Research in Economics 70 (2016) 304–319

Contents lists available at ScienceDirect

Research in Economics

journal homepage: www.elsevier.com/locate/rie

# A comparison of some out-of-sample tests of predictability in iterated multi-step-ahead forecasts[☆]

Pablo M. Pincheira[a], Kenneth D. West[b,*]

[a] School of Business, Adolfo Ibáñez University, Diagonal Las Torres 2640, Peñalolén, Santiago, Chile
[b] Department of Economics, University of Wisconsin-Madison, 1180 Observatory Drive, Madison, WI 53706-1393, USA

ARTICLE INFO

*Article history:*
Received 3 February 2016
Accepted 15 March 2016
Available online 1 April 2016

*Keywords:*
Causality
Random walk
Out-of-sample
Prediction
Multistep forecasts

ABSTRACT

We consider tests of equal population forecasting ability when mean squared prediction error is the metric for forecasting ability, the two competing models are nested, and the iterated method is used to obtain multistep forecasts. We use Monte Carlo simulations to explore the size and power of the MSPE-adjusted test of Clark and West (2006, 2007) (CW) and the Diebold–Mariano–West (DMW) test. The empirical size of the CW test is almost always tolerable: across a set of 252 simulation results that span 5 DGPs, 9 horizons, and various sample sizes, the median size of nominal 10% tests is 8.8%. The comparable figure for the DMW test, which is generally undersized, is 2.2%. An exception for DMW occurs for long horizon forecasts and processes that quickly revert to the mean, in which case CW and DMW perform comparably. We argue that this is to be expected, because at long horizons the two competing models are both forecasting the process to have reverted to its mean. An exception for CW occurs with a nonlinear DGP, in which CW is usually oversized. CW has greater power and greater size adjusted power than does DMW in virtually all DGPs, horizons and sample sizes. For both CW and DMW, power tends to fall with the horizon, reflecting the fact that forecasts from the two competing models both converge towards the mean as the horizon grows. Consistent with these results, in an empirical exercise comparing models for inflation, CW yields many more rejections of equal forecasting ability than does DMW, with most of the rejections occurring at short horizons.

## 1. Introduction

Forecast comparisons have long played a role in evaluation of economic models. A prominent early example was in exchange rate economics. Meese and Rogoff's (1983) demonstration that then-popular exchange rate models forecast no better than a random walk stimulated a huge literature, and made forecast evaluation a common element of evaluation of both reduced form and structural exchange rate models (see Engel et al., 2007). A number of recent examples may be found in volume 2 of the Handbook of Forecasting, whose chapters use forecast comparisons to evaluate, for example, everything from DSGE macro-models (Del Negro and Schorfheide, 2013) to macro-finance models for interest rates (Duffee, 2013) to Phillips curves for inflation (Faust and Wright, 2013).

[☆] We thank Michael McCracken for helpful comments and Jose Luis Nolazco for assistance with the data.
[*] Corresponding author.
*E-mail addresses:* pablo.pincheira@uai.cl (P.M. Pincheira), kdwest@wisc.edu (K.D. West).

http://dx.doi.org/10.1016/j.rie.2016.03.002

Forecast comparisons potentially involve not only point estimates of measures of forecast quality but also standard errors on cross-model differences in forecast quality. Our paper is concerned with the accuracy of inference about forecast quality once a standard error and *t*-statistic are constructed. In the literature, two leading ways to construct standard errors are the methods proposed in Clark and West (2006, 2007) (hereafter CW) and Diebold and Mariano (1995) and West (1996) (hereafter, DMW).[1] Our aim in this paper is to use Monte Carlo simulations to explore the size and power of the CW and DMW tests at multistep horizons. We are interested in their performance both absolutely and relative to one another.

We use the conventional measure of forecast quality, i.e., mean squared prediction error (MSPE). In our simulations, we calibrate our parameters and sample sizes to macro and financial applications such as weekly or monthly exchange rates or stock returns, quarterly GDP or monthly CPI inflation. Four of our five models are linear, and one is nonlinear. Our artificial data generating processes all involve what are called "nested" models, which compare a simple stripped down null model to an alternative model that adds on regressors whose coefficients are presumed to be zero in the null model. We make multistep forecasts using what is called the "iterated" method. As explained in detail in the next section, this method relies on textbook procedures to make multistep forecasts.

The empirical size of the CW test is almost always tolerable: across a set of 252 simulation results that span 5 DGPs, 9 horizons, and various sample sizes, the median size of nominal 10% tests is 8.8%. The comparable figure for the DMW test, which is generally undersized, is 2.2%. An exception for DMW occurs for long horizon forecasts and processes that quickly revert to the mean, in which case the fact that forecasts from both models have reverted to the mean leads DMW to perform as well as CW. An exception for CW occurs with a nonlinear DGP, in which CW is usually oversized; DMW is also oversized but less so than CW. CW has greater power and greater size adjusted power than does DMW in virtually all DGPs, horizons and sample sizes. Power tends to fall with the horizon, consistent with the fact that both models converge towards forecasting the mean.

Two implications for applied work are to use CW in preference to DMW, and to focus on short horizons, because that is where power is greatest. Indeed, in our empirical exercise comparing two models for inflation, CW yields many more rejections of equal forecasting ability than does DMW, with most of the rejections occurring at short horizons.

The simulation results are broadly similar to those in Clark and West (2006, 2007), who focus on one step ahead rather than multistep predictions. They are also similar to the results in Clark and McCracken (2013b), who also compare multistep forecasts using the iterated method.

Our use of the iterated method to construct long horizon forecasts distinguishes our study from most earlier ones. Most of the research evaluating multistep forecast tests assume predictions are constructed using the "direct" rather than iterated method to forecast (the next section briefly explains the direct method). See for instance, Clark and McCracken (2005a) and Busetti and Marcucci (2013). We distinguish ourselves from the aforementioned Clark and McCracken (2013b) paper via use of different DGPs, horizons and sample sizes. Since, as well, some recent empirical literature (e.g., Faust and Wright (2013) and Pincheira and Gatty (2016)) employs the iterated method for multistep forecasts, there is a need for econometric evaluation of forecast inference techniques when the iterated forecasts are used.

We emphasize that we are testing equal population forecasting ability. That is, the relevant set of applications are ones that use forecast comparisons as a model evaluation technique. This is to be distinguished from tests of equal forecasting ability conditional on a given sample, where one is simply looking for a good forecast. See Clark and McCracken (2013a) for further discussion. While DMW can be used to compare equal population forecasting ability when comparing non-nested models (West, 1996), our application is to nested models. So our simulation results are of questionable relevance to comparisons of non-nested models or comparison of forecasting ability conditional on a given sample.

The rest of the paper is organized as follows. Section 2 outlines CW and DMW and the general econometric environment. Section 3 describes our DGPs and our simulation setup. Section 4 presents simulation evidence showing the size and power performance of the two tests. Section 5 illustrates the use of these tests in an empirical application. Section 6 concludes. An on-line appendix available from the authors contains some additional results omitted from the published paper to save space.

## 2. Econometric setup and forecast evaluation framework

### 2.1. Construction of forecasts

Our linear econometric setup considers nested specifications for a scalar dependent variable $y_{t+1}$ as follows:

$$y_{t+1} = X_t' \beta + e_{1t+1} \quad \text{(model 1: null model)}, \tag{2.1}$$

$$y_{t+1} = X_t' \beta + Z_t' \gamma + e_{2t+1} \quad \text{(model 2: alternative model)}, \tag{2.2}$$

where $e_{1t+1}$ and $e_{2t+1}$ are mean zero and i.i.d.

[1] See West (2006) and Clark and McCracken (2013a) for a discussion of some other methods to construct standard errors.

Under the null, $\gamma=0$. In that case, model 2 reduces to model 1. In population (i.e., abstracting from sampling error), forecasts, forecast errors and mean squared forecast errors are the same for both models, for forecasts at any horizon. Under the alternative, $\gamma\neq 0$.[2] Thus, under the alternative, forecasts will be different for the two models. Since model 2 includes information useful in explaining $y_t$ , the population forecasts from model 2 will be superior to those of model 1. As noted above, we used mean squared prediction error as our measure of whether one forecast is superior to another.

Here is how we generate our forecasts. Let $\hat{y}_{1,t+h|t}$ and $\hat{y}_{2,t+h|t}$ be $h$ period ahead forecasts from each of the two models, with $\hat{X}_{t+h|t}$ and $\hat{Z}_{t+h|t}$ the corresponding forecasts of $X$ and $Z$. Let $\hat{\beta}_{1t}$ be a least squares estimate of model 1 that only uses data up to period $t$, with $\hat{\beta}_{2t}$ and $\hat{\gamma}_{2t}$ the model 2 counterparts. Then

$$\hat{y}_{1,t+h|t}=\hat{X}'_{t+h-1|t}\hat{\beta}_{1t}, \hat{y}_{2,t+h|t}=\hat{X}'_{t+h-1|t}\hat{\beta}_{2t}+\hat{Z}'_{t+h-1|t}\hat{\gamma}_{2t}. \tag{2.3}$$

To make these formulas operational, we must construct $\hat{X}_{t+h-1|t}$ and $\hat{Z}_{t+h-1|t}$. We fit univariate or vector autoregressions to the variables in $X_t$ and $Z_t$ and use standard textbook formulas to construct forecasts. Suppose, for example, that $X_t$ is absent and that $Z_t=r_t$ is a scalar that is modeled as a zero mean AR(1) with parameter $\varphi$: $r_t=\varphi r_{t-1}+u_t$, $u_t\sim$i.i.d. Let $\hat{\varphi}_t$ be a least squares estimate of the autoregression that only uses data up to period $t$. Then

$$\hat{r}_{t+h|t}=\hat{\varphi}_t^h r_t. \tag{2.4}$$

A similar, though multivariate, model is used when $Z_t$ is a vector or when model 1 involves regressors $X_t$ .

Readers familiar with the forecasting literature will recognize this as the *iterated* method of generating a multistep forecast. In this method, a single set of regression estimates is used to generate forecasts for all horizons. The alternative is to use the *direct* method. In this method, one runs distinct regressions for each horizon, for example for model 1 estimating

$$y_{t+h}=x_t'\beta_h+\eta_{1t+h}. \tag{2.5}$$

The direct forecast is $x_t'\hat{\beta}_{th}$, where $\hat{\beta}_{th}$ is an estimate of $\beta_h$ that only relies on data up to period $t$. Note that the slope coefficient $\beta_h$ is subscripted by $h$, as is the MA($h-1$) disturbance $\eta_{1t+h}$.

Analytical comparisons of the two methods may be found in Ing (2003) and Schorfheide (2005). A comprehensive empirical comparison is in Marcellino et al. (2006). Our reading is that neither the theoretical nor empirical literature endorses one approach over the other. Despite the lack of clear superiority of one method over another, virtually all previous literature that has considered questions similar to ours has assumed use of the direct method. Hence our decision to focus on the iterated method.

