Received: 8 April 2022 | Revised: 22 March 2023 | Accepted: 3 August 2023

DOI: 10.1002/jae.3050

WILEY Journal of APPLIED ECONOMETRICS

RESEARCH ARTICLE

Tests for equal forecast accuracy under heteroskedasticity

David I. Harvey[1] | **Stephen J. Leybourne**[1] | **Yang Zu**[2]

[1]School of Economics, University of Nottingham, Nottingham, UK

[2]Department of Economics, University of Macau, Macau, China

Correspondence
David I. Harvey, School of Economics, University of Nottingham, University Park, Nottingham NG7 2RD, UK.
Email: dave.harvey@nottingham.ac.uk

Summary

Heteroskedasticity is a common feature in empirical time series analysis, and in this paper, we consider the effects of heteroskedasticity on statistical tests for equal forecast accuracy. In such a context, we propose two new Diebold–Mariano-type tests for equal accuracy that employ nonparametric estimation of the loss differential variance function. We demonstrate that these tests have the potential to achieve power improvements relative to the original Diebold–Mariano test in the presence of heteroskedasticity, for a quite general class of loss differential series. The size validity and potential power superiority of our new tests are studied theoretically and in Monte Carlo simulations. We apply our new tests to competing forecasts of changes in the dollar/sterling exchange rate and find the new tests provide greater evidence of differences in forecast accuracy than the original Diebold–Mariano test, illustrating the value of these new procedures for practitioners.

KEYWORDS

Diebold–Mariano test, forecast accuracy, nonparametric volatility estimation

1 | INTRODUCTION

Forecasting economic and financial time series plays a central role in decision making, both in the public sector policy-making context and in private sector environments, with the quality of the forecasts being a key ingredient in making effective and appropriate decisions. Consequently, evaluation of the quality of competing forecasts is of great importance, in order to determine which of a number of forecasting approaches will deliver the best indicator of the future state of the world. In this context, it is crucial to have available techniques that allow discernment as to whether one set of forecasts is more accurate than another according to some measure of forecast error loss. In statistical terms, this translates into the need to have size-controlled but powerful tests of the null of equal forecast accuracy, and a number of procedures have been developed to this end. Early contributions to this problem were provided by, inter alia, Granger and Newbold (1977) and Meese and Rogoff (1988), focusing on testing equal mean squared forecast error under restricted assumptions concerning the forecast errors. In a key development to the literature, Diebold and Mariano (1995) (DM) proposed a test for equal forecast accuracy based on general loss differentials. Subsequent work on forecast evaluation testing has focused on cases where the forecasts have been produced by estimated models, either non-nested or nested, with major contributions to this strand of the literature being West (1996) and Giacomini and White (2006) (GW); for reviews of this literature, see West (2006) and Clark and McCracken (2013).

Heteroskedasticity is a common feature in macroeconomic and financial data. If heteroskedasticity exists in the series being forecast during the evaluation period, that will likely be transferred into loss differential series based on the forecast errors over that evaluation period. For example, suppose we consider two competing q-step-ahead forecasts f_{1t} and f_{2t} of a variable y_{t+q} and evaluate the forecasts using quadratic loss. Then, the loss differential can be expressed as $\Delta L_{t,q} = (y_{t+q} - f_{1t})^2 - (y_{t+q} - f_{2t})^2 = f_{1t}^2 - f_{2t}^2 - 2y_{t+q}(f_{1t} - f_{2t})$, and it is clear that any heteroskedasticity in y_{t+q} will be translated

into heteroskedasticity in the loss differential series. That heteroskedasticity is a pertinent feature of loss differential series in practical applications can clearly be seen from the exchange rate examples considered in our empirical application (see Figures 7 and 8). Consequently, the effects of heteroskedasticity are a very important consideration when analysing the performance of tests of competing forecast accuracy. The asymptotic size of the DM test is robust to the presence of heteroskedasticity in the loss differential, but it is not necessarily efficient in terms of power.

In this paper, in the context of a constant mean model for the loss differential series, we propose new DM-type tests, exploiting the unconditional heteroskedasticity structure present in the data (while also permitting the existence of rather flexible conditional heteroskedasticity), which can achieve power gains relative to the original DM test. The new tests are constructed using unconditional heteroskedasticity-adjusted loss differential series, where the heteroskedasticity structure is estimated by a kernel-type nonparametric estimator. We derive the limit distributions of the test statistics under the null of equal forecast accuracy, and these limits are found to be standard normal, in line with the original DM test. We further derive the distributions of the new statistics under a local alternative, and we show that under heteroskedasticity, the new tests dominate the original DM test in terms of local asymptotic power for a quite general class of loss differential series.

Our approach largely follows the GW framework of testing hypotheses about *forecasting methods*, with the 'method' encapsulating both the model and the parameter estimation procedure used to generate the forecasts, assuming a rolling window model estimation scheme. Clearly, the performance of model-based forecasts relies on both the adequacy of the model and the precision of the estimation procedure. In particular, the GW framework offers a context to discuss technical conditions imposed on the models, which has practical value in comparing model-based forecasts. In contrast, the DM framework abstracts from forecast models and imposes assumptions directly on the loss differential series, thereby being able to compare non-model-based forecasts (e.g., survey-based forecasts). However, as discussed in Patton (2015) and Diebold (2015b), the two frameworks are closely related, and conditions imposed on loss differentials directly as in the DM framework can be viewed as high-level conditions to be satisfied by forecasting methods. Therefore, our approach can be equally applied to the DM forecast comparison environment.

A highly relevant problem for many forecast evaluators is using forecasts to compare *forecasting models*. West (1996) makes an important contribution in this direction, showing that in order to compare models based on forecasts, one has to take into account the model estimation error and modify the DM statistic to achieve correct inference. We discuss how our new DM-type statistics might also be adjusted to test hypotheses about models, along the lines of West (1996) and West and McCracken (1998).

The rest of the paper is organised as follows. In Section 2, we set up our framework, and in Section 3, we introduce two new DM-type tests for equal accuracy of competing forecast methods, demonstrating their asymptotic validity under quite general assumptions on the data and models. Section 4 studies the local asymptotic powers of our tests and compares them with the corresponding local asymptotic power of the DM test. In Section 5, we present finite sample size and power simulation results. Section 6 discusses the issue of comparing forecasting models and also considers the implications of a time-varying loss differential mean for equal accuracy testing. In Section 7, we present the results of an empirical illustration, evaluating competing forecasts of changes in the dollar/sterling exchange rate across both the pre-European Exchange Rate Mechanism (ERM) and post-ERM periods. Section 8 concludes. Proofs of our asymptotic results are provided in Appendix S1. In the remainder of the paper, we use the following notation: $\xrightarrow{d}$ denotes convergence in distribution, $\xrightarrow{p}$ convergence in probability and $\lfloor . \rfloor$ the integer part of its argument.

2 | MODELLING FRAMEWORK

We first introduce the notation and discuss assumptions about the data, the models and the estimation procedures, where we largely follow the framework associated with the unconditional predictive ability test of GW. Consider a sequence of random data vectors $W_t \equiv \{W_t : \Omega \to \mathbb{R}^{s+1}, s \in \mathbb{N}, t = 1, 2, \dots\}$ defined on a complete probability space $(\Omega, \mathcal{F}, P)$. The data vector W_t is partitioned as $W_t = (y_t, X_t')'$, where y_t is the variable being forecast and X_t is a vector of predictors. Define the filtration $\mathcal{F}_t = \sigma(W_1, W_2, \dots, W_t)$. Suppose two competing models are used to forecast the variable of interest q steps ahead, that is, y_{t+q}. The forecasts formulated at time t are based on the information set $\mathcal{F}_t$ and are denoted by $f_i(W_t, W_{t-1}, \dots, W_{t-w_i+1}; \hat{\beta}_{i,t})$ for model i, $i = 1, 2$. Here, f_i is a measurable function, and $\hat{\beta}_{i,t}$ is the estimated parameter for model i using a fixed number of observations w_i over the period $t - w_i + 1, \dots, t$, for $i = 1, 2$. As in GW, this formulation is rather general and the forecasts can be point, interval, probability or density predictions. With this formulation, it is explicit that the two forecasts are *mappings* which actually map past data, through models chosen and the parameters

estimated, to forecasts of the future time period $t+q$. We follow GW and refer to $f_i(W_t, W_{t-1}, \ldots, W_{t-w_i+1}; \hat{\beta}_{i,t})$, $i = 1, 2$ as *forecast methods*.

Let $t = 1, \ldots, R$ with $R \geq \max(w_1, w_2)$ denote the sample of data prior to the beginning of the forecasting exercise, such that the forecasts are generated at times $t = R+1, \ldots, T$. For a measurable loss function L, we thus have a sequence of $T-R$ forecast loss differentials:

$$\{\Delta L_{t,q}\}_{t=R+1}^{T} \equiv \left\{L\left(y_{t+q}, f_1\left(W_t, \ldots, W_{t-w_1+1}; \hat{\beta}_{1,t}\right)\right) - L\left(y_{t+q}, f_2\left(W_t, \ldots, W_{t-w_2+1}; \hat{\beta}_{2,t}\right)\right)\right\}_{t=R+1}^{T}.$$

The loss function L can take many forms, for example quadratic loss, absolute loss or others such as those discussed in Chapter 2 of Elliott and Timmermann (2016). In the case of evaluating density forecasts, the loss function usually takes the form of the negative of a scoring rule—see Gneiting and Raftery (2007) for more discussion and the references therein. In what follows, to ease exposition, we take $q = 1$ and suppress the dependence on q in all our notation for ΔL_t. This gives rise to a sample of $n = T - R$ loss differentials

$$\Delta L_t, \; t = R+1, \ldots, T,$$

from which to compute test statistics for equal forecast accuracy.

In this paper, we consider a model where the loss differential series has a constant mean. That is, $E(\Delta L_t) = c$, where c is a constant. We are interested in testing between the following hypotheses:

$$H_0 : c = 0 \; vs. \; H_1 : c \neq 0.$$

One-sided alternatives can also be considered in the usual way. Note that our framework is in line with DM and West (1996) who also assume a constant mean for the ΔL_t series (these papers also assume stationarity). This differs from the GW framework where the loss differential mean is allowed to be time varying; in section 6.2 we discuss issues surrounding testing within a framework that permits a time-varying mean for ΔL_t.

As in GW, we make the following assumption on the data vector $\{W_t\}$:

Assumption 1. $\{W_t\}$ is α-mixing of size $-r/(r-2)$ with $r > 2$.

Given our focus on considering unconditional heteroskedasticity in the loss differential series, we now specify possible unconditional heteroskedasticity in the ΔL_t sequence through the following assumption:

Assumption 2. Let $\text{Var}(\Delta L_t) = \sigma_t^2 = \sigma^2((t-R)/n)$, where $\sigma(.)$ is a deterministic, positive function.

Assumption 1 specifies dependence conditions permitted in the data and does not require stationarity; these dependence conditions are sufficient for a central limit theorem (CLT) to hold and are comparable with the assumptions made in Theorem 4 of GW. Within our testing framework, the loss differential series $\{\Delta L_t\}$ will inherit the same dependence conditions as the data. Assumption 2 imposes very little on the unconditional variance of the loss differential series $\{\Delta L_t\}$. Further assumptions for the $\sigma(.)$ function will be relevant and imposed later when discussing the estimation of this function. Notice that the conditional variance of the $\{\Delta L_t\}$ series is not explicitly specified in our framework, thus can be very flexible. To be more explicit, any conditional heteroskedasticity structure is permitted in our model, provided the dependence conditions in Assumption 1 are satisfied. From Carrasco and Chen (2002), it is known that many commonly used conditional heteroskedasticity models, such as the ARCH model of Engle (1982), the GARCH model of Bollerslev (1986) and the log normal stochastic volatility model of Andersen (1994), when stationary, are all α-mixing with coefficients decaying exponentially fast and are thus permitted in our model.

