This research was conducted in part while I was a PhD Student in the Department of Economics at the University of Oxford and as a dissertation intern at the Federal Reserve Bank of Cleveland. An earlier version was published as a Federal Reserve Bank of Cleveland Working Paper No. 17-17 (`https://doi.org/10.26509/frbc-wp-201717`). This research was also supported in part by a grant from the Robertson Foundation (grant 9907422). I am grateful for comments and suggestions from Vanessa Berenguer-Rico, Jennifer L. Castle, Todd E. Clark, Michael P. Clements, David F. Hendry, Xiyu Jiao, Kevin Sheppard and participants at multiple Federal Reserve Bank of Cleveland brown bag seminars. Thanks to Maik Wolters for sharing his forecasts. All numerical results were obtained using OxMetrics 7.2 (OSX/U); see Doornik (2013). The manuscript was prepared using LyX 2.2.3. All errors are my own.

"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 help guide many important policy decisions. For example, central banks care about the expected trajectory of the economy due to lags in policy implementation and the monetary transmission mechanism. The Federal Reserve incorporates future economic paths through optimal control monetary policy rules (see Yellen, 2012), while several central banks, including Sveriges Riksbank and Norges Bank, publish expected future paths of policy rates.[1] Fiscal policies are judged in part by their expected impact on the trajectories of the deficit and the debt (see Martinez, 2015), whereas emergency management agencies use a hurricane's projected path to decide when and where to order evacuations. Thus, it is crucial that forecast paths are as accurate as possible.

While forecast accuracy is typically assessed at each forecast horizon individually, this is not sufficient to ensure that the entire path is accurate. A path forecast is a joint statement about multiple forecast horizons. It includes forecasts for each horizon as well as the dynamics across horizons. While forecast dynamics are often ignored, Faust and Wright (2013) and Schorfheide and Song (2015) find that nowcast accuracy matters for longer-horizon forecast accuracy. This indicates that forecast dynamics can propagate forecast inaccuracy across horizons. Therefore, assessments of path forecast accuracy need to account for these interactions and for the dependencies across horizons.

Joint uncertainty bands can capture the dynamic dependencies along the forecast path. Kolsrud (2007) was the first to propose methods for computing simultaneous prediction bands. Since then, a substantial literature has developed; see Jordà and Marcellino (2010), Staszewska-Bystrova (2011), Jordà et al. (2013), Kolsrud (2015), Wolf and Wunderli (2015), and Knüppel (2018) among others. These methods incorporate the dynamics across forecast horizons into the uncertainty around the forecast path.

[1]For example, see the Riksbank and Norges Bank Monetary Policy Reports in September 2019.

Despite recent advances in measuring path uncertainty, joint tests of forecast accuracy do not fully capture dependencies across horizons. The multi-horizon joint forecast accuracy tests by Capistrán (2006) and Quaedvlieg (2019) are simple or weighted averages of single horizon loss differentials, which ignore the dynamics between forecast errors across horizons.

I fill this gap in the literature by proposing a joint test of equal path forecast accuracy. I start by illustrating the link between path forecast-error loss functions and joint forecast-error densities. Building on this insight, I extend the framework of Giacomini and White (2006) to formulate a joint test of equal path forecast accuracy. This test is linked to existing joint multi-horizon tests while also capturing forecast dynamics across horizons. It is also related to Hungnes (2018)'s multi-horizon forecast encompassing test in that better path forecast accuracy is necessary but not sufficient for one path to forecast-encompass another path; see Ericsson (1992). I also propose a modified version of the joint path test to circumvent the heteroskedasticity and autocorrelation consistent (HAC) estimator of the variance, which has poor small sample performance when there is large persistence; see Müller (2014).

Monte Carlo simulations illustrate the trade-offs associated with these tests. Path forecast accuracy tests are less able to detect differences in forecast biases which tend to be highly correlated across horizons and so are offset by higher forecast-error covariances. However, joint path tests are better able to capture differences in variances, covariances, and dynamics across forecast models. This is especially true for differences in forecast dynamics across horizons, which other tests do not capture.

I apply the path tests in two applications. First, I test for differences in the accuracy of the Federal Reserve Board's Greenbook path forecasts of real GDP growth, inflation and interest rates and real-time forecasts from four Dynamic Stochastic General Equilibrium (DSGE) models, the Survey of Professional Forecasters (SPF), and a Vector Equilibrium Correction Model (VEqCM). I decompose the differences to show that the joint forecast-error dynamics across horizons drive the overall differences by exacerbating them for some models and ameliorating them for others. In the second application, I test for differences in the accuracy of path forecasts for Hurricane Irma in 2017. The tests show that while the official forecast dominates, other methods can match or exceed its accuracy depending on how the forecast path is weighted.

The rest of the paper is structured as follows. The next section introduces measures of path forecast accuracy. Section 3 formulates the joint path forecast accuracy test and compares it against other approaches. It also develops a special case of the general path test. Section 4 evaluates of the performance of the test statistics against alternative loss functions in various simulations. Section 5 applies the tests to evaluate the performance of macroeconomic path forecasts. Section 6 uses the tests to evaluate alternative path forecasts of Hurricane Irma in 2017. Section 7 concludes. Proofs of the main results follow in a Mathematical Appendix.

2 Measuring Path Forecast Accuracy

Let $\mathbf{y}_t$ be a K-dimensional random vector. Denote point forecasts for horizon h of this vector as $\widehat{\mathbf{y}}_t^m(h) = \widehat{\mathbb{E}}_t(\mathbf{y}_{t+h} \mid \mathbf{y}_t, \mathbf{y}_{t-1}, \ldots)$ where $\widehat{\mathbb{E}}_t(\cdot|\cdot)$ is the empirical conditional expectations operator, 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 $K \times 1$ vector of point forecast errors is

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

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

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

where $N = T-m+1$ is the sample of forecast-error observations. Summarizing the information within the MSE matrix using the trace of $\widehat{\mathbf{\Sigma}}_{N,h}^m$ ignores any error covariances across variables. It is possible to capture the dependencies between variables using the (log) determinant: $\left|\widehat{\mathbf{\Sigma}}_{h,N}^m\right|$. The log determinant was popularized by Doan et al. (1984) and is commonly used in evaluations of vector autoregressive (VAR) and DSGE model forecasts; see Adolfson et al. (2007), Del Negro et al. (2007), Schorfheide and Song (2015), and Berg (2016). Alternatively, see Komunjer and Owyang (2012) for a class of asymmetric multivariate loss functions.

If $\widehat{\mathbf{Y}}_t^m(H)$ and $\mathbf{Y}_{t,H}$ are the 1 to H-step predicted and observed paths, the path errors are

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

I use the general MSE loss function proposed by Clements and Hendry (1993) to evaluate the path forecast errors. Their general forecast-error second-moment matrix extends the MSE matrix in (2.2) to multiple forecast horizons

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

where each K-dimensional block along the main diagonal of (2.4) represents $\widehat{\mathbf{\Sigma}}^{m}_{N,h}$ for $h = 1, ..., H$, the off diagonals are co-movements between horizons and variables, and the determinant $\left|\widehat{\mathbf{\Phi}}^{m}_{N,H}\right|$, i.e. the GFESM, summarizes this information across variables and horizons. Clements and Hendry (1995, 1997, 1998) show that the GFESM captures changes in forecast dynamics, whereas trace MSE metrics do not. Christoffersen and Diebold (1998) propose an alternative metric that captures cointegrating relationships and forecast dynamics.

2.1 The Joint Density as a General Loss Function

Predictive distributions can be thought of as general forecast error loss functions. However, the standard marginal predictive distribution is unable to handle loss functions across multiple horizons (or transformations thereof). Granger (1999, p. 171) argues that it is therefore necessary to consider a joint predictive distribution. Choosing an elliptically contoured joint forecast-error distribution, as in Jordà et al. (2013), generates a joint forecast-error density which is directly related to the GFESM:

$$f_{t,U_H}\left(\widetilde{\mathbf{U}}^{m}_{t,H}\right) = C\left|\widehat{\mathbf{\Phi}}^{m}_{N,H}\right|^{-1/2} exp\left\{-g_t\left(\widetilde{\mathbf{U}}^{m\prime}_{t,H}\left\{\widehat{\mathbf{\Phi}}^{m}_{N,H}\right\}^{-1}\widetilde{\mathbf{U}}^{m}_{t,H}\right)\right\}, \tag{2.5}$$

where C is a constant, $g_t(\cdot)$ is a measurable density function, and $\widehat{\mathbf{\Phi}}^{m}_{N,H}$ is positive definite.

Elliptical densities encompass a broad class of relevant distributions including the multivariate normal and t distributions. They also impose a symmetric loss function but allow for a general correlation structure across variables and horizons while accommodating fat-tails. From (2.5) it is also possible to construct a predictive likelihood following Clements and Hendry (1998). Thus, assumptions made about the joint forecast-error density can also be interpreted as choices about the joint loss function and vice versa. This allows for tests of equal path forecast accuracy in terms of either general loss or joint density functions. The next section formulates a general test of equal path forecast accuracy using the joint forecast-error density as a general loss function.

3 A Test of Equal Path Forecast Accuracy

In this section I show how to construct a path forecast accuracy test using the joint forecast-error density as a general loss function. I start by describing the notation and the environment. I then describe the test and the necessary assumptions. Finally, I interpret it against other joint multi-horizon tests using a specific loss function.

Consider a stochastic process $\mathbf{Z} \equiv \{\mathbf{z}_t : \Omega \to \mathbb{R}^s,\ S \in \mathbb{N}.\ t = 1, 2, ...\}$ defined on a complete probability space (Ω, F, P). Partition the observed vector as $\mathbf{z}_t \equiv \left(\mathbf{y}_t, \mathbf{x}_t'\right)'$, where $\mathbf{y}_t : \Omega \to \mathbb{R}^K$ is the vector of variables of interest and $\mathbf{x}_t : \Omega \to \mathbb{R}^{S-K}$ is a vector of predictors. Define the information set at time t as $\mathbf{F}_t = \sigma\left(\mathbf{z}_t, \mathbf{z}_{t-1}, \ldots, \mathbf{z}_{t-m+1}; \mathbf{\Pi}_{m,t}\right)$ where $\mathbf{\Pi}$ is the matrix of population parameters for the $\mathbf{z}_t$ process and suppose that methods 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 $\mathbf{F}_t$. Denote the path forecasts for two sets of methods $j \in \{1, 2\}$ by $\widehat{\mathbf{Y}}^m_{t,j}(H) \equiv l_j\left(\mathbf{z}_t, \mathbf{z}_{t-1}, \ldots, \mathbf{z}_{t-m+1}; \hat{\mathbf{\Pi}}_{j,m,t}\right)$ where $l_j(\cdot)$ represents a measurable function that is allowed to vary across variables and horizons. Subscripts indicate that the time t forecasts are measurable functions of the m most recent observations where m is finite.[2]

The $S(t - m + 1) \times HK$ matrix $\hat{\mathbf{\Pi}}_{j,m,t}$ collects the parameter estimates for the set of methods in j. The only requirement for how the forecasts are produced is that they are measurable functions estimated over the same finite estimation window. This allows them to be generated using a range of methods which may differ for each variable and / or horizon as long as they are estimated over the same window and only rely on information in $\mathbf{F}_t$.

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 $\widetilde{\mathbf{U}}^m_{m,H,j} = \mathbf{Y}_{m,H} - \widehat{\mathbf{Y}}^m_{m,j}(H)$. The estimation window is rolled forward one observation and the second set of forecasts are obtained using $2, \ldots, m+1$ where the path forecast errors are generated as $\widetilde{\mathbf{U}}^m_{m+1,H,j} = \mathbf{Y}_{m+1,H} - \widehat{\mathbf{Y}}^m_{m+1,j}(H)$. This procedure is iterated, so that the last set of forecasts are obtained using $T-m, ..., T$ and the path forecast errors are generated as $\widetilde{\mathbf{U}}^m_{T,H,j} = \mathbf{Y}_{T,H} - \widehat{\mathbf{Y}}^m_{T,j}(H)$. This yields a sequence of $N = T - m + 1$ path forecast-error observations.

[2]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.

This closely follows the setup for Giacomini and White (2006)'s unconditional predictive ability test with two important differences.[3] First, I focus on the vector of path forecasts across multiple horizons and variables. Second, I use the log joint forecast-error density as a general loss function, which is similar to the log score; see Gneiting and Raftery (2007). This aligns my approach with the forecast density evaluation literature; see Berkowitz (2001), Mitchell and Hall (2005), Bao et al. (2007) and Amisano and Giacomini (2007). A key difference is that the joint density is chosen by the forecast evaluator. This choice is based on an acceptable loss profile and not necessarily related to the underlying data or forecast generating process. Once the forecast errors are filtered through the joint loss, it is possible to use the methods proposed by Giacomini and Rossi (2010) and Hansen et al. (2011).