### 2.2. The CW and DMW tests

Let $\hat{e}_{1,t+h|t}\equiv\hat{y}_{1,t+h|t}-\hat{X}'_{t+h-1|t}\hat{\beta}_{1t}$ and $\hat{e}_{2,t+h|t}\equiv\hat{y}_{2,t+h|t}-\hat{X}'_{t+h-1|t}\hat{\beta}_{2t}-\hat{Z}'_{t+h-1|t}\hat{\gamma}_{2t}$ denote the forecast errors at horizon $h$. For simplicity drop the $h$ subscript on $P(h)$ so that $P$ is the number of predictions and prediction errors. The estimates of $h$ period ahead mean squared prediction errors (MSPE) from the two models are

$$\hat{\sigma}^2_{1,h}=\frac{1}{P}\sum_{t=R}^{R+P-1}\hat{e}^2_{1,t+h|t}, \hat{\sigma}^2_{2,h}=\frac{1}{P}\sum_{t=R}^{R+P-1}\hat{e}^2_{2,t+h|t}. \tag{2.6}$$

$$\hat{\sigma}^2_{1,h}=\frac{1}{P}\sum_{t=R}^{R+P-1}\hat{e}^2_{1,t+h|t}, \hat{\sigma}^2_{2,h}=\frac{1}{P}\sum_{t=R}^{R+P-1}\hat{e}^2_{2,t+h|t}. \tag{2.7}$$

Under the null, $\sigma^2_{1,h}=\sigma^2_{2,h}$; under the alternative $\sigma^2_{1,h}>\sigma^2_{2,h}$. Let $\hat{V}$ be an estimate of the long run variance of $e^2_{1,t+h|t}-e^2_{2,t+h|t}$, constructed from the time series on $\hat{e}^2_{1,t+h|t}-\hat{e}^2_{2,t+h|t}$. The DMW test constructs the $t$-statistic

$$\frac{\hat{\sigma}^2_{1,h}-\hat{\sigma}^2_{2,h}}{\sqrt{\hat{V}}} \tag{2.8}$$

and rejects the null if the $t$-statistic exceeds the relevant one-sided critical value – 1.28 for a 10% test, for example. Because it is convention to use standard normal critical values in evaluating (2.8), we refer to DMW as MSPE-normal.[3]

One indication that this procedure is potentially troubled is that, as noted above, under the null, $e_{1,t+h|t}$ and $e_{2,t+h|t}$ are the same random variable, rendering problematic the notion of constructing the long run variance of $e^2_{1,t+h|t}-e^2_{2,t+h|t}$. This is an indication of a deeper problem. Holding the regression sample size $R$ fixed, under the null the DMW statistic converges to a negative value rather than zero as the number of predictions $P\rightarrow\infty$ (Clark and West 2006, 2007). The reason is that the inclusion of the additional regressors $Z_t$ in the alternative model inflates the finite sample value of $\hat{\sigma}^2_{2,h}$: in finite samples the act of estimating coefficients on those regressors introduces noise into the forecasting equation even though the corresponding population coefficient vector $\gamma$ is zero. Since we are conducting a one sided test, the implication is that DMW will be undersized.

[2] See Clark and McCracken (2005b) for an analysis of power of out of sample tests of predictive ability.

[3] Various Clark and McCracken papers (e.g., Clark and McCracken, 2013b) refer to DMW as MSE-t.

The CW statistic adjusts downward the estimate of the MSPE from model 2 to account for the inflation noted in the previous paragraph. Specifically, construction of CW begins by producing an adjusted estimate for the MSPE from model 2,

$$\hat{\sigma}^2_{2,h}-adj.=\frac{1}{P}\sum_{t=R}^{R+P-1}\left[\hat{e}^2_{2,t+h|t}-\left(\hat{y}_{1,t+h|t}-\hat{y}_{2,t+h|t}\right)^2\right] \tag{2.9}$$

(See Clark and West (2006, 2007) for the logic that leads to this adjustment.) Now redefine $\hat{V}$ to be an estimate of the long run variance of $e^2_{1,t+h|t}-[e^2_{2,t+h|t}-(y_{1,t+h|t}-y_{2,t+h|t})^2]$, constructed from the time series on $\hat{e}^2_{1,t+h|t}-\left[\hat{e}^2_{2,t+h|t}-(\hat{y}_{1,t+h|t}-\hat{y}_{2,t+h|t})^2\right]$. The CW test relies on the $t$-statistic

$$\frac{\hat{\sigma}^2_{1,h}-\left(\hat{\sigma}^2_{2,h}-adj.\right)}{\sqrt{\hat{V}}} \tag{2.10}$$

The CW test can also be considered an encompassing test akin to the test proposed by Harvey et al. (1998). This alternative interpretation implies that the CW test is evaluating whether a particular combination between the null and alternative model generates a forecasting strategy with the lowest RMSPE between the following options: (A) to generate forecasts with the null model, (B) to generate forecasts with the alternative model, or (C) to generate forecasts with an average between the strategies in (A) and (B). Let us elaborate.[4]

With some algebra (see Clark and West (2007)), the numerator of CW may be shown to be equal to

$$\frac{2}{P}\sum_{t=R}^{R+P-1}\hat{e}_{1,t+h|t}\left(\hat{e}_{1,t+h|t}-\hat{e}_{2,t+h|t}\right). \tag{2.11}$$

Now, for a given positive scalar $\lambda$ (smaller than one) we could build the following convex forecast combination:

$$y^C_{t+h|t}=\lambda\hat{y}_{2,t+h|t}+(1-\lambda)\hat{y}_{1,t+h|t} \tag{2.12}$$

with forecast error given by

$$e^C_{t+h|t}=\lambda\hat{e}_{2,t+h|t}+(1-\lambda)\hat{e}_{1,t+h|t}=\lambda\left(\hat{e}_{2,t+h|t}-\hat{e}_{1,t+h|t}\right)+\hat{e}_{1,t+h|t} \tag{2.13}$$

The corresponding MSPE is given by

$$E(e^C_{t+h|t})^2=\lambda^2E\left(\hat{e}_{2,t+h|t}-\hat{e}_{1,t+h|t}\right)^2+E(\hat{e}_{1,t+h|t})^2+2\lambda E\left(\hat{e}_{2,t+h|t}-\hat{e}_{1,t+h|t}\right)\hat{e}_{1,t+h|t} \tag{2.14}$$

Since

$$E\left(\hat{e}_{2,t+h|t}-\hat{e}_{1,t+h|t}\right)^2>0 \tag{2.15}$$

Expression (2.14) is a strictly convex quadratic function with a unique global minimum given by

$$\lambda^*=-\frac{E\left(\hat{e}_{2,t+h|t}-\hat{e}_{1,t+h|t}\right)\hat{e}_{1,t+h|t}}{E\left(\hat{e}_{2,t+h|t}-\hat{e}_{1,t+h|t}\right)^2} \tag{2.16}$$

We notice that under mild conditions the numerator of the CW statistic (2.11) converges in probability to twice the numerator of $\lambda^*$. As long as $\lambda^*$ is different from one and zero, the MSPE of the optimal combination should be lower than the MSPE of the two individual forecast in the combination. Rejection of the null hypothesis of the CW test indicates that a combination with a positive weight on the forecast coming from the alternative model should be preferable to either individual forecast.[5]

### *2.3. Asymptotic justification*

To motivate (2.9), and to help interpret our findings, consider the special case in which the null and alternative models are

$$y_{t+1}=e_{1t+1}, \tag{2.17}$$

$$y_{t+1}=y_t\gamma+e_{2t+1}. \tag{2.18}$$

This is a special case of (2.1)–(2.2) with $\beta=0$ and $Z_t=y_t$. Under the null, $\gamma=0$. So in population we may drop the 1 and 2 subscripts and write

$$y_{t+1}=e_{t+1}. \tag{2.19}$$

[4] The argument about to be given assumes rolling windows, with asymptotics done as in Giacomini and White (2006) or Clark and West (2006): $R$ fixed, $P\rightarrow\infty$. Hence this discussion departs from our general focus on population predictive ability rather than predictive ability in a given sample.

[5] If the CW statistic cannot reject the null then we have two possibilities: combination gains are negligible or they might be obtained with a negative weight on the forecasts from the alternative model, which has no simple interpretation.

In (2.18), let $\hat{\gamma}_t$ denote an estimate of $\gamma$ that relies on data going from either $t-R+1$ to $t$ (rolling samples) or from 1 to $t$ (recursive samples).[6] We have

$$\hat{y}_{1,t+h|t}=0, \hat{e}_{1,t+h|t}=y_{t+h}, \hat{e}_{2,t+h|t}=y_{t+h}-y_t\hat{\gamma}_t^h,$$
$$\hat{e}^2_{1,t+h|t}-\hat{e}^2_{2,t+h|t}=y^2_{t+h}-\left(y_{t+h}-y_t\hat{\gamma}_t^h\right)^2=2y_{t+h}y_t\hat{\gamma}_t^h-\left(y_t\hat{\gamma}_t^h\right)^2. \tag{2.20}$$

Thus the numerators of the DMW and CW statistics are

$$\hat{\sigma}^2_{1,h}-\hat{\sigma}^2_{2,h}=\frac{2}{P}\sum_{t=R}^{R+P-1}y_{t+h}y_t\hat{\gamma}_t^h-\frac{1}{P}\sum_{t=R}^{R+P-1}\left(y_t\hat{\gamma}_t^h\right)^2,$$
$$\hat{\sigma}^2_{1,h}-\hat{\sigma}^2_{2,h}-adj.=\frac{2}{P}\sum_{t=R}^{R+P-1}y_{t+h}y_t\hat{\gamma}_t^h. \tag{2.21}$$

Since $y_{t+h}=e_{t+h}$, ~i.i.d., and $\hat{\gamma}_t$ relies only on data that ends in $t$, $Ey_{t+h}y_t\hat{\gamma}_t^h=0$. Thus the expectation of the numerators of the DMW and CW statistics are

$$E(\hat{\sigma}^2_{1,h}-\hat{\sigma}^2_{2,h})=-\tfrac{1}{P}\sum_{t=R}^{R+P-1}E\left(y_t\hat{\gamma}_t^h\right)^2,\quad E\left(\hat{\sigma}^2_{1,h}-\hat{\sigma}^2_{2,h}-adj.=0.\right) \tag{2.22}$$

Although $\gamma=0$ in population, in a sample of finite size, $\hat{\gamma}_t$ and thus $\hat{\gamma}_t^h$ have positive probability of being non-zero. Thus $E\left(y_t\hat{\gamma}_t^h\right)^2>0$ for each $t$, yielding

$$E(\hat{\sigma}^2_{1,h}-\hat{\sigma}^2_{2,h})<0. \tag{2.23}$$

By contrast, the fact that $E(\hat{\sigma}^2_{1,h}-\hat{\sigma}^2_{2,h}-adj.)=0$ means the CW statistic is centered at zero.

Under reasonable conditions, asymptotic normality of the CW statistic follows easily if the rolling scheme is used so that the sample size used to estimate $\gamma$ is held fixed. For in this case, per the logic in Giacomini and White (2006), $y_{t+h}y_t\hat{\gamma}_t^h$ is a stationary random variable and under suitable technical conditions the usual central limit theorem applies.

