Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Multi-Horizon Forecast Comparison

Rogier Quaedvlieg

To cite this article: Rogier Quaedvlieg (2021) Multi-Horizon Forecast Comparison, Journal of Business & Economic Statistics, 39:1, 40-53, DOI: 10.1080/07350015.2019.1620074

To link to this article: https://doi.org/10.1080/07350015.2019.1620074

Multi-Horizon Forecast Comparison

Rogier QUAEDVLIEG
Department of Business Economics, Erasmus School of Economics, PO Box 1738, 3000 DR Rotterdam, The Netherlands (*quaedvlieg@ese.eur.nl*)

We introduce tests for multi-horizon superior predictive ability (SPA). Rather than comparing forecasts of different models at multiple horizons individually, we propose to jointly consider all horizons of a forecast path. We define the concepts of uniform and average SPA. The former entails superior performance at each individual horizon, while the latter allows inferior performance at some horizons to be compensated by others. The article illustrates how the tests lead to more coherent conclusions, and how they are better able to differentiate between models than the single-horizon tests. We provide an extension of the previously introduced 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.

KEY WORDS: Forecasting; Long-horizon; Multiple testing; Path forecasts; Superior predictive ability.

1. INTRODUCTION

Forecasts at multiple horizons should rarely be judged in isolation. The full forecast path plays an important role in many policy decisions. For instance, in the context of macroeconomic 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 intermittent path the variable takes between now and sometime in the future. 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 path forecasts, it is useful to compare the accuracy of the complete path.

The standard approach is to compare various models at different horizons independently, potentially leading to incoherent conclusions. For example, in a given sample, we might find that a first model is significantly better at predicting two and five periods ahead, 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. Indeed, when we compare performance at multiple horizons, we implicitly face a multiple testing problem. As such, 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 (SPA). There are at least three reasons why one might be interested in such a test. First, it entails a more robust definition of a model's SPA. Second, 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 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 clearly far more stringent, but by properly controlling the family-wise error rate using bootstrap methods, 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 et al. (2011), which allows the comparison of 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.

The tests proposed in this article fall into the framework implicitly defined in Diebold and Mariano (1995), and explicitly set out in, among others, Hansen (2005) and Giacomini

Journal of Business & Economic Statistics
January 2021, Vol. 39, No. 1
DOI: 10.1080/07350015.2019.1620074
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jbes*.

and White (2006). 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), and greatly expanded on by amongst others Clark and McCracken (2005, 2012) and Clark and West (2007), 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) provided an excellent overview of the literature. The asymptotic theory in this finite-sample 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.

In practice, the proposed tests 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 model has uniform SPA across all horizons that can reasonably be forecasted, which is strong evidence in favor of a specification. However, in many cases, users may have different models for different ranges, that is, short-, mid-, and long-term 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. In the context of model comparison, Capistrán (2006) introduces an unweighted version of the average SPA test. Subsequent research by Martinez (2017) provides a generalization of the unweighted average SPA test in a GFESM context (Clements and Hendry 1993), explicitly allowing for differences in covariance dynamics of the various models, while we target the loss-differential directly as a primitive. Finally, the literature on vector forecasts, concerning multiple variables rather than multiple horizons, faces the similar problem of forecast comparison in the presence of correlated forecast errors (e.g., Clements and Hendry 1993; Komunjer and Owyang 2012).

We analyze the finite sample properties of the tests in simulation studies. We consider the two pairwise tests and the multi-horizon MCS. 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 comparisons will lead to more frequent rejection than a test applied to a subset of the same paths. Naturally, this is determined by the relative increases in average loss differentials and the variance of the loss differential as a function of horizon.

As an empirical illustration, 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 2–24 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.

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 illustration, and finally Section 5 concludes.

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 $t = 1, \ldots, T$. We are interested in point forecasts $\hat{\boldsymbol{y}}^h_{i,t}$ at multiple horizons, $h = 1, \ldots, H$. The forecasts may come from econometric models, professional forecasters, or any other alternative. Whenever the forecasts are derived from models, the forecasts $\hat{\boldsymbol{y}}^h_{i,t} = \hat{\boldsymbol{y}}^h_{i,t}(\hat{\boldsymbol{\theta}}^h_{i,t})$ 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 and they may be based on nested or non-nested models. 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 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 $i = 1, \ldots, M$, we have forecasts $\hat{\boldsymbol{y}}_{i,t} = [\hat{\boldsymbol{y}}^1_{i,t}, \ldots, \hat{\boldsymbol{y}}^H_{i,t}]'$, where $\hat{\boldsymbol{y}}^h_{i,t}$ is model i's forecast of $\boldsymbol{y}_t$ based on the information set $\mathcal{F}_{t-h}$. 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 an H-dimensional vector, with elements $L^h_{i,t} = L(\boldsymbol{y}_t, \hat{\boldsymbol{y}}^h_{i,t})$.

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 vector, with elements $d^h_{ij,t}$. Our hypotheses are defined in terms of the expected loss differentials, $E(\boldsymbol{d}_{ij,t})$ and as such we focus on the properties of $\boldsymbol{d}_{ij,t}$. In particular, we make the following assumption.

Assumption 1. The vector of loss differences $\boldsymbol{d}_{ij,t}$ is $L_{2+\delta}$ near epoch dependent (NED) on $\{V_t\}$ with NED coefficients v_k of size $-2(r-1)/(r-2)$, where $\{V_t\}$ is α-mixing of size $-(2+\delta)(r+\delta)/(r-2)$, for some $r > 2$ and $0 < \delta \leq 2$, and $var(d^h_{ij,t}) > 0$ for all $h = 1, \ldots, H$.

The assumption allows for considerable heterogeneity in the mean $E(\boldsymbol{d}_{ij,t}) = \boldsymbol{\mu}_{ij,t}$, as well as dependence. However, our object of interest remains $\boldsymbol{\mu}_{ij} \equiv \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{\mu}_{ij,t}$, although conditional tests in the spirit of Giacomini and White (2006) could be developed. We make the following assumption on the amount of time-variation, where $\ell = o(T^{1/2})$ is the block-length parameter of the bootstrap, defined below in Section 2.1.2.

