Reality Checks and Comparisons of Nested Predictive Models

Reality Checks and Nested Forecast Model Comparisons *

Todd E. Clark
Federal Reserve Bank of Kansas City

Michael W. McCracken
Federal Reserve Bank of St. Louis

September 2010

Abstract

This paper develops a novel and effective bootstrap method for simulating asymptotic critical values for tests of equal forecast accuracy and encompassing among many nested models. The bootstrap, which combines elements of fixed regressor and wild bootstrap methods, is simple to use. We first derive the asymptotic distributions of tests of equal forecast accuracy and encompassing applied to forecasts from multiple models that nest the benchmark model – that is, reality check tests applied to nested models. We then prove the validity of the bootstrap for these tests. Monte Carlo experiments indicate that our proposed bootstrap has better finite-sample size and power than other methods designed for comparison of non-nested models. We conclude with empirical applications to multiple-model forecasts of commodity prices and GDP growth.

JEL Nos.: C53, C12, C52
Keywords: Prediction, forecast evaluation, equal accuracy

* *Clark (corresponding author)*: Economic Research Dept.; Federal Reserve Bank of Kansas City; 1 Memorial Drive; Kansas City, MO 64198; todd.e.clark@kc.frb.org. *McCracken*: Research Division; Federal Reserve Bank of St. Louis; P.O. Box 442; St. Louis, MO 63166; michael.w.mccracken@stls.frb.org. We gratefully acknowledge helpful discussions with Peter Hansen. The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Banks of Kansas City or St. Louis.

1 Introduction

With improvements in computational power, it has become increasingly straightforward to mine economic and financial datasets for variables that might be useful for forecasting. This point is made clearly in any number of papers, including Denton (1985), Lo and MacKinlay (1990), and Hoover and Perez (1999). In light of this data mining, it is clear that the search itself must be taken into account when trying to validate whether any of the findings are statistically significant. Various methods exist to manage this multiple testing problem, including the traditional use of Bonferroni bounds as well as more recent methods involving q-values (Storey 2002). Each of these methods has strengths and weaknesses in terms of applicability to a given situation.

The bootstrap is one particularly tractable method of managing this multiple testing problem. Specifically, in the context of out-of-sample tests of predictive ability, White (2000) develops a bootstrap method of constructing an asymptotically valid test of the null hypothesis that forecasts from a potentially large number of predictive models are as accurate as those from a baseline model. Since then, some research has extended the applicability of the results in White (2000). Hansen (2005) shows that normalizing and re-centering the test statistic in a specific manner can lead to a more accurately sized and powerful test. Corradi and Swanson (2007) provide important modifications that permit situations in which parameter estimation error enters the asymptotic distribution.

One thing these existing bootstrap "reality check's" have in common is the assumption that the baseline model is non-nested with at least one of the competing models. More precisely, they require that the $(M \times 1)$ vector of out-of-sample averages of the loss differentials (each element of the vector is the average difference in forecast accuracy between the baseline and a distinct competing model) is asymptotically normal with a positive semi-definite long-run covariance matrix. Along with additional moment and mixing conditions, this condition on the long-run covariance matrix is satisfied if at least one of the models is non-nested with the baseline.

However, in many forecast comparisons, the baseline model is a simpler, nested version of all of the competing models. The purpose of such a comparison is to determine whether the additional predictors associated with the larger models improve forecast accuracy relative to the restricted baseline. In general, in asset return applications, economic theories of efficient markets imply that excess returns form a martingale difference and a

null hypothesis under which all predictive models nest the baseline model of zero expected return. In macroeconomic applications, it is standard to examine the predictive content of some variable for output or inflation by comparing a baseline autoregressive model to alternative models that add in lags of other potential predictors. Examples in the literature include Cheung, Chinn, and Garcia's (2005) examination of various exchange rate models, Goyal and Welch's (2003, 2008) studies of the predictive content of a variety of business fundamentals for stock returns, and Stock and Watson's (1999, 2003) analyses of output and inflation forecasting models.

In such evaluations of multiple forecasting models that nest the benchmark model, the results in studies such as White (2000) on the asymptotic and finite-sample properties of bootstrap-based joint tests of equal forecast accuracy, based on non-nested models, may not apply. Intuitively, with nested models, the null hypothesis that the restrictions imposed in the benchmark model are true implies the population errors of all of the competing forecasting models are exactly the same. This in turn implies that the population difference between the competing models' mean square forecast errors is exactly zero, with zero variance. As a result, the distribution of a t–statistic for equal MSE may be non–standard. Indeed, Clark and McCracken (2001, 2005) and McCracken (2007) show that, for pairs of forecasts from nested models, the distributions of tests for equal forecast accuracy are typically not Normally distributed.

Motivated by the frequency with which researchers compare predictions from a baseline nested model with many alternative, nesting models, this paper develops a novel, simple bootstrap that can be used to construct joint tests of equal out-of-sample forecast accuracy or forecast encompassing among such forecasts. This new bootstrap allows for both conditional heteroskedasticity and serial correlation in the forecast errors.

Before proceeding, we should make clear our contribution to the literature on simulation-based methods for multiple testing. Hansen (2005) notes explicitly that his and the results in White (2000) do not apply when the baseline model is nested by all of its competitors. For the case of a small set of nested models, Hubrich and West (2010) propose comparing an adjusted t-test for equal mean square error (equivalently, a t-test for forecast encompassing) against the simulated maximum of a set of standard normal random variables. Their approach is based on the observation of Clark and West (2007) that, while the true asymptotic distributions of the t-tests are not normally distributed under general conditions, the dis-

tributions can often be reasonably approximated by the standard normal. Inoue and Kilian (2004) explicitly derive the asymptotic distribution of two tests for equal population-level out-of-sample predictive ability when the baseline model is nested by many competing models — both of which we consider here. Our paper extends Inoue and Kilian (2004) in three dimensions. First, we provide analytics for two more tests. Second, we extend the results in Inoue and Kilian (2004) to environments that allow for forecast horizons greater than one period and conditionally heteroskedastic errors. Finally, and most importantly, we develop a fixed regressor bootstrap method for obtaining asymptotic critical values under the null of equal accuracy at the population level, and we prove the validity of the bootstrap.

Although our results apply only to a setup that some might see as restrictive — direct, multi–step (DMS) forecasts from nested models — the list of studies analyzing such forecasts suggests our results should be useful to many researchers. Recent applications considering a variety of DMS forecasts from nested linear models include, among others: the studies cited at the beginning of this section; Hong and Lee (2003); Hubrich (2005); Mark (1995); Kilian (1999); Butler, Grullon and Weston (2005); Sarno, et al. (2005); Cooper and Gulen (2006); Guo (2006); Rapach and Wohar (2006); Bruneau, et al. (2007); Rapach and Strauss (2007); Moench (2008); Billmeier (2009); Chen, et al. (2009); Hendry and Hubrich (2009); and Molodtsova and Papell (2009).

The remainder proceeds as follows. After introducing essential notation and the forecast-based tests considered, Section 2 presents the tests considered and our proposed bootstrap method. Section 3 derives the asymptotic distributions of the tests and proves the validity of the bootstrap. Section 4 presents Monte Carlo results on the finite–sample performance of our proposed bootstrap compared to the methods of White (2000), Hansen (2005), and Hubrich and West (2010). Section 5 applies our tests to forecasts of commodity prices and GDP growth in the U.S. Section 6 concludes.

2 A bootstrap for nested model reality checks

After describing essential notation and the forecast test statistics considered, this section presents our proposed bootstrap algorithm.

2.1 Essential notation

At each forecast origin $t = T, ..., T + P - \tau$, we observe the sequence $\{y_s, x'_s\}_{s=1}^{t}$. τ-step ahead forecasts of the scalar $y_{t+\tau}$, $\tau \geq 1$, are generated using a $(k \times 1, k = k_0 + k_{\bar{M}})$ vector of covariates $x_t = (x'_{0,t}, x'_{\bar{M},t})'$ that consists of a baseline set of predictors $x_{0,t}$ that occur in each model as well as a vector of additional covariates $x_{\bar{M},t}$. The set of predictors x_t may include lags of y_t. Sub-vectors of $x_{\bar{M},t}$ differentiate the competing unrestricted models. There therefore exists $2^{k_{\bar{M}}} - 1$ unique unrestricted linear regression models that nest the baseline model within it. Let $j = 1, ..., M$ denote an index of the collection of $M \leq 2^{k_{\bar{M}}} - 1$ unique models $x'_{j,t}\beta_j$ to be compared to the baseline nested model $x'_{0,t}\beta_0$.[1]

At every forecast origin, each of the forecasting models is estimated by OLS, yielding coefficients $\hat{\beta}_{j,t}$. The τ-step ahead forecast errors for model j are $\hat{u}_{j,t+\tau} = (y_{t+\tau} - x'_{j,t}\hat{\beta}_{j,t})$, $j = 0, 1, ..., M$. Because the models are nested, the null hypothesis that the additional predictors do not provide predictive content implies the population forecast errors $u_{j,t+\tau} \equiv (y_{t+\tau} - x'_{j,t}\beta_j)$ satisfy $u_{j,t+\tau} = u_{0,t+\tau} \equiv u_{t+\tau}$ for all $j = 1, ..., M$.

2.2 Test statistics

We consider a total of four forecast–based tests: two tests of equal forecast accuracy and two tests for forecast encompassing. In particular, we consider multiple-model variants of the t–statistic for equal MSE developed by Diebold and Mariano (1995) and West (1996) and the F–statistic proposed by McCracken (2007). We also consider multiple-model variants of the t–statistic for encompassing developed in Harvey, Leybourne, and Newbold (1998) and West (2001) and the variant proposed by Clark and McCracken (2001). Application of the tests to multiple models involves taking the maximum of test statistics formed for each alternative model forecast to the benchmark model forecast.

The two tests of equal MSE are based upon the sequence of vectors of loss differentials $(\hat{d}_{1,t+\tau}, ..., \hat{d}_{M,t+\tau})'$, where $\hat{d}_{j,t+\tau} = \hat{u}^2_{0,t+\tau} - \hat{u}^2_{j,t+\tau}$. If we define $MSE_j = (P - \tau + 1)^{-1} \sum_{t=T}^{T+P-\tau} \hat{u}^2_{j,t+\tau}$ $(j = 0, 1, ..., M)$, $\bar{d}_j = (P - \tau + 1)^{-1} \sum_{t=T}^{T+P-\tau} \hat{d}_{j,t+\tau} = MSE_0 - MSE_j$, $\hat{\gamma}_{d_j d_j}(l) = (P - \tau + 1)^{-1} \sum_{t=T+l}^{T+P-\tau} (\hat{d}_{j,t+\tau} - \bar{d}_j)(\hat{d}_{j,t+\tau-l} - \bar{d}_j)$, $\hat{\gamma}_{d_j d_j}(-l) = \hat{\gamma}_{d_j d_j}(l)$, and (for a kernel $K(\cdot)$ and truncation parameter L defined later) $\hat{S}_{d_j d_j} = \sum_{l=-\bar{l}}^{\bar{l}} K(l/L)\hat{\gamma}_{d_j d_j}(l)$, the statistics take the

[1] Note that while we do not require that $M = 2^{k_{\bar{M}}} - 1$, we do require that all $k_{\bar{M}}$ of the $x's$ are used in at least one of the unrestricted models.

form

$$\max_{j=1,...,M}(\text{MSE-}t_j) = \max_{j=1,...,M}((P-\tau+1)^{1/2} \times \frac{\bar{d}_j}{\sqrt{\hat{S}_{d_j d_j}}}). \quad (1)$$

$$\max_{j=1,...,M}(\text{MSE-}F_j) = \max_{j=1,...,M}((P-\tau+1) \times \frac{\bar{d}_j}{MSE_j}). \quad (2)$$

Similarly, the two tests of forecast encompassing are based upon the sequence of vectors $(\hat{c}_{1,t+\tau}, ..., \hat{c}_{M,t+\tau})'$, where $\hat{c}_{j,t+\tau} = \hat{u}_{0,t+\tau}(\hat{u}_{0,t+\tau} - \hat{u}_{j,t+\tau})$. If we define $\bar{c}_j = (P-\tau+1)^{-1}\sum_{t=T}^{T+P-\tau}\hat{c}_{j,t+\tau}$, $\hat{\gamma}_{c_j c_j}(l) = (P-\tau+1)^{-1}\sum_{t=T+l}^{T+P-\tau}(\hat{c}_{j,t+\tau} - \bar{c}_j)(\hat{c}_{j,t+\tau-l} - \bar{c}_j)$, $\hat{\gamma}_{c_j c_j}(-l) = \hat{\gamma}_{c_j c_j}(l)$, and $\hat{S}_{c_j c_j} = \sum_{l=-\bar{l}}^{\bar{l}} K(l/L)\hat{\gamma}_{c_j c_j}(l)$, the statistics take the form

$$\max_{j=1,...,M}(\text{ENC-}t_j) = \max_{j=1,...,M}((P-\tau+1)^{1/2} \times \frac{\bar{c}_j}{\sqrt{\hat{S}_{c_j c_j}}}). \quad (3)$$

$$\max_{j=1,...,M}(\text{ENC-}F_j) = \max_{j=1,...,M}((P-\tau+1) \times \frac{\bar{c}_j}{MSE_j}). \quad (4)$$

2.3 The bootstrap

Our new, bootstrap-based method of approximating the asymptotically valid critical values for multiple-model comparisons between nested models is an extension of that discussed in Clark and McCracken (2009) in the context of pairwise comparisons. Specifically, we use a variant of the wild fixed regressor bootstrap developed in Goncalves and Kilian (2004) but adapted to the direct multi-step nature of the forecasts. In this framework the x's are held fixed across the artificial samples and the dependent variable is generated using the direct multi-step equation $y^*_{s+\tau} = x'_{0,s}\hat{\beta}_{0,T} + \hat{v}^*_{s+\tau}$, $s = 1, ..., T+P-\tau$, for a suitably chosen artificial error term $\hat{v}^*_{s+\tau}$ designed to capture both the presence of conditional heteroskedasticity and an assumed $MA(\tau-1)$ serial correlation structure in the τ-step ahead forecasts. Specifically, we construct the artificial samples and bootstrap critical values using the following algorithm.[2]

1. Estimate the parameter vector $\beta_{\bar{M}}$ associated with the unrestricted model that includes all $k_{\bar{M}}$ predictors using OLS and store the residuals $\hat{v}_{s+\tau} = y_{s+\tau} - x'_s\hat{\beta}_{\bar{M}}$, $s = 1, ..., T+P-\tau$.

2. Using NLLS, estimate an $MA(\tau-1)$ model for the OLS residuals $\hat{v}_{s+\tau}$ such that $v_{s+\tau} = \varepsilon_{s+\tau} + \theta_1\varepsilon_{s+\tau-1} + ... + \theta_{\tau-1}\varepsilon_{s+1}$. Let $\eta_{s+\tau}$, $s = 1, ..., T+P-\tau$, denote an $i.i.d\ N(0,1)$

[2] Our approach to generating artificial samples of multi-step forecast errors builds on a sampling approach proposed in Hansen (1996).

sequence of simulated random variables. Define $\hat{v}^*_{s+\tau} = (\eta_{s+\tau}\hat{\varepsilon}_{s+\tau} + \hat{\theta}_1\eta_{s-1+\tau}\hat{\varepsilon}_{s+\tau-1} + ... + \hat{\theta}_{\tau-1}\eta_{s+1}\hat{\varepsilon}_{s+1})$, $s = 1, ..., T+P-\tau$.

3. Estimate the parameter vector β_0 associated with the restricted model using OLS and store the fitted values $x'_{0,s}\hat{\beta}_{0,T}$ $s = 1, ..., T+P-\tau$. Form artificial samples of $y^*_{s+\tau}$ using the fixed regressor structure, $y^*_{s+\tau} = x'_{0,s}\hat{\beta}_{0,T} + \hat{v}^*_{s+\tau}$.

4. Using the artificial data, construct an estimate of the test statistics (e.g., $\underset{j=1,...,M}{max}$(MSE-$F_j$)) as if this were the original data.

5. Repeat steps $2-4$ (note, however, that the NLLS estimation of an MA model is not repeated) a large number of times: $j = 1, ..., N$.

6. Reject the null hypothesis, at the α% level, if the test statistic is greater than the $(100-\alpha)$%-ile of the empirical distribution of the simulated test statistics.

3 Theoretical results on distributions and bootstrap validity

This section proves the validity of the bootstrap proposed above, after providing further detail on the econometric environment and deriving the asymptotic distributions of the tests presented in section 2.2. The proofs are provided in Appendix 1.

3.1 Environment details

We denote the loss associated with the τ-step ahead forecast errors as $\hat{u}^2_{j,t+\tau} = (y_{t+\tau} - x'_{j,t}\hat{\beta}_{j,t})^2$, $j = 0, 1, ..., M$. As indicated above, under the null, the population forecast errors $u_{j,t+\tau} \equiv (y_{t+\tau} - x'_{j,t}\beta_j)$ satisfy $u_{j,t+\tau} = u_{0,t+\tau} \equiv u_{t+\tau}$ for all $j = 1, ..., M$.