For two alternative sets of path forecasting methods let

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

I allow for different weights at each horizon using the fact that the joint density is the product of the conditional and marginal densities. The unique temporal ordering means that the log joint density for each set of methods j can be decomposed as

$$ln\left\{f_{t,U_{H,j}}\left(\widetilde{\mathbf{U}}^m_{t,H,j}\right)\right\} = \sum_{h=1}^{H} ln\left\{f_{t,u_{h,j}|U_{h-1,j}}\left(\tilde{\mathbf{u}}^m_{t,j}(h) \mid \tilde{\mathbf{u}}^m_{t,j}(0),\ldots,\tilde{\mathbf{u}}^m_{t,j}(h-1)\right)\right\}, \tag{3.2}$$

where $\tilde{\mathbf{u}}^m_{t,j}(0) = \mathbf{0}$ and the forecast error density at each horizon is conditional on all previous horizons. Allowing for fixed weights at each horizon, (3.1) becomes

$$WLR_{m,t,H,f} = \sum_{h=1}^{H} w_h \nabla ln\left\{f_{t,u_h|U_{h-1}}\left(\tilde{\mathbf{u}}^m_t(h) \mid \tilde{\mathbf{u}}^m_t(0),\ldots,\tilde{\mathbf{u}}^m_t(h-1)\right)\right\}, \tag{3.3}$$

where w_h is the weight assigned to each horizon. For example, this allows longer horizons to receive less weight than shorter horizons while also accounting for the dependence between them. Re-weighting the conditional and marginal error densities differs from Amisano and Giacomini (2007) who re-weight the density forecast to focus on different aspects of the distribution. Since the optimal choice of weights is generally unknown, the sensitivity to this choice could be explored further using the procedures in Barendse and Patton (2019).

[3]The unconditional predictive ability test is similar to Diebold and Mariano (1995) and West (1996). However, as discussed in Clark and McCracken (2013), the former focuses on finite sample results while the latter focus on the population, which has implications for the null hypothesis.

A test for equal performance of the weighted path forecasts can be formulated as

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

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

where $\overline{WLR}_{m,N,H,f} = \frac{1}{N}\sum_{t=m}^{T} WLR_{m,t,H,f}$ which implies that the alternative hypothesis does not require stationarity. A test of the null is based on the following statistic

$$\tau_{m,N,H.f} = \frac{\sqrt{N} \times \overline{WLR}_{m,N,H,f}}{\hat{\sigma}_{N,f}}, \tag{3.6}$$

where $\hat{\sigma}^2_{N,f}$ is a heteroskedasticity and autocorrelation consistent (HAC) estimator of the asymptotic variance $\sigma^2_N = \mathbb{V}\left[\sqrt{N} \times \overline{WLR}_{m,N,H,f}\right]$; see Andrews (1991), Müller (2014) and Lazarus et al. (2018) among others.

A level a test rejects the null hypothesis of equal performance of the path forecasts of systems 1 and 2 whenever $abs\left(\tau_{m,N,H,f}\right) > z_{a/2}$, where $z_{a/2}$ is the $(1 - a/2)$ quantile of a standard normal distribution. The following theorem provides asymptotic justification:

Theorem 1. For a finite estimation window $m < \infty$, and $0 \leq w_h < \infty$, suppose

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. $\mathbb{E}\left[ln\left\{f_{t,U_{H,j}}\left(\widetilde{\mathbf{U}}^m_{t,H,j}\right)\right\}\right]^{2r} < \infty$ for all t and $j \in \{1, 2\}$;

3. $\sigma^2_N > 0$ for all N sufficiently large.

Then (a) under H_0 in (3.4), $\tau_{m,N,H,f} \xrightarrow{D} N(0,1)$ as $N \to \infty$ and (b) under H_A in (3.5), for some constant $c \in \mathbb{R}$, $P\left[abs\left(\tau_{m,N,H,f}\right) > c\right] \to 1$ as $N \to \infty$.

Assumption 2 requires the existence of at least four moments of the log forecast-error densities, and must be verified on a case-by-case basis depending on the forecast errors and loss function. For example, the normal density requires the existence of at least eight moments. This effectively requires that the forecast errors do not behave too erratically and limits the flexibility of Assumption 1, which allows the data to be non-stationary and characterized by heterogeneity, dependence, and structural breaks.

Assumption 3 allows for the test to consider both nested and non-nested models in sample, but implies that they are not nested in the population; see Giacomini and White (2006, p. 1546). If this assumption does not hold, then the true forecast-error densities are identical and the test statistic will converge to a non-standard distribution; see Clark and McCracken

(2013). Hansen and Timmermann (2015) show that for $H = K = 1$, this distribution is related to the distribution of the difference between Wald test statistics from the forecast sample (N) and the estimation sample (m). Thus, it is also related to the difference between likelihood ratio test statistics from the respective samples, see Engle (1984), and so it may be possible to test this assumption following Vuong (1989).

3.1 Comparing Other Approaches

Any number of densities can be used as a loss function. Choosing $f_t(\cdot)$ from the class of elliptical densities as in (2.5) gives a symmetric loss function that is linked to the standard MSE loss and to the GFESM. In particular, for the multivariate normal density, $f_t(\cdot) = \varphi_t(\cdot)$, with fixed weights so that $w_h = w \; \forall \; h \in H$, (3.6) is a test of differences in the log GFESM

$$\tau_{m,N,H,\varphi} = -\frac{\sqrt{N}}{2\hat{\sigma}_{N,\varphi}} w \nabla ln\left(\left|\widehat{\mathbf{\Phi}}^{m}_{N,H}\right|\right). \tag{3.7}$$

Allowing for different weights across horizons gives

$$\tau_{m,N,H,\varphi} = -\frac{\sqrt{N}}{2\hat{\sigma}^{G}_{N,\varphi}} \sum_{h=1}^{H} w_h \nabla ln\left(\left|\widehat{\mathbf{\Sigma}}^{m}_{N,h|(0,\ldots,h-1)}\right|\right), \tag{3.8}$$

where $ln\left|\widehat{\mathbf{\Sigma}}^{m}_{N,h|(0,\ldots,h-1)}\right|$ is the log determinant of the conditional MSE matrix at each horizon, see (2.2), which is orthogonalized by all prior horizons.[4]

Expanding (3.8) to facilitate comparisons with other multi-horizon approaches gives

$$\begin{aligned}\tau_{m,N,H.\varphi} = &-\frac{\sqrt{N}}{2\hat{\sigma}_{N,\varphi}}\left[\sum_{h=1}^{H} w_h \nabla\left\{ln\left(\left|\widehat{\mathbf{\Sigma}}^{m}_{N,h}\right|\right) - ln\left(\frac{\left|\widehat{\mathbf{\Sigma}}^{m}_{N,h}\right|}{\left|\widehat{\mathbf{\Sigma}}^{m}_{N,h|(0,\ldots,h-1)}\right|}\right)\right\}\right. \\ &+ \frac{1}{N}\sum_{t=m}^{T} \nabla\left\{\widetilde{\mathbf{U}}^{m\prime}_{t,H} W^{\frac{1}{2}\prime}_{H,K}\left(\left\{\widehat{\mathbf{\Phi}}^{m}_{N,H}\right\}^{-1} - \mathbf{I}_{HK}\right) W^{\frac{1}{2}}_{H,K} \widetilde{\mathbf{U}}^{m}_{t,H} + C\left(W_{H,K}\right)\right\} \\ &\left.+ \sum_{h=1}^{H} w_h \sum_{k=1}^{K} \nabla\left\{\frac{1}{N}\sum_{t=m}^{T} \tilde{u}^{m}_{t}(h,k)^2\right\}\right],\end{aligned} \tag{3.9}$$

where $\tilde{u}^{m}_{t}(h,k)$ is the kth element of $\tilde{\mathbf{u}}^{m}_{t}(h)$ as defined in (2.1), $W_{H,K} = diag(\{w_1,\ldots,w_H\}) \otimes \mathbf{I}_K$, and $C\left(W_{H,K}\right)$ captures interactions between the weights and covariance terms. $C\left(W_{H,K}\right) = 0$ when $w_h = w \; \forall h \in H$ or when $\widehat{\mathbf{\Phi}}^{m}_{N,H} = diag\left(\widehat{\mathbf{\Phi}}^{m}_{N,H}\right)$. The first two lines of (3.9) capture the weighted differences in the forecast-error covariances and dynamics (i.e. weighted differences

[4]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. Also note that $\widehat{\mathbf{\Sigma}}_{1|0} = \widehat{\mathbf{\Sigma}}_1$.

in the Gaussian copulas). The third line captures the weighted differences in the marginal densities across variables and horizons; i.e. the weighted MSEs.

This illustrates how the joint path test relates to existing tests. The unweighted trace test statistic proposed by Capistrán (2006) or the weighted average superior predictive ability (aSPA) test discussed in Quaedvlieg (2019) is captured in the last line of (3.9). The log determinant metric, popularized by Doan et al. (1984), is captured in the first term of the first line of (3.9). The joint test incorporates this information while also including additional information which captures differences in the dependence and dynamics of the forecasts. This is important, especially when differences in the forecast-error covariances can offset or exacerbate the differences between variables and across horizons.

3.2 A Special Case: The Normal Distribution

Although the general path test has many advantages, it is limited by its reliance on the HAC estimator of the asymptotic variance. While HAC estimators allow for considerable flexibility, they exhibit poor performance in small samples when there is high persistence; see Müller (2014). Quaedvlieg (2019) proposes a bootstrap estimator which allows for high persistence but also requires large samples. I show that under specific distributional assumptions about the forecast errors, it is possible to circumvent the use of HAC or bootstrap estimators. This provides further insight into the properties of path tests.

Until now I have treated the joint density as a general loss function, which does not necessarily impose a strict distributional assumption on the forecast errors. However, going one step further and imposing that the forecast errors follow a multivariate normal distribution, then the unweighted version of (3.6) is

$$\tau_{m,N,H,\varphi} = \frac{\sqrt{N} \times \overline{LR}_{m,N,H,\varphi}}{\tilde{\sigma}_N}, \tag{3.10}$$

where $\tilde{\sigma}_N^2 = 4H \times tr\left[\left(\left\{\mathbf{I}_{HK} - \frac{1}{2}\left(\frac{(H-1)^2+\mathbb{1}_{H>1}}{H^2}\right)\hat{\mathbf{\Theta}}_{N,H}\hat{\mathbf{\Theta}}_{N,H}\right\}(1-\hat{\gamma}_N^2) + 2\hat{\mathbf{\Theta}}_{N,H}(1-\hat{\gamma}_N)\right)\left(\mathbf{I}_{HK} + \hat{\mathbf{\Theta}}_{N,H}\right)^{-2}\right]$ is a consistent estimator of the asymptotic variance, $\hat{\mathbf{\Theta}}_{N,H}$ is a consistent estimator of the squared stacked non-centrality, and $\hat{\gamma}_N < 1$ is a consistent estimator of the correlation across forecasting methods.[5] Asymptotic justification for (3.10) comes from the following theorem:

[5]For simplicity this formulation assumes that the estimators of the non-centrality are identical across forecasting methods. This is relaxed in the simulations and the application.

Theorem 2. For fixed K and H, large m, and for $j \in \{1, 2\}$ suppose

1. $\widetilde{\mathbf{U}}^{m}_{t,H,j} \sim N_{HK}\left[\mathbf{\Xi_{H,j}}, \mathbf{\Omega}_{H,j}\right]$ with $\mathbf{\Omega}_{H,j}$ positive definite;
2. $\widetilde{\mathbf{u}}^{m}_{t,j}(h) = \sum_{i=0}^{s} \mathbf{\Pi}_{j,h,i}\mathbf{v}_{j,t+h-i} + o_p(1)$ for $s \leq h-1$.

Then (a) under H_0 in (3.4), $\tau_{m,N,H,\varphi} \xrightarrow{D} N(0,1)$ as $N \to \infty$ and (b) under $H_A: \mathbb{E}\left[LR_{m,t,H,\varphi}\right] \neq 0 \quad t = 1, 2, ...$, for some constant $c \in \mathbb{R}$, $P\left[abs\left(\tau_{m,N,H,\varphi}\right) > c\right] \to 1$ as $N \to \infty$.

Assumption 1 requires that the forecast errors are jointly gaussian. This is a strong assumption given that forecast errors often suffer from large outliers and heavy tails. However, it is possible to reformulate potential heteroskedasticity, outliers, and fat tails as abrupt changes in the time-varying bias; see Hendry and Martinez (2017). Therefore, the validity of this assumption depends on how the non-centrality is treated.

Assumption 2 requires that the stacked forecast errors each follow a $MA(h-1)$ process. Diebold and Mariano (1995) find that $MA(h-1)$ dependence works well in practice and it has since become a standard assumption; see Giacomini and White (2006, footnote 5) and Clark and McCracken (2013). It is possible to allow for higher order dependence but requires that the initial forecast horizon is treated separately from the rest of the path.[6]

Assumption 2 also requires that any estimation error collapses. This aligns more closely with Diebold and Mariano (1995) rather than Giacomini and White (2006) by focusing on the population results. As a result, nested models are dealt with differently in that the underlying shock processes are assumed to be correlated, which is why it is necessary to estimate the correlation across forecasting methods.