Assumption 2. $\frac{1}{T}\sum_{t=1}^{T}|\mu_{ij,t}^h - \mu_{ij}^h|^{2+\delta} = o(\ell^{-1-\delta/2})$ for some $0 < \delta \leq 2$ and all $h = 1, \ldots, H$.

Assumption 2 limits the potential degree of heterogeneity, but still allows for, for instance, a case with a finite number of properly behaved breaks in the mean. See Gonçalves and White (2002) for details.

The assumptions are 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; Gonçalves and White 2002), such that

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

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

Note that $\boldsymbol{\mu}_{ij,t}$ is implicitly defined as a function of estimated parameters. Indeed, our focus is on finite-sample predictive ability. This contrasts with the population-level framework, first analyzed by West (1996), where the hypotheses are defined in terms of expected loss at the population values of the parameters. Construction of such tests requires a different asymptotic framework, extensively discussed in West (2006).

While the finite-sample predictive ability hypothesis is practically appealing, seeing as we typically only have the estimated parameters, it does come with some restrictions. In particular, the framework permits parameters that are estimated on a (bounded) rolling window, or just once (fixed scheme), but it prohibits the use of forecasts generated by recursive parameter estimates, or (asymptotically) expanding windows. It can however handle both nested and nonnested models, as nonvanishing estimation error prevents the singularity that may occur in nested models when parameters are at their probability limits. See Giacomini and White (2006) for a broad discussion of this framework.

The assumption on $\boldsymbol{d}_{ij,t}$ is sufficient for validity of one of the most common tests for comparing two models' forecasting performance at a single horizon h, the Diebold and Mariano (1995) test. They test the null hypothesis that

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

using a standard t-test:

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

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 hth horizon. In such a setting, taking into account the heterogeneity, 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.

2.1. Multi-Horizon Hypotheses

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 at each horizon, or at the very least, not significantly worse than the other model. Another potential outcome that tells a consistent story, is that one model works well for short horizons, while the other model performs better at longer horizons. 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 nonconsecutive 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.

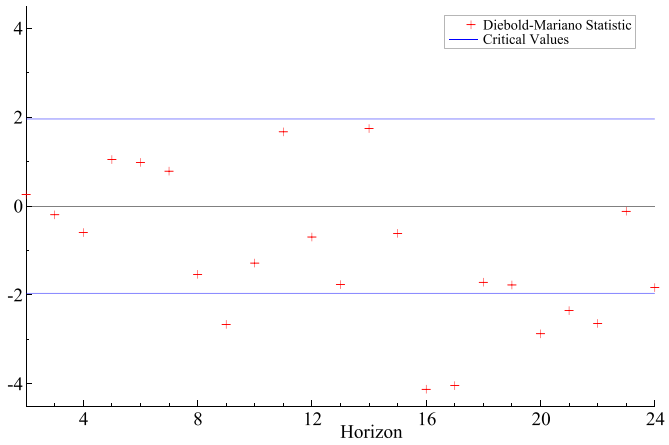

Figure 1. Diebold–Mariano tests at different horizons for earnings of production workers.

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 2–24 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 SPA. The most natural, and strongest, notion is that a superior model should have better forecasts at each individual horizon. To that effect, define

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

We refer to a situation with $\mu_{ij}^{(\text{Unif})} > 0$ 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. A

milder definition of multi-horizon SPA is *average* superior predictive ability (aSPA). Here, we compare models based on their weighted average loss difference

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

with weights $\boldsymbol{w} = [w_1, \ldots, w_H]'$ summing to one. Obvious candidates for $\boldsymbol{w}$ are equal-weighted or weights decaying in horizon. Note that we take the average loss, which is distinct from the loss of the average, which is just one aspect of the forecast 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, aSPA requires the user to take a stand on 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) proposed 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 increase our ability to differentiate models. The variance of loss differences typically increases with horizon, and as such adding an additional horizon may actually decrease power. Moreover, forecasts beyond a certain limiting horizon may become uninformative (Breitung and Knüppel 2017). 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 necessarily 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 $\mu_{ij}^{(\text{Unif})}$. 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

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

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

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

For validity of our procedures $\hat{\omega}_{ij}^h$ can be estimated using any consistent HAC-type estimator. We use the quadratic spectral kernel (Andrews 1991) for reasons elaborated on below, but the more standard Bartlett kernel of Newey and West (1987) is also consistent.

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 $\bar{\boldsymbol{d}}_{ij}$ 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 the statistic will be nonpivotal, as its distribution does depend on the full covariance matrix, which makes $\boldsymbol{\Omega}_{ij}$ 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). 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

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

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

$$t_{\text{aSPA},ij} = \frac{\sqrt{T}\bar{d}_{ij}}{\hat{\zeta}_{ij}}, \tag{10}$$

where $\bar{d}_{ij} = \boldsymbol{w}'\bar{\boldsymbol{d}}_{ij}$. Similar to the uSPA statistic, we avoid estimating the full covariance matrix $\boldsymbol{\Omega}_{ij}$, and choose to estimate $\zeta_{ij} \equiv \sqrt{\boldsymbol{w}'\boldsymbol{\Omega}_{ij}\boldsymbol{w}}$ directly based on $\boldsymbol{w}'\boldsymbol{d}_{ij,t}$ using the HAC estimator.

Throughout the article 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 $\boldsymbol{d}_{ij,t}$, $\boldsymbol{\Omega}_{ij}$. Weighting may be of particular importance in the scenario where one makes aggregate h-period ahead forecasts, that is, $\sum_{h=1}^{H} Y_{t+h}$, which results in clear scale differences that should be inversely weighted.