The following additional notation will be used. Define the selection matrices J_j satisfying $x_{j,s} = J'_j x_s$ for all $j = 0, 1, ..., M$. Let $H(t) = (t^{-1}\sum_{s=1}^{t-\tau} x_s u_{s+\tau}) = (t^{-1}\sum_{s=1}^{t-\tau} h_{s+\tau})$, $B_j(t) = (t^{-1}\sum_{s=1}^{t-\tau} x_{j,s}x'_{j,s})^{-1}$, $B_j = (Ex_{j,s}x'_{j,s})^{-1}$, $B(t) = (t^{-1}\sum_{s=1}^{t-\tau} x_s x'_s)^{-1}$, and $B = (Ex_s x'_s)^{-1}$. For $H(t)$ and J_j defined above, $\sigma^2 = Eu^2_{t+\tau}$, and a $((k_j - k_0)\times k, k_j = \dim(x_{j,s}))$ matrix $\tilde{A}_j$ satisfying $\tilde{A}'_j\tilde{A}_j = B^{-1/2}(-J_0B_0J'_0 + J_jB_jJ'_j)B^{-1/2}$, let $\tilde{h}_{j,t+\tau} = \sigma^{-1}\tilde{A}_jB^{1/2}h_{t+\tau}$ and $\tilde{H}_j(t) = \sigma^{-1}\tilde{A}_jB^{1/2}H(t)$. If we define $\gamma_{\tilde{h}\tilde{h},j}(i) = E\tilde{h}_{j,t+\tau}\tilde{h}'_{j,t+\tau-i}$, then $S_{\tilde{h}\tilde{h},j} = \gamma_{\tilde{h}\tilde{h},j}(0) + \sum_{i=1}^{\tau-1}(\gamma_{\tilde{h}\tilde{h},j}(i) + \gamma'_{\tilde{h}\tilde{h},j}(i))$. Let $W(s)$ denote a $k \times 1$ vector standard Brownian motion.

To derive the asymptotic distributions of the tests, we need four assumptions.

Assumption 1: The vector of coefficients β_j in each forecasting model j are estimated recursively by OLS: $\hat{\beta}_{j,t} = \arg\min_{\beta_j} t^{-1}\sum_{s=1}^{t-\tau}(y_{t+\tau} - x'_{j,t}\beta_j)^2$, $j = 0, 1, ..., M$.

Assumption 2: (a) $U_{t+\tau} = [h'_{t+\tau}, vec(x_t x'_t - Ex_t x'_t)']'$ is covariance stationary. (b) $EU_{t+\tau} = 0$. (c) For all $l > \tau - 1$, $Eh_{t+\tau}h'_{t+\tau-l} = 0$. (d) $Ex_t x'_t < \infty$ and is positive definite. (e) For some $r > 8$, $U_{t+\tau}$ is uniformly L^r bounded. (f) For some $r > d > 2$, $U_{t+\tau}$ is strong mixing with coefficients of size $-rd/(r-d)$. (g) $lim_{T\to\infty} T^{-1} E(\sum_{s=1}^{T-\tau} U_{s+\tau})(\sum_{s=1}^{T-\tau} U_{s+\tau})' = \Omega < \infty$ is positive definite.

Assumption 3: (a) Let $K(x)$ be a continuous kernel such that for all real scalars x, $|K(x)| \leq 1$, $K(x) = K(-x)$ and $K(0) = 1$. (b) For some bandwidth L and constant $i \in (0, 0.5)$, $L = O(P^i)$. (c) The number of covariance terms $\bar{l}$, used to estimate the long–run covariances $S_{c_j c_j}$ and $S_{d_j d_j}$ defined in Section 2.2, satisfies $\tau - 1 \leq \bar{l} < \infty$.

Assumption 4: $\lim_{P,T\to\infty} P/T = \lambda_P \in (0, \infty)$.

The assumptions provided here are very similar to those provided in Clark and McCracken (2005). We restrict attention to forecasts generated using parameters estimated recursively by OLS (Assumption 1)[3] and we do not allow for processes with either unit roots or time trends (Assumption 2).[4] We provide asymptotic results for situations in which the in-sample and out-of-sample sizes T and P are of the same order (Assumption 4).

Assumption 3 is necessitated by the serial correlation in the multi-step (τ-step) forecast errors — errors from even well-specified models exhibit serial correlation, of an $MA(\tau - 1)$ form. Typically, researchers constructing a t-statistic utilizing the squares of these errors account for serial correlation of at least order $\tau - 1$ in forming the necessary standard error estimates. Meese and Rogoff (1988), Groen (1999), and Kilian and Taylor (2003), among other applications to forecasts from nested models, use kernel-based methods to estimate the relevant long-run covariance.[5] We therefore impose conditions sufficient to cover applied practices. Parts (a) and (b) are not particularly controversial. Part (c), however, imposes

[3] Assumption 1 could easily be weakened to allow for parameters estimated using a rolling window of T observations at each forecast origin t. Under the rolling scheme, after redefining the Γ_i, $i = 1 - 3$ appropriately, Theorems 3.1 and 3.2 continue to hold. More importantly, Theorems 3.3 and 3.4 also continue to hold and hence our bootstrap is valid under both the rolling and recursive schemes.

[4] Our assumptions do, however, allow y_t and x_t to be stationary differences of trending variables. As to other technical aspects of Assumption 2, (a) and (d) together ensure that in large samples, sample averages of the outer product of the predictors will be invertible and hence the least squares estimate will be well defined. Part (c) enables the use of Markov inequalities when showing certain terms are asymptotically negligible. Along with (c), (e) and (f) allow us to use results in Hansen (1992) and Davidson (1994) regarding the weak convergence of partial sums to Brownian motion and that functionals of these partial sums converge in distribution to stochastic integrals.

[5] For similar uses of kernel–based methods in analyses of non–nested forecasts, see, for example, Diebold and Mariano (1995) and West (1996).

the restriction that the known $MA(\tau-1)$ structure of the model errors (from Assumption 2c) is taken into account when constructing the MSE-t and ENC-t statistics discussed in Section 2. Although Assumption 3 and our theoretical results admit a range of kernel and bandwidth approaches, in our Monte Carlo experiments and empirical application we compute the variances required by the MSE-t and ENC-t statistics (for $\tau > 1$) using the Newey and West (1987) estimator with a lag length of $1.5 * \tau$.

3.2 Asymptotic Distributions

Under the null that $x_{\bar{M},t}$ has no population level predictive power for $y_{t+\tau}$, for all $j = 1, ..., M$ the population difference in MSEs, $E(u^2_{0,t+\tau} - u^2_{j,t+\tau})$, and the moment condition $Eu_{0,t+\tau}(u_{0,t+\tau} - u_{j,t+\tau})$ will equal 0. Clark and McCracken (2005) show that for a given j, MSE-t_j, MSE-F_j, ENC-t_j and ENC-F_j are bounded in probability. Since the max operator is continuous and M is finite, we therefore expect the maxima of each of the four statistics to also be bounded in probability. In contrast, when $x_{\bar{M},t}$ has predictive power, the population difference in MSEs, $E(u^2_{0,t+\tau} - u^2_{j,t+\tau})$, and the moment condition $Eu_{0,t+\tau}(u_{0,t+\tau} - u_{j,t+\tau})$ will be positive for at least one j. In the case where j is known, Clark and McCracken (2005) show that MSE-t_j, MSE-F_j, ENC-t_j and ENC-F_j diverge to positive infinity and hence each test is consistent. But again, in the present environment with multiple models, we do not know which model j has predictive content. However, since the max operator is continuous and M is finite, we expect the maxima of each of the four statistics to also be consistent regardless of whether we know which model j has predictive content.

For a given j, Clark and McCracken (2005) and McCracken (2007) show that, for τ–step ahead forecasts, the MSE-F_j, MSE-t_j, ENC-F_j, and ENC-t_j statistics converge in distribution to functions of stochastic integrals of quadratics of Brownian motion, with limiting distributions that depend on the sample split parameter π, the number of exclusion restrictions $k_j - k_0$, and the unknown nuisance parameter $S_{\tilde{h}_j\tilde{h}_j}$. This continues to hold in the presence of multiple-model comparisons. If we define $\Gamma_{1,j} = \int_1^{1+\lambda_P} s^{-1}W'(s)J_jS_{\tilde{h}_j\tilde{h}_j}J_j'dW(s)$, $\Gamma_{2,j} = \int_1^{1+\lambda_P} s^{-2}W'(s)J_jS_{\tilde{h}_j\tilde{h}_j}J_j'W(s)ds$, and $\Gamma_{3,j} = \int_1^{1+\lambda_P} s^{-2}W'(s)J_jS^2_{\tilde{h}_j\tilde{h}_j}J_j'W(s)ds$, the following two theorems provide the asymptotic distributions of the multiple model variants of these four statistics.

Theorem 3.1: Maintain Assumptions 1, 2, and 4. Under the null hypothesis, it follows that: (a) $\underset{j=1,...,M}{max}(\text{MSE-}F_j) \rightarrow_d \underset{j=1,...,M}{max}(2\Gamma_{1,j} - \Gamma_{2,j})$, and (b) $\underset{j=1,...,M}{max}(\text{ENC-}F_j) \rightarrow_d \underset{j=1,...,M}{max}(\Gamma_{1,j})$.

Theorem 3.2: Maintain Assumptions $1-4$. Under the null hypothesis, it follows that: (a) $\underset{j=1,...,M}{max}(\text{MSE-}t_j) \rightarrow_d \underset{j=1,...,M}{max}((\Gamma_{1,j}-.5\Gamma_{2,j})/\Gamma_{3,j}^{1/2})$, and (b) $\underset{j=1,...,M}{max}(\text{ENC-}t_j) \rightarrow_d \underset{j=1,...,M}{max}(\Gamma_{1,j}/\Gamma_{3,j}^{1/2})$.

Theorem 3.1 extends the results in Inoue and Kilian (2004) to environments in which the model errors are allowed to be not only conditionally heteroskedastic but serially correlated of an $MA(\tau-1)$ form. Theorem 3.2 is new and extends the asymptotics for the MSE-t and ENC-t statistics of Clark and McCracken (2005) to an environment where there are many nested comparisons. The results in both theorems exploit some asymptotic theory derived in Clark and McCracken (2005), the rank condition on Ω, and a simple application of the Continuous Mapping Theorem applied to the max functional.

Unfortunately, neither of these theorems provides us with a closed form for the asymptotic distribution of the test statistic and hence tabulating critical values is infeasible. Simulating critical values may be feasible but will be context-specific due to having to estimate the unknown nuisance parameters $S_{\tilde{h}_j\tilde{h}_j}$. Our simple and novel bootstrap method for computing asymptotically valid critical values overcomes this problem.

Proving the asymptotic validity of this bootstrap requires a modest strengthening of the moment conditions on the model residuals.

Assumption $2'$: (a) $U_{t+\tau} = [h'_{t+\tau}, vec(x_t x'_t - Ex_t x'_t)']'$ is covariance stationary. (b) $\text{E}(\varepsilon_{s+\tau}|\varepsilon_{s+\tau-j}, x_{s-j}$ $j \geq 0) = 0$. (c) Let $\gamma = (\beta'_{\bar{M}}, \theta_1, ..., \theta_{\tau-1})'$, $\hat{\gamma} = (\hat{\beta}'_{\bar{M}}, \hat{\theta}_1, ..., \hat{\theta}_{\tau-1})'$, and define the function $\hat{\varepsilon}_{s+\tau} = \hat{\varepsilon}_{s+\tau}(\hat{\gamma})$ such that $\hat{\varepsilon}_{s+\tau}(\gamma) = \varepsilon_{s+\tau}$. In an open neighborhood N around γ, there exists $r > 8$ such that $\sup_{1\leq s\leq T} ||\sup_{\gamma\in N}(\hat{\varepsilon}_{s+\tau}(\gamma), \nabla\hat{\varepsilon}'_{s+\tau}(\gamma), x_s)'||_r \leq c$. (d) $Ex_t x'_t < \infty$ and is positive definite. (e) For some $r > d > 2$, $U_{t+\tau}$ is strong mixing with coefficients of size $-rd/(r-d)$, (f) $lim_{T\rightarrow\infty} T^{-1}E(\sum_{s=1}^{T-\tau} U_{s+\tau})(\sum_{s=1}^{T-\tau} U_{s+\tau})' = \Omega < \infty$ is positive definite.

Assumption $2'$ differs from Assumption 2 in two ways. First, in (b) it emphasizes the point that the forecast errors, and by implication $h_{t+\tau}$, form an $MA(\tau-1)$ process. Second, in (c) it bounds the second moments not only of $h_{t+\tau} = (\varepsilon_{s+\tau} + \theta_1\varepsilon_{s+\tau-1} + ... + \theta_{\tau-1}\varepsilon_{s+1})x_s$ but also the functions $\hat{\varepsilon}_{s+\tau}(\gamma)x_s$, and $\nabla\hat{\varepsilon}_{s+\tau}(\gamma)x_s$ for all γ in an open neighborhood of γ. These assumptions are primarily used to show that the bootstrap-based artificial samples, which are a function of the estimated errors $\hat{\varepsilon}_{s+\tau}$, adequately replicate the time series properties of the original data in large samples. Specifically, we must insure that the bootstrap analog of $h_{s+\tau}$ is not only zero mean but has the same long-run variance V.

Such an assumption is not needed for our earlier results since the model forecast errors $\hat{u}_{j,s+\tau}$, are linear functions of $\hat{\beta}_j$ and Assumption 2 already imposes moment conditions on $\hat{u}_{j,s+\tau}$ via moment conditions on $h_{s+\tau}$.

In the following let MSE-F_j^*, ENC-F_j^*, MSE-t_j^*, and ENC-t_j^* denote statistics generated using the artificial samples from our bootstrap and let $=^{d^*}$ and $\rightarrow^{d^*}$ denote equality and convergence in distribution with respect to the bootstrap-induced probability measure P^*. Similarly, let $\Gamma_{i,j}^*$, $i = 1 - 3$, denote random variables generated using the artificial samples satisfying $\Gamma_{i,j}^* =^{d^*} \Gamma_{i,j}$ for $\Gamma_{i,j}$ defined in Theorems 3.1 and 3.2.

Theorem 3.3: Maintain Assumptions $1, 2'$, and 4. (a) $\underset{j=1,...,M}{max}(\text{MSE-}F_j^*) \rightarrow^{d^*} \underset{j=1,...,M}{max}(2\Gamma_{1,j}^* - \Gamma_{2,j}^*)$. (b) $\underset{j=1,...,M}{max}(\text{ENC-}F_j^*) \rightarrow^{d^*} \underset{j=1,...,M}{max}(\Gamma_{1,j}^*)$.

Theorem 3.4: Maintain Assumptions $1, 2', 3$, and 4. (a) $\underset{j=1,...,M}{max}(\text{MSE-}t_j^*) \rightarrow^{d^*} \underset{j=1,...,M}{max}((\Gamma_{1,j}^* - .5\Gamma_{2,j}^*)/\Gamma_{3,j}^{*1/2})$. (b) $\underset{j=1,...,M}{max}(\text{ENC-}t_j^*) \rightarrow^{d^*} \underset{j=1,...,M}{max}(\Gamma_{1,j}^*/\Gamma_{3,j}^{*1/2})$.

In Theorems 3.3 and 3.4 we show that our fixed-regressor bootstrap provides an asymptotically valid method of estimating the critical values associated with the null of equal (population-level) forecast accuracy among many nested models. To understand this, note that nowhere in either Theorem do we make an assumption about whether the null or alternative hypothesis is true. Hence regardless of whether or not the null hypothesis holds, the bootstrapped critical values are consistent for the appropriate percentiles of the asymptotic distributions in Theorems 3.1 and 3.2 associated with the specific statistic being used for inference. This result follows directly from how we generate the artificial data using our bootstrap. First, the null is imposed by modeling the conditional mean component of $y_{t+\tau}$ as the restricted model $x_{0,t}'\beta_0$ and hence none of the other predictors $x_{\bar{M},t}$ exhibit any predictive content. Second, we insure that all of the predictors are orthogonal to the pseudo-residuals (and exhibit the appropriate degree of serial correlation and conditional heteroskedasticity) by using a wild form of the bootstrap based on those residuals estimated using all of the predictors and not just those in the restricted model.

As we will show in the next section, our fixed regressor bootstrap provides reasonably sized tests in our Monte Carlo simulations, outperforming other bootstrap-based methods for estimating the asymptotically valid critical values necessary to test the null of equal predictive ability among many nested models.

4 Monte Carlo evidence

We use simulations of multivariate DGPs with some of the features of common finance and macroeconomic applications to evaluate the finite sample properties of the above approaches to testing for equal forecast accuracy. In these simulations, the benchmark forecasting model is a univariate model of the predictand y; the alternative models add combinations of other variables of interest. The null hypothesis is that, in population, the forecasts are all equally accurate. The alternative hypothesis is that at least one of the other forecasts is more accurate than the benchmark. All of the forecasts are generated under the recursive scheme.

We proceed by summarizing the test statistics and inference approaches considered, detailing the data generating processes (DGPs), and presenting the size and power of the tests, for a nominal size of 10% (results for 5% are qualitatively the same).

4.1 Tests and inference approaches

We evaluate the size and power properties of the MSE-F, MSE-t, ENC-F, and ENC-t test statistics given in section 2.2. More specifically, we compare these test statistics to critical values obtained from the fixed regressor bootstrap detailed in section 2.3. Under our asymptotics, the bootstrapped critical values are asymptotically valid.

For comparison to our suggested approach, we include in the analysis size and power results for test statistics compared against alternative sources of critical values. Specifically, we compare the MSE-F and MSE-t tests to critical values obtained from a non-parametric bootstrap patterned on White's (2000) method: we create bootstrap samples of forecast errors by sampling (with replacement) from the time series of sample forecast errors, and construct test statistics for each sample draw. However, as noted above and in White (2000), this procedure is not, in general, asymptotically valid when applied to nested models. While not established in any existing literature, this non-parametric approach may be valid under the alternative asymptotics of Giacomini and White (2006), which treat the estimation sample size T as finite and the forecast sample size P as limiting to ∞. We include the method in part for its computational simplicity and in part to examine the potential pitfalls of using the approach.

In our non-parametric implementation, we follow White (2000) in using the stationary bootstrap of Politis and Romano (1994) and centering the bootstrap distributions around

the sample values of the test statistics (specifically, the numerators of the t-statistics). The stationary bootstrap is parameterized to make the average block length equal to twice the forecast horizon. As to centering of test statistics, under the non–parametric approach, the relevant null hypothesis is that the MSE difference (benchmark MSE less alternative model MSE) is at most 0. Following White (2000), each bootstrap draw of a given test statistic is re-centered around the corresponding sample test statistic. Bootstrapped critical values are computed as percentiles of the resulting distributions of re–centered test statistics.