We are not aware of a general set of conditions in which asymptotic normality results when one uses the recursive scheme or when $\beta\neq0$ so that the null model includes at least one regressor. See Clark and West (2007) and Clark and McCracken (2013a). But Clark and West (2007) argue analytically and with simulations that for the direct scheme approximate normality results under more general circumstances. In particular, MSPE-adjusted (2.10) is the same as Clark and McCracken's (2001) Enc-t statistic. Simulations completed by Clark and McCracken indicate that for the direct scheme, use of (2.10) with (say) 10% tests will result in actual sizes of between 5% and 10%. Hence the statistic can reasonably be thought of as approximately normal, though a bit undersized. We conjecture the same holds for iterated forecasts.[7] This conjecture that seems to be upheld in our simulations.

Four final points. First, since the numerator of DMW is centered in negative territory (see Eq. (2.23)), we expect one sided DMW tests, which at the 10% level reject only if the $t$-statistic is greater than +1.28, to be undersized. Second, for larger $P$ we expect the undersizing to be worse: holding $R$ fixed and letting $P\rightarrow\infty$, the first term on the right hand side of (2.21) converges in probability to zero, the second term to a negative constant. So for large $P$, DMW will pile up around a negative value.

The third and fourth points reflect the observation that as the forecast horizon $h$ gets big, the forecasts from both the null and alternative model will tend towards the mean of $y$. The sample MSPE from each model will therefore tend to be similar, and each will tend to be near the unconditional variance. Further, the CW adjustment will tend to be near zero. (To put this in terms of the example above: $\hat{\gamma}_t^h$ will be near zero for large $h$. Thus the CW adjustment (the final term in (2.21)) will tend to be near zero, and DMW ≈ CW. As well (see (2.20)), $\hat{\gamma}_t^h\approx0$ means $\hat{\sigma}^2_{1,h}\approx\hat{\sigma}^2_{2,h}\approx\frac{1}{P}\sum_{t=R}^{R+P-1}y^2_{t+h}$.) Thus (our third point), we expect DMW to behave like CW and thus be less undersized for larger $h$. Our fourth and final point is that this convergence to the mean implies that differences in power will fade as $h$ gets big.

## 3. Monte Carlo simulations

Our five simulation DGPs include four linear ones stimulated by empirical work in asset pricing (DGPs 1–3) and macroeconomics (DGP 4). Our fifth and final DGP is stimulated by recent work on the CPI by Pincheira et al. (2016). All driving shocks ($e_{t+1}$ and $v_{t+1}$ in DGPs 1–4, $u_{t+1}$ and $\mu_{t+1}$ in DGP 5) are i.i.d. normal. In all simulations we experimented with both rolling and recursive samples, a single value of initial regression sample size $R$ and four values of the number of one step

[6] For overviews of the relevant forecasting literature, including definition and discussion of rolling vs. recursive, see West (2006) and Clark and McCracken (2013a).

[7] See Clark and McCracken (2013b) for a brief exposition of the complications involved in extending the results for direct forecasts to iterated forecasts.

ahead predictions $P$.[8] For conciseness, we report results in detail for only one value of $P$, with results for other values of $P$ detailed in the on-line appendix.

### 3.1. Experimental design

*DGPs 1–2*: For a case where the null is a martingale model (possibly with drift), we consider DGPs such as the ones used in Clark and West (2006), Mankiw and Shapiro (1986), Nelson and Kim (1993), Stambaugh (1999), Campbell (2001) and Tauchen (2001). The general setup is the following:

Null model:

$$y_{t+1}=e_{t+1}\quad \text{(model1)}. \tag{3.1}$$

Alternative model:

$$y_{t+1}=\alpha_y+\gamma r_t+e_{t+1}\quad \text{(model2)}, \tag{3.2a}$$

$$r_{t+1}=\alpha_r+\varphi_1 r_t+\varphi_2 r_{t-1}+\ldots+\varphi_p r_{t-p}+v_{t+1}. \tag{3.2b}$$

This simple setup maps into the notation of (2.1)–(2.2) via: the term $X_t'\beta$ is absent and $Z_t=(1\ r_t\ )'$. In all our simulations, $\alpha_y=\alpha_r=\varphi_3=\ldots=\varphi_p=0$. Let

$$\text{var}(e_{t+1})=\sigma_e^2;\text{var}(v_{t+1})=\sigma_v^2;\text{corr}(e_{t+1},v_{t+1})=\rho. \tag{3.3}$$

We parameterize this as follows (rationale for these values is given below):

|  | $\varphi_1$ | $\varphi_2$ | $\sigma_e^2$ | $\sigma_v^2$ | $\rho$ | $\gamma$, under $H_0$ | $\gamma$, under $H_A$ |  |
|---|---|---|---|---|---|---|---|---|
| DGP 1 | 1.19 | $-0.25$ | $(1.75)^2$ | $(0.075)^2$ | 0 | 0 | $-2$ | (3.4) |
| DGP 2 | 0.5 | 0 | $(0.06)^2$ | $(0.06)^2$ | $-0.4$ | 0 | $-0.9$ |  |

In both DGPs, the null forecast (model 1) imposes $\alpha_y=\gamma=0$, thus assuming $y_{t+1}=e_{t+1}$. The null yields simply the martingale difference or "no change" forecast of 0 for all $t$ and all forecasting horizons. (In terms of the notation above, $\hat{y}_{1,t+h|t}=0$ for all $t$ and $h$.) In both DGPs, the alternative forecast (model 2) for multistep horizons is obtained in part from equation (3.2b), i.e., a regression of $r_{t+1}$ on its own lags and a constant. ("In part" because (3.2a) is of course required as well.) In some simulations we imposed the correct lag order in (3.2b) ($=2$ in DGP 1, $=1$ in DGP 2); in others we use BIC to choose the lag length with maximum lag $p=8$. For the alternative, we compute forecasts using the iterated method and OLS estimates of our parameters, yielding for horizon $h=2$ and lag length $p=2$ in (3.2b), for example,

$$\hat{r}_{t+1|t}=\hat{\alpha}_{rt}+\hat{\varphi}_{1t}r_t+\hat{\varphi}_{2t}r_{t-1}, \tag{3.4$'$}$$

$$\hat{y}_{t+2|t}=\hat{\alpha}_{yt}+\hat{\gamma}_t\hat{r}_{t+1|t}. \tag{3.5}$$

Here, the $t$ subscripts on the coefficients $\hat{\alpha}_{rt}$, $\hat{\varphi}_{1t}$,$\hat{\varphi}_{2t}$, $\hat{\alpha}_{yt}$ and $\hat{\gamma}_t$ emphasize that they are estimated from a sample that ends at date $t$.

The first parameterization, labeled DGP 1, is based roughly on estimates from the exchange rate application considered in the empirical work reported in Clark and West (2006), in which $y_{t+1}$ is the monthly percentage change in a US dollar bilateral exchange rate and $r_t$ is the corresponding interest differential. The parameters were obtained from monthly data. For this DGP we consider an initial estimation window of 120 observations ($R=120$) and report results for $P=300$ predictions. The initial window of $R=120$ corresponds to a reasonable (to us) initial 10 year sample size to estimate regression parameters; the implied sample size (420 months, or 35 years) is one consistent with studies of the modern floating era. The on-line appendix also presents results for $P=100$, $P=200$, and $P=740$. The first two are sizes seen in studies applied to the current floating era; the last is for comparison. We consider experiments in which the number of lags in expression (3.2) is known and ones in which the number of lags in each estimation window is selected using BIC. The maximum lag length is 8 (i.e., $p=8$).

The second parameterization, DGP 2, is calibrated to monthly returns in the copper price $y_{t+1}$ and the Chilean Peso-Dollar exchange rate $r_t$ , using monthly data 1990–2015. The exchange rate was monthly average of daily observations, which accounts for the serial correlation coefficient of $\phi_1=0.5$. According to Chen et al. (2010) commodity currencies should have the ability to predict commodity returns. Accordingly, we set $\gamma=-0.9$ in experiments evaluating power. For this DGP we consider an initial estimation window of 100 observations ($R=100$) and again report results for $P=300$ months. The on-line appendix also considers $P=100$, $P=200$, and $P=400$. The implied sample size ($R+P$) is in the range found in relevant studies.

*DGP 3*: Like DGP 2, DGP 3 is motivated by the literature on commodity currencies. DGP 3 is calibrated to monthly returns of the Non-Fuel Price Index of the IMF $y_{t+1}$ and three commodity currencies versus the U.S. dollar: $r_{1t}$=Australia, $r_{2t}$=South

[8] For a horizon $h$, the number of predictions is $P-h+1$.

Africa and $r_{3t}$=Chile. Null model:

$$y_{t+1} = \alpha_y + \delta y_t + e_{t+1} \quad \text{(model1)}. \tag{3.6}$$

Alternative model:

$$y_{t+1} = \alpha_y + \gamma_1 r_{1t} + \gamma_2 r_{2t} + \gamma_3 r_{3t} + \delta y_t + e_{t+1} \quad \text{(model2)}, \tag{3.7a}$$

$$r_{it+1} = \alpha_{ir} + \varphi_i r_{it} + v_{it+1}, i = 1, 2, 3. \tag{3.7b}$$

In the notation of (2.1)–(2.2), $X_t$=$y_t$ and $Z_t$=(1 $r_{1t}$ $r_{2t}$ $r_{3t}$ )′. In contrast to DGPs 1 and 2, we generate forecasts of future $X_t$'s (=future $y_t$'s) with a vector rather than univariate autoregression. Parameters:

$$\alpha_y = \alpha_{1r} = \alpha_{2r} = \alpha_{3r} = 0, \delta = 0.3, \varphi_1 = \varphi_2 = 0.33, \varphi_3 = 0.5;$$
$$\text{under } H_0, \gamma_1 = \gamma_2 = \gamma_3 = 0; \text{ under } H_A, \gamma_1 = -0.12; \gamma_2 = -0.03\gamma_3 = -0.12. \tag{3.8}$$

These parameters were calibrated to 1990–2015 monthly data, with the three currencies monthly average of daily values. The variance–covariance structure of the shocks ($e_{t+1,}v_{1t+1}, v_{2t+1}, v_{3t+1}$) is given by $10^{-3}$ times the following matrix:

$$\begin{pmatrix} 0.536 & -0.296 & -0.229 & -0.221 \\ -0.296 & 0.666 & 0.352 & 0.251 \\ -0.229 & 0.352 & 1.09 & 0.251 \\ -0.221 & 0.251 & 0.251 & 0.478 \end{pmatrix}$$

We consider an initial estimation window of 120 observations ($R$=120) and $P$=240 predictions. The on-line appendix also presents results for $P$=80, $P$=160, and $P$=320.