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 nonstandard critical values. Moreover, critical values obtained via bootstrap techniques may lead to

better finite sample 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% one-sided 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 moving block bootstrap of Künsch (1989) and Liu and Singh (1992). In the moving block bootstrap (MBB), a pseudo time-series of length T is generated by means of randomly drawn blocks of length ℓ from the original data. Assume for simplicity that $T = \ell K$. Let $I_1, \ldots, I_K$ be iid random variables uniformly distributed on $\{1, \ldots, T-\ell+1\}$, and define the array $\tau_t \equiv \{I_1+1, \ldots, I_1+\ell, \ldots, I_K+1, \ldots, I_K+\ell\}$. The pseudo time-series is therefore $\boldsymbol{d}^b_{ij,t} = \boldsymbol{d}_{ij,\tau_t}$, with elements $d^{hb}_{ij,t}$.

By computing either of the test statistics on many MBB resamples, we approximate the distribution of the original statistics under the null. Validity of the bootstrap for studentized statistics requires careful choice of the variance estimators of both the original statistic and the bootstrapped statistics. Regarding the original statistic, for first-order validity, the variance estimator merely needs to be consistent, which is true for most HAC-type estimators. But as Götze and Künsch (1996) noted, for asymptotic refinements the kernel weights need to be chosen more carefully. In particular, triangular weights should be avoided in favor of rectangular or quadratic weights, which motivates our choice of the quadratic spectral kernel.

For the bootstrapped statistics, the appropriate estimator differs from both the HAC-estimator above and the closed-form expression, which is known for the MBB (Künsch 1989). Instead, Götze and Künsch (1996) and Gonçalves and White (2004) demonstrate the validity of the block bootstrap for studentized statistics using the "natural" estimator, which uses the fact that each block's means are conditionally iid

$$(\hat{\omega}^{hb}_{ij})^2 \equiv \frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{\ell}\left(\sum_{t=1}^{\ell} d^{hb}_{ij,(k-1)\ell+t} - \bar{d}^{hb}_{ij}\right)^2\right], \tag{11}$$

where $\bar{d}^{hb}_{ij} = \frac{1}{T}\sum_{t=1}^{T} d^{hb}_{ij,t}$.

Based on the above, we summarize how to obtain the critical values of the test for uSPA and aSPA under the null:

Algorithm 1 (Multi-horizon SPA bootstrap).
For $b = 1, \ldots, B$:

1. Resample $\boldsymbol{d}_{ij,t}$ using a MBB with block length ℓ, to obtain $\boldsymbol{d}^b_{ij,t}$, with elements $d^{hb}_{ij,t}$.
2. uSPA: Compute $\bar{d}^{hb}_{ij} = \frac{1}{T}\sum_{t=1}^{T} d^{hb}_{ij,t}$ for each h.
   Compute $\hat{\omega}^{hb}_{ij}$ using (11) applied to $d^{hb}_{ij,t}$ for each h.
   Compute the uSPA statistic:
   $t^b_{\text{uSPA},ij} = \min_h[\sqrt{T}(\bar{d}^{hb}_{ij} - \bar{d}^{h}_{ij})/\hat{\omega}^{hb}_{ij}]$
   aSPA: Compute $\bar{d}^b_{ij} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{w}'\boldsymbol{d}^b_{ij,t}$.
   Compute $\hat{\zeta}^b_{ij}$ using (11) applied to $\boldsymbol{w}'\boldsymbol{d}^b_{ij,t}$.
   Compute the aSPA statistic:
   $t^b_{\text{aSPA},ij} = \sqrt{T}(\bar{d}^b_{ij} - \bar{d}_{ij})/\hat{\zeta}^b_{ij}$.

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

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 $\boldsymbol{D}_{ij} \equiv \text{diag}(\omega^1_{ij}, \ldots, \omega^H_{ij})$ and $\hat{\boldsymbol{D}}_{ij}, \hat{\boldsymbol{D}}^b_{ij}$ analogously defined using $\hat{\omega}^h_{ij}$ and $\hat{\omega}^{hb}_{ij}$. Let Assumption 1 hold, and moreover, assume that $\ell_T \equiv \ell$, $\ell_T \to \infty$ and $\ell_T = o(T^{1/2})$, then

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

where P^b denotes the bootstrap probability measure.

The proof is provided in Appendix A and mostly follows from the results of Gonçalves and White (2004), who prove validity of the MBB for Wald statistics under similar assumptions. From Theorem 1 we obtain the following corollary.

Corollary 1. Let the assumptions from Theorem 1 hold. Then,

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

and

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

The corollary demonstrates that the bootstrap may be used to obtain the critical values for both the uSPA and aSPA, test statistics. It 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 MCS approach of Hansen, Lunde, and Nason (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 multi-horizon MCS, analogous to Hansen et al. (2011), we define the MCS as the subset of models for which we find no statistical support to differentiate them

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

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

The associated null hypotheses are

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

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

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

The multi-horizon MCS, based on either uSPA or aSPA, 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) proposed the maximum of all pairwise $t_{\text{DM},ij}$ statistics to test for equivalence, but since the critical value of the $t_{\bullet\text{SPA},ij}$ 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 $\max_{i,j\in\mathcal{M}}[t_{\bullet\text{SPA},ij} - c^{\alpha}_{\bullet\text{SPA},ij}]$. To obtain the distribution of this maximum statistic, we require the use of a double bootstrap. The computational cost is therefore relatively high, but the multi-horizon MCS remains feasible as it merely involves bootstrapping studentized means, without re-estimation of models.

Algorithm 2 (Multi-horizon MCS bootstrap).