We also include Hansen's (2005) SPA test (the centered version, SPA_c), which modifies the reality check of White (2000) to improve power when the model set includes some that forecast poorly. The SPA test is compared against the bootstrap approach of Hansen (2005), based on the non-parametric sampling we use for the MSE-t test.[6]

For the DGPs involving small numbers of models (DGPs 2 and 3), we also provide results for the testing approach of Hubrich and West (2010). They suggest comparing an adjusted test for equal MSE — which is the same as the ENC-t test — against critical values obtained from Monte Carlo simulations of the distribution of the maximum of a set of correlated normal random variables. The suggestion is based on the observation in Clark and West (2007) that, while the ENC-t test applied to forecasts from nested models has a non-standard asymptotic distribution under large T, P asymptotics, critical values from that distribution are often quite similar to standard normal critical values. Hubrich and West emphasize that their approach has the advantage of requiring Monte Carlo simulations of a normal distribution that may be simpler than bootstrap simulations.

4.2 Monte Carlo design

For all DGPs, we generate data using independent draws of innovations from the normal distribution and the autoregressive structure of the DGP. The initial observations necessitated by the lag structure of each DGP are generated with draws from the unconditional normal distribution implied by the DGP. We consider forecast horizons (τ) of one and four steps. With quarterly data in mind, we also consider a range of sample sizes (T, P), reflecting those commonly available in practice: 40,80; 40,120; 80,40; 80,80; 80,120; 120,40; and 120,80. In all cases, our reported results are based on 2000 Monte Carlo draws and 499 bootstrap replications.

[6] For consistency with the rest of our analysis, for multi-step forecasts we compute the HAC variance that enters the SPA test statistic with the Newey and West (1987) estimator. In re-running a subset of our experiments with the HAC estimator used by Hansen (2005), we obtained essentially the same results.

4.2.1 DGP 1

DGP 1 is based loosely on the empirical properties of the change in core U.S. inflation ($y_{t+\tau}$, where τ denotes the forecast horizon) and various economic indicators, some persistent and some not. With this DGP, we examine the properties of tests applied to a large number of forecasts — 128 — based on combinations of predictors.

In DGP 1, the true model for $y_{t+\tau}$ includes y_t and up to one other predictor, $x_{1,t}$:

$$\begin{aligned} y_{t+\tau} &= -0.3y_t + bx_{1,t} + u_{t+\tau} \\ u_{t+\tau} &= \varepsilon_{t+\tau} \text{ if } \tau = 1, \quad u_{t+\tau} = \varepsilon_{t+\tau} + 0.95\varepsilon_{t+\tau-1} + 0.9\varepsilon_{t+\tau-2} + 0.8\varepsilon_{t+\tau-3} \text{ if } \tau = 4 \\ x_{i,t} &= \gamma_i x_{i,t-1} + v_{i,t}, \quad \gamma_i = 0.8 - 0.1(i-1), \quad i = 1, \ldots, 7 \\ & \quad E(\varepsilon_t v_{i,t}) = 0 \ \forall \ i, \quad E(v_{i,t}v_{j,t}) = 0 \ \forall \ i \neq j \\ & \quad var(\varepsilon_{t+\tau}) = 2.0, \quad var(v_{i,t}) = (1 - \gamma_i^2), \end{aligned} \tag{5}$$

where the coefficient b varies across experiments, as indicated below. Note that, for a forecast horizon of 4 steps, the error term $u_{t+\tau}$ in the equation for $y_{t+\tau}$ follows an MA(3) process.

In DGP 1 experiments, forecasts are generated from models that include all possible combinations of $x_{1,t}$ and six other ($i = 2, \ldots, 7$) $x_{i,t}$ variables. The null forecasting model is

$$y_{t+\tau} = \beta_0 + \beta_1 y_t + u_{0,t+\tau}. \tag{6}$$

We consider a total of 127 alternative models, each of which includes a constant, y_t, and a different combination of the $x_{i,t}$, $i = 1, \ldots, 7$, variables:

$$y_{t+\tau} = \beta_0 + \beta_1 y_t + \tilde{\beta}_j' \tilde{x}_{\bar{M},j,t} + u_{j,t+\tau}, \ j = 1, \ldots, 127, \tag{7}$$

where $\tilde{x}_{\bar{M},j,t}$ contains a unique combination of the $x_{i,t}$, $i = 1, \ldots, 7$, variables.

To evaluate the size properties of tests of equal (population-level) forecast accuracy, the coefficient b is set to 0, such that the benchmark model is the true one and, in population, forecasts from all of the models are equally accurate. To evaluate power, we set $b = 0.4$ in experiments at the 1-step ahead horizon and $b = 0.8$ in experiments at the 4-step ahead horizon.[7] In these power experiments, in population the best forecast model is the one that includes y_t and $x_{1,t}$; other models that include these and other variables will be just as accurate, in population.

[7] We obtained qualitatively similar results in power experiments in which the DGP for $y_{t+\tau}$ included multiple x variables (specifically, $x_{1,t}$, $x_{2,t}$, and $x_{3,t}$).

4.2.2 DGP 2

DGP 2 is based on the empirical properties of the quarterly stock return (corresponding to our predictand $y_{t+\tau}$) and predictor ($x_{i,t}$) data of Goyal and Welch (2008). With this DGP, we examine the properties of tests applied to a modest number of forecasts — 17 — obtained by adding a different single indicator to each alternative model.

In DGP 2, the true model for $y_{t+\tau}$ includes a constant and up to one other predictor, $x_{1,t}$:

$$\begin{aligned} y_{t+\tau} &= 1.0 + bx_{1,t} + u_{t+\tau} \\ u_{t+\tau} &= \varepsilon_{t+\tau} \text{ if } \tau = 1, \quad u_{t+\tau} = \varepsilon_{t+\tau} + 0.95\varepsilon_{t+\tau-1} + 0.9\varepsilon_{t+\tau-2} + 0.8\varepsilon_{t+\tau-3} \text{ if } \tau = 4 \\ & \quad var(\varepsilon_{t+\tau}) = 2.0 \\ x_{i,t} &= \gamma_i x_{i,t-1} + v_{i,t}, \quad i = 1, \ldots, 16, \end{aligned} \tag{8}$$

where the coefficient b varies across experiments, as indicated below, and the coefficients γ_i, $i = 1, \ldots, 16$, and the variance-covariance matrix of the error terms are set on the basis of estimates obtained from the Goyal-Welch data. The AR(1) coefficients on the x variables range from 0.99 to -0.13, with most above 0.95. At the 1-step horizon, the correlations between u_t and $v_{i,t}$ range from 0.04 to -0.8 (across i). Again, for a forecast horizon of 4 steps, the error term $u_{t+\tau}$ in the equation for $y_{t+\tau}$ follows an MA(3) process.

In DGP 2 experiments, forecasts are generated from a total of 17 models, similar in form to those used in such studies as Goyal and Welch (2008). The null forecasting model is

$$y_{t+\tau} = \beta_0 + u_{0,t+\tau}. \tag{9}$$

The 16 alternative models each include a single $x_{i,t}$:

$$y_{t+\tau} = \beta_0 + \beta_i x_{i,t} + u_{i,t+\tau}, \ i = 1, \ldots, 16. \tag{10}$$

To evaluate the size properties of tests of equal (population-level) forecast accuracy, the coefficient b is set to 0, such that the benchmark model is the true one and, in population, forecasts from all of the models are equally accurate. To evaluate power, we set $b = 0.8$ in experiments at the 1-step ahead horizon and $b = 2.0$ in experiments at the 4-step ahead horizon. In these power experiments, in population the best forecast model is the one that includes y_t and $x_{1,t}$.

4.2.3 DGP 3

Finally, we also report size and power results for **DGP 3**, which is the same as DGP 2 except that all of the covariances among the error terms u_t and $v_{i,t}$ are set to 0. Our rationale is that, as highlighted in Stambaugh (1999), the combination of (1) a significant correlation between u_t and $v_{i,t}$ and (2) high persistence in $x_{i,t}$ can lead to significant bias in the regression estimate of the coefficient on $x_{i,t}$. The presence of such a bias may adversely affect the properties of the forecast tests. By setting to 0 the covariances between u_t and each $v_{i,t}$ in DGP 3, we eliminate such potential distortions.

4.3 Size results

The results in Tables 1 and 2 indicate that tests of equal MSE and forecast encompassing based on our proposed fixed regressor bootstrap have good size properties in a range of settings. For example, with DGPs 1 and 3, the empirical size of the MSE-t test based on our bootstrap critical values averages 10.8 percent (range of 9.1 to 12.3 percent) at the 1-step horizon and averages 11.3 percent (range of 9.6 to 13.4 percent) at the 4-step horizon. In the same experiments, the empirical size of the ENC-t test compared against fixed regressor bootstrap critical values averages 9.8 percent (range of 8.9 to 10.7 percent) at the 1-step horizon and averages 10.4 percent (range of 8.9 to 12.3 percent) at the 4-step horizon. In most, although not all, cases, the tests of forecast encompassing have slightly lower size than tests of equal MSE. In broad terms, the F- and t-type tests have comparable size.

With DGP 2, the tests compared against critical values from our proposed bootstrap are prone to modest over-sizing. For example, in these experiments, the size of the MSE-t test averages 15.0 percent (range of 14.0 to 16.7 percent) at the 1-step horizon and averages 17.0 percent (range of 14.7 to 18.8 percent) at the 4-step horizon. In these DGP 2 experiments, the rejection rate for the ENC-t test averages 13.1 percent (range of 11.6 to 14.0 percent) at the 1-step horizon and averages 15.5 percent (range of 13.7 to 16.8 percent, at the 4-step horizon. As these examples indicate, with DGP 2, the over-sizing is a little greater at the 4-step horizon than the 1-step horizon.

Tests of forecast encompassing (or equality of adjusted MSEs) based on the approach of Hubrich and West (2010) have reasonable size properties at the 1-step horizon, but not the 4-step horizon. With DGP 3 and 1-step ahead forecasts, the size of the ENC-t test compared against critical values obtained by simulating the maximum of normal random

variables ranges from 5.4 to 9.8 percent. For most T, P settings with DGP 3, the Hubrich-West approximation yields a slightly to modestly undersized test, consistent with simulation results in Hubrich and West (2010). But with DGP 2 and 1-step ahead forecasts, the ENC-t test compared against Hubrich-West critical values can be slightly undersized or slightly oversized, with a rejection rate ranging from 8.1 to 12.1 percent. At the 4-step horizon, the Hubrich-West approach yields too high a rejection rate, especially for small P. For example, with DGP 3, the rejection rate ranges from 16.7 to 35.6 percent. The clear tendency of size to improve as P increases suggests the over-sizing is due to small-sample imprecision of the autocorrelation-consistent estimated variance of the normal random variables.[8]

The results in Tables 1 and 2 also indicate that tests of equal MSE based on critical values obtained from a non-parametric bootstrap are generally unreliable for the null of equal accuracy at the population level. Rejection rates based on the non-parametric bootstrap are systematically too low. In particular, across all three DGPs and the two forecast horizons, the size of the MSE-t test peaks at 1.2 percent. At the 1-step horizon, the size of the MSE-F test is modestly better, peaking at 4.2 percent. At the 4-step horizon, the MSE-F test based on the non-parametric bootstrap is generally undersized for DGPs 1 and 3 but ranges from modestly undersized to slightly oversized for DGP 2. Consistent with the results in Hansen (2005), the SPA test is slightly more accurately sized than the MSE-t based on the non-parametric bootstrap. The biggest improvement occurs at the smallest forecast sample size, of $P = 40$. For example, with DGP 2 and 1-step forecasts, the rejection rate of the SPA test is 8.5 percent, compared to 1.2 percent for the non-parametric version of the MSE-t test. At the 4-step horizon, the spike in the rejection rate at $P = 40$ is large enough to make the SPA test over-sized: for instance, with DGP 2, the rejection rates of the SPA and MSE-t tests are 34.8 and 0.4 percent, respectively. This spike that occurs with a small sample and a multi-step horizon suggests the problem rests in the autocorrelation-consistent variance that enters the test statistic.

4.4 Power results

Broadly, the Monte Carlo results on finite-sample power reflect the size results. For testing the null of equal forecast accuracy at the population level, power is much better for tests based on our proposed fixed regressor bootstrap than the non-parametric bootstrap. For a t-test of forecast encompassing, power based on the Hubrich-West approach to inference is

[8] In some additional experiments, incorporating pre-whitening in the HAC estimator did not improve size.

similar to, although generally a bit lower, than power based on the fixed regressor bootstrap.

More specifically, based on critical values from the fixed regressor bootstrap, the powers of the MSE-F, MSE-t, ENC-F, and ENC-t tests are generally consistent with the patterns Clark and McCracken (2001, 2005) summarize for pairwise forecast comparisons. The empirical powers of these tests can be ranked as follows: ENC-F > MSE-F, ENC-t > MSE-t. MSE-F is often more powerful than ENC-t, but the ranking of these two tests varies with τ and the T, P setting. For example, with 1-step ahead forecasts, DGP 1, and T, P = 80,80, the powers of the MSE-F, MSE-t, ENC-F, and ENC-t tests are, respectively, 67.0, 43.3, 76.2, and 62.1 percent, respectively. As might be expected, power increases with the size of the forecast sample. For instance, with 1-step ahead forecasts, DGP 1, and T = 80, the rejection rate of the MSE-F test rises from 48.5 percent at $P = 40$ to 78.9 percent at $P = 120$.

At the 1-step horizon, using the Hubrich and West (2010) approach to simulating critical values for the ENC-t test yields modestly lower power than does the fixed regressor bootstrap. For example, in the DGP 3 results, the power of the ENC-t test based on the fixed regressor bootstrap ranges from 32.2 to 64.9 percent; power based on Hubrich-West critical values ranges from 30.5 to 58.9 percent. However, at the 4-step horizon, the Hubrich-West approach yields higher rejection rates than the fixed regressor bootstrap method (for ENC-t), due to the size distortions of the Hubrich-West approach at the multi-step horizon. For instance, with DGP 3 and $\tau = 4$, the power of the ENC-t test based on the fixed regressor bootstrap ranges from 16.6 to 36.7 percent; power based on Hubrich-West critical values ranges from 35.2 to 48.1 percent.

Rejection rates based on the non-parametric bootstrap are much lower. For the MSE-t test, in most cases power is trivial, in the sense that it is below the nominal size of the test (10 percent). For example, under DGP 1, the rejection rate of the MSE-t test ranges from 0.2 to 6.0 percent (across forecast horizons and sample sizes). Power is quite a bit higher for the MSE-F and SPA tests than the MSE-t test, but still well below the power of the tests based on the fixed regressor bootstrap. For example, with 1-step ahead forecasts, DGP 1, and T, P = 80,80, the powers of the MSE-F and SPA tests based on the non-parametric bootstrap are 21.4 and 12.9 percent, respectively, compared to the 6.0 percent power of the MSE-t test based on the non-parametric bootstrap. Using critical values from the fixed regressor bootstrap raises the powers of the MSE-F and MSE-t tests to 67.0 and

43.3 percent, respectively. While power based on the non-parametric bootstrap tends to be slightly to modestly higher with DGPs 2 and 3 than with DGP 1, the same patterns prevail.

5 Applications

In this section we illustrate the use of the tests and inference approaches described above with two applications. In the first, based on Chen, Rogoff, and Rossi (2010), we examine the predictive content of exchange rates for commodity prices, at a monthly frequency. In the second, patterned after studies such as Stock and Watson (2003), we apply our tests to forecasts of quarterly U.S. GDP growth based on a range of potential leading indicators.

More specifically, in the commodity price application, we examine forecasts of monthly growth in commodity prices from a total of 28 models. Commodity prices are measured with the spot price for industrials published by the Commodities Research Bureau (CRB). The null model includes a constant and one lag of growth in commodity prices. The alternative models add various combinations of a commodity futures price (the CRB index for industrial commodities) and exchange rates, all in growth rates and lagged one month (and with all exchange rates relative to the U.S. dollar). Drawing on Chen, Rogoff, and Rossi (2010), we use exchange rates for a few important commodity economies with relatively long histories of floating exchange rates (Australia, Canada, and New Zealand), some other industrialized economies (U.K. and Japan), one index of exchange rates for major U.S. trading partners, and another index of exchange rates for other important U.S. trading partners. To ensure some heterogeneity in predictive content, we have deliberately included some exchange rates (e.g., for Japan and the U.K.) that, based on the conceptual framework of Chen, Rogoff, and Rossi, should not be expected to have predictive content for commodity prices.

The full set of variables and models for the commodity price application is listed in Table 5. Appendix 2 provides further descriptions of the data. Our model estimation sample begins with January 1987, and we examine recursive 1-month ahead forecasts (that is, our estimation sample expands as forecasting moves forward in time) for 1997 through 2008.

In the GDP application, we examine 1-quarter and 4-quarter ahead forecasts of real GDP growth from 14 models. The null model includes a constant and one lag of GDP

growth, where GDP growth is measured as $(400/\tau)\ln(\text{GDP}_{t+\tau}/\text{GDP}_t)$:

$$y_{t+\tau} = b_0 + b_1 y_t + u_{t+\tau}. \tag{11}$$

Each of 13 alternative models adds in one lag of a (potential) leading indicator x_t:

$$y_{t+\tau} = b_0 + b_1 y_t + b_2 x_t + u_{t+\tau}, \tag{12}$$

where the set of leading indicators includes the change in consumption's share in GDP (measured with nominal data), weekly hours worked in manufacturing, building permits, purchasing manager indexes for supplier delivery times and orders, new claims for unemployment insurance, growth in real stock prices (real price = S&P 500 index/core PCE price index), the change in the 3-month Treasury bill rate, the change in the 1-year Treasury bond yield, the change in the 10-year Treasury bond yield, the 3-month to 10-year yield spread, the 1-year to 10-year yield spread, and the spread between Aaa and Baa corporate bond yields (from Moody's).