While the asymptotic variance is largely driven by the non-centrality, if the forecast errors are unbiased and uncorrelated across models, then the asymptotic variance of (3.10) simplifies to $\tilde{\sigma} = 2H\sqrt{K}$. This relates to Anderson (2003, Theorem 7.5.4) and Clements and Hendry (1993, equation 38) and effectively shows that in the unbiased case, the test statistic under the null is an average of the differences in the conditional mean square forecast errors across horizons. Note also that $\tilde{\sigma}$ is not an estimate and could serve as a null hypothesis for testing the HAC estimate of the long-run variance.

[6]For example, the entire path can be decomposed as $ln\left|\widehat{\mathbf{\Phi}}_{N,H}\right| = \sum_{h=2}^{H} ln\left|\widehat{\mathbf{\Sigma}}_{N+h|1\ldots h-1}\right| + ln\left|\widehat{\mathbf{\Sigma}}_{N+1}\right|$ where $\widehat{\mathbf{\Sigma}}_{N+h|1\ldots h-1}$ is the estimated MSE matrix at horizon h conditional on all previous horizons.

4 Simulations

Monte Carlo experiments are used to compare the properties of the path forecast accuracy tests with existing joint tests. I start by describing the path forecast-error generating process. I then describe alternative tests and present the results. All numerical results were obtained using OxMetrics Version 7.2; see Doornik (2013).

4.1 The Path Forecast-Error Generating Process

I generate the path forecast errors as follows:

$$\widetilde{\mathbf{U}}_{t,H,\{b,v,c_k,c_h\}} \equiv \boldsymbol{\theta}_b + \boldsymbol{\Psi}_{\Pi,H}\boldsymbol{\zeta}^{1/2'}_{v,c_k,c_h}\mathbf{V}_{t,H}, \tag{4.1}$$

where $\mathbf{V}_{t,H}$ is an asymmetric Hankel matrix where each unique element is $\mathbf{v}_{t,h} \overset{\text{iid}}{\sim} N_K(\boldsymbol{\mu}, \mathbf{I}_K)$ which is correlated across forecast models by γ. I set $\boldsymbol{\mu} = \mathbf{0}$ and $\gamma = 0.1$. The forecast errors follow a $MA(h-1)$ process and exhibit dependence and biases across variables and horizons. They 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 $\boldsymbol{\zeta}_{v,c_k,c_h}$. Note that this simulation set-up is similar to Quaedvlieg (2019) with the addition that the forecast errors follow a $MA(h-1)$ process and the specification of a richer correlation structure that allows me to explore the impact of changes in cross-horizon forecast-error dynamics.

Since forecast errors typically 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. I use the correlation matrix C, with elements ρ_{g,h,l,k,c_k,c_h}:

$$\rho_{g,h,l,k,c_k,c_h} = \begin{cases} 1 & \text{if } g = h,\ l = k \\ \exp\left(-1.2 + 0.025max\left(g,h\right) - 0.125abs\left(h-g\right)\right) + c_h & \text{if } g \neq h,\ l = k \\ \exp\left(-1.8\right) + c_k & \text{if } g = h,\ l \neq k \\ \exp\left(-1 - \sqrt{abs((k-l)(h-g))}\right) + \frac{c_k + c_h}{2} & \text{if } g \neq h,\ l \neq k \end{cases} \tag{4.2}$$

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 the simulations. c_k and c_h are set equal to zero while under the alternative they vary across models. I allow the variance to change across horizons so that $\sigma_{v,h,k} = v\left(1 + \frac{\sqrt{h-1}}{2}\right)$ where $v = 1$. The variance and correlation are

combined so that $\boldsymbol{\zeta}_{v,c_k,c_h} = diag\left(\boldsymbol{\sigma}_v\right) C_{c_k,c_h} diag\left(\boldsymbol{\sigma}_v\right)$.

Now define the model-specific bias at each horizon as

$$\theta_{b,h} = b\left(1 + \sqrt{h-1}\right), \tag{4.3}$$

where the bias increases with the horizon but is similar across variables. b governs the degree of bias across models where in the baseline case $b = 1$. $\boldsymbol{\theta}_b$ shifts all horizons or variables up or down by a fixed spread, while changes in $\boldsymbol{\mu}$ increase the spread across horizons.

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

$$\boldsymbol{\Psi}_{\Pi,H} = \begin{pmatrix} \mathbf{I}_K & \mathbf{0} & \cdots & \mathbf{0} \\ \boldsymbol{\Pi} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \boldsymbol{\Pi}^{H-1} & \cdots & \boldsymbol{\Pi} & \mathbf{I}_K \end{pmatrix}, \tag{4.4}$$

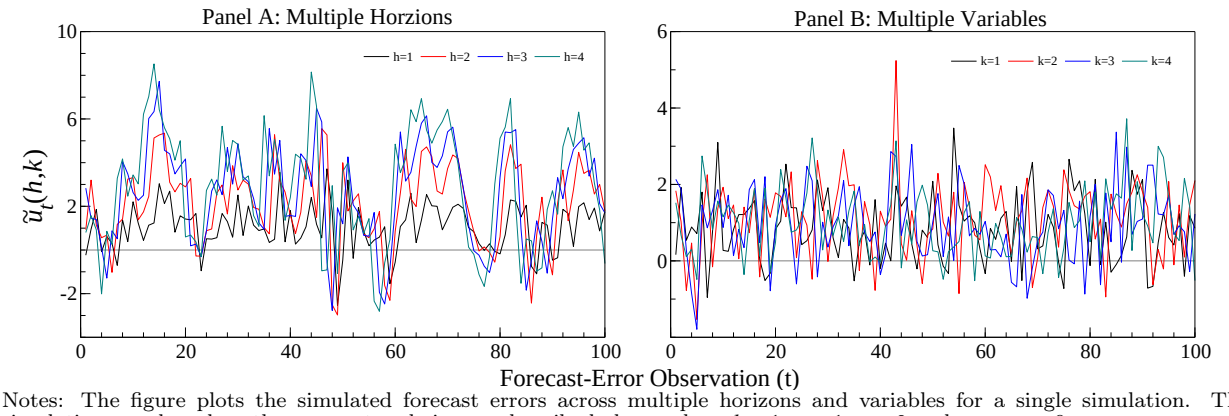
so that serial correlation accumulates across horizons as a $MA(h-1)$ process. I define the $\pi_{j,k}$ elements for each $K \times K$ matrix $\boldsymbol{\Pi}$ as:

$$\pi_{j,k} = \begin{cases} 0.4 + min\left(\frac{k}{10}, 0.5\right) & j = k \\ 0.2 & j \neq k \end{cases}. \tag{4.5}$$

Note that $\boldsymbol{\Psi}_{\Pi,H}$ does not matter when $\boldsymbol{\theta}_b = \boldsymbol{\mu} = \mathbf{0}$ since $\left|\boldsymbol{\Psi}_{\Pi,H}\right| = 1$. Figure 4.1 visualizes the choice of the DGP by plotting the forecast errors across 100 observations. The pattern of the forecast errors across horizons is similar to the U.S. Congressional Budget Office's path forecast errors for the gross federal debt; see Martinez (2015). Note that alternative parameterizations do not substantively change the overall results.

I compare the performance of four test statistics: (1) the average univariate test (i.e. aSPA), (2) the average multivariate test (i.e. aMSPA), (3) the equally weighted path test from (3.7), and (4) the normal path test from (3.10). The first represents the approach in Capistrán (2006) which does not explicitly capture differences in covariances or dynamics. The second is an average of the log determinant measure from Doan et al. (1984) across horizons which captures the covariances but not the dynamics.

The standard errors for the first three test statistics are computed using the HAC estimator from Andrews (1991). Harvey et al. (1997)'s small sample correction is applied to all test statistics. The block bootstrap procedure from Quaedvlieg (2019) can also be used.



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 4.1: Illustration of Forecast-Error Data Generating Process

4.2 Null Rejection Frequency

I first evaluate the tests under the null hypothesis where the forecast accuracy from the two forecast methods is essentially the same. Since tests are conducted based on a nominal size of 5%, then the expected null rejection frequency for each test statistic is 5%.

Table 4.1: Null Rejection Frequencies for Tests of Equal Path Forecast Accuracy

	aSPA (HAC)				aMSPA (HAC)				General Path (HAC)				Normal Path			
N↓ ‖ H→	2	4	12	24	2	4	12	24	2	4	12	24	2	4	12	24
32	1.92	0.13	0.02	0.02	2.81	0.87	0.29	0.13	2.81	0.53	0.08	0.25	5.48	5.27	4.87	4.71
64	3.41	1.12	0.08	0.02	4.30	2.83	0.62	0.26	4.51	2.46	0.17	0.14	5.22	5.05	5.12	4.96
128	4.29	3.05	0.18	0.01	4.72	4.60	1.89	0.40	4.91	4.21	0.59	0.13	4.99	4.90	4.86	5.03
256	4.90	4.17	1.55	0.73	5.24	4.90	4.28	2.08	5.14	4.60	2.81	1.41	5.32	4.75	5.00	5.08
512	5.12	4.58	3.38	1.89	5.35	5.05	4.79	3.60	5.36	5.24	4.17	2.98	5.14	5.23	5.03	5.23
1000	5.26	4.85	4.13	2.82	5.33	5.10	5.01	4.42	5.13	4.82	4.73	4.29	4.94	4.62	4.90	5.00

Notes: The nominal size is 5%, K=1, 20,000 Replications.

Table 4.1 illustrates how the empirical rejection frequencies for each test statistic change with the number of forecast-error observations and the length of the forecast horizon. The test statistics which utilize the HAC estimator all start off undersized, especially for longer path forecasts. At longer forecast paths they remain undersized even as the number of forecast-error observations increases to 1000. Simulations also indicate that the block bootstrap procedure by Quaedvlieg (2019) does not produce a substantive improvement in this case; see Table A.1. The normal path test, which does not rely on a HAC or bootstrap estimator, remains close to the nominal size across all forecast error observations and for all lengths of the forecast path even as $H \to N$. This illustrates the large gains from not relying on the HAC estimator when there is a large amount of persistence.

Table 4.2: Null Rejection Frequencies Across Horizons and Variables

	aSPA (HAC)				aMSPA (HAC)				General Path (HAC)				Normal Path			
N↓ ‖ H=K→	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
32	3.50	1.81	1.19	1.12	4.59	3.26	2.62	2.53	4.59	3.33	2.56	2.96	6.27	5.61	4.30	4.27
64	4.13	3.67	2.87	2.75	4.72	4.91	4.91	5.23	4.72	5.06	4.79	5.01	5.41	5.25	4.63	4.43
128	4.50	4.35	3.73	3.95	4.76	4.96	5.00	5.80	4.76	4.89	5.36	5.70	5.04	4.85	4.41	4.43
256	4.88	4.86	4.61	4.72	4.99	5.44	4.89	5.43	4.99	5.32	5.19	5.76	5.03	4.93	4.25	4.54
512	4.97	5.02	4.78	4.89	5.05	5.25	5.12	5.17	5.05	5.12	5.18	5.46	5.09	4.87	4.64	4.48
1000	5.08	4.99	5.23	5.15	5.13	5.06	5.26	5.16	5.13	5.06	5.10	5.30	5.09	4.87	4.58	4.59

Notes: The nominal size is 5%, 20,000 Replications.

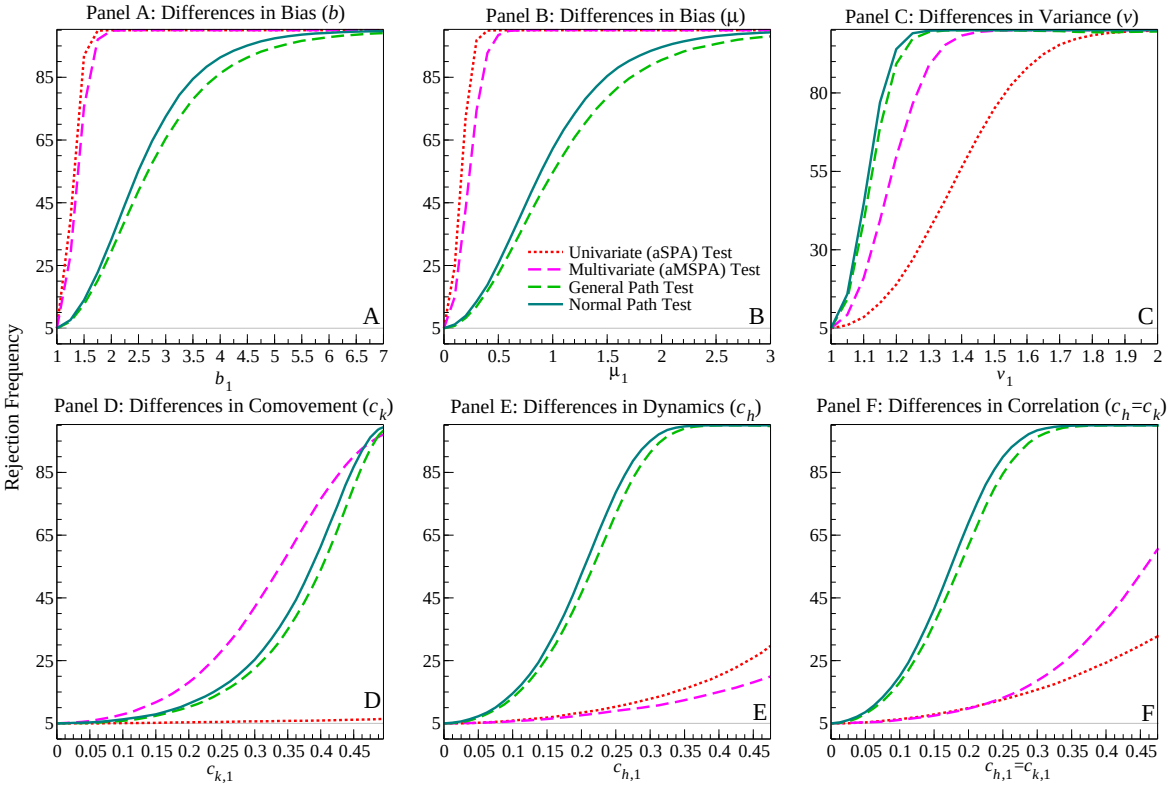
Table 4.2 shows that the short-comings of the HAC estimator are lessened when both the number of variables and the forecast horizons increase. However, the normal path test still outperforms when K increases with H, with the null rejection frequency remaining close to the nominal size. Note that when $H = 1$ the aMSPA and the general path tests are identical since they both use the multivariate normal density as a loss function.

