Comparing Forecasting Performance with Panel Data

Ritong Qu[*] Allan Timmermann[†] Yinchu Zhu[‡]

August 27, 2023

Abstract

We develop new methods for testing equal predictive accuracy for panels of forecasts, exploiting information in both the time series and cross-sectional dimensions of the data. We examine general tests of equal forecasting performance averaged across all time periods and individual units along with tests that focus on subsets (clusters) of time or units. Our tests are demonstrated in an empirical application that compares IMF forecasts of country-level real GDP growth and inflation to private-sector survey forecasts and forecasts from a simple time-series model.

Key words: Economic forecasting; Panel Data; Panel Diebold-Mariano Test

[*] International Monetary Fund, 709 19th Street NW, Washington, D.C. 20431, U.S.A.; rqu@imf.org. The views in this paper represent only the authors' and should therefore not be reported as representing the views of the International Monetary Fund, its Executive Board, or IMF management.

[†] Rady School of Management, University of California, San Diego, 9500 Gilman Dr, La Jolla, CA 92093, U.S.A.; atimmermann@ucsd.edu

[‡] ‡Department of Economics, Brandeis University, 415 South Street, Waltham, MA 02453, U.S.A.; yinchuzhu@brandeis.edu

1 Introduction

Forecasts of economic and financial variables are increasingly recorded not only for a single outcome variable but across multiple variables and at many points in time, giving rise to panels of forecasts. A prominent example is survey data with individual survey respondents or organizations reporting forecasts for multiple countries, industries or macroeconomic variables over extensive periods of time. Panels are also common in comparisons of the forecasting performance of alternative prediction models fitted to different variables.[1]

The presence of both a cross-sectional and a time-series dimension in panel data creates unique opportunities for testing economic hypotheses and comparing the predictive accuracy of different forecasts. At the most aggregate level, one can test whether two sets of forecasts have the same predictive accuracy "on average", i.e., when averaged both cross-sectionally and over time. This hypothesis does not rule out that one forecast dominates another in expectation for *some* time periods or for *some* units. Rather, it states that such differences in forecasting performance average out across time and units. Tests of this hypothesis may therefore overlook differences that arise only during some periods or affect only certain units.

To address this point, one can instead compare two forecasts' accuracy either by averaging along the time-series dimension (e.g., years) for individual variables or groups of variables or, alternatively, by averaging along the cross-sectional dimension for a cluster of units, in both cases testing whether a pair of forecasts are equally accurate (in expectation) within each cluster.[2] Tests of the resulting hypotheses can yield important insights into the economic sources of rejections of equal predictive accuracy. For example, a test that exploits cross-sectional information but uses only a short time-series record might find that model-based forecasts are inferior to survey forecasts only during financial crises or economic recessions, while the two sets of forecasts are equally accurate during more normal times. This might indicate that the model-based forecasts adapt too slowly to sharp shifts in the underlying state of the economy, while conversely survey participants can exploit forward-looking information to improve their forecasts during such periods. Alternatively, one could use a longer time-series record to separately compare the predictive accuracy of two competing

[1] Baltagi (2013) provides an extensive review of forecast applications that use panel data.

[2] In fact, Qu et al. (2021) establish conditions under which comparisons of predictive accuracy can be conducted on a single cross-section of data.

approaches for predicting company earnings, clustering forecasts and outcomes into pre-specified groups defined by industry, region, or country to gain insights into the forecasts' relative accuracy for different types of firms.

Considerations such as these lead us to study tests of equal predictive accuracy that average over time for pre-specified cross-sectional groups of variables or pre-specified blocks of time. These tests use the results of Ibragimov and Müller (2010, 2016) and so require normality assumptions for the average loss differential computed for the individual clusters of forecast errors along with independence across clusters. Such assumptions generally require invoking a Central Limit Theorem (CLT) for the clusters and so restrict the kind of dependencies across forecast errors that can be accommodated. Alternatively, one can use use the randomization test recently proposed by Canay et al. (2017). We consider both types of tests and establish conditions under which their usage is valid with panel data. Moreover, to address situations in which the assumption of block diagonality in the forecast errors from different clusters fails to hold, we develop a simple approach to correct for such correlations.

To provide practical guidance on which approach to testing the null of equal predictive accuracy works best, we undertake an extensive set of Monte Carlo simulations that allow for serial correlation in forecast errors as well as spatial dependencies and factors both within and across clusters of forecast errors. In these simulations, we also compare our test statistics to a variety of tests recently proposed by Akgun et al. (2022). We find that it is important to account for these features as many tests display important size distortions and tend to overreject–sometimes by a very large amount–in the presence of temporal or cross-sectional dependencies in forecast errors.

We further illustrate the new tests in an empirical application to the International Monetary Fund's (IMF) World Economic Outlook (WEO) forecasts of annual real GDP growth and inflation for a large cross-section of countries covering 30 annual observations and four forecast horizons over the period from 1990 to 2019. We compare these forecasts to private-sector survey forecasts reported by the Consensus Economics organization in addition to forecasts generated by a simple autoregressive time-series model.

Empirically, for GDP growth forecasts, we mostly find that we cannot reject the null that the IMF and Consensus Economics forecasts are equally accurate, except during the peak of the Global Financial Crisis (2008) at which point the IMF forecasts became relatively more accurate.

Conversely, we find that the IMF current-year inflation forecasts are significantly more accurate than their Consensus Economics counterparts, although they seem to be equally accurate at the one-year horizon. This finding can be attributed mostly to the accuracy of the IMF inflation forecasts for non-advanced economies along with relatively accurate forecasts for advanced economies during the global financial crisis.

Looking at the term structure of squared forecast errors across different forecast horizons, our tests allow us to identify the horizons at which the IMF forecasts gain in precision. We find that the accuracy of the IMF's GDP growth forecasts only begins to improve in the fall of the previous year. This suggests that information that facilitates more accurate forecasts of GDP growth tends to be quite short-lived and that improvements in real GDP growth forecasts more than 15 months out from the target date are relatively minor. Conversely, inflation forecasts tend to improve both at longer and shorter forecast horizons.

A related literature has focused on evaluating the efficiency of forecasts with panel data; see, e.g., Keane and Runkle (1990), Davies and Lahiri (1995), and Patton and Timmermann (2012). However, this literature does not provide methods for systematically comparing the relative accuracy of different forecasts or for conducting tests of the null of equal predictive accuracy across different forecasts.

In work that is complementary to ours, Akgun et al. (2022) provide a comprehensive analysis of the properties of tests of equal predictive accuracy with panel data under assumptions of a strong factor structure in loss differentials. While these authors focus on exploiting such linear factor structures in loss differentials, we propose inference methods under generic forms of cross-sectional dependencies. For example, we do not impose strong factor conditions or assume that the factors and idiosyncratic error terms follow strictly stationary linear processes. Another advantage is that our proposal can be applied in settings in which the factor structure is not directly imposed on the loss differential; in fact, we allow for cross-sectional dependence structures that are more general than those associated with linear factor models. Finally, we also do not require hac standard errors to compute our test statistics, a point that again differentiates our analysis from that in Akgun et al. (2022).

The outline of the paper is as follows. Section 2 introduces tests of equal predictive accuracy for panels of forecasts conducted on the pooled average (pooling both cross-sectionally and across time) or pooled separately across time clusters or cross-sectional clusters. Section 3 describes approaches for testing equal predictive accuracy within clusters of time or for groups of variables. Using these tests, Section 4 conducts an empirical analysis that compares the IMF forecasts of GDP growth and inflation to the equivalent Consensus Economics and autoregressive forecasts. Section 5 concludes.

2 Panel Tests of Equal Predictive Accuracy

Consider a panel of data with y_{it} denoting the realized value of unit i at time t, where $i = 1, ..., n$ refers to the cross-sectional dimension and $t = 1, ..., T$ refers to the time-series dimension.[3] Further, suppose we observe a series of h-step-ahead forecasts of the outcome, $y_{it|t-h}$, generated conditional on information available to the forecaster at time $t-h$. We denote these by $\hat{y}_{it|t-h,m}$, where $m = 1, ..., M$ indexes the individual forecasts (e.g., forecasting models) and $h \geq 0$ is the forecast horizon. To keep the analysis simple, we focus on the case with a pair of competing forecasts, $M = 2$. However, our approach can easily be generalized to a setting with an arbitrary (and growing) number of forecasts, M.

To compare the predictive accuracy of different forecasts we must have a loss function that quantifies the cost of different forecast errors. Following Diebold and Mariano (1995), define the loss associated with forecast m as $L_{it|t-h,m} = L(y_{it}, \hat{y}_{it|t-h,m})$. Consistent with most empirical work, we assume that the loss is a quadratic function of the forecast error, $e_{it|t-h,m} = y_{it} - \hat{y}_{it|t-h,m}$, and thus takes the form[4]

$$L(y_{it}, \hat{y}_{it|t-h,m}) = e^2_{it|t-h,m}. \tag{1}$$

Following Diebold and Mariano (1995) and Giacomini and White (2006), we treat the forecasts as given and make high-level assumptions on the distribution of the forecast errors or, more generally, the sequence of losses $L_{it|t-h,m}$. In particular, we do not consider the effect of estimation error on the distribution of the test statistics which we derive.[5]

[3] To simplify notations, we assume that n does not depend on time, but our analysis readily allows for unbalanced panels.

[4] See Elliott et al. (2005) for a more general loss function that nests squared error loss as a special case.

[5] Estimation error and its effect on tests for equal predictive accuracy features prominently in the analysis of West (1996), Clark and McCracken (2001), McCracken (2007), and Hansen and Timmermann (2015).

2.1 Tests for the Pooled Average

We can consider many different ways to aggregate the loss for panels of forecasts and outcomes. A natural starting point is the pooled average loss associated with forecast m averaged across the T time-series observations and n cross-sectional units:

$$\overline{L}_m \equiv \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} L(y_{it}, \hat{y}_{it|t-h,m}). \tag{2}$$

Our first hypothesis is that the pooled average loss is equal in expectation for a pair of forecasts m_1 and m_2:

$$H_0^{pool} : E[\overline{L}_{m_1}] = E[\overline{L}_{m_2}]. \tag{3}$$

The null in (3) does not rule out that the expected predictive accuracy of a pair of forecasts, m_1 and m_2, is different for a particular time period, t. It also does not rule out that forecast m_1 is more accurate than m_2 for some units, i, while being less accurate for others. Rather, it states that such differences average out across the cross-sectional and time-series dimensions.