The full set of variables and models for the GDP growth application is listed in Table 6. Appendix 2 provides further descriptions of the data. Our model estimation sample begins with 1961:Q2, and we examine recursive forecasts for 1985:Q1+τ-1 through 2009:Q4.

Tables 5-7 provide the results of the applications. Tables 5 and 6 report RMSE ratios for each alternative model forecast relative to the benchmark (a number less than 1 indicates the alternative is more accurate than the benchmark) and p-values for tests of equal MSE and forecast encompassing, applied on a pairwise basis for each alternative compared to the benchmark. The models are listed in order of forecast accuracy (most to least accurate relative to the benchmark model) as measured by RMSE. Table 7 provides p-values of the reality check tests, along with a listing of the best model identified by each test. We use 9999 replications in computing the bootstrap p-values.

The pairwise forecast comparisons of Table 5 indicate that both exchange rates and the commodity futures price have predictive content for spot commodity prices: nearly all of the alternative models forecast commodity prices more accurately — although only slightly so, in terms of RMSEs — than the benchmark AR(1). The model with the lowest RMSE includes the constant and commodity price lag of the benchmark model, the futures price, and the Australian dollar exchange rate. Ranked by RMSE, the next few models include various combinations of the futures price, Australian dollar exchange rate, major country exchange rate index, and other important trading partners exchange rate index. According

to the pairwise tests based on the fixed regressor bootstrap, more than one-half of the models are significantly better than the benchmark at the 10 percent significance level. Consistent with our Monte Carlo evidence, using a non-parametric bootstrap consistently yields higher p-values and implies fewer models to have predictive content on a pairwise basis.

The reality check results provided in the top panel of Table 7 show that, taking the search for a best model into account, most of the tests compared against our proposed bootstrap critical values continue to reject the null of equal forecast accuracy. In particular, the lowest MSE model remains significantly better than the benchmark: the reality check version of the MSE-F test has a fixed regressor bootstrap p-value of 1.6 percent; the reality check version of the ENC-F test identifies the same model as best and rejects equal accuracy with a p-value of 4.1 percent. The t-tests for equal MSE and encompassing identify other models as best, with the reality check version of the MSE-t test indicating the model with the futures price is significantly more accurate than the benchmark but the ENC-t test failing to reject the null of equal accuracy. Again, consistent with the Monte Carlo results, p-values are considerably higher with the non-parametric bootstrap than our fixed regressor approach. Under the non-parametric approach to the reality check, neither MSE-t nor SPA reject the null of equal accuracy.

The pairwise forecast comparisons of Table 6 show that, at the 1-quarter horizon, a handful of variables have significant predictive content for real GDP growth, while evidence of predictive content is much weaker at the 4-quarter horizon. At the shorter horizon, most of the tests based on a fixed regressor bootstrap indicate that five models — the ones including (in addition to the constant and GDP growth lag of the benchmark), respectively, the change in the consumption share, growth in building permits, growth in stock prices, the Baa-Aaa interest rate spread, and the purchasing manager index of new orders — forecast significantly better than the AR(1) benchmark (using a 10 percent significance level). As expected, p-values based on the non-parametric bootstrap are considerably higher and yield weaker evidence of predictive content, with both the MSE-F and MSE-t tests rejecting the null for just the model including the change in the consumption share.

At the longer horizon, only three models have RMSEs lower than the benchmark, and only one — the model including growth in building permits — is significantly better according to the pairwise tests. However, the ENC-F and ENC-t tests indicate that, at the population level, stock prices have significant predictive content for GDP growth, even

though, in this sample, the model yields an RMSE equal to that of the benchmark model. The ENC-F test yields the same result for the purchasing manager index of new orders.

The reality check test results provided in the second and third panels of Table 7 show that some of the evidence of predictive content in leading indicators for GDP growth hold up (using fixed regressor bootstrap p-values) once the search for a best model is taken into account. As should be expected, the p-values are generally higher for reality check tests than pairwise tests. But the significance in the pairwise case mostly holds up under the reality check microscope. In particular, at the 1-quarter horizon, the MSE-F test continues to indicate that the consumption share significantly improves forecasts of GDP growth — using fixed regressor critical values, but not non-parametric critical values. At the 4-quarter horizon, the reality check p-values of the MSE-F and ENC-F tests are 0.1 and 0.3 percent, respectively, indicating building permits have significant predictive content for GDP growth even once model search is taken into account. However, the reality check p-values for MSE-t and ENC-t tests (which generally have lower power than their F-type counterparts, according to our Monte Carlo results) are above the 10 percent threshold, failing to reject the null.

6 Conclusion

This paper develops a new bootstrap method for simulating asymptotic critical values for tests of equal forecast accuracy and encompassing among many nested models. The bootstrap, which combines elements of fixed regressor and wild bootstrap methods, is simple to use. We first derive the asymptotic distributions of tests of equal forecast accuracy and encompassing applied to forecasts from multiple models that nest the benchmark model – that is, reality check tests applied to nested models. These distributions are non-standard and involve unknown nuisance parameters. We then prove the validity of the bootstrap for simulating critical values from these distributions.

Using our new fixed regressor bootstrap, we then conduct a range of Monte Carlo simulations to examine the finite-sample properties of the tests. These experiments indicate our proposed bootstrap has good size and power properties, especially in comparison to the non-parametric methods of White (2000) and Hansen (2005) developed for use with non-nested models. In the final part of our analysis, we illustrate the use of our tests with applications to forecasts of commodity prices and GDP growth.

7 Appendix 1: Proofs

The following additional notation will be used. For any matrix A, let $|A|$ denote the max norm, let $\sup_t$ denote $\sup_{T\leq t\leq T+P-\tau}$, let Ω_{11} denote the upper $k\times k$ block-diagonal element of Ω, and define both $Q_j(t) = -J_jB_j(t)J_j' + B(t)$ and $Q_j = -J_jB_jJ_j' + B$. Define $v^*_{s+\tau} = (\eta_{s+\tau}\varepsilon_{s+\tau} + \theta_1\eta_{s+\tau-1}\varepsilon_{s+\tau-1}+\cdots+\theta_{\tau-1}\eta_{s+1}\varepsilon_{s+1}), \widehat{v}^*_{s+\tau} = (\eta_{s+\tau}\widehat{\varepsilon}_{s+\tau}+\widehat{\theta}_1\eta_{s+\tau-1}\widehat{\varepsilon}_{s+\tau-1}+\cdots+\widehat{\theta}_{\tau-1}\eta_{s+1}\widehat{\varepsilon}_{s+1}), h^*_{s+\tau} = x_sv^*_{s+\tau}$, $\widehat{h}^*_{s+\tau} = x_s\widehat{v}^*_{s+\tau}$, $H^*(T) = (T^{-1}\sum_{t=1}^{T-\tau} h^*_{s+\tau})$ and $\hat{H}^*(T) = (T^{-1}\sum_{t=1}^{T-\tau}\hat{h}^*_{s+\tau})$. More generally, we let * denote a statistical property, such as convergence in distribution $\rightarrow^{d^*}$, defined with respect to the bootstrap-induced probability measure P^*.

Proof of Theorems 3.1 and 3.2: In both cases, the proof follows almost directly from Theorems 3.1-3.4 in Clark and McCracken (2005) adjusted for a few details. Specifically, for a fixed model j, if we redefine $\tilde{h}_{j,s+\tau}$ as $\tilde{h}_{s+\tau}$ the algebra follows identically from Clark and McCracken (2005) and hence we obtain the asymptotic distributions for the pairwise comparisons. That we obtain the appropriate joint distribution of all M of the pairwise test statistics follows from the rank condition on Ω in Assumption 2 and the fact that under our assumptions, Corollary 29.19 of Davidson (1994) suffices for $T^{-1/2}\sum_{s=1}^{t} S_{\tilde{h}\tilde{h}}^{-1/2}\tilde{h}_{s+\tau}$ to converge weakly to a standard Brownian motion $W(s)$. The result then follows from an application of the Continuous Mapping Theorem to the max functional.

Lemma 1: Maintain Assumptions $1, 2$, and 4. (a) $\sup_t |T^{1/2}(B_j(t)-B_j)| = O_p(1)$. Maintain Assumptions $1, 2'$ and 4. (b) $T^{-1/2}\sum_{s=1}^{t-\tau} h^*_{s+\tau} \Rightarrow^* \Omega_{11}^{1/2}W^*(s)$. (c) $\sum_{t=T}^{T+P-\tau}(T^{-1/2}\tilde{h}'^*_{j,t+\tau})(T^{1/2}\tilde{H}^*_j(t)) \rightarrow^{d^*} \Gamma^*_{1,j}$. (d) $\sup_t |T^{1/2}H^*(t))| = O_{p^*}(1)$. (e) $\sup_t |T^{1/2}(\hat{H}^*(t) - H^*(t))| = o_{p^*}(1)$, (f) $\hat{\sigma}_j^{2*} \rightarrow^{p^*} \sigma^2$.

Proof of Lemma 1: (a) The proof is given in Lemma A1 of Clark and McCracken (2005).

(b) First note that relative to the bootstrap-induced probability measure P^*, $h^*_{s+\tau}$ is a heteroskedastic vector $MA(\tau-1)$ sequence with independent, zero mean, Normally distributed increments. As such it is a L_r-bounded k-vector sequence with each element L_2-NED of size -1 on an α-mixing process of size $-r/(r-2)$. Second, note that since the increments η_s are $i.i.d.$ $N(0,1)$, we obtain $E^*(T^{-1/2}\sum_{s=1}^{t-\tau} h^*_{s+\tau})(T^{-1/2}\sum_{s=1}^{t-\tau} h^*_{s+\tau})' = (T^{-1/2}\sum_{s=1}^{t-\tau} h_{s+\tau})(T^{-1/2}\sum_{s=1}^{t-\tau} h_{s+\tau})'$. Finally, since Assumption $2'$ implies $\lim_{T\rightarrow\infty}(T^{-1/2}\sum_{s=1}^{T} h_{s+\tau})(T^{-1/2}\sum_{s=1}^{T} h_{s+\tau})' \rightarrow^p \Omega_{11} < \infty$, Corollary 29.19 of Davidson (1994) implies $T^{-1/2}\sum_{s=1}^{t-\tau} h^*_{s+\tau} \Rightarrow^* \Omega_{11}^{1/2}W^*(s)$ and the proof is complete.

(c) Given the proof of (b) (notably the delineation of the properties associated with $h^*_{s+\tau}$), Theorem 30.14 of Davidson (1994) implies $\sum_{t=T}^{T+P-\tau}(T^{-1/2}\tilde{h}'^*_{j,t+\tau})(T^{1/2}\tilde{H}^*_j(t)) \rightarrow^{d^*} \Gamma^*_{1,j}$. Note that the typical drift term associated with a stochastic integral based on correlated increments is zero because of the $\tau-$lag between $\tilde{h}^*_{j,t+\tau}$ and $\tilde{H}^*_j(t)$ – that is, $E^*(\tilde{h}^*_{j,t+\tau}|\tilde{H}^*_j(t)) = 0$ for all t. A detailed argument is given in Lemma A1 of Clark and McCracken (2005).

(d) First note that since $T \leq t \leq T+P-\tau$, $\sup_t |T^{1/2}H^*(t))| \leq \sup_t |T^{-1/2}\sum_{s=1}^{t-\tau} h^*_{s+\tau}|$. The Continuous Mapping Theorem, Lemma 1 (b), and Assumption 4 then imply $\sup_t |T^{-1/2}\sum_{s=1}^{t-\tau} h^*_{s+\tau}| \rightarrow^{d^*} \sup_{1\leq s\leq 1+\lambda_P} |\Omega_{11}^{1/2}W^*(s)| = O_{p^*}(1)$ and the proof is complete.

(e) For ease of presentation, we show the result assuming $\tau = 2$ and hence $\widehat{v}^*_{s+2} = \eta_{s+2}\widehat{\varepsilon}_{s+2} + \widehat{\theta}\eta_{s+1}\widehat{\varepsilon}_{s+1}$ and $v^*_{s+2} = \eta_{s+2}\varepsilon_{s+2} + \theta\eta_{s+1}\varepsilon_{s+1}$. Rearranging terms gives us

$$\begin{aligned} T^{1/2}(\widehat{H}^*(t) - H^*(t)) &= T^{-1/2}\sum_{s=1}^{t-\tau}(\widehat{v}^*_{s+2} - v_{s+2})x_s = T^{-1/2}\sum_{s=1}^{t-\tau}(\eta_{s+2}(\widehat{\varepsilon}_{s+2} - \varepsilon_{s+2}) \\ &\quad +\theta\eta_{s+1}(\widehat{\varepsilon}_{s+1} - \varepsilon_{s+1}) + (\widehat{\theta} - \theta)\eta_{s+1}(\widehat{\varepsilon}_{s+1} - \varepsilon_{s+1}) + (\widehat{\theta} - \theta)\eta_{s+1}\varepsilon_{s+1})x_s. \end{aligned}$$

If we take a first order Taylor expansion of both $\widehat{\varepsilon}_{s+2}$ and $\widehat{\varepsilon}_{s+1}$, then for some $\overline{\gamma}_{s+2}$ and $\overline{\gamma}_{s+1}$ in the closed cube with opposing vertices $\widehat{\gamma}$ and γ we obtain

$$\begin{aligned} T^{1/2}(\widehat{H}^*(t) - H^*(t)) &= T^{-1/2}\sum_{s=1}^{t-\tau}(\eta_{s+2}\nabla\widehat{\varepsilon}_{s+2}(\overline{\gamma}_{s+2})(\widehat{\gamma} - \gamma) + \theta\eta_{s+1}\nabla\widehat{\varepsilon}_{s+1}(\overline{\gamma}_{s+1})(\widehat{\gamma} - \gamma) \\ &\quad +(\widehat{\theta} - \theta)\eta_{s+1}\nabla\widehat{\varepsilon}_{s+1}(\overline{\gamma}_{s+1})(\widehat{\gamma} - \gamma) + (\widehat{\theta} - \theta)\eta_{s+1}\varepsilon_{s+1})x_s \end{aligned}$$

and hence

$$\begin{aligned} \sup_t |T^{1/2}(\widehat{H}^*(t) - H^*(t))| &\leq (k+1)\sup_t |T^{-1}\sum_{s=1}^{t-\tau}\eta_{s+2}x_s\nabla\widehat{\varepsilon}_{s+2}(\overline{\gamma}_{s+2})||T^{1/2}(\widehat{\gamma} - \gamma)| \\ &\quad +|\theta|(k+1)\sup_t |T^{-1}\sum_{s=1}^{t-\tau}\eta_{s+1}x_s\nabla\widehat{\varepsilon}_{s+1}(\overline{\gamma}_{s+1})||T^{1/2}(\widehat{\gamma} - \gamma)| \\ &\quad +|\widehat{\theta} - \theta|(k+1)\sup_t |T^{-1}\sum_{s=1}^{t-\tau}\eta_{s+1}x_s\nabla\widehat{\varepsilon}_{s+1}(\overline{\gamma}_{s+1})||T^{1/2}(\widehat{\gamma} - \gamma)| \\ &\quad +|T^{1/2}(\widehat{\theta} - \theta)|\sup_t |T^{-1}\sum_{s=1}^{t-\tau}\eta_{s+1}x_s\varepsilon_{s+1}|. \end{aligned}$$

Assumptions 1 and 2′ suffice for both $T^{1/2}(\widehat{\gamma} - \gamma)$ and $T^{1/2}(\widehat{\theta} - \theta)$ to be $O_p(1)$. In addition since, for large enough samples, Assumption 2′ bounds the second moments of $\nabla\widehat{\varepsilon}_{s+2}(\overline{\gamma}_{s+2})$ and $\nabla\widehat{\varepsilon}_{s+1}(\overline{\gamma}_{s+1})$ as well as x_s, the fact that the $\eta_{s+\tau}$ are *i.i.d.* $N(0,1)$ then implies $T^{-1}\sum_{s=1}^{T-\tau}\eta_{s+2}x_s\nabla\widehat{\varepsilon}_{s+2}(\overline{\gamma}_{s+2})$, $T^{-1}\sum_{s=1}^{T-\tau}\eta_{s+1}x_s\nabla\widehat{\varepsilon}_{s+1}(\overline{\gamma}_{s+1})$, and $T^{-1}\sum_{s=1}^{T-\tau}\eta_{s+1}x_s\varepsilon_{s+1}$ are all $o_{a.s.^*}(1)$. This in turn, (along with Assumption (4)) implies that $sup_t|.|$ of each of the partial sums is $o_{p^*}(1)$ and the proof is complete.