4.3 Non-Null Rejection Frequency

Next I evaluate the tests when the null hypothesis of equal path forecast accuracy does not hold. Although the literature typically only considers differences in forecast bias, I examine differences in bias, variance, and correlation across variables and horizons. This provides a holistic picture of how the tests perform for a set of different hypotheses. I focus on the case where $K = 2$, $H = 4$ and $N = 200$ and adjust the nominal size of all of the tests under the null so that the empirical size is 5% to facilitate comparisons.

Figure 4.2 illustrates the results. Panels A and B indicate that the path forecast tests are less able to distinguish differences in forecast bias when compared against other joint accuracy tests. This is due to the fact that the bias persists across multiple horizons. This persistence implies an increase in the forecast-error covariance, which offsets increases in forecast biases at individual horizons. Thus, path forecast accuracy tests require that biases are idiosyncratic across horizons or are very large in order to detect differences.

Path forecast accuracy tests are better at detecting most other forecast-error differences; see panels C, D, E and F of Figure 4.2. Panel C shows that path tests are better at detecting differences in the forecast-error variances. However, the real strength of the path tests stems from their ability to detect differences in dynamics (panel E) across forecast horizons. In each of these cases, the normal path test has the highest rejection frequency followed by the general path test. This illustrates that one of the main advantages to using a joint multi-



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%. 20,000 replications.

Figure 4.2: Rejection Frequencies for Different Alternatives

horizon loss function is its ability to capture differences in forecast-error covariances (i.e. dynamics) across horizons. This is especially important for understanding whether forecast errors propagate across horizons differently in alternative models.

5 Evaluating Macroeconomic Path Forecasts

I now use the path tests to compare the accuracy of the Federal Reserve Board's Greenbook path forecasts against other path forecasts of GDP growth, inflation, and interest rates. While it is natural to focus on the path of the level of GDP, analyzing the path of GDP growth rates is identical to the levels when using the GFESM, since it is invariant to this transformation. The data and forecasts used in this analysis are real GDP growth (hereafter referred to as GDP), the GDP deflator (hereafter referred to as inflation) and the federal funds rate (hereafter referred to as interest rate) spanning the so-called Great Moderation period from 1985Q4-2000Q4. The 'actual' values are those published in the Federal Reserve Board's Greenbook two quarters after the quarter to which the data refer. This follows the common practice of using the third (final) data estimates.

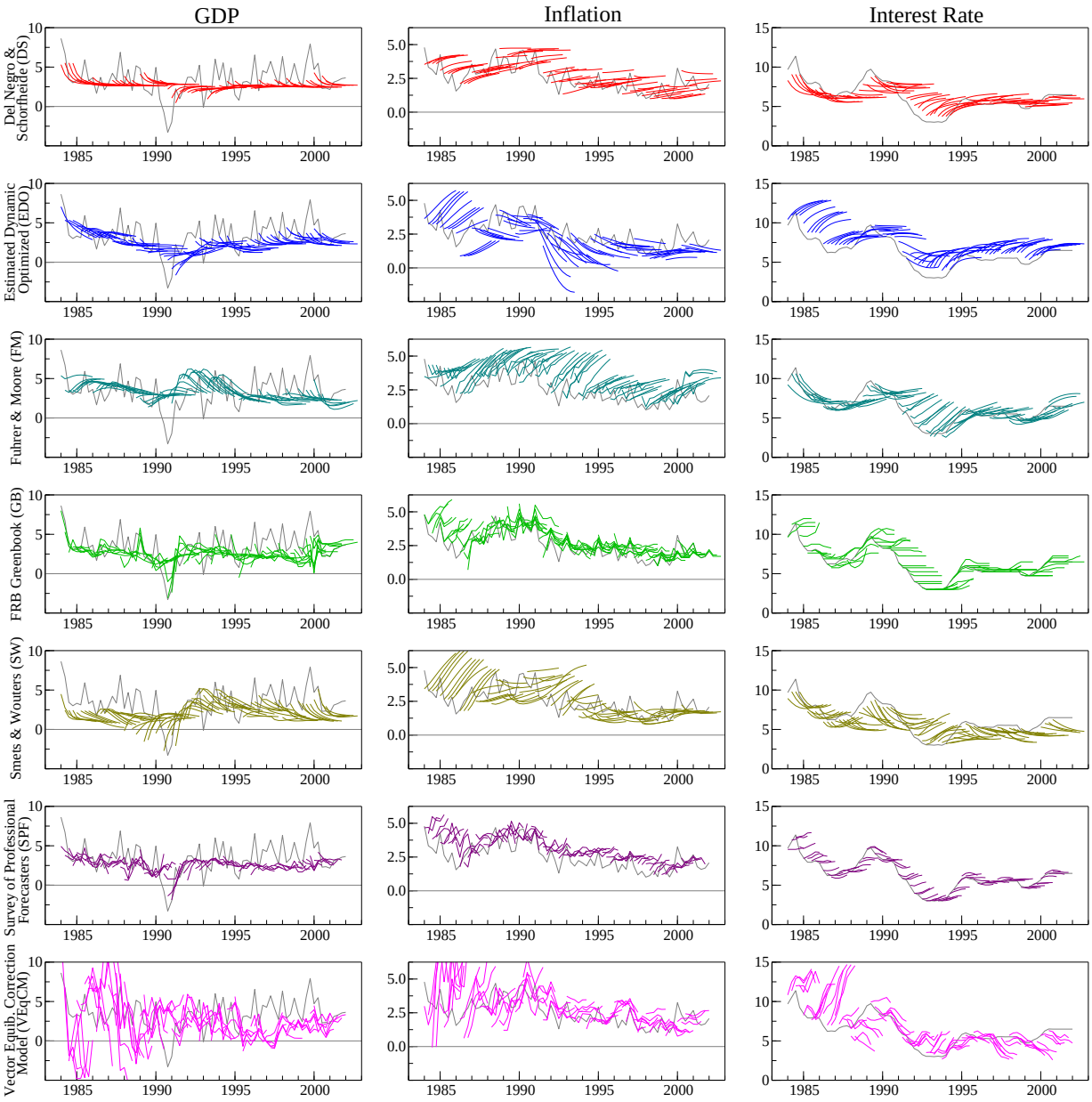


Figure 5.1: 'Path' Forecasts and Actuals, 1984-2002

I compare the Greenbook (GB) forecasts against four DSGE models, a Vector Equilibrium Correction Model (VEqCM), and the Survey of Professional Forecasters (SPF). The DSGE model forecasts were generated by Wolters (2015) using the real-time dataset from Faust and Wright (2009).[7] They were chosen because of their academic and policy relevance as well as their wide-spread use before the Great Recession and include Del Negro and Schorfheide (2004, "DS"), Edge et al. (2008, "EDO"), Fuhrer (1997, "FM"), and Smets and Wouters (2007, "SW"). I generate the VEqCM model forecasts following Anderson et al. (2002) using the real-time datasets from Croushore and Stark (2001) and Faust and Wright (2009). I also include the median SPF forecast. While the SPF doesn't produce a forecast of the federal funds rate, I adjust the 3 month Treasury bill rate path forecast using the real-time gap between the T-bill rate and the federal funds rate when the forecast was generated.

Figure 5.1 plots the actual values and the paths of the forecasts made in each quarter for all models and variables. The columns are spanned by different variables, while the rows are spanned by different models. Looking at the forecasts for GDP, the DSGE models and the SPF do not capture the dynamics but adhere closely to the long-run mean of the series. This is particularly true for forecasts from the DS model. The Greenbook path forecasts and the VEqCM do attempt to capture some of the dynamics away from the mean, however the VEqCM produces erratic forecasts early on in the sample.

The DSGE models have monotonic path forecasts of inflation. The DS model's inflation forecasts are constant across horizons, while the FM model's forecasts trend upwards towards a much higher equilibrium, which is likely influenced by the high inflation period prior to the start of the forecast sample. The EDO and SW inflation forecasts are more reactive to the business cycle despite having periods of both substantial over and under prediction.

Forecasts of the interest rate are similar across models. The Greenbook 'forecasts' of the interest rate over this period represent the interest rate path, typically no change, upon which the Greenbook forecasts are conditioned. The DS model's forecasts adhere closely to the mean, while the EDO and FM forecasts over predict the interest rate which is consistent with the fact that they also over-predict inflation. The SW model's forecasts capture some of the dynamics despite under-predicting the interest rate for long periods.

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

Although the forecasts extend out through eight-quarters-ahead, I only evaluate them up through four-quarters-ahead due to data limitations. I include the nowcasts despite the fact that there is a potentially large information advantage for the Greenbook and SPF. This is because the model forecasts only use data from the previous quarter and are not augmented by higher frequency data. Despite this, the results are robust to excluding the nowcasts or to adjusting the forecasts to account for the Greenbook nowcast.

It is useful to examine the stacked forecast-error second moment matrices to get a preliminary view on whether there is value to evaluating the path jointly. These are depicted in Figure A.1 for all of the forecast methods. It shows the rich covariance structure in the forecast-errors and indicates that there are vast differences in this structure across the methods, especially for the EDO and the VEqCM models.

I take the Greenbook as the baseline against which I compare the other path forecasts. I measure the accuracy of the path forecasts using the GFESM and test for differences using the general and the normal path tests. For the normal path test I allow for time-varying bias following Hendry and Martinez (2017), which captures any deviations from normality.

Table 5.1 presents the results. Both tests present fairly consistent results. However, the normal path test tends to reject the null hypothesis of equal path accuracy with a higher level of confidence than the general path test. For example, for GDP, the general path test rejects the null hypothesis that the DS model and the Greenbook have equal path forecast accuracy with a p-value of 4.5 percent, whereas the normal path test rejects the hypothesis with a p-value of 0 percent. This is consistent with the simulation results where the normal path test has uniformly higher rejection frequencies than the general path test.

The tests disagree when considering all three variables jointly as a system. For example, the general path test rejects the null hypothesis of equal path forecast accuracy for the DS model and the Greenbook with a p-value of 4.3 percent whereas the normal path test fails to reject the null hypothesis with a p-value of over 25 percent. Overall, the tests and metrics indicate that the DS model performs best, however that is due to its performance for GDP, which dominates the other variables in the system due to its larger variation. While the Greenbook path forecasts do not perform best in any case, their accuracy is not statistically distinguishable from the best forecast method (SPF) for inflation and interest rates.