*DGP 4*: For DGPs calibrated to macro data, we consider two final DGPs. DGP 4 is the very same DGP 2 in Clark and West (2007). This data generating process is based on models estimated with quarterly data exploring the relationship between US GDP growth and the Federal Reserve Bank of Chicago's factor index of economic activity. This DGP takes the following form:

Null:

$$y_{t+1} = \alpha_y + \delta y_t + e_{t+1} \quad \text{(model1)}. \tag{3.9}$$

Alternative:

$$y_{t+1} = \alpha_y + \delta y_t + \gamma_1 r_t + \gamma_2 r_{t-1} + \gamma_3 r_{t-2} + \gamma_4 r_{t-3} + \ldots + \gamma_p r_{t-p} + e_{t+1} \quad \text{(model 2)}, \tag{3.10a}$$

$$r_{t+1} = \alpha_r + 0.804 r_t - 0.221 r_{t-1} + 0.226 r_{t-2} - 0.205 r_{t-3} + v_{t+1}. \tag{3.10b}$$

In the notation of (2.1)–(2.2), $X_t$=(1 $y_t$ )′ and $Z_t$=($r_t$ … $r_{t-p}$ )′. In all simulations, $\gamma_5$= … =$\gamma_p$=0. Other parameters

$$\alpha_y = 2.237, \delta = 0.261, \alpha_r = 0, \quad \sigma_e^2 = 10.505, \sigma_v^2 = 0.366, \rho = 0.528;$$

$$\text{under } H_0, \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 0; \text{under } H_A, \gamma_2 = -0.633, \gamma_3 = -0.377 \text{ and } \gamma_4 = -0.52. \tag{3.11}$$

In some simulations we imposed the correct lag order of 4 in (3.10a); in others we use BIC to choose the lag length with maximum lag $p$=8. We consider an initial estimation window of 80 observations ($R$=80) and $P$=120 forecasts. The sample sizes are smaller than in DGPs 1–3 because we are calibrating to quarterly data; the implied quarterly sample of 200 observations corresponds to 50 years of data. The on-line appendix also presents results for $P$=40, $P$=80 and $P$=160.

*DGP 5:* Our last DGP (DGP 5) is based on recent work in which traditional measures of monthly CPI core inflation are used to forecast monthly CPI headline inflation (see Pincheira et al., 2016, for details). This is our only nonlinear model. For clarity, we relabel $y_t$ as $\pi_t$ and $r_t$ as $\pi_t^{core}$. The DGP is as follows. Let $u_t$ and $\mu_t$ be i.i.d. shocks. Null:

$$\pi_{t+1} = \alpha + \varphi_\pi \pi_t + \epsilon_{t+1} \quad \text{(model 1)}, \tag{3.12a}$$

$$\varepsilon_{t+1} = u_{t+1} - \theta u_t - \tau u_{t-11} + \tau\theta u_{t-12}. \tag{3.12b}$$

Alternative:

$$\pi_{t+1} = \alpha + \varphi_\pi \pi_t + \gamma \pi_t^{core} + \varepsilon_{t+1} \quad \text{(model 2)}, \tag{3.13a}$$

$$\varepsilon_{t+1} = u_{t+1} - \theta u_t - \tau u_{t-11} + \tau\theta u_{t-12}, \tag{3.13b}$$

$$\pi_{t+1}^{core} = \delta + \omega_{t+1}, \tag{3.13c}$$

$$\omega_{t+1} = \varphi_\omega \omega_t + \mu_{t+1} - b\mu_{t-11}. \tag{3.13d}$$

This DGP does not quite map into (2.1)–(2.2) because the disturbances $\varepsilon_{t+1}$ and $\omega_{t+1}$ are serially correlated. Specifically, for i.i.d. $u_{t+1}$ and $\mu_{t+1}$, $\varepsilon_{t+1} = \left(1 - \tau L^{12}\right)(1 - \theta L)u_{t+1}$ and $(1 - \varphi_\omega L)\omega_{t+1} = \left(1 - bL^{12}\right) \mu_{t+1}$.

**Table 1**
Empirical size, nominal 10% tests, DGPs 1 and 2.

| **MSPE-adjusted/CW** | | | | **MSPE-normal/DMW** | | |
|---|---|---|---|---|---|---|
| **(1)** | **(2)** | **(3)** | **(4)** | **(5)** | **(6)** | **(7)** |
| **Horizon** | **DGP 1** | | **DGP 2** | **DGP 1** | | **DGP 2** |
| | **$p=1$** | **BIC** | | **$p=1$** | **BIC** | |
| **Panel A: Rolling regressions** | | | | | | |
| **$h=1$** | 0.071 | 0.071 | 0.074 | 0.000 | 0.000 | 0.000 |
| **$h=2$** | 0.071 | 0.071 | 0.072 | 0.000 | 0.000 | 0.001 |
| **$h=3$** | 0.077 | 0.078 | 0.066 | 0.000 | 0.000 | 0.002 |
| **$h=6$** | 0.078 | 0.077 | 0.068 | 0.002 | 0.002 | 0.002 |
| **$h=9$** | 0.074 | 0.074 | 0.070 | 0.003 | 0.003 | 0.002 |
| **$h=12$** | 0.071 | 0.072 | 0.069 | 0.003 | 0.003 | 0.002 |
| **$h=18$** | 0.073 | 0.074 | 0.072 | 0.005 | 0.005 | 0.003 |
| **$h=24$** | 0.070 | 0.069 | 0.074 | 0.007 | 0.007 | 0.004 |
| **$h=36$** | 0.071 | 0.071 | 0.078 | 0.009 | 0.009 | 0.008 |
| **Panel B: Recursive regressions** | | | | | | |
| **$h=1$** | 0.070 | 0.070 | 0.063 | 0.004 | 0.004 | 0.004 |
| **$h=2$** | 0.071 | 0.071 | 0.066 | 0.005 | 0.005 | 0.008 |
| **$h=3$** | 0.069 | 0.069 | 0.057 | 0.007 | 0.007 | 0.008 |
| **$h=6$** | 0.072 | 0.072 | 0.058 | 0.009 | 0.008 | 0.010 |
| **$h=9$** | 0.066 | 0.067 | 0.057 | 0.010 | 0.010 | 0.009 |
| **$h=12$** | 0.066 | 0.066 | 0.059 | 0.010 | 0.010 | 0.010 |
| **$h=18$** | 0.062 | 0.062 | 0.060 | 0.016 | 0.015 | 0.011 |
| **$h=24$** | 0.063 | 0.063 | 0.061 | 0.015 | 0.014 | 0.011 |
| **$h=36$** | 0.065 | 0.065 | 0.068 | 0.016 | 0.015 | 0.017 |

*Note*: 1. The table presents empirical sizes for two tests for equality of population mean squared prediction errors (MSPEs) against the one-sided alternative that the alternative model has lower MSPE. Columns (2)–(4) present the test proposed in Clark and West (2006), (5)–(7) the test proposed in Diebold and Mariano (1995) and West (1996). The CW test (2.10) adjusts the difference in sample MSPEs for noise that results because the alternative model's forecast relies on estimates of parameters whose population values are zero. The DMW test (2.8) simply uses differences in MSPEs. Multistep forecasts are computed with the iterated method; see Section 2.

2. The null is that the predictand $y_t$ is white noise, the alternative that $y_t$ depends on a constant and a variable $r_t$ that follows an autoregression. Section 3 of the text gives exact specifications and parameter values. In columns (2), (4), (5) and (7), the alternative model uses the population lag length in the autoregression for $r_t$ ; in columns (3) and (6) the alternative model uses BIC to select the lag length. All models are estimated by least squares.

3. Results are based on 5000 replications. A figure of 0.071 in column (2), $h=1$, for example, indicates that about 350 of the 5000 $t$-statistics (2.10) were greater than 1.28, where 1.28 is the 10% critical value for a one-sided test.

4. Let $R$ be the rolling sample size (panel A) or the smallest recursive sample used to estimate parameters needed under the alternative to make a forecast (panel B). Then $R=120$ in DGP 1, $R=100$ in DGP 2. In both DGPs, the number of predictions is $P=300$. Results for other values of $P$, and for nominal 0.05 and 0.01 tests, are available in the on-line Appendix.

We calibrate these two processes to match in-sample estimates for monthly demeaned U.S. CPI data:

$$\begin{aligned}&\alpha=0, \varphi_\pi=0.96, \theta=-0.45, \tau=0.93, \delta=0, \varphi_\omega=0.99, b=0.93, \sigma_u^2=0.075, \sigma_\mu^2=0.011;\\ &\quad \mathrm{cov}(u_t,\mu_t)=0.007; \text{under } H_0, \gamma=0; \text{under } H_A, \gamma=0.05.\end{aligned} \tag{3.14}$$

In contrast to our previous DGPs, DGP 5 requires nonlinear estimation, because of the seasonal serial correlation in $\varepsilon_{t+1}$ and $\omega_{t+1}$. As a consequence we estimate this DGP with nonlinear least squares. We consider an initial estimation window of 250 observations ($R=250$) and $P=375$ forecasts. The implied sample of 625 months is about 52 years. The on-line appendix also presents results for: $P=125$, $P=250$ and $P=500$.

### 3.2. Some details about our simulations

For each DGP we consider 5000 independent replications. In each replication, we generate 1500 observations on our dependent and independent variables. We discard the first 500 values to ensure stationarity. We evaluate the CW and DMW tests using a variety of combinations of the number of observations used in the first estimation window $R$ and the number of one-step-ahead forecasts $P(1)$. We consider both recursive and rolling schemes. In addition to forecasts $h=1$ periods ahead, we compute iterated forecasts at several forecasting horizons $h=2$, 3, 6, 9, 12, 18, 24 and 36. For all these forecasting horizons we compute the CW and DMW $t$-statistics to compare them with standard normal critical values at the 10%, 5% and 1% significance level for one sided tests. We only show results at the 10% significance level, but the rest of the tables are available upon request.

Here is how we divided up our artificial samples into a segment used for estimation of parameters needed to make forecasts and to a segment used for prediction and prediction errors. Let us assume that we have a total of $T+1$ observations on $y_t$. The end point of the first sample used to estimate regression parameters is observation $R$ (as in $\underline{\text{r}}$egression). We