Heteroskedasticity in the loss differential might arise from heteroskedasticity in the target variable being predicted. By way of an illustration, consider a simple forecast accuracy comparison where the target variable is given by $y_t = \mu + \omega_t \eta_t$, with ω_t deterministic and $\eta_t \sim IIDN(0,1)$, so that the variable being predicted is heteroskedastic ($\text{Var}(y_t) = \omega_t^2$). Suppose two methods for forecasting y_{t+q} are being evaluated: (i) a forecast based on the mean plus an irrelevant putative predictor x_t, $f_{1t} = \mu + x_t$, where $x_t \sim IIDN(0, c)$ and x_t, η_t independent; (ii) a naive forecast based on the mean of the target variable, $f_{2t} = \mu$. In this scenario, abstracting from estimation of μ and using a quadratic loss function, we have

$$\begin{aligned}\Delta L_t &= f_{1t}^2 - f_{2t}^2 - 2y_{t+q}(f_{1t} - f_{2t}) \\ &= x_t^2 - 2x_t \omega_{t+q} \eta_{t+q}.\end{aligned}$$

Here, $E(\Delta L_t) = c$ and $\text{Var}(\Delta L_t) = 2c^2 + 4c\omega_{t+q}^2$; hence, the heteroskedasticity in y_t translates into heteroskedasticity in ΔL_t. Given the prevalence of heteroskedasticity in target variables being predicted in economics and finance, it is to be expected that heteroskedastic loss differentials will be commonplace in forecast evaluation exercises.

Further motivation for the heteroskedastic loss differential framework that we adopt can be seen through considering the impact of heteroskedasticity in forecast errors. For example, consider the following simple illustrative model for forecast errors, using the notation $e_{i,t+q} = y_{t+q} - f_{it}$, $i = 1, 2$:

$$e_{1,t+q} = \omega_{t+q} u_{1,t+q} + z_t,$$
$$e_{2,t+q} = \omega_{t+q} u_{2,t+q},$$

with ω_t deterministic, $(u_{1t}, u_{2t}) \sim IIDN(0, I_2)$ and $z_t \sim IIDN(0, c)$, with z_t independent of (u_{1t}, u_{2t}). In this setup, the forecast errors are clearly heteroskedastic, with $\text{Var}(e_{1,t+q}) = \omega_{t+q}^2 + c$ and $\text{Var}(e_{2,t+q}) = \omega_{t+q}^2$. Using a quadratic loss function, we have

$$\begin{aligned}\Delta L_t &= e_{1,t+q}^2 - e_{2,t+q}^2 \\ &= \omega_{t+q}^2 u_{1,t+q}^2 + z_t^2 + 2\omega_{t+q} u_{1,t+q} z_t - \omega_{t+q}^2 u_{2,t+q}^2.\end{aligned}$$

Here, $E(\Delta L_t) = c$ and $\text{Var}(\Delta L_t) = 2c^2 + 4c\omega_{t+q}^2 + 8\omega_{t+q}^4$, and the heteroskedasticity in the forecast errors translates into heteroskedasticity in ΔL_t.

In the context of model-based density forecast evaluation based on scoring rules, such as those discussed in Gneiting and Raftery (2007), heteroskedasticity could also emerge as a result of model misspecification. Consider forecast evaluation for macroeconomic series as discussed in Clark (2011). For illustrative purposes, we consider the special case of a density forecast for a scalar variable. In Clark's model, the forecast for the one-step-ahead conditional density function of the target series is the normal density

$$\hat{f}(y) = \frac{1}{\sqrt{2\pi\hat{\sigma}_t^2}} \exp\left(-\frac{1}{2}\left(\frac{y - \hat{\mu}_t}{\hat{\sigma}_t}\right)^2\right),$$

for $y \in \mathbb{R}$, where $\hat{\mu}_t$ is the one-step-ahead forecast of the conditional mean implied by Clark's conditional mean autoregressive model and $\hat{\sigma}_t$ is the corresponding one-step-ahead conditional volatility estimator implied by Clark's stochastic volatility model for the error volatility. When the logarithmic scoring rule is used, the loss series takes the form

$$L(y_{t+1}) = -\log \hat{f}(y_{t+1}) = \frac{1}{2}\left(\log(2\pi) + \log\left(\hat{\sigma}_t^2\right)\right) + \frac{1}{2}\left(\frac{y_{t+1} - \hat{\mu}_t}{\hat{\sigma}_t}\right)^2, \; t = R+1, \ldots, T.$$

From this representation, we see that if Clark's stochastic volatility model is correctly specified, then both $\log\left(\hat{\sigma}_t^2\right)$ and $\left(\frac{y_{t+1} - \hat{\mu}_t}{\hat{\sigma}_t}\right)^2$ will be homoskedastic and the loss series will also be homoskedastic. If two forecasters both adopt the correctly specified model, but estimate the parameters with different model estimation strategies, the resulting loss differentials will be homoskedastic, even if the original data are heteroskedastic. However, other than this specific case, any misspecification in the dynamics of the conditional mean and/or conditional volatility models in one or both forecasters will likely result in heteroskedasticity in the loss differential series.

3 | TEST STATISTICS AND THEIR ASYMPTOTIC NULL DISTRIBUTIONS

The test statistic given in DM and GW is defined as

$$DM = \sqrt{n}\frac{n^{-1}\sum_{t=R+1}^{T} \Delta L_t}{\sqrt{\hat{\Omega}(\Delta L)}}, \tag{1}$$

where $\hat{\Omega}(\Delta L)$ is a suitable heteroskedasticity and autocorrelation consistent (HAC) estimator of the variance of $n^{-1/2}\sum_{t=R+1}^{T} \Delta L_t$. Throughout our analysis, the HAC estimator used takes the form

$$\hat{\Omega}(x) = n^{-1}\sum_{t=R+1}^{T}\sum_{s=R+1}^{T} x_t x_s k\left(\frac{t-s}{b}\right), \tag{2}$$

where $k(.)$ denotes a kernel function and b the lag truncation parameter. In (1), we have the HAC estimator obtained by setting $x_t = \Delta L_t$ in (2).

3.1 | Infeasible heteroskedasticity-adjusted statistics

We now motivate two DM-type test statistics that explicitly account for the unconditional heteroskedastic component σ_t, initially assuming the function $\sigma(.)$ is known. First, notice that the null hypothesis $H_0 : E(\Delta L_t) = 0$ is equivalent to a null hypothesis $H_0' : E(\Delta L_t/\sigma_t) = 0$ given that σ_t is deterministic. That is, the equal forecast accuracy hypothesis expressed in terms of the $\{\Delta L_t\}$ series is equivalent to equal forecast accuracy expressed in terms of the heteroskedasticity-adjusted $\{\Delta L_t/\sigma_t\}$ series. Applying the DM statistic to the heteroskedasticity-adjusted loss differential $\{\Delta L_t/\sigma_t\}$, we obtain the first of our new DM-type statistics, which is infeasible at this stage due to the assumed knowledge of $\sigma(.)$:

$$DM' = \sqrt{n}\frac{n^{-1}\sum_{t=R+1}^{T}\frac{\Delta L_t}{\sigma_t}}{\sqrt{\hat{\Omega}\left(\frac{\Delta L}{\sigma}\right)}}, \tag{3}$$

where $\hat{\Omega}\left(\frac{\Delta L}{\sigma}\right)$ denotes the HAC estimator (2) evaluated using $x_t = \frac{\Delta L_t}{\sigma_t}$, that is, an estimator of the variance of $n^{-1/2}\sum_{t=R+1}^{T}\frac{\Delta L_t}{\sigma_t}$.

Alternatively, as noted in Diebold (2015a), the DM statistic can be viewed as a HAC standard error-corrected t statistic for testing $H_0 : c = 0$ in the following regression:

$$\Delta L_t = c + \varepsilon_t,$$

where ε_t is a heterogeneous dependent error sequence satisfying $E(\varepsilon_t) = 0$. Under Assumption 2, we also have that $\mathrm{Var}(\varepsilon_t) = \mathrm{Var}(\Delta L_t) = \sigma_t^2$, hence on making a weighted least squares (WLS) transformation of the above regression, we have

$$\frac{\Delta L_t}{\sigma_t} = c\frac{1}{\sigma_t} + v_t, \tag{4}$$

and the new 'error' series $\{v_t\} = \{\varepsilon_t/\sigma_t\}$ satisfies $E(v_t) = 0$ and $\mathrm{Var}(v_t) = 1$. Notice that (4) is a regression model of the series $\{\Delta L_t/\sigma_t\}$ on a regressor $\{1/\sigma_t\}$. Using a standard sandwich-form Wald-type statistic for testing the restriction $c = 0$ in the regression (4), after some simple algebra, we obtain our second DM-type infeasible statistic

$$DM^* = \sqrt{n}\frac{n^{-1}\sum_{t=R+1}^{T}\frac{\Delta L_t}{\sigma_t^2}}{\sqrt{\hat{\Omega}\left(\frac{\Delta L}{\sigma^2}\right)}}, \tag{5}$$

where $\hat{\Omega}\left(\frac{\Delta L}{\sigma^2}\right)$ denotes the HAC estimator (2) evaluated using $x_t = \frac{\Delta L_t}{\sigma_t^2}$. Notice that our HAC estimator in the denominator is constructed under the null $c = 0$.

3.2 | Feasible heteroskedasticity-adjusted statistics

In practice, we require feasible versions of the infeasible statistics (3) and (5), which depend on the unknown quantity σ_t. To this end, we propose estimating σ_t using the following nonparametric estimator for the function $\sigma^2(\tau)$, $\tau \in [0, 1]$:

$$\hat{\sigma}^2(\tau) = \sum_{j=R+1}^{T} w_{\tau,j}\Delta L_j^2, \tag{6}$$

where $w_{\tau,j} = K\left(\frac{(j-R)/n-\tau}{h}\right) / \sum_{j=R+1}^{T} K\left(\frac{(j-R)/n-\tau}{h}\right)$ with $K(.)$ a kernel function and h the bandwidth. Notice that the kernel function $K(.)$ and the bandwidth parameter h are distinct from the kernel function $k(.)$ and the lag truncation parameter b in (2). The corresponding estimator $\hat{\sigma}_t^2$, $t = R+1, \ldots, T$, is given by $\hat{\sigma}_t^2 = \hat{\sigma}^2((t-R)/n)$ and $\hat{\sigma}_t = \sqrt{\hat{\sigma}_t^2}$.

The feasible version of the DM' statistic (3) then becomes

$$DM' = \sqrt{n}\frac{n^{-1}\sum_{t=R+1}^{T}\frac{\Delta L_t}{\hat{\sigma}_t}}{\sqrt{\hat{\Omega}\left(\frac{\Delta L}{\hat{\sigma}}\right)}}, \tag{7}$$

where $\hat{\Omega}\left(\frac{\Delta L}{\hat{\sigma}}\right)$ denotes the HAC estimator (2) evaluated using $x_t = \frac{\Delta L_t}{\hat{\sigma}_t}$. Note that although the HAC estimator $\hat{\Omega}\left(\frac{\Delta L}{\hat{\sigma}}\right)$ takes the standard form (see, e.g., De Jong & Davidson, 2000), it is actually applied to a series that has been adjusted by the nonparametrically estimated $\hat{\sigma}_t$. Therefore, its statistical properties are unknown and need to be established as part of our analysis.

Analogously, the feasible version of the DM^* statistic (5) is given by

$$DM^* = \sqrt{n}\frac{n^{-1}\sum_{t=R+1}^{T}\frac{\Delta L_t}{\hat{\sigma}_t^2}}{\sqrt{\hat{\Omega}\left(\frac{\Delta L}{\hat{\sigma}^2}\right)}},$$

where $\hat{\Omega}\left(\frac{\Delta L}{\hat{\sigma}^2}\right)$ denotes the HAC estimator (2) evaluated using $x_t = \frac{\Delta L_t}{\hat{\sigma}_t^2}$.

Next, we derive the asymptotic null distributions of the two new feasible statistics DM' and DM^*, for which we make the following further assumptions:

Assumption 3. The forecast methods are based on the rolling window scheme with fixed window sizes w_i satisfying $w_i \leq \bar{w} < \infty$, $i = 1, 2$.