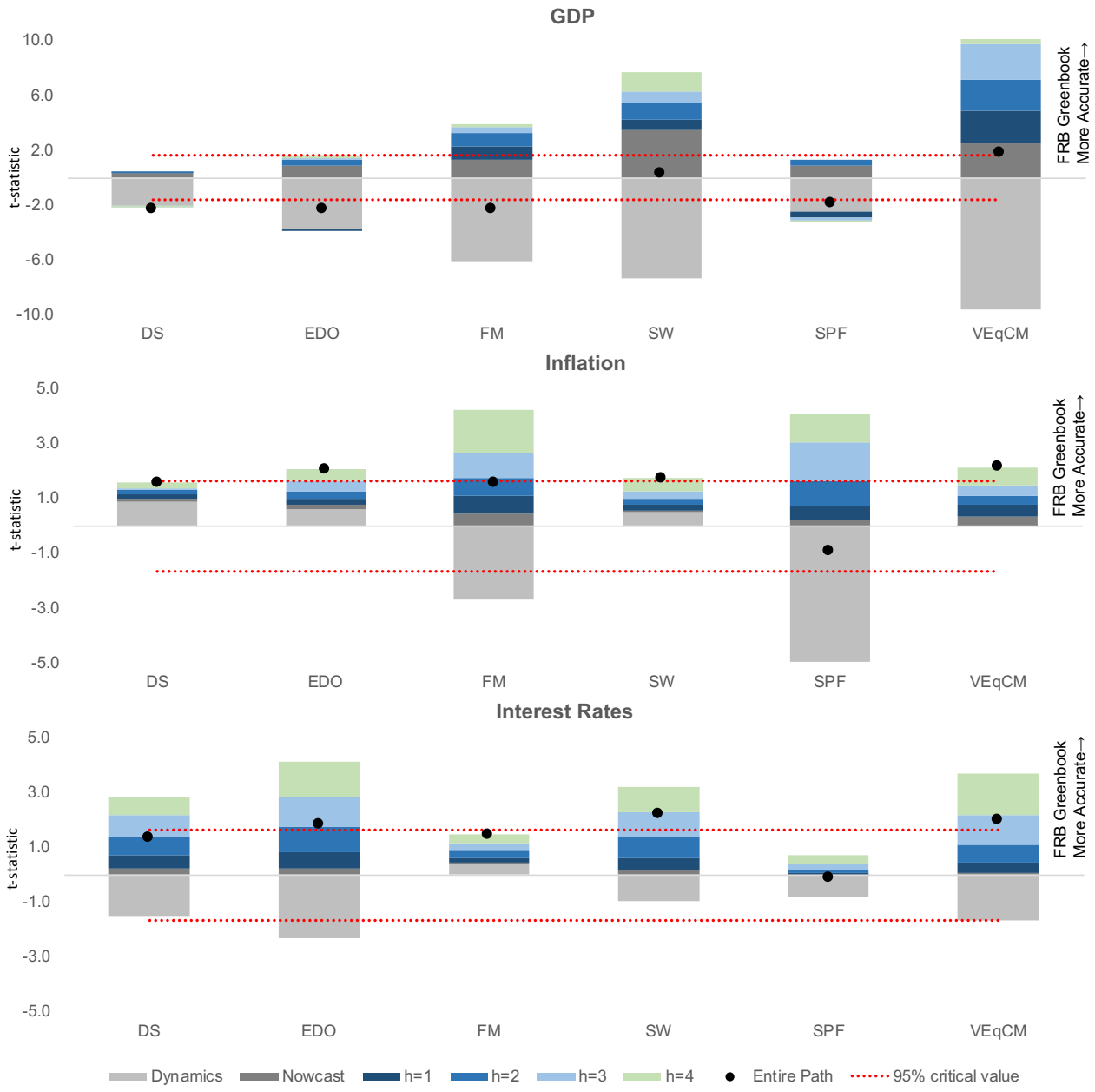
Table 5.1: Testing for Equal Macroeconomic Path Accuracy

	GB	DS	EDO	FM	SW	SPF	VEqCM
GDP							
GFESM (in percentage points):	0.95	**0.36**	0.66	0.67	0.98	0.76	1.93
General path test		1.70**	2.30**	2.28**	0.34	1.91**	1.83**
		[0.045]	[0.011]	[0.011]	[0.369]	[0.028]	[0.033]
Normal path test (time-varying bias)		10.83***	3.85***	3.51***	0.34	2.66***	7.79***
		[0.000]	[0.000]	[0.000]	[0.368]	[0.004]	[0.000]
Inflation							
GFESM (in percentage points):	0.48	0.63	0.73	0.56	0.66	**0.46**	0.96
General path test		1.56*	2.06**	1.55*	1.73**	0.91	2.14**
		[0.060]	[0.019]	[0.061]	[0.042]	[0.182]	[0.016]
Normal path test (time-varying bias)		3.15***	4.99***	1.78**	4.06***	0.53	9.07***
		[0.000]	[0.000]	[0.037]	[0.000]	[0.297]	[0.000]
Interest Rates							
GFESM (in percentage points):	0.36	0.46	0.59	0.49	0.56	**0.36**	0.95
General path test		1.34*	1.81**	1.47*	2.22**	0.10	2.02**
		[0.090]	[0.035]	[0.071]	[0.013]	[0.461]	[0.022]
Normal path test (time-varying bias)		3.06***	6.93***	3.98***	5.95***	0.18	13.80***
		[0.001]	[0.000]	[0.000]	[0.000]	[0.429]	[0.000]
3-Variable System							
GFESM (in percentage points):	0.50	**0.34**	0.49	0.46	0.56	0.45	0.97
General path test		1.72**	0.41	1.08	1.31*	2.52***	1.66**
		[0.043]	[0.340]	[0.141]	[0.094]	[0.006]	[0.048]
Normal path test (time-varying bias)		0.62	0.13	0.55	2.38***	0.35	9.46***
		[0.268]	[0.449]	[0.289]	[0.009]	[0.363]	[0.000]

Notes: Tests include all forecast horizons through 4-quarters-ahead. Bold values indicate the best path forecast. The p-value associated with the tail probability of the null hypothesis is in brackets: *p< 0.1 **p< 0.05 ***p< 0.01.

Next I decompose the path tests to better understand what aspect of the path forecasts is driving the differences in the forecasts. This is done using the decomposition of the (unweighted) general test in (3.9) in order to obtain the relative contribution of each horizon as well as the joint dynamics across all horizons.

Figure 5.2 presents this decomposition. It indicates that the forecast dynamics (or common error components across horizons) play an important role across all models and variables. Differences in forecast dynamics can exacerbate differences between the models (e.g. inflation), ameliorate, or even flip the differences (e.g. GDP or interest rates). For example, the FM model performs worst across all horizons in terms of the GDP point forecasts. However, when accounting for differences in forecast dynamics, FM's path forecast is significantly more accurate than the Greenbook path forecast. The figure also shows that the information disadvantages in the nowcasts are primarily evident in the GDP forecasts and most significant for the FM, SW, and VEqCM models.

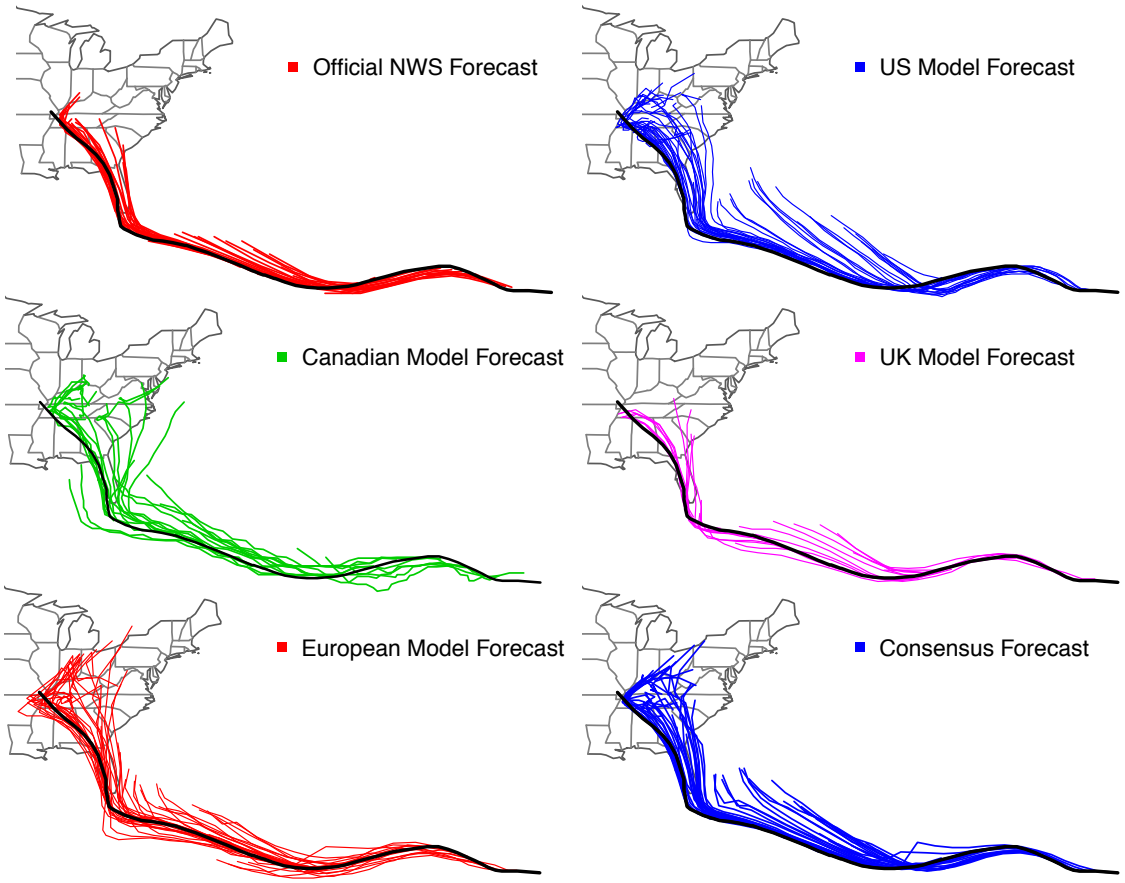


Notes: The decomposition follows from (3.9). See text for model and data definitions.

Figure 5.2: Decomposition of Differences in Path Accuracy (Relative to GB)

6 Path Forecast Accuracy for Hurricane Irma

Now I test for differences in forecasts of the path of Hurricane Irma in 2017 using alternative weighting schemes. Irma formed from a tropical wave near Cape Verde in late August, crossed the Atlantic in early September, and caused an estimated 65 billion dollars in damages in the Caribbean and along the gulf coast of Florida.



Notes: The Official NWS forecast (OFCL) comes from the National Hurricane Center. The US model (AEMN) comes from NOAA's Global Forecasting System. The Canadian model (CMC) is from the Canadian Meteorological Center's global model. The UK Model (EGRR) is from the UK Met's global model. The Consensus model (TVCN) a simple average of the US, Canadian, UK, and European models where available. The European model (EMXI) is the ECMWF model forecast. It was derived from the other forecasts. All plots are restricted to a maximum forecast horizon of 7-days-ahead.

Figure 6.1: Alternative Path forecasts for Hurricane Irma

Although there were over 100 different forecast methods for Hurricane Irma, I focus on six prominent methods: (1) the 'Official' National Weather Service forecast, (2) the 'US model'; (3) the 'Canadian model'; (4) the 'UK model'; the (5) 'European model'; and (6) the 'Consensus' which is a simple combination of the model forecasts. The Official forecast is a subjective / judgmental forecast while the European model is widely considered to be the most accurate model-based forecast. Each forecast was updated every six hours with horizons ranging from 6-hours-ahead to over 7-days-ahead. Figure 6.1 plots the forecasts from each method over time and space in order to provide insight into the differences across the methods. Many of the forecasts including the Official, US, Canadian, UK and Consensus have a clear north-easterly bias at the longest horizons.

I compute the forecast errors using methods for calculating distance on the surface of a spheroid by Vincenty (1975). There are 37 forecast-error observations for each method across seven forecast horizons (12, 24, 36, 48, 72, 96 and 120-hours-ahead) with which to measure and test for differences in forecast accuracy. I compute the unconditional, conditional and joint measures of forecast accuracy across horizons. Figure A.2 illustrates that relative model performance depends on which metric is considered.

I test for differences across models relative to the Official forecast and check how sensitive these differences are to alternative path weights based on the National Hurricane Center's measure of ex-ante forecast uncertainty; i.e. the 'cone of uncertainty'. To discount horizons with more uncertainty, I set weights based on the inverse of the radius of the cone. Alternatively, to focus on horizons with more uncertainty, I set weights based on the length of the radius. For the exact weights see columns 4 and 5 respectively in Table A.2.

Table 6.1: Testing for Irma's Path Accuracy

	OFCL	US	CAN	UK	EU	Con
Equally-weighted GFESM (in miles):	18.4	20.3	22.7	21.6	24.4	18.5
General path test statistic		1.92**	2.27**	1.63*	2.54***	0.11
		[0.027]	[0.012]	[0.051]	[0.006]	[0.457]
Short-weighted GFESM (in miles):	16.8	19.9	20.8	19.4	22.5	17.0
General path test statistic		2.61***	2.59***	1.39*	2.59***	0.23
		[0.005]	[0.005]	[0.082]	[0.005]	[0.409]
Long-weighted GFESM (in miles):	21.7	22.6	26.5	25.8	27.3	21.9
General path test statistic		0.77	1.70**	1.44*	2.43***	0.20
		[0.220]	[0.045]	[0.075]	[0.008]	[0.422]

Notes: See notes for Figure 6.1. Tests include all forecast horizons through 120-hours-ahead. The p-value associated with the tail probability of the null hypothesis is in brackets: *p< 0.1 **p< 0.05 ***p< 0.01.

Table 6.1 presents the results. The equally weighted tests indicates that all of the forecasts, except for the consensus model, perform significantly worse than the official forecast. The differences are larger and more significant when focusing on shorter horizons and are smaller and less significant when focusing on longer horizons. In fact, he European model outperforms the official forecast if all of the weight goes to the longest horizon (not shown); which is coincides with the conditional RMSE rankings in Figure A.2.

