Multi-Horizon Forecast Comparison

Rogier Quaedvlieg*
Erasmus School of Economics, Erasmus University Rotterdam

February 28, 2018

Abstract

We introduce tests for multi-horizon superior predictive ability. Rather than comparing forecasts of different models at multiple horizons individually, we propose to jointly consider all relevant horizons within a forecast path. We define the concepts of uniform and average superior predictive ability. The former entails superior performance at each individual horizon, while the latter allows inferior performance at some horizons to be compensated by others. We show that the tests lead to more coherent conclusions, and have greater power to differentiate models than the single-horizon tests. We provide an extension of the Model Confidence Set to allow for multi-horizon comparison of more than two models. Simulations demonstrate appropriate size and high power. An illustration of the tests on a large set of macroeconomic variables demonstrates the empirical benefits of multi-horizon comparison.

Keywords: Forecasting, Long-Horizon, Multiple Testing, Path Forecasts, Superior Predictive Ability
JEL: C22, C52, C53, C58

*Corresponding author: Department of Business Economics, Erasmus School of Economics, PO Box 1738, 3000 DR Rotterdam, The Netherlands.

1 Introduction

Forecasts at multiple horizons should rarely be judged in isolation. The full forecasts path plays an important role in many policy decisions. For instance, in the context of typical macro-economic variables such as unemployment and inflation, policymakers require forecasts at different horizons to make informed decisions; the user does not only care about the value many periods from now, but the full path the series takes between now and that time. The importance of the path is not restricted to economics, as evidenced by for instance the large literature on forecasting climate data. As such, when comparing two or more different models in terms of their ability to make forecast paths, it is useful to compare the complete path.

The standard approach is to compare various models at different horizons independently, potentially leading to incoherent conclusions. For example, we might find that a first model is significantly better at predicting two and five ahead, while the second model has significantly better predictions three periods ahead, while the difference in forecasting performance is insignificant at all other horizons. The fact that either model performed worse at a single horizon, should not necessarily disqualify the model, and neither should the fact that the difference between the two models is insignificant at some horizons. When we compare performance at multiple horizons, we implicitly face a multiple testing problem. In finite samples we are likely to find that a mis-specified model will outperform even the population model at one of the many horizons one could consider. Comparing all horizons jointly guards us against this problem.

We therefore propose a test for multi-horizon superior predictive ability. There are at least three reasons why one might be interested in such a test. First, it entails a more robust definition of superior predictive ability. Second, while the hypotheses differ, jointly considering multiple horizons, allows us to construct a powerful test to disentangle models. Finally, as stated before, it guards us against spurious results induced by the multiple testing issues arising from considering multiple horizons individually.

We introduce two bootstrap-based test statistics, which can be used to test for two alternative definitions of multi-horizon superior predictive ability (SPA). The first statistic considers *uniform* multi-horizon SPA, which is defined as a model with lower loss at each

individual horizon. The second statistic is used to test for *average* multi-horizon SPA, which allows poor performance at some horizons to be compensated by superior performance at other horizons. The first definition is obviously far more stringent, but by properly controlling the family-wise error rate, equality of the models' forecast performance may still be rejected, even if the resulting superior model's empirical performance is inferior at some horizons. Importantly, both uniform and average multi-horizon SPA, as well as their respective tests, are defined in such a way that they reduce to the standard Diebold and Mariano (1995) test when only considering a single horizon.

In addition to the pairwise tests, we propose a multi-horizon version of the Model Confidence Set (MCS) of Hansen, Lunde, and Nason (2011), in order to compare more than two models at once. The multi-horizon MCS contains the set of models that have the best joint performance across horizons with given probability. Other multiple-model comparison techniques, such as those of White (2000) and Hansen (2005) can also easily be adapted to the multi-horizon framework.

In practice, the tests proposed in this paper should be viewed as applicable to a spectrum of potential hypotheses. On the one extreme, a potential user may be interested in just a single horizon, in which case the proposed tests reduce to the standard Diebold and Mariano (1995) test. On the other extreme, the test can be used to show that a new model has uniform SPA across all horizons, which is strong evidence in favor of a new specification. However, in many cases, users may have different models for different ranges of environments, i.e. short-term, mid-term and long-run forecasts. In such a scenario the tests may equally be applied to subsets of horizons.

There is a large empirical literature that reports forecasts at multiple horizons. Typically, these forecasts are evaluated and compared based on tests applied to each horizon separately. Exceptions are the work of Patton and Timmermann (2012), who propose a test for multi-horizon forecast optimality, and Jordà and Marcellino (2010), who call it path forecast evaluation. Their tests regard internal consistency of a single model, rather than comparing the performance of multiple models across horizons.[1]

[1]The literature on vector forecasts, concerning multiple variables rather than multiple horizons, faces similar problems of forecast comparison in the presence of correlated forecast error (e.g. Komunjer and Owyang, 2012).

The tests proposed in this paper fall into the framework implicitly defined in Diebold and Mariano (1995), and explicitly set out in, amongst others, Giacomini and White (2006) and Hansen (2005). We test for finite-sample multi-horizon predictive ability; the accuracy of forecasts at estimated values of parameters. This is in contrast to the literature set out by West (1996) whose aim is to use the forecasts to learn something about population-level predictive ability; accuracy of forecasts at the population value of the parameters. Clark and McCracken (2013) provide an excellent overview of the literature. The asymptotic theory in this setting requires non-vanishing estimation error, and as such a limitation of our tests is that they do not accommodate forecasts derived from models with recursively estimated parameters. We do permit the common rolling-window forecasting scheme, and a situation where parameters are estimated once at the beginning of the forecasting period.

We analyze the finite sample properties of the tests in simulation studies. We consider the two pairwise tests and the multi-horizon model confidence set. We demonstrate that the tests have appropriate size and good power, even in moderately sized samples. In addition, the simulations are used to investigate the conditions under which the multi-horizon comparison will lead to a more powerful test. This will turn out to be determined by the relative increases in average loss differentials and the variance of the loss differential as a function of horizon.

Empirically we consider two different datasets to illustrate the contributions of this paper. To highlight the pairwise tests, we revisit Marcellino, Stock, and Watson (2006), who investigate the relative merits of iterated and direct long-horizon forecasts. We test for both uniform and average SPA using 2 to 24 month horizon forecasts on their dataset of 170 macroeconomic time-series. By jointly considering all horizons, we find stronger evidence of iterated forecasts outperforming direct forecasts. When looking at individual series, we find that many of the incoherent results across horizons can be attributed to the multiple testing issues and lack of power.[2]

We proceed as follows. Section 2 sets out our theoretical framework and introduces the tests. Section 3 provides simulation evidence of size and power of the tests. Section 4 provides the empirical illustrations of the various tests, and finally Section 5 concludes.

[2]The Supplemental Appendix contains an application of the multi-horizon model confidence set in the context of Realized Volatility forecasting.

2 Setup

In this section we discuss the general setup. We consider the problem of comparing forecasts for potentially multivariate time series $\boldsymbol{y}_t$ over the time-period $t = 1, ..., T$. We are interested in point forecasts $\hat{\boldsymbol{y}}_{i,t}^h$ at multiple horizons, $h = 1, ..., H$. The forecasts may come from econometric models, professional forecasters, or any other alternative. Whenever the forecasts are derived from models, the forecasts $\hat{\boldsymbol{y}}_{i,t}^h = \hat{\boldsymbol{y}}_{i,t}^h(\hat{\boldsymbol{\theta}}_{i,t}^h)$ are based on estimated parameters $\hat{\boldsymbol{\theta}}$. We have two or more competing sets of forecasts, which may be based on different information sets, they may be based on nested or non-nested models, or come from any other source. We will use the term 'model' loosely to refer to all potential sources of forecasts.

The main contribution of this paper is to not 'only' consider the one-step ahead, or only at the h-step ahead forecast in isolation, but to jointly compare the quality of the full path of 1 to H-step ahead forecasts. That is, for model $i = 1, ..., M$, we have forecasts $\hat{\boldsymbol{y}}_{i,t} = [\hat{\boldsymbol{y}}_{i,t}^1, ..., \hat{\boldsymbol{y}}_{i,t}^H]$, where $\hat{\boldsymbol{y}}_{i,t}^h$ is model i's forecast of $\boldsymbol{y}_t$ based on information up until time $\mathcal{F}_{t-h}$. We define a general loss function $\boldsymbol{L}_{i,t} = L(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_{i,t})$, which maps the forecast errors into a H-dimensional row vector, with elements $L_{i,t}^h = L(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_{i,t})$.

For any loss function, and any two sets of forecasts, we compare models in terms of their loss differential

$$\boldsymbol{d}_{ij,t} \equiv \boldsymbol{L}_{i,t} - \boldsymbol{L}_{j,t}, \tag{1}$$

which is an H-dimensional row vector, with elements $d_{ij,t}^h$. Note that $d_{ij,t}^h$ is implicitly defined as a function of estimated parameters, and our focus is on finite-sample predictive ability. Our hypotheses are defined in terms of the expected loss differentials, $\mu_{ij}^h \equiv E(d_{ij,t}^h)$, and as such we focus on the properties of $\boldsymbol{d}_{ij,t}$, which have some implications on how the forecasts may be generated.

In particular, we make the following assumption on $\boldsymbol{d}_{ij,t}$.

Assumption 1. *The vector of loss differences $\boldsymbol{d}_{ij,t}$ is (strictly) stationary and α-mixing of size $-(2+\delta)(r+\delta)/(r-2)$, for some $r > 2$ and $\delta > 0$, where $E|\boldsymbol{d}_t|^{r+\delta} < \infty$ and $Var(d_{ij,t}^h) > 0$ for all $h = 1, ..., H$.*

This assumption is needed to ensure that population moments of $\boldsymbol{d}_{ij,t}$ are well defined,

and to justify the bootstrap techniques introduced in Section 2.1.2. Under the stated assumption a central limit theorem applies (e.g. De Jong, 1997), such that

$$\sqrt{T}(\bar{\boldsymbol{d}}_{ij} - \boldsymbol{\mu}_{ij}) \rightarrow^{d} N_m(0, \boldsymbol{\Omega}_{ij}), \tag{2}$$

where $\boldsymbol{\Omega}_{ij} \equiv \text{avar}(\sqrt{T}(\bar{\boldsymbol{d}}_{ij} - \boldsymbol{\mu}_{ij}))$.

The assumption on $\boldsymbol{d}_{ij}$ is sufficient for validity of one of the most common tests for comparing two models', i and j, forecasting performance at a single horizon h, the Diebold and Mariano (1995) test. They test the null hypothesis that

$$H_{DM} : \mu_{ij}^{h} = 0, \tag{3}$$

using a standard t-test:

$$t_{DM}^{h} = \frac{\sqrt{T}\bar{d}_{ij}^{h}}{\hat{\omega}_{ij}^{h}}, \tag{4}$$

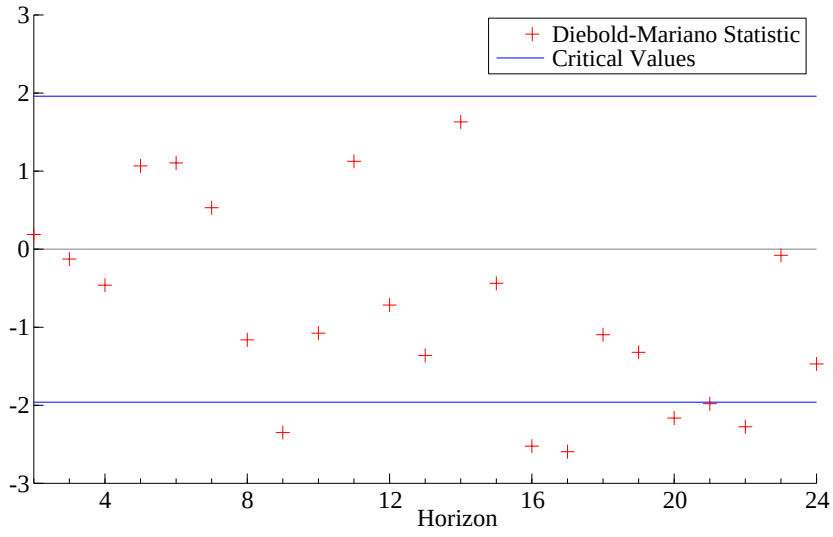
where $\bar{d}_{ij}^{h} = \frac{1}{T}\sum d_{ij,t}^{h}$, and $\omega_{ij}^{h} = \boldsymbol{\Omega}_{ij,hh}^{1/2}$, the square root of the diagonal element corresponding to the $h-$th horizon. In such a setting, the variance can be estimated using a HAC-type estimator, as in for instance Giacomini and White (2006), or, following Hansen et al. (2011), it may be obtained using bootstrap methods.