Assumption 4. The volatility function $\sigma(.)$ is continuously differentiable. It is uniformly bounded by a constant M on $[0, 1]$ and $\int_0^1 \sigma(\tau)d\tau < \infty$.

Assumption 5. The kernel function $k(.)$ is symmetric, satisfies $|k(.)| \leqslant 1$ and $k(0) = 1$ and is continuous at zero and almost everywhere else. The kernel function also satisfies $\int_{-\infty}^{\infty}|k(x)|dx < \infty$ and $\int_{-\infty}^{\infty}|\phi_k(x)|dx < \infty$, where $\phi_k(x) = (2\pi)^{-1}\int_{-\infty}^{\infty}k(y)e^{-ixy}dy$.

Assumption 6. The lag truncation parameter b satisfies $b \to \infty$ and $n^{-1/2}b \to 0$ as $n \to \infty$.

Assumption 7. The kernel function $K(.) : \mathbb{R} \to \mathbb{R}^+$ is bounded, Lipschitz continuous and satisfies $\int_{-\infty}^{\infty}K(x)dx > 0$, $\int_{-\infty}^{\infty}|K(x)x|dx < \infty$, $\int_{-\infty}^{\infty}|K(x)|dx < \infty$, $xK(x) \to 0$ as $x \to \infty$, and $\int_{-\infty}^{\infty}|\phi_K(x)|dx < \infty$, where $\phi_K(x) = (2\pi)^{-1}\int_{-\infty}^{\infty}K(y)e^{-ixy}dy$. K is differentiable, with $K'(.)$ bounded over $\mathbb{R}$, and $\int_{-\infty}^{\infty}|\phi_{K'}(x)|dx < \infty$, where $\phi_{K'}(x) = (2\pi)^{-1}\int_{-\infty}^{\infty}K'(y)e^{-ixy}dy$.

Assumption 8. The bandwidth h satisfies $h \to 0$ and $nh^4 \to \infty$ as $n \to \infty$.

Our Assumption 3 only admits a rolling window scheme, thereby excluding a recursive estimation scheme. This is comparable with GW's corresponding assumptions in view of the discussion in McCracken (2020). Assumption 4 imposes a smoothness condition for the unconditional volatility function; the commonly used logistic smooth transition function and trend function for unconditional volatility both satisfy Assumption 4. Notice that the smoothness assumption does not necessarily imply slow and gradual changes in the unconditional volatility function. For example, a logistic transition function with a large speed parameter permits very rapid volatility changes, while still satisfying Assumption 4.[1] Assumptions 5 and 6 are largely from De Jong and Davidson (2000) and are needed for consistent estimation of the long-run

[1]We conjecture that it is possible to extend our theory to allow for a finite number of discontinuities in the volatility function, by utilising and adapting the proof strategy in Xu and Phillips (2008), Cavaliere et al. (2022) and Boswijk and Zu (2022). However, such an extension is out of the scope of this

variances. Our Assumption 6 is stronger than that used in De Jong and Davidson (2000)[2]; this is needed to prove the validity of our HAC estimators involving adjustments made by the nonparametric estimator $\hat{\sigma}_t$.

In the following theorem, the asymptotic distributions of the statistics are established under the null hypothesis.

Theorem 1.

a. *Under Assumptions 1–8, and assuming also that* $\text{Var}\left(n^{-1/2}\sum_{t=R+1}^{T}\frac{\Delta L_t}{\sigma_t}\right) > 0$ *and* $\text{Var}\left(n^{-1/2}\sum_{t=R+1}^{T}\frac{\Delta L_t}{\sigma_t^2}\right) > 0$ *for large n, and* $E(|\Delta L_t|^{2r}) < \infty$ *for all* $R+1 \leqslant t \leqslant T$, *then under* H_0, *as* $n \to \infty$,

$$DM' \xrightarrow{d} N(0,1),\ DM^* \xrightarrow{d} N(0,1).$$

b. *Under Assumptions 1–8, and assuming also that* $\text{Var}\left(n^{-1/2}\sum_{t=R+1}^{T}\Delta L_t\right) > 0$ *for large n, and* $E(|\Delta L_t|^{r}) < \infty$ *for all* $R+1 \leqslant t \leqslant T$, *then under* H_0, *as* $n \to \infty$,

$$DM \xrightarrow{d} N(0,1).$$

The result for *DM* is the same as that in Theorem 4 of GW, but we make a slightly less stringent moment assumption $E(|\Delta L_t|^{r}) < \infty$ here as GW require $E(|\Delta L_t|^{2r}) < \infty$.

4 | LOCAL ASYMPTOTIC POWER ANALYSIS

To study the power performance of the tests, in this section, we look at their asymptotic powers under the local alternative hypothesis H_1, where we apply the relevant Pitman drift to c:

$$H_1 : E(\Delta L_t) = \frac{c}{\sqrt{n}} \tag{8}$$

where, without loss of generality, we consider $c > 0$.[3] The limits of our new DM-type statistics DM' and DM^*, and also the original DM statistic *DM*, under the local alternative (8), are given in the following theorem.

Theorem 2.

a. *Under Assumptions 1–8, denoting* $\xi_n^2 = \text{Var}\left(n^{-1/2}\sum_{t=R+1}^{T}\frac{\Delta L_t}{\sigma_t}\right)$ *and* $\zeta_n^2 = \text{Var}\left(n^{-1/2}\sum_{t=R+1}^{T}\frac{\Delta L_t}{\sigma_t^2}\right)$, *then*

$$\lim_{n\to\infty}\xi_n^2 = \xi^2,\ \lim_{n\to\infty}\zeta_n^2 = \zeta^2,$$

where ξ^2 *and* ζ^2 *are finite. Further, if* $\xi_n^2 > 0$ *and* $\zeta_n^2 > 0$ *for large n, and* $E(|\Delta L_t|^{2r}) < \infty$ *for all* $R+1 \leqslant t \leqslant T$, *then under the local alternative* H_1 *of (8), as* $n \to \infty$,

$$DM' \xrightarrow{d} \frac{c}{\xi}\int_0^1 \frac{1}{\sigma(\tau)}\mathrm{d}\tau + N(0,1),\ DM^* \xrightarrow{d} \frac{c}{\zeta}\int_0^1 \frac{1}{\sigma^2(\tau)}\mathrm{d}\tau + N(0,1).$$

b. *Under Assumptions 1–8, denoting* $\psi_n^2 = \text{Var}\left(n^{-1/2}\sum_{t=R+1}^{T}\Delta L_t\right)$, *then*

$$\lim_{n\to\infty}\psi_n^2 = \psi^2,$$

paper and thus left for future research. Instead, we provide some additional Monte Carlo simulation evidence that the performance of our tests is not affected when the volatility function is discontinuous—see Appendix S1 for details.

[2]The condition imposed on the lag truncation parameter in De Jong and Davidson (2000) translates to $n^{-1}b \to 0$ if their near-epoch dependent array assumption is specialised to our α-mixing case.

[3]Clark and McCracken (2015) also consider a local alternative hypothesis when comparing equal forecast accuracy between nested models.

where ψ^2 is finite. Further, if $\psi_n^2 > 0$ for large n, and $E(|\Delta L_t|^r) < \infty$ for all $R+1 \leqslant t \leqslant T$, then under the local alternative H_1 of (8), as $n \to \infty$,

$$DM \xrightarrow{d} \frac{c}{\psi} + N(0,1).$$

In this theorem, we first note that under our mixing conditions, the long-run variances ξ^2, ζ^2 and ψ^2 are finite. This allows us to obtain analytical expressions for the limiting local alternative distributions of the test statistics. Note also that the local asymptotic distribution of *DM* is a new result.

From the results of Theorem 2, we see that under the local alternative model (8), the limit distributions of the DM', DM^* and *DM* statistics have different location shifts relative to their common standard normal limit null distribution. Because the magnitude of the location shift dictates the local asymptotic powers of the tests, it is informative to compare the relative location shift magnitudes.

Under the current set of assumptions, it is not straightforward to directly compare the magnitudes of the different location shifts. However, in the case when the sequence $\{(\Delta L_t - cn^{-1/2})/\sigma_t\}$ is covariance stationary, we can derive explicit expressions for ξ^2, ζ^2 and ψ^2 and the relationships between them. From the derived relationships, we can further show that there is a fixed order between the magnitudes of the three location shifts. This is detailed in the following proposition. Note that the sequence $\{(\Delta L_t - cn^{-1/2})/\sigma_t\}$ already has a constant mean 0 and variance 1 under our assumptions; hence, the restriction that the sequence is covariance stationary only additionally imposes that the covariance structure does not change over time.

Proposition 1. *Under the conditions of Theorem 2 and under the local alternative H_1 of (8), if the sequence $\{(\Delta L_t - cn^{-1/2})/\sigma_t\}$, $t = R+1, \ldots, T$, is covariance stationary with lth-order autocovariance γ_l, then its long-run variance, $\gamma_0 + 2\sum_{l=1}^{\infty}\gamma_l$, is finite. Further, as $n \to \infty$,*

$$\xi^2 = \gamma_0 + 2\sum_{l=1}^{\infty}\gamma_l,\ \psi^2 = \xi^2\int_0^1 \sigma^2(\tau)\mathrm{d}\tau,\ \zeta^2 = \xi^2\int_0^1 \frac{1}{\sigma^2(\tau)}\mathrm{d}\tau.$$

Denoting the location shifts for DM', DM^* and DM in Theorem 2 by $L_{DM'}$, L_{DM^*} and L_{DM}, respectively, then

$$L_{DM'} = \frac{c}{\xi}\int_0^1 \frac{1}{\sigma(\tau)}\mathrm{d}\tau,\ L_{DM^*} = \frac{c}{\xi}\sqrt{\int_0^1 \frac{1}{\sigma^2(\tau)}\mathrm{d}\tau},\ L_{DM} = \frac{c}{\xi}\frac{1}{\sqrt{\int_0^1 \sigma^2(\tau)\mathrm{d}\tau}}, \tag{9}$$

and

$$L_{DM^*} \geqslant L_{DM'} \geqslant L_{DM},$$

with the equalities holding when the $\sigma(.)$ function is a constant.

The implication of Proposition 1 is that when $\{(\Delta L_t - cn^{-1/2})/\sigma_t\}$ is covariance stationary, a power ranking exists between the tests under the local alternatives when unconditional heteroskedasticity exists, with the DM^* test being the most powerful, followed by DM', and *DM* being the least powerful.