7 Conclusions

The standard focus on point forecasts ignores forecast dynamics which are crucial to understanding the accuracy of the forecast path. I show that the GFESM can be thought of as a path forecast accuracy metric and is equivalent to choosing a joint multivariate normal density as a loss function. I propose a general test for differences in path forecast accuracy using the link with the joint density. I illustrate that for a multivariate normal loss function, the test is closely linked to existing joint multi-horizon tests, except that it explicitly captures differences in error covariances and dynamics, which other tests do not. I also consider a special case of this test which does not require the use of a heteroskedasticity and autocorrelation consistent estimator of the variance.

Monte Carlo simulations illustrate the benefits and trade-offs associated with using joint tests. There are large gains from avoiding the HAC estimator, particularly for long forecast paths. Although path forecast accuracy tests are less able to detect differences in biases, since they are highly correlated across horizons, they are more likely to capture differences in variances, covariances and dynamics across forecast models. This is particularly true for differences in forecast-error dynamics, which other tests cannot capture.

I apply the path tests in two relevant policy settings. I start by comparing the Federal Reserve Board's Greenbook path forecasts with several model forecasts. Decomposing the differences along the path, I find that forecast dynamics play an important role and that there are common differences across models. In the second application, I evaluate the path forecast accuracy of models for Hurricane Irma in 2017. While the official forecast tends to dominate, the tests show that other models can match its accuracy depending on how weights are assigned along the path.

Overall, the difference in path forecast accuracy provides a new and unique perspective. This is because, unlike point forecasts, the path includes dependencies across horizons. The applications show that these dynamics can play an important role in determining whether there are persistent differences across horizons. These findings are reinforced by the simulation exercise and demonstrate the value of using path forecast accuracy tests to assess the accuracy of multi-horizon forecasting systems.

Bibliography

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, R. G., Hoffman, D. L., and Rasche, R. H. (2002). A Vector Error-Correction Forecasting Model of the U.S. Economy. *Journal of Macroeconomics*, 24(4):569 – 598.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. New York: Wiley, third edition.

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

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

Barendse, S. and Patton, A. J. (2019). Comparing predictive accuracy in the presence of a loss function shape parameter. mimeo.

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

Berkowitz, J. (2001). Testing Density Forecasts, with Applications to Risk Management. *Journal of Business & Economic Statistics*, 19(4):465–474.

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.

Capistrán, C. (2006). On Comparing Multi-horizon Forecasts. *Economics Letters*, 93(2):176–181.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Duxbury Advanced Series. Pacific Grove, CA: Duxbury, second edition.

Christoffersen, P. F. and Diebold, F. X. (1998). Cointegration and Long-Horizon Forecasting. *Journal of Business & Economic Statistics*, 16(4):450–456.

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

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.

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

Del Negro, M. and Schorfheide, F. (2004). Priors From General Equilibrium Models for VARs. *International Economic Review*, 45(2):643–673.

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*. London: Timberlake Consultants Ltd.

Edge, R. M., Kiley, M. T., and Laforte, J.-P. (2008). Natural Rate Measures in an Estimated DSGE Model of the U.S. Economy. *Journal of Economic Dynamics and Control*, 32(8):2512–2535.

Engle, R. F. (1984). Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics. In Griliches, Z. and Michael D. Intriligator, editors, *Handbook of Econometrics*, volume 2, chapter 13, pages 775–826. New York, NY: North Holland.

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.

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 2A, chapter 1, pages 3–56. Amsterdam: 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.

Giacomini, R. and Rossi, B. (2010). Forecast Comparisons in Unstable Environments. *Journal of Applied Econometrics*, 25(4):595–620.

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. J. (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.

Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the Equality of Prediction Mean Squared Errors. *International Journal of Forecasting*, 13(2):281–291.

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.

Hungnes, H. (2018). Encompassing Tests for Evaluating Multi-Step System Forecasts Invariant to Linear Transformations. Discussion Paper No. 871, Statistics Norway, Research Department, Oslo, Norway.

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.

Knüppel, M. (2018). Forecast-error-based Estimation of Forecast Uncertainty When the Horizon is Increased. *International Journal of Forecasting*, 34(1):105–116.

Kolsrud, D. (2007). Time-Simultaneous Prediction Band for a Time Series. *Journal of Forecasting*, 26(3):171–188.

Kolsrud, D. (2015). A Time-Simultaneous Prediction Box for a Multivariate Time Series. *Journal of Forecasting*, 34(8):675–693.

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., Stock, J. H., and Watson, M. W. (2018). HAR Inference: Recommendations for Practice. *Journal of Business & Economic Statistics*, 36(4):541–559.

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

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.

Müller, U. K. (2014). HAC Corrections for Strongly Autocorrelated Time Series. *Journal of Business & Economic Statistics*, 32(3):311–322.

Quaedvlieg, R. (2019). Multi-Horizon Forecast Comparison. *Journal of Business & Economic Statistics*, Forthcoming:1–14.

Schorfheide, F. and Song, D. (2015). Real-Time Forecasting with a Mixed-Frequency VAR. *Journal of Business & Economic Statistics*, 33(3):366–380.

Silvester, J. R. (2000). Determinants of Block Matrices. *Mathematical Gazette*, 84(501):460–467.

Smets, F. and Wouters, R. (2007). Shocks and Frictions in U.S. Business Cycles: A Bayesian DSGE Approach. *American Economic Review*, 97(3):586–606.

Staszewska-Bystrova, A. (2011). Bootstrap Prediction Bands for Forecast Paths From Vector Autoregressive Models. *Journal of Forecasting*, 30(8):721–735.

Vincenty, T. (1975). Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations. *Survey Review*, 23(176):88–93.

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57(2):307–333.

West, K. D. (1996). Asymptotic Inference About Predictive Ability. *Econometrica*, 64(5):1067–1084.

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.

Yellen, J. L. (2012). Perspectives on Monetary Policy. Remarks by Vice Chair Janet L. Yellen At the Boston Economic Club Dinner, Boston, Massachussetts. Available Online At: `https://www.federalreserve.gov/newsevents/speech/yellen20120606a.pdf` [last Accessed: 16th May, 2018], Board of Governors of the Federal Reserve System, Washington, D.C.

Mathematical Appendix

Proof of Theorem 1

From assumption 2, note that

$$\mathbb{E}\left[ln\left\{f_{t,U_{H,j}}\left(\widetilde{\mathbf{U}}^{m}_{t,H,j}\right)\right\}\right]^{2r} = \sum_{h=1}^{H}\mathbb{E}\left[ln\left\{f_{t,u_{h,j}|U_{h-1,j}}\left(\tilde{\mathbf{u}}^{m}_{t,j}(h) \mid \tilde{\mathbf{u}}^{m}_{t,j}(0),\ldots,\tilde{\mathbf{u}}^{m}_{t,j}(h-1)\right)\right\}\right]^{2r} < \infty$$

where the equality follows from the indpendence of conditional densities. The proof then follows from Amisano and Giacomini (2007) who show that assumptions 1-3 satisfy Theorem 4 of Giacomini and White (2006) when $\mathbf{W}_t \equiv \mathbf{z}_t$ and $\Delta L_{m,t} \equiv WLR_{m,t,H}$.

Proof of Theorem 2

I first consider the central case by extending Cai et al. (2015)'s central limit theorem for log determinants to stacked $MA(h-1)$ processes using the following Lemmas:

Lemma 3. Let $\boldsymbol{X}_{H,T+1},\ldots,\boldsymbol{X}_{H,T+N} \sim N_{HK}\left(\boldsymbol{\Xi}_{\mathbf{H}},\boldsymbol{\Omega}_H\right)$ where $\boldsymbol{\Omega}_H$ has dimension $KH\times KH$ with $KH \leq N$ and $\boldsymbol{X}_{H,T+n} = \left\{X'_{T+n},\ldots,X'_{T+H+n-1}\right\}'$ where each $X_{t+n+h-1}$ is of dimension $K\times 1$ and follows a MA$(h-1)$ process. Denote the sample second moment matrix by $\widehat{\boldsymbol{\Phi}}_{N,H}$ and the sample squared non-centrality by $\hat{\boldsymbol{\Theta}}_{N,H}$. Then

$$ln\left\{det\left(\widehat{\Phi}_{N,H}\right)\right\} - ln\left\{det\left(\Phi_H\right)\right\} = ln\left\{det\left(\hat{\mathbf{I}}_{N,KH} + \hat{\boldsymbol{\Theta}}_{N,H} + o_p(1)\right)\right\} - ln\left\{det\left(\mathbf{I}_{KH} + \boldsymbol{\Theta}_H\right)\right\},$$

where $\hat{\mathbf{I}}_{H,N,K} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{Y}_{H,T+n}\boldsymbol{Y}'_{H,T+n}$ is the sample matrix of two stacked vectors where $\boldsymbol{Y}_{H,T+n} = \left\{Y'_{T+n},\ldots,Y'_{T+H+n-1}\right\}'$ and $Y_{T+1},\ldots,Y_{T+H+N-1} \sim IN_K\left(\mathbf{0},\mathbf{I}_K\right)$.

Proof of Lemma 3. Define $\boldsymbol{Y}_{H,T+n} = \boldsymbol{\Omega}_H^{-1/2}\left(\boldsymbol{X}_{H,T+n} - \boldsymbol{\Xi}_{\mathbf{H}}\right)$ which removes any covariance and most of the dependence given the assumed structure so that

$$\begin{aligned}
ln\left\{\frac{det\left(\widehat{\Phi}_{N,H}\right)}{det\left(\Phi_H\right)}\right\} &= ln\left\{\frac{det\left(\widehat{\Phi}_{N,H}\right)}{det\left(\boldsymbol{\Omega}_H + \boldsymbol{\Xi}_{\mathbf{H}}\boldsymbol{\Xi}'_{\mathbf{H}}\right)}\right\} \\
&= ln\left\{\frac{det\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{Y}_{H,T+n}\boldsymbol{Y}'_{H,T+n} + \boldsymbol{\Omega}_H^{-1/2}\hat{\boldsymbol{\Xi}}_{\mathbf{N,H}}\hat{\boldsymbol{\Xi}}'_{\mathbf{N,H}}\boldsymbol{\Omega}_{\mathbf{H}}^{\mathbf{-1/2}} + o_p(1)\right)}{det\left(\mathbf{I}_{KH} + \boldsymbol{\Omega}_{\mathbf{H}}^{\mathbf{-1/2}}\boldsymbol{\Xi}_{\mathbf{H}}\boldsymbol{\Xi}'_{\mathbf{H}}\boldsymbol{\Omega}_H^{-1/2}\right)}\right\} \\
&= ln\left\{det\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{Y}_{H,T+n}\boldsymbol{Y}'_{H,T+n} + \hat{\boldsymbol{\Theta}}_{N,H} + o_p(1)\right)\right\} - det\left(\mathbf{I}_{KH} + \boldsymbol{\Theta}_H\right) \\
&= ln\left\{det\left(\hat{\mathbf{I}}_{N,KH} + \hat{\boldsymbol{\Theta}}_{N,H} + o_p(1)\right)\right\} - det\left(\mathbf{I}_{KH} + \boldsymbol{\Theta}_H\right).
\end{aligned}$$

Lemma 2. The distribution of $ln\left\{det\left(N\hat{\mathbf{I}}_{N,KH}\right)\right\}$ is the same as H times the sum of

K-independent log χ^2 distributions such that

$$ln\left\{det\left(N\hat{\mathbf{I}}_{N,KH}\right)\right\} \overset{D}{=} H\sum_{k=1}^{K} ln\left(\chi^2_{N-k+1}\right),$$

where $\chi^2_N, \ldots, \chi^2_{N-K+1}$ are independent χ^2 distributions with $N-k+1$ degrees of freedom.

Proof of Lemma 2. Define $\mathbf{C}_{h,N,K} = \left\{\frac{1}{N}\sum_{n=1}^{N} Y_{T+n+h-1}Y'_{T+n}, \ldots, \frac{1}{N}\sum_{n=1}^{N} Y_{T+n+h-1}Y'_{T+n+h-2}\right\}'$ where $\mathbf{C}_{1,N,K} = \mathbf{0}$ and $\hat{\mathbf{I}}_{T,N,K,h} = \frac{1}{N}\sum_{n=1}^{N} Y_{T+n+h-1}Y'_{T+n+h-1}$. Then by the properties of block determinants it is possible to decompose as

$$\begin{aligned} ln\left\{det\left(N\hat{\mathbf{I}}_{N,KH}\right)\right\} = &\sum_{h=1}^{H} ln\left\{det\left(N\hat{\mathbf{I}}_{T,N,K,h}\right)\right\} \\ &+ \sum_{h=2}^{H} ln\left\{det\left(\mathbf{I}_K - \hat{\mathbf{I}}^{-1}_{T,N,K,h}\left[\mathbf{C}'_{h,N,K}\hat{\mathbf{I}}^{-1}_{h-1,N,K}\mathbf{C}_{h,N,K}\right]\right)\right\}, \end{aligned}$$

where $\hat{\mathbf{I}}_{h-1,N,K}$ is invertible as long as $K(h-1) \leq N$. The intuition behind this decomposition follows from the properties of determinants of block matrices; see Silvester (2000).