To test H_0^{pool}, define the squared-error loss differential between forecasts m_1 and m_2 for unit i at time t as

$$\Delta L_{i,t|t-h} = e^2_{it|t-h,m_1} - e^2_{it|t-h,m_2}. \tag{4}$$

We can then test the null in (3) using the test statistic

$$J^{DM}_{n,T} = (nT)^{-1/2} \frac{\sum_{t=1}^{T} \sum_{i=1}^{n} \Delta L_{i,t|t-h}}{\hat{\sigma}(\Delta L_{t|t-h})}, \tag{5}$$

where $\hat{\sigma}(\Delta L_{t|t-h})$ is a consistent estimator for $\sqrt{\text{Var}\left((nT)^{-1/2} \sum_{t=1}^{T} \sum_{i=1}^{n} \Delta L_{i,t|t-h}\right)}$.

The test statistic in (5) pools information across both the time-series and cross-sectional dimensions and, as such, is naturally viewed as a Diebold-Mariano panel test for equal predictive accuracy (Diebold and Mariano (1995)). Pooling information across both dimensions can potentially provide greater statistical power in empirical work.

Letting $\overline{\Delta L}_{t|t-h} = n^{-1} \sum_{i=1}^{n} \Delta L_{i,t|t-h}$ be the cross-sectional average loss differential at time t, we can define the (scaled) average loss at time t as

$$R_{t|t-h} = n^{1/2}\overline{\mathit{\Delta} L}_{t|t-h}. \tag{6}$$

Under standard assumptions of weak serial dependence in the sequence of forecast losses, we can compute the standard error in the denominator of (5) using a Newey and West (1987) estimator:

$$\hat{\sigma}(\Delta L_{t|t-h}) = \sqrt{\sum_{j=-J}^{J}(1-j/J)\hat{\gamma}_h(j)}, \tag{7}$$

where $J > 0$ is the maximum lag length and $\hat{\gamma}_h(j) = T^{-1}\sum_{t=j+1}^{T}\tilde{R}_{t-j|t-j-h}\tilde{R}_{t|t-h}$ with $\tilde{R}_{t|t-h} = R_{tt-h} - \bar{R}_h$ and $\bar{R}_h = T^{-1}\sum_{s=1}^{T}R_{s-h}$. For $j < 0$, we set $\hat{\gamma}(j) = \hat{\gamma}(-j)$.

Assuming that T is large, under standard conditions we can invoke a central limit theorem (CLT) for the time series data $\{R_t\}_{t=1}^{T}$. We summarize these arguments in the following result:

Theorem 1. *Suppose that* $\max_{1\le t\le T} E|R_{t|t-h}|^r$ *is bounded with* $r > 2$ *and that* $\{R_{t|t-h}\}_{t=1}^{T}$ *is* α*-mixing of size* $-r/(r-2)$*. Also assume that* $\hat{\sigma}(\Delta L_{t|t-h}) = \bar{\sigma}_{n,T} + o_P(1)$ *and* $\bar{\sigma}_{n,T} > 0$ *is bounded away from zero, where* $\bar{\sigma}_{n,T}^2 = \mathrm{Var}\left((nT)^{-1/2}\sum_{t=1}^{T}\sum_{i=1}^{n}\Delta L_{i,t|t-h}\right)$*. Then under* H_0^{pool} *in (3),* $J_{n,T}^{DM} \xrightarrow{d} N(0,1)$*.*

Theorem 1 follows by exploiting the assumption of weak serial dependence. Specifically, by Theorem 5.20 of White (2014), we have

$$J_{n,T}^{DM}\frac{\hat{\sigma}(\Delta L_{t|t-h})}{\bar{\sigma}_{n,T}} \xrightarrow{d} N(0,1).$$

By the consistency of $\hat{\sigma}(\Delta L_{t|t-h})$, the desired result follows from Slutzky's theorem.

Note that we do not require restrictions on the degree of cross-sectional dependence and can allow for arbitrary cross-sectional dependence in the loss differentials. The panel data structure naturally has two dimensions (cross-sectional and temporal) and we only need to exploit one dimension to establish the asymptotic normality. Since the weak temporal dependence is a widely accepted assumption, we have the luxury of being agnostic about the nature of cross-sectional dependence. Even if we have decided to adopt a linear factor model, it might not be immediately clear whether we should impose this model on the forecast error or on the loss differential (such as in Akgun et al. (2022)); it is possible that the final conclusion is sensitive to this subtle modeling choice. Of course, the factor model might not be linear at all; for example, consider $\Delta L_{i,t|t-h} = g(f_t, \lambda_i) + u_{i,t}$, where $g(\cdot,\cdot)$ is a non-parametric function, f_t is the factor and λ_i is the factor loading, see Section 2.4 of Chatterjee (2015). Here, the test statistic $J^{DM}_{n,T}$ avoids having to take a stand on the exact structure of cross-sectional dependence. [6]

In practice, we may be interested in knowing if a particular forecast (m_1 or m_2) was significantly more accurate than an alternative forecast in some time periods or for some variables even if this does not carry over to other periods or hold for all variables. To address this issue, we next develop test statistics that can be used to identify differences across pre-defined groups of time or groups of units.

3 Testing Equal Predictive Accuracy for Subgroups

In many situations, the relative accuracy of a set of economic forecasts can be expected to differ across time or across variables either due to their use of different information sets or due to differences in modeling approaches. The Federal Reserve may, for example, have superior information relative to private forecasters about the state of the economy or the likely future path of interest rates that is particularly useful during financial crises. During normal times, this informational advantage may be smaller. Under this scenario, the economic forecasts of the Federal Reserve could be more accurate than private sector forecasts during financial crises but not during normal times. As a second example, the IMF may have superior expertise and information about developing economies and program countries in particular, whereas information is more symmetric–vis-a-vis private sector forecasters–for advanced economies. As a third example, two forecasts could be equally accurate "on average" with one forecast being better for advanced economies but worse for developing economies.

In situations such as these, the null in (3) of equal "average" predictive accuracy is of less interest as we might be specifically interested in testing whether two forecasts

[6] Specifically, provided that R_{t+h} is weakly serially dependent and satisfies β-mixing, we can establish asymptotic normality for $J = (T)^{-1/2} \sum_{t=1}^{T} R_{t+h}/\hat{\sigma}(\Delta L_{t+h|t})$ in (5) without imposing restrictions on the cross-sectional dependence in the loss differentials.

are equally accurate either across certain periods of time or for different cross-sectional groups or clusters. This section develops a framework for conducting such tests.

3.1 Time Clusters

We first consider testing whether a pair of forecasts are equally accurate during certain pre-defined blocks of time. To this end, we partition the panel of loss differentials along the time-series dimension into a set of K clusters $\{t_1, t_2, ..., t_K\}$ which are assumed to be mutually exclusive and exhaustive so that $\cup_{j=1}^{K} t_j = [1 : T]$. Denote the associated test statistics by $\{R_{t_1}, R_{t_2}, ..., R_{t_K}\}$.[7] For example, if each cluster has equal length, q, the test statistic for the jth cluster can be computed as $R_{t_j} = q^{-1} \sum_{t=(j-1)q}^{jq-1} R_{t|t-h}$.[8] When $q = 1$, each time period is a separate cluster.

Suppose we are interested in testing that the null of equal predictive accuracy for two forecasts holds within each of the time clusters:

$$H_0^{Tcluster} : \; ER_{t_1} = ER_{t_2} = \cdots = ER_{t_K} = 0. \tag{8}$$

The null in (8) does not test whether the loss differential averaged across the K clusters equals zero, i.e., $K^{-1} \sum_{j=1}^{K} ER_{t_j} = 0$. This would be identical to testing the null in (3) which arises as a special case with a single cluster, i.e., $K = 1$. Clearly this null is less restrictive than, and indeed implied by, $H_0^{Tcluster}$ in (8) which tests that equal predictive accuracy holds for *each* time cluster.

Suppose that n is large and assume that a CLT applies to the cross-section of forecast errors so R_{t_j} is Gaussian.[9] Then we can test the null in (8) using the framework for inference with clusters developed by Ibragimov and Müller (2010, 2016). In the present context, this approach offers several advantages. Besides arising naturally as a way of testing (8), the approach does not require stationarity of the underlying loss differentials. Moreover, it can be used with as little as $T = 2$ time periods and gives rise to a t-test that is easily computed:

[7] For simplicity, we suppress the h subscript in our notations here, but it is implicit that all underlying forecasts use a horizon of h periods.

[8] More generally, $q = \lfloor T/K \rfloor$ is the average length of each cluster and we can let the cluster length vary across the K clusters.

[9] This assumption also rules out strong serial dependence among the loss differentials.

$$J_n^R = \frac{\sqrt{K}\bar{R}}{\sqrt{(K-1)^{-1}\sum_{j=1}^{K}(R_{t_j}-\bar{R})^2}}, \tag{9}$$

where $\bar{R} = K^{-1}\sum_{j=1}^{K} R_{t_j}$ is the loss differential averaged across the K clusters. We can establish the distributional properties of the test statistic in (9) under the following assumption:

Assumption 1. *Let $R_{(n)} = (R_{t_1}, ..., R_{t_K})' \in \mathbb{R}^K$. Suppose that $R_{(n)} - ER_{(n)} \to^d N(0, \Omega)$ as $n \to \infty$, where Ω is a diagonal matrix.*

The diagonal matrix in Assumption 1 requires that R_{t_j} is asymptotically independent across the K clusters. This is a very mild assumption that is satisfied under the usual mixing conditions or other weak temporal dependence assumptions. As discussed in Section 3.1 of Ibragimov and Müller (2010), one way to achieve this is by separating the subsamples (t_j) by a sufficient number of time periods.[10]

Importantly, Assumption 1 refers to the properties of the loss differentials and so we do not require that the data generating process for the outcome variable be stationary provided that any non-stationary components either are incorporated in both forecasts or, if this does not hold, affect both forecasts equally and so vanish from the loss differentials.