Based on the result of Proposition 1, we now proceed to evaluate numerically the local asymptotic powers of the different tests for a range of volatility specifications, in order to illustrate the relative performance of the tests. We consider the following four representative volatility functions for $\sigma(.)$:

(i) Constant volatility: $\sigma(\tau) = \sigma_1\ \ \forall\tau$.
(ii) Smooth transition in volatility from σ_1 to σ_2:

$$\sigma(\tau) = \sigma_1 + \frac{\sigma_2 - \sigma_1}{1 + \exp\{-30(\tau - 0.4)\}}.$$

(iii) Smooth transition in volatility from σ_2 to σ_1:

$$\sigma(\tau) = \sigma_2 + \frac{\sigma_1 - \sigma_2}{1 + \exp\{-30(\tau - 0.4)\}}.$$

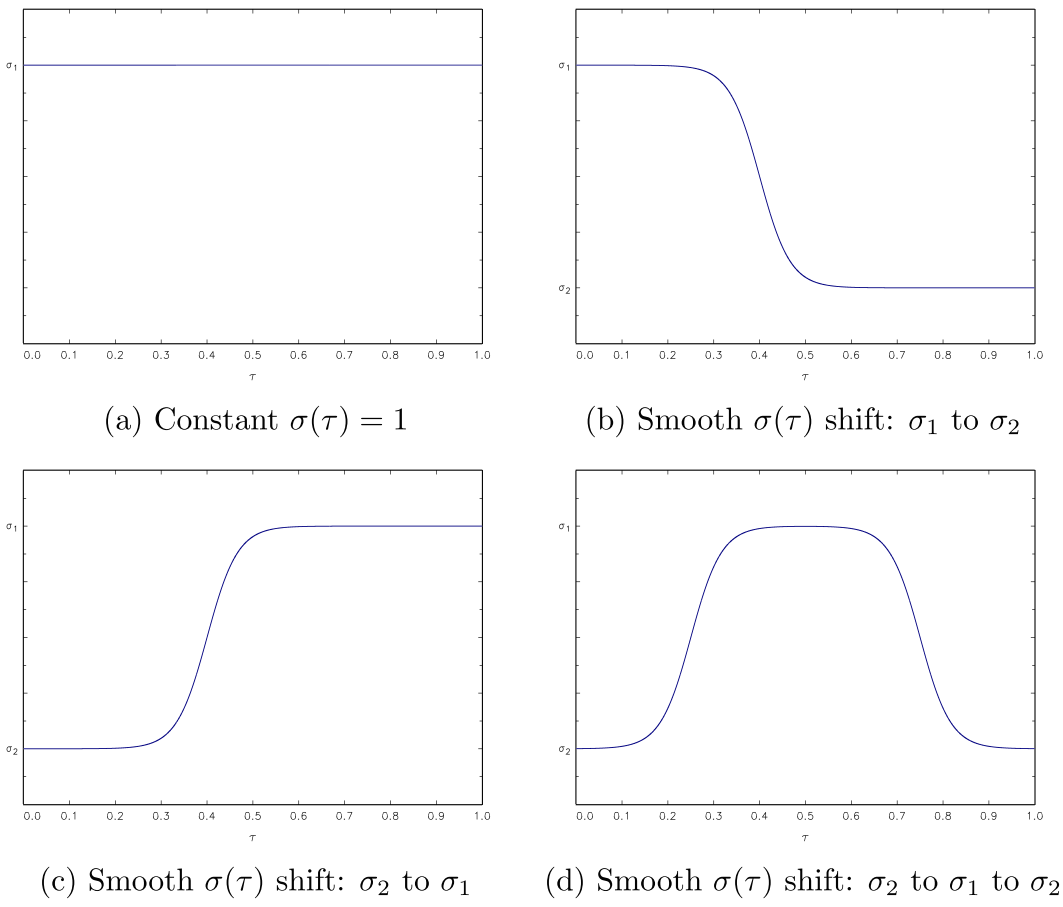
(iv) Smooth double transition in volatility from σ_2 to σ_1 to σ_2:

$$\sigma(\tau) = \sigma_2 + \frac{\sigma_1 - \sigma_2}{1 + \exp\{-30(\tau - 0.25)\}} + \frac{\sigma_2 - \sigma_1}{1 + \exp\{-30(\tau - 0.75)\}}.$$

Here, we adopt the logistic function to model smooth transitions in volatility, and initially, we set $\sigma_1 = 1$ and $\sigma_2 = 1/5$, so that, for example, (ii) comprises a smooth transition from 1 to 1/5 with transition speed 30 and transition mid-point 0.4. Figure 1 gives plots of the four volatility functions. Using these $\sigma(\tau)$ functions, we evaluate the local asymptotic power of the tests using the offset representations in (9) for a grid of c values ranging from 0 to 4 with step size 0.1 ($c = 0$ representing the null). The integrals in (9) are calculated by numerical approximation involving 10,000 discretised steps, and we set $\xi = \sqrt{2.696}$, motivated by the ARMA(1,1) specification adopted in the finite sample simulations that follow in the next section. For a given c, ξ and $\sigma(\tau)$ function, the offsets in (9) can be calculated, and the local asymptotic powers of the tests evaluated using the cumulative distribution function of the normal distribution.

Figure 2 presents the local power results for two-sided nominal 0.05-level tests for the four volatility functions. In panel (a), the local power curves of DM', DM^* and DM coincide, because the offsets reduce to $\frac{c}{\xi}$ for all three tests. Note that this implies no loss of power through using the heteroskedasticity-adjusted DM' and DM^* tests relative to DM. In panels (b)–(d), substantial differences are observed between the three local power profiles, and a similar pattern is observed across all three volatility functions. In line with the result in Proposition 1, we see that DM has the lowest powers of the three tests under heteroskedasticity. The new tests DM' and DM^* offer considerable power gains relative to DM through their direct accounting of the heteroskedastic features of the loss function. Between DM' and DM^*, the power differences are more modest, but DM^* dominates DM' in terms of power, again in line with Proposition 1.

In order to compare the local powers across different magnitudes of volatility change, we next present plots of the local powers of two-sided nominal 0.05-level tests across a range of σ_2 values, for the same three time-varying volatility functions (ii), (iii), and (iv) above. Specifically, we consider $\sigma_2 = 1/x$ for $x = \{5, 4.9, \ldots, 1\}$, and to aid comparison across the different tests, we calibrate c for a given value of σ_2 so that the power of DM is equal to 0.50, that is, for each σ_2, we



(a) Constant $\sigma(\tau) = 1$ (b) Smooth $\sigma(\tau)$ shift: σ_1 to σ_2

(c) Smooth $\sigma(\tau)$ shift: σ_2 to σ_1 (d) Smooth $\sigma(\tau)$ shift: σ_2 to σ_1 to σ_2

FIGURE 1 Volatility function specifications.

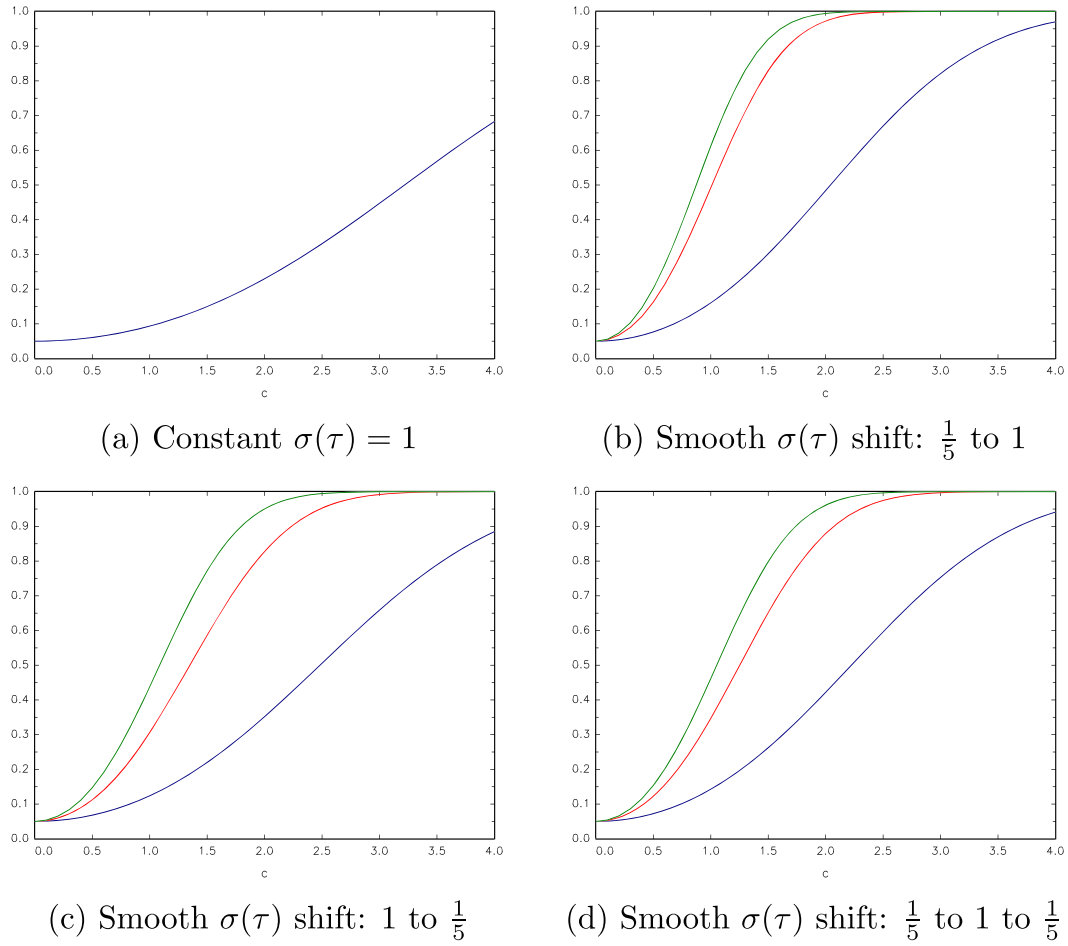


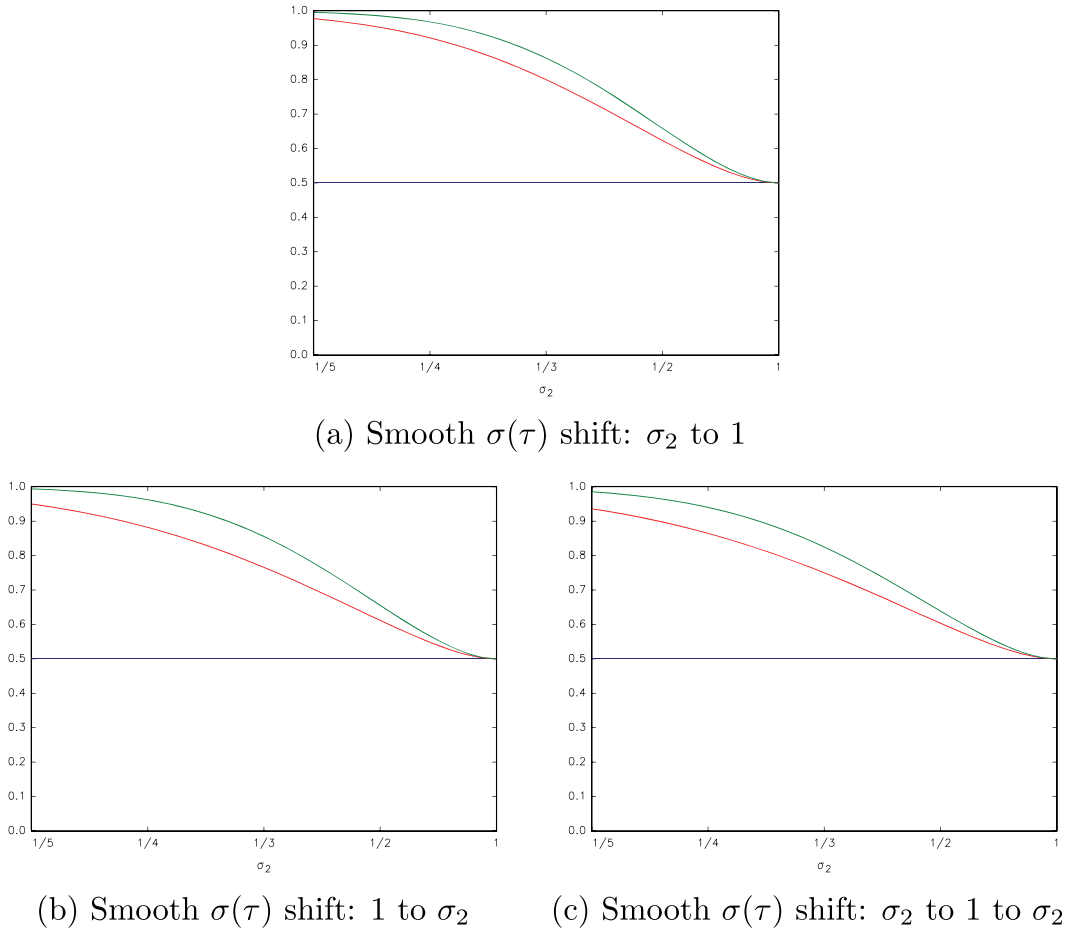
FIGURE 2 Local asymptotic powers of two-sided nominal 0.05-level tests: DM: —; DM': —; DM^*: —.

set $c = 1.96\xi\sqrt{\int_0^1 \sigma^2(\tau)\mathrm{d}\tau}$ so that $L_{DM} = 1.96$. The results are given in Figure 3, and it is clear that the power differences between the different tests become more exaggerated as the magnitude of the volatility change increases. The powers of the tests coincide when $\sigma_2 = 1$ because in that case, no volatility change occurs, but then as σ_2 decreases towards the largest change considered (1/5), the power gains of DM' and DM^* over DM quickly become evident, with marked differences in power levels apparent even for relatively small changes in volatility. Overall, the limiting power results suggest a valuable role for the new tests, and DM^* in particular, when heteroskedasticity is present in the forecast evaluation sample.