From the Law of Large Numbers for iid variables we know that as $N \to \infty$ then $\hat{\mathbf{I}}_{T,N,K,h} \xrightarrow{P} \mathbf{I}_K$ and $\frac{1}{N}\sum_{n=1}^{N} Y_{T+n+h}Y'_{T+n+j} \xrightarrow{P} \mathbf{0}\ \forall j \neq h$ so that $\hat{\mathbf{I}}_{h-1,N,K} \xrightarrow{P} \mathbf{I}_{K(h-1)}$ and $\mathbf{C}_{h,N,K} \xrightarrow{P} \mathbf{0}$. Since the log determinant is a continuous function, then by the Continuous Mapping Theorem

$$\sum_{h=2}^{H} ln\left\{det\left(\mathbf{I}_K - \hat{\mathbf{I}}^{-1}_{T,N,K,h}\left[\mathbf{C}'_{h,N,K}\hat{\mathbf{I}}^{-1}_{h-1,N,K}\mathbf{C}_{h,N,K}\right]\right)\right\} \xrightarrow{P} 0\ as\ N \to \infty.$$

The first term can be decomposed further as

$$\begin{aligned} \sum_{h=1}^{H} ln\left\{det\left(N\hat{\mathbf{I}}_{T,N,K,h}\right)\right\} = &\sum_{h=1}^{H} ln\left\{det\left(N\hat{\mathbf{I}}_{T,N,K,1} + N\hat{\mathbf{I}}_{T,N,K,h} - N\hat{\mathbf{I}}_{T,N,K,1}\right)\right\} \\ = &H \times ln\left\{det\left(N\hat{\mathbf{I}}_{T,N,K,1}\right)\right\} \\ &+ \sum_{h=2}^{H} ln\left\{det\left(\mathbf{I}_K + \hat{\mathbf{I}}^{-1}_{T,N,K,1}\left[\hat{\mathbf{I}}_{T,N,K,h} - \hat{\mathbf{I}}_{T,N,K,1}\right]\right)\right\}. \end{aligned}$$

Repeatedly applying the Law of Large Numbers for iid variables to each of the individual terms, then by the Continuous Mapping Theorem

$$\sum_{h=2}^{H} ln\left\{det\left(\mathbf{I}_K + \hat{\mathbf{I}}^{-1}_{T,N,K,1}\left[\hat{\mathbf{I}}_{T,N,K,h} - \hat{\mathbf{I}}_{T,N,K,1}\right]\right)\right\} \xrightarrow{P} 0\ as\ N \to \infty.$$

Finally, for fixed H as $N \to \infty$

$$H \times ln\left\{det\left(N\hat{\mathbf{I}}_{T,N,K,1}\right)\right\} \overset{D}{=} H\sum_{k=1}^{K} ln\left(\chi^2_{N-k+1}\right),$$

which follows from the Bartlett decomposition where each variable is independent as $N \to \infty$; see Anderson (2003). Combining each of these results using the Continuous Mapping Theorem for fixed H and K as $N \to \infty$ proves the Lemma.

Theorem 2a. Let $\boldsymbol{X}_{H,T+1}, \ldots, \boldsymbol{X}_{H,T+N} \sim N_{HK}(0, \boldsymbol{\Omega}_H)$ where $\boldsymbol{\Omega}_H$ has dimension $KH \times KH$ with $KH \leq N$ and $\boldsymbol{X}_{H,T+n} = \{X_{T+n}, \ldots, X_{T+H+n-1}\}'$ where each $X_{t+n+h-1}$ is of dimension $K \times 1$ and follows a MA$(h-1)$ process. Suppose that $N \to \infty$. Then the log determinant of the sample second-moment matrix $\widehat{\boldsymbol{\Phi}}_H$ satisfies

$$\frac{ln\left\{det\left(\widehat{\boldsymbol{\Phi}}_{N,H}\right)\right\} - \tau_{N,K,H} - ln\left\{det\left(\boldsymbol{\Phi}_H\right)\right\}}{\sigma_{N,K,H}} \to N(0,1) \ \ as \ N \to \infty,$$

where constants $\tau_{N,K,H}$ and $\sigma_{N,K,H}$ are given as

$$\tau_{N,K,H} := \sum_{k=1}^{K} H\left[\Psi\left(\frac{N-k+1}{2}\right) - ln\left(\frac{N}{2}\right)\right],$$

where $\Psi(x) = \frac{\partial}{\partial z} ln\left(\Gamma(z)\right)|_{z=x}$, where $\Gamma(\cdot)$ is the gamma function and

$$\sigma_{N,K,H} := H\left(\sum_{k=1}^{K} \frac{2}{N-k+1}\right)^{\frac{1}{2}}.$$

Proof of Theorem 2a. It follows from **Lemma 1** and **Lemma 2** that

$$\begin{aligned}
z &= ln\left\{det\left(\widehat{\boldsymbol{\Phi}}_{N,H}\right)\right\} - \tau_{N,K,H} - ln\left\{det\left(\boldsymbol{\Phi}_H\right)\right\} \\
&= ln\left\{det\left(\hat{\mathbf{I}}_{N,KH}\right)\right\} - \sum_{k=1}^{K} H\left[\Psi\left(\frac{N-k+1}{2}\right) - ln\left(\frac{N}{2}\right)\right] \\
&= ln\left\{det\left(N\hat{\mathbf{I}}_{N,KH}\right)\right\} - \sum_{k=1}^{K} H\left[\Psi\left(\frac{N-k+1}{2}\right) + ln(2)\right] \\
&\overset{D}{=} H\sum_{k=1}^{K}\left[ln\left(\chi^2_{N-k+1}\right) - \Psi\left(\frac{N-k+1}{2}\right) - ln(2)\right].
\end{aligned}$$

Using the characteristic function of the log χ^2 distribution gives

$$\begin{aligned}
\phi_z(t) &= \prod_{k=1}^{K} \phi_{H \times ln\left(\chi^2_{N-k+1}\right)}(t) \, \frac{1}{exp\left\{it \times H\left[\Psi\left(\frac{N-k+1}{2}\right) + ln(2)\right]\right\}} \\
&= \prod_{k=1}^{K} \mathbb{E}\left(\chi^2_{N-k+1}\right)^{it \times H} \frac{1}{exp\left\{it \times H\left[\Psi\left(\frac{N-k+1}{2}\right)\right]\right\} 2^{it \times H}}.
\end{aligned}$$

Then applying the formula for moments from the moment generating function of the χ^2

distribution:

$$\begin{aligned}\phi_z(t) &= \prod_{k=1}^{K}\left[\frac{\Gamma\left(\frac{1}{2}(N-k+1)+it\times H\right)}{\Gamma\left(\frac{1}{2}(N-k+1)\right)}\right]2^{it\times H}\frac{1}{exp\left\{it\times H\left[\Psi\left(\frac{N-k+1}{2}\right)\right]\right\}2^{it\times H}}\\ &= \prod_{k=1}^{K}\left[\frac{\Gamma\left(\frac{1}{2}(N-k+1)+it\times H\right)}{\Gamma\left(\frac{1}{2}(N-k+1)\right)}\right]\frac{1}{exp\left(it\times H\left(\Psi\left(\frac{N-k+1}{2}\right)\right)\right)}.\end{aligned}$$

Taking logs gives

$$ln\left\{\phi_z(t)\right\}\overset{D}{=}\sum_{k=1}^{K}\left\{ln\left\{\Gamma\left(\frac{1}{2}(N-k+1)+it\times H\right)\right\}-ln\left\{\Gamma\left(\frac{1}{2}(N-k+1)\right)\right\}-it\times H\left(\Psi\left(\frac{N-k+1}{2}\right)\right)\right\}.$$

Cai et al. (2015) show that when $H=1$ this can be approximated as

$$ln\left\{\phi_z(t)\right\}\overset{D}{\approx}(it)^2\frac{1}{\sum_{k=1}^{K}(N-k+1)}+O\left((t)^2\frac{1}{\sum_{k=1}^{K}(N-k+1)^2}\right).$$

Extending this approximation to $H\geq 1$ gives

$$\begin{aligned}\phi_o(t) &= \phi_z\left(\frac{t}{\sigma_{N,K,H}}\right)\\ &\overset{D}{\approx} exp\left\{\frac{(it\times H)^2}{\sigma^2_{N,K,H}}\frac{1}{\sum_{k=1}^{K}(N-k+1)}+O\left(\frac{(t\times H)^2}{\sigma^2_{N,K,H}}\frac{1}{\sum_{k=1}^{K}(N-k+1)^2}\right)\right\}\\ &= exp\left\{\frac{(it)^2}{\sum_{k=1}^{K}\frac{2}{(N-k+1)}}\frac{1}{\sum_{k=1}^{K}(N-k+1)}+O\left(\frac{t^2}{\sum_{k=1}^{K}\frac{2}{(N-k+1)}}\frac{1}{\sum_{k=1}^{K}(N-k+1)^2}\right)\right\}\\ &= exp\left\{\frac{(it)^2}{2}+O\left(\frac{t^2}{2}\frac{\frac{1}{\sum_{k=1}^{K}(N-k+1)^2}}{\sum_{k=1}^{K}\frac{1}{(N-k+1)}}\right)\right\}\\ &= exp\left\{\frac{(it)^2}{2}+O\left(\frac{t^2}{2}r_{N,K}\right)\right\},\end{aligned}$$

where Cai et al. (2015, Lemma 3) shows that $r_{N,K}\to 0$ as $N\to\infty$ when $K\leq N$. As a result, when $N\to\infty$ then this gives the characteristic function of the standard normal distribution:

$$\phi_o(t)\to e^{\frac{(it)^2}{2}}.$$

Now I consider the non-central case with an additional Lemma:

Lemma 3. Let $\boldsymbol{X}_{T+1},\ldots,\boldsymbol{X}_{T+N}\overset{iid}{\sim}N_K(\boldsymbol{\Xi},\boldsymbol{\Omega})$ where $\boldsymbol{\Omega}$ and $\boldsymbol{\Xi}$ have dimension $K\times K$ with $K\leq N$. Let the sample second-moment matrix $\hat{\boldsymbol{\Phi}}_N$ have a non-central Wishart distribution where $N\hat{\boldsymbol{\Phi}}_N\sim W_K(N,\boldsymbol{\Omega},\boldsymbol{\Theta})$ and $\boldsymbol{\Theta}=\boldsymbol{\Omega}^{-1/2'}\boldsymbol{\Xi}'\boldsymbol{\Xi}\boldsymbol{\Omega}^{-1/2}$ is the squared non-centrality. Then

the log determinant of the sample second-moment matrix satisfies

$$\frac{\sqrt{N}\left\{ln\left\{det\left(\hat{\mathbf{\Phi}}_N\right)\right\} ln\left\{det\left(\mathbf{\Phi}\right)\right\}\right\}}{\sqrt{2\times tr\left[\left(\mathbf{I}_K+2\mathbf{\Theta}\right)\left(\mathbf{I}_K+\mathbf{\Theta}\right)^{-2}\right]}}\rightarrow N\left(0,1\right)\ as\ N\rightarrow\infty.$$

Proof of Lemma 3. The proof follows from Fujikoshi (1968, Theorem 1) when applying the delta-method from Casella and Berger (2002, Theorem 5.5.24) where the natural log is a continuous function.