By Theorem 1 of Ibragimov and Müller (2010) and the continuous mapping theorem, we have the following result:

Theorem 2. *Suppose that Assumption 1 and one of the following conditions hold:*
(1) $K \geq 2$ and $\alpha \leq 0.08326$.
(2) $2 \leq K \leq 14$ and $\alpha \leq 0.1$.
(3) $K \in \{2, 3\}$ and $\alpha \leq 0.2$.
Then under $H_0^{T cluster}$ we have

$$\limsup_{n\to\infty} P\left(|J_n^R| > t_{K-1,1-\alpha/2}\right) \leq \alpha.$$

Here $t_{K-1,1-\alpha}$ denotes the $1-\alpha$ quantile of the Student-t distribution with $K-1$ degrees of freedom. When the test statistics computed for the individual clusters do not have the same variance, using critical values from the student-t distribution can

[10] Provided that the data are weakly dependent, by a CLT it follows that the cluster averages will be Gaussian with diagonal variance-covariance matrix.

lead to conservative inference. With a conventional test size ($\alpha \leq 0.05$), we only need $K \geq 2$ clusters to apply the test. However, if $\alpha = 0.10$, we can have at most $K = 14$ clusters.

An alternative strategy for testing the null in (8) is to use the randomization test proposed by Canay et al. (2017). Define the randomization p-value:

$$\hat{p}_R = 2^{-K} \sum_{\xi_1,\ldots,\xi_K \in \{-1,1\}} \mathbf{1}\left\{ \left| \sum_{j=1}^{K} R_{t_j} \xi_j \right| > \left| \sum_{j=1}^{K} R_{t_j} \right| \right\}, \tag{10}$$

where $\xi_1, \ldots, \xi_K \in \{-1, 1\}$ are all possible combinations of the K variables $\xi_1, \ldots, \xi_K$, each of which takes a value of ±1.

Under the conditions stated in Assumption 1, we have the following result:

Theorem 3. *Suppose Assumption 1 holds. Then under $H_0^{Tcluster}$ we have*

$$\limsup_{n\to\infty} \left| P\left(\hat{p}_R > \alpha\right) - \alpha \right| \leq (3/2) \times 2^{-K}.$$

The formal proof of this result is available in the Appendix. The result here is not the same as Theorem 3.1 of Canay et al. (2017) since we do not resolve ties by a random coin flip. Although the assumptions of the randomization test in (10) are the same as those used by the t-test in (9), the two tests have different properties. For example, the t-test might be more accurate when K is very small. In empirical applications with less than five clusters, inference at the 5% significance level only rejects when the p-value is exactly zero. As pointed out in Canay et al. (2017), the bound on the size distortion in Theorem 3 implies that, provided the number of clusters is not too small, the null rejection probability will be at least $\alpha - (3/2) \times 2^{-K}$. On the other hand, for larger values of K, Theorem 3 implies a similarity property of the randomization test.[11]

We note that the proposal here for testing the time clusters can accommodate cross-sectional dependencies. An important case is the factor models considered in Akgun et al. (2022). Under the data-generating process in Equation (1) and Assumptions 1 and 6 therein, we can show (in their notations) that $n^{-1}\sum_{i=1}^{n} E(\Delta L_{it}) = n^{-1}\sum_{i=1}^{n} \mu_i =: \bar{\mu}_n$, which does not depend on t. Hence, to test the null hypothesis of $\bar{\mu}_n = 0$ (or $H_{0,1}$ therein), we can divide the time dimension into K clusters and apply

[11] A test is similar if its rejection probability is the same across all parameter values that satisfy the null hypothesis.

our proposed method above.

3.2 Cross-sectional Clusters

In addition to testing the null of equal predictive accuracy for a pair of forecasts, averaged cross-sectionally for different blocks in time, we can also test whether the forecasts are equally accurate within each of a set of pre-specified cross-sectional clusters. This type of test typically averages over the full time-series sample, as opposed to the test in (9) which performs cross-sectional averaging. For example, we may be interested in testing whether two forecasts are equally accurate for advanced as well as for developing economies. This null does *not* amount to testing whether the predictive accuracy is the same for advanced and developing economies–we would generally expect forecasts to be less accurate for the more volatile developing economies. Rather, it amounts to separately testing whether a pair of forecasts have the same expected accuracy among developing economies as well as among advanced economies even though, in absolute terms, their predictive accuracy could be different across the two sets of economies.

To set up such a test, suppose that the individual units have been categorized into K cross-sectional clusters, denoted by $H_1, ..., H_K$. Let $|H_j|$ denote the number of elements in the jth cluster, i.e., the cardinality of H_j, with $\sum_{j=1}^{K} |H_j| = n$ and define

$$D_j = |H_j|^{-1/2} T^{-1/2} \sum_{i \in H_j} \sum_{t=1}^{T} \Delta L_{i,t|t-h}, \tag{11}$$

The null hypothesis of equal predictive accuracy within each cross-sectional cluster takes the form

$$H_0^{Ccluster} : \; ED_1 = ED_2 = \cdots = ED_K = 0. \tag{12}$$

This setup is equivalent to that in Section 3.1. However, here we rely on the time-series dimension T being sufficiently large to ensure that the K time-series averages of loss differentials are approximately Gaussian and the goal is to test that their means are all zero.

Let $\bar{D} = K^{-1} \sum_{j=1}^{K} D_j$ be the average of the loss differences across the K cross-sectional clusters and consider the test statistic

$$J_n^D = \frac{\sqrt{K}\bar{D}}{\sqrt{(K-1)^{-1}\sum_{j=1}^{K}(D_j-\bar{D})^2}}. \tag{13}$$

Analogous to the result for the time-series clusters, we make the following assumption:[12]

Assumption 2. *Let $D_{n,T} = (D_1, ..., D_K)' \in \mathbb{R}^K$. Suppose that $D_{n,T} - E(D_{n,T}) \to^d N(0, \Omega)$ as $n, T \to \infty$, where Ω is a diagonal matrix.*

Assumption 2 relies on a CLT for time-series averages and so rules out situations with either a small T or strong serial dependency. By Theorem 1 of Ibragimov and Müller (2010) and the continuous mapping theorem, we have the following result:

Theorem 4. *Suppose Assumption 2 and one of the following conditions hold:*
(1) $K \geq 2$ and $\alpha \leq 0.08326$.
(2) $2 \leq K \leq 14$ and $\alpha \leq 0.1$.
(3) $K \in \{2, 3\}$ and $\alpha \leq 0.2$.
Then under $H_0^{Ccluster}$ the following holds

$$\limsup_{n,T\to\infty} P\left(|J_n^D| > t_{K-1,1-\alpha/2}\right) \leq \alpha.$$

Theorem 4 establishes conditions under which the simple test procedure of Ibragimov and Müller (2010) can be applied to test the null of equal predictive accuracy within clusters of units formed as subsets of the cross-sectional data.

Similarly, we can establish a result that is equivalent to Theorem 3 for the cross-sectional clusters. To this end, define the randomization p-value:

$$\hat{p}_D = 2^{-K} \sum_{\xi_1,...,\xi_K \in \{-1,1\}} \mathbf{1}\left\{\left|\sum_{j=1}^{K} D_j \xi_j\right| > \left|\sum_{j=1}^{K} D_j\right|\right\}. \tag{14}$$

Using Assumption 2, we have

Theorem 5. *Suppose Assumption 2 holds. Then under $H_0^{Ccluster}$ we have*

$$\limsup_{n,T\to\infty} |P(\hat{p}_D > \alpha) - \alpha| \leq (3/2) \times 2^{-K}.$$

[12] A sufficient condition for $D_{n,T} - E(D_{n,T}) \to^d N(0, \Omega)$ is that $|H_j| \to \infty$ for each j along with weak serial dependence for $\Delta L_{i,t+h|t}$.

Provided that the conditions in Assumption 2 hold, this result means that we can apply the randomization test in (14) to the cross-sectional clusters.

3.3 De-correlating Clusters

Assumption 2 requires that the covariance matrix Ω is (block) diagonal. This may well be a good approximation for some empirical applications but is likely to fail in many other cases. In what follows we therefore consider both cases.

We begin with a data generating process (DGP) that satisfies Assumption 2. Let

$$e_{t,i} = \lambda_i f_{t,g(i)} + u_{t,i}, \tag{15}$$

where $u_{t,i}$ is i.i.d $N(0, \sigma_u^2)$ across i and t and $g(i)$ is the cluster of variable i, i.e., $g(i) \in \{1, ..., K\}$. We assume that $f_{t,k}$ is i.i.d across t and $k \in \{1, ..., K\}$ so each cluster maps into a unique factor consistent with the block-diagonal structure of Ω. Further, we can allow for serial correlation in $f_{t,k}$ and $u_{t,i}$.

Next, consider a DGP that fails to satisfy Assumption 2:

$$\begin{aligned} e_{t,i} &= \lambda_i f_t + u_{t,i}, \\ f_t &= \phi f_{t-1} + \xi_t, \\ u_{t,i} &= \rho u_{t,i} + v_{t,i} \end{aligned} \tag{16}$$

with i.i.d ξ_t and $v_{t,i}$ across t and i. In this case, the covariance matrix of the forecast errors is not block-diagonal and tests that assume such a structure are unlikely to work well.

A possible approach for handling deviations from block diagonality in forecast errors is to first decorrelate the clusters of forecast errors and then apply the method to the decorrelated data. We next explain how to implement these steps. Let $A_t = (A_{1,t}, A_{2,t}, ..., A_{K,t})'$ be the loss differentials averaged within each of the K clusters:

$$A_{j,t} = |H_j|^{-1} \sum_{i \in H_j} \Delta L_{i,t|t-h}.$$

Next, define the estimated second-moment matrix $\hat{\Omega} = T^{-1} \sum_{t=1}^{T} A_t A_t'$, and let

$$B_{n,T} = \hat{\Omega}^{-1/2} \sum_{t=1}^{T} A_t.$$

We can now implement the tests in Theorems 4 and 5, except that $D_{n,T}$ is replaced by $B_{n,T}$. We consider this procedure both in the Monte Carlo simulations and in the empirical work.

4 Monte Carlo Simulations

In this section we compare the finite-sample performance of the test statistics through a set of Monte Carlo simulation experiments. We consider two new DGPs to evaluate the size and power of our panel tests. The first (DGP 1) assumes that the cross-sectional forecast errors are independent but serially persistent. Conversely, the second (DGP 2) posits an hierarchical factor structure in the forecast errors, allowing for both cluster-specific as well as a global factor in the forecast errors. We also consider two DGPs from Akgun et al. (2022): DGP 3 assumes a spatial AR(1) process in forecast errors, while DGP 4 assumes that loss differentials follow a two-factor model.