Importantly, a variety of common forecasting schemes do not satisfy Assumption 1, and therefore asymptotic normality. In particular, the framework permits parameters that are estimated on a rolling window, or just once (fixed scheme), but it prohibits the use of forecasts generated by recursive parameter estimates, as that violates the stationarity assumption. It can however handle both nested and non-nested models, as nonvanishing estimation error prevents singularity that may occur in nested models and parameters are at their probability limits. See Giacomini and White (2006) for a broad discussion of this framework.

2.1 Multi-Horizon Hypotheses of Interest

The Diebold and Mariano (1995) test can be used to compare model performance at each horizon individually. This can lead to a number of different conclusions. In an ideal situation this procedure finds significant evidence that a single model performs best on each horizon, or at the very least, not significantly worse than the other model. Another

Figure 1: Diebold-Mariano Tests at different Horizons for Earnings of Production Workers.



Note: This Figure presents the forecast comparison of the LEHM time-series of Marcellino et al. (2006). It plots the Diebold-Mariano test statistics as a function of forecast horizon ($h = 2, ..., 24$), for the loss differential of iterated minus direct forecasts. Lag lengths of the autoregressive models are selected based on BIC.

potential outcome that tells a consistent story, is that one model works well for short horizons, while the other model performs better at longer horizon. However, we may also come across situations in which the individual tests do not lead to coherent results. For instance, we may encounter a situation in which model i performs better than model j at most horizons, except for two or three non-consecutive horizons. This lack of coherency is most likely due to simple sampling error, which may cause even the population model to be beaten by a mis-specified model at some horizons.

To illustrate such a situation, consider Figure 1, which presents a preview of the empirical analysis in Section 4. We plot the Diebold-Mariano statistics over horizons 2 to 24 of the mean square forecast error comparison between direct and iterated autoregressive forecasts for a series of earnings of production workers. The statistic at the majority of horizons is negative indicating that direct forecasts outperform the iterated ones. However, all but six of the statistics are individually insignificant, and out of the insignificant ones, six have a positive statistic. Similar results can be found all throughout the forecasting literature.

The question arises whether this picture may provide joint evidence to conclude that either model significantly outperforms across all horizons. The negative point estimates may simply be due to sampling error, and the insignificance of the remaining horizons may potentially be attributed to lack of power. Alternatively, perhaps we can at least find statistical evidence for the claim that the average loss across horizons is either positive or negative.

We therefore propose the notion of multi-horizon superior predictive ability. The most natural, and strongest, notion is that a superior model should have better forecasts at each individual horizon. To that effect, define

$$\mu_{ij}^{(Unif)} = \min_h \mu_{ij}^h. \tag{5}$$

We refer to a situation with $\mu_{ij}^{(Unif)} > 0$ as *uniform* superior predictive ability (uSPA) of model j.

The definition of uSPA is strict, and we may often fail to find evidence for such relative forecasting performance. Therefore, we present a milder definition of superior predictive ability which we refer to as *average* superior predictive ability (aSPA). Here, we compare models based on their weighted average loss difference[3]

$$\mu_{ij}^{(Avg)} = \boldsymbol{w}'\boldsymbol{\mu}_{ij} = \sum_{h=1}^{H} w_h \mu_{ij}^h, \tag{6}$$

with weights $\boldsymbol{w} = [w_1, ..., w_H]$ summing to one. Obvious candidates for $\boldsymbol{w}$ are equal-weighted or weights decaying in the horizon. Note that we take the average loss, which is distinct from the loss of the average, which is just one aspect of the path.

The concepts of uniform and average SPA have clear links to the concepts of first- and second order forecast dominance respectively, and the tests in the next section also bear resemblance to tests for stochastic dominance (e.g. Linton, Maasoumi, and Whang, 2005; Linton, Song, and Whang, 2010). Similar to those concepts, uSPA implies aSPA, while the reverse is not necessarily true. We may be able to determine a ranking based on aSPA, even if uSPA fails to do so. However, average SPA requires the user to take a stand on

[3]Weighting may be of particular importance in the scenario where one makes aggregate $h-$period ahead forecasts, i.e. $\sum_{h=1}^{H} Y_{t+h}$, which results in clear scale differences that should be inversely weighted.

the relative importance of under-performance at one horizon against out-performance at another. More generally, the tests are closely related to work on multivariate inequality tests (e.g. Bartholomew, 1961; Wolak, 1987). In particular, Patton and Timmermann (2010) propose a solution similar to our uSPA test in the context of testing for monotonicity in asset pricing relationships.

A couple of remarks need to be made regarding testing multiple horizons jointly. First, increasing the number of horizons will not always result in a more powerful test. The variance of loss differences typically increases with horizon, and as such adding an additional horizon may actually decrease power.[4] Figure 1 shows however, that the single-horizon statistics are hardly affected by increasing variance, as the mean loss differential also tends to increase in horizon. The relative speed of accumulation across horizons will play an important role in the power of multi-horizon tests, which will be studied in the simulations.

Second, since forecast errors tend to be correlated across both horizon and time, the increase of information from considering, say, two horizons rather than one, does not provide a similar increase in information as doubling the out-of-sample period length. The tests introduced below should therefore mostly be interpreted as a guard against the implicit multiple testing issue, with the increase of power through H times as many loss observations being a secondary benefit.

2.1.1 Choice of Test Statistic

First, we consider a test on the minimum loss differential $\mu_{ij}^{(Unif)}$. If model j is better than model i, the minimum loss difference over all h should be greater than zero. Here we test the null hypothesis

$$H_{0,\text{uSPA}} : \mu_{ij}^{(Unif)} \leq 0, \tag{7}$$

against the alternative that $\mu_{ij}^{(Unif)} > 0$. We consider one-sided hypotheses, as models i and j can easily be switched. In order to test this hypothesis, we simply consider the minimum over all the individual Diebold-Mariano statistics DM_t^h:

[4] Moreover, forecasts beyond a certain limiting horizon may become uninformative, see Breitung and Knüppel (2017), which provides a natural stopping point.

$$t_{\text{uSPA},ij} = \min_h \frac{\sqrt{T}\bar{d}_{ij}^h}{\hat{\omega}_{ij}^h}. \tag{8}$$

Note that we take the minimum of the studentized test statistic, rather than studentizing the minimum. The main advantage of this is that we only require estimates of the diagonal of the covariance matrix of $\bar{d}_{ij,t}$ rather than the full matrix. This is of particular importance when H grows too large to obtain a sensible estimate of the covariance matrix. The downside is that as a result the statistic will be non-pivotal, as its distribution does depend on the full covariance matrix, which makes $\mathbf{\Omega}_{ij}$ a nuisance parameter. As discussed before, this nuisance parameter problem is handled by the bootstrap methods, which implicitly deal with these problems. This feature has previously been used by White (2000), Hansen (2005), Clark and McCracken (2005) and Hansen et al. (2011).[5] For a related discussion on the relative merits of nonquadratic statistics, see Hansen (2005) in the context of loss differences between a benchmark model and many alternative competing models.

Next, we consider a simple test for average SPA, based on the weighted-average loss differential. The associated null is

$$H_{0,\text{aSPA}} : \mu_{ij}^{(Avg)} \leq 0, \tag{9}$$

with alternative $\mu_{ij}^{(Avg)} > 0$. A simple studentized statistic takes the form

$$t_{\text{aSPA},ij} = \frac{\sqrt{T}\boldsymbol{w}'\bar{\boldsymbol{d}}_{ij}}{\hat{\zeta}_{ij}}. \tag{10}$$

Similar to the uSPA statistic, we avoid estimating the full covariance matrix $\mathbf{\Omega}_{ij}$, and choose to estimate $\zeta_{ij} \equiv \sqrt{\boldsymbol{w}'\mathbf{\Omega}_{ij}\boldsymbol{w}}$, directly based on $\boldsymbol{w}'\boldsymbol{d}_{ij,t}$.[6]

[5] There are some situations in which the joint distribution of competing models may be known. For instance, between direct and iterated forecasts of the same AR(p) model (Ing, 2003), but general results cannot be obtained. As such the bootstrap provides a reasonable method to obtain the distribution of the statistics.

[6] An unweighted version of the aSPA test statistic was also considered in Capistrán (2006). That test on asymptotic critical values, while our bootstrap critical values work distinctly better in small samples. Subsequent research by Martinez (2017) provides a generalization of the unweighted aSPA test in a GFESM context (Clements and Hendry, 1993), explicitly allowing for differences in covariance dynamics of the various models. We target the loss-differential directly as a primitive.

Throughout the paper we will simply use an equal weighted average with $w_h = 1/H$, for all h. Different weights would correspond to different utility functions of the forecaster. Alternatively, one could use 'efficient' weights to minimize ζ_{ij} by setting the weights for each horizon inversely proportional to their variance $(\omega_{ij}^h)^2$, or more generally the inverse of an estimate of the full covariance matrix of $\boldsymbol{d}_{ij,t}$, $\boldsymbol{\Omega}_{ij}$.

Note that the aSPA test is simply a Diebold-Mariano test on the weighted average loss-series, $\boldsymbol{w}'\boldsymbol{d}_{ij,t}$. Moreover, the test for uSPA is in fact a special case of aSPA, with $w_h = 1$ for h equal to the 'minimum' horizon, and zero otherwise. Typically, the weighted averages will converge to a standard normal distribution, such that standard critical values may be used. Special choices of weights, such as those amounting to quantiles of the distribution will require non-standard critical values. Obtaining the critical value using bootstrap techniques may lead to better finite sample properties properties in the equal-weighted case as well, and as a result we suggest obtaining bootstrapped critical values regardless of the choice of weights.

2.1.2 Bootstrap Implementation

The minimum over multiple t-statistics will not follow a student distribution, and is dependent on the number of statistics H. Rather than the standard 95% critical value of 1.645, the appropriate critical value will be lower and may actually be negative for large H. As a result, depending on the degree of sampling variation, observing a negative statistic at any of the horizons may not be sufficient evidence to stop us from rejecting the null in favor of uSPA, and shows the need for appropriate multiple testing techniques.

We obtain the distribution of the statistics under the null using bootstrap techniques. The chosen method needs to take into account the dependence across horizons and the likely serial correlation in forecast errors. Throughout the paper we will use the stationary block bootstrap of Politis and Romano (1994) and Gonçalves and de Jong (2003). We set the parameter of the stationary bootstrap to $q = 0.05$ corresponding to an average block length of 20 observations, and all results are based on $B = 999$ re-samples.

First, the t_{uSPA} and t_{aSPA} statistics require estimates of the variance of individual and weighted-average loss differentials respectively, $\hat{\omega}_{ij}^h$ and $\hat{\zeta}_{ij}$. The stationary bootstrap vari-

ance estimator of an average is known in closed form, so our recommendation is to use the bootstrap population value directly, given by

$$(\hat{\omega}_{ij})^2 \equiv \hat{\gamma}_{0,ij} + 2\sum_{k=1}^{T-1} \kappa(T,k)\hat{\gamma}_{k,ij}, \tag{11}$$

where γ_k is the usual autocovariance function applied to d_{ij}^h or $\boldsymbol{w}'\boldsymbol{d}_{ij,t}$, and kernel weights (under the stationary bootstrap) are given by $\kappa(T,k) \equiv \frac{T-k}{T}(1-q)^k + \frac{k}{T}(1-q)^{T-k}$ (Politis and Romano, 1994). Using the formula prevents the need for a double bootstrap.

Next, we use a bootstrap to obtain the critical values of the test for uSPA and aSPA under the null:

Algorithm 1 (Multi-Horizon SPA Bootstrap)**.**
For bootstrap sample $b = 1, ..., B$:

1. *Re-sample centered $\boldsymbol{d}_{ij,t} - \bar{\boldsymbol{d}}_{ij}$ with replacement, using a stationary bootstrap with parameter q, to obtain $\boldsymbol{d}_{ij,t}^b$, with elements $d_{ij,t}^{hb}$.*

2. *uSPA: Compute $\bar{d}_{ij}^{hb} = \frac{1}{T}\sum_{t=1}^{T} d_{ij,t}^{hb}$ for each h.*
   Compute $\hat{\omega}_{ij}^{hb}$ for each h using (11).
   Compute the uSPA statistic: $t_{uSPA,ij}^b = \min_h[\sqrt{T}\bar{d}_{ij}^{hb}/\hat{\omega}_{ij}^{hb}]$
   aSPA: Compute $\bar{d}_{ij}^b = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{w}'\boldsymbol{d}_{ij,t}^b$.
   Compute $\hat{\zeta}_{ij}^b$ using (11).
   Compute the aSPA statistic: $t_{aSPA,ij}^b = \sqrt{T}\bar{d}_{ij}^b/\hat{\zeta}_{ij}^b$.