(f) Straightforward algebra shows that

$$\begin{aligned} \hat{\sigma}_j^{2*} &= P^{-1}\sum_{t=T}^{T+P-\tau}\hat{u}^{*2}_{j,t+\tau} = \{P^{-1}\sum_{t=T}^{T+P-\tau}\hat{v}^{*2}_{t+\tau}\} \\ &\quad -\{2P^{-1}\sum_{t=T}^{T+P-\tau}\hat{h}'^*_{t+\tau}(Q_0(t) - Q_j(t))B^{-1}(t)J_0\hat{\beta}_{0,T} - 2P^{-1}\sum_{t=T}^{T+P-\tau}\hat{h}'^*_{t+\tau}B_j(t)J_j{}'\hat{H}^*(t) \\ &\quad +\hat{\beta}'_{0,T}J_0[P^{-1}\sum_{t=T}^{T+P-\tau}B^{-1}(t)(Q_0(t) - Q_j(t))x_tx_t'(Q_0(t) - Q_j(t))B^{-1}(t)]J_0'\hat{\beta}_{0,T} \\ &\quad -2\hat{\beta}'_{0,T}J_0[P^{-1}\sum_{t=T}^{T+P-\tau}B^{-1}(t)(Q_0(t) - Q_j(t))x_tx_t'J_jB_j(t)J_j'\hat{H}^*(t)] \\ &\quad +P^{-1}\sum_{t=T}^{T+P-\tau}\hat{H}^*(t)J_jB_j(t)J_j'x_tx_t'J_jB_j(t)J_j'\hat{H}^*(t)\}. \end{aligned}$$

We first show that $P^{-1}\sum_{t=T}^{T+P-\tau}\hat{v}^{*2}_{t+\tau} \rightarrow^{p^*} \sigma^2$. If we take a first order Taylor expansion of $\hat{v}^{*2}_{t+\tau}$ then for some $\overline{\gamma}_{t+\tau}$ in the closed cube with opposing vertices $\hat{\gamma}$ and γ we obtain $\hat{v}^{*2}_{t+\tau} =$

$v_{t+\tau}^{*2} + 2\hat{v}_{t+\tau}^{*}(\overline{\gamma}_{t+\tau})(\partial\hat{v}_{t+\tau}^{*}(\overline{\gamma}_{t+\tau})/\partial\gamma)(\hat{\gamma}-\gamma)$. That $P^{-1}\sum_{t=T}^{T+P-\tau} v_{t+\tau}^{*2} \rightarrow^{p^*} \sigma^2$ follows from the fact that $E^*(P^{-1}\sum_{t=T}^{T+P-\tau} v_{t+\tau}^{*2}) = P^{-1}\sum_{t=T}^{T+P-\tau} v_{t+\tau}^{2} \rightarrow^{p} \sigma^2$ and $\lim_{T\rightarrow\infty} V^*(P^{-1}\sum_{t=T}^{T+P-\tau} v_{t+\tau}^{*2}) = 0$. Since Assumptions 1 and 2′ suffice for both $P^{-1}\sum_{t=T}^{T+P-\tau} \hat{v}_{t+\tau}^{*}(\overline{\gamma}_{t+\tau})(\partial\hat{v}_{t+\tau}^{*}(\overline{\gamma}_{t+\tau})/\partial\gamma) = O_{p^*}(1)$ and $\hat{\gamma}-\gamma = o_p(1)$, the proof is complete.

We now must show that each element of the second bracketed right-hand side term is $o_{p^*}(1)$. For brevity we only show the result for the first and third terms. For the first bracketed term note that since the η_t's are *i.i.d.* zero mean increments, conditional on the observables $\hat{h}_{t+\tau}^{\prime *}(Q_0(t) - Q_j(t))B^{-1}(t)J_0$ is a heteroskedastic $MA(\tau-1)$ process with finite variance and hence $P^{-1}\sum_{t=T}^{T+P-\tau} \hat{h}_{t+\tau}^{\prime *}(Q_0(t) - Q_j(t))B^{-1}(t)J_0 = o_{p^*}(1)$. Since $\hat{\beta}_{0,T} = O_p(1)$ the result is complete. For the third bracketed term, algebra along the lines of that in Clark and McCracken (2005) implies that $P^{-1}\sum_{t=T}^{T+P-\tau} J_0B^{-1}(t)(Q_0(t)-Q_j(t))x_tx_t'(Q_0(t)-Q_j(t))B^{-1}(t)J_0' \rightarrow^{p} J_0B^{-1}[Q_0-Q_j]B^{-1}J_0'$. Since $\hat{\beta}_{0,T} = O_p(1)$ and $J_0B^{-1}[Q_0 - Q_j]B^{-1}J_0' = 0$ the proof is complete.

Proof of Theorem 3.3: Given Lemma 1(f), throughout we will ignore the denominator term $\hat{\sigma}_j^{2*}$ in both the MSE-F_j^* and ENC-F_j^* statistics. In parts (a) and (b) below we focus on the asymptotic distributions for a fixed pairwise comparison between models 0 and j. That we obtain the appropriate joint distribution of all M of the pairwise test statistics follows from the rank condition on Ω in Assumption 2′, Lemma 1(b), and the fact that for each j, each of the statistics have asymptotic representations as functionals of the same standard Brownian motion $W^*(s)$. The result then follows from an application of the Continuous Mapping Theorem to the `max` functional.

(a) Straightforward algebra implies that for each j,

$$\begin{array}{l}
\sum_{t=T}^{T+P-\tau}(\hat{u}_{0,t+\tau}^{*2} - \hat{u}_{j,t+\tau}^{*2}) = \sum_{t=T}^{T+P-\tau}\{2h_{t+\tau}^{\prime *}(Q_0(t)-Q_j(t))H^*(t) \\
-H'^*(t)(-J_0B_0(t)J_0'x_tx_t'J_0B_0(t)J_0' + J_jB_j(t)J_j'x_tx_t'J_jB_j(t)J_j')H^*(t)\} \\
+2\sum_{t=T}^{T+P-\tau}\{h_{t+\tau}^{\prime *}(Q_0(t)-Q_j(t))(\hat{H}^*(t)-H^*(t)) + (\hat{h}_{t+\tau}^{*} - h_{t+\tau}^{*})'(Q_0(t)-Q_j(t))H^*(t) \\
-H'^*(t)(-J_0B_0(t)J_0'x_tx_t'J_0B_0(t)J_0' + J_jB_j(t)J_j'x_tx_t'J_jB_j(t)J_j')(\hat{H}^*(t)-H^*(t)) \\
+(\hat{h}_{t+\tau}^{*} - h_{t+\tau}^{*})'(Q_0(t)-Q_j(t))(\hat{H}^*(t)-H^*(t)) \\
-(0.5)(\hat{H}^*(t)-H^*(t))'(-J_0B_0(t)J_0'x_tx_t'J_0B_0(t)J_0' + J_jB_j(t)J_j'x_tx_t'J_jB_j(t)J_j')(\hat{H}^*(t)-H^*(t))\}
\end{array} \tag{13}$$

Note that there are 2 bracketed $\{.\}$ terms in (13). We will show that the first of these has as its limit $2\Gamma_{1,j}^* - \Gamma_{2,j}^*$ where $\Gamma_i^* =^{d^*} \Gamma_i$, for Γ_i, $i = 1, 2$, defined in the text. We then show that the remaining bracketed term is $o_{p^*}(1)$.

Proof of bracket 1: Consider the first part of the bracket. Rearranging terms gives us

$$\begin{array}{l}
\sum_{t=T}^{T+P-\tau} 2h_{t+\tau}^{\prime *}(Q_0(t)-Q_j(t))H^*(t) = 2\sum_{t=T}^{T+P-\tau} h_{t+\tau}^{\prime *}(Q_0-Q_j)H^*(t) \\
+\sum_{t=T}^{T+P-\tau} h_{t+\tau}^{\prime *}\{(Q_0(t)-Q_j(t)) - (Q_0-Q_j)\}H^*(t) \\
= 2\sigma^2\sum_{t=T}^{T+P-\tau}(T^{-1/2}\tilde{h}_{j,t+\tau}^{\prime *})(T^{1/2}\tilde{H}_j^*(t)) + \\
T^{-1}\sum_{t=T}^{T+P-\tau} h_{t+\tau}^{\prime *}\{T^{1/2}((Q_0(t)-Q_j(t)) - (Q_0-Q_j))\}(T^{1/2}H^*(t)),
\end{array}$$

where $\tilde{h}_{j,t+\tau}^{*}$ and $\tilde{H}_j^*(t)$ are the bootstrap equivalents of $\tilde{h}_{j,t+\tau}$ and $\tilde{H}_j(t)$ defined in section 3.1.

That $\sigma^2 \sum_{t=T}^{T+P-\tau}(T^{-1/2}\tilde{h}'^*_{j,t+\tau})(T^{1/2}\tilde{H}^*_j(t)) \rightarrow^{d^*} \sigma^2\Gamma^*_{1,j}$ follows from Lemma 1 (c). To show that the remainder term is $o_{p^*}(1)$ note that Lemmas 1 (a) and (d) imply $\sup_t |T^{1/2}H^*(t)| = O_{p^*}(1)$ and $\sup_t |T^{1/2}((Q_0(t) - Q_j(t)) - (Q_0 - Q_j))| = O_p(1)$. Since $E^*(h^*_{t+\tau}|H^*(t), B_0(t), B_j(t)) = 0$ for all t and $\lim_{T\rightarrow\infty} Var^*(T^{-1}\sum_{t=T}^{T+P-\tau} h'^*_{t+\tau}\{T^{1/2}((Q_0(t) - Q_j(t)) - (Q_0 - Q_j))\}(T^{1/2}H^*(t))) = 0$ the proof is complete.

Now consider the second part of the bracket. Rearranging terms gives us

$$\begin{array}{l}
\sum_{t=T}^{T+P-\tau} H'^*(t)(-J_0B_0(t)J_0'x_tx_t'J_0B_0(t)J_0' + J_jB_j(t)J_j'x_tx_t'J_jB_j(t)J_j')H^*(t) \\
= \sum_{t=T}^{T+P-\tau} H'^*(t)(Q_0 - Q_j)H^*(t) \\
+ \sum_{t=T}^{T+P-\tau} H'^*(t)\{(-J_0B_0(t)J_0'x_tx_t'J_0B_0(t)J_0' + J_jB_j(t)J_j'x_tx_t'J_jB_j(t)J_j') - (Q_0 - Q_j)\}H^*(t) \\
= \sigma^2 \sum_{t=T}^{T+P-\tau} \tilde{H}'^*(t)\tilde{H}^*(t) \\
+ \sum_{t=T}^{T+P-\tau} H'^*(t)\{(-J_0B_0(t)J_0'x_tx_t'J_0B_0(t)J_0' + J_jB_j(t)J_j'x_tx_t'J_jB_j(t)J_j') - (Q_0 - Q_j)\}H^*(t).
\end{array}$$

That $\sigma^2 \sum_{t=T}^{T+P-\tau} \tilde{H}'^*_j(t)\tilde{H}^*_j(t) \rightarrow^{d^*} \sigma^2\Gamma^*_{2,j}$ follows from the Continuous Mapping Theorem and Lemma 1 (b). To show that the remaining term is $o_{p^*}(1)$ note that by adding and subtracting terms we obtain

$$\begin{array}{l}
\sum_{t=T}^{T+P-\tau} H'^*(t)\{(-J_0B_0(t)J_0'x_tx_t'J_0B_0(t)J_0' + J_jB_j(t)J_j'x_tx_t'J_jB_j(t)J_j') - (Q_0 - Q_j)\}H^*(t) \\
= \sum_{(\text{all } m,n,o)=1,2} \sum_{t=T}^{T+P-\tau} H'^*(t)\{(-J_0a_{m,t}J_0'b_{n,t}J_0a_{o,t}J_0' + J_jc_{m,t}J_j'b_{n,t}J_jc_{o,t}J_j') - (Q_0 - Q_j)\}H^*(t),
\end{array}$$

where $a_{1,t} = B_0$, $a_{2,t} = B_0(t) - B_0$, $b_{1,t} = B^{-1}$, $b_{2,t} = x_tx_t' - B^{-1}$, $c_{1,t} = B_j$, and $c_{2,t} = B_j(t) - B_j$. If the indices m, n, o are all 1 then the remainder term is numerically zero and hence it suffices to show that for all other permutations of the indices that the elements of the remainder term are $o_{p^*}(1)$. The proofs of each are very similar and hence we show the result for the case when the indices are all equal to 2. To do so note that

$$\begin{array}{rl}
& |\sum_{t=T}^{T+P-\tau} H'^*(t)(-J_0a_{2,t}J_0'b_{2,t}J_0a_{2,t}J_0' + J_jc_{2,t}J_j'b_{2,t}J_jc_{2,t}J_j')H^*(t)| \\
\leq & 2k^4T^{-1}(\sup_t |T^{1/2}H^*(t)|)^2(\max_i \sup_t |T^{1/2}(B_i(t) - B_i)|)^2(T^{-1}\sum_{t=T}^{T+P-\tau} |x_tx_t' - B^{-1}|)
\end{array}$$

Lemma 1 (d) implies $\sup_t |T^{1/2}H^*(t)|$ is $O_{p^*}(1)$ while Lemma 1 (a) implies $\max_i \sup_t |T^{1/2}(B_i(t) - B_i)|$ is $O_p(1)$. Since Assumption 2′ is sufficient for $T^{-1}\sum_{t=T}^{T+P-\tau} |x_tx_t' - B^{-1}|$ to be $O_p(1)$ the result follows from the fact that T^{-1} is $o(1)$.

Proof of bracket 2: We must show that each of the five components of the second bracketed term in (13) are $o_{p^*}(1)$. The proofs of each are similar and as such we only show the result for the fourth component. If we take the absolute value of this term we find that

$$\begin{array}{l}
|\sum_{t=T}^{T+P-\tau} (\hat{h}^*_{t+\tau} - h^*_{t+\tau})'(Q_0(t) - Q_j(t))(\hat{H}^*(t) - H^*(t))| \\
\leq k^2(T^{-1/2}\sum_{t=T}^{T+P-\tau} |\hat{h}^*_{t+\tau} - h^*_{t+\tau}|)(\sup_t |Q_0(t) - Q_j(t)|)(\sup_t T^{1/2}|\hat{H}^*(t) - H^*(t)|).
\end{array}$$

Lemma 1(e) implies $\sup_t T^{1/2}|\hat{H}^*(t) - H^*(t)| = o_{p^*}(1)$ while Assumption 2′ suffices for $\sup_t |Q_0(t) - Q_j(t)| = O_p(1)$.

The result will follow if $T^{-1/2}\sum_{t=T}^{T+P-\tau} |\hat{h}^*_{t+\tau} - h^*_{t+\tau}| = O_{p^*}(1)$. For simplicity we assume, as in the proof of Lemma 1(e), that $\tau = 2$ and hence the forecast errors form an $MA(1)$ process. If we

then take a Taylor expansion in precisely the same fashion as in the proof of Lemma 1(e) we have

$$\begin{aligned} T^{-1/2}\sum_{t=T}^{T+P-\tau}|\hat{h}^*_{t+\tau}-h^*_{t+\tau}| \quad \leq \quad & (k+1)T^{-1}\sum_{t=T}^{T+P-\tau}|\eta_{t+2}x_t\nabla\widehat{\varepsilon}_{t+2}(\overline{\gamma}_{t+2})||T^{1/2}(\widehat{\gamma}-\gamma)| \\ & +|\theta|(k+1)T^{-1}\sum_{t=T}^{T+P-\tau}|\eta_{t+1}x_t\nabla\widehat{\varepsilon}_{t+1}(\overline{\gamma}_{t+1})||T^{1/2}(\widehat{\gamma}-\gamma)| \\ & +|\widehat{\theta}-\theta|(k+1)T^{-1}\sum_{t=T}^{T+P-\tau}|\eta_{t+1}x_t\nabla\widehat{\varepsilon}_{t+1}(\overline{\gamma}_{t+1})||T^{1/2}(\widehat{\gamma}-\gamma)| \\ & +|T^{1/2}(\widehat{\theta}-\theta)|T^{-1}\sum_{t=T}^{T+P-\tau}|\eta_{t+1}x_t\varepsilon_{t+1}|. \end{aligned}$$

Assumptions 1 and 2′ suffice for both $T^{1/2}(\widehat{\gamma}-\gamma)$ and $T^{1/2}(\widehat{\theta}-\theta)$ to be $O_p(1)$. Since, for large enough samples, Assumption 2′ bounds the second moments of $\nabla\widehat{\varepsilon}_{t+2}(\overline{\gamma}_{t+2})$, $\nabla\widehat{\varepsilon}_{t+1}(\overline{\gamma}_{t+1})$, and x_t, that $\eta_{t+\tau}$ is distributed *i.i.d.* $N(0,1)$ implies $T^{-1}\sum_{t=T}^{T+P-\tau}|\eta_{t+2}x_t\nabla\widehat{\varepsilon}_{t+2}(\overline{\gamma}_{t+2})|$, $T^{-1}\sum_{t=T}^{T+P-\tau}|\eta_{t+1}x_t\nabla\widehat{\varepsilon}_{t+1}(\overline{\gamma}_{t+2})|$, and $T^{-1}\sum_{t=T}^{T+P-\tau}|\eta_{t+1}x_t\varepsilon_{t+1}|$ are all $O_{p^*}(1)$, and the proof is complete.

(b) Straightforward algebra implies that for each j,

$$\begin{aligned} & \textstyle\sum_{t=T}^{T+P-\tau}\hat{u}^*_{0,t+\tau}(\hat{u}^*_{0,t+\tau}-\hat{u}^*_{j,t+\tau}) = \sum_{t=T}^{T+P-\tau}\{h'^*_{t+\tau}(Q_0(t)-Q_j(t))H^*(t)\} \\ & -\textstyle\sum_{t=T}^{T+P-\tau}\{H'^*(t)(-J_0B_0(t)J_0'x_tx_t'J_0B_0(t)J_0'+J_0B_0(t)J_0'x_tx_t'J_jB_j(t)J_j')H^*(t)\} \\ & +\textstyle\sum_{t=T}^{T+P-\tau}\{h'^*_{t+\tau}(Q_0(t)-Q_j(t))(\hat{H}^*(t)-H^*(t)) \\ & +(\hat{h}^*_{t+\tau}-h^*_{t+\tau})'(Q_0(t)-Q_j(t))H^*(t) \\ & -H'^*(t)(-J_0B_0(t)J_0'x_tx_t'J_0B_0(t)J_0'+J_0B_0(t)J_0'x_tx_t'J_jB_j(t)J_j')(\hat{H}^*(t)-H^*(t)) \\ & -H'^*(t)(-J_0B_0(t)J_0'x_tx_t'J_0B_0(t)J_0'+J_jB_j(t)J_j'x_tx_t'J_0B_0(t)J_0')(\hat{H}^*(t)-H^*(t)) \\ & +(\hat{h}^*_{t+\tau}-h^*_{t+\tau})'(Q_0(t)-Q_j(t))(\hat{H}^*(t)-H^*(t)) \\ & -(\hat{H}^*(t)-H^*(t))'(-J_0B_0(t)J_0'x_tx_t'J_0B_0(t)J_0'+J_0B_0(t)J_0'x_tx_t'J_jB_j(t)J_j')(\hat{H}^*(t)-H^*(t))\}. \end{aligned} \tag{14}$$

Note that there are 3 bracketed $\{.\}$ terms in (14). We will show that the first of these has as its limit $\Gamma^*_{1,j}$ where $\Gamma^*_{1,j} =^{d^*} \Gamma_{1,j}$, for $\Gamma_{1,j}$ defined in the text. We then show that the remaining two bracketed terms are $o_{p^*}(1)$.