4.1 Data Generating Processes

We begin by explaining how we set up our Monte Carlo experiments. Throughout the analysis, we consider squared error loss $L(e_{t,i}) = e_{t,i}^2$ for a pair of forecast errors denoted $e_{t,i,1}$ and $e_{t,i,2}$ with associated loss differentials $\Delta L_{t,i} = e_{t,i,1}^2 - e_{t,i,2}^2$.[13] Throughout the analysis, we focus on a test size of 5% and report results based on $1,000$ MC simulations.

Our main MC simulations consider four values for the time-series dimension: $T =$50, 100, 500, 1,000 and two values of the cross-sectional dimension: $N = 50, 100$. Our implementation of the permutation tests sets the number of blocks (K) equal to 5 and 10.[14] Each set of forecast errors, $e_{t,i,1}$ of model 1 and $e_{t,i,2}$ of model 2, forms a $T \times N$-dimensional panel.

[13] For simplicity we drop references to the forecast horizon, h, of the forecast errors.

[14] Fixing the value of K is consistent with the assumptions in Theorem 2 and 3.

Turning to the specific assumptions, DGP 1 assumes that the time series of individual forecast errors follow AR(1) processes that are independent in the cross-section:

$$\begin{aligned} e_{t,i,1} &= \mu_1(1-\phi) + \phi e_{t-1,i,1} + \varepsilon_{t,i,1}, \quad \varepsilon_{t1} \sim N(0,\sigma_\varepsilon^2), \\ e_{t,i,2} &= \mu_2(1-\phi) + \phi e_{t-1,i,2} + \varepsilon_{t,i,2}, \quad \varepsilon_{t2} \sim N(0,\sigma_\varepsilon^2), \end{aligned} \tag{17}$$

where $\varepsilon_{t,i,1}$ and $\varepsilon_{t,i,2}$ are independent across time, variables and models and $\sigma_\varepsilon = 1$.

We consider two values of the persistence parameter, ϕ, namely $\phi = 0.5$ (DGP 1.1) and $\phi = 0.8$ (DGP 1.2), corresponding to modest and fairly high persistence in the forecast errors.[15]

In the simulations used to evaluate the size of our tests, we impose the null of equal predictive accuracy by setting $\mu_1 = \mu_2 = 0$. For all simulation exercises, we compute Newey-West standard errors for the DM tests based on a Bartlett kernel with a maximum lag length set at $T^{1/3}$.

DGP 2 adopts a hierarchical factor structure in the forecast errors:[16]

$$\begin{aligned} e_{t,i,m} &= \lambda h_{t,m} + f_{t,g(i),m} + u_{t,i,m}, \\ h_{t,m} &= \phi_h h_{t-1,m} + \varepsilon_{t,m}, \\ f_{t,g(i),m} &= \phi f_{t-1,g(i),m} + \xi_{t,g(i),m}, \\ u_{t,i,m} &= (1-\rho)\mu_m + \rho u_{t,i,m} + v_{t,i,m}. \end{aligned} \tag{18}$$

Here, $g(i)$ is the cluster that variable i belongs to and $g(i) \in \{1, ..., K\}$. We further assume that the factors $f_{t,k,m}$ are i.i.d across time t, models $m \in \{1,2\}$ and clusters $k \in \{1, ..., K\}$. Thus, each cluster is affected by its own factor. To generate dependencies across units belonging to different clusters, we also include a global factor, h_t, that affect all units. We allow for serial correlation in h_t, $f_{t,k}$ and $u_{t,i}$. For the simulation exercise, we assume $\phi_h = \phi = \rho = 0.5$ while ε_t, $\xi_{t,k}$ and $v_{t,i}$ are standard Gaussian variables that are i.i.d. across t, k and i. The size of each cluster is the same, N/K, and $g(i) = \lceil iK/N \rceil$.

We consider two values for λ in equation (18): $\lambda = 0$ for DGP 2.1 and $\lambda = 1$ for DGP 2.2. Hence, errors are independent across groups ($\lambda = 0$) under DGP

[15] While it is not uncommon for economic variables to be highly persistent, forecast errors can be expected to be far less persistent, at least under squared error loss.

[16] By assuming a factor structure in the underlying forecast errors with separate factors for the two sets of models, our setup allows loss differentials to be correlated within clusters.

2.1. DGP 2.2 ($\lambda = 1$) represents the case where errors are not independent across groups. DGP 2.3 assumes identical factors across forecast errors of two models by putting further constraints on DGP 2.2: $h_{t,1} = h_{t,2}$ and $f_{t,g(i),1} = f_{t,g(i),2}$ for all i and t. As proved in Qu et al. (2021), the sum of squared factors are offset in the squared loss differentials, such that the DGP 2.3 satisfies the null hypothesis of equal expected squared loss *conditional* on factor realizations, at each single cross-section. To examine the size of the tests, again we set $\mu_1 = \mu_2 = 0$. To evaluate their power, we set $\mu_1 = \kappa/(NT)^{0.25}, \mu_2 = 0$.

Our third specification (DGP 3) is based on the model in equations (28) and (29) in Akgun et al. (2023) applied to a pair of forecast errors:

$$e_{t,i,m} = \rho \sum_{j=1}^{n} w_{i,j} e_{t,j,m} + u_{t,i,m}, \ m \in \{1,2\}, \tag{19}$$

where $u_{t,i,m}$ are i.i.d. Gaussian $N(0, \sigma_m^2)$. and w_{ij} is an element from a spatial matrix W_n. Letting $e_{t,m} = (e_{t,1,m}, e_{t,2,m}, ..., e_{t,n,m})'$ and $u_{t,m} = (u_{t,1,m}, u_{t,2,m}, ..., u_{t,n,m})'$, it can be shown that $e_{t,m}$ can be sampled as

$$e_{t,m} = (I_n - \rho W_n)^{-1} u_{t,m}.$$

Spatial interactions between units are created with a row-normalized rooktype weight matrix, W, whose units lie on a $p_1 \times p_2$ rectangular grid such that the first p_1 units are located in the first column of the grid, the second p_1 units located in the second column, and so on with $n = p_1 p_2$. For each $n \in \{10, 20, 30, 50, 100\}$, we choose p_1 as 2, 3, 6, 10 and 50, respectively. Following Akgun et al. (2022), we set $\rho = 0.5$. To evaluate the size of the tests, we set $\sigma_1 = \sigma_2 = 1$ and to estimate their power, we set $\sigma_1 = 1$ and $\sigma_2 = 1 + 0.05\kappa/(NT)^{0.25}$ and vary κ.

Our fourth and final specification (DGP 4) uses the model in equation (30) of Akgun et al. (2022) and so directly specifies the loss differentials as a two-factor model:

$$\Delta L_{t,i} = \mu + \lambda_{i1} f_{t1} + \lambda_{i1} f_{t2} + \varepsilon_{t,i}, \tag{20}$$

where ε_{it} follows (19) scaled by a constant:

$$\varepsilon_{ti} = \sqrt{N \text{tr}\left((I - \rho W)(I - \rho W)'\right)} e_{t,i,1}. \tag{21}$$

Factor loadings are generated independently as $\lambda_{i1}, \lambda_{i2} \sim N(1, 0.2)$ and the factors are drawn from independent standard Gaussian distributions, $f_{t1}, f_{t2} \sim N(0, 1)$. We set $\mu = 0$ when evaluating size. To compute the power of the tests, we set $\mu = 10\kappa/\sqrt{NT}$.

4.2 Akgun et al. (2023) Test Statistics

Using a somewhat different modeling framework than ours, Akgun et al. (2022) propose a variety of test statistics to test the null hypothesis of equal predictive accuracy of panel forecasts on average, (3). For comparison, and to illustrate the importance of how cross-sectional dependency is treated, we consider both their test statistics that assume no cross-sectional dependence as well as two of their tests that allow for strong cross-sectional dependence. Their S1 test statistic is suitable for situations with cross-sectional independence and is computed as

$$S_{nT}^{(1)} = \frac{\sqrt{nT}\overline{\mathit{\Delta} L_t}}{\hat{\sigma}_{1,nT}}, \tag{22}$$

where

$$\begin{aligned}
\hat{\sigma}^2_{1,nT} &= \frac{\sum_{i=1}^{n}\sum_{t,s=1}^{T} k_T\left(|t-s|/d_T\right)\Delta\tilde{L}_{it}\Delta\tilde{L}_{is}}{NT}, \\
\Delta\tilde{L}_{it} &= \Delta L_{it} - T^{-1}\sum_{t=1}^{T}\Delta L_{it}, \\
\overline{\mathit{\Delta} L}_{nT} &= \sum_{i=1}^{n}\sum_{t=1}^{T}\Delta L_{it}/nT.
\end{aligned}$$

Here $k_T(\cdot)$ is a kernel function, and d_T is its width such that $\lim_{T\longrightarrow\infty} d_T^2/T \longrightarrow 0$. For the simulation exercise, we use a Bartlett kernel for $k_T(\cdot)$ and set $d_T = T^{1/3}$.

The $S3$ test of Akgun et al. (2022) allows for strong cross-sectional dependence and is computed as

$$S_{nT}^{(3)} = \frac{\sqrt{nT}\overline{\mathit{\Delta} L_t}}{\hat{\sigma}_{3,nT}}, \tag{23}$$

where

$$\hat{\sigma}^2_{3,nT} = \frac{\sum_{i,j=1}^{n}\sum_{t,s=1}^{T} k_T\left(|t-s|/d_T\right)\Delta\tilde{L}_{it}\Delta\tilde{L}_{js}}{nT}.$$

Akgun et al. (2022) also develop statistics to test hypothesis (12). Let $\overrightarrow{D} = (D_1, D_2, ..., D_K)'$, where D_k is defined in equation (11) Their C1 statistic requires cross-sectional independence and is computed as

$$C_{NT}^{(1)} = \overrightarrow{D}' \hat{\Omega}_{1,nT}^{-1} \overrightarrow{D}, \tag{24}$$

where

$$\hat{\Omega}_{1,nT} = \frac{1}{T} \sum_{i=1}^{n} \sum_{t,s=1}^{T} \frac{k_T \left(|t-s|/d_T\right) \iota_{g_i} \iota_{g_i}' \Delta \tilde{L}_{it} \Delta \tilde{L}_{is}}{n_{g_i}}$$

with $g_i \in \{1, 2, ..., K\}$ indicating the group index that variable i belongs to. n_{g_i} is the number of variables in group g_i. ι_{g_i} is a column vector with n elements whose jth element equals one if j belongs to group g_i and otherwise equals zero.