Finally, obtain an appropriate critical value $c_{\bullet SPA,ij}^{\alpha}$ as the α-quantile of the bootstrap distribution of either of the two $t_{\bullet SPA,ij}^b$. Rejection occurs if $t_{\bullet SPA,ij} > c_{\bullet SPA,ij}^{\alpha}$. Alternatively, a p-value may be computed as $\sum_{b=1}^{B} \mathbf{1}_{\{t_{\bullet SPA,ij}^b > t_{\bullet SPA,ij}\}}/B$.

The following Theorem provides the foundation for the validity of the bootstrap algorithm for both the test for uSPA and aSPA.

Theorem 1 (Bootstrap Validity Studentized Statistics)**.** *Let $\boldsymbol{D} \equiv diag(\omega^1, ..., \omega^H)$ and $\hat{\boldsymbol{D}}$, $\boldsymbol{D}^b$ analogously defined using $\hat{\omega}^h$ and $\hat{\omega}^{hb}$. Let Assumption 1 hold, and moreover, assume that $q_T = q$ satisfies $q_T \to 0$ and $Tq_T^2 \to \infty$ as $T \to \infty$, then*

$$\sup_{x \in \mathbb{R}^H} \left| P^b\left[\sqrt{T}(\boldsymbol{D}^b)^{-1}(\bar{\boldsymbol{d}}^b - \bar{\boldsymbol{d}}) \leq x\right] - P\left[\sqrt{T}\hat{\boldsymbol{D}}^{-1}(\bar{\boldsymbol{d}} - \boldsymbol{\mu}) \leq x\right] \right| \to_p 0, \tag{12}$$

P^b denotes the bootstrap probability measure.

The proof is provided in Appendix A. From Theorem 1 we obtain the following Corollary.

Corollary 1. *Let the Assumptions from Theorem 1 hold. Then,*

$$\sup_z \left| P^b \left[\min_h \sqrt{T} \frac{\bar{d}_{ij}^{hb} - \bar{d}_{ij}^h}{\hat{\omega}_{ij}^{hb}} \leq z \right] - P \left[\min_h \sqrt{T} \frac{\bar{d}_{ij}^h - \mu_{ij}}{\hat{\omega}_{ij}^h} \leq z \right] \right| \to_p 0, \tag{13}$$

and

$$\sup_z \left| P^b \left[\sqrt{T} \frac{\boldsymbol{w}' \bar{\boldsymbol{d}}_{ij}^b - \boldsymbol{w}' \bar{\boldsymbol{d}}_{ij}}{\hat{\zeta}_{ij}^b} \leq z \right] - P \left[\sqrt{T} \frac{\boldsymbol{w}' \bar{\boldsymbol{d}}_{ij} - \boldsymbol{w}' \boldsymbol{\mu}_{ij}}{\hat{\zeta}_{ij}} \leq z \right] \right| \to_p 0. \tag{14}$$

The Corollary demonstrates that the bootstrap may be used to obtain the critical values for both the uSPA, (13), and aSPA, (14), test statistics. If follows directly from Theorem 1 and the continuous mapping theorem combined with the fact that the average and minimum are smooth functions of the elements of the vector $\boldsymbol{d}_{ij,t}$. Weighted averages are obviously smooth functions, and, as shown in Proposition 2.2 of White (2000), the minimum of a vector of differences is a continuous function of the elements of the vector.

2.2 The Multi-Horizon Model Confidence Set

The two tests introduced in the previous section can only be used for a pairwise comparison of models. In this section we extend this to a general M-dimensional set of models $\mathcal{M}$, by adapting the Model Confidence Set (MCS) approach of Hansen et al. (2011) to allow for joint multi-horizon testing. They propose an algorithm that selects a subset of $\mathcal{M}$ that contains the set of best models with a given probability, which we denote $\tilde{\alpha}$. The standard MCS can broadly be interpreted as a sequential Diebold-Mariano test, and as such it readily extends to the case with either the $t_{\text{uSPA},ij}$ or $t_{\text{aSPA},ij}$ statistics.

For the uSPA multi-horizon MCS, analogous to Hansen et al. (2011), we define the MCS as the set of models for which we find no statistical support to differentiate the models within the set:

$$\mathcal{M}^*_{\text{uSPA}} \equiv \{i \in \mathcal{M}^0 : \min_h \mu_{ij}^h \leq 0, \forall j \in \mathcal{M}^0\} \tag{15}$$

$$\mathcal{M}^*_{\text{aSPA}} \equiv \{i \in \mathcal{M}^0 : \boldsymbol{w}' \bar{\boldsymbol{d}}_{ij} \leq 0, \forall j \in \mathcal{M}^0\} \tag{16}$$

The associated null hypothesis are

$$H_{\mathcal{M},\text{uSPA}} : \min_h \mu_{ij}^h \leq 0, \text{ for all } i,j \in \mathcal{M} \tag{17}$$

$$H_{\mathcal{M},\text{aSPA}} : \boldsymbol{w}'\bar{\boldsymbol{d}}_{ij} \leq 0, \text{ for all } i,j \in \mathcal{M} \tag{18}$$

with $\mathcal{M} \subseteq \mathcal{M}^0$.

The multi-horizon model confidence set is obtained sequentially as

1. Set $\mathcal{M} = \mathcal{M}^0$.
2. Test $H_{\mathcal{M},\bullet\text{SPA}}$ using an equivalence test at level $\tilde{\alpha}$.
3. If $H_{\mathcal{M},\bullet\text{SPA}}$ is not rejected, define $\widehat{\mathcal{M}}_{\bullet\text{SPA},1-\tilde{\alpha}} = \mathcal{M}$.
   If the null is rejected, use the elimination rule to remove a model from $\mathcal{M}$, and go back to Step 2.

The equivalence test has to be adapted to the multi-horizon setting. Hansen et al. (2011) propose the maximum of all pairwise t_{DM} statistics to test for equivalence, but since the critical value of the $t_{\bullet\text{SPA}}$ statistics are not necessarily the same for all pairs $\{i,j\}$, we cannot simply consider the maximum of the $t_{\bullet\text{SPA},ij}$. Due to the fact that the critical values can be both positive and negative, we instead consider the maximum of the centered statistics $\max_{i,j\in\mathcal{M}}[t_{\bullet\text{SPA}} - c^{\alpha}_{\bullet\text{SPA}}]$. Unfortunately, we require the use of a double bootstrap to obtain its distribution. The computational cost is high, but feasible as we simply bootstrap studentized means, and require no re-estimation of models.

Algorithm 2 (Multi-Horizon MCS Bootstrap)**.**

1. *For each pair $\{i,j\} \in \mathcal{M}$, compute the statistic $t_{\bullet SPA,ij}$ and use Algorithm 1 to obtain an estimate of the associated critical value $c^{\alpha}_{\bullet SPA,ij}$.*
2. *Define $t_{Max,uSPA} \equiv \max_{i,j\in\mathcal{M}} \left[t_{\bullet SPA,ij} - c^{\alpha}_{\bullet SPA,ij}\right]$, i.e. the test statistic furthest from its critical value.*
3. *For each of the bootstrap samples $\boldsymbol{d}^b_{ij,t}$, $b = 1,...,B$, in Step 1:*
   a. *For each pair $\{i,j\} \in \mathcal{M}$, apply Algorithm 1 to the bootstrap sample $\boldsymbol{d}^b_{ij,t}$ directly, to obtain $c^{\alpha b}_{\bullet SPA,ij}$.*

b. Compute the bootstrapped $t^b_{Max,uSPA} \equiv \max_{i,j\in\mathcal{M}} \left[t^b_{uSPA,ij} - c^{\alpha b}_{ij}\right]$

4. *Obtain the appropriate critical value as the* $\tilde{\alpha}$*-quantile of the bootstrap distribution* $t^b_{Max,uSPA}$*, or define the p-value as* $p \equiv \frac{1}{B}\sum_{b=1}^{B} I_{\{t_{Max,uSPA} < t^b_{Max,uSPA}\}}$.

The combination of equivalence test and elimination rule adhere to the definition of coherency of Hansen et al. (2011). To obtain reasonable p-values we follow Hansen et al. (2011) in imposing that a p-value for a model can not be lower than any previously eliminated model, and follow the convention that the last remaining model obtains a p-value of one. Also note that the level of the critical values of the pairwise tests, α, and the one for the MCS $\tilde{\alpha}$, may differ. In large samples, the choice of α is of little importance as all $t_{\bullet\text{SPA},ij}$ are approximately normally distributed with unit variance. However, in small samples, the choice of α may impact the ordering of the different models.

3 Simulations

In this section we report the results of Monte Carlo experiments to demonstrate appropriate size and good power of the single tests, as well as desirable properties of the Multi-Horizon Model Confidence Set.[7]

3.1 Data Generating Process

First we describe how we generate 'losses' of a given model i. Our design closely resembles that of the simulation section of Hansen et al. (2011), where losses are simulated directly, rather than obtained indirectly through the forecasting performance of various models on generated data. This allows us to easily increase the number of models, to control their relative performance directly, and to impose the notions of uniform and average SPA. However, in contrast to Hansen et al. (2011) who simulate one-step-ahead losses, we need to simulate forecast-path losses, which requires a certain dependence structure. We calibrate

[7]All results reported in this paper are based on programs written in Ox version 7.0 (Doornik, 2012). Example code detailing the implementation of the various tests and simulations is available on Quaedvlieg's website.

our DGP to obtain similar properties to the loss differential between an AR(1) and AR(2) when the true model is the latter.[8]

We consider simulation set-ups with two and ten models. For the ten-model setup, the average loss of each model is proportional to the $H-$dimensional vector $\boldsymbol{\theta}$, which governs the loss differentials. We will consider two different definitions pertaining to the uSPA and aSPA below. Each model i has average loss equal to $\boldsymbol{\theta}_i = \frac{(i-1)}{9}\boldsymbol{\theta}$, with $i = 1, ..., 10$, and therefore $\boldsymbol{\mu}_{ij} = \boldsymbol{\theta}_i - \boldsymbol{\theta}_j$. For the two-model setting we will only consider $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, such that the population difference between the models equals $\boldsymbol{\mu}_{12} = \boldsymbol{\theta}/9$.

The elements of $\boldsymbol{\theta} = [\theta^1, ..., \theta^h]$, determine how loss varies across horizons. A misspecified model is expected to lead to greater divergence at longer horizons, and as such we assume loss is increasing in horizon. We consider two different definitions in order to highlight the tests for uSPA and aSPA. First, we set

$$\theta^{h(Unif)} = (1 + \phi\sqrt{h-1})\lambda/\sqrt{T}. \tag{19}$$

The loss differential is non-negative at all horizons, implying that the superior model has both uniform and average superior predictive ability. λ governs the size of the loss-differential, while ϕ governs how fast the average loss increases as a function of horizon. When $\phi = 0$ the loss is equal at all horizons, while for $\phi > 0$ loss is increasing in horizon.

Next, we set

$$\theta^{h(NonUnif)} = \begin{cases} -\lambda/\sqrt{T} & \text{if } h = 1 \\ c(1 + \phi\sqrt{h-1})\lambda/\sqrt{T} & \text{if } h > 1, \end{cases}, \tag{20}$$

with $c = 1+2/\sum_{h=2}^{H}(1+\phi\sqrt{h-1})$, such that $\sum_{h=1}^{H}\theta^{h(NonUnif)} = \sum_{h=1}^{H}\theta^{h(Unif)}$. We impose non-uniformity through the first horizon, to ensure that the single negative differential is included in all multi-horizon tests. Note that under this definition, the first model does have aSPA for $H > 1$, but no uniform SPA for any horizon.

We generate the losses as follows:

$$\begin{aligned} \boldsymbol{L}_{i,t} &\equiv \boldsymbol{\theta}_i + \boldsymbol{Y}_{i,t} \\ \boldsymbol{Y}_{i,t} &= \boldsymbol{\varrho} \circ \boldsymbol{Y}_{i,t-1} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}_t, \end{aligned} \tag{21}$$

[8] A visual illustration which highlights the properties of the DGP is available in the supplemental appendix.

where $\boldsymbol{\epsilon}_t \sim i.i.d.\ \mathcal{N}_H(0, \boldsymbol{I})$ and $\circ$ denotes the Hadamard product. The losses are serially correlated through ϱ and correlated across horizons through $\boldsymbol{\Sigma}$. While for $h = 1$, a case can be made that forecast errors will be uncorrelated over time if the model is well-specified, long horizon forecasts are likely to be strongly autocorrelated, even for a perfectly specified model. We set the first order autocorrelation to $\varrho_h = 0.2\sqrt{h-1}$, which ranges between 0 for $h = 1$ and 0.87 for $h = 20$.