1. For each pair $\{i,j\} \in \mathcal{M}$, compute the statistic $t_{\bullet\text{SPA},ij}$. Apply Algorithm 1, with a common set of indices, τ_t, for all pairs, to obtain estimates of the associated critical values $c^{\alpha}_{\bullet\text{SPA},ij}$.
2. Define $t_{\text{Max},\bullet\text{SPA}} \equiv \max_{i,j\in\mathcal{M}}\left[t_{\bullet\text{SPA},ij} - c^{\alpha}_{\bullet\text{SPA},ij}\right]$, that is, the test statistic furthest from its critical value.
3. For each of the bootstrap samples $\boldsymbol{d}^b_{ij,t}$, $b = 1,\ldots,B$, obtained 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 $c^{\alpha b}_{\bullet\text{SPA},ij}$.
   (b) Compute the bootstrapped $t^b_{\text{Max},\bullet\text{SPA}} \equiv \max_{i,j\in\mathcal{M}}\left[t^b_{\bullet\text{SPA},ij} - c^{\alpha b}_{\bullet\text{SPA},ij}\right]$.
4. Obtain the appropriate critical value as the $\tilde{\alpha}$-quantile of the bootstrap distribution $t^b_{\text{Max},\bullet\text{SPA}}$, or define the p-value as $p \equiv \frac{1}{B}\sum_{b=1}^{B}\mathbf{I}_{\{t_{\text{Max},\bullet\text{SPA}} < t^b_{\text{Max},\bullet\text{SPA}}\}}$.

The combination of equivalence test and elimination rule adhere to the definition of coherency of Hansen et al. (2011). Algorithm 2 is a standard application of the double bootstrap, and therefore we conjecture validity follows by extension of Theorem 1 and validity of the bootstrap in the original MCS of Hansen et al. (2011, Appendix 1.1).

To obtain reasonable p-values we follow Hansen et al. (2011) in imposing that a p-value for a model cannot 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 MCS. Throughout the remainder of the paper, we set the block length to $\ell = 3$, and we use $B = 999$ bootstrap resamples. All results reported in this paper are based on programs written in Ox version 7.0 (Doornik 2012). Ox and Matlab code detailing the implementation of the various tests, simulations and empirical results, is available on Quaedvlieg's website.

3.1. Data-Generating Process

First, we describe how we generate "losses" of a given model i. Our design closely resembles that of the simulations in 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 this dependence to that of the loss differential between an AR(1) and AR(2) when the true model is the latter.

We consider simulation set-ups with two and ten models. For the 10-model setup, the average loss of each model is parameterized by an 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 $i = 1,\ldots,10$, 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 $\boldsymbol{\theta} = [\theta^1, \ldots, \theta^h]'$, 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 to highlight the tests for uSPA and aSPA. First, we set

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

The loss differential is nonnegative 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

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

with $c = 1+2/\sum_{h=2}^{H}(1+\phi\sqrt{h-1})$, such that $\sum_{h=1}^{H}\theta^{h(\text{NonUnif})} = \sum_{h=1}^{H}\theta^{h(\text{Unif})}$. We impose nonuniformity 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 uSPA at any horizon.

We generate the losses as follows

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

where $\boldsymbol{\epsilon}_t \sim \text{iid}N(0, \boldsymbol{I})$ and $\circ$ denotes the Hadamard product. The losses are serially correlated through $\boldsymbol{\varrho}$ 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 $\text{cov}(L^h_{i,t}, L^g_{j,t}) = 0$, 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 $S = 10{,}000$ simulations, and vary the parameters of the DGP. We take three sample sizes $T = 250, 500, 1000$. 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 at the 5% level, that is, we test whether model 1 outperforms model 2 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 $\boldsymbol{\theta}^{(\text{Unif})}$ and $\boldsymbol{\theta}^{(\text{NonUnif})}$, referred to as uniform and nonuniform alternative, displayed in the top and bottom panel, respectively.

First consider the top panel, which is based on $\boldsymbol{\theta}^{(\text{Unif})}$. 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%, irrespective of horizon. When $\lambda > 0$, the loss differential at each horizon is positive. For 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 all 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 upward, all tests are different. The tests for uSPA and aSPA use the loss-differentials of all horizons, which results in increasing rejection frequencies in H. In line

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											
	0	0.054	0.052	0.053	0.054	0.053	0.054	0.054	0.054	0.053	0.050	0.051	0.052
	5	0.118	0.199	0.185	0.164	0.115	0.198	0.221	0.239	0.113	0.217	0.251	0.258
250	10	0.203	0.478	0.440	0.356	0.199	0.425	0.504	0.548	0.199	0.526	0.612	0.616
	20	0.478	0.933	0.900	0.804	0.468	0.808	0.890	0.929	0.467	0.961	0.984	0.988
	40	0.934	1.000	1.000	1.000	0.930	0.994	0.998	0.999	0.929	1.000	1.000	1.000
	0	0.051	0.051	0.053	0.051	0.049	0.052	0.052	0.052	0.050	0.051	0.050	0.052
	5	0.110	0.203	0.180	0.162	0.109	0.191	0.221	0.241	0.109	0.224	0.254	0.259
500	10	0.197	0.486	0.441	0.355	0.195	0.422	0.500	0.546	0.195	0.535	0.617	0.619
	20	0.478	0.941	0.909	0.804	0.476	0.814	0.891	0.933	0.474	0.966	0.989	0.989
	40	0.936	1.000	1.000	1.000	0.934	0.994	0.997	0.999	0.933	1.000	1.000	1.000
	0	0.052	0.056	0.049	0.051	0.051	0.055	0.051	0.054	0.051	0.053	0.054	0.054
	5	0.105	0.189	0.189	0.157	0.104	0.191	0.215	0.227	0.106	0.217	0.250	0.253
1000	10	0.195	0.467	0.444	0.347	0.196	0.421	0.499	0.543	0.197	0.532	0.613	0.618
	20	0.471	0.932	0.906	0.802	0.469	0.808	0.893	0.931	0.468	0.966	0.987	0.988
	40	0.933	1.000	1.000	0.999	0.934	0.994	0.998	1.000	0.933	1.000	1.000	1.000
		Nonuniform alternative											
	0	0.056	0.056	0.054	0.055	0.056	0.054	0.058	0.055	0.055	0.055	0.053	0.056
	5	0.023	0.213	0.199	0.167	0.022	0.099	0.134	0.166	0.022	0.183	0.243	0.254
250	10	0.009	0.500	0.463	0.363	0.009	0.073	0.133	0.200	0.009	0.426	0.581	0.616
	20	0.001	0.943	0.918	0.819	0.001	0.013	0.031	0.064	0.001	0.890	0.978	0.985
	40	0.000	1.000	1.000	1.000	0.000	0.000	0.000	0.001	0.000	1.000	1.000	1.000
	0	0.053	0.052	0.053	0.052	0.052	0.052	0.054	0.053	0.053	0.052	0.052	0.052
	5	0.022	0.210	0.204	0.164	0.022	0.099	0.138	0.166	0.022	0.183	0.242	0.254
500	10	0.009	0.487	0.450	0.366	0.010	0.069	0.127	0.183	0.009	0.419	0.577	0.614
	20	0.001	0.947	0.918	0.828	0.001	0.013	0.030	0.066	0.001	0.901	0.981	0.989
	40	0.000	1.000	1.000	1.000	0.000	0.000	0.000	0.001	0.000	1.000	1.000	1.000
	0	0.049	0.051	0.053	0.053	0.048	0.053	0.054	0.056	0.050	0.051	0.052	0.054
	5	0.022	0.211	0.198	0.159	0.021	0.095	0.134	0.159	0.021	0.184	0.242	0.253
1000	10	0.008	0.494	0.459	0.354	0.008	0.071	0.130	0.193	0.008	0.429	0.585	0.618
	20	0.001	0.942	0.922	0.820	0.001	0.015	0.036	0.071	0.001	0.895	0.981	0.988
	40	0.000	1.000	1.000	1.000	0.000	0.000	0.001	0.001	0.000	1.000	1.000	1.000