Proof of bracket 1: This term is identical to the first component of the first bracketed term in the proof of Theorem 3.3 (a) (multiplied by 1/2) and hence the result is immediate.

Proof of bracket 2: We must show that this term is $o_{p^*}(1)$. Note however, that this term is nearly identical to that in equation (13) from the proof of Theorem 3.3 (a) and hence nearly identical arguments show that

$$\begin{aligned} & \textstyle\sum_{t=T}^{T+P-\tau}H'^*(t)(-J_0B_0(t)J_0'x_tx_t'J_0B_0(t)J_0'+J_0B_0(t)J_0'x_tx_t'J_jB_j(t)J_j')H^*(t) \\ & =\textstyle\sum_{t=T}^{T+P-\tau}H'^*(t)(-J_0B_0J_0'B^{-1}J_0B_0J_0'+J_0B_0J_0'B^{-1}J_jB_jJ_j')H^*(t)+o_{p^*}(1). \end{aligned}$$

The result then follows since $-J_0B_0J_0'B^{-1}J_0B_0J_0'+J_0B_0J_0'B^{-1}J_jB_jJ_j'=0$.

Proof of bracket 3: We must show that each of the six components of the third bracketed term in (14) are $o_{p^*}(1)$. However, these terms are nearly identical to those in the second bracketed term from the proof of Theorem 3.3 (a) and hence the proofs are omitted for brevity.

Proof of Theorem 3.4: In parts (a) and (b) below we focus on the asymptotic distributions for a fixed pairwise comparison between models 0 and j. That we obtain the appropriate joint distribution

of all M of the pairwise test statistics follows from the rank condition on Ω in Assumption $2'$, Lemma 1(b), and the fact that for each j, each of the statistics have asymptotic representations as functionals of the same standard Brownian motion $W^*(s)$. The result then follows from an application of the Continuous Mapping Theorem to the `max` functional.

(a) Given Theorem 3.3 (a) and the Continuous Mapping Theorem it suffices to show that, for each j, $P\sum_{l=-\bar{l}}^{\bar{l}} K(l/M)\hat{\gamma}^*_{dd,j}(l) \to^{d^*} 4\sigma^4\Gamma^*_{3,j}$ where $\Gamma^*_{3,j} =^{d^*} \Gamma_{3,j}$ for $\Gamma_{3,j}$ defined in the text. Before doing so it is convenient to redefine the two bracketed terms terms from (13) used in the main decomposition of the loss differential in Theorem 3.3(a) (absent the summations, but keeping the brackets) as

$$(\hat{u}^{*2}_{0,t+\tau} - \hat{u}^{*2}_{j,t+\tau}) = \{2A^*_{1,t} - A^*_{2,t}\} + 2\{B^*_{1,t} + B^*_{2,t} + B^*_{3,t} + B^*_{4,t} + B^*_{5,t}\}.$$

With this in mind, if we ignore the finite sample difference between P and $P-\tau+1$, we obtain

$$\begin{array}{l} P\sum_{l=-\bar{l}}^{\bar{l}} K(l/M)\hat{\gamma}^*_{dd,j}(l) = \sum_{l=-\bar{l}}^{\bar{l}} K(l/M)\sum_{t=T+l}^{T+P-\tau}(\hat{u}^{*2}_{0,t+\tau} - \hat{u}^{*2}_{j,t+\tau})(\hat{u}^{*2}_{0,t-l+\tau} - \hat{u}^{*2}_{j,t-l+\tau}) \\ = 4\{\sum_{l=-\bar{l}}^{\bar{l}} K(l/M)\sum_{t=T+l}^{T+P-\tau} A^*_{1,t}A^*_{1,t-l}\} \\ +4\sum_{l=-\bar{l}}^{\bar{l}} K(l/M)\sum_{t=T+l}^{T+P-\tau}\{ \text{ other cross products of } A^*_{1,t}, A^*_{2,t}, B^*_{1,t}, B^*_{2,t}, B^*_{3,t}, B^*_{4,t}, B^*_{5,t}, \\ \text{with } A^*_{1,t-l}, A^*_{2,t-l}, B^*_{1,t-l}, B^*_{2,t-l}, B^*_{3,t-l}, B^*_{4,t-l}, B^*_{5,t-l}\}. \end{array} \tag{15}$$

In the remainder we show that the bracketed term converges to σ^4 times $\Gamma^*_{3,j} =^{d^*} \Gamma_{3,j}$ and that each of the cross product terms are each $o_{p^*}(1)$.

Proof of bracket 1: Straightforward algebra implies that

$$\begin{array}{l} \sum_{l=-\bar{l}}^{\bar{l}} K(l/M)\sum_{t=T+l}^{T+P-\tau} A^*_{1,t}A^*_{1,t-l} = \\ \sigma^4\sum_{l=-\bar{l}}^{\bar{l}} K(l/M)\sum_{t=T+l}^{T+P-\tau}(T^{1/2}\tilde{H}^{*\prime}_j(t))E^*\tilde{h}^*_{j,t+\tau}\tilde{h}^{*\prime}_{j,t-l+\tau}(T^{1/2}\tilde{H}^*_j(t-l)) \\ +\sigma^4\sum_{l=-\bar{l}}^{\bar{l}} K(l/M)\sum_{t=T+l}^{T+P-\tau}\{H^{*\prime}(t)((Q_0(t)-Q_j(t)) - (Q_0-Q_j))E^*h^*_{t+\tau}h^{*\prime}_{t-l+\tau}(Q_0-Q_j)H^*(t-l) \\ +H^{*\prime}(t)((Q_0(t)-Q_j(t)) - (Q_0-Q_j))(h^*_{t+\tau}h^{\prime *}_{t-l+\tau} - E^*h^*_{t+\tau}h^{*\prime}_{t-l+\tau})(Q_0-Q_j)H^*(t-l) \\ +H^{*\prime}(t)((Q_0(t)-Q_j(t)) - (Q_0-Q_j))(h^*_{t+\tau}h^{\prime *}_{t-l+\tau} - E^*h^*_{t+\tau}h^{*\prime}_{t-l+\tau})((Q_0(t-l)-Q_j(t-l)) - (Q_0-Q_j))H^*(t-l) \\ +H^{*\prime}(t)(Q_0-Q_j)(h^*_{t+\tau}h^{\prime *}_{t-l+\tau} - E^*h^*_{t+\tau}h^{*\prime}_{t-l+\tau})(Q_0-Q_j)H^*(t-l) \\ +H^{*\prime}(t)(Q_0-Q_j)(h^*_{t+\tau}h^{\prime *}_{t-l+\tau} - E^*h^*_{t+\tau}h^{*\prime}_{t-l+\tau})((Q_0(t-l)-Q_j(t-l)) - (Q_0-Q_j))H^*(t-l) \\ +H^{*\prime}(t)(Q_0-Q_j)E^*h^*_{t+\tau}h^{*\prime}_{t-l+\tau}((Q_0(t-l)-Q_j(t-l)) - (Q_0-Q_j))H^*(t-l) \\ +H^{*\prime}(t)((Q_0(t-l)-Q_j(t-l)) - (Q_0-Q_j))E^*h^*_{t+\tau}h^{*\prime}_{t-l+\tau}((Q_0(t-l)-Q_j(t-l)) - (Q_0-Q_j))H^*(t-l)\}, \end{array}$$

where $\tilde{H}^*(t)$ is the bootstrap equivalent of $\tilde{H}(t)$ defined in section 3.1. Since l is finite, the Continuous Mapping Theorem and Lemma 1(b) imply

$$\sigma^4\sum_{l=-\bar{l}}^{\bar{l}} K(l/M)\sum_{t=T+l}^{T+P-\tau}(T^{1/2}\tilde{H}^{\prime *}_j(t))E^*\tilde{h}^*_{j,t+\tau}\tilde{h}^{*\prime}_{j,t-l+\tau}(T^{1/2}\tilde{H}^*_j(t-l)) \to^{d^*} \sigma^4\Gamma^*_{3,j}.$$

We must now show that each element of the second bracketed right-hand side term in (15) is $o_{p^*}(1)$. The proof of each is similar and as such we provide the result for the first and fourth elements.

For the first, after taking the absolute value we obtain

$$\begin{array}{ll} & |\sum_{l=-\bar{l}}^{\bar{l}} K(l/M)\sum_{t=T+l}^{T+P-\tau} H^{*\prime}(t)((Q_0(t) - Q_j\ (t)) - (Q_0\ -Q_j\ ))E^*\ h^*_{t+\tau}\ h^{*\prime}_{t-l+\tau}\ (Q_0\ -Q_j)H^*\ (t-l)| \\ \leq & 2T^{-1/2}\ \bar{l}k^4\ (\sup_t T^{1/2}\ |H^*(t)|)^2(|Q_0\ -Q_j\ |)(\sup_t\ T^{1/2}\ |(Q_0\ (t) - Q_j\ (t)) - (Q_0\ -Q_j\ )|)\times \\ & (\max_{|l|\leq\bar{l}}\ (T^{-1}\sum_{t=T+l}^{T+P-\tau}\ |E^*\ h^*_{t+\tau}\ h^{*\prime}_{t-l+\tau}\ |)). \end{array}$$

Lemmas 1(a) and (d) imply both $\sup_t T^{1/2}|H^*(t)| = O_{p^*}(1)$ and $\sup_t T^{1/2}|(Q_0(t)-Q_j(t))-(Q_0-Q_j)| = O_p(1)$. Since for each $|l| \leq \bar{l}$, Assumption 2′ is sufficient for $T^{-1}\sum_{t=T+l}^{T+P-\tau}|E^*h^*_{t+\tau}h^{*\prime}_{t-l+\tau}|$ to be $O_p(1)$ the result follows from the fact that $T^{-1/2} = o(1)$.

For the fourth term note that after rearranging terms we obtain

$$\begin{aligned}
&\sum_{l=-\bar{l}}^{\bar{l}} K(l/M)\sum_{t=T+l}^{T+P-\tau} H^{*\prime}(t)(Q_0-Q_j)(h^*_{t+\tau}h^{\prime *}_{t-l+\tau} - E^*h^*_{t+\tau}h^{*\prime}_{t-l+\tau})(Q_0-Q_j)H^*(t-l)\\
= \;& T^{-1/2}\sum_{l=-\bar{l}}^{\bar{l}} K(l/M)\sum_{t=T+l}^{T+P-\tau}(T^{1/2}H^{*\prime}(t-l)(Q_0-Q_j)\otimes T^{1/2}H^{*\prime}(t)(Q_0-Q_j))\times\\
&(T^{-1/2}vec(h^*_{t+\tau}h^{\prime *}_{t-l+\tau} - E^*h^*_{t+\tau}h^{*\prime}_{t-l+\tau})).
\end{aligned}$$

Recall that by Lemma 1 (b), $T^{1/2}H^*(t) \Rightarrow^* \Omega_{11}^{1/2}W^*(s)$. Moreover, note that conditional on the observables, $h^*_{t+\tau}h^{\prime *}_{t-l+\tau} - E^*h^*_{t+\tau}h^{*\prime}_{t-l+\tau}$ forms a heteroskedastic L^2-bounded $MA(\tau-1)$ process with increments that are uncorrelated with $H^*(t)$ and $H^*(t-l)$. Hence conditional on the observables and for each $|l| \leq \bar{l}$, Theorem 30.14 of Davidson (1994) suffices for $\sum_{t=T+l}^{T+P-\tau}(T^{1/2}H^{*\prime}(t-l)(Q_0 - Q_j)\otimes T^{1/2}H^{*\prime}(t)(Q_0-Q_j))(T^{-1/2}vec(h^*_{t+\tau}h^{\prime *}_{t-l+\tau} - E^*h^*_{t+\tau}h^{*\prime}_{t-l+\tau}) = O_{p^*}(1)$. The result follows since $T^{-1/2} = o(1)$ and $K(x) < 1$ for all x.

Proof of bracket 2: We must show each of the remaining cross-products of $A^*_{1,t}$, $A^*_{2,t}$, and $B^*_{i,t}$ with $A^*_{1,t-l}$, $A^*_{2,t-l}$, and $B^*_{i,t-l}$ (all $i = 1, ..., 5$) in (15) are $o_{p^*}(1)$. The proofs of each are similar and as such we only show the result for that associated with the cross-product of $A^*_{1,t}$ and $B^*_{4,t-l}$. If we take the absolute value of this term we find that

$$\begin{aligned}
&|\textstyle\sum_{l=-\bar{l}}^{\bar{l}} K(l/M)\sum_{t=T}^{T+P-\tau} h^{\prime *}_{t+\tau}(Q_0(t)-Q_j(t))H^*(t)(\hat{h}^*_{t-l+\tau} - h^*_{t-l+\tau})'\times\\
&(Q_0(t-l)-Q_j(t-l))(\hat{H}^*(t-l) - H^*(t-l))|\\
&\leq 2\bar{l}k^5(T^{-1}\textstyle\sum_{t=T}^{T+P-\tau}|\hat{h}^*_{t+\tau} - h^*_{t+\tau}||h^*_{t+\tau}|)(\sup_t|Q_0(t)-Q_j(t)|)^2(\sup_t T^{1/2}|\hat{H}^*(t) - H^*(t)|)\times\\
&(\sup_t T^{1/2}|H^*(t)|).
\end{aligned}$$

Lemma 1(e) implies $\sup_t T^{1/2}|\hat{H}^*(t) - H^*(t)| = o_{p^*}(1)$ while Lemmas 1(a) and (d) imply both $\sup_t|Q_0(t) - Q_j(t)| = O_p(1)$ and $(\sup_t T^{1/2}|H^*(t)|) = O_{p^*}(1)$.

The result will follow if $T^{-1/2}\sum_{t=T}^{T+P-\tau}|\hat{h}^*_{t+\tau} - h^*_{t+\tau}||h^*_{t+\tau}| = O_{p^*}(1)$. For simplicity we assume, as in the proof of Lemma 1(e), that $\tau = 2$ and hence the forecast errors form an $MA(1)$ process. If we then take a Taylor expansion in precisely the same fashion as in the proof of Lemma 1(e) we have

$$\begin{aligned}
&T^{-1/2}\sum_{t=T}^{T+P-\tau}|\hat{h}^*_{t+\tau} - h^*_{t+\tau}||h^*_{t+\tau}|\\
\leq\;& (k+1)T^{-1}\sum_{t=T}^{T+P-\tau}|\eta_{t+2}x_t\nabla\widehat{\varepsilon}_{t+2}(\overline{\gamma}_{t+2})||h^*_{t+\tau}||T^{1/2}(\widehat{\gamma}-\gamma)|\\
&+|\theta|(k+1)T^{-1}\sum_{t=T}^{T+P-\tau}|\eta_{t+1}x_t\nabla\widehat{\varepsilon}_{t+1}(\overline{\gamma}_{t+1})||h^*_{t+\tau}||T^{1/2}(\widehat{\gamma}-\gamma)|\\
&+|\widehat{\theta}-\theta|(k+1)T^{-1}\sum_{t=T}^{T+P-\tau}|\eta_{t+1}x_t\nabla\widehat{\varepsilon}_{t+1}(\overline{\gamma}_{t+1})||h^*_{t+\tau}||T^{1/2}(\widehat{\gamma}-\gamma)|\\
&+|T^{1/2}(\widehat{\theta}-\theta)|T^{-1}\sum_{t=T}^{T+P-\tau}|\eta_{t+1}x_t\varepsilon_{t+1}||h^*_{t+\tau}|.
\end{aligned}$$

Assumptions 1 and 2′ suffice for both $T^{1/2}(\widehat{\gamma}-\gamma)$ and $T^{1/2}(\widehat{\theta}-\theta)$ to be $O_p(1)$. Since, for large enough samples, Assumption 2′ bounds the second moments of $\nabla\widehat{\varepsilon}_{t+2}(\overline{\gamma}_{t+2})$, $\nabla\widehat{\varepsilon}_{t+1}(\overline{\gamma}_{t+1})$, and x_t, that $\eta_{s+\tau}$ is distributed *i.i.d.* $N(0,1)$ implies that $T^{-1}\sum_{t=T}^{T+P-\tau}|\eta_{t+2}x_t\nabla\widehat{\varepsilon}_{t+2}(\overline{\gamma}_{t+2})||h^*_{t+2}|$, $T^{-1}\sum_{t=T}^{T+P-\tau}|\eta_{t+1}x_t\nabla\widehat{\varepsilon}_{t+1}(\overline{\gamma}_{t+1})||h^*_{t+2}|$, and $T^{-1}\sum_{t=T}^{T+P-\tau}|\eta_{t+1}x_t\varepsilon_{t+1}||h^*_{t+2}|$ are all $O_{p^*}(1)$, and the proof is complete.