The forecast errors at different horizons are not independent. First, we define the covariance structure across horizons, at a single point in time. Since most models will converge to the unconditional mean when h becomes large, the correlations should be close to one for adjacent horizons when h is large, and smaller for short-horizons. We define the correlation matrix $\boldsymbol{R}$, with elements $\rho_{g,h}$:

$$\rho_{g,h} = \begin{cases} 1 & \text{if } g = h \\ \exp(-0.4 + 0.025 \max(g,h) - 0.125|g-h|) & \text{if } g \neq h. \end{cases} \tag{22}$$

Our simulations will use $H = 20$, so the corner points of the correlation matrix are $\rho_{1,2} = 0.60$, $\rho_{1,20} = 0.10$ and $\rho_{19,20} = 0.95$. Next, the variance should be increasing in horizon. For simplicity we set it to $\sigma_h = 1 + \psi\sqrt{h-1}$. The variance plays a crucial role in the multi-horizon tests. If the variance is increasing too quickly, adding additional horizons may actually decrease the power of the test, rather than increasing it. We combine the variance and correlation to $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\sigma})\boldsymbol{R}\,\text{diag}(\boldsymbol{\sigma})$.

Note that in our simulation set-up $Cov(L^h_{i,t}, L^g_{j,t}) = 0$, for all models i and j and all horizons g and h. A positive correlation, holding individual variances fixed, would decrease the variance of the loss-difference and make it easier to differentiate models. A negative correlation would conversely increase the variance of the difference, but is unlikely to occur in this particular setting. The results below can thus be interpreted as a lower bound.

3.2 Pairwise Tests

In this section we investigate the properties of tests for the comparison of two models. The main goals of this section are to analyze the power and size of the newly introduced tests based on t_{uSPA} and t_{aSPA}. We report results over $S = 1000$ simulations, and vary

the parameters of the DGP. We take three sample sizes $T = 250, 500, 1000$. In order to investigate the trade-off of adding additional horizons, we analyze the effect of the parameters that govern how average loss (ϕ) and its variance (ψ) depend on horizon h. We set $\phi = 0, 1, 2$ and $\psi = 0, 0.125, 0.25$. The parameter that governs the magnitude of the loss differential is set to $\lambda = 0, 5, 10, 20, 40$. Throughout, we consider one-sided tests $\mu_{12} < 0$ at the 5% level, i.e. we test whether model one outperforms model two at multiple individual horizons, in uSPA, or in aSPA. We report results for different horizons $H = 1, 5, 10$ and 20. The DM test uses that specific horizon only, while the uniform and average SPA tests use all horizons up to and including H.

We start by establishing appropriate size and good power of the three tests in Table 1. We vary T and λ, and keep $\phi = 1$ and $\psi = 0.125$ fixed at their middle levels. We consider both loss differentials $\theta^{(Unif)}$ and $\theta^{(NonUnif)}$, referred to as Uniform and Non-Uniform alternative, displayed in the top and bottom panel respectively.

First consider the top panel, which is based on $\theta^{(Unif)}$. When $\lambda = 0$, we are under the null, as the average loss of the two models is identical. We see that all three tests have size close to the nominal 5%. The small size distortions appear to be most severe for the basic Diebold-Mariano test, are typically increasing in H, and decreasing in T.

When $\lambda > 0$, the loss differential at each horizon is positive. First consider the standard Diebold-Mariano test. We see that power is increasing in λ, while the influence of the sample size T is minimal. It is evident that the horizon also plays a significant role in the power of the test. Given our choice of ϕ the loss differential is increasing in h, which leads to higher power. On the other hand, the variance of the loss differential is also increasing in h, decreasing the ability to differentiate models. In this case this results in the highest power at $h = 5$ for the single-horizon test, with slightly lower power for longer horizons.

Under the alternative in the top panel, Model 1 has both uniform and average superior predictive ability, and as such both tests should reject. For $h = 1$, all three tests are identical, and the slight differences in rejection frequencies are simulation noise. For $h = 5$ and upwards, all tests are different. The tests for uSPA and aSPA use the loss-differential at all horizons, which results in typically slightly higher rejection frequencies at the same DGP. While the DM test has most power at $H = 5$, the tests for uSPA and aSPA become

Table 1: Univariate Simulation Results: Size and Power

		Diebold-Mariano Test				Test for uSPA				Test for aSPA			
	H	1	5	10	20	1	5	10	20	1	5	10	20
T	λ	Uniform Alternative											
250	0	0.083	0.067	0.082	0.085	0.076	0.066	0.058	0.051	0.076	0.065	0.060	0.072
250	5	0.154	0.239	0.217	0.193	0.137	0.203	0.234	0.238	0.135	0.241	0.273	0.278
250	10	0.235	0.537	0.488	0.416	0.218	0.434	0.522	0.558	0.211	0.533	0.606	0.631
250	20	0.548	0.944	0.905	0.843	0.500	0.825	0.899	0.936	0.490	0.952	0.973	0.972
250	40	0.946	1.000	1.000	0.999	0.901	0.996	0.999	0.999	0.902	1.000	1.000	1.000
500	0	0.054	0.055	0.053	0.055	0.054	0.055	0.060	0.044	0.051	0.052	0.055	0.056
500	5	0.111	0.220	0.189	0.173	0.110	0.192	0.224	0.232	0.107	0.217	0.254	0.255
500	10	0.222	0.498	0.464	0.398	0.199	0.429	0.501	0.541	0.206	0.520	0.598	0.608
500	20	0.498	0.940	0.911	0.833	0.473	0.810	0.893	0.929	0.468	0.955	0.982	0.987
500	40	0.936	1.000	1.000	1.000	0.919	0.993	0.996	0.999	0.925	1.000	1.000	1.000
1000	0	0.063	0.072	0.055	0.052	0.056	0.058	0.063	0.072	0.060	0.068	0.061	0.063
1000	5	0.115	0.187	0.203	0.155	0.115	0.186	0.206	0.201	0.109	0.206	0.247	0.215
1000	10	0.217	0.481	0.478	0.366	0.213	0.441	0.513	0.566	0.216	0.548	0.621	0.631
1000	20	0.463	0.925	0.920	0.821	0.454	0.804	0.879	0.926	0.449	0.955	0.981	0.983
1000	40	0.936	1.000	1.000	0.998	0.925	0.994	0.997	0.998	0.930	1.000	1.000	1.000
		Non-Uniform Alternative											
250	0	0.062	0.069	0.079	0.087	0.053	0.069	0.066	0.052	0.053	0.067	0.072	0.077
250	5	0.034	0.256	0.247	0.206	0.030	0.114	0.154	0.189	0.029	0.210	0.263	0.291
250	10	0.011	0.548	0.500	0.410	0.013	0.070	0.125	0.172	0.014	0.406	0.571	0.599
250	20	0.000	0.942	0.939	0.851	0.001	0.014	0.032	0.067	0.000	0.878	0.973	0.990
250	40	0.000	1.000	1.000	1.000	0.000	0.000	0.000	0.002	0.000	1.000	1.000	1.000
500	0	0.072	0.054	0.055	0.059	0.066	0.050	0.047	0.058	0.073	0.060	0.045	0.057
500	5	0.025	0.237	0.221	0.184	0.019	0.108	0.146	0.164	0.020	0.200	0.253	0.264
500	10	0.011	0.526	0.491	0.397	0.010	0.077	0.148	0.204	0.010	0.437	0.597	0.639
500	20	0.000	0.958	0.936	0.853	0.000	0.011	0.038	0.073	0.000	0.899	0.979	0.981
500	40	0.000	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000
1000	0	0.051	0.055	0.054	0.047	0.049	0.048	0.051	0.057	0.048	0.051	0.048	0.056
1000	5	0.026	0.232	0.217	0.147	0.027	0.107	0.133	0.154	0.027	0.189	0.252	0.264
1000	10	0.011	0.516	0.459	0.341	0.010	0.074	0.118	0.197	0.012	0.450	0.593	0.596
1000	20	0.001	0.949	0.934	0.830	0.001	0.022	0.042	0.072	0.001	0.908	0.983	0.990
1000	40	0.000	1.000	1.000	0.999	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000

Note: This table provides rejection frequencies over $S = 1000$ simulations according to the DGP outlined in Section 3.1. The parameters ϕ and ψ are fixed at 1 and 0.125 respectively, while the other parameters vary as indicated. In the panel denoted Uniform Alternative, the losses are generated according to $\theta^{(Unif)}$, while the Non-Uniform Alternative panel results are generated using $\theta^{(NonUnif)}$.

increasingly more powerful when the number of horizons increases.

Now consider the bottom panel, which is based on $\theta^{(NonUnif)}$. Under this alternative, Model 2 has lower loss than Model 1 at $h = 1$, but higher loss for all horizons $h \geq 2$. As a result, Model 1 has average SPA for horizons $h > 1$, but never uniform SPA.

First again consider the Diebold-Mariano test. For $h = 1$, the number of rejections when $\lambda = 0$ shows appropriate size, but when $\lambda > 0$, the number of rejections of our one-sided test appropriately converge to 0, as the second model is actually superior to the first. Recall that $\theta^{(NonUnif)}$ is chosen such that over the 20 horizons, the average $\theta^{(NonUnif)}$ is equal to $\theta^{(Unif)}$. As a result, relative to the top panel, for $h > 1$ we see that the univariate tests have higher power in the bottom panel, as the loss differential is slightly larger to compensate the negative differential at $H = 1$. We observe similar results for the aSPA test, which converges to zero rejections at $H = 1$ when $\lambda > 0$. For $H = 5$ and $H = 10$ is has slightly lower power than under the uniform alternative, as indeed the average loss differential is only equal at $H = 20$. At $H = 20$ the power is the same as under the uniform alternative, up to simulation noise.

The test for uSPA however shows very different results, as under this alternative, no model has uSPA. This is clearly reflected in the rejection frequencies, as the results show that the test indeed does not reject the null in most cases. For small λ the single negative loss differential is sometimes deemed within the range of random variation, and we see rejections of up to 20% when $\lambda = 10$. However, when λ increases the test rightfully fails to reject in almost all iterations.

In Table 1 we analyzed the properties of the tests keeping ϕ and ψ fixed. Next, Table 2 reports on the performance of the test for average SPA, under the uniform alternative, whilst varying ϕ and ψ, keeping $T = 500$ fixed. The aim of this simulation is to demonstrate that the test may not always become more powerful as the number of horizons increases. In particular, their properties depend on the degree to which the average loss differential and its variance evolve as a function of horizon.