Finally, the C3 test statistic of Akgun et al. (2022) allows for strong cross-sectional dependence by considering the correlation between $\Delta \tilde{L}_{it}$ and $\Delta \tilde{L}_{js}$ to compute the standard deviation of the loss differentials:

$$C_{nT}^{(3)} = \overrightarrow{D}' \hat{\Omega}_{3,nT}^{-1} \overrightarrow{D}, \tag{25}$$

where

$$\hat{\Omega}_{3,nT} = \frac{1}{T} \sum_{i,j=1}^{n} \sum_{t,s=1}^{T} \frac{k_T \left(|t-s|/d_T\right) \iota_{g_i} \iota_{g_j}' \Delta \tilde{L}_{it} \Delta \tilde{L}_{js}}{\sqrt{n_{g_i} n_{g_j}}}.$$

4.3 Simulation Results

Our first DGP (DGP1) allows forecast errors to be serially correlated though there is no cross-sectional correlation, no group structure and no common factors affecting the forecast errors. For this "plan vanilla" process, we would expect our tests to have the right size and this is generally what we find. For the version with modest serial correlation (DGP 1.1 in Table 1), our time-block and group tests all have excellent size properties with close to the correct size of 5% across different combinations of T, N, and K. Hence there is no need to use the decorrelated test statistics in this scenario though these tests also appear to be properly sized.

In contrast, the S1 and S3 tests proposed by Akgun et al. (2022) are somewhat oversized in the two small samples ($T = 50, 100$) with a rejection rates equal to roughly twice the nominal size of 5%. These size distortions tend to disappear, however, as T increases to $T = 500, 1000$. The C1 and C3 test statistics proposed by Akgun et al. (2022) are both seriously oversized with rejection rates for C1 between 15% and 23% and rejection rates of C3 between 25% and 57% in the smallest sample ($T = 50$) .

Although these rejection rates are reduced in the largest sample ($T = 1000$), the C1 and C3 tests remain oversized with rejection rates close to twice the nominal size.

For the more persistent time series (DGP 1.2 in Table 2), our test statistics continue to have excellent size properties, though the time-block permuation test is somewhat oversized in the smallest sample ($T = 50$). Similarly, for both DGP 1.1 and 1.2, the single cross-section test of Qu, Timmermann, and Zhu (2023) has roughly the correct size in both cross-sections ($N = 50, 100$).

Size distortions are pronounced for the S1 and S3 test statistics which have rejection rates around 25% for $T = 50$ and 11-13% for $T = 1,000$. Thus, even in data with a large time-series dimension, these tests over-reject. Similarly, the C1 and C3 tests now hugely overreject with rejection rates around 60-80% for the C1 test statistic and rejection rates in the 70-97% range for the C3 test for the data with the smallest time-series dimension ($T = 50$). Even for the longest data ($T = 1,000$), rejection rates for both of these test statistics exceed 24% across different combinations of $N$and K.

Next, consider the second DGP (DGP 2) which assumes a cluster dependence structure with each cluster having its own separate factor (Table 3). For DGP 2.1, which imposes that the errors are dependent within clusters but independent across clusters ($\lambda = 0$), our test statistics continue to have excellent size properties as the only systematic evidence of over-sizing occurrs for the permutation tests when $T = 50$ and $K = 5$. In contrast, the S1 test statistic of Akgun et al. (2022) is hugely oversized with rejection rates typically exceeding 40% when $N = 50$ and $K = 5$ and rejection rates that exceed 70% for $N = 100$ and $K = 5$. This holds even for the largest values of T and so size distortions do not disappear even in data with a large time-series dimension. The S3 test statistic performs much better with a size around 10% for $T = 50$ and a size closer to 6% when $T = 1,000$. The C1 test displays even greater size distortions than the S1 test with rejection rates exceeding 90% in almost all cases. The C3 test statistic also is greatly oversized in the samples with a smaller time-series dimension ($T = 50, 100$) but has a rejection rate close to 10% when $T = 1,000$.

DGP 2.2 allows the forecast errors to be correlated both within groups and across groups ($\lambda = 1$). Under this setup (Table 4), the time-block t-test and time-block permutation test continue to have good size properties. Now, however, the group t-test and group permutation tests are seriously oversized with rejection rates around 15-20% for $K = 5$ and rejection rates of 27-29% for $K = 10$. Importantly, however, the decorrelated group t-test and permutation test succeeds in effectively correcting these size distortions, leading to tests with approximately correct rejection rates. In contrast, the S1 and C1 tests are even more oversized for DGP 2.2 than for DGP 2.1, while the size distortions for the S3 and C3 test statistics are comparable to those found for DGP 2.1.

DGP 2.3 (Table 5) assumes identical factors across the two panels of forecast errors and the resulting size of the time-block and group tests is similar to those found under DGP 1. The key difference is that, consistent with theory, the size of the single cross-section test now is around 5%. The size of the S1 and S3 tests exceeds 10% when T is smaller than 100 while the C1 and C3 tests are even more oversized than the S1 and S3 tests.

Next, we turn to the two DGPs proposed by Akgun et al. (2022). Recall that DGP 3 assumes a spatial dependence structure in the forecast errors.[17] For this process (Table 6), our time-block t-test and permutation test have no material size distortions across all values of N, K, and T. The group t-test is mildly oversized for $T = 50$ if N is small but oversizing in the rejection rate gets reduced as the T and N dimensions grow larger. The group permutation test tends to be mildly oversized, particularly when $N = K = 10$ with rejection rates around 10%. Moreover, while the upward bias in the rejection rate tends to decrease when N increases, it does not disappear when T grows. Once again, however, the decorrelated group t-test or permutation test effectively gets rid of the over-rejection of these tests.

Due to the spatial correlation in forecast errors, the S1 test overrejects with a size a little above 10% for most combinations of T and N. Conversely, the S3 test statistic, which accounts for cross-sectional dependencies in forecast errors, has the right size even for the smallest sample sizes such as $T = 50, N = 10$. For most combinations of T, N the C1 test is oversized and rejects in about 10% of cases, though it has a better size when $N = K = 10$ and $T = 1,000$. The C3 test is somewhat oversized mainly for small values of T ($T = 50$), particularly when $K = 10$ and N is small. This tendency to overreject disappears, however, for large values of T.

The fourth DGP (DGP 4 reported in Table 7) is quite different from the first three DGPs as it specifies factors directly for loss differentials with spatial dependencies in the residuals. Under this DGP, the time-block t-test has near-perfect size while the time-block permutation test is mildly oversized but generally has size close to the nominal value. The Group t-test and group permutation tests are hugely oversized,

[17] Following Akgun et al. (2023), report results on a finer grid of values of N.

however, with rejection rates between 50% and 80%. The decorrelated group t-test and group permutation tests handle the correlation in loss differentials much better, and are close to having the correct size.

For this DGP, the S1 test is hugely oversized with rejection rates between 45% and 80%, while the S3 test comes close to being correctly sized, albeit with a modest amount of over-rejection (1-2%) for $T = 50$. Again, this difference can be explained by how the tests handle cross-sectional dependencies either by ignoring it (S1) or accounting for it (S3). The C1 test is also strongly oversized with rejection rates that hardly vary across different values of T but increase in N, e.g., from around 25% for $N = 10$ to around 70% for $N = 100$ when $K = 5$. The C3 test is somewhat biased when $T = 50, 100$ with rejection rates around 10%, but the size of this test approaches the correct level for larger values of T such as $T = 1,000$. Biases in size do not depend much on the value of N but tend to increase a little when we move from $K = 5$ to $K = 10$ blocks.

The null of equal conditionally expected loss differentials entertained by the single cross-section test fails to hold under DGP 3 and DPG 4. We skip examining the rejection rates of the single cross-section test under DGP 3 and 4.

4.4 Power of Tests

To evaluate the power of the tests for DGP 1.1 and 1.2, we set $\mu_1 = \kappa/(NT)^{1/4}$ and $\mu_2 = 0$ and examine how the power shifts as κ moves away from zero. We set $N = 100$ and $T = 50$. The x axis in Figures 1 and 2 is κ. When evaluating the power of the single cross-section test, we take the last period as sample. Figure 1 shows the results. The time-block t-test and time-block permutation test tend to have slightly higher power than the grouped tests. Conversely, using decorrelated data can reduce the power of the test. Unsurprisingly, ignoring time-series information and using the single cross-section test leads to the largest decline in power. Stronger serial correlation in forecast errors also tends to weaken the power of the tests, as can be seen by comparing the top and bottom panels in Figure 1.[18]

Similar results are found for the tests implemented under DGP 2.1 and 2.2. Figure 2 shows that the time-block t-test and permutation tests produce power that is on a

[18] We omit results for the Akgun et al. (2022) tests which tend to be oversized in small samples for these DGPs.

par or better than the other tests - at least those that have the right size.[19]

5 Empirical Results

To illustrate the economic insights that can be gained from the test statistics introduced in sections 2 and 3, we next conduct an empirical analysis that focuses on the predictive accuracy of the International Monetary Fund's (IMF) World Economic Outlook (WEO) forecasts of real GDP growth and inflation across the world's economies.[20] We compare the WEO forecasts to forecasts from a private-sector organization (Consensus Economics) as well as forecasts from a simple autoregressive model.

5.1 Data

The IMF WEO forecasts are reported twice each year, namely in April (labeled Spring, or S) and October (Fall, or F), for the current-year ($h = 0$) and next year ($h = 1$) horizons. As illustrated in Figure 3, this produces a set of four forecast horizons, listed in decreasing order: $\{h = 1, S;\ h = 1, F;\ h = 0, S;\ h = 0, F\}$.[21] For a subset of (mostly advanced) countries, current-year forecasts go back to 1990, while next-year forecasts start in 1991. For other countries the forecasts start later, giving a shorter data sample. For all countries, the last outcome is recorded for 2019. In total, the WEO forecasts cover 182 countries.