(b) Given Theorem 3.3 (b) and the Continuous Mapping Theorem it suffices to show that, for each j, $P\sum_{l=-\bar{l}}^{\bar{l}}K(l/M)\hat{\gamma}^*_{cc,j}(l)\rightarrow^{d^*}\sigma_u^4\Gamma^*_{3,j}$ where $\Gamma^*_{3,j}=^{d^*}\Gamma_{3,j}$ for $\Gamma_{3,j}$ defined in the text. Before doing so it is convenient to redefine the three bracketed terms terms from (14) used in the main decomposition of the product of the forecast error from the baseline model with the difference in the baseline and model j forecast errors in Theorem 3.3 (b) (absent the summations, but keeping the brackets) as

$$\hat{u}^*_{0,t+\tau}(\hat{u}^*_{0,t+\tau}-\hat{u}^*_{j,t+\tau})=\{A^*_{1,t}\}+\{B^*_t\}+\{C^*_{1,t}+C^*_{2,t}+C^*_{3,t}+C^*_{4,t}+C^*_{5,t}+C^*_{6,t}\}.$$

With this in mind, if we ignore the finite sample difference between P and $P-\tau+1$, we obtain

$$\begin{array}{l}
P\sum_{l=-\bar{l}}^{\bar{l}}K(l/M)\hat{\gamma}^*_{cc,j}(l)\\
=\sum_{l=-\bar{l}}^{\bar{l}}K(l/M)\sum_{t=T+l}^{T+P-\tau}\hat{u}^*_{0,t+\tau}(\hat{u}^*_{0,t+\tau}-\hat{u}^*_{j,t+\tau})\hat{u}^*_{0,t-l+\tau}(\hat{u}^*_{0,t-l+\tau}-\hat{u}^*_{j,t-l+\tau})\\
=\{\sum_{l=-\bar{l}}^{\bar{l}}K(l/M)\sum_{t=T+l}^{T+P-\tau}A^*_{1,t}A^*_{1,t-l}\}\\
+\sum_{l=-\bar{l}}^{\bar{l}}K(l/M)\sum_{t=T+l}^{T+P-\tau}\{\text{ other cross products of } A^*_{1,t},B^*_t,C^*_{1,t},C^*_{2,t},C^*_{3,t},C^*_{4,t},C^*_{5,t},C^*_{6,t},\\
\text{with } A^*_{1,t-l},B^*_{t-l},C^*_{1,t-l},C^*_{2,t-l},C^*_{3,t-l},C^*_{4,t-l},C^*_{5,t-l},C^*_{6,t-l}\}.
\end{array}\tag{16}$$

In the remainder we show that the bracketed term converges to σ^4 times $\Gamma^*_{3,j}=^{d^*}\Gamma_{3,j}$ and that each of the cross product terms are each $o_{p^*}(1)$.

Proof of bracket 1: This term is identical to that in the proof of Theorem 3.4 (a) and hence the result is immediate.

Proof of bracket 2: We must show each of the remaining cross-products of $A^*_{1,t}$, B^*_t, and $C^*_{i,t}$ with $A^*_{1,t-l}$, B^*_{t-l}, and $C^*_{i,t-l}$ (all $i=1,...,6$) in (16) are $o_{p^*}(1)$. Nearly all of these cross products are identical to those from the proof of Theorem 3.4 (a). The only ones that are distinct are those that contain B^*_t (or B^*_{t-l}). As such we will only show the result for the cross product of B^*_t with $A^*_{1,t-l}$. If we take the absolute value of this term we find that

$$\begin{array}{l}
|\sum_{l=-\bar{l}}^{\bar{l}}K(l/M)\sum_{t=T+l}^{T+P-\tau}h'^*_{t-l+\tau}(Q_0(t-l)-Q_j(t-l))H^*(t-l)\times\\
H'^*(t)(-J_0B_0(t)J_0'x_tx_t'J_0B_0(t)J_0'+J_0B_0(t)J_0'x_tx_t'J_jB_j(t)J_j')H^*(t)|\\
\leq(T^{-1/2})2\bar{l}k^4(\sup_t T^{1/2}|H^*(t)|)^3(\sup_t|Q_0(t)-Q_j(t)|)\times\\
(T^{-1}\sum_{t=T}^{T+P-\tau}|h^*_{t-l+\tau}||-J_0B_0(t)J_0'x_tx_t'J_0B_0(t)J_0'+J_0B_0(t)J_0'x_tx_t'J_jB_j(t)J_j'|).
\end{array}$$

Since Assumption 2′ suffices for

$$T^{-1}\sum_{t=T}^{T+P-\tau}|h^*_{t-l+\tau}||-J_0B_0(t)J_0'x_tx_t'J_0B_0(t)J_0'+J_0B_0(t)J_0'x_tx_t'J_jB_j(t)J_j'|=O_{p^*}(1)$$

and Lemmas 1(a) and (d) imply both $\sup_t|Q_0(t)-Q_j(t)|=O_p(1)$ and $\sup_t T^{1/2}|H^*(t)|=O_{p^*}(1)$ the result follows from the fact that $T^{-1/2}=o(1)$.

8 Appendix 2: Data

All data were obtained from the FAME database of the Federal Reserve Board of Governors. The tables below describe each series, in some cases using the acronym PCE to denote personal consumption expenditures. As indicated in the text, the data from the commodity price application are monthly, with the commodity price series constructed as averages of weekly data (from Tuesday of each week) and the exchange rates obtained as monthly averages of daily data. In the case of the GDP application, the series on GDP, nominal PCE, nominal GDP, and the PCE price index ex food and energy are source quarterly data; for all other series, the quarterly data were constructed as within-quarter averages of source monthly data. The transformations to changes or growth rates were applied to the quarterly levels.

Data for commodity price application

variable	description
CRB spot	CRB index of spot prices, raw industrials
CRB futures	CRB index of futures prices, raw industrials
XR-AUS	spot exchange rate, Australia (U.S. $)
XR-CAN	spot exchange rate, Canada (U.S. $)
XR-JAP	spot exchange rate, Japan (U.S. $)
XR-NZ	spot exchange rate, New Zealand (U.S. $)
XR-UK	spot exchange rate, U.K. (U.S. $)
XR major	major currencies dollar index
XR other	other important trading partners dollar index

Data for GDP application

variable	description
GDP	GDP, chain dollar
C/Y	nominal PCE/nominal GDP
Hours	average weekly hours of production workers in manufacturing
Unemp. claims	initial claims for unemployment insurance
Permits	new privately-owned housing units authorized, single-family
PMI orders	Purchasing Managers Index (manuf.) of supplier deliveries
PMI deliveries	Purchasing Managers Index (manuf.) of new orders
S&P 500	S&P index of 500 common stocks/PCE price index ex food and energy
3-month Treasury	3-month Treasury bill rate (secondary market)
1-year Treasury	Yield on U.S. Treasury securities, 1-year constant maturity
10-year Treasury	Yield on U.S. Treasury securities, 10-year constant maturity
AAA	Moody's yield on Aaa corporate bonds
BAA	Moody's yield on Baa corporate bonds

References

Billmeier, Andreas (2009), "Ghostbusting: Which Output Gap Measure Really Matters?" *International Economics and Economic Policy* 6, 391-419.

Bruneau, C., O. De Bandt, A. Flageollet, and E. Michaux (2007), "Forecasting Inflation Using Economic Indicators: The Case of France," *Journal of Forecasting* 2, 1-22.

Butler, Alexander W., Gustavo Grullon, and James P. Weston (2005), "Can Managers Forecast Aggregate Market Returns?" *Journal of Finance* 60, 963-986.

Cheung, Yin-Wong, Menzie D. Chinn, and Antonio Garcia Pascual (2005), "Empirical Exchange Rate Models of the Nineties: Are any Fit to Survive?" *Journal of International Money and Finance* 24, 1150-1175.

Chen, Yu-Chin, Kenneth Rogoff, and Barbara Rossi (2010), "Can Exchange Rates Forecast Commodity Prices?" *Quarterly Journal of Economics* 125, 1145-1194.

Clark, Todd E., and Michael W. McCracken (2001), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics* 105, 85-110.

Clark, Todd E., and Michael W. McCracken (2005), "Evaluating Direct Multistep Forecasts," *Econometric Reviews* 24, 369-404.

Clark, Todd E., and Michael W. McCracken (2009), "Nested Forecast Model Comparisons: A New Approach to Testing Equal Accuracy," manuscript, Federal Reserve Bank of St. Louis.

Clark, Todd E., and Kenneth D. West (2007), "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models," *Journal of Econometrics* 138, 291-311.

Cooper, Michael, and Huseyin Gulen (2006), "Is Time-Series-Based Predictability Evident in Real Time?" *Journal of Business* 79, 1263-1292.

Corradi, Valentina, and Norman R. Swanson (2007), "Nonparametric Bootstrap Procedures for Predictive Inference Based on Recursive Estimation Schemes," *International Economic Review* 48, 67-109.

Davidson, Russell (1994), *Stochastic Limit Theory*, New York: Oxford University Press.

Denton, Frank T. (1985), "Data Mining as an Industry," *Review of Economics and Statistics* 67, 124-127.

Diebold, Francis X., and Roberto S. Mariano (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13, 253-263.

Giacomini, Rafaella, and Halbert White (2006), "Tests of Conditional Predictive Ability," *Econometrica* 74, 1545-1578.

Goncalves, Sylvia, and Lutz Kilian (2004), "Bootstrapping Autoregressions with Conditional Heteroskedasticity of Unknown Form," *Journal of Econometrics* 123, 89-120.

Goyal, Amit, and Ivo Welch (2003), "Predicting the Equity Premium with Dividend Ratios," *Management Science* 49, 639-654.

Goyal, Amit, and Ivo Welch (2008), "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction, *Review of Financial Studies* 21, 1455-1508.

Groen, Jan J. (1999), "Long Horizon Predictability of Exchange Rates: Is it for Real?" *Empirical Economics* 24, 451-469.

Guo, Hui (2006), "On the Out-of-Sample Predictability of Stock Market Returns," *Journal of Business* 27, 645-670.

Hansen, Bruce E. (1992), "Convergence to Stochastic Integrals for Dependent Heterogeneous Processes, *Econometric Theory* 8, 489-500.

Hansen, Bruce E. (1996), "Erratum: The Likelihood Ratio Test Under Nonstandard Conditions: Testing the Markov Switching Model of GNP," *Journal of Applied Econometrics* 11, 195-198.

Hansen, Peter R. (2005), "A Test for Superior Predictive Ability," *Journal of Business and Economic Statistics* 23, 365-380.

Harvey, David I., Stephen J. Leybourne, and Paul Newbold (1998), "Tests for Forecast Encompassing," *Journal of Business and Economic Statistics* 16, 254-259.

Hendry David F., and Kirstin Hubrich (2009), "Combining Disaggregate Forecasts or Combining Disaggregate Information to Forecast an Aggregate," *Journal of Business and Economic Statistics*, forthcoming.

Hong, Yongmiao, and Tae-Hwy Lee (2003), "Inference on Predictability of Foreign Exchange Rates via Generalized Spectrum and Nonlinear Time Series Models," *Review of Economics and Statistics* 85, 1048-1062.

Hoover, Kevin D., and Stephen J. Perez (1999), "Data Mining Reconsidered: Encompassing and the General-to-Specific Approach to Specification Search," *Econometrics Journal* 2, 167-191.

Hubrich Kirstin (2005), "Forecasting Euro Area Inflation: Does Aggregating Forecasts by HICP Component Improve Forecast Accuracy?" *International Journal of Forecasting* 21, 119-136.

Hubrich, Kirstin, and Kenneth D. West (2010), "Forecast Evaluation of Small Nested Model Sets," *Journal of Applied Econometrics* 25, 574-594.

Inoue, Atsushi, and Lutz Kilian (2004), "In–Sample or Out–of–Sample Tests of Predictability? Which One Should We Use?" *Econometric Reviews* 23, 371-402.

Kilian, Lutz (1999), "Exchange Rates and Monetary Fundamentals: What Do We Learn from Long-Horizon Regressions?" *Journal of Applied Econometrics* 14, 491-510.

Kilian, Lutz, and Mark P. Taylor (2003), "Why Is it So Difficult to Beat the Random Walk Forecast of Exchange Rates?" *Journal of International Economics* 60, 85-107.

Lo, Andrew W., and A. Craig MacKinlay (1990), "Data-Snooping Biases in Tests of Financial Asset Pricing Models," *Review of Financial Studies* 3, 431-467.

Mark, Nelson C. (1995), "Exchange Rates and Fundamentals: Evidence on Long-Horizon

Predictability," *American Economic Review* 85, 201-218.

McCracken, Michael W. (2007), "Asymptotics for Out-of-Sample Tests of Granger Causality," *Journal of Econometrics* 140, 719-752.

Meese, Richard, and Kenneth Rogoff (1988), "Was it Real? The Exchange Rate–Interest Differential Relation Over the Modern Floating-Rate Period," *Journal of Finance* 43, 933-948.

Molodtsova, Tanya, and David H. Papell (2009), "Out-of-Sample Exchange Rate Predictability with Taylor Rule Fundamentals," *Journal of International Economics* 77, 167-180.

Moench, Emanuel (2008), "Forecasting the Yield Curve in a Data-Rich Environment: A No-Arbitrage Factor-Augmented VAR Approach," *Journal of Econometrics* 146, 26-43.

Newey, Whitney K., and Kenneth D. West (1987), "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55, 703-708.

Politis, Dimitris N., and Joseph P. Romano (1994), "The Stationary Bootstrap," *Journal of the American Statistical Association* 89, 1303-1313.

Rapach, David, and Jack K. Strauss (2007), "Bagging or Combining (or Both)? An Analysis Based on Forecasting U.S. Employment Growth," *Econometric Reviews*, forthcoming.

Rapach, David, and Mark E. Wohar (2006), "In-Sample vs. Out-of-Sample Tests of Stock Return Predictability in the Context of Data Mining," *Journal of Empirical Finance* 13, 231-247.

Sarno, Lucio, Daniel L. Thornton, and Giorgio Valente (2005), "Federal Funds Rate Prediction," *Journal of Money, Credit and Banking* 37, 449-472.

Stambaugh, Robert F. (1999), "Predictive Regressions," *Journal of Financial Economics* 54, 375-421.

Stock, James H., and Mark W. Watson (1999), "Forecasting Inflation," *Journal of Monetary Economics* 44, 293-335.

Stock, James H., and Mark W. Watson (2003), "Forecasting Output and Inflation: The Role of Asset Prices," *Journal of Economic Literature* 41, 788-829.

Storey, John D. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society*, Series B, 64, 479-498.

West, Kenneth D. (1996), "Asymptotic Inference About Predictive Ability," *Econometrica* 64, 1067-1084.

West, Kenneth D. (2001), "Tests for Forecast Encompassing When Forecasts Depend on Estimated Regression Parameters," *Journal of Business and Economic Statistics* 19, 29-33.

White, Halbert (2000), "A Reality Check For Data Snooping," *Econometrica* 68, 1097-1127.

Table 1: Monte Carlo Results on Size, 1-Step Horizon

(nominal size = 10%)

statistic	*source of critical values*	T=40 P=80	T=40 P=120	T=80 P=40	T=80 P=80	T=80 P=120	T=120 P=40	T=120 P=80
DGP 1								
MSE-F	fixed regressor	0.125	0.121	0.105	0.120	0.115	0.118	0.113
MSE-t	fixed regressor	0.108	0.110	0.113	0.111	0.104	0.123	0.112
ENC-F	fixed regressor	0.118	0.105	0.093	0.115	0.104	0.104	0.104
ENC-t	fixed regressor	0.099	0.104	0.097	0.100	0.093	0.104	0.104
MSE-F	non-parametric	0.002	0.000	0.012	0.004	0.005	0.023	0.009
MSE-t	non-parametric	0.001	0.000	0.004	0.005	0.002	0.009	0.005
SPA	non-parametric	0.005	0.000	0.025	0.009	0.005	0.044	0.013
DGP 2								
statistic	*source of critical values*	T=40 P=80	T=40 P=120	T=80 P=40	T=80 P=80	T=80 P=120	T=120 P=40	T=120 P=80
MSE-F	fixed regressor	0.161	0.145	0.125	0.159	0.148	0.120	0.126
MSE-t	fixed regressor	0.148	0.142	0.167	0.146	0.140	0.159	0.147
ENC-F	fixed regressor	0.138	0.147	0.102	0.125	0.132	0.100	0.109
ENC-t	fixed regressor	0.134	0.126	0.140	0.138	0.128	0.137	0.116
MSE-F	non-parametric	0.009	0.007	0.022	0.023	0.014	0.042	0.028
MSE-t	non-parametric	0.001	0.002	0.004	0.004	0.004	0.012	0.005
SPA	non-parametric	0.011	0.005	0.058	0.024	0.014	0.085	0.033
ENC-t	Hubrich-West	0.096	0.081	0.121	0.093	0.084	0.116	0.085
DGP 3								
statistic	*source of critical values*	T=40 P=80	T=40 P=120	T=80 P=40	T=80 P=80	T=80 P=120	T=120 P=40	T=120 P=80
MSE-F	fixed regressor	0.100	0.113	0.102	0.114	0.101	0.096	0.094
MSE-t	fixed regressor	0.096	0.102	0.123	0.097	0.091	0.121	0.103
ENC-F	fixed regressor	0.095	0.101	0.079	0.099	0.112	0.089	0.081
ENC-t	fixed regressor	0.090	0.096	0.100	0.097	0.089	0.107	0.095
MSE-F	non-parametric	0.011	0.008	0.021	0.011	0.009	0.035	0.017
MSE-t	non-parametric	0.000	0.000	0.004	0.002	0.002	0.005	0.002
SPA	non-parametric	0.003	0.001	0.044	0.009	0.008	0.048	0.016
ENC-t	Hubrich-West	0.066	0.067	0.092	0.072	0.054	0.098	0.072

Notes:
1. The data generating processes are defined in equations (5) and (8). In all of these experiments, the coefficient b in the DGPs is set to 0, such that, in population, all of the models are equally accurate.
2. For each artificial data set, forecasts of $y_{t+\tau}$ (where τ denotes the forecast horizon) are formed recursively using estimates of the forecasting equations described in section 4.2. These forecasts are then used to form the indicated test statistics, given in section 2.2. T and P refer to the number of in–sample observations and 1-step ahead forecasts, respectively.
3. In each Monte Carlo replication, the simulated test statistics are compared against bootstrapped critical values, using a significance level of 10%. Sections 2.3 and 4.1 describe the bootstrap procedures.
4. The number of Monte Carlo simulations is 2000; the number of bootstrap draws is 499.