NOTES: This table provides rejection frequencies over $S = 10{,}000$ 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 $\boldsymbol{\theta}^{(\text{Unif})}$, while the nonuniform alternative panel results are generated using $\boldsymbol{\theta}^{(\text{NonUnif})}$.

with the results from the DM test, the largest increase in power is between $H = 1$ and $H = 5$.

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

For the Diebold–Mariano test, when $h = 1$, 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 $\boldsymbol{\theta}^{(\text{NonUnif})}$ is chosen such that over the 20 horizons, the average $\boldsymbol{\theta}^{(\text{NonUnif})}$ is equal to $\boldsymbol{\theta}^{(\text{Unif})}$. As a result, compared to the top panel, for $h > 1$ we see that the univariate tests typically have higher power in the bottom panel, as the loss differential is slightly larger to compensate for 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 $H = 10$ it has slightly lower power than under the uniform alternative, as indeed the average loss differential is only equal at $H = 20$, at which point they coincide.

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.

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	Uniform alternative											
$\phi = 0$	0	0.054	0.054	0.053	0.052	0.056	0.053	0.055	0.054	0.047	0.048	0.049	0.049
	5	0.109	0.123	0.128	0.138	0.103	0.103	0.092	0.086	0.107	0.101	0.085	0.082
	10	0.200	0.238	0.252	0.274	0.197	0.183	0.162	0.149	0.198	0.156	0.132	0.114
	20	0.470	0.567	0.616	0.671	0.476	0.425	0.355	0.300	0.473	0.336	0.256	0.198
	40	0.930	0.971	0.982	0.991	0.926	0.869	0.756	0.618	0.932	0.738	0.551	0.400
$\phi = 1$	0	0.49	0.052	0.052	0.052	0.053	0.050	0.052	0.052	0.053	0.053	0.052	0.049
	5	0.101	0.233	0.338	0.475	0.109	0.191	0.221	0.241	0.107	0.164	0.166	0.155
	10	0.199	0.501	0.655	0.780	0.195	0.422	0.500	0.546	0.204	0.359	0.374	0.359
	20	0.472	0.820	0.900	0.950	0.476	0.814	0.891	0.933	0.473	0.772	0.805	0.774
	40	0.932	0.994	0.998	0.999	0.934	0.994	0.997	0.999	0.926	0.992	0.996	0.996
$\phi = 2$	0	0.048	0.051	0.050	0.050	0.050	0.050	0.052	0.057	0.050	0.049	0.049	0.047
	5	0.103	0.330	0.480	0.627	0.105	0.274	0.355	0.413	0.109	0.231	0.261	0.273
	10	0.201	0.539	0.689	0.808	0.195	0.537	0.677	0.775	0.203	0.505	0.592	0.624
	20	0.464	0.816	0.903	0.951	0.471	0.817	0.904	0.953	0.481	0.815	0.899	0.941
	40	0.928	0.995	0.998	0.999	0.929	0.993	0.997	0.999	0.936	0.994	0.998	0.999

NOTES: This table provides rejection frequencies for the test for uniform superior predictive ability over $S = 10{,}000$ simulations according to the DGP outlined in Section 3.1. The losses are generated according to $\boldsymbol{\theta}^{(\text{Unif})}$, and the sample size $T = 500$ for all results.

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 uSPA, under the uniform alternative, while 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, power is slightly increasing in h, which is due to the fact that our sample size increases. 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 than 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.

3.3. Model Confidence Sets

In this section, we evaluate the ability of the multi-horizon MCS to distinguish between models. We base our conclusions on the ten-model scenario. We use $\boldsymbol{\theta}^{(\text{Unif})}$ to generate the loss differentials. Recall that 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 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. However, the table shows that potency is always close to the expected 80% for all T and H, which means that for around 80% of our simulations, not a single model was removed from the set. When $\lambda > 0$, there is a single superior model,

Table 3. Multivariate simulation results: potency and gauge

		Potency				Gauge			
	H	1	5	10	20	1	5	10	20
T	λ								
	0	0.787	0.793	0.797	0.807				
	5	0.966	0.939	0.922	0.916	4.481	1.941	1.448	1.166
250	10	0.937	0.970	0.979	0.973	1.554	0.379	0.214	0.120
	20	0.950	0.998	1.000	1.000	0.165	0.012	0.002	0.000
	40	1.000	1.000	1.000	1.000	0.000	0.000	0.000	0.000
	0	0.781	0.790	0.792	0.797				
	5	0.960	0.946	0.929	0.921	4.441	2.038	1.494	1.238
500	10	0.934	0.959	0.967	0.981	1.533	0.412	0.208	0.080
	20	0.953	1.000	1.000	1.000	0.183	0.008	0.002	0.000
	40	0.998	1.000	1.000	1.000	0.002	0.000	0.000	0.000
	0	0.787	0.799	0.806	0.802				
	5	0.953	0.938	0.926	0.923	4.346	1.952	1.468	1.166
1000	10	0.909	0.955	0.975	0.979	1.472	0.447	0.195	0.114
	20	0.960	0.998	0.999	1.000	0.166	0.010	0.003	0.000
	40	0.999	1.000	1.000	1.000	0.001	0.000	0.000	0.000