The middle quadrant is equivalent to the set-up in Table 1, and for this table we mainly discuss the four extreme quadrants. When $\phi = \psi = 0$, the average and variance of the loss differentials are constant across horizons. Here we see that without exception,

Table 2: Univariate Simulation Results: Varying loss properties at different horizons

		$\psi = 0$				$\psi = 0.125$				$\psi = 0.25$			
	H	1	5	10	20	1	5	10	20	1	5	10	20
λ	ϕ	Uniform Alternative											
250	0	0.050	0.054	0.054	0.053	0.066	0.052	0.068	0.053	0.053	0.050	0.046	0.047
250	5	0.125	0.133	0.137	0.165	0.117	0.121	0.096	0.089	0.126	0.109	0.094	0.095
250	10	0.216	0.256	0.274	0.295	0.220	0.206	0.168	0.156	0.202	0.157	0.138	0.120
250	20	0.487	0.592	0.628	0.659	0.475	0.425	0.370	0.297	0.467	0.349	0.259	0.200
250	40	0.904	0.968	0.980	0.991	0.929	0.865	0.774	0.608	0.941	0.737	0.579	0.418
500	0	0.051	0.061	0.054	0.055	0.054	0.055	0.060	0.044	0.051	0.060	0.061	0.053
500	5	0.104	0.233	0.317	0.457	0.110	0.192	0.224	0.232	0.117	0.166	0.156	0.162
500	10	0.210	0.491	0.635	0.777	0.199	0.429	0.501	0.541	0.221	0.373	0.389	0.390
500	20	0.481	0.819	0.888	0.941	0.473	0.810	0.893	0.929	0.465	0.760	0.789	0.771
500	40	0.934	0.998	1.000	1.000	0.919	0.993	0.996	0.999	0.917	0.992	0.998	0.997
1000	0	0.058	0.056	0.080	0.067	0.060	0.053	0.055	0.061	0.061	0.053	0.052	0.052
1000	5	0.105	0.337	0.484	0.636	0.093	0.265	0.349	0.415	0.109	0.238	0.267	0.283
1000	10	0.190	0.535	0.671	0.795	0.187	0.552	0.683	0.789	0.182	0.487	0.591	0.621
1000	20	0.476	0.814	0.889	0.946	0.477	0.810	0.897	0.951	0.482	0.818	0.894	0.930
1000	40	0.918	0.995	0.998	0.999	0.926	0.997	0.999	1.000	0.931	0.994	0.997	0.999

Note: This table provides rejection frequencies for the test for uniform superior predictive ability over $S = 1000$ simulations according to the DGP outlined in Section 3.1. The losses are generated according to $\theta^{(Unif)}$, and the sample size $T = 500$ for all results.

power is increasing in h, as we simply add more information on the model's performance. When $\phi = 0$ but $\psi = 0.25$, the average loss differential remains fixed, but its variance is increasing. As a result, adding more horizons decreases power drastically, such that the number of rejections at $h = 20$ is less then half those at $h = 1$. When $\phi = 2$ and $\psi = 0$, the mean loss differential is increasing, while the variance is fixed, and power is large. Even with $\lambda = 5$, the test using all 20 horizons rejects in over 60% of samples. Finally, when $\phi = 2$ and $\psi = 0.25$, for $h > 1$, the power of the test is only marginally increasing across horizons. As such it presents a setting in which adding more or fewer horizons mainly adds in terms of interpretation and robustness of conclusions, but not in terms of increasing power.

Table 3: Multivariate Simulation Results: Potency and Gauge

		Potency				Gauge			
	H	1	5	10	20	1	5	10	20
T	λ								
250	0	0.788	0.692	0.631	0.552				
250	5	0.963	0.963	0.963	0.962	4.198	1.516	1.284	1.202
250	10	0.977	0.990	0.990	0.994	1.671	0.413	0.323	0.292
250	20	0.994	1.000	1.000	1.000	0.452	0.026	0.010	0.005
250	40	1.000	1.000	1.000	1.000	0.031	0.000	0.000	0.000
500	0	0.870	0.827	0.797	0.781				
500	5	0.978	0.980	0.981	0.985	3.997	1.401	1.183	1.080
500	10	0.983	0.991	0.997	0.999	1.501	0.377	0.323	0.262
500	20	0.997	0.999	0.999	1.000	0.449	0.012	0.004	0.003
500	40	1.000	1.000	1.000	1.000	0.038	0.000	0.000	0.000
1000	0	0.903	0.883	0.871	0.858				
1000	5	0.975	0.987	0.989	0.988	3.562	1.269	1.017	0.866
1000	10	0.987	0.996	0.996	0.997	1.458	0.317	0.236	0.185
1000	20	0.996	0.999	1.000	1.000	0.382	0.016	0.007	0.004
1000	40	1.000	1.000	1.000	1.000	0.021	0.000	0.000	0.000

Note: This table provides the potency and gauge of the multi-horizon MCS over $S = 1000$ simulations according to the DGP outlined in Section 3.1. The potency is defined as the fraction of correct superior models in the MCS. The gauge is defined as the number of models incorrectly included in the MCS. The parameters ϕ and ψ are fixed at 1 and 0.125 respectively, while the other parameters vary as indicated. The losses are generated based on the uniform alternative $\theta^{(Unif)}$.

3.3 Model Confidence Sets

In this section we evaluate the ability of the Multi-Horizon Model Confidence Set to distinguish models. We base our conclusions on the ten-model scenario. We use $\theta^{h(Unif)}$ to generate the loss differentials. Recall this means that the average loss of model i equals $\theta_i = \frac{(i-1)}{9}\theta$. As such there is a single superior model, and the loss differential between the first and the ith model increases linearly for the remaining nine models.

As in Table 1, we investigate the effect of T and λ, and use the middle scenarios, $\phi = 1$ and $\psi = 0.125$ throughout the analysis. The effects of changing ϕ and ψ on the ability of the Multi-Horizon MCS to differentiate models is similar to the gain and loss of power in the pairwise setting.

We summarize the Multi-Horizon MCS performance by two simple measures, potency and gauge. These concepts were used by Hendry and Doornik (2014) in the setting of model selection. The notions are similar, but distinct, from the usual size and power. Potency is defined as the fraction of appropriately selected models in the MCS. For $\lambda = 0$, all models are equal, and therefore defined as average fraction of models in the MCS. For $\lambda > 0$, Model 1 is the single best model, and hence the reported number is the fraction of times this model is in the MCS. The MCS is defined in such a way that the potency should at least equal one minus the level of the MCS, which we set at $\tilde{\alpha} = 0.20$. Gauge is the number of inferior models wrongly included in the MCS. For obvious reasons, we only report the gauge for $\lambda > 0$. Ideally, the MCS should remove the remaining nine models, and identify Model 1 as the unique best model. Of course, potency and gauge are strongly interlinked, through the level of the MCS. A higher level will make the procedure more potent, but will worsen the gauge.

Results are reported in Table 3. First consider $\lambda = 0$ for the various T. Recall that when $\lambda = 0$, all models are identical. In this case, the MCS procedure should not remove any model. This is a very stringent test, especially for the multi-horizon MCS. When $H = 1$, we see that the potency is close to the 80% for all T. The potency is even higher than nominal for the larger sample sizes. However, for larger H we see that potency is reduced significantly. For $T = 250$ with $H = 20$ it even reduces to 55%. The decrease in potency can be explained by the increase in loss variance and the strong correlation structure we generate. If by pure chance a model does have slightly higher average loss, it will have higher loss across all horizons. As a result the multi-horizon tests accumulate this information and removes the models. Importantly, it is completely in line with the standard potency-gauge trade-off.

Indeed, when looking at $\lambda > 0$, the MCS is a single model, and potency is well above the required 80% for all T and H. The gauge is decreasing in all parameters H, T and λ. That is, the MCS is better able to remove inferior models the more horizons we consider, the more time-series observation we have, and the greater the loss differentials between the models. Note that the effect of the number of horizons is large. The decrease in gauge of going from $H = 1$ to $H = 5$ is of an entirely different magnitude than increasing

the number of observations from $T = 250$ to $T = 1000$. As such, when a model truly has superior multi-horizon SPA, using multiple horizons is a powerful, and almost always feasible, way to differentiate the models.

4 Multi-Horizon Comparison of Direct and Iterated Forecasts

In this section we revisit the results of Marcellino et al. (2006), who investigate the performance of iterated versus direct forecasts using 170 monthly U.S. macroeconomic time series spanning 1959 to 2002.[9] They find that iterated forecasts tend to outperform direct forecasts, and the relative performance improves with the forecast of horizon. In their empirical analysis, they only consider four different horizons, $h = 3, 6, 12$ and 24. Based on the example in Figure 1, it is clear that picking just four out of all possible horizons may lead to unrepresentative, and potentially wrong, conclusions.

Here we test for significant superior predictive ability using the two tests developed in this paper. We test for uniform and average SPA across horizons $h = 2, ..., 24$. We exclude the first horizon since iterated and direct forecasts are equivalent for $h = 1$. For the sake of comparison, we also report the single-horizon Diebold-Mariano results.

We use the data provided on Mark Watson's website. The data consists of 170 series divided up into five different categories. We apply their suggested data transformation to deal with the non-stationary nature of some of the series, such that models are estimated in levels, log-levels, differences or log-differences. Forecasts are similarly evaluated on the transformed series. The number of observations per series varies between 412 and 528, with an average of 510 observations. For more details, we refer to Marcellino et al. (2006).

We mostly follow the forecasting methodology of Marcellino et al. (2006). We perform direct and iterated AR(p) forecasts, with four different choices of lag orders. First, we set p equal to either 4 or 12. Second, every period, we choose the optimal lag-length

[9]The supplemental appendix contains an additional application of the the multi-horizon MCS to evaluate a wide range of realized volatility forecasting models on multiple horizons jointly, using the dataset of Bollerslev, Patton, and Quaedvlieg (2016).

between 1 and 12, based on either AIC or BIC using the estimation sample. Note that it is entirely possible that on any given period the lag selection based on AIC or BIC results in different lag-lengths for the direct and iterated models. We then compare the direct and iterated forecasts per lag selection procedure. Our parameter estimates are based on a rolling window of 120 observations, rather than the expanding window used in Marcellino et al. (2006), since our framework requires non-vanishing estimation error.

For the iterated forecasts, we estimate the parameters of the following model using OLS.

$$y_{t+1} = \theta_0 + \sum_{i=1}^{p} \theta_i y_{t+1-i} + \epsilon_{t+1}. \tag{23}$$

The iterated $h-$step ahead forecasts are constructed recursively as

$$\hat{y}^{It}_{t+h|t} = \hat{\theta}_0 + \sum_{i=1}^{p} \hat{\theta}_i y_{t+h-i|t}. \tag{24}$$

For the direct forecasts, we estimate a model on the $h-$step ahead observation,

$$y_{t+h} = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t+1-i} + \epsilon_{t+h}. \tag{25}$$

To remain strictly out-of-sample, we only use data from the 120 observations of our rolling window, i.e. the last observation on the left-hand side is part of those 120 observations. Note that this does reduce the actual number of observations used for parameter estimation. We then obtain direct $h-$step ahead forecasts as

$$\hat{y}^{Dir}_{t+h|t} = \hat{\phi}_0 + \sum_{i=1}^{p} \hat{\phi}_i y_{t+1-i}. \tag{26}$$

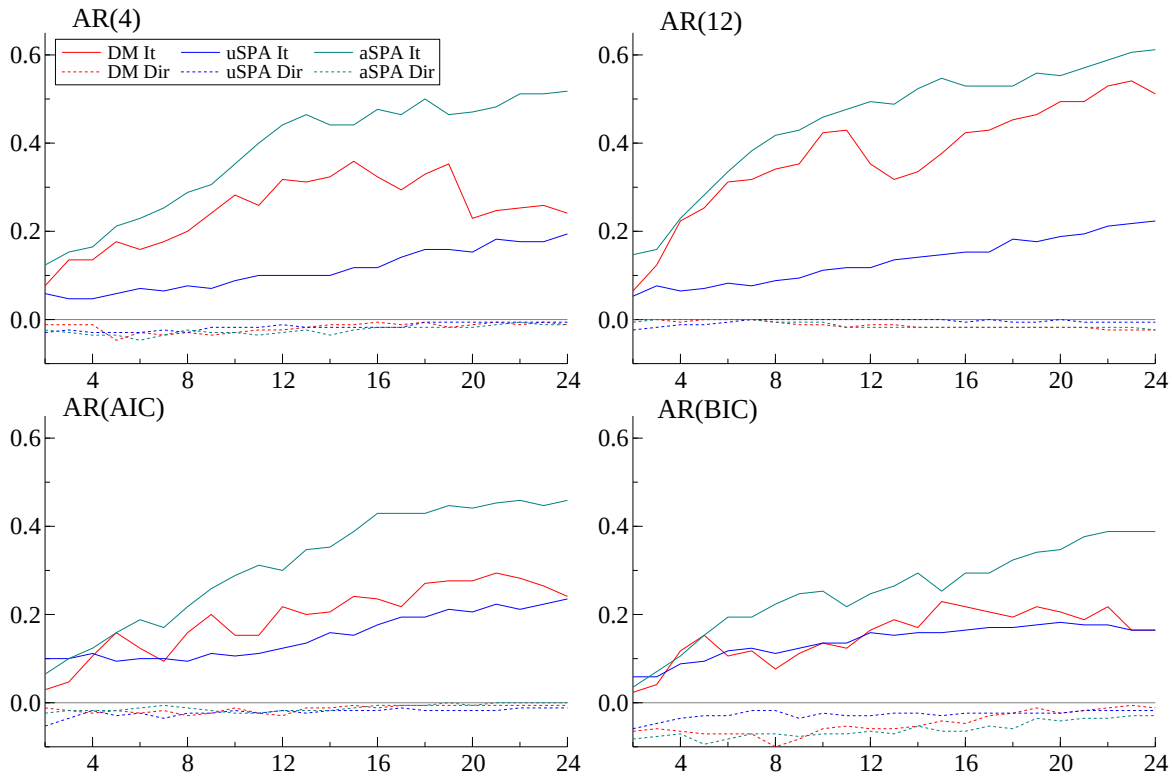
The forecasts are evaluated using the mean square forecasting error (MSFE)

$$L^{MSFE}(\hat{y}_{t+h|t}, y_{t+h}) = (\hat{y}_{t+h|t} - y_{t+h})^2. \tag{27}$$

4.1 Aggregate Results

Throughout this section we will report results of the multi-horizon tests for the range of maximum horizons $H = 2, ..., 24$. This should be interpreted as illustration of the tests, while in practice it is recommended to choose a single long-term horizon H, which includes all relevant horizons h.