While the GW framework focuses on evaluating forecasting methods using a rolling window model estimation scheme, the original DM forecast evaluation framework abstracts from forecast models and imposes assumptions directly on the loss differential series. It is straightforward to translate our asymptotic treatment of the DM and newly proposed statistics DM' and DM^* to this case: we would simply impose Assumption 1 on $\{\Delta L_t\} \equiv \{L(y_{t+q}, f_{1t}) - L(y_{t+q}, f_{2t})\}$ instead of $\{W_t\}$ (note that Assumption 3 then becomes irrelevant). Our results in Theorems 1 and 2 would continue to apply, and under the additional conditions of Proposition 1, the same local asymptotic power rankings of the tests would arise. Hence, while our primary analysis is set within the GW framework, the central results are equally applicable to the DM forecast comparison environment.

5 | FINITE SAMPLE SIMULATIONS

In this section, we perform Monte Carlo simulations to study the finite sample performance of the DM', DM^* and DM tests. Under the null hypothesis, we directly simulate $\Delta L_t = \sigma_t z_t$, while under the local alternative, we simulate $\Delta L_t = cn^{-1/2} + \sigma_t z_t$ for $t = 1, \ldots, n$ with $n = \{100, 200, 400\}$ and the same grid of c values considered in the previous section. For z_t, we use the following normalised stationary invertible ARMA(1,1) specification:



(a) Smooth $\sigma(\tau)$ shift: σ_2 to 1

(b) Smooth $\sigma(\tau)$ shift: 1 to σ_2

(c) Smooth $\sigma(\tau)$ shift: σ_2 to 1 to σ_2

FIGURE 3 Local asymptotic powers of two-sided nominal 0.05-level tests, c calibrated across σ_2: DM: —; DM': —; DM^*: —.

$$z_t = \frac{u_t}{\sqrt{\text{Var}(u_t)}},\ u_t = 0.3u_{t-1} + \varepsilon_t + 0.5\varepsilon_{t-1},\ \varepsilon_t \sim IIDN(0,1).$$

Here, z_t is normalised to have unit variance, so the loss differential ΔL_t has variance σ_t^2. Note that for this process, the long-run variance of z_t is $\xi^2 = 2.696$; hence, our specification for z_t is consistent with the setting for ξ used in the local asymptotic power calculations above. For σ_t, we use the same four volatility specifications as in the previous section, focusing on the case $\sigma_2 = 1/5$ and discretising $\sigma(\tau)$ as $\sigma(t/n)$ in the obvious way. Using these data generating processes (DGPs) ensures that our finite sample analysis will be directly comparable with our preceding local asymptotic work.

For the nonparametric estimator $\hat{\sigma}^2(.)$, we use the Gaussian kernel throughout. For the bandwidth h, we use a cross-validation method. Given the dependent nature of the z_t series, the classical leave-one-out cross-validation procedure performs poorly, often leading to a (left) boundary solution and thereby selecting a very small bandwidth. This phenomenon is well known in the statistics literature, and Hardle and Vieu (1992) (see also Chu & Marron, 1991; Hart & Vieu, 1990; Tong & Yao, 1998) propose instead using leave $2l + 1$ out cross-validation to deal with the dependence. That is, in addition to giving zero weight to the current observation t, it also assigns zero weight to l observations before and after time t, to compute a leave $2l + 1$ out estimator for time t, which is then further used to evaluate the leave $2l + 1$ out cross-validation criterion function. We apply this criterion to select h using $l = 2$, using a 100 point grid of possible h values over the range $5/n$ to 0.5, as this was found to provide good bandwidth selection. For the long-run variance estimators $\hat{\Omega}(.)$, we use a Bartlett kernel with lag truncation parameter $b = \lfloor 1.2n^{1/3} \rfloor$, with $n^{1/3}$ being the optimal rate for the Bartlett kernel, as discussed in Andrews (1991). Figures 4–6 present results for the sizes and powers of two-sided nominal 0.05-level tests for $n = 100$, 200 and 400, respectively, using 10,000 replications.

The stand-out feature from Figures 4–6 is that the patterns of power behaviour bear a very close resemblance to the corresponding local asymptotic results in Figure 2, and a clear movement towards the limit results is seen as the sample size n increases. When $c = 0$, we observe very little in the way of finite sample size distortion, even for the smallest sample

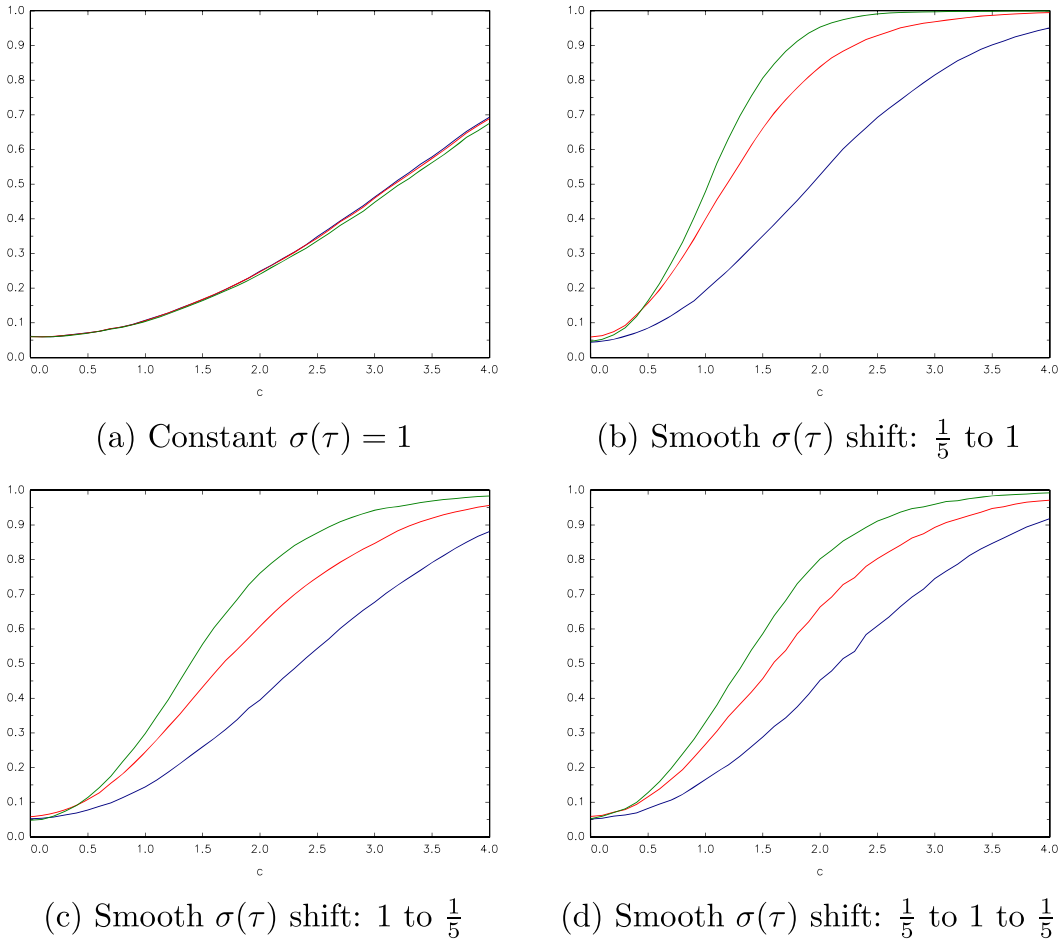


FIGURE 4 Finite sample powers of two-sided nominal 0.05-level tests, $n = 100$: DM: —; DM': —; DM^*: —.

size $n = 100$, implying that the feasible DM' and DM^* tests behave reliably under the null in finite samples, along with DM. For $c > 0$, the power gains of DM' and DM^* over DM observed in the asymptotic context are seen to carry over into finite samples, with substantial gains again available under heteroskedasticity. Moreover, DM^* dominates in terms of power in all the finite samples considered. These results confirm the usefulness of the new heteroskedasticity-adjusted procedures in delivering more powerful tests for equal forecast accuracy.[4]

6 | DISCUSSIONS AND EXTENSIONS

6.1 | Comparing forecasting models

West (1996) and West and McCracken (1998) consider a closely related problem of testing equal performance between *forecasting models*. In particular, their analysis, adapted to our context, focuses on testing for a zero mean of the loss differential

$$\{\Delta L_t(\beta^*)\}_{t=R+1}^{T} \equiv \{L(y_{t+q}, f_{1t}(\beta_1^*)) - L(y_{t+q}, f_{2t}(\beta_2^*))\}_{t=R+1}^{T},$$

where $\beta^* = (\beta_1^{*\prime}, \beta_2^{*\prime})'$ with β_1^* and β_2^* the pseudo-true values of the forecasting models, obtained by projecting the DGP on the two considered models, respectively. Note that the forecasts $f_{1t}(\beta_1^*)$ and $f_{2t}(\beta_2^*)$ depend on data at time t or earlier. Because the hypothesis concerns the pseudo-true models (characterised by the model specification and the pseudo-true

[4]In Appendix S1, we also provide additional finite sample simulation evidence to illustrate the size robustness and power properties of the procedures when (i) conditional heteroskedasticity is present in the data and (ii) instantaneous level shifts occur in the volatility function. We find that the size remains close to the nominal level, and the power rankings of the tests are similar to those reported in this section.

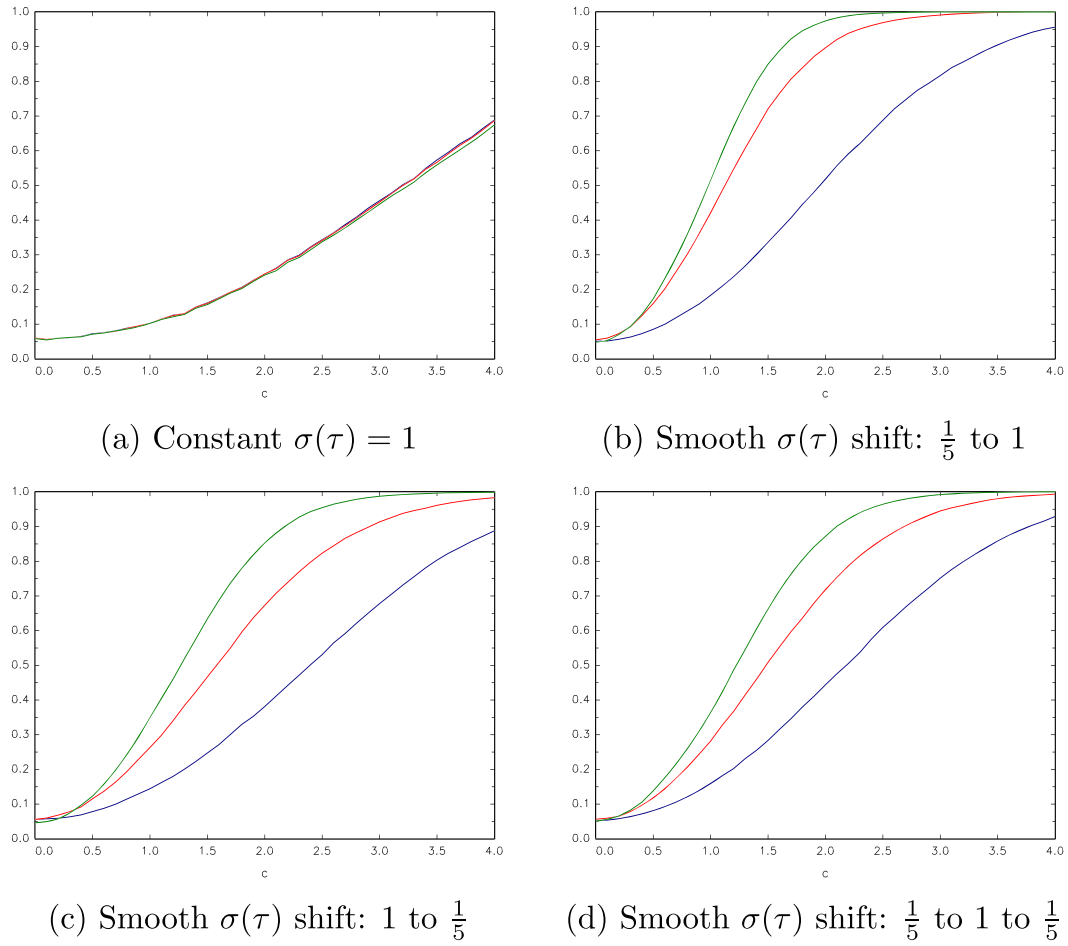


FIGURE 5 Finite sample powers of two-sided nominal 0.05-level tests, $n = 200$: DM: —; DM': —; DM^*: —.

parameter values), it is interpreted as a hypothesis for the equality of forecasting performance of the two models. In contrast to the GW approach, where the comparison of forecasting methods includes the effect of both model specification *and* model parameter estimation, the West-type hypothesis considers the model specification part only. The two approaches can be complementary; for example, in the case that we find a difference in the forecasting performance of two forecast methods using a GW approach, testing a West-type hypothesis can help us in further understanding whether the difference comes from the models being used, excluding the effects of the parameter estimation uncertainty.