NOTES: This table provides the potency and gauge of the multi-horizon MCS over $S = 10{,}000$ 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 $\boldsymbol{\theta}^{(\text{Unif})}$.

which is easier to select, and potency is close to 100% for all combinations of 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 observations 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 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, Stock, and Watson (2006), who investigated the performance of iterated versus direct forecasts using 170 monthly U.S. macroeconomic time series spanning 1959–2002. 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. Therefore, we test for multi-horizon superior predictive ability across horizons $h = 2, \ldots, 24$ using the two tests developed in this paper. 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 consist of 170 series divided up into five different categories. We apply their suggested data transformation to deal with the nonstationary 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, Stock, and Watson (2006).

We mostly follow the forecasting methodology of Marcellino, Stock, and Watson (2006), with one exception; our parameter estimates are based on a rolling window of 120 observations, rather than an expanding window, which is required for validity of our tests. 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 between 1 and 12, based on either AIC or BIC using the estimation sample. Note that it is entirely possible that in 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.

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

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

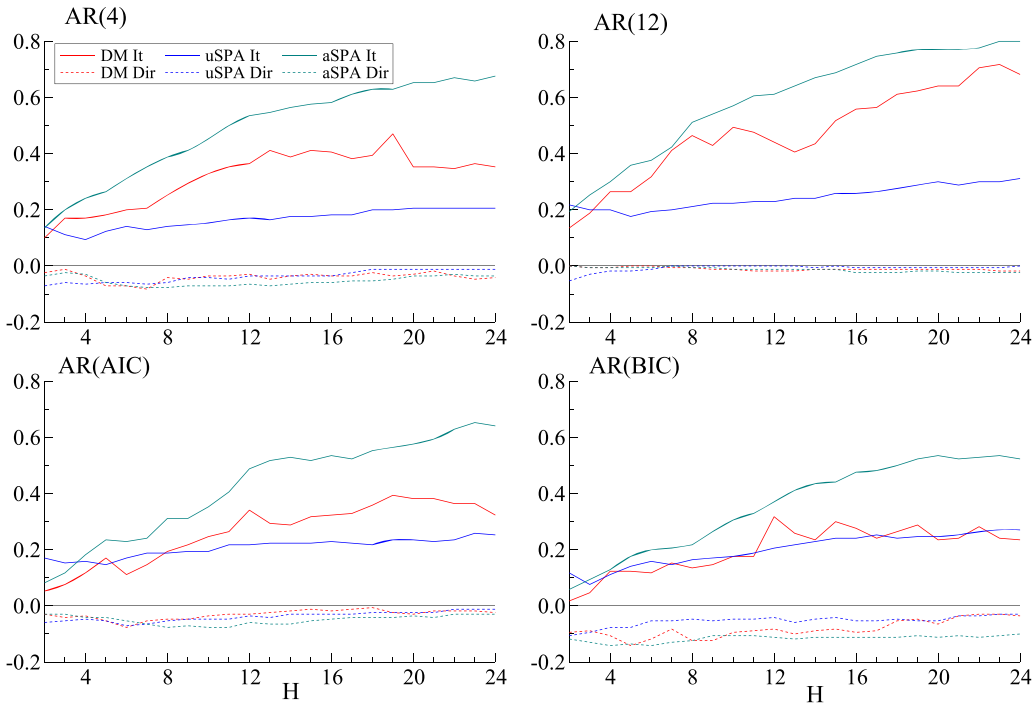

Figure 2. Rejection frequencies equal forecasting performance across horizons.

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, that is, 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

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

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

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

4.1. Aggregate Results

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

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 results are mostly in line white those of Marcellino, Stock, and Watson (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. Especially for the single-horizon and uSPA tests, the rejection frequencies in favor of direct forecasts decrease 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 30% for the two-year ahead forecast. Based on the AR(12) model, the number of rejections is significantly higher at about 60%. Importantly, the number of rejections is unstable across horizons. For instance, based on AR(4), looking at just horizon $h = 19$ we would reject for almost 50% of the series, while for horizon $h = 20$ the percentage would be closer to 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 nearly monotonically 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.