Figure 2: Rejection Frequencies equal forecasting performance across horizons.



Note: This figure plots fraction of rejections out of 170 series, as a function of horizon. The tests in either direction are performed at the 2.5% significance level. Positive, solid lines plot the rejections in favor of Iterated forecasts, while the negative, dotted lines plot rejections in favor of Direct forecasts. The different plots depict the fractions for different lag-selection methods.

We formally test for superior predictive ability using the Diebold-Mariano, uSPA and aSPA tests on each of the 170 series and each of the 23 horizons. Figure 2 summarizes the rejection frequencies for one-sided tests in either direction at 2.5% level. Each of the four panels corresponds to one of the lag selections. The positive solid lines are the rejection frequencies in favor of iterated forecasts, while the negative dotted lines are the negative of the rejection frequencies in favor of direct forecasts.

The significance tests are mostly in line the results of Marcellino et al. (2006). Across the three tests, we find convincing evidence in favor of iterated forecasts. Rejection frequencies in favor of direct forecasts are typically at, or below, the level of the test, suggesting that iterated forecasts are no worse than direct forecasts. Only for lag-selection based on BIC, which tends to select the smallest models, we find rejection frequencies higher than the

level of the tests for small H. Interestingly, these higher rejection frequencies converge to close to zero when H grows.

Of course none of the three tests are directly comparable, but the rejection frequencies at different horizons serve to highlight the merits of joint multi-horizon tests. The Diebold-Mariano test hardly ever rejects for short horizons, which rises to about 20% for the two-year ahead forecast. Importantly, the number of rejections is unstable across horizons. For instance, based on AR(12), looking at just horizon $h = 11$ we would reject for over 40% of the series, while horizon $h = 13$ would reject less than 30%.

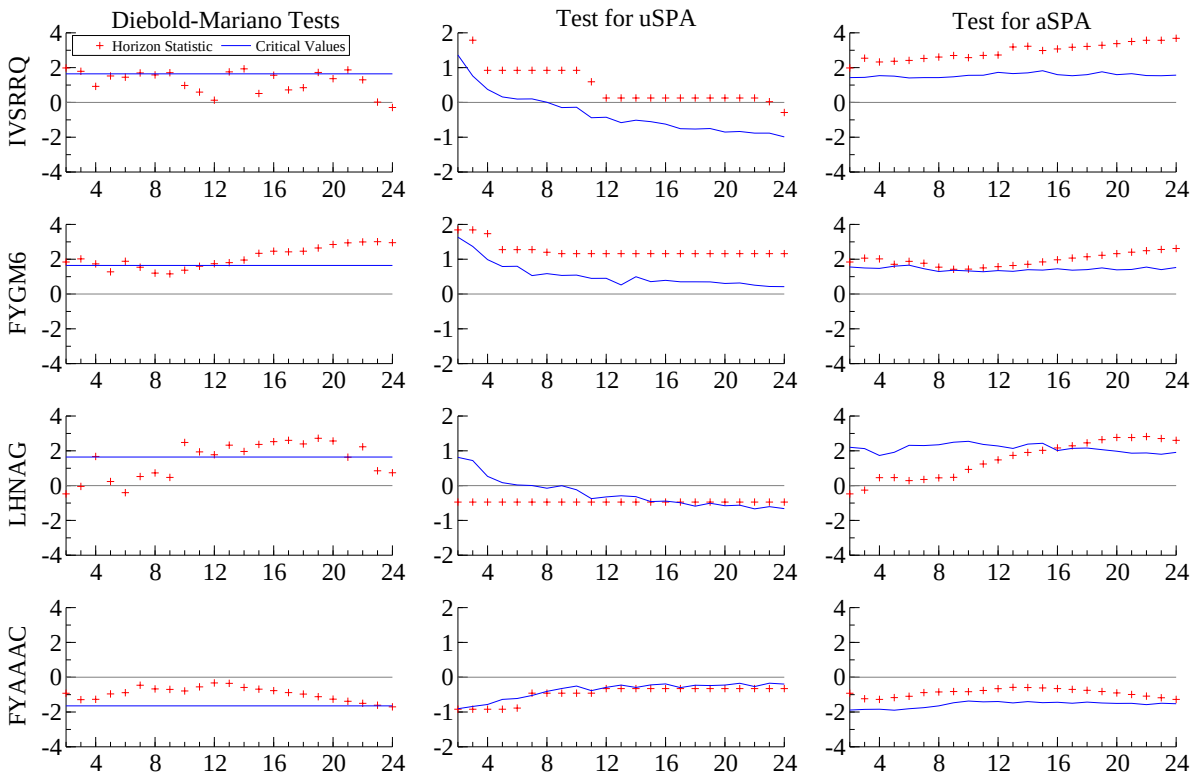
Naturally, we typically find fewer rejections based on the test for uSPA, settling at about 20% of the series for $H = 24$. The total amount of rejections is however naturally increasing in the number of horizons under consideration H, suggesting coherent conclusions irrespective of number of the actual chosen horizon. In contrast to the DM-test, the rejection rates are also mostly stable across the four panels.

Of course, even if the test for uSPA fails to differentiate models, the test for aSPA still may, as it is the weaker hypothesis. We find that the rejection fractions of the test for aSPA are indeed higher than those of uSPA, but also consistently higher than those of the single-horizon Diebold-Mariano tests. Similar to the test for uSPA, the rejection frequencies are almost monotonically increasing in the horizon H. We find that the across the 23 horizons, iterated forecasts provide average superior predictive ability relative to direct forecasts for between 40% and 60% of the series. The contrast with the DM test is easy to understand. Mechanically, a small loss differential at a single horizon results in a failure to reject for the univariate test, while the multi-horizon test may find that the evidence at shorter horizons is sufficient to compensate.

4.2 Individual Results

To better illustrate the relative merits of the various hypotheses and tests, we zoom in on a number of individual series in Figure 3. Each column corresponds to one of the three tests, Diebold-Mariano, uSPA and aSPA. The red cross denotes the test-statistic at, or up to, horizon h. The blue line provides the one-sided critical value at 5%. For the DM-test this is based on the Gaussian quantiles, while for the •SPA tests we report $c_{ij}^{5\%}$ based on

Figure 3: Individual Test Results



Note: This figure plots test statistics with critical values for the univariate, uniform and average SPA tests as a function of horizon. We highlight four series from the Marcellino et al. (2006) dataset. IVSRRQ is log-difference of the Inventory over Sales Ratio of retail trade. FYGM6 is the log of the 6 month US treasury bill interest rate. LHNAG is the log-difference of non-agricultural employed civilian labor force. FYAAC is the log of bond yield on AAA securities.

Bootstrap Algorithm 1. Each row corresponds to a different time-series, chosen to highlight various facets of the tests.

We observe a number of different patterns. For instance, IVSRRQ has a positive Diebold-Mariano test-statistic at each horizons but $h = 24$. The single-horizon test is only significant at a small number of horizons and insignificant at all others. The test for aSPA however, aggregates the information over multiple horizons, which are all positive, and finds sufficient evidence at all horizons to conclude that the iterated forecasts outperform the direct forecasts. The statistics are actually increasing in horizon, due to reduced variance ζ_{ij}. The single negative loss differential at $h = 24$ clearly does not provide sufficient evidence to reject aSPA. Moreover, it does not even provide sufficient evidence to reject uSPA of the iterated forecasts. As the bootstrapped critical values clearly illustrate,

when we consider more than a single horizon, we might reasonable expect to observe a negative differential, even if the true loss differential μ_{ij}^h is positive for all h. As a result, we conclude that iterated forecast provide both uSPA and aSPA, despite only finding significant evidence of superior predictive ability at a four horizons using the Diebold-Mariano test.

FYGM6 shows a similar picture, but with more consistent relative performance. The iterated forecasts perform better at every horizon, and the single-horizon test find significant evidence for most horizons. Again, we find evidence for aSPA at all horizons, although this time the test statistics hardly increase for longer horizons H. More interesting is that we are now in a situation where limited variability in loss-differentials results in a case where the critical value of uSPA remains positive, even at $H = 24$. Due to the consistent performance of the iterated forecasts we still find evidence for uSPA.

The third series, LHNAG, has no clear winner at short horizons, but iterated forecasts appear to dominate direct forecasts at longer horizons. The single-horizon statistic picks up on this, with significant differentials at twelve, non-consecutive horizons. The test for aSPA combines the joint evidence and rejects the null from $H \geq 16$. The test for uSPA is severely impacted by the negative statistic at $h = 2$. However, this negative statistic was small, and is not surpassed at higher horizons. As a result, starting from $H = 18$ and up, we conclude that the negative short-horizon statistic was likely sampling error, and find support for uSPA of iterated forecasts.

The final example, FYAAAC is a series where the direct forecasts appear to mostly outperform the iterated ones. All forecast differentials are negative but small. Their level results in a situation in which the univariate and average statistic are insignificant at all horizons. However, its consistently negative values results in the fact that the uniform statistic does reject at all horizons $H \geq 8$. Hence, we find evidence for uSPA, but not for aSPA. While the definition of uSPA implies aSPA, in any given sample, the tests may of course not reach this conclusion. A result like this occurs rarely though, and across the 170 series we perform both these tests, we only find evidence for uSPA and not for aSPA a negligible two times, while the reverse is pervasive throughout.

Overall, Figure 3 makes it clear that comparing forecast path accuracy by looking

at individual horizons is often insufficient to understand whether a model has superior predictive ability or not. The joint performance over multiple horizons provides a clearer and more consistent picture then the single-horizon statistics.

5 Conclusion

We introduce the notion of multi-horizon forecast comparison. We propose to jointly evaluate multiple horizons when testing for superior predictive ability, rather than considering multiple horizons individually. We argue this has three advantages. First, multi-horizon superior predictive ability provides a more complete definition of a model's superior performance. Second, tests that involve multiple horizons are generally more powerful than tests that only consider a single horizon, allowing us to disentangle models more easily. Finally, it guards us against the implicit multiple testing issue arising from picking and choosing (potentially multiple) individual horizons.

We propose two bootstrap-based tests that evaluate different hypotheses of multi-horizon forecasting performance. The first tests for uniform superior predictive ability, which is defined as superior forecasts at each individual horizon. The second tests the weaker hypothesis that the average loss across horizons is lower. Both tests reduce to the standard Diebold-Mariano test when only considering a single horizon. We demonstrate that the ability to differentiate models empirically increases with the number of horizons under consideration. While forecast error variance increases in horizon, model mis-specification also tends to increase the average forecast loss as a function of horizon, which is the main driver of the increased power.

The basic tests allow the statistical comparison of two models. In addition, in order to compare a greater number of models directly, we extend the Model Confidence Set methodology to allow for multiple-horizon evaluation. The procedure allows us to find the set of models that contains the model with multi-horizon superior predictive ability with a certain confidence level. Both the pairwise tests and the Model Confidence Set are shown to be properly sized and powerful in simulations.

The pairwise comparison is illustrated on the comparison of direct and iterated of macro-economic variables, based on the data in Marcellino et al. (2006). We find that

despite conflicting evidence when looking at individual horizons, we are often able to find statistical evidence for either average SPA or uniform SPA, or both, when considering multiple horizons jointly. This suggests that the conflicting evidence is typically the results from the implicit multiple-testing issue of picking and choosing a few horizons.

While there are situations in which we may not be interested in a single model for all horizons, many forecasting problems are expected to have a unique best model at all horizons; the model that best approximates the data generating process. The tests for uniform and average superior predictive ability therefore have wide applicability in many fields. In these situations, the tests provide a more consistent, more reliable and more powerful method of distinguishing models.

References

Bartholomew, D., 1961. A test of homogeneity of means under restricted alternatives. Journal of the Royal Statistical Society. Series B (Methodological), 239–281.

Bollerslev, T., Patton, A. J., Quaedvlieg, R., 2016. Exploiting the errors: A simple approach for improved volatility forecasting. Journal of Econometrics 192 (1), 1–18.

Breitung, J., Knüppel, M., 2017. How far can we forecast? statistical tests of the predictive content. Working Paper.