**Table 2**
Empirical size, nominal 10% tests, DGPs 3–5.

| MSPE-adjusted/CW | | | | | MSPE-normal/DMW | | | |
|---|---|---|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Horizon | DGP 3 | DGP 4 | | DGP 5 | DGP 3 | DGP 4 | | DGP 5 |
| | | $p=3$ | BIC | | | $p=3$ | BIC | |
| **Panel A: Rolling regressions** | | | | | | | | |
| $h=1$ | 0.075 | 0.089 | 0.073 | 0.099 | 0.000 | 0.000 | 0.003 | 0.005 |
| $h=2$ | 0.088 | 0.091 | 0.090 | 0.112 | 0.013 | 0.001 | 0.012 | 0.010 |
| $h=3$ | 0.094 | 0.094 | 0.093 | 0.124 | 0.036 | 0.002 | 0.026 | 0.013 |
| $h=6$ | 0.109 | 0.103 | 0.115 | 0.155 | 0.094 | 0.039 | 0.086 | 0.025 |
| $h=9$ | 0.111 | 0.122 | 0.120 | 0.187 | 0.103 | 0.086 | 0.102 | 0.038 |
| $h=12$ | 0.107 | 0.120 | 0.119 | 0.208 | 0.101 | 0.101 | 0.105 | 0.053 |
| $h=18$ | 0.109 | 0.113 | 0.115 | 0.185 | 0.103 | 0.103 | 0.109 | 0.082 |
| $h=24$ | 0.118 | 0.121 | 0.116 | 0.170 | 0.111 | 0.108 | 0.111 | 0.093 |
| $h=36$ | 0.107 | 0.115 | 0.118 | 0.149 | 0.101 | 0.107 | 0.113 | 0.104 |
| **Panel B: Recursive regressions** | | | | | | | | |
| $h=1$ | 0.069 | 0.081 | 0.060 | 0.078 | 0.003 | 0.001 | 0.008 | 0.010 |
| $h=2$ | 0.087 | 0.083 | 0.075 | 0.089 | 0.025 | 0.005 | 0.022 | 0.012 |
| $h=3$ | 0.090 | 0.081 | 0.088 | 0.095 | 0.052 | 0.010 | 0.042 | 0.015 |
| $h=6$ | 0.107 | 0.102 | 0.105 | 0.117 | 0.101 | 0.054 | 0.089 | 0.025 |
| $h=9$ | 0.115 | 0.108 | 0.108 | 0.147 | 0.112 | 0.088 | 0.099 | 0.040 |
| $h=12$ | 0.115 | 0.119 | 0.117 | 0.162 | 0.113 | 0.108 | 0.113 | 0.054 |
| $h=18$ | 0.120 | 0.127 | 0.112 | 0.149 | 0.117 | 0.119 | 0.109 | 0.068 |
| $h=24$ | 0.121 | 0.133 | 0.115 | 0.130 | 0.120 | 0.129 | 0.113 | 0.073 |
| $h=36$ | 0.125 | 0.131 | 0.124 | 0.121 | 0.123 | 0.126 | 0.121 | 0.089 |

*Note*: 1. See notes of Table 1.
2. In DGPs 3 and 4, the null is that $y_t$ follows an AR(1), the alternative that $y_t$ is driven by a multivariate VAR that implies that the univariate process for $y_t$ is not an AR(1). In DGP 5, the null is that $y_t$ follows a certain univariate seasonal ARMA process, the alternative that $y_t$ follows a certain multivariate seasonal ARMA process that implies that the univariate process for $y_t$ is not the seasonal ARMA assumed under the null. Section 3 of the text gives exact specifications. In columns (2), (3), (6), (7) and (9), the alternative uses population lag lengths; in columns (3) and (7) the alternative uses BIC to pick lags in the equation for $y_t$ . DGPs 3 and 4 are estimated by least squares, DGP 5 by nonlinear least squares.
3. The initial regression size $R$ is $R=120$ (DGP 3), $R=80$ (DGP 4) and $R=250$ (DGP 5). The number of predictions is $P=300$ (DGP 3), $P=120$ ( DGP 4) and $P=375$ (DGP 5). Results for other values of $P$, and for nominal 0.05 and 0.01 tests, are available in the on-line Appendix.

generate a sequence of $P(h)$ $h$-step-ahead predictions estimating the models in either rolling windows of fixed size $R$ or recursive windows of size equal or greater than $R$.

For rolling windows, to generate the first set of forecasts we estimate our models with the first $R$ observations of our sample. Thus, these forecasts are built with information available only at time $R$ and are compared to the observation$y_{R+h}$ for each value of $h$. (Here, $R$ is playing the role of $t$ in Section 2 above.) Next, we estimate our models with the second rolling window of size $R$ that includes observations 2 through $R+1$. These $h$-step-ahead forecasts are compared to the observation $y_{R+h+1}$ for each value of $h$. We continue until the last forecasts are built using the last $R$ available observations for estimation. These forecasts are compared to the observation $y_{T+1}$.

When recursive or expanding windows are used instead, the only difference with the procedure described in the previous paragraph relates to the size of the estimation windows. In the recursive scheme, the estimation window size grows with the number of available observations for estimation. For instance, the first set of forecasts is constructed estimating the models in a window of size $R$, whereas the final set of forecasts is constructed based on models estimated in a window of size $T+1-h$. Thus, we generate a total of $P(h)$ forecasts, with $P(h)$ satisfying $R+(P(h)-1)+h=T+1$. So $P(h)=T+2-h-R$.

As indicated above, our choices of $P$ and $R$ vary with the DGP, so as to align with samples used in the applications that motivate the DGP. But throughout we consider forecasting horizons $h=1$, 2, 3, 6, 9, 12, 18, 24 and 36 periods.

We construct estimates of the long run variance ($\hat{V}$ in (2.8) and (2.10)) using Newey and West (1987, 1994).

## 4. Simulation results

To save space, for each DGP, we report results only for (1) a single value of the number of predictions $P$, and (2) one sided tests of nominal size 0.10. Results for the full range of values of $P$ described above, as well as results for tests of nominal size 0.05 and 0.01 are reported in the on-line Appendix.

**Table 3**
Size adjusted power, nominal 10% tests, DGPs 1 and 2.

| **MSPE-adjusted/CW** | | | | **MSPE-normal/DMW** | | |
|---|---|---|---|---|---|---|
| **(1)** | **(2)** | **(3)** | **(4)** | **(5)** | **(6)** | **(7)** |
| **Horizon** | **DGP 1** | | **DGP 2** | **DGP 1** | | **DGP 2** |
| | **$p=1$** | **BIC** | | **$p=1$** | **BIC** | |
| **Panel A: Rolling regressions** | | | | | | |
| **$h=1$** | 0.995 | 0.995 | 1.000 | 0.990 | 0.990 | 1.000 |
| **$h=2$** | 0.991 | 0.991 | 0.892 | 0.982 | 0.982 | 0.768 |
| **$h=3$** | 0.981 | 0.981 | 0.375 | 0.956 | 0.955 | 0.302 |
| **$h=6$** | 0.890 | 0.889 | 0.110 | 0.792 | 0.776 | 0.094 |
| **$h=9$** | 0.703 | 0.702 | 0.110 | 0.540 | 0.526 | 0.091 |
| **$h=12$** | 0.512 | 0.515 | 0.107 | 0.343 | 0.334 | 0.092 |
| **$h=18$** | 0.297 | 0.303 | 0.111 | 0.155 | 0.154 | 0.097 |
| **$h=24$** | 0.218 | 0.222 | 0.111 | 0.115 | 0.115 | 0.094 |
| **$h=36$** | 0.178 | 0.178 | 0.111 | 0.094 | 0.093 | 0.104 |
| **Panel B: Recursive regressions ($P=300$)** | | | | | | |
| **$h=1$** | 0.999 | 0.999 | 1.000 | 0.992 | 0.992 | 1.000 |
| **$h=2$** | 0.998 | 0.998 | 0.945 | 0.982 | 0.982 | 0.805 |
| **$h=3$** | 0.992 | 0.992 | 0.508 | 0.962 | 0.961 | 0.420 |
| **$h=6$** | 0.939 | 0.938 | 0.110 | 0.827 | 0.824 | 0.102 |
| **$h=9$** | 0.797 | 0.796 | 0.109 | 0.624 | 0.616 | 0.094 |
| **$h=12$** | 0.599 | 0.603 | 0.108 | 0.449 | 0.446 | 0.095 |
| **$h=18$** | 0.326 | 0.330 | 0.112 | 0.218 | 0.215 | 0.103 |
| **$h=24$** | 0.219 | 0.219 | 0.114 | 0.146 | 0.145 | 0.108 |
| **$h=36$** | 0.172 | 0.173 | 0.110 | 0.115 | 0.114 | 0.103 |

*Note*: 1. See notes of Table 1.

### *4.1. Simulation results: size*

For the case of a martingale difference sequence, details of our simulations for DGP 1 and DGP 2 are in Table 1. From Table 1, columns (2)–(4), we see that the CW test is modestly undersized. Actual sizes of nominal 0.10 tests range from about 0.06 to 0.08. Performance is slightly worse in recursive (panel B) than in rolling (panel A) samples. In columns (5)–(7) we see that DMW is seriously undersized, with actual sizes of 0.00 to about 0.02.

Both the modest undersizing of CW and the extreme undersizing of DMW is consistent with Clark and West (2006) and the logic described in Section 2 above. So, too, is the fact that the DMW test is less undersized as the horizon increases; this is consistent with the point made above that the alternative model tends to forecast the unconditional mean at longer horizons, or more generally that forecasts from the null and alternative models become increasingly similar at longer horizons. Other than DMW dependence on horizon, there is remarkably little variation across DGPs, horizon and whether or not the alternative model uses BIC to estimate the lag length (3.2b) (columns (3) and (6)).

How does variation in the number of predictions $P$ affect results? Here are the numbers for DGP 1, $h=12$:

$$\begin{array}{cccccccc} & \text{MSPE}-\text{adjusted/CW} & & & & \text{MSPE}-\text{normal/DMW} & & \\ P=100 & P=200 & P=300 & P=740 & P=100 & P=200 & P=300 & P=740 \\ 0.075 & 0.070 & 0.071 & 0.084 & 0.017 & 0.007 & 0.003 & 0.000 \end{array} \tag{4.1}$$

In (4.1), the figures for $P=300$ repeat the $h=12$ values in columns (2) and (5) of Table 1; the other values come from our on-line appendix. CW is more or less insensitive to $P$. This is consistent with Clark and McCracken (2001), whose analytical results and simulations for the direct method indicate that for a nominal .10 test CW should have size between 0.05 and 0.10. In our simulations, we find that this result also applies to the iterated method. By contrast, DMW is increasingly undersized as $P$ increases. This is consistent with the logic noted above: as $P$ increases, DMW will increasingly pile up around the negative expected value of the second term on the right hand side of (2.21). (4.1) reflects a pattern in the results so ubiquitous that in our subsequent discussion of size we will not have occasion to show values for any $P$ other than the baseline one reported in the tables.