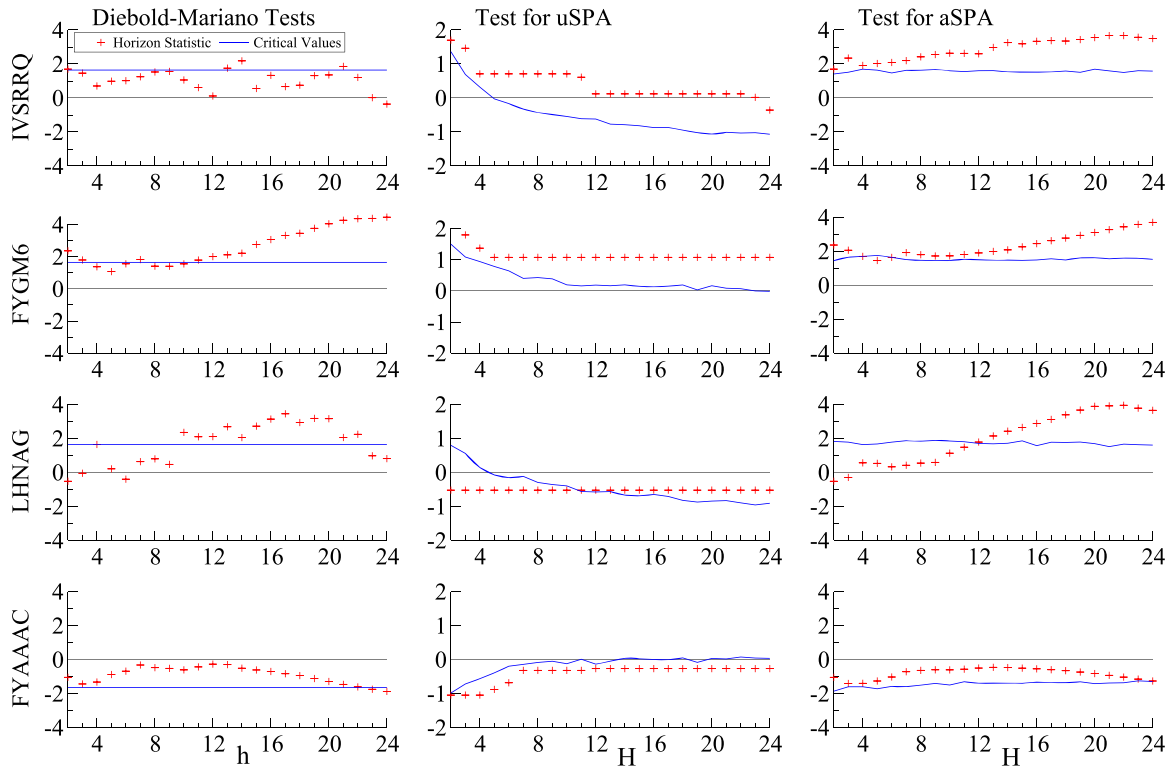

Figure 3. Results for individual series.

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 rates of the test for aSPA are indeed higher than those for uSPA, but also consistently higher than those for 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 across the 23 horizons, iterated forecasts provide average superior predictive ability relative to direct forecasts for between 50% and 70% 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. Results for Individual Series

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 crosses denote the test-statistics at, or up to, horizon h. The lines provide the one-sided critical value at 5%. For the DM-test this is based on the Gaussian quantiles, while for the multi-horizon tests we report $c^{5\%}_{\bullet\text{SPA},ij}$ based on 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, except $h = 24$. 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 reasonably expect to observe a negative differential, even if the true loss differential μ^h_{ij} is positive for all h. As a result, we conclude that iterated forecasts provide both uSPA and aSPA, despite only finding significant evidence of superior predictive ability at 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 $H = 24$.

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 thirteen consecutive horizons starting at $h = 10$. The test for aSPA combines the joint evidence and rejects the null from $H \geq 12$. 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 $H = 11$ 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, but $h = 24$. However, its consistently negative values results in the fact that the uniform statistic does reject at all horizons $H \geq 3$. Hence, we find evidence for uSPA, but not for aSPA until we consider all 24 horizons. 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. Across the 170 series we perform both these tests, we only find evidence for uSPA and not for aSPA a negligible three 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 coherent 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 that this has three advantages. First, multi-horizon superior predictive ability provides a more complete definition of a model's superior performance. Second, by using multiple horizons we can construct a powerful test, 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 (weighted) 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 misspecification 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, to compare a larger number of models directly, we extend the MCS methodology to allow for multi-horizon comparison. 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 MCS are shown to be properly sized and powerful in simulations.

The pairwise comparison is illustrated by means of a comparison between direct and iterated forecasts of macroeconomic variables, based on the data in Marcellino, Stock, and Watson (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 incoherence is typically the result of the implicit multiple-testing issue of picking and choosing a few horizons.

APPENDIX A: BOOTSTRAP VALIDITY

Proof Theorem 1. Under either null hypothesis $\sqrt{T}\boldsymbol{D}_{ij}^{-1}\bar{\boldsymbol{d}}_{ij} \to_d N(0, \boldsymbol{R}_{ij})$, where $\boldsymbol{R}_{ij} = \boldsymbol{D}_{ij}^{-1}\boldsymbol{\Omega}_{ij}\boldsymbol{D}_{ij}^{-1}$, and $\to_d$ denotes convergence in distribution. By standard arguments, the quadratic spectral HAC estimator (Andrews 1991) is consistent for $\boldsymbol{D}_{ij}$ and therefore, $\sqrt{T}\hat{\boldsymbol{D}}_{ij}^{-1}\bar{\boldsymbol{d}}_{ij} \to_d N(0, \boldsymbol{R}_{ij})$.

Next, we show that the bootstrap consistently estimates the distribution of $\sqrt{T}\hat{\boldsymbol{D}}_{ij}^{-1}\bar{\boldsymbol{d}}_{ij}$. Under the stated assumptions, it follows from Theorem 2.2 of Gonçalves and White (2002) that

$$\sup_{x\in\mathbb{R}^H} |P^b[\sqrt{T}(\bar{\boldsymbol{d}}_{ij}^b - \bar{\boldsymbol{d}}_{ij}) \leq x] - P[\sqrt{T}(\bar{\boldsymbol{d}}_{ij} - \boldsymbol{\mu}_{ij}) \leq x]| \to_p 0, \quad \text{(A.1)}$$

where P^b denotes the bootstrap distribution. While this demonstrates that the bootstrap distribution can be used to approximate the distribution of $\sqrt{T}(\bar{\boldsymbol{d}}_{ij} - \boldsymbol{\mu}_{ij})$, it does not immediately justify the validity of the bootstrap for the studentized statistics, just valid bootstrap confidence intervals. Theorem 3.1 of Gonçalves and White (2004) applied to studentized statistics shows that under the null, the studentized statistic is approximated by the bootstrap

$$\sup_{x\in\mathbb{R}^H} |P^b[\sqrt{T}(\hat{\boldsymbol{D}}_{ij}^b)^{-1}(\bar{\boldsymbol{d}}_{ij}^b - \bar{\boldsymbol{d}}_{ij}) \leq x] - P[\sqrt{T}\hat{\boldsymbol{D}}_{ij}^{-1}(\bar{\boldsymbol{d}}_{ij} - \boldsymbol{\mu}_{ij}) \leq x]| \to_p 0, \quad \text{(A.2)}$$

provided that for any $\epsilon > 0$, $P^b[|\hat{\boldsymbol{D}}_{ij}^b - \boldsymbol{D}_{ij}| > \epsilon] \to_p 0$. This condition is established for the estimator in Equation (11) under Assumptions 1 and 2 of this article for the MBB in Lemma B.1 of Gonçalves and White (2004). □

ACKNOWLEDGMENTS

The author would like to thank the editor, Todd Clark, the associate editor, and three anonymous referees, as well as Sébastien Laurent, Andrew Patton, Alessandro Pollastri, Stephan Smeekes, as well as seminar and conference participants at Duke University, UNC Chapel Hill and QFFE 2018, for helpful comments and suggestions.