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

Clark, T., McCracken, M., 2013. Advances in forecast evaluation. In: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting, Vol. 2. North Holland, Amsterdam, pp. 1107–1201.

Clark, T. E., McCracken, M. W., 2005. Evaluating direct multistep forecasts. Econometric Reviews 24 (4), 369–404.

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

De Jong, R. M., 1997. Central limit theorems for dependent heterogeneous random variables. Econometric Theory 13 (3), 353–367.

Diebold, F. X., Mariano, R. S., 1995. Comparing predictive accuracy. Journal of Business & Economic Statistics 13 (3), 134–144.

Doornik, J. A., 2012. An Object-Oriented Matrix Programming Language Ox 7. Timberlake Consultants Ltd.

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

Gonçalves, S., de Jong, R., 2003. Consistency of the stationary bootstrap under weak moment conditions. Economics Letters 81 (2), 273–278.

Gonçalves, S., White, H., 2002. The bootstrap of the mean for dependent heterogeneous arrays. Econometric Theory 18 (6), 1367–1384.

Gonçalves, S., White, H., 2005. Bootstrap standard error estimates for linear regression. Journal of the American Statistical Association 100 (471), 970–979.

Hansen, P. R., 2005. A test for superior predictive ability. Journal of Business & Economic Statistics 23 (4), 365–380.

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

Hendry, D. F., Doornik, J. A., 2014. Empirical model discovery and theory evaluation: automatic selection methods in econometrics. Cambridge, Massachusetts: MIT Press.

Ing, C.-K., 2003. Multistep prediction in autoregressive processes. Econometric theory 19 (02), 254–279.

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

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

Linton, O., Maasoumi, E., Whang, Y.-J., 2005. Consistent testing for stochastic dominance under general sampling schemes. The Review of Economic Studies 72 (3), 735–765.

Linton, O., Song, K., Whang, Y.-J., 2010. An improved bootstrap test of stochastic dominance. Journal of Econometrics 154 (2), 186–202.

Marcellino, M., Stock, J. H., Watson, M. W., 2006. A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. Journal of Econometrics 135 (1), 499–526.

Martinez, A., 2017. Testing for differences in path forecast accuracy: Forecast-error dynamics matter. Working Paper.

Patton, A. J., Timmermann, A., 2010. Monotonicity in asset returns: New tests with applications to the term structure, the capm, and portfolio sorts. Journal of Financial Economics 98 (3), 605–625.

Patton, A. J., Timmermann, A., 2012. Forecast rationality tests based on multi-horizon bounds. Journal of Business & Economic Statistics 30 (1), 1–17.

Politis, D. N., Romano, J. P., 1994. The stationary bootstrap. Journal of the American Statistical Association 89 (428), 1303–1313.

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

White, H., 2000. A reality check for data snooping. Econometrica 68 (5), 1097–1126.

Wolak, F. A., 1987. An exact test for multiple inequality and equality constraints in the linear regression model. Journal of the American Statistical Association 82 (399), 782–793.

A Bootstrap Validity

Proof Theorem 1: Under either null hypothesis $\boldsymbol{D}^{-1}\bar{\boldsymbol{d}} \to_d N(0, \boldsymbol{R})$, where $\boldsymbol{R} = \boldsymbol{D}^{-1}\boldsymbol{\Omega}\boldsymbol{D}^{-1}$. First, to prevent the need for a double bootstrap, we use the closed-form expression for

the stationary bootstrap variance of the mean, presented in Equation (11), to estimate $\hat{\boldsymbol{D}}$. Under the stated assumptions on q_T, and assumptions slightly weaker than those presented in 1, Theorem 1 of Gonçalves and de Jong (2003) proves that $\hat{\boldsymbol{D}}$ is consistent for $\boldsymbol{D}$. As a result, by standard arguments we have that $\hat{\boldsymbol{D}}^{-1}\bar{\boldsymbol{d}} \to N(0, \boldsymbol{R})$.

Next, we show that the bootstrap consistently estimates the distribution of $\hat{\boldsymbol{D}}^{-1}\bar{\boldsymbol{d}}$. From Theorem 2 of Gonçalves and de Jong (2003) it follows that

$$\sup_{x \in \mathbb{R}^H} |P^b(\bar{\boldsymbol{d}}^b - \bar{\boldsymbol{d}} \leq x) - P(\bar{\boldsymbol{d}} - \boldsymbol{\mu} \leq x)| \to_p 0, \tag{28}$$

where P^b denotes the bootstrap distribution. This demonstrates that the bootstrap distribution can be used to approximate the distribution of $(\bar{\boldsymbol{d}} - \boldsymbol{\mu})$. It however does not immediately justify the validity of the bootstrap for the studentized statistics (see e.g. Gonçalves and White, 2005).

For

$$\sup_{x \in \mathbb{R}^H} |P^b(\boldsymbol{D}^b)^{-1}(\bar{\boldsymbol{d}}^b - \bar{\boldsymbol{d}}) \leq x) - P(\hat{\boldsymbol{D}}^{-1}(\bar{\boldsymbol{d}} - \boldsymbol{\mu}) \leq x)| \to_p 0, \tag{29}$$

to hold, we need that $\boldsymbol{D}^b \to_p \boldsymbol{D}$, which follows from Gonćalves and White's (2002) Corollary 2.1.

Online Appendix to:

Multi-Horizon Forecast Comparison

February 28, 2018

Rogier Quaedvlieg

1 Volatility Forecasting

To illustrate the comparison of more than two models, we consider volatility forecasting. The widespread availability of intraday returns has resulted in a large and well-developed literature in reduced form volatility modeling. While the latent volatility historically needed to be filtered from data using for instance GARCH-type models, the sum of squared intraday data can be used to obtain accurate daily estimates of volatility. As a result, volatility essentially becomes observable. This means that we can model volatility directly in reduced form, making forecasting and evaluation simple.

We use the dataset of Bollerslev et al. (2016), which provides Realized Volatility estimates based on five-minute returns, as well as related measures for, the S&P500 futures and 27 individual Dow Jones stocks. The sample period is between April 1997 to December 2013, such that we obtain between 3096 and 3202 daily observations.

The literature on forecasting realized volatility is large, and we could not hope to cover every important contender. We therefore stick to eleven of the most prominent models, provided in Table S.1. We use six models that produce direct forecasts: the pure AR(1) and AR(22) models, as well as the four HAR-type specifications.

HAR models were first advocated in Corsi (2009) as a simple approximation of the long-memory ARFIMA models. The model is essentially a restricted AR(22) model which imposes common parameters for the weekly and monthly lags. The model has wide appeal and can be estimated simply using OLS. As a result the literature has built on the HAR model and many extension were proposed. Here we consider three of these extensions. First, we consider the HARQ of Bollerslev et al. (2016), who propose to let the autoregressive

Table S.1: Overview of Volatility Models

Model Name	Specification
Direct Forecasts	
AR(1)	$RV_{t+h-1\|t} = \phi_0 + \phi_1 RV_{t-1} + \epsilon_t$
AR(22)	$RV_{t+h-1\|t} = \phi_0 + \sum_{i=1}^{22} \phi_i RV_{t-i} + \epsilon_t$
HAR	$RV_{t+h-1\|t} = \phi_0 + \phi_1 RV_{t-1} + \phi_2 RV_{t-1\|t-5} + \phi_3 RV_{t-1\|t-22} + \epsilon_t$
HARQ	$RV_{t+h-1\|t} = \phi_0 + (\phi_1 + \phi_{1Q} RQ_{t-1}^{1/2}) RV_{t-1} + \phi_2 RV_{t-1\|t-5} + \phi_3 RV_{t-1\|t-22} + \epsilon_t$
CHAR	$RV_{t+h-1\|t} = \phi_0 + \phi_1 BPV_{t-1} + \phi_2 BPV_{t-1\|t-5} + \phi_3 BPV_{t-1\|t-22} + \epsilon_t$
SHAR	$RV_{t+h-1\|t} = \phi_0 + \phi_{1+} RS_{t-1}^{+} + \phi_{1-} RS_{t-1}^{-} + \phi_2 RV_{t-1\|t-5} + \phi_3 RV_{t-1\|t-22} + \epsilon_t$
Iterated Forecasts	
AR(1)	$RV_t = \phi_0 + \phi_1 RV_{t-1} + \epsilon_t$
AR(22)	$RV_t = \phi_0 + \sum_{i=1}^{22} \phi_i RV_{t-i} + \epsilon_t$
ARMA(1,1)	$RV_t = \phi_0 + \phi_1 RV_{t-1} + \epsilon_t + \theta_1 \epsilon_{t-1}$
ARFIMA$(0,d,0)$	$(1-L)^d RV_t = \phi_0 + \epsilon_t$
ARFIMA$(1,d,0)$	$(1-L)^d RV_t = \phi_0 + \phi_1 RV_{t-1} + \epsilon_t$

Note: This table provides the models used for the Volatility Forecasting Multi-Horizon MCS. The multi-period average is defined as $RV_{t-1|t-k} = \frac{1}{k}\sum_{i=1}^{k} RV_{t-i}$. RQ is the realized Quarticity, BPV is the Bi-Power Variation introduced in Barndorff-Nielsen and Shephard (2004), and RS^+ and RS^- are the semi-variances of Barndorff-Neilsen, Kinnebrouk, and Shephard (2010). Estimates and forecasts of the ARFIMA models are obtained using the ARFIMA package for Ox (Doornik and Ooms, 2012).

parameters vary over time as a function of the measurement error variance in the realized volatility estimates. The C(ontinuous-)HAR model decomposes Realized Volatility in a predictable continuous variation part, bi-power variation (BPV), and an unpredictable part due to jumps. Finally, we consider the S(emi-variance)HAR of Patton and Sheppard (2015). They decompose realized volatility into two semi-variances stemming from positive and negative returns (RS^+ and RS^-). The negative semi-variance has greater predictive power than the positive one, and as a result the decomposition leads to improved forecasts. All these measures are available in the dataset provided by Bollerslev et al. (2016).

Next, we use five models that provide iterated forecasts. We again take the pure autoregressive models, AR(1) and AR(22). We also consider an ARMA(1,1), which would,

similar to the HARQ, account for measurement error in RV if it were homoskedastic. Finally, we consider two ARFIMA models that account for the long-memory in realized volatility, which was documented in, amongst others, Andersen, Bollerslev, Diebold, and Labys (2003). Both the pure ARFIMA(0,d,0) and the augmented ARFIMA(1,d,0) are used in the volatility literature (e.g. Asai, McAleer, and Medeiros, 2012).

In contrast to the literature, we compute h-step ahead forecasts, RV_{t+h}, rather than cumulative forecasts, $RV_{t+h|t+1}$. The reason is that we already jointly consider all horizons, and using cumulative forecasts would strongly overweight short-term forecasts. We use rolling window parameter estimates based on 1000 observations. The maximum horizon we consider is $H = 20$, which corresponds to volatility for roughly the next month.

We consider two loss-functions. Based on the results of Patton (2011), we have to be careful in choosing the loss function, since we are evaluating the forecasts based on estimated RV_t, rather than the true, still latent, volatility. Therefore we use the consistent loss function MSFE:

$$L^{MSFE}(\widehat{RV}_{t+h}, RV_{t+h}) = (\widehat{RV}_{t+h} - RV_{t+h})^2 \tag{1.30}$$

as in the previous section, as well as the QLIKE:

$$L^{QLIKE}(\widehat{RV}_{t+h}, RV_{t+h}) = \frac{RV_{t+h}}{\widehat{RV}_{t+h}} - \log\left(\frac{RV_{t+h}}{\widehat{RV}_{t+h}}\right) - 1, \tag{1.31}$$

where $\widehat{RV}_{t+h}$ denotes the $h-$step ahead forecast and RV_{t+h} the ex-post estimate.

We summarize the results of this empirical test by looking at four horizons, $H = 1, 5, 10, 20$. First, we report the forecasting loss at the these individual horizons in Table S.2. We present the results for the S&P500 separately, as well as the average loss over the remaining Dow Jones stocks. The forecasting performance of the various models varies widely, both within a single horizon and across horizons. Across all stocks and both loss functions, at $H = 1$ the ARFIMA models appear to perform best, closely followed by the HARQ model. For the market the SHAR model has lowest loss. The AR(22) performs very poorly in terms of MSFE loss, but has better relative performance in terms of QLIKE, where the AR(1) model is kept at distance.

For longer horizons, some of the relative model performance is shifted around, but the most striking feature is that the distance between the losses of the ARFIMA model and

all other specifications increases in h. While the approximate long-memory provided by the HAR-type models provides a reasonable approximation at short horizons, it appears that the actual long-memory provided by the fractionally integrated becomes important at longer horizons.