We compare the WEO forecasts at the four forecast horizons to current-year and next-year forecasts reported by the Consensus Economics organization in their April and October surveys. Consensus Economics (CE) is a London-based organization which each month surveys a range of private forecasters. Their forecasts are carefully checked and are known to be of high quality. Moreover, their forecasts have been used in prior studies such as Loungani (2001), Patton and Timmermann (2010) and Patton and Timmermann (2011). The list of countries covered by CE is somewhat smaller than that covered by the WEO forecasts, restricting the cross-sectional dimension of

[19] We have omitted results for the single cross-section test which is over-sized for these DGPs.

[20] The WEO forecasts are extensively followed by the public and have been the subject of a number of academic studies, as summarized in Timmermann (2007).

[21] The WEO forecasts cover forecast horizons up to five years but we do not use the longer forecast horizons due to the relatively short time span of our data.

our comparison. In total, we can compare the WEO and CE forecasts for 85 (real output growth) or 86 (inflation) countries.

We also compare the one-year-ahead ($h = 1$) WEO forecasts to forecasts generated by an AR(1) model estimated separately for each country. Though this is a very simple approach, parsimonious models have often proven difficult to beat in empirical analyses of out-of-sample forecasting performance, see, e.g., Faust and Wright (2013). Autoregressive forecasts of the outcome variable in year t, y_{it}, are based on a forecasting model that uses data on the outcome for the previous year, y_{it-1}, to estimate the intercept and AR(1) coefficient. This gives an advantage to the AR(1) model because, in practice, the previous year's GDP growth and inflation are not observed until well into year t. Data on actuals extend back to 1985 and we use the 10-year window 1985-1994 as a warm-up period, adopting a recursively expanding estimation window to produce subsequent forecasts. Thus, the first AR(1) forecast uses data from 1985-1994 to predict the outcome for 1995. The second forecast uses data from 1985-1995 to predict the outcome for 1996, and so on.

5.2 Comparisons of GDP Growth Forecasts

The top row in Table 8 reports values of the test statistic for the null of equal (pooled average) predictive accuracy (H_0^{pool} in (3)) which averages squared-error loss differences both cross-sectionally and across time. We set up the tests so that positive values indicate that the WEO forecasts are, on average, more accurate than the CE or autoregressive forecasts, while negative values suggest the opposite.[22]

First, consider the forecasts of real GDP growth (Panel A). The pooled average t-test in equation (5) is positive or zero across all forecast horizons. However, the tests comparing the accuracy of the WEO forecasts to the CE forecasts (four left-most columns) fail to be significant for any of the individual horizons. Conversely, the comparisons of the one-year-ahead WEO forecasts to the autoregressive forecasts (listed in the two right-most columns in Table 8) show that the WEO Fall next-year forecasts (though not the Spring forecasts) are significantly more accurate, on average, than the AR forecasts with a t-statistic of 2.45 and a p-value of 0.01.

Next, consider testing the null $H_0^{Tcluster}$ in (8) that the forecasts are equally accurate for all time clusters. We first treat individual calendar years as separate time

[22] In particular, this means that m_1 refers to the CE or autoregressive forecasts while m_2 refers to the WEO forecasts.

clusters so that the current-year forecasts ($h = 0$) are based on $K = 30$ one-year time clusters, while next-year forecasts ($h = 1$) are based on $K = 29$ one-year clusters. Rows 3 and 4 in Table 8 report the t-statistic from equation (9) along with the p-value for a one-sided test against the alternative that the WEO forecasts are more accurate. None of the t-tests comparing the WEO and CE forecasts is statistically significant as evidenced by the p-values which all exceed 0.30. The randomization test (10) reported in the fifth row leads to identical conclusions with p-values ranging from 0.34 to 1.00, suggesting that there is no statistically significant differences in the average predictive accuracy of the WEO vs CE forecasts for any of the individual years during our sample.

We also consider an alternative time-clustering scheme that uses three time clusters arranged around the Global Financial Crisis (GFC), namely 1995-2006, 2007-2009, and 2010-2019.[23] The results, listed in line six for the randomization p-value, show that we cannot reject the null that the CE and WEO forecasts of GDP growth were equally accurate before, during and after the GFC. Conversely, with p-values below 0.01, there is very strong evidence that the one-year-ahead Spring and Fall WEO forecasts of GDP growth were significantly more accurate than the autoregressive forecasts for at least one of these time clusters. This stands in contrast to the results from applying the same test statistic to the individual years and suggests that additional power can be gained from grouping time periods based on economic characteristics–in this case the unfolding of a major global crisis.

We also consider results that cluster the country observations along a set of IMF classifications which consider geographical regions and economic development stages. Specifically, we use a partition of seven clusters of countries, namely (i) Advanced Economies (labeled ae and containing 36 countries in 2016), (ii) Emerging and Developing Europe (eeur, 9), (iii) Emerging and Developing Asia (dasia, 27), (iv) Latin America and the Caribbean (lac, 32), (v) Middle East, North Africa, Afghanistan, and Pakistan (menap, 21), (vi) Commonwealth of Independent States (cis, 12), and (vii) Sub-Sahara Africa (ssa, 45), with the acronym for the cluster and the number of countries within each cluster listed in parentheses. Consensus Economics cover fewer countries in their forecasts–particularly among developing economies. To ensure that we have a sufficiently large number of members in each cluster in the WEO vs. CE comparison, we therefore merge the Emerging and Developing Asia, Middle

[23] See https://www.stlouisfed.org/financial-crisis/full-timeline for a timeline of the financial crisis.

East, North Africa, Afghanistan, and Pakistan, and Sub-Sahara Africa clusters into one cluster labeled DMS.[24] This leaves us with $K = 5$ clusters for the WEO vs. CE comparisons.

Row seven in Table 8 reports the t-statistic from equation (13) with p-values in rows eight and nine, the latter based on the randomization test (14). Again we fail to find evidence of significant differences in the accuracy of the WEO versus CE forecasts. Conversely, the one-year Spring and Fall WEO forecasts are significantly more accurate than the autoregressive forecast with p-values of 0.06 or 0.04, respectively. Finally, the randomization test based on country group clusters finds no statistically significant differences in the predictive accuracy of the WEO and CE GDP growth forecasts but again finds that the WEO forecasts are significantly more accurate than the autoregressive forecasts.

The corresponding p-values based on decorrelated data listed in the rows ten and eleven of Panel A lead to fewer rejections of the null of equal predictive accuracy. For example, the p-values for the decorrelated group cluster tests are 0.16 and 0.04 for the Spring and Fall WEO forecasts versus 0.06 and 0.04 for the original data. Still, using the decorrelated data, we still reject the null of equal predictive accuracy of the WEO Fall and autoregressive forecasts at the 5% significance level.

The bottom two rows in Panel A of Table 8 report p-values of Sup tests proposed in Qu et al. (2019). The first row of Sup tests (row 12) examines the null hypothesis that, across all countries, the WEO forecasts are worse than either the CE or AR(1) forecasts. Against the CE forecasts, this null is not rejected for one-year Spring and Fall forecasts and Spring current-year forecasts but it is rejected with a p-value of 0.01 for current-year Fall forecasts. This suggests that there exists at least one country for which the WEO current-year Fall GDP forecasts are significantly more accurate than the corresponding CE forecasts. In addition, there are countries whose one-year Spring and Fall WEO forecasts are significantly more accurate than the autoregressive forecast, as indicated by the p-values of 0.00 for the Sup test applied to these cases. Row thirteen reports results from testing the reverse null hypothesis, i.e., that the WEO forecasts are superior to the CE or AR(1) forecasts for all countries. This null is not rejected in a single case, indicating that the CE and AR(1) benchmarks fail

[24] Our comparison of the WEO and autoregressive inflation forecasts combines the Middle East, North Africa, Afghanistan, and Pakistan and sub-Sahara Africa groups into a single cluster, yielding a total of six clusters.

to significantly outperform the WEO forecasts for even a single country at a single forecast horizon.

These results show that we cannot reject the null that the WEO and CE forecasts of GDP growth are equally accurate for the pooled average as well as within the clusters formed along time-series or cross-sectional dimensions. However, we find strong evidence that the one-year-ahead WEO Spring and Fall forecasts of GDP growth are significantly more accurate than autoregressive forecasts for at least one period (time cluster) and one cross-sectional group (regional cluster).

To help interpret the aggregate test statistics reported in Table 8, in Table 9 we break down the comparisons of predictive accuracy by country clusters and, thus, report the test statistic (5) applied to the countries in the individual clusters. Interestingly, there is no evidence that the accuracy of the WEO and CE forecasts of GDP growth differ significantly for any of the five clusters that we use to compare these forecasts (Panel A). In contrast, the WEO forecasts are significantly more accurate than the AR(1) forecasts for Emerging and Developing Europe and Latin America and the Caribbean (Fall forecasts only).

5.3 Comparisons of Inflation Forecasts

Turning next to the inflation forecasts, the top row of Panel B in Table 8 shows that the pooled average squared-error losses of current-year ($h = 0$) Spring and Fall WEO inflation forecasts are significantly smaller than those of the CE forecasts with p-values of 0.00 and 0.01, respectively. Similar conclusions hold when we test the null of equal predictive accuracy for the individual-year or GFC time clusters (rows three through six). Interestingly, while the country cluster tests (rows seven through nine) for the current-year WEO Spring inflation forecasts continue to be significantly more accurate than their CE counterparts at the 10% level or below, current-year Fall forecasts fail to reject the null of equal predictive accuracy. Overall, though, our tests suggest a strong rejection of the null of equal predictive accuracy of the WEO and CE current-year inflation forecasts both across time and across economic groups against the alternative that the WEO forecasts are more accurate. The Sup test rejects the null that current-year Fall WEO forecasts for all countries are worse than the corresponding CE forecasts, showing the existence of at least one country for which the WEO forecasts are significantly more accurate than the CE forecasts.

For the next-year forecasts, i.e., at forecast horizons exceeding a year, the t-statistic for the pooled average or time clusters as well as the randomization p-value based on one-year clusters reject the null of equal predictive accuracy at either the 5% level ($h = 1, S$) or at the 10% level ($h = 1, F$). The t-statistic or randomization p-value based on country group clusters, as well as their decorrelated counterparts, fail to reject the null of equal predictive accuracy. The Sup test fails to reject the null that WEO forecasts is worse than CE forecasts for all country series, as well as the reverse hypothesis.