[Received July 2017. Revised December 2018.]

REFERENCES

Andrews, D. W. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 3, 817–858. [43,52]

Bartholomew, D. (1961), "A Test of Homogeneity of Means Under Restricted Alternatives," *Journal of the Royal Statistical Society*, Series B, 23, 239–281. [43]

Breitung, J., and Knüppel, M. (2017), "How far can We Forecast? Statistical Tests of the Predictive Content," Working Paper. [43]

Capistrán, C. (2006), "On Comparing Multi-Horizon Forecasts," *Economics Letters*, 93, 176–181. [41]
Clark, T. E., and McCracken, M. W. (2005), "Evaluating Direct Multistep Forecasts," *Econometric Reviews*, 24, 369–404. [41,43]
——— (2012), "Reality Checks and Comparisons of Nested Predictive Models," *Journal of Business & Economic Statistics*, 30, 53–66. [41]
——— (2013), "Advances in Forecast Evaluation," in *Handbook of Economic Forecasting* (Vol. 2), eds. G. Elliott and A. Timmermann, Amsterdam: North Holland, pp. 1107–1201. [41]
Clark, T. E., and West, K. D. (2007), "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models," *Journal of Econometrics*, 138, 291–311. [41]
Clements, M. P., and Hendry, D. F. (1993), "On the Limitations of Comparing Mean Square Forecast Errors," *Journal of Forecasting*, 12, 617–637. [41]
De Jong, R. M. (1997), "Central Limit Theorems for Dependent Heterogeneous Random Variables," *Econometric Theory*, 13, 353–367. [42]
Diebold, F. X., and Mariano, R. S. (1995), "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 134–144. [40,41,42]
Doornik, J. A. (2012), *An Object-Oriented Matrix Programming Language Ox 7*, Timberlake Consultants Ltd. [45]
Giacomini, R., and White, H. (2006), "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545–1578. [41,42]
Gonçalves, S., and White, H. (2002), "The Bootstrap of the Mean for Dependent Heterogeneous Arrays," *Econometric Theory*, 18, 1367–1384. [42,52]
——— (2004), "Maximum Likelihood and the Bootstrap for Nonlinear Dynamic Models," *Journal of Econometrics*, 119, 199–219. [44,52]
Götze, F., and Künsch, H. R. (1996), "Second-Order Correctness of the Blockwise Bootstrap for Stationary Observations," *The Annals of Statistics*, 24, 1914–1933. [44]
Hansen, P. R. (2005), "A Test for Superior Predictive Ability," *Journal of Business & Economic Statistics*, 23, 365–380. [40,43]
Hansen, P. R., Lunde, A., and Nason, J. M. (2011), "The Model Confidence Set," *Econometrica*, 79, 453–497. [40,42,43,45]
Hendry, D. F., and Doornik, J. A. (2014), *Empirical Model Discovery and Theory Evaluation: Automatic Selection Methods in Econometrics*, Cambridge, MA: MIT Press. [48]
Jordà, O., and Marcellino, M. (2010), "Path Forecast Evaluation," *Journal of Applied Econometrics*, 25, 635–662. [41]
Komunjer, I., and Owyang, M. T. (2012), "Multivariate Forecast Evaluation and Rationality Testing," *Review of Economics and Statistics*, 94, 1066–1080. [41]
Künsch, H. R. (1989), "The Jackknife and the Bootstrap for General Stationary Observations," *The Annals of Statistics*, 17, 1217–1241. [44]
Linton, O., Maasoumi, E., and Whang, Y.-J. (2005), "Consistent Testing for Stochastic Dominance Under General Sampling Schemes," *The Review of Economic Studies*, 72, 735–765. [43]
Linton, O., Song, K., and Whang, Y.-J. (2010), "An Improved Bootstrap Test of Stochastic Dominance," *Journal of Econometrics*, 154, 186–202. [43]
Liu, R. Y., and Singh, K. (1992), "Moving Blocks Jackknife and Bootstrap Capture Weak Dependence," in *Exploring the Limits of Bootstrap* (Vol. 225), eds. R. Lepage and L. Billiard, New York: Wiley, p. 248. [44]
Marcellino, M., Stock, J. H., and Watson, M. W. (2006), "A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series," *Journal of Econometrics*, 135, 499–526. [41,49,50,52]
Martinez, A. (2017), "Testing for Differences in Path Forecast Accuracy: Forecast-Error Dynamics Matter," Working Paper. [41]
Newey, W. K., and West, K. D. (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708. [43]
Patton, A. J., and 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, 605–625. [43]
——— (2012), "Forecast Rationality Tests Based on Multi-Horizon Bounds," *Journal of Business & Economic Statistics*, 30, 1–17. [41]
West, K. D. (1996), "Asymptotic Inference About Predictive Ability," *Econometrica*, 64, 1067–1084. [41,42]
——— (2006), "Forecast Evaluation," in *Handbook of Economic Forecasting* (Vol. 1), Amsterdam: Elsevier, pp. 99–134. [42]
White, H. (2000), "A Reality Check for Data Snooping," *Econometrica*, 68, 1097–1126. [40,43,44]
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, 782–793. [43]