To test the West-type hypothesis that $E(\Delta L_t(\beta^*)) = 0$ in our heteroskedastic setting, we could consider extending our new statistics along the same lines as West (1996) and West and McCracken (1998). First, consider the infeasible statistic (3) where σ_t is treated as known. Let the model estimates be denoted by $\hat{\beta}_t = \left(\hat{\beta}'_{1t}, \hat{\beta}'_{2t}\right)'$, where these estimates are obtained from a fixed, rolling window or recursive estimation scheme. It is straightforward to see that one can make a West-type Taylor expansion of the average

$$n^{-1} \sum_{t=R+1}^{T} \frac{\Delta L_t(\hat{\beta}_t)}{\sigma_t},$$

around the loss differential evaluated at the pseudo-true value β^*, leading to the following West-type asymptotic null distribution result under a set of high-level assumptions for the parametric estimation stage, and assuming stationarity in the data:

$$n^{-1/2} \sum_{t=R+1}^{T} \frac{\Delta L_t(\hat{\beta}_t)}{\sigma_t} \xrightarrow{d} N(0, \Omega_\sigma),$$

where $\Omega_\sigma = S_{1\sigma} + \Pi_1 \left(F_\sigma B S'_{2\sigma} + S_{2\sigma} B' F'_\sigma\right) + \Pi_2 F_\sigma V_\beta F'_\sigma$. Here, $S_{1\sigma}$ is the long-run variance of $\Delta L_t(\beta^*)/\sigma_t$, $S_{2\sigma}$ is the long-run cross-covariance matrix between $\Delta L_t(\beta^*)/\sigma_t$ and the moment function used for parameter estimation and

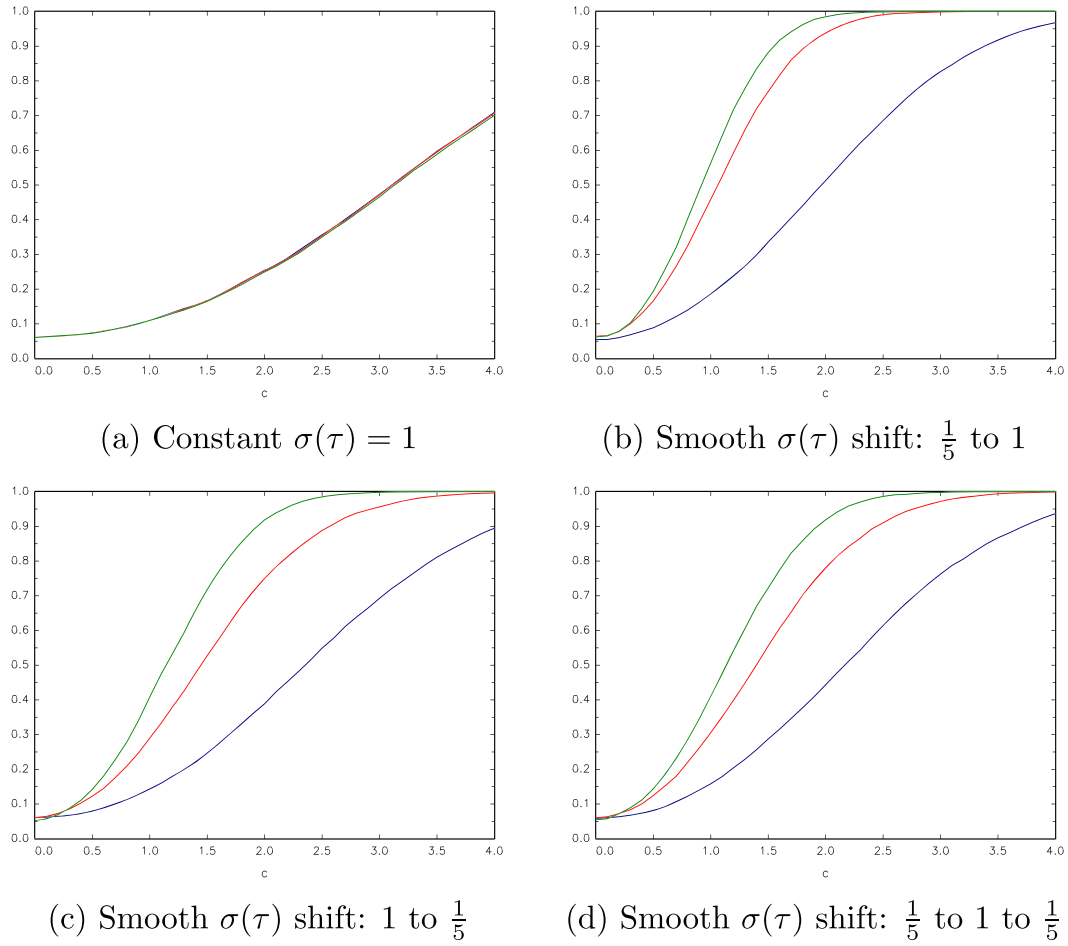


FIGURE 6 Finite sample powers of two-sided nominal 0.05-level tests, $n = 400$: DM: —; DM': —; DM^*: —.

$F_\sigma = E\left(\frac{\partial\{\Delta L_t(\beta^*)/\sigma_t\}}{\partial\beta}\right)$. The remaining terms are the same as in West (1996) and West and McCracken (1998) as they only involve the parametric estimation stage: V_β denotes the limiting variance–covariance matrix of the parameter estimator, B is the inverse of the limit of the Hessian matrix associated with the moment function used to estimate the model and Π_1 and Π_2 are constants, the values of which depend on which of the fixed, rolling window or recursive schemes are employed for model estimation. Further, note that the West (1996) framework already accommodates possible estimation methods that are robust to any heteroskedasticity in the model estimation stage; hence, no further heteroskedasticity adjustment would be needed in that part of the environment.

In practice, using our nonparametric estimator $\hat{\sigma}_t$, a variant of DM' in (7) could be constructed to compare forecast models using $\Delta L_t\left(\hat{\beta}_t\right)$ in place of ΔL_t and replacing $\hat{\Omega}\left(\frac{\Delta L}{\hat{\sigma}}\right)$ with an estimate of Ω_σ based on $\Delta L_t\left(\hat{\beta}_t\right)/\hat{\sigma}_t$. A variant of DM^* could also be considered in a similar manner. Of course, the introduction of the estimator $\hat{\sigma}_t$ would require a non-trivial development of the West-type theory, because the potential effects of using $\hat{\sigma}_t$ instead of σ_t would need to be evaluated. We leave such an extension for future research.

6.2 | Time-varying loss differential mean

In this paper, we consider a constant mean model $E(\Delta L_t) = c$ for the loss differential series and test the hypothesis $H_0 : c = 0$ against $H_1 : c \neq 0$. The constant mean is a maintained *assumption* under both the null and the alternative *hypotheses*. Here, we make a distinction between the *model* and the *assumption* we make on the model, which characterises the framework in which we perform our tests, and the *hypotheses* we wish to test within this framework. The stationary testing framework considered by DM and West (1996) also implies a constant mean for the loss differential series.

Although the assumption of a constant mean is widely made in practical applications, as evidenced by the popularity of the DM and West tests, there are certainly cases where a time-varying mean might be more relevant. Consider forecast evaluation testing within a model where the loss differential has a time-varying mean:

$$E(\Delta L_t) = c_t. \tag{10}$$

In this context, the equal performance hypothesis between the two forecasts becomes

$$H_0 : c_t = 0, \ t = R+1, \ \dots, T. \tag{11}$$

Testing (11) within model (10) has been the main subject of more recent research in the forecast evaluation literature. For example, Odendahl et al. (2023) consider testing the null hypothesis (11) against the alternative that the relative performance is state dependent.[5] Earlier, Giacomini and Rossi (2010) proposed use of a local relative forecast performance measure, adopting a fluctuation type statistic to test the null hypothesis (11), and subsequently track changes in relative forecast performance when the null is rejected. Rossi and Sekhposyan (2010) also apply the Giacomini and Rossi (2010) test to study the relative performance of forecasts from various economic models.

A fact that is perhaps less well documented is that GW were actually the first to consider testing the hypothesis (11) in the time-varying mean model (10). The unconditional test of GW proposes use of the DM *test statistic* to test the null hypothesis (11). Amisano and Giacomini (2007) apply the same unconditional test to the problem of evaluating density forecasts.

It is important to note that once a more general model of the form (10) is entertained, the parameter space under consideration is substantially enlarged relative to the constant mean model. In the constant mean case, interest focuses on whether $c = 0$ within a one-dimensional real parameter space $c \in \mathbb{R}$, while in the time-varying mean model (10), the approach of GW is to test if $(c_{R+1}, \dots, c_T)$ is a zero vector within an n-dimensional real parameter space $(c_{R+1}, \dots, c_T) \in \mathbb{R}^n$, which will become infinite-dimensional as $n \to \infty$. When this null hypothesis is violated, there are many possibilities for the behaviour of the c_t, and the DM statistic may not have power against all possible departures from the null. For example, consider the case where $c_t \neq 0$ (for at least some $t = R+1, \dots, T$) but the average loss differential $n^{-1}\sum_{t=R+1}^{T} c_t = 0$. This is certainly a deviation from the null hypothesis (11), but the DM statistic, the construction of which is based on the average of loss differentials, will be unlikely to have power because the average loss differential is zero. Hence, when two different forecasts perform *equally on average*, this is unlikely to be detected by the unconditional test of GW (this is referred to as a power 'blind spot' in the terminology of Li et al., 2022). This is perhaps why GW only establish consistency for their unconditional test under a non-exhaustive alternative hypothesis of different *average* forecast performance, leaving other possible deviations from their null hypothesis unstudied.