We present summary results for DGPs 3–5 in Table 2. Begin with DGPs 3 and 4. Table 3, columns (2)–(4) indicates that the CW test has size ranging from about 0.07 to 0.13. It tends to be undersized at shorter horizons ($h \le 3$), oversized at longer horizons ($h \ge 6$). Columns (6)–(8) indicate that DMW is, once again, undersized for $h \le 6$ but is well-sized for $h \ge 9$. It appears that forecasts from both models have converged to the unconditional mean by $h=9$. Hence both CW and DMW have similar sizes. Faster convergence here than in DGP 1 is unsurprising, given the relevant AR processes have smaller roots. Convergence is slower in DGP 2 than in these two DGPs because in DGP 2, only one model (the alternative) includes a constant. The forecast from the null model is exactly zero at all horizons, whereas for large $h$ forecasts from the alternative

**Table 4**
Size adjusted power, nominal 10% tests, DGPs 3–5.

| **MSPE-adjusted/CW** | | | | | **MSPE-normal/DMW** | | | |
|---|---|---|---|---|---|---|---|---|
| **(1)** | **(2)** | **(3)** | **(4)** | **(5)** | **(6)** | **(7)** | **(8)** | **(9)** |
| **Horizon** | **DGP 3** | **DGP 4** | | **DGP 5** | **DGP 3** | **DGP 4** | | **DGP 5** |
| | | $\boldsymbol{p=3}$ | **BIC** | | | $\boldsymbol{p=3}$ | **BIC** | |
| **Panel A: Rolling regressions** | | | | | | | | |
| $\boldsymbol{h=1}$ | 0.866 | 1.000 | 1.000 | 0.248 | 0.831 | 1.000 | 0.999 | 0.236 |
| $\boldsymbol{h=2}$ | 0.619 | 0.992 | 0.994 | 0.246 | 0.498 | 0.975 | 0.913 | 0.243 |
| $\boldsymbol{h=3}$ | 0.334 | 0.710 | 0.639 | 0.250 | 0.270 | 0.626 | 0.348 | 0.235 |
| $\boldsymbol{h=6}$ | 0.127 | 0.483 | 0.390 | 0.248 | 0.116 | 0.293 | 0.191 | 0.250 |
| $\boldsymbol{h=9}$ | 0.121 | 0.201 | 0.196 | 0.254 | 0.116 | 0.134 | 0.119 | 0.252 |
| $\boldsymbol{h=12}$ | 0.110 | 0.145 | 0.137 | 0.254 | 0.106 | 0.125 | 0.111 | 0.249 |
| $\boldsymbol{h=18}$ | 0.108 | 0.159 | 0.154 | 0.233 | 0.106 | 0.143 | 0.129 | 0.212 |
| $\boldsymbol{h=24}$ | 0.107 | 0.175 | 0.166 | 0.188 | 0.106 | 0.160 | 0.149 | 0.173 |
| $\boldsymbol{h=36}$ | 0.112 | 0.173 | 0.155 | 0.146 | 0.109 | 0.157 | 0.139 | 0.136 |
| **Panel B: Recursive regressions** | | | | | | | | |
| $\boldsymbol{h=1}$ | 0.939 | 1.000 | 1.000 | 0.404 | 0.863 | 1.000 | 0.999 | 0.379 |
| $\boldsymbol{h=2}$ | 0.689 | 0.998 | 0.998 | 0.396 | 0.524 | 0.978 | 0.933 | 0.371 |
| $\boldsymbol{h=3}$ | 0.373 | 0.787 | 0.693 | 0.398 | 0.274 | 0.654 | 0.396 | 0.379 |
| $\boldsymbol{h=6}$ | 0.129 | 0.503 | 0.435 | 0.395 | 0.120 | 0.325 | 0.235 | 0.365 |
| $\boldsymbol{h=9}$ | 0.107 | 0.224 | 0.220 | 0.380 | 0.105 | 0.159 | 0.149 | 0.358 |
| $\boldsymbol{h=12}$ | 0.109 | 0.119 | 0.117 | 0.378 | 0.106 | 0.105 | 0.098 | 0.351 |
| $\boldsymbol{h=18}$ | 0.112 | 0.135 | 0.146 | 0.347 | 0.111 | 0.127 | 0.136 | 0.311 |
| $\boldsymbol{h=24}$ | 0.116 | 0.141 | 0.151 | 0.293 | 0.114 | 0.130 | 0.143 | 0.251 |
| $\boldsymbol{h=36}$ | 0.104 | 0.131 | 0.136 | 0.182 | 0.103 | 0.125 | 0.127 | 0.154 |

*Note*: 1. See note of Table 2.

model converge to a sample mean which is close to, but not exactly, zero. Simulations not reported for the sake of brevity show that when forecasts of the alternative model are constructed imposing $\hat{\alpha}_{yt}=0$ in (3.5), convergence of both forecasts to the mean is as rapid in DGP 2 as in DGPs 3 and 4.

In DGPs 1 and 4, lag selection by BIC has little effect on size, except in DGP 4, where DMW is better sized when using BIC.

Recall that DGP 4 is calibrated to quarterly data. We note that in applications for which DGP 4 is representative, one would almost certainly not forecast more than 12 quarters ahead. Hence the relevant results are for short horizons ($h<9$), where DMW is consistently undersized, and medium horizons ($h$=9,12), where DMW performs comparably to CW.

Turn now to DGP 5, the model estimated by nonlinear least squares. The CW test (column (5)) now behaves quite differently. It is distinctly oversized except at short ($h\le 3$) horizons. The oversizing increases with horizon until $h$=12, at which point it declines, presumably reflecting forecasts that are closer and closer to the unconditional mean. The DMW test (column (9)) displays the familiar pattern of undersizing that diminishes with the horizon. We do not have an explanation for the behavior of CW. Possibilities include sample sizes that are too small, simulations that by chance are unrepresentative, or a failure of the theory to apply.

In the end we ran simulations over 9 horizons × 4 values of $P$ × 7 DGPs=252 sets of simulations. (We get 7 rather than 5 for the number of DGPs by counting the use of BIC for DGPs 1 and 4 as two additional DGPs.) Let us sort the empirical size from lowest to highest. Over those 252 results, here are the values at the first quartile (the 63rd smallest value), the median, and the third quartile (63rd largest value):

$$\begin{array}{ccc@{\quad}ccc} \multicolumn{3}{c}{\mathrm{MSPE-adjusted/CW}} & \multicolumn{3}{c}{\mathrm{MSPE-normal/DMW}} \\ \mathrm{Q1} & \mathrm{median} & \mathrm{Q3} & \mathrm{Q1} & \mathrm{median} & \mathrm{Q3} \\ 0.071 & 0.088 & 0.118 & 0.007 & 0.022 & 0.101 \end{array} \tag{4.2}$$

Eq. (4.2) and the results in Tables 1 and 2 lead to the following summary on size. Consistent with the theory described in Section 2, DMW is undersized, CW is adequately sized. In DGPs 1–4 the CW test displays adequate size at all forecasting horizons, much better than the DMW test with the exception of long horizon forecasts for DGPs 3 and 4 where CW and DMW are comparable. For our model estimated by nonlinear least squares (DGP 5), CW is preferable at short horizons but choosing between CW and DMW at long horizons involves comparing oversized (CW) and undersized (DMW) tests.

DMW behaves like CW, and hence is adequately sized, when the forecasts from the null and alternative models are similar. In this case the adjustment term $P^{-1}\sum_t(\hat{y}_{1,t+h|t}-\hat{y}_{2,t+h|t})^2$ (see (2.9)) is quantitatively small. This happens, for

**Table 5**
Empirical power, nominal 10% tests, DGPs 1 and 2.

| MSPE-adjusted/CW | | | | MSPE-normal/DMW | | |
|---|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Horizon | DGP 1 | | DGP 2 | DGP 1 | | DGP 2 |
| | $p=1$ | BIC | | $p=1$ | BIC | $p=1$ |
| **Panel A: Rolling regressions** | | | | | | |
| $h=1$ | 0.993 | 0.993 | 1.000 | 0.707 | 0.707 | 0.960 |
| $h=2$ | 0.986 | 0.986 | 0.850 | 0.647 | 0.645 | 0.192 |
| $h=3$ | 0.976 | 0.976 | 0.289 | 0.556 | 0.552 | 0.012 |
| $h=6$ | 0.865 | 0.866 | 0.080 | 0.267 | 0.255 | 0.002 |
| $h=9$ | 0.649 | 0.649 | 0.079 | 0.106 | 0.100 | 0.003 |
| $h=12$ | 0.450 | 0.460 | 0.077 | 0.042 | 0.038 | 0.002 |
| $h=18$ | 0.242 | 0.250 | 0.080 | 0.017 | 0.016 | 0.004 |
| $h=24$ | 0.182 | 0.186 | 0.083 | 0.015 | 0.014 | 0.006 |
| $h=36$ | 0.144 | 0.145 | 0.086 | 0.018 | 0.015 | 0.008 |
| **Panel B: Recursive regressions ($P=300$)** | | | | | | |
| $h=1$ | 0.998 | 0.998 | 1.000 | 0.845 | 0.845 | 0.981 |
| $h=2$ | 0.996 | 0.996 | 0.914 | 0.793 | 0.793 | 0.404 |
| $h=3$ | 0.989 | 0.988 | 0.373 | 0.728 | 0.725 | 0.072 |
| $h=6$ | 0.916 | 0.914 | 0.064 | 0.452 | 0.450 | 0.010 |
| $h=9$ | 0.736 | 0.734 | 0.064 | 0.256 | 0.256 | 0.011 |
| $h=12$ | 0.517 | 0.516 | 0.066 | 0.130 | 0.128 | 0.011 |
| $h=18$ | 0.252 | 0.250 | 0.068 | 0.055 | 0.054 | 0.013 |
| $h=24$ | 0.159 | 0.161 | 0.070 | 0.034 | 0.033 | 0.014 |
| $h=36$ | 0.129 | 0.130 | 0.076 | 0.031 | 0.031 | 0.019 |

*Note*: 1. See note of Table 1.

**Table 6**
Empirical power, nominal 10% tests, DGPs 3–5.

| MSPE-adjusted/CW | | | | | MSPE-normal/DMW | | | |
|---|---|---|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Horizon | DGP 3 | DGP 4 | | DGP 5 | DGP 3 | DGP 4 | | DGP 5 |
| | | $p=3$ | BIC | | | $p=3$ | BIC | |
| **Panel A: Rolling regressions** | | | | | | | | |
| $h=1$ | 0.837 | 1.000 | 1.000 | 0.246 | 0.138 | 0.952 | 0.936 | 0.028 |
| $h=2$ | 0.590 | 0.991 | 0.992 | 0.265 | 0.165 | 0.623 | 0.635 | 0.044 |
| $h=3$ | 0.323 | 0.699 | 0.622 | 0.281 | 0.134 | 0.158 | 0.150 | 0.058 |
| $h=6$ | 0.137 | 0.485 | 0.417 | 0.324 | 0.111 | 0.164 | 0.172 | 0.093 |
| $h=9$ | 0.129 | 0.229 | 0.226 | 0.360 | 0.120 | 0.118 | 0.121 | 0.129 |
| $h=12$ | 0.116 | 0.168 | 0.157 | 0.384 | 0.108 | 0.126 | 0.114 | 0.159 |
| $h=18$ | 0.116 | 0.182 | 0.173 | 0.348 | 0.108 | 0.148 | 0.139 | 0.179 |
| $h=24$ | 0.125 | 0.203 | 0.185 | 0.282 | 0.117 | 0.173 | 0.159 | 0.161 |
| $h=36$ | 0.118 | 0.193 | 0.173 | 0.210 | 0.111 | 0.165 | 0.151 | 0.139 |
| **Panel B: Recursive regressions** | | | | | | | | |
| $h=1$ | 0.917 | 1.000 | 1.000 | 0.356 | 0.362 | 0.976 | 0.965 | 0.093 |
| $h=2$ | 0.657 | 0.998 | 0.997 | 0.375 | 0.255 | 0.759 | 0.748 | 0.108 |
| $h=3$ | 0.353 | 0.754 | 0.666 | 0.386 | 0.174 | 0.280 | 0.247 | 0.121 |
| $h=6$ | 0.144 | 0.509 | 0.448 | 0.426 | 0.121 | 0.218 | 0.219 | 0.164 |
| $h=9$ | 0.122 | 0.233 | 0.236 | 0.461 | 0.118 | 0.145 | 0.148 | 0.212 |
| $h=12$ | 0.123 | 0.141 | 0.138 | 0.485 | 0.119 | 0.113 | 0.113 | 0.246 |
| $h=18$ | 0.131 | 0.164 | 0.160 | 0.432 | 0.128 | 0.146 | 0.145 | 0.240 |
| $h=24$ | 0.142 | 0.183 | 0.168 | 0.344 | 0.139 | 0.169 | 0.154 | 0.200 |
| $h=36$ | 0.135 | 0.167 | 0.158 | 0.207 | 0.132 | 0.154 | 0.146 | 0.137 |

*Note*: 1. See note of Table 2.

example, for horizons long enough so that both null and alternative forecasts have converged to the mean. This will only happen for very large horizons for persistent DGPs such as DGP 1; it will happen for more modest horizons for rapidly mean reverting processes such as DGPs 3 and 4.