Table 2: Monte Carlo Results on Size, 4-Step Horizon

(nominal size = 10%)

DGP 1								
	source of	T=40	T=40	T=80	T=80	T=80	T=120	T=120
statistic	*critical values*	P=80	P=120	P=40	P=80	P=120	P=40	P=80
MSE-F	fixed regressor	0.135	0.115	0.119	0.135	0.119	0.116	0.104
MSE-t	fixed regressor	0.115	0.113	0.133	0.120	0.114	0.134	0.096
ENC-F	fixed regressor	0.127	0.124	0.119	0.120	0.108	0.117	0.107
ENC-t	fixed regressor	0.115	0.111	0.123	0.109	0.103	0.118	0.094
MSE-F	non-parametric	0.012	0.002	0.061	0.019	0.009	0.072	0.025
MSE-t	non-parametric	0.002	0.001	0.007	0.001	0.001	0.014	0.003
SPA	non-parametric	0.027	0.009	0.190	0.056	0.019	0.221	0.054
DGP 2								
	source of	T=40	T=40	T=80	T=80	T=80	T=120	T=120
statistic	*critical values*	P=80	P=120	P=40	P=80	P=120	P=40	P=80
MSE-F	fixed regressor	0.208	0.199	0.201	0.181	0.178	0.144	0.166
MSE-t	fixed regressor	0.188	0.175	0.167	0.163	0.176	0.147	0.171
ENC-F	fixed regressor	0.162	0.145	0.157	0.144	0.146	0.115	0.136
ENC-t	fixed regressor	0.164	0.163	0.150	0.143	0.168	0.137	0.158
MSE-F	non-parametric	0.054	0.030	0.130	0.066	0.050	0.127	0.074
MSE-t	non-parametric	0.000	0.001	0.005	0.002	0.001	0.004	0.004
SPA	non-parametric	0.110	0.052	0.334	0.137	0.081	0.348	0.177
ENC-t	Hubrich-West	0.290	0.227	0.423	0.261	0.230	0.413	0.276
DGP 3								
	source of	T=40	T=40	T=80	T=80	T=80	T=120	T=120
statistic	*critical values*	P=80	P=120	P=40	P=80	P=120	P=40	P=80
MSE-F	fixed regressor	0.115	0.114	0.120	0.125	0.104	0.103	0.110
MSE-t	fixed regressor	0.096	0.116	0.118	0.100	0.109	0.114	0.107
ENC-F	fixed regressor	0.106	0.109	0.114	0.114	0.107	0.086	0.103
ENC-t	fixed regressor	0.093	0.101	0.104	0.089	0.099	0.105	0.098
MSE-F	non-parametric	0.030	0.018	0.088	0.048	0.026	0.098	0.059
MSE-t	non-parametric	0.000	0.000	0.004	0.001	0.001	0.002	0.001
SPA	non-parametric	0.044	0.019	0.258	0.077	0.039	0.288	0.113
ENC-t	Hubrich-West	0.207	0.167	0.335	0.193	0.169	0.356	0.220

Notes:
1. See the notes to Table 1.

Table 3: Monte Carlo Results on Power, 1-Step Horizon

(nominal size = 10%)

DGP 1								
	source of	T=40	T=40	T=80	T=80	T=80	T=120	T=120
statistic	*critical values*	P=80	P=120	P=40	P=80	P=120	P=40	P=80
MSE-F	fixed regressor	0.613	0.748	0.485	0.670	0.789	0.531	0.714
MSE-t	fixed regressor	0.438	0.606	0.292	0.433	0.573	0.279	0.421
ENC-F	fixed regressor	0.643	0.768	0.547	0.762	0.860	0.634	0.829
ENC-t	fixed regressor	0.579	0.738	0.367	0.621	0.798	0.383	0.645
MSE-F	non-parametric	0.056	0.102	0.097	0.141	0.214	0.141	0.215
MSE-t	non-parametric	0.011	0.030	0.026	0.037	0.060	0.038	0.059
SPA	non-parametric	0.061	0.095	0.119	0.099	0.129	0.138	0.139
DGP 2								
	source of	T=40	T=40	T=80	T=80	T=80	T=120	T=120
statistic	*critical values*	P=80	P=120	P=40	P=80	P=120	P=40	P=80
MSE-F	fixed regressor	0.706	0.821	0.567	0.763	0.871	0.642	0.822
MSE-t	fixed regressor	0.510	0.668	0.388	0.554	0.679	0.370	0.567
ENC-F	fixed regressor	0.758	0.883	0.617	0.844	0.938	0.722	0.902
ENC-t	fixed regressor	0.672	0.824	0.452	0.701	0.863	0.460	0.735
MSE-F	non-parametric	0.108	0.191	0.110	0.194	0.286	0.152	0.247
MSE-t	non-parametric	0.040	0.065	0.043	0.083	0.145	0.070	0.136
SPA	non-parametric	0.104	0.142	0.188	0.201	0.235	0.232	0.273
ENC-t	Hubrich-West	0.598	0.770	0.429	0.645	0.804	0.440	0.670
DGP 3								
	source of	T=40	T=40	T=80	T=80	T=80	T=120	T=120
statistic	*critical values*	P=80	P=120	P=40	P=80	P=120	P=40	P=80
MSE-F	fixed regressor	0.445	0.594	0.381	0.554	0.678	0.439	0.616
MSE-t	fixed regressor	0.336	0.469	0.285	0.394	0.515	0.284	0.414
ENC-F	fixed regressor	0.455	0.595	0.412	0.578	0.717	0.481	0.654
ENC-t	fixed regressor	0.404	0.569	0.322	0.502	0.649	0.357	0.554
MSE-F	non-parametric	0.088	0.170	0.105	0.194	0.290	0.158	0.248
MSE-t	non-parametric	0.017	0.036	0.026	0.060	0.087	0.049	0.098
SPA	non-parametric	0.045	0.075	0.125	0.120	0.150	0.174	0.189
ENC-t	Hubrich-West	0.353	0.520	0.305	0.451	0.589	0.337	0.508

Notes:
1. The data generating processes are defined in equations (5) and (8). In all of these experiments, the coefficient b in the DGPs is set to the non-zero values given in section 4.2, such that, in population, the most accurate model is one of the alternatives.
2. See the notes to Table 1.

Table 4: Monte Carlo Results on Power, 4-Step Horizon

(nominal size = 10%)

DGP 1								
	source of	T=40	T=40	T=80	T=80	T=80	T=120	T=120
statistic	*critical values*	P=80	P=120	P=40	P=80	P=120	P=40	P=80
MSE-F	fixed regressor	0.404	0.452	0.333	0.436	0.526	0.379	0.476
MSE-t	fixed regressor	0.256	0.301	0.193	0.239	0.313	0.201	0.251
ENC-F	fixed regressor	0.450	0.499	0.360	0.495	0.604	0.408	0.552
ENC-t	fixed regressor	0.281	0.373	0.202	0.280	0.401	0.203	0.312
MSE-F	non-parametric	0.066	0.059	0.127	0.104	0.121	0.166	0.163
MSE-t	non-parametric	0.004	0.002	0.017	0.011	0.015	0.022	0.013
SPA	non-parametric	0.092	0.064	0.264	0.131	0.120	0.314	0.182
DGP 2								
	source of	T=40	T=40	T=80	T=80	T=80	T=120	T=120
statistic	*critical values*	P=80	P=120	P=40	P=80	P=120	P=40	P=80
MSE-F	fixed regressor	0.607	0.707	0.529	0.627	0.747	0.515	0.661
MSE-t	fixed regressor	0.430	0.520	0.276	0.384	0.494	0.255	0.377
ENC-F	fixed regressor	0.584	0.708	0.505	0.637	0.779	0.535	0.712
ENC-t	fixed regressor	0.438	0.595	0.266	0.433	0.618	0.264	0.456
MSE-F	non-parametric	0.180	0.184	0.220	0.212	0.236	0.239	0.231
MSE-t	non-parametric	0.005	0.004	0.015	0.015	0.018	0.017	0.025
SPA	non-parametric	0.290	0.228	0.508	0.331	0.300	0.506	0.377
ENC-t	Hubrich-West	0.627	0.699	0.637	0.635	0.718	0.626	0.652
DGP 3								
	source of	T=40	T=40	T=80	T=80	T=80	T=120	T=120
statistic	*critical values*	P=80	P=120	P=40	P=80	P=120	P=40	P=80
MSE-F	fixed regressor	0.328	0.412	0.298	0.397	0.490	0.305	0.436
MSE-t	fixed regressor	0.206	0.287	0.173	0.227	0.301	0.176	0.263
ENC-F	fixed regressor	0.320	0.406	0.294	0.383	0.513	0.311	0.462
ENC-t	fixed regressor	0.200	0.309	0.166	0.241	0.367	0.185	0.302
MSE-F	non-parametric	0.081	0.100	0.172	0.138	0.177	0.195	0.191
MSE-t	non-parametric	0.001	0.001	0.005	0.003	0.006	0.008	0.007
SPA	non-parametric	0.116	0.088	0.350	0.170	0.147	0.384	0.244
ENC-t	Hubrich-West	0.352	0.415	0.481	0.391	0.465	0.479	0.446

Notes:
1. The data generating processes are defined in equations (5) and (8). In all of these experiments, the coefficient b in the DGPs is set to the non-zero values given in section 4.2, such that, in population, the most accurate model is one of the alternatives.
2. See the notes to Table 1.

Table 5: Pairwise Tests of Equal Accuracy for Monthly Commodity Prices (1-Month Forecast Horizon)

alternative model variables	*RMSE(alt.)/ RMSE(null)*	**Bootstrap p–values** MSE-F fix. reg.	MSE-t fix. reg.	ENC-F fix. reg.	ENC-t fix. reg.	MSE-F non-par.	MSE-t non-par.
CRB futures, XR-AUS	0.970	0.004	0.010	0.007	0.022	0.048	0.075
XR-AUS, XR-major, XR-other	0.974	0.004	0.010	0.009	0.026	0.046	0.068
XR-AUS, XR-other	0.975	0.005	0.014	0.009	0.025	0.050	0.072
CRB futures, XR-other	0.976	0.003	0.005	0.012	0.026	0.022	0.046
XR-AUS	0.977	0.008	0.016	0.010	0.018	0.046	0.070
CRB futures, XR-CAN	0.980	0.011	0.013	0.021	0.033	0.069	0.093
CRB futures	0.981	0.009	0.006	0.019	0.022	0.018	0.048
CRB futures, XR-JAP	0.981	0.010	0.017	0.027	0.054	0.055	0.084
CRB futures, XR-NZ	0.982	0.010	0.012	0.025	0.036	0.056	0.099
CRB futures, XR-major	0.983	0.012	0.008	0.030	0.033	0.028	0.061
XR-NZ, XR-AUS, XR-CAN	0.983	0.024	0.036	0.035	0.057	0.127	0.142
CRB futures, XR-UK	0.984	0.016	0.020	0.031	0.049	0.059	0.098
CRB futures, all 7 XR's	0.989	0.037	0.040	0.064	0.089	0.254	0.262
XR-CAN, XR-other	0.991	0.043	0.049	0.068	0.074	0.158	0.166
XR-CAN, XR-major, XR-other	0.992	0.049	0.055	0.085	0.094	0.186	0.184
XR-NZ, XR-other	0.993	0.042	0.032	0.081	0.061	0.125	0.173
XR-JAP, XR-major, XR-other	0.993	0.072	0.087	0.130	0.177	0.238	0.240
XR-other	0.993	0.032	0.015	0.071	0.034	0.039	0.063
XR-JAP, XR-other	0.994	0.071	0.068	0.126	0.126	0.159	0.172
XR-CAN	0.995	0.084	0.102	0.095	0.096	0.222	0.229
XR-NZ, XR-major, XR-other	0.995	0.064	0.053	0.116	0.087	0.217	0.249
XR-UK, XR-other	0.997	0.114	0.118	0.166	0.161	0.272	0.283
XR-NZ	0.998	0.107	0.102	0.131	0.108	0.317	0.330
XR-major	1.001	0.410	0.691	0.534	0.768	0.866	0.912
XR-JAP	1.002	0.353	0.331	0.400	0.389	0.614	0.616
XR-UK	1.003	0.801	0.483	0.704	0.473	0.721	0.755
XR-UK, XR-major, XR-other	1.009	0.682	0.439	0.544	0.478	0.744	0.783

Notes:
1. As described in section 5, monthly forecasts of the growth rate of commodity prices in period $t+1$ are generated from a null model that includes a constant and growth in prices in period t and alternative models that include the baseline model variables and the period t values of the growth rates of the futures price and various exchange rates. The table lists the additional variables included in each alternative model. Forecasts from January 1997 to December 2008 are obtained from models estimated with a data sample starting in January 1987.
2. This table provides pairwise tests of equal forecast accuracy. For each alternative model, the table reports the ratio of the alternative model's RMSE to the null model's forecast RMSE and bootstrapped p-values for the null hypothesis of equal accuracy, for the test statistics indicated in the columns. Sections 2.3 and 4.1 describe the bootstrap procedures. The RMSE of the null model is 2.408 (the predictand is defined as 100 times the log change in the price level).

Table 6: Pairwise Tests of Equal Accuracy for GDP

alternative model variables	*RMSE(alt.)/ RMSE(null)*	**Bootstrap p–values** MSE-F fix. reg.	MSE-t fix. reg.	ENC-F fix. reg.	ENC-t fix. reg.	MSE-F non-par.	MSE-t non-par.
1-quarter horizon							
$\Delta(C/Y)$	0.921	0.000	0.011	0.000	0.003	0.079	0.047
Δ ln Permits	0.930	0.000	0.026	0.000	0.002	0.106	0.090
Δ ln S&P 500	0.941	0.000	0.044	0.000	0.004	0.173	0.146
Spread, Baa − Aaa	0.982	0.017	0.079	0.023	0.079	0.185	0.163
PMI orders	0.987	0.041	0.169	0.000	0.002	0.437	0.435
Unemp. claims	0.997	0.126	0.139	0.164	0.153	0.268	0.298
Δ 3-month Treasury	0.997	0.140	0.084	0.246	0.132	0.167	0.177
Δ 1-year Treasury	1.002	0.455	0.772	0.614	0.858	0.934	0.929
Hours	1.002	0.331	0.602	0.523	0.729	0.846	0.896
PMI deliveries	1.004	0.796	0.809	0.824	0.799	0.891	0.907
Δ 10-year Treasury	1.008	0.645	0.614	0.738	0.639	0.884	0.898
Spread, 10y − 3m	1.108	0.998	0.991	0.913	0.568	0.999	1.000
Spread, 10y − 1y	1.206	1.000	1.000	0.973	0.686	1.000	1.000
4-quarter horizon							
Δ ln Permits	0.843	0.000	0.043	0.000	0.015	0.010	0.092
Hours	0.992	0.147	0.167	0.135	0.163	0.432	0.444
$\Delta(C/Y)$	0.998	0.216	0.195	0.239	0.165	0.293	0.304
Δ ln S&P 500	1.000	0.283	0.277	0.000	0.011	0.597	0.596
Unemp. claims	1.001	0.371	0.367	0.341	0.337	0.593	0.590
PMI orders	1.004	0.317	0.255	0.026	0.104	0.559	0.557
Spread, Baa − Aaa	1.005	0.398	0.780	0.498	0.831	0.866	0.935
PMI deliveries	1.010	0.790	0.420	0.116	0.294	0.655	0.657
Δ 3-month Treasury	1.027	0.951	0.737	0.946	0.683	0.997	0.927
Δ 1-year Treasury	1.030	0.960	0.813	0.937	0.672	0.992	0.965
Δ 10-year Treasury	1.055	0.967	0.944	0.867	0.641	0.964	0.993
Spread, 10y − 3m	1.233	0.998	0.986	0.296	0.346	0.993	0.998
Spread, 10y − 1y	1.387	1.000	0.996	0.752	0.607	1.000	1.000

Notes:
1. As described in section 5, quarterly forecasts of GDP growth in period $t+\tau$ are generated from a null model that includes a constant and GDP growth in period t and alternative models that include the baseline model variables and the period t value of one additional predictor, listed in the table. Forecasts from 1985:Q1+τ-1 through 2009:Q4 are obtained from models estimated with a data sample starting in 1961:Q2.
2. This table provides pairwise tests of equal forecast accuracy. For each alternative model, the table reports the ratio of the alternative model's RMSE to the null model's forecast RMSE and bootstrapped p-values for the null hypothesis of equal accuracy, for the test statistics indicated in the columns. Sections 2.3 and 4.1 describe the bootstrap procedures. The RMSE of the null model is 2.275 at the 1-quarter horizon and 1.832 at the 4-quarter horizon (the predictand is defined as $(400/\tau)\ln(\text{GDP}_{t+\tau}/\text{GDP}_t)$).

Table 7: Reality Check (Best Model) Tests of Equal Accuracy

	MSE-F	MSE-t	ENC-F	ENC-t
1-month ahead commodity price forecasts				
variables in best model	CRB futures, XR-AUS	CRB futures	CRB futures, XR-AUS	XR-AUS, XR-major, XR-other
fix. reg. p-value	0.016	0.060	0.041	0.178
non-par. p-value	0.080	0.274	NA	NA
SPA p-value	NA	0.245	NA	NA
Hubrich-West p-value	NA	NA	NA	0.238
1-quarter ahead GDP forecasts				
variables in best model	$\Delta(C/Y)$	$\Delta(C/Y)$	Δ ln S&P 500	Δ ln Permits
fix. reg. p-value	0.000	0.137	0.000	0.047
non-par. p-value	0.242	0.427	NA	NA
SPA p-value	NA	0.321	NA	NA
Hubrich-West p-value	NA	NA	NA	0.040
4-quarter ahead GDP forecasts				
variables in best model	Δ ln Permits	Δ ln Permits	Δ ln Permits	Δ ln S&P 500
fix. reg. p-value	0.001	0.377	0.003	0.298
non-par. p-value	0.011	0.846	NA	NA
SPA p-value	NA	0.523	NA	NA
Hubrich-West p-value	NA	NA	NA	0.224

Notes:
1. See the notes to Tables 5 and 6. The variables in the best models (the model that maximizes each test statistic) are given in the top rows of each panel.
2. The table reports p-values for various tests of the null that, in population, all of alternative models are as accurate as the null model (for each application).