In order to determine whether the differences in loss are statistically significant, we compute the Multi-Horizon MCS based on forecasts up to these same four horizons. The $H = 1$ case corresponds to the standard MCS for one-step ahead forecasts, while the longer horizons are based on the contributions in this paper, and use the forecasts from horizons 1 to H. The results are reported in Table S.3. For the S&P500 we report the p-values, relating to the probability that the respective model is in the multi-horizon MCS. In the bottom panel we report the fraction of stocks where each respective model was in the 80% multi-horizon MCS. That is, the models that p-values greater than 0.2. First consider the top panel.

As has often been noted, the more volatile nature of MSFE makes it more difficult to distinguish models. Indeed, for the S&P500 at $H = 1$, the forecasting performance of all eleven models is statistically equivalent, which also largely holds for the individual stocks. For larger H, the MSFE-based MCS quickly shrinks to include only the two ARFIMA models at $H = 20$. Based on QLIKE, the HAR, HARQ and SHAR can keep up with the ARFIMA models up until horizon $H = 10$, with HARQ being the single best model at $H = 5$. However, when considering the full forecasting path over the next month, the ARFIMA models jointly provide the best forecasts for both loss-functions.

For the individual stocks we see a similar picture, although the multi-horizon MCS appears to have more gauge for small H, and less for large H, compared to the S&P500 results. Based on MSFE, the ARFIMA models again dominate, as one of the two models is included in the MCS for all stocks at all horizons. While gauge is clearly increasing in H, with the average size of the MCS almost halving, the MCS has more difficulty eliminating candidate models for individual stocks. For instance, both the HARQ and SHAR remain in over half the MCS at $H = 20$. The QLIKE results are similar to the MSFE results, in that the ARFIMA models are in the MCS for almost all the series. Across H, we again observe that out of the HAR-type specifications, HARQ and SHAR appear to be the best

models, and remain in the MCS for over half the series, even at $H = 20$. Between the two ARFIMA specifications, adding the AR component does appear to give the slight edge, especially when we consider longer horizons forecast paths.

Table S.2: Realized Volatility: Forecast Loss

	$h=1$		$h=5$		$h=10$		$h=20$	
	MSFE	QLIKE	MSFE	QLIKE	MSFE	QLIKE	MSFE	QLIKE
				S&P500				
Direct								
AR(1)	2.982	0.215	3.946	0.331	5.102	0.410	6.027	0.537
AR(22)	5.884	0.191	6.550	0.340	6.268	0.403	7.207	0.570
HAR	3.628	0.148	6.157	0.248	7.900	0.346	7.976	0.488
HARQ	2.911	0.138	6.111	0.248	8.438	0.365	9.123	0.538
CHAR	3.589	0.150	6.053	0.256	8.299	0.377	8.285	0.498
SHAR	2.719	0.132	5.273	0.260	6.615	0.349	7.040	0.483
Iterated								
AR(1)	2.982	0.215	5.608	0.552	6.303	0.740	6.463	0.832
AR(22)	5.884	0.191	6.050	0.350	6.913	0.427	6.904	0.533
ARFIMA$(0,d,0)$	2.686	0.135	3.585	0.237	4.183	0.309	4.957	0.408
ARFIMA$(1,d,0)$	2.759	0.135	3.565	0.230	4.150	0.304	4.887	0.406
ARMA(1,1)	2.878	0.147	3.933	0.276	4.597	0.406	5.667	0.591
				Average across stocks				
Direct								
AR(1)	19.784	0.233	25.424	0.324	27.031	0.369	29.467	0.441
AR(22)	24.602	0.190	32.010	0.262	35.004	0.323	36.813	0.386
HAR	18.093	0.168	26.282	0.242	30.593	0.296	31.227	0.376
HARQ	16.299	0.160	24.946	0.237	30.068	0.301	30.764	0.377
CHAR	18.305	0.168	26.768	0.242	30.997	0.297	31.220	0.379
SHAR	17.568	0.160	25.023	0.235	28.816	0.290	30.812	0.371
Iterated								
AR(1)	19.784	0.233	30.538	0.477	32.642	0.563	33.875	0.605
AR(22)	24.602	0.190	32.140	0.264	34.714	0.323	35.675	0.405
ARFIMA$(0,d,0)$	15.669	0.156	20.457	0.229	22.328	0.272	25.264	0.333
ARFIMA$(1,d,0)$	16.071	0.156	20.738	0.224	22.501	0.268	25.320	0.328
ARMA(1,1)	16.857	0.169	24.191	0.262	27.012	0.343	30.986	0.448

Note: This table provides the empirical forecasting loss for the various volatility forecasting models. The top panel presents results for the realized volatility of S&P500 futures, while the bottom panel provides the average loss across 27 DJIA stocks.

Table S.3: Realized Volatility: Multi-Horizon Model Confidence Set Results

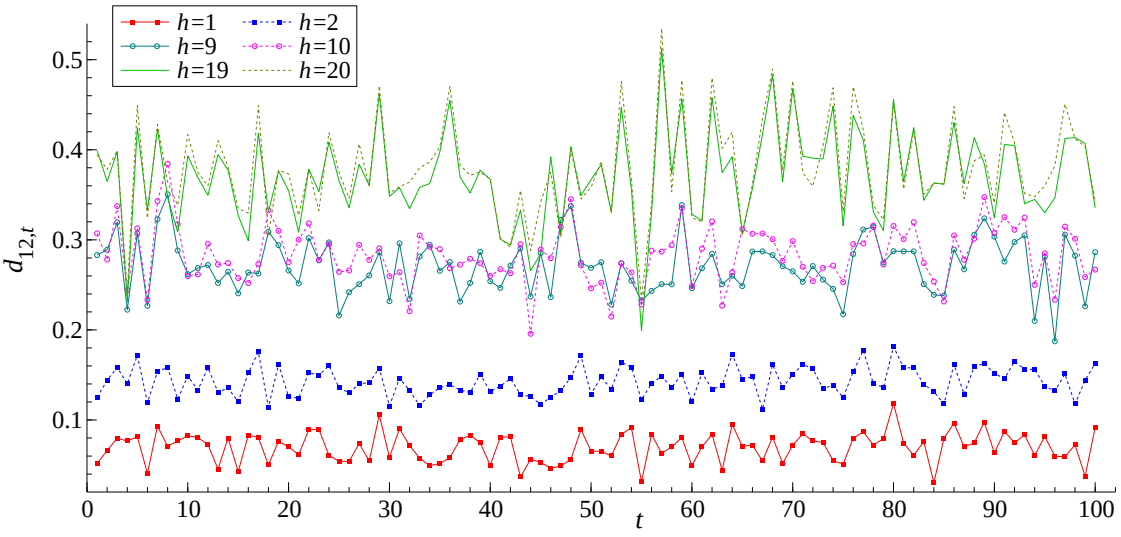
	$H=1$		$H=5$		$H=10$		$H=20$	
	MSFE	QLIKE	MSFE	QLIKE	MSFE	QLIKE	MSFE	QLIKE
	S&P500 Multi-Horizon MCS p-values							
Direct								
AR(1)	**0.385**	0.000	0.093	0.000	0.097	0.000	0.088	0.001
AR(22)	**0.385**	0.000	0.093	0.001	0.097	0.003	0.194	0.002
HAR	**0.385**	**0.424**	0.093	**0.592**	0.097	**0.435**	0.194	0.182
HARQ	**0.437**	**0.547**	0.118	**1.000**	0.097	**0.435**	0.143	0.182
CHAR	**0.437**	0.000	0.118	0.002	0.097	0.006	0.143	0.003
SHAR	**0.676**	**1.000**	**0.789**	**0.629**	**0.441**	**0.435**	0.194	0.182
Iterated								
AR(1)	**0.437**	0.000	**0.789**	0.000	0.080	0.000	0.063	0.001
AR(22)	**0.385**	0.000	0.093	0.000	0.080	0.004	0.194	0.030
AFRIMA$(0,d,0)$	**1.000**	**0.547**	**0.789**	**0.629**	**0.441**	**1.000**	**0.257**	**1.000**
ARFIMA$(1,d,0)$	**0.676**	**0.547**	**1.000**	**0.629**	**1.000**	**0.435**	**1.000**	**0.328**
ARMA(1,1)	**0.466**	0.001	0.093	0.000	0.080	0.000	0.063	0.001
	Fraction of stocks included in 80% Multi-Horizon MCS							
Direct								
AR(1)	0.556	0.000	0.111	0.000	0.111	0.000	0.037	0.000
AR(22)	0.630	0.111	0.185	0.296	0.222	0.333	0.148	0.185
HAR	0.852	0.667	0.852	0.593	0.519	0.481	0.407	0.259
HARQ	0.889	0.741	0.889	0.741	0.852	0.815	0.556	0.519
CHAR	0.815	0.148	0.667	0.037	0.593	0.259	0.296	0.111
SHAR	0.704	0.407	0.815	0.630	0.778	0.593	0.519	0.519
Iterated								
AR(1)	0.556	0.000	0.148	0.000	0.148	0.000	0.185	0.000
AR(22)	0.556	0.000	0.185	0.000	0.259	0.037	0.259	0.000
AFRIMA$(0,d,0)$	1.000	0.889	1.000	0.741	1.000	0.704	0.963	0.704
ARFIMA$(1,d,0)$	0.926	0.852	1.000	0.889	1.000	0.926	1.000	0.926
ARMA(1,1)	0.778	0.148	0.593	0.037	0.556	0.074	0.519	0.037
Average size $\widehat{\mathcal{M}}_{0.2}$	8.259	3.963	6.444	3.963	6.037	4.222	4.889	3.259

Note: This table provides the multi-horizon Model Confidence Set results. The top panel provides p-values for the S&P500 Realized Variance forecasts. Boldface denotes the model is part of the 80% multi-horizon MCS. The bottom panel presents the fraction of the 27 firms for which the model was included in the 80% multi-horizon MCS. The final row provides the average number of models in the 80% multi-horizon MCS.

2 Illustration Data Generating Process Simulation

In order to visualize the choice of DGP in Section 3 we plot several loss differentials $d_{ij,t}^h$ across 100 time periods in Figure S.1. The losses are generated based on $\theta^{(Unif)}$, and averaged across 10,000 simulations to reduce the variance. We plot the loss at three pairs of adjacent horizons, $h = 1, 2, 9, 10, 19$ and 20. The figure illustrates all the major facets of the data generating process. First, driven by ϕ, the increase in loss differential between horizons $h = 1$ and 2 is greater than that between the other two adjacent pairs. Next, as governed by ψ, the variance of the loss differential increases with h. Third, due to Equation (22), the correlation between adjacent pairs is increasing with h, as evidenced by the clearly stronger correlation between $h = 19$ and 20, compared to the other two pairs.

Figure S.1: Illustration of Simulation Data Generating Process



Note: This figure plots average simulated loss differential across 10,000 simulations. The simulations are based on parameter choices $\phi = 1$, $\psi = 0.125$ and $\lambda = 10$, which is the median parameter choice for the remainder of the simulations.

Additional References

Andersen, T. G., Bollerslev, T., Diebold, F. X., Labys, P., 2003. Modeling and forecasting realized volatility. Econometrica 71 (2), 579–625.

Asai, M., McAleer, M., Medeiros, M. C., 2012. Modelling and forecasting noisy realized volatility. Computational Statistics & Data Analysis 56 (1), 217–230.

Barndorff-Neilsen, O. E., Kinnebrouk, S., Shephard, N., 2010. Measuring downside risk: realised semivariance. Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle,(Edited by T. Bollerslev, J. Russell and M. Watson) 117 (136), 117–136.

Barndorff-Nielsen, O. E., Shephard, N., 2004. Power and bipower variation with stochastic volatility and jumps. Journal of Financial Econometrics 2 (1), 1–37.

Bollerslev, T., Patton, A. J., Quaedvlieg, R., 2016. Exploiting the errors: A simple approach for improved volatility forecasting. Journal of Econometrics 192 (1), 1–18.

Corsi, F., 2009. A simple approximate long-memory model of realized volatility. Journal of Financial Econometrics 7 (2), 174–196.

Doornik, J. A., Ooms, M., 2012. A package for estimating, forecasting and simulating Arfima models: Arfima package 1.06 for Ox.

Patton, A. J., 2011. Volatility forecast comparison using imperfect volatility proxies. Journal of Econometrics 160 (1), 246–256.

Patton, A. J., Sheppard, K., 2015. Good volatility, bad volatility: Signed jumps and the persistence of volatility. Review of Economics and Statistics 97 (3), 683–697.