Theorem 2b. Let $\boldsymbol{X}_{H,T+1},\ldots,\boldsymbol{X}_{H,T+N}\sim N_{HK}\left(\mathbf{\Xi}_H,\mathbf{\Omega}_H\right)$ where $\mathbf{\Omega}_H$ has dimension $KH\times KH$ with $KH\leq N$ and $\boldsymbol{X}_{H,T+n}=\left\{X_{T+n},\ldots,X_{T+H+n-1}\right\}'$ where each $X_{t+n+h-1}$ is of dimension $K\times 1$ and follows a MA$(h-1)$ process. Let the sample second-moment matrix $\widehat{\mathbf{\Phi}}_{N,H}$ have a non-central Wishart distribution where $N\widehat{\mathbf{\Phi}}_{N,H}\sim W_K\left(N,\mathbf{\Omega}_H,\mathbf{\Theta}_H\right)$ and $\mathbf{\Theta}_H=\mathbf{\Omega}_H^{-1/2'}\mathbf{\Xi}_H'\mathbf{\Xi}_H\mathbf{\Omega}_H^{-1/2}$ is the squared non-centrality. Then $ln\left\{det\left(\widehat{\mathbf{\Phi}}_{N,H}\right)\right\}$ satisfies

$$\frac{\sqrt{N}\left\{ln\left\{det\left(\widehat{\mathbf{\Phi}}_{N,H}\right)\right\}-ln\left\{det\left(\mathbf{\Phi}_H\right)\right\}\right\}}{\sqrt{2H\times tr\left[\left(\mathbf{I}_{HK}+2\mathbf{\Theta}_H-\frac{1}{2}\left(\frac{(H-1)^2+\mathbb{1}_{H>1}}{H^2}\right)\mathbf{\Theta}_H'\mathbf{\Theta}_H\right)\left(\mathbf{I}_{HK}+\mathbf{\Theta}_H\right)^{-2}\right]}}\rightarrow N\left(0,1\right)\ as\ N\rightarrow\infty.$$

Proof of Theorem 2b. I prove the theorem for the case where $H=2$ and then generalize the result using Lemma 3. It follows from Lemma 1 that

$$ln\left\{det\left(\widehat{\mathbf{\Phi}}_{N,2}\right)\right\}-ln\left\{det\left(\mathbf{\Phi}_2\right)\right\}=ln\left\{det\left(\hat{\mathbf{I}}_{N,2}+\hat{\mathbf{\Theta}}_{N,2}+o_p(1)\right)\right\}-ln\left\{det\left(\mathbf{I}_2+\mathbf{\Theta}_2\right)\right\},$$

where assuming that the squared non-centrality is identical across horizons gives

$$\hat{\mathbf{I}}_{N,2}+\hat{\mathbf{\Theta}}_{N,2}=\begin{pmatrix}\hat{\sigma}_{N,1}^2 & \tilde{\sigma}_{N,1,2}\\ \tilde{\sigma}_{N,1,2} & \hat{\sigma}_{N,1}^2+o_p(1)\end{pmatrix}+\begin{pmatrix}\hat{\theta}_N^2 & \hat{\theta}_N^2\\ \hat{\theta}_N^2 & \hat{\theta}_N^2\end{pmatrix}$$

where $\hat{\sigma}_{N,1}^2\sim N\left(1,2\right)$, $\hat{\theta}_N\sim N\left(\theta,1\right)$, and $\tilde{\sigma}_{N,1,2}\sim N\left(0,1\right)$ so that

$$\hat{\eta}-\eta=\begin{pmatrix}\hat{\theta}_N\\ \hat{\sigma}_{N,1}^2\\ \tilde{\sigma}_{N,1,2}\end{pmatrix}-\begin{pmatrix}\theta\\ 1\\ 0\end{pmatrix}\sim N\left[\begin{pmatrix}0\\ 0\\ 0\end{pmatrix},\begin{pmatrix}1 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 1\end{pmatrix}\right].$$

The multivariate delta method, see Casella and Berger (2002, Theorem 5.5.28), implies that

$$\sqrt{N}\left(h\left(\hat{\boldsymbol{\eta}}\right)-h\left(\boldsymbol{\eta}\right)\right)\sim N\left[\mathbf{0},\frac{\partial h\left(\boldsymbol{\eta}\right)}{\partial\boldsymbol{\eta}}'\begin{pmatrix}1 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 1\end{pmatrix}\frac{\partial h\left(\boldsymbol{\eta}\right)}{\partial\boldsymbol{\eta}}\right].$$

Choosing the function as

$$h(\boldsymbol{\eta}) = ln\left\{det\begin{pmatrix} \sigma_1^2+\theta^2 & \sigma_{1,2}+\theta^2 \\ \sigma_{1,2}+\theta^2 & \sigma_1^2+\theta^2 \end{pmatrix}\right\} = ln\left\{\left(\sigma_1^2+\theta^2\right)^2 - \left(\sigma_{1,2}+\theta^2\right)^2\right\}$$
$$= ln\left\{\left(\sigma_1^2\right)^2 + 2\sigma_1^2\theta^2 - \left(\sigma_{1,2}\right)^2 - 2\sigma_{1,2}\theta^2\right\},$$

then the vector of partial derivatives is

$$\frac{\partial h(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \left\{\frac{4\theta}{1+2\theta^2}, \frac{2\left(1+\theta^2\right)}{1+2\theta^2}, \frac{-2\theta^2}{1+2\theta^2}\right\}',$$

which gives

$$\sqrt{N}\left(ln\left\{det\left(\widehat{\Phi}_{N,2}\right)\right\} - ln\left\{det\left(\Phi_2\right)\right\}\right) \sim N\left[0, \frac{8\left(1+2\theta^2+\frac{3}{2}\theta^4\right)}{\left(1+2\theta^2\right)^2}\right].$$

This can be written more generally as

$$\sqrt{N}\left(ln\left\{\frac{det\left(\widehat{\Phi}_{N,2}\right)}{det\left(\Phi_2\right)}\right\}\right) \sim N\left[0, 4\times tr\left\{\left(\mathbf{I}_2 + 2\boldsymbol{\Theta}_2 - \frac{1}{4}\boldsymbol{\Theta}_2'\boldsymbol{\Theta}_2\right)\left(\mathbf{I}_2+\boldsymbol{\Theta}_2\right)^{-2}\right\}\right],$$

which generalizes for multiple H and K following Lemma 3 so that

$$\frac{\sqrt{N}\left(ln\left\{det\left(\widehat{\Phi}_{N,H}\right)\right\} - ln\left\{det\left(\Phi_H\right)\right\}\right)}{\sqrt{2H\times tr\left\{\left(\mathbf{I}_{KH} + 2\boldsymbol{\Theta}_H - \frac{1}{2}\left(\frac{(H-1)^2+\mathbb{1}_{H>1}}{H^2}\right)\boldsymbol{\Theta}_H'\boldsymbol{\Theta}_H\right)\left(\mathbf{I}_{KH}+\boldsymbol{\Theta}_H\right)^{-2}\right\}}} \rightarrow N[0,1].$$

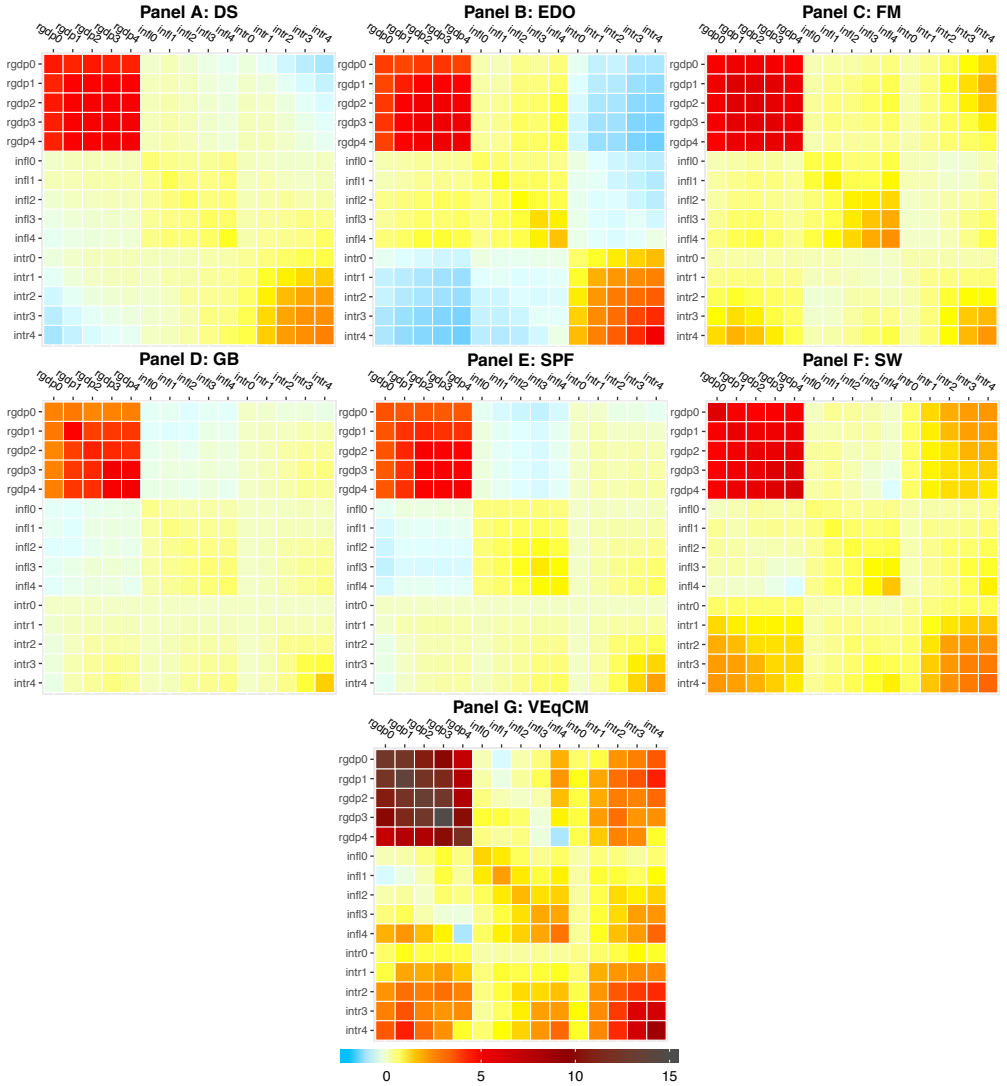
The rest of the result follows directly from the multivariate delta method when considering the forecast errors from both methods simultaneously and allowing them to be correlated.

A Additional Tables and Figures

Table A.1: Null Rejection Frequencies for Tests of Equal Path Forecast Accuracy

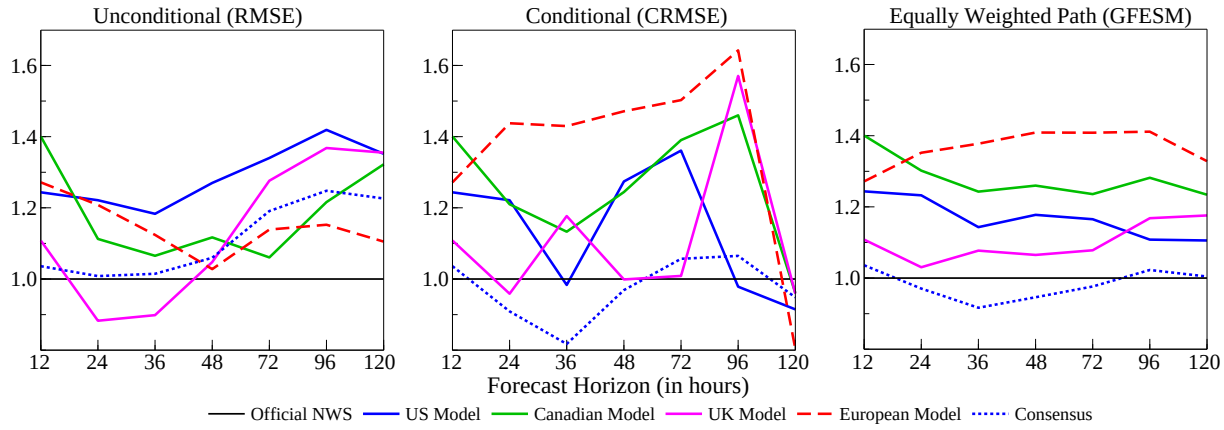
	aSPA (bootstrap)				aMSPA (bootstrap)				Path (bootstrap)				Normal Path			
N↓ ‖ H→	2	4	12	24	2	4	12	24	2	4	12	24	2	4	12	24
32	1.86	0.88	0.87	1.54	2.42	1.48	1.34	2.15	2.44	1.04	0.90	2.23	5.48	5.27	4.87	4.71
64	3.14	1.27	0.42	0.62	3.85	2.66	0.97	1.06	3.79	2.12	0.61	0.66	5.22	5.05	5.12	4.96
128	4.02	2.71	0.49	0.40	4.65	3.94	2.08	0.94	4.73	3.71	0.88	0.88	4.99	4.90	4.86	5.03
256	4.85	3.93	1.52	0.99	5.17	4.65	3.86	2.26	4.90	4.71	2.82	1.68	5.32	4.75	5.00	5.08
512	4.89	4.65	3.23	2.00	5.01	5.10	4.86	3.69	5.25	4.73	4.05	3.04	5.14	5.23	5.03	5.23
1000	4.96	4.83	4.30	3.24	5.11	5.01	5.11	4.69	5.11	4.87	4.37	4.13	4.94	4.62	4.90	5.00

Notes: The nominal size is 5%, K=1, 20,000 Replications.



Notes: rgdp = GDP, infl = inflation, intr = interest rate and where the numbers denote the forecast horizon from 0 to 4 quarters ahead. The matrices are symmetric by construction where the mean square forecast errors for each variable and horizon fall along the main diagonal of each matrix. See text for model definitions.

Figure A.1: Forecast-Error Second-Moment Matrices by Forecast Method



Notes: All metrics are estimated using 37 observations. Values less than 1 indicate that the method is more accurate than the 'Official' NWS forecast while values greater than 1 indicate that the method is less accurate. The Conditional RMSE (CRMSE) is computed as the RMSE for a given horizon conditional on the accuracy of all prior horizons.

Figure A.2: Relative Forecast Accuracy Rankings for Hurricane Irma

Table A.2: Alternative Weights for Hurricane Irma

Horizon	Radius	Equal Weight	Short Weighted	Long Weighted
12	29	1	2.282	0.293
24	45	1	1.471	0.455
36	63	1	1.050	0.637
48	78	1	0.848	0.789
72	107	1	0.618	1.082
96	159	1	0.416	1.608
120	211	1	0.314	2.134

Notes: Radius is in miles. Weights sum to the number of horizons. The radius comes from the National Hurricane Center in 2017: `https://web.archive.org/web/20170906033729/https://www.nhc.noaa.gov/aboutcone.shtml`