The two DM-type statistics proposed in this paper are based on the unconditional volatility (or variance) function-reweighted average of the loss differential series, that is, $n^{-1}\sum_{t=R+1}^{T} c_t/\sigma_t \left(\text{or } n^{-1}\sum_{t=R+1}^{T} c_t/\sigma_t^2\right)$. In a homoskedastic setting, if used to test the hypothesis (11) in a time-varying mean model, these statistics will have similar properties to the DM statistic, as outlined above, and therefore also lack power against an alternative of the form $n^{-1}\sum_{t=R+1}^{T} c_t = 0$. In a heteroskedastic environment, however, the behaviour of the new statistics DM' and DM^* diverges from that of DM, and hence, there are different regions of the alternative hypothesis parameter space where the new statistics will have power and different regions where 'blind spots' may exist. For example, it would be expected that the new tests will lack power against alternatives to (11) for which $n^{-1}\sum_{t=R+1}^{T} c_t/\sigma_t = 0 \left(\text{or } n^{-1}\sum_{t=R+1}^{T} c_t/\sigma_t^2 = 0\right)$, but they may well have power against the alternative where $n^{-1}\sum_{t=R+1}^{T} c_t = 0$, in contrast to the DM statistic. Therefore, the two new DM-type statistics that we propose could be viewed as complementary to DM (i.e., to the GW unconditional test), removing some of the 'blind spots' where their test lacks power. In fact, any DM-type statistic computed from a non-trivially 'reweighted' original loss differential series could be used for this purpose. One only has to be careful with the interpretation of a rejection observed using such reweighted test statistics: for example, in the context of a time-varying mean model with heteroskedasticity, a rejection of (11) within model (10) by our new statistics can only be interpreted as 'a difference in performance at some point(s) in time', rather than 'a certain forecast is better or worse', as the signs of the statistics, being based on $n^{-1}\sum_{t=R+1}^{T} c_t/\sigma_t$ or $n^{-1}\sum_{t=R+1}^{T} c_t/\sigma_t^2$, do not carry relevant information about the average

[5]The Odendahl et al. (2023) framework is general and can also be used to test state-dependence for forecast errors and other moments of losses.

relative performance between two forecasts, i.e. the sign of $n^{-1}\sum_{t=R+1}^{T} c_t$. Note that such differences in the signs of the statistics do not occur in the constant mean model $E(\Delta L_t) = c$, which is the primary focus of our paper, because then $sign\left(n^{-1}\sum_{t=R+1}^{T} c/\sigma_t\right) = sign\left(n^{-1}\sum_{t=R+1}^{T} c/\sigma_t^2\right) = sign(c)$.

6.3 | Comparing multiple forecasts

In our main exposition, we consider evaluating two forecasts. In Appendix S1, we discuss issues surrounding extensions of our analysis to a setting where multiple forecasts are compared with a common benchmark forecast.

7 | EMPIRICAL ILLUSTRATION

To illustrate the potential benefits of the new testing procedure, we consider an empirical example of comparing exchange rate forecasts. The performance of competing forecasts of changes in exchange rates has been a topic of extensive study in the literature; see, for example, Rossi (2013) for a comprehensive review. A common finding from such work, beginning with Meese and Rogoff (1983), is that the random walk 'no change' forecast typically outperforms predictors based on economic models. In line with a similar application in DM, we consider evaluating the accuracy of two sets of exchange rate forecasts. The first forecast, denoted f_{1t}, is the prediction implicit in the 3-month forward rate, that is, the difference between the 3-month forward rate and the spot rate. The second forecast, denoted f_{2t}, is the no change forecast implied by a random walk model. The variable to be predicted is the q-month change in the dollar/sterling exchange rate, with the forecasts f_{1t} and f_{2t} made on a monthly basis from 1979:1–2020:12 at four horizons: $q = 1$, 3, 6 and 12. During the period under consideration, the United Kingdom started with a managed floating exchange rate system, before joining

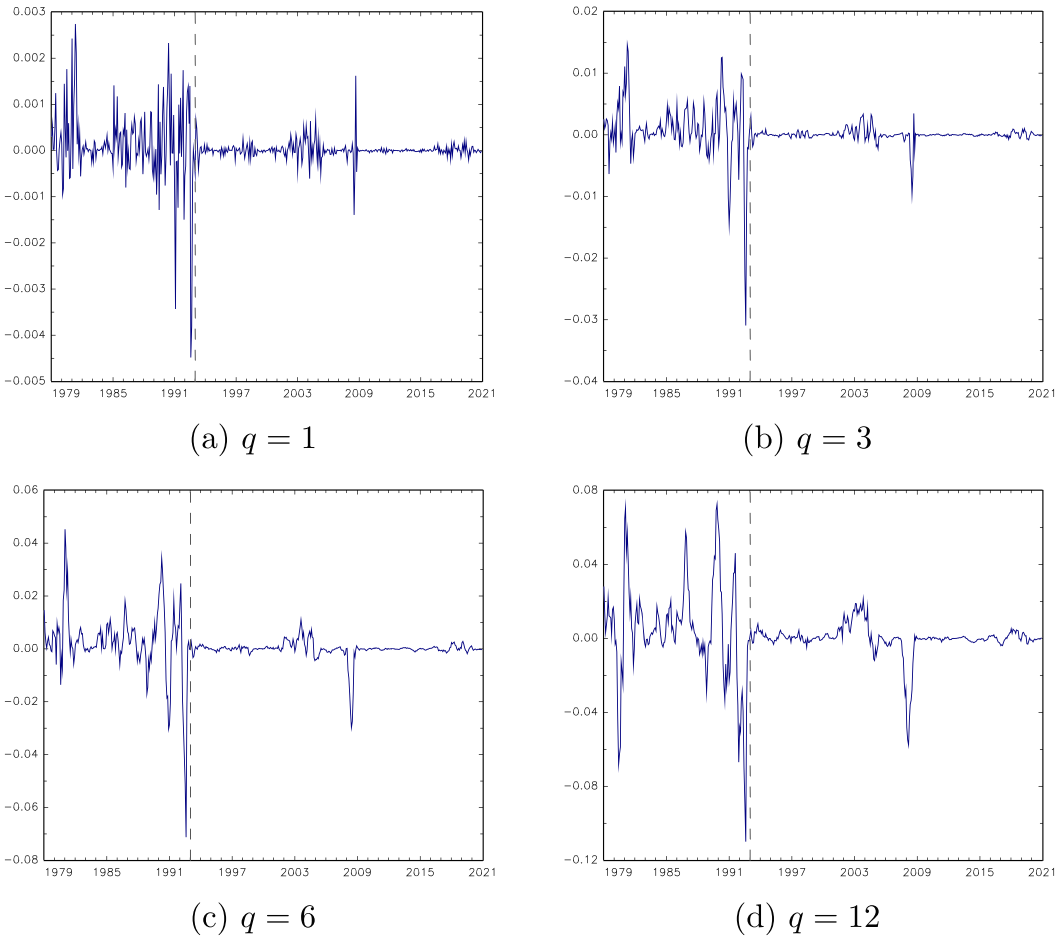


FIGURE 7 Loss differential for q-month-ahead dollar/sterling exchange rate change forecasts (forward–random walk): squared error loss.

the ERM in October 1990. The United Kingdom subsequently left the ERM after two years in September 1992, after which a floating exchange rate system was adopted. Denoting the q-month-ahead forecast errors by $e_{1,t+q}$ and $e_{2,t+q}$, we assess whether there is evidence against equal accuracy for the two sets of forecasts, using two loss measures: (i) squared error loss, that is, $\Delta L_t = e_{1,t+q}^2 - e_{2,t+q}^2$, and (ii) absolute error loss, that is, $\Delta L_t = |e_{1,t+q}| - |e_{2,t+q}|$. Data are obtained from the Bank of England's website.

In applying the DM, DM' and DM^* tests, the kernel function and lag truncation parameter used in the long-run variance estimator are the same as in the Monte Carlo simulations in Section 4, that is, we use the Bartlett kernel for $k(.)$ with $b = \lfloor 1.2n^{1/3} \rfloor$. For the nonparametric variance estimator, we again use the Gaussian kernel for $K(.)$ and determine h by leave $2l+1$ out cross-validation. Here we use $l = 20$ and a 100 point grid of possible h values over an interval of width 0.1, this interval being chosen from initial inspection over a wider range to locate a region where the cross-validation function appears convex.

In Figures 7 and 8, we show the loss differential series over time for the two loss measures and the four forecast horizons. It is clear that, in all cases, the loss differential series exhibit changes in volatility over the sample period, most noticeably following the United Kingdom's departure from the ERM in September 1992, with a considerable decrease in volatility clearly apparent in the post-ERM part of the sample period. The dashed line on each plot separates the full sample period into two parts: forecasts made over the period 1979:1–1992:12 and forecasts made over the period 1993:1–2020:12, with the split point chosen to be a few months after the ERM departure. Within each sub-sample period, changes in volatility are again apparent, suggesting heteroskedasticity is a feature of these loss differentials in both the pre-ERM and post-ERM periods, in addition to differences in volatility between the two periods.

We first consider the results under the squared error loss measure. The values of the test statistics, the associated (two-sided) p-values, based on the $N(0,1)$ distribution, and the selected bandwidth h are given in Table 1. For the full sample period in Panel A, there is very little evidence against the equal accuracy null hypothesis, regardless of which test

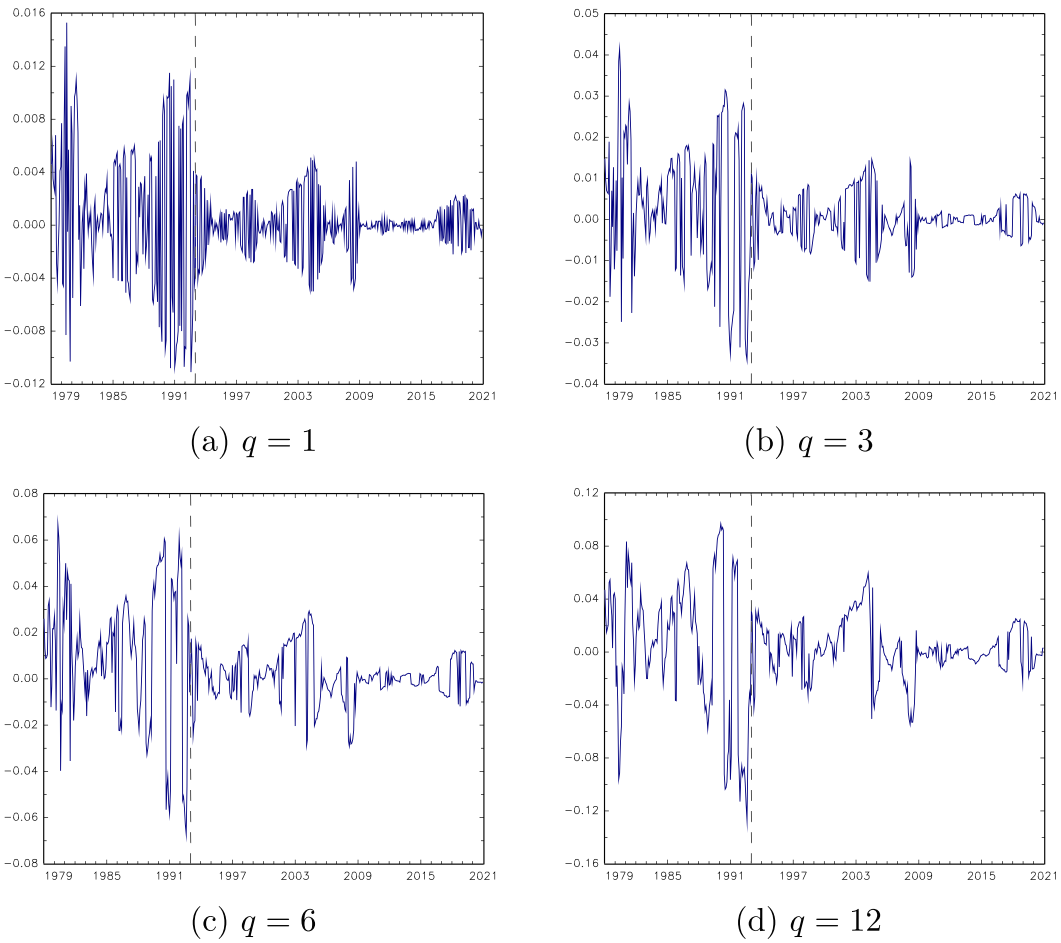


FIGURE 8 Loss differential for q-month-ahead dollar/sterling exchange rate change forecasts (forward–random walk): absolute error loss.