Comparisons of the average predictive accuracy of the WEO and autoregressive forecasts of inflation unequivocally lead to strong rejections of the null of equal predictive accuracy. The null is strongly rejected for the pooled average (top row) with p-values below one percent for both sets of next-year forecasts. Similar conclusions are obtained from the time-series cluster and regional cluster tests. Only the t-statistic based on country-group clusters and the decorrelated test statistics fail to reject the null of equal predictive accuracy. As shown in the MC simulation section, these tests tend to have weaker power than the other tests. Overall, this evidence suggests that there are both time periods and regions for which the WEO inflation forecasts are significantly more accurate than the autoregressive forecasts of inflation. The sup test strongly rejects rejects the null that the next-year WEO forecasts are worse than the autoregressive forecasts for all countries. Conversely, this test fails to reject the null that the WEO forecasts are better than the autoregressive forecasts for all countries.

To examine in more detail the reasons for these aggregate test results, Panel B in Table 9 reports results for the individual country clusters. We find that current-year WEO forecasts are significantly more accurate than the CE forecasts only for the group of Latin America and Caribbean, Comonwealth of Independent States and DMS countries. This finding is consistent with the notion that the IMF possesses special expertise when it comes to predicting inflation rates in less developed economies. Compared to the autoregressive inflation forecasts, we see large and significant improvements in the WEO one-year forecasts across most clusters included in our analysis, with the only exception of emerging Europe and Latin America and Carribean economies.

Figure 4 plots the single-period cross-sectional test statistics that compare the predictive accuracy of the WEO and CE inflation forecasts. Each point on the graphs is computed for the corresponding year in the sample. The values shown in Figure 4 is the t-statistics defined in equation (10) in Qu et al. (2021). Positive values indicates WEO forecasts have lower average loss than consensus economics forecasts. The dashed lines are 5% thresholds of the test statistics. Ignoring multiple test concerns, we find that the null of equal predictive accuracy is rejected for two years in our sample at the two longest horizons ($h = 1, F$ and $h = 1, F$) and for five or six years in our sample at the two shortest (current-year) horizons ($h = 0, S$ and $h = 0, F$). In contrast, the CE inflation forecasts are not significantly more accurate than the WEO forecasts in any given year.

5.4 Comparisons Across Forecast Horizons

Because we observe WEO forecasts of the same outcome (real GDP growth or inflation for country i in year t) reported at four different horizons, we can measure whether the accuracy of the forecasts improves as the time of the outcome draws closer and the forecast horizon shrinks. We would expect predictive accuracy to improve as the forecast horizon is reduced and more information about the outcome becomes available. Ordering the WEO forecasts from the longest ($h = 1, S$) to the shortest ($h = 0, F$) horizon, this means that we would expect

$$H_0^{horizon} : E[e^2_{h=0,F}] \leq E[e^2_{h=0,S}] \leq E[e^2_{h=1,F}] \leq E[e^2_{h=1,S}]. \tag{26}$$

To test if this holds, following Patton and Timmermann (2011) we consider the following four squared error loss differences:

$$\begin{aligned}
\Delta L_{i,t+h}(h = 1, S; h = 1, F) &= e^2_{i,t+1,S} - e^2_{i,t+1,F} \\
\Delta L_{i,t+h}(h = 1, F; h = 0, S) &= e^2_{i,t+1,F} - e^2_{i,t+0,S}, \\
\Delta L_{i,t+h}(h = 0, S; h = 0, F) &= e^2_{i,t+0,S} - e^2_{i,t+0,F}, \\
\Delta L_{i,t+h}(h = 1, S; h = 0, F) &= e^2_{i,t+1,S} - e^2_{i,t+0,F}
\end{aligned} \tag{27}$$

The last difference in (27) is used to measure whether, on average, the WEO current-year Fall forecasts ($h = 0, F$) are more accurate than the Spring forecasts for the same outcome computed one year previously ($h = 1, S$) and, thus accumulates any gains in accuracy over the three preceding intervals. Generally, we would expect current-year forecasts to be more accurate than next-year forecasts and a failure to reject the null of equal predictive accuracy against the alternative $E[e^2_{h=0,F}] < E[e^2_{h=1,S}]$ would suggest that the IMF does not learn any new forecast-relevant information during the 18 months leading up to and including most of the current year whose outcome is being predicted.

Identifying the points in time at which forecast-improving information arrives is economically important but also inherently difficult as such information often is not directly observed. Our tests can help address this issue as they directly reflect changes in the accuracy of forecasts of the same outcome variable computed at different points as the "event date" draws closer.

Table 10 shows test results comparing the predictive accuracy of the WEO forecasts at the four different horizons. Positive values of the test statistics (small p-values) indicate that the forecast computed at the shorter horizon is more accurate than the forecasts computed at the longer horizon.

5.4.1 GDP Growth Forecasts

We start again with the forecasts of real GDP growth (Panel A) and first compare the accuracy for the next-year spring and fall WEO forecasts ($h = 1, S$ versus $h = 1, F$) shown in column 1. Regardless of whether we use the pooled average (top row), single-year time-series cluster or group cluster test statistics, we find no evidence of a statistically significant improvement in the accuracy of the Spring versus Fall one-year-ahead WEO forecasts in these comparisons. Interestingly, however, the time-series cluster test that focuses on the performance prior to, during and after the Global Financial Crisis strongly (line 6) rejects the null that the WEO one-year-ahead Spring and Fall GDP growth forecasts are equally accurate against the alternative that the Fall forecasts are more accurate. This suggests that the IMF did improve on the accuracy of their one-year-ahead inflation forecasts between the spring and fall WEO issues either before, during or after the Global Financial Crisis.

The pooled average and region-cluster tests both reject the null of no improvement in predictive accuracy when moving from the prior-year Fall WEO to the current-year Spring WEO forecasts ($h = 1, F$ vs $h = 0, S$ in column 2), with the group-cluster tests indicating particularly strong rejections. Interestingly, the tests based on the individual-year time clusters do not reject the null in this case, suggesting that the rejection is driven by differences in the average predictive accuracy of the two sets of forecasts within one or more economic (country) groups. The Sup test strongly rejects the null that shorter-horizon forecasts are less accurate than longer-horizon forecasts for all country series. Conversely, this test fails to reject the null that shorter-horizon forecasts are more accurate than the longer-horizon forecasts.

Comparing the average predictive accuracy of the current-year WEO forecasts produced in the spring and fall ($h = 0, S$ versus $h = 0, F$), all test statistics strongly reject the null, producing p-values below 0.05 except for a single case (country group clusters) whose p-value is 0.10. This is unsurprising since a considerable amount of information relevant for forecasting current-year GDP growth gets released between April and October of the current year, the two dates at which these WEO forecasts are reported. All tests also strongly reject the null of no improvement between the points of the longest ($h = 1, S$) and shortest ($h = 0, F$) forecast horizons. This shows that, on a cumulative basis, the predictive accuracy of the WEO forecasts improves both in specific years, and for some economic groups.

Table 11 presents results broken down by individual economic clusters. There is strong evidence that the accuracy of the GDP growth forecasts improves across all individual horizons for the Advanced Economies, Emerging and Developing Europe, Developing Asia, and Latin America and the Caribbean. Conversely, we mainly see significant improvements in predictive accuracy on a cumulative basis for the MENAP, CIS and SSA economies.

5.4.2 Inflation Forecasts

Turning to the predictive accuracy of the inflation forecasts (Table 10, panel B), the pooled average and time- and regional cluster t-tests computed for the next-year Spring and Fall forecasts all generate p-values below 0.05. Forecasts of next-year inflation thus become significantly more accurate between the points where the prior-year Spring and Fall WEOs are computed. We also see see large improvements in the comparisons of the WEO next-year fall and current-year spring forecasts of inflation (second column) and when comparing current-year spring and fall inflation forecasts (third column). Unsurprisingly, this evidence of significant improvements in the accuracy of the inflation forecasts at each step of the forecast revision process translates into mostly significant rejections of the null of equal predictive accuracy for the one-year-ahead spring forecast ($h = 1, S$) and the current-year fall forecasts ($h = 0, F$). The group cluster tests based on the decorrelated data also lead to a rejection of the null of equal predictive accuracy in the comparisons of $h = 1, F$ vs $h = 0, S$ and $h = 0, S$ vis $h = 0, F$.

The results disaggregated by regional cluster (Panel B in Table 11) show evidence of broad-based and consistent improvements in the accuracy of the WEO inflation forecasts as the forecast horizon is reduced both across time and across different groups of economies, with the weakest evidence again materializing for the comparison of the two longest horizons, i.e., $h = 1, S$ versus $h = 1, F$.

Figure 5 shows plots the value of our single cross-sectional test for equal predictive accuracy of long and short WEO forecasts in a given year. The values shown in Figure 5 is the t-statistics defined in equation (10) in Qu et al. (2021). Positive values indicates shorter horizon forecasts have lower average loss than longer-horizon forecasts. The dashed lines are 5% thresholds of the test statistics. For the $h = 1, F$ versus $h = 1, S$ comparison, the cross-sectional test statistic is positive for almost all years in our sample and statistically significant in six years. Interestingly, this test is significantly negative test in 2009. Though rare, this can happen when a sharp shock leads to a reversal in the trends assumed by the forecast computed at the shorter horizon. For all other horizon comparisons we find that the single cross-section test statistics tend to be positive and highly significant in most years, consistent with improvements to predictive accuracy as the forecast horizon is shrunk and the event window is shortened. The Sup test strongly rejects the null that shorter-horizon forecasts are less accurate than longer-horizon forecaster for all countries except for $h = 1, S$ versus $h = 1, F$. Conversely, the Sup test does not reject the reverse null hypothesis that shorter-horizon forecasts are more accurate than longer-horizon forecasts for all countries.

We conclude from these results that improvements in the accuracy of the WEO inflation forecasts as the target date draws closer are more widespread across time, regions and forecast horizons than the improvements observed for the WEO forecasts of real GDP growth. The strong improvements in predictive accuracy observed in the next-year Fall versus Spring forecasts suggest that forecast-relevant information arrives further back in time for the inflation process than for GDP growth and that the IMF incorporates this information to improve their forecasts.