### *4.2. Simulation results: power*

Tables 3 and 4 present results for size adjusted power, Tables 5 and 6 for power.

In DGPs 1–4, CW shows good power and size adjusted power at low horizons, with power falling towards 0.10 as the horizon increases. For example, for DGP 1, in Panel A, column (2) of Table 3, size adjusted power is 0.995 for $h=1$, falling to 0.178 for $h=36$; for DGP 3, in Panel A, column (2) of Table 4 the comparable figures are 0.866 and 0.112 . The fall is slowest in DGP 1, where the alternative exploits a very persistent regressor. That the fall is slowest in DGP 1, and the fact that power falls as $h$ increases, is consistent with long horizon forecasts under both null and alternative approaching the unconditional mean. This means that in all DGPs MSPEs from both forecasts are numerically similar for large $h$, but the value of $h$ that qualifies as "large" is biggest for DGP 1 where the alternative relies on a very persistent regressor.

In DGP 5, power is not good for CW even at low horizons. Of course this is a direct reflection of the calibration of the alternative (as is the good power at low horizons for DGPs 1–4): in (3.13a) the key parameter $\gamma$ is set to a very small value (see Eq. (3.14)). But as well, the qualitative pattern of power declining towards nominal size does not apply in this DGP. Instead power rises and then falls with the horizon, mimicking the rise and fall of empirical size in Table 1. We do not have an intuitive explanation for this.

Unsurprisingly, power increases as $P$ increases. Here are results for various $P$ for DGP 1, $h=12$:

$$\begin{array}{cccccccc} \multicolumn{4}{c}{\text{MSPE}-\text{adjusted/CW}} & \multicolumn{4}{c}{\text{MSPE}-\text{normal/DMW}} \\ P=100 & P=200 & P=300 & P=740 & P=100 & P=200 & P=300 & P=740 \\ 0.343 & 0.439 & 0.512 & 0.707 & 0.237 & 0.290 & 0.343 & 0.488 \end{array} \tag{4.3}$$

The figures for $P=300$ repeat the $h=12$ values in columns (2) and (5) of Table 3. The pattern of power increasing with $P$ was ubiquitous, characterizing all DGPs.

In general, for CW, power tends to be higher for recursive than for rolling regressions, though there are some exceptions. The differences between recursive and rolling are, however, small.

DMW is more erratic. Overall, the qualitative pattern is as in CW, with high size adjusted power for small horizons that declines as the horizon increases – see for example column (5) in Table 3. But raw power sometimes declines well below 0.10. See, for example the figures for $h\geq 3$ for DGP 2 in column (7) of Table 5. For this DGP, for both CW and DMW, power is barely above size for $h\geq 3$. The implication is that behavior under the null and alternative are very similar for such horizons – that is, under both null and alternative forecasts have converged to the mean for $h\geq 3$.

Here are the quartiles for size adjusted power and for power, analogous to the figures for size in Eq. (4.2). Size adjusted power

$$\begin{array}{cccccc} \multicolumn{3}{c}{\text{MSPE}-\text{adjusted/CW}} & \multicolumn{3}{c}{\text{MSPE}-\text{normal/DMW}} \\ Q1 & \text{Median} & Q3 & Q1 & \text{Median} & Q3 \\ 0.138 & 0.255 & 0.708 & 0.119 & 0.197 & 0.557 \end{array} \tag{4.4}$$

Power:

$$\begin{array}{cccccc} \multicolumn{3}{c}{\text{MSPE}-\text{adjusted/CW}} & \multicolumn{3}{c}{\text{MSPE}-\text{normal/DMW}} \\ Q1 & \text{Median} & Q3 & Q1 & \text{Median} & Q3 \\ 0.161 & 0.282 & 0.682 & 0.090 & 0.145 & 0.229 \end{array} \tag{4.5}$$

Eqs. (4.4) and (4.5) and Tables 3–6 lead to the following summary on power. In every single entry in Tables 3–6, CW has higher power or size adjusted power than does DMW. Sometimes the gap is large. (Example in Table 5: DGP 1, $h=6$: 0.865 for CW vs. 0.267 for DMW (rolling) and 0.916 for CW vs. 0.452 for DMW (recursive).) Sometimes the gap is small. But the fact that CW has higher power in each simulation leads to CW having higher power everywhere in the distribution in Eqs. (4.4) and (4.5). Higher raw power (Eq. (4.5), Tables 5 and 6) is consistent with the theory outlined in Section 2.

## 5. Empirical Illustration

We consider predicting inflation with an international factor. A relatively recent literature has explored the predictive linkages between domestic and international inflation concluding that, at least for some countries, this linkage is important both at the core and headline level. See for instance Ciccarelli and Mojon (2010), Morales-Arias and Moura (2013), Hakkio (2009), Pincheira and Gatty (2016) and Medel et al. (forthcoming).

Let $\pi_{it}$ be year-on-year domestic inflation rates in country $i$. Following the literature cited in the previous paragraph, we build an international inflation factor (IIF) as the simple average of $\pi_{it}$ measured using monthly CPI data, with $i$ ranging over

**Table 7**
Forecasts of year-on-year headline CPI Inflation.

| (1) | MSPE-adjusted/CW | | | | | MSPE-normal/DMW | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) |
| Country | $h=1$ | $h=3$ | $h=6$ | $h=12$ | $h=24$ | $h=1$ | $h=3$ | $h=6$ | $h=12$ | $h=24$ |
| | 0.00 | 0.01 | 0.02 | 0.04 | 0.09 | 0.00 | 0.00 | 0.00 | −0.01 | 0.02 |
| Austria | (0.00) | (0.01) | (0.03) | (0.06) | (0.12) | (0.00) | (0.01) | (0.03) | (0.06) | (0.12) |
| | 1.75** | 0.88 | 0.68 | 0.58 | 0.77 | 0.59 | 0.30 | 0.09 | -0.08 | 0.17 |
| | 0.01 | 0.03 | 0.08 | 0.27 | 0.23 | 0.00 | 0.01 | 0.05 | 0.20 | 0.13 |
| Belgium | (0.00) | (0.02) | (0.07) | (0.16) | (0.23) | (0.00) | (0.02) | (0.06) | (0.15) | (0.22) |
| | 1.98** | 1.16 | 1.17 | 1.73** | 1.03 | 1.09 | 0.67 | 0.73 | 1.32* | 0.59 |
| | 0.11 | 0.25 | 0.36 | −0.50 | −2.14 | 0.06 | 0.15 | 0.16 | −0.93 | −3.03 |
| Chile | (0.04) | (0.13) | (0.28) | (0.65) | (1.32) | (0.03) | (0.12) | (0.28) | (0.63) | (1.35) |
| | 2.79*** | 1.99** | 1.27 | −0.77 | −1.63 | 1.98** | 1.24 | 0.57 | −1.46 | −2.25 |
| | 0.01 | 0.03 | 0.03 | 0.17 | 0.36 | 0.01 | 0.01 | −0.02 | 0.03 | −0.03 |
| Italy | (0.01) | (0.02) | (0.04) | (0.12) | (0.31) | (0.00) | (0.02) | (0.04) | (0.12) | (0.31) |
| | 2.34*** | 1.31* | 0.68 | 1.44* | 1.17 | 1.52* | 0.47 | −0.42 | 0.26 | -0.10 |
| | 0.01 | 0.06 | 0.16 | 1.14 | 2.63 | −0.03 | −0.21 | −0.59 | −0.63 | −1.37 |
| Mexico | (0.01) | (0.09) | (0.18) | (0.46) | (0.97) | (0.01) | (0.11) | (0.28) | (0.55) | (0.81) |
| | 0.83 | 0.73 | 0.86 | 2.46*** | 2.71*** | −2.50 | −1.94 | −2.09 | −1.12 | −1.69 |
| | 0.04 | 0.09 | 0.30 | 0.51 | 0.29 | 0.02 | 0.03 | 0.15 | 0.27 | 0.04 |
| USA | (0.02) | (0.10) | (0.27) | (0.27) | (0.45) | (0.02) | (0.10) | (0.24) | (0.21) | (0.34) |
| | 2.09** | 0.83 | 1.08 | 1.88** | 0.64 | 1.37* | 0.33 | 0.65 | 1.29* | 0.10 |

*Note*:
1. In this table forecasts from a simple AR(1) (null model, or model 1) for year-on-year monthly CPI inflation rate are compared to forecasts coming from an alternative model (model 2) that augments model 1 with a measure of international inflation. See Section 5 for details.
2. International inflation is defined as the simple average of monthly year-on-year domestic CPI inflation rates for 31 OECD economies.
3. The first row for each country is $\hat{\sigma}^2_{1,h}-\hat{\sigma}^2_{2,h}$, the difference in sample MSPE between the null and the alternative model. A positive value means the null model (model 1) had a larger sample MSPE than did the alternative model (model 2). A negative value means that the null model had a larger sample MSPE.
4. Newey-West (1987, 1994) standard errors are in parentheses.
5. * Means statistically significant at the 10% significant level. ** Means statistically significant at the 5% level and *** denotes statistically significance at the 1% level.
6. Data are described in the text. See notes of earlier tables for additional definitions.

the current 31 OECD countries[9]

$$\pi_t^{IIF}=\frac{1}{31}\sum_{i=1}^{31}\pi_{it}. \tag{5.1}$$

We use data ranging from January 1995 to December 2015 (252 observations). We focus on headline inflation. For the out-of-sample analysis we estimate our models by OLS in recursive windows with an initial window length of 100 observations ($R=100$, from January 1995 to April 2003). This means that our first one-step-ahead forecast is made for May 2003, while the last one is made for December 2015. We consider forecasts for the following horizons: $h=1$, 3, 6, 12 and 24 months ahead. We analyze if the IIF has the ability to predict inflation for $i=$Austria, Belgium, Chile, Italy, Mexico and the US. For each country, we consider the following nested models:

$$\pi_{it+1}=\alpha_\pi+\beta\pi_{it}+e_{t+1} \quad (\text{model 1: null model}), \tag{5.2}$$

$$\pi_{it+1}=\alpha_\pi+\beta\pi_{it}+\gamma(L)\pi_t^{IIF}+e_{t+1} \quad (\text{model 2: alternative model}), \tag{5.3}$$

$$\pi_{t+1}^{IIF}=\alpha_r+\varphi\pi_t^{IIF}+v_{t+1}. \tag{5.4}$$

Here, $\gamma(L)=\sum_{j=0}^{q}\gamma_j L^j$ represents a lag polynomial and $L$ is the lag operator such that $L^jX_t=X_{t-j}$. In contrast to our DGP 5, but consistent with our linear DGPs 1–4, the disturbances $e_{t+1}$ and $v_{t+1}$ are i.i.d.