TABLE 1 Application of tests to q-month-ahead dollar/sterling exchange rate change forecasts (forward–random walk): squared error loss.

q	DM	DM'	DM^*	h
Panel A. 1979:1–2020:12				
1	1.745 (0.081)	1.697 (0.090)	1.632 (0.103)	0.291
3	1.350 (0.177)	1.182 (0.237)	0.953 (0.341)	0.300
6	1.087 (0.277)	0.825 (0.410)	0.511 (0.609)	0.300
12	0.714 (0.476)	0.440 (0.660)	0.125 (0.901)	0.272
Panel B. 1979:1–1992:12				
1	1.615 (0.106)	2.201 (0.028)	2.497 (0.013)	0.136
3	1.344 (0.179)	2.274 (0.023)	2.640 (0.008)	0.106
6	1.305 (0.192)	2.291 (0.022)	2.732 (0.006)	0.102
12	0.890 (0.373)	1.500 (0.134)	2.016 (0.044)	0.134
Panel C. 1993:1–2020:12				
1	0.892 (0.373)	1.016 (0.310)	1.051 (0.293)	0.168
3	0.177 (0.859)	0.512 (0.609)	0.916 (0.360)	0.172
6	−0.276 (0.783)	−0.020 (0.984)	0.356 (0.722)	0.187
12	−0.096 (0.924)	0.276 (0.783)	0.820 (0.412)	0.180

Note: p-values are given in parentheses.

TABLE 2 Application of tests to q-month-ahead dollar/sterling exchange rate change forecasts (forward–random walk): absolute error loss.

q	DM	DM'	DM^*	h
Panel A. 1979:1–2020:12				
1	1.815 (0.070)	1.764 (0.078)	1.627 (0.104)	0.261
3	2.373 (0.018)	2.224 (0.026)	1.893 (0.058)	0.261
6	2.681 (0.007)	2.429 (0.015)	1.974 (0.048)	0.265
12	0.980 (0.327)	0.873 (0.383)	0.725 (0.468)	0.262
Panel B. 1979:1–1992:12				
1	1.731 (0.084)	2.160 (0.031)	2.305 (0.021)	0.065
3	2.231 (0.026)	2.828 (0.005)	3.070 (0.002)	0.054
6	2.617 (0.009)	3.143 (0.002)	3.195 (0.001)	0.063
12	0.680 (0.496)	1.516 (0.130)	2.124 (0.034)	0.066
Panel C. 1993:1–2020:12				
1	0.500 (0.617)	0.563 (0.574)	0.685 (0.493)	0.151
3	0.861 (0.389)	0.833 (0.405)	0.761 (0.447)	0.144
6	0.979 (0.327)	0.883 (0.377)	0.728 (0.466)	0.137
12	0.870 (0.384)	0.836 (0.403)	0.775 (0.438)	0.151

Note: p-values are given in parentheses

is considered. This is true across the different forecast horizons, except for $q = 1$ where there is modest evidence at the 0.10 level that the random walk forecast has greater accuracy than the forward rate forecast according to both DM and DM', with DM^* also very close to rejection at this significance level.

In addition to evaluating the forecasts over the full sample period, we also consider the two sub-sample periods 1979:1–1992:12 and 1993:1–2020:12, given the contrasting behaviour of the loss differentials between these pre- and post-ERM departure periods. Panel B provides the results for the pre-1993 period, and here, we find a noticeable difference between the different tests. While DM fails to reject the null of equal accuracy for all four horizons, we find that the new tests provide considerably more evidence in favour of the alternative. Specifically, DM' rejects the null at (less than) the 0.05 level when $q = 1$, 3 and 6, while DM^* rejects at around the 0.01 level in these three horizons and also at the 0.05 level for $q = 12$. This suggests that once tests are used that potentially offer greater power in the presence of heteroskedasticity in the loss differential, evidence is uncovered to indicate a systematic accuracy gain of the random walk forecasts relative to those based on the forward rate. The pattern of rejection/non-rejection across the different tests is in line with what we might expect from our asymptotic and finite sample Monte Carlo analysis, where we showed that DM' and DM^* can offer greater levels of power than DM under heteroskedasticity. We also note that a stronger rejection of the null is associated with DM^* than for DM', again as would be expected from our earlier analysis.

For the post-1993 period in Panel C, all tests indicate no evidence against the null at any forecast horizon. This suggests that, in contrast to the pre-1993 period, the mean squared forecast errors of the two sets of forecasts are equal in this later period of time. It appears, then, that the United Kingdom's departure from the ERM had a substantial effect on the relative predictive accuracy of these competitor forecasts.

Next, we consider the case of absolute error loss, with Table 2 reporting the results in the same format as Table 1. For the full sample, Panel A shows that more evidence exists in favour of the superiority of random walk forecasts when considering absolute error loss compared with squared error loss. While results for $q = 1$ and $q = 12$ are similar across the two loss measures, we now observe rejections of the null by all three tests at conventional significance levels for $q = 3$ and 6. For the earlier pre-ERM sub-sample, results in Panel B show that all three tests now reject for $q = 1$, 3 and 6, but, as was seen in Panel B of Table 1, the p-values decrease as we move from DM to DM' to DM^*, with rejections obtained at lower significance levels for the new DM' and DM^* tests than for DM, particularly for $q = 1$ and 3. When $q = 12$, the same pattern of rejections is obtained as in Panel B of Table 1, with rejection indicated by the DM^* test alone. Finally, for the later post-ERM sub-sample, Panel C shows no rejections of the null by any test, in line with the corresponding results for squared error loss.

Overall, our results show fairly clearly that the random walk forecasts outperform the forward rate forecasts in the pre-ERM period, while the different forecasts appear to have equal accuracy following the change induced by the United Kingdom's ERM departure. The evidence for the random walk forecast superiority in the earlier period is delivered most forcefully by the newly proposed DM' and DM^* tests, with no evidence provided by DM under squared error loss, and more limited evidence under absolute error loss. The rejections obtained from the DM' and DM^* tests are fairly consistent across both loss measures and are obtained for all forecast horizons by DM^* and all except the longest horizon by DM'.

8 | CONCLUSION

In this paper, we have considered the effects of heteroskedasticity on statistical tests for equal forecast accuracy. We proposed two new DM-type tests which explicitly take account of heteroskedasticity through nonparametric estimation of the volatility function. For a quite general class of loss differential series, the heteroskedasticity-adjusted tests we proposed are able to deliver (often substantially) higher levels of power relative to the unmodified DM test when heteroskedasticity is present in the loss differential, while retaining the same levels of power as the original test under homoskedasticity. We demonstrated these features theoretically using a local asymptotic power analysis of the new and original tests. Monte Carlo simulations confirmed a close association between the finite sample and local limit results, implying that our modified tests should work well in empirical settings. This was supported through an empirical illustration using forecasts of changes in the dollar/sterling exchange rate, with the new tests providing greater evidence of differences in the accuracy of competing forecasts than the original DM test. The new procedures should therefore make a valuable addition to the suite of forecast evaluation tests available to practitioners, specifically offering the potential for more reliable detection of departures from the equal accuracy null when heteroskedasticity is a feature of the loss differential.

OPEN RESEARCH BADGES

This article has been awarded Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. Data is available at https://doi.org/10.15456/jae.2023347.1628818346.

DATA AVAILABILITY STATEMENT

The data and code used in the empirical application of this paper have been made available at the Journal of Applied Econometrics Data Archive: https://journaldata.zbw.eu/dataset/tests-for-equal-forecast-accuracy-under-heteroskedasticity-replication-data.

REFERENCES

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

Andersen, T. G. (1994). Stochastic autoregressive volatility: A framework for volatility modeling. *Mathematical Finance*, *4*, 75–102.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, *59*, 817–858.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*, 307–327.

Boswijk, H. P., & Zu, Y. (2022). Adaptive testing for cointegration with nonstationary volatility. *Journal of Business and Economic Statistics*, *40*, 744–755.

Carrasco, M., & Chen, X. (2002). Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory*, *18*, 17–39.

Cavaliere, G., Nielsen, M., & Taylor, A. M. R. (2022). Adaptive inference in heteroscedastic fractional time series models. *Journal of Business and Economic Statistics*, *40*, 50–65.
Chu, C. K., & Marron, J. S. (1991). Comparison of two bandwidth selectors with dependent errors. *Annals of Statistics*, *19*, 1906–1918.
Clark, T. E. (2011). Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility. *Journal of Business and Economic Statistics*, *29*, 327–341.
Clark, T. E., & McCracken, M. (2013). Advances in forecast evaluation. In Elliott, G., & Timmermann, A. (Eds.), *Handbook of economic forecasting*, Part B, Vol. 2: Elsevier, pp. 1107–1201.
Clark, T. E., & McCracken, M. (2015). Nested forecast model comparisons: A new approach to testing equal accuracy. *Journal of Econometrics*, *186*, 160–177.
De Jong, R. M., & Davidson, J. (2000). Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. *Econometrica*, *68*, 407–423.
Diebold, F. X. (2015a). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold-Mariano tests. *Journal of Business and Economic Statistics*, *33*, 1–9.
Diebold, F. X. (2015b). Rejoinder to comments on 'Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold-Mariano tests'. *Journal of Business and Economic Statistics*, *33*, 24.
Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*, 253–263.
Elliott, G., & Timmermann, A. (2016). *Economic forecasting*. Princeton University Press.
Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*, 987–1007.
Giacomini, R., & Rossi, B. (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics*, *25*, 595–620.
Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*, 1545–1578.
Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*, 359–378.
Granger, C. W. J., & Newbold, P. (1977). *Forecasting economic time series*. Academic Press.
Hardle, W., & Vieu, P. (1992). Kernel regression smoothing of time series. *Journal of Time Series Analysis*, *13*, 209–232.
Hart, J. D., & Vieu, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *Annals of Statistics*, *18*, 873–890.
Li, J., Liao, Z., & Quaedvlieg, R. (2022). Conditional superior predictive ability. *Review of Economic Studies*, *89*, 843–875.
McCracken, M. W. (2020). Diverging tests of equal predictive ability. *Econometrica*, *88*, 1753–1754.
Meese, R. A., & Rogoff, K. (1983). Empirical exchange rate models of the seventies. Do they fit out of sample? *Journal of International Economics*, *14*, 3–24.
Meese, R. A., & Rogoff, K. (1988). Was it real? The exchange rate-interest differential relation over the modern floating-rate period. *Journal of Finance*, *43*, 933–948.
Odendahl, F., Rossi, B., & Sekhposyan, T. (2023). Evaluating forecast performance with state dependence. *Journal of Econometrics*, *237*, 105220.
Patton, A. J. (2015). Comment on 'Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold-Mariano tests'. *Journal of Business and Economic Statistics*, *33*, 22–24.
Rossi, B. (2013). Exchange rate predictability. *Journal of Economic Literature*, *51*, 1063–1119.
Rossi, B., & Sekhposyan, T. (2010). Have economic models' forecasting performance for US output growth and inflation changed over time, and when? *International Journal of Forecasting*, *26*, 808–835.
Tong, H., & Yao, Q (1998). Cross-validatory bandwidth selection for regression estimation based on dependent data. *Journal of Statistical Planning and Inference*, *68*, 387–415.
West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, *64*, 1067–1084.
West, K. D. (2006). Forecast evaluation. In Elliott, G., Granger, C. W. J., & Timmermann, A. (Eds.), *Handbook of economic forecasting*, Vol. 1: Elsevier, pp. 99–134.
West, K. D., & McCracken, M. W. (1998). Regression-based tests of predictive ability. *International Economic Review*, *39*, 817–840.
Xu, K. L., & Phillips, P. C. (2008). Adaptive estimation of autoregressive models with time-varying variances. *Journal of Econometrics*, *142*, 265–280.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of the article.

How to cite this article: Harvey D. I., Leybourne S. J., & Zu Y. (2024). Tests for equal forecast accuracy under heteroskedasticity. *Journal of Applied Econometrics*, *39*(5), 850–869. https://doi.org/10.1002/jae.3050