6 Conclusion

This paper develops new methods for testing the null of equal predictive accuracy of pairs of forecasts in the context of panel data in which we observe time series of forecasts for outcomes of multiple units. Such data structures allow us to compare the (relative) performance of alternative forecasts in a way that exploits both the time series and cross-sectional dimensions.

Our paper undertakes an extensive set of Monte Carlo simulations from which we conclude the following. First, our time-block t-test and permutation tests have the right size across a wide variety of data generating processes displaying features such as serial correlation, spatial correlation, and factor structure in clusters or in the full set of forecast errors. The grouped t-test and permutation test also work well in settings with serial correlation in forecast errors, though tend to become oversized in the presence of strong cross-sectional correlation in forecast errors. For this latter case, we propose simple decorrelated group and permultation tests which are properly sized. Among the tests of Akgun et al. (2023), we find that serial correlation in forecast errors can lead to serious size distortions. Unsurprisingly, their simplest tests that ignore cross-sectional correlation in loss differentials tend to be strongly oversized when such dependencies exist. However, their more sophisticated tests that account for cross-sectional dependencies in loss differentials (S3 and C3) perform much better with reduced size distortions. Overall, the Monte Carlo simulations suggest that our time-block t-test and permutation tests provide robust inference for testing the null of equal predictive accuracy in a variety of scenarios that allow for autocorrelation and cross-sectional dependencies in forecast errors.

We illustrate our tests in an empirical analysis that compares the accuracy of the World Economic Outlook forecasts reported by the IMF to forecasts from a private organization (Consensus Economics) as well as forecasts generated by a simple autoregressive model. Our new tests identify important differences in predictive accuracy and have the ability to pinpoint for which groups of countries or which periods in time one forecast is more accurate than other forecasts.

References

Akgun, O., Pirotte, A., Urga, G., and Yang, Z. (2022). Equal predictive ability tests for panel data with an application to oecd and imf forecasts. *arXiv preprint arXiv:2003.02803*.

Baltagi, B. H. (2013). Panel data forecasting. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2, part B, pages 995–1024. Elsevier.

Canay, I. A., Romano, J. P., and Shaikh, A. M. (2017). Randomization tests under an approximate symmetry assumption. *Econometrica*, 85(3):1013–1030.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214.

Clark, T. E. and McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of econometrics*, 105(1):85–110.

Davies, A. and Lahiri, K. (1995). A new framework for analyzing survey forecasts using three-dimensional panel data. *Journal of Econometrics*, 68(1):205–227.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, pages 253–263.

Elliott, G., Komunjer, I., and Timmermann, A. (2005). Estimation and testing of forecast rationality under flexible loss. *Review of Economic Studies*, 72(4):1107–1125.

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

Hansen, P. R. and Timmermann, A. (2015). Equivalence between out-of-sample forecast comparisons and wald statistics. *Econometrica*, 83(6):2485–2505.

Ibragimov, R. and Müller, U. K. (2010). t-Statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics*, 28(4):453–468.

Ibragimov, R. and Müller, U. K. (2016). Inference with few heterogeneous clusters. *Review of Economics and Statistics*, 98(1):83–96.

Keane, M. P. and Runkle, D. E. (1990). Testing the rationality of price forecasts: New evidence from panel data. *American Economic Review*, 80(4):714–735.

Loungani, P. (2001). How accurate are private sector forecasts? Cross-country evidence from consensus forecasts of output growth. *International journal of forecasting*, 17(3):419–432.

McCracken, M. W. (2007). Asymptotics for out of sample tests of granger causality. *Journal of Econometrics*, 140(2):719–752.

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

Patton, A. J. and Timmermann, A. (2010). Why do forecasters disagree? lessons from the term structure of cross-sectional dispersion. *Journal of Monetary Economics*, 57(7):803–820.

Patton, A. J. and Timmermann, A. (2011). Predictability of output growth and inflation: A multi-horizon survey approach. *Journal of Business & Economic Statistics*, 29(3):397–410.

Patton, A. J. and Timmermann, A. (2012). Forecast rationality tests based on multi-horizon bounds. *Journal of Business & Economic Statistics*, 30(1):1–40.

Qu, R., Timmermann, A., and Zhu, Y. (2019). Do any economists have superior forecasting skills? *Working paper.*

Qu, R., Timmermann, A., and Zhu, Y. (2021). Comparing forecasting performance in cross-sections. *Journal of Econometrics*, forthcoming.

Timmermann, A. (2007). An evaluation of the world economic outlook forecasts. *IMF Staff Papers*, 54(1):1–33.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer Science & Business Media.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica: Journal of the Econometric Society*, pages 1067–1084.

White, H. (2014). *Asymptotic theory for econometricians.* Academic press.

Appendix: proof of Theorems 3 and 5

Proof of Theorem 3. Let $b = (b_1, ..., b_K)' \in \mathbb{R}^K$. For $\xi = (\xi_1, ..., \xi_K)' \in \{-1, 1\}^K$, define $m(b, \xi) = \left|\sum_{j=1}^K \xi_j b_j\right|$. For any integer $j \in \{1, ..., 2^K\}$, define $m^{(j)}(b)$ to be the j-th largest number (counting repeated values) in $\{m(b,\xi)\}_{\xi\in\{-1,1\}^K}$, i.e., $m^{(1)}(b) \leq m^{(2)}(b) \leq \cdots \leq m^{(2^K)}(b)$. Define $k_* = \left\lceil 2^K(1-\alpha)\right\rceil$, where $\lceil t\rceil$ denotes the smallest integer no smaller than t.

We observe that $2^{-K}\sum_{\xi_1,...,\xi_K\in\{-1,1\}} \mathbf{1}\left\{\left|\sum_{j=1}^K \xi_j b_j\right| > \left|\sum_{j=1}^K b_j\right|\right\} > \alpha$ if and only $\sum_{\xi_1,...,\xi_K\in\{-1,1\}} \mathbf{1}\left\{\left|\sum_{j=1}^K \xi_j b_j\right| \leq \left|\sum_{j=1}^K b_j\right|\right\} < 2^K(1-\alpha)$, which means that $\left|\sum_{j=1}^K b_j\right| < m^{(k_*)}(b)$.

Define the function $h : \mathbb{R}^K \mapsto \mathbb{R}$ by $h(b) = \sum_{j=1}^K b_j - m^{(k_*)}(b)$. Therefore, $\hat{p}_R > \alpha$ if and only if $h(R_{(n)}) > 0$. By Assumption 1 and the null hypothesis, $R_{(n)} \xrightarrow{d} R \sim N(0, \Omega)$.

Let $B = \{r \in \mathbb{R}^K : \ h(r) > 0\}$. Notice that $P(R \in \partial B) = 0$ since $R \sim N(0, \Omega)$, where ∂B denotes the boundary of B in the usual topology on $\mathbb{R}^K$. By Theorem 1.3.4 of van der Vaart and Wellner (1996),

$$P(h(R_{(n)}) > 0) \to P(h(R) > 0). \tag{28}$$

Let $\xi \odot R$ denote the entrywise multiplication, i.e., $\xi \odot R = (\xi_1 R_1, ..., \xi_K R_K)'$. We observe that $m^{(k_*)}(R) = m^{(k_*)}(\xi \odot R)$; $\{m(R, z)\}_{z\in\{-1,1\}^K}$ and $\{m(\xi \odot R, z)\}_{z\in\{-1,1\}^K}$ are the same numbers in a different order so the ranked version is exactly the same. Notice that

$$\begin{aligned}\sum_{\xi=(\xi_1,...,\xi_n)\in\{-1,1\}^K} \mathbf{1}\{h(\xi\odot R) > 0\} &= \sum_{\xi\in\{-1,1\}^K} \mathbf{1}\left\{\sum_{j=1}^K \xi_j R_j - m^{(k_*)}(\xi \odot R) > 0\right\} \\ &= \sum_{\xi\in\{-1,1\}^K} \mathbf{1}\left\{\sum_{j=1}^K \xi_j R_j - m^{(k_*)}(R) > 0\right\}.\end{aligned}$$

By definition, $\sum_{\xi\in\{-1,1\}^K} \mathbf{1}\left\{\sum_{j=1}^K \xi_j R_j > m^{(k_*)}(R)\right\} \leq 2^K - k_*$. We notice that $m(R, \xi) = \left|\sum_{j=1}^K \xi_j b_j\right| = \left|\sum_{j:\ \xi_j=1} b_j - \sum_{l:\ \xi_l=-1} b_l\right|$. Therefore, for a generic R, $\{m(R, \xi)\}_{\xi\in\{-1,1\}^K}$ have repeated numbers and each distinct number is repeated exactly twice.

It follows that $\sum_{\xi\in\{-1,1\}^K} \mathbf{1}\left\{\sum_{j=1}^K \xi_j R_j > m^{(k_*)}(R)\right\} \geq 2^K - (k_* + 2)$. Therefore,

$$-2 \leq \sum_{\xi\in\{-1,1\}^K} \mathbf{1}\left\{\sum_{j=1}^K \xi_j R_j > m^{(k_*)}(R)\right\} - (2^K - k_*) \leq 0.$$

Since $\xi \odot R$ and R have the same distribution (due to the diagonality of Ω), we have $P(h(\xi \odot R) > 0) = P(h(R) > 0)$ for any ξ. Taking expectations on the above display, we obtain $-2 \leq 2^K \cdot P(h(R) > 0) - (2^K - k_*) \leq 0$, which means that $-2^{1-K} \leq P(h(R) > 0) - (1 - 2^{-K}k_*) \leq 0$ and thus

$$\left|P(h(R) > 0) - (1 - 2^{-K}k_*)\right| \leq 2^{1-K}. \tag{29}$$

By (28) and (29), we have that $\limsup_{n\to\infty}\left|P(h(R_{(n)}) > 0) - (1 - 2^{-K}k_*)\right| \leq 2^{1-K}$. Since $k_* = \left\lceil 2^K(1-\alpha)\right\rceil$, we have that $2^K(1-\alpha) \leq k_* \leq 2^K(1-\alpha) + 1$, which means that $|1 - 2^{-K}k_* - \alpha| \leq 2^{-K}$. Hence,

$$\limsup_{n\to\infty}\left|P(h(R_{(n)}) > 0) - \alpha\right| \leq 2^{1-K} + 2^{-K} = (3/2) \times 2^{-K}.$$

The proof is complete. □

Proof of Theorem 5. The