The lag order $q$ is selected in each estimation window with BIC with $1\le q\le 12$.

Table 7 shows our results. For CW, in columns (2)–(6), the point estimate is the numerator of MSPE adjusted, given in (2.10). For DMW, the point estimate is simply the difference between model 1 and model 2 MSPEs, i.e., $\sigma^2_{1,h}–\sigma^2_{2,h}$ (see (2.8)). Thus in columns (7)–(11), a negative value means the null model has a smaller sample MSPE than did the alternative model ($\hat{\sigma}^2_{1,h}<\hat{\sigma}^2_{2,h}$) while a positive value means the converse ($\hat{\sigma}^2_{1,h}>\hat{\sigma}^2_{2,h}$). By construction, the point estimates for CW are algebraically larger than those for DMW.

[9] We consider the following countries: Austria, Belgium, Canada, Chile, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, The Netherlands, Norway, Poland, Portugal, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, U.K. and the U.S. Data source: OECD Main Economic Indicators.

**Table 8**
RMSPE for Mexico.

| | RMSPE for Mexico | | | | |
|---|---|---|---|---|---|
| Model | $h=1$ | $h=3$ | $h=6$ | $h=12$ | $h=24$ |
| Model without the IIF (model 1: null model) | 0.299 | 0.636 | 0.775 | 1.162 | 1.286 |
| Model with IIF (model 2: alternative model) | 0.352 | 0.785 | 1.093 | 1.406 | 1.739 |
| 0.8*(model 1)+0.2*(model 2) | 0.298 | 0.635 | 0.774 | 1.092 | 1.135 |

Note:
1. In this table forecasts from a simple AR(1) (null model) for year-on-year monthly CPI inflation rate are compared to forecasts coming from the same model but augmented with lags of international inflation. (the alternative model), and with a linear combination between the forecasts of these two models with weights 0.8 on the null model and 0.2 on the alternative model.
2. Figures in the table represent RMSPE for headline year-on-year CPI inflation in Mexico. For example, for $h=1$, the $-0.03$ figure for the MSPE difference in column (7) in Table 7 for Mexico aligns with the figures in the first two rows of the $h=1$ column in the present table via: $-0.03=(0.299)^2-(0.352)^2$.
3. See notes to the previous table for additional definitions.

Consistent with the simulations on both size and power, DMW rejects less frequently than does CW. In columns 7–11 in Table 7 there are 5 rejections at the 10% level using DMW (Belgium and the US, when forecasting one year ahead ($h=12$)), and for Chile, Italy and the US when forecasting $h=1$ month ahead). In columns (2)–(6) of that table there are 12 rejections using the CW ($h=1$ for all countries but Mexico, $h=3$ for Chile and Italy, $h=12$ for Belgium, Italy, Mexico and the USA, and $h=24$ for Mexico).

A country with striking results is Mexico, where the CW test strongly rejects the null hypothesis for horizons of one and two years ($h=12$ and $h=24$). For the same country we do not reject the null with the DMW test. Furthermore, the unadjusted MSPE differences in columns (9) and (10) are negative, meaning $\hat{\sigma}^2_{1,h}$, the sample MSPE from the null model, is less than $\hat{\sigma}^2_{2,h}$, the sample MSPE of the alternative model that exploits international inflation. We can interpret this result according to expression (2.14) that we rewrite for the particular case in which $\lambda = 1$:

$$E(\hat{e}_{2,t+h|t})^2 - E(\hat{e}_{1,t+h|t})^2 = E(\hat{e}_{2,t+h|t} - \hat{e}_{1,t+h|t})^2 + 2E(\hat{e}_{2,t+h|t} - \hat{e}_{1,t+h|t})\hat{e}_{1,t+h|t} \tag{5.5}$$

Results for Mexico have a positive left hand side, but a negative second term in the right hand side of expression (5.5). This results from the positive first term on the right hand side, which is the term that Clark and West (2007) propose to remove. The contrast between CW and DMW indicates that a combination between the alternative and null models with a positive but small weight on the alternative model should outperform either of the individual models. The optimal weight, computed for each horizon $h$ using (2.16) and the obvious sample analogs to the expectations in (2.16), yields $\lambda^*$ above 0.2 at each horizon. We use this ex-post information to forecast using a linear combination between the alternative and null models that gives a weight of $\lambda$=20% to the model with the IIF and a weight of 80% to the null model. We obtain more accurate forecasts at every single horizon. See Table 8.[10]

## 6. Summary and concluding remarks

In this paper we explore the behavior of two tests commonly used to compare forecasts from competing nested models: the MSPE adjusted statistic of Clark and West (2006, 2007) (CW) and the Diebold–Mariano–West (DMW) tests, which we call MSPE-normal. The focus of interest is multistep ahead forecasts computed using the iterated method. Our Monte Carlo simulations for linear models indicate that CW tests are reasonably well sized across all horizons, while DMW is quite undersized except, in some DGPs, at horizons long enough that forecasts from competing models are similar. Our simulations for a nonlinear model indicated that neither test was very well sized. In terms of power, the CW is preferred to DMW test at all horizons. Power is an increasing function of the sample size (i.e., the number of forecasts $P$). Longer horizons mean less power, presumably because at longer horizons both null and alternative converge to forecasting the mean.

An application in the context of inflation forecasts is consistent with our simulation results.

Future research could explore in more detail the behavior of the CW test in nonlinear DGPs or compare its performance against other benchmarks, like the tests proposed by Clark and McCracken (2001). Similarly, the analysis of a joint test of predictability across all possible forecasting horizons could represent a natural extension of the present work.

[10] Note that Table 8 is not a forecasting comparison in the sense of all the other comparisons in this paper: we used the results of Table 7's forecasting comparison to pick $\lambda$=20%. Rather, Table 8 illustrates the fact that when CW and DMW conflict, an implication is that one can do better by combining forecasts.

## References

Busetti, F., Marcucci, J., 2013. Comparing forecast accuracy: a monte carlo investigation. Int. J. Forecast. 29 (1), 13–27.

Campbell, J.Y., 2001. Why long horizons? A study of power against persistent alternatives. J. Empir. Financ. 8, 459–491.

Ciccarelli, M., Mojon, B., 2010. Global inflation. Rev. Econ. Stat. 92 (3), 524–535.

Chen, Y., Rogoff, K.S., Rossi, B., 2010. Can exchange rates forecast commodity prices? Q. J. Econ. 125 (3), 595–620.

Clark, T., McCracken, M., 2001. Tests of equal forecast accuracy and encompassing for nested models. J. Econom. 105, 85–110.

Clark, T., McCracken, M., 2005a. Evaluating direct multistep forecasts. Econom. Rev. 24, 369–404.

Clark, T., McCracken, M., 2005b. The power of tests of predictive ability in the presence of structural breaks. J. Econom. 124 (1), 1–31.

Clark, T., McCracken, M., 2013a. Advances in forecast evaluation. Handbook of Economic Forecasting, vol. 2B. , Elsevier, Amsterdam, pp. 1107–1201.

Clark, T., McCracken, M., 2013b. Evaluating the accuracy of forecasts from vector autoregressions. In: Fomby, T., Killian, L., Murphy, A. (Eds.), Vector Autoregressive Modeling – New Developments and Applications: Essays in Honor of Christopher A. Sims, Emerald Group Publishing, Bingley.

Clark, T., West, K.D., 2006. Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. J. Econom. 135 (1–2), 155–186.

Clark, T., West, K.D., 2007. Approximately normal tests for equal predictive accuracy in nested models. J. Econom. 138, 291–311.

Del Negro, M., Schorfheide, F., 2013. DSGE based model forecasting. Handbook of Economic Forecasting, vol. 2B. , Elsevier, Amsterdam, pp. 57–140.

Diebold, F., Mariano, R., 1995. Comparing predictive accuracy. J. Bus. Econ. Stat. 13 (3), 253–263.

Duffee, G., 2013. Forecasting interest rates. Handbook of Economic Forecasting, vol. 2B. , Elsevier, Amsterdam, pp. 385–426.

Engel, C., Mark, N., West, K., 2007. Exchange rate models are not as bad as you think. In: Acemoglu, D., Rogoff, K., Woodford, M. (Eds.), NBER Macroeconomics Annual, University of Chicago Press, Chicago, pp. 381–443.

Faust, J., Wright, J., 2013. Forecasting inflation. Handbook of Economic Forecasting, vol. 2B. , Elsevier, Amsterdam, pp. 2–56.

Giacomini, R., White, H., 2006. Tests of conditional predictive ability. Econometrica 74, 1545–1578.

Hakkio, C., 2009. Global inflation dynamics. Research Working Paper 09-01, Federal Reserve Bank of Kansas City.

Harvey, D.I., Leybourne, S.J., Newbold, P., 1998. Tests for forecast encompassing. J. Bus. Econ. Stat. 16, 254–259.

Ing, C.K., 2003. Multistep prediction in autoregressive processes. Econom. Theory, 254–279.

Mankiw, N.G., Shapiro, M.D., 1986. Do we reject too often? Small sample properties of tests of rational expectations models. Econ. Lett. 20, 139–145.

Marcellino, M., Stock, J., Watson, M., 2006. A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. J. Econom. 127 (1–2), 499–526.

Medel, C., Pedersen, M., Pincheira, P., 2016. The elusive predictive ability of global inflation. Int. Financ. (forthcoming).

Meese, R.A., Rogoff, K., 1983. Empirical exchange rate models of the seventies: do they fit out of sample? J. Int. Econ. 14, 3–24.

Morales-Arias, L., Moura, G.V., 2013. A conditional heteroskedastic global inflation model. J. Econ. Stud. 40 (4), 572–596.

Nelson, C.R., Kim, M.J., 1993. Predictable stock returns: the role of small sample bias. J. Financ. 48, 641–661.

Newey, W.K., West, K., 1987. A simple, positive, semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55 (3), 703–708.

Newey, W.K., West, K.D., 1994. Automatic lag selection in covariance matrix estimation. Rev. Econ. Stud. 61, 631–654.

Pincheira, P., Gatty, A., 2016. Forecasting chilean inflation with international factors. Empir. Econ. http://dx.doi.org/10.1007/s00181.

Pincheira, P., Selaive, J., Nolazco, J.L., 2016. The evasive predictive ability of core inflation. BBVA Research Working Paper No. 15-34, January.

Schorfheide, F., 2005. VAR forecasting under misspecification. J. Econom. 128, 99–136.

Stambaugh, R.F., 1999. Predictive regressions. J. Financ. Econ. 54, 375–421.

Tauchen, G., 2001. The bias of tests for a risk premium in forward exchange rates. J. Empir. Financ. 8, 695–704.

West, K.D., 1996. Asymptotic inference about predictive ability. Econometrica 64 (5), 1067–1084.

West, K.D., 2006. Forecast evaluation. Elliott, G., Granger, C., Timmerman, A. (Eds.), Handbook of Economic Forecasting, vol. 1. , Elsevier, Amsterdam, pp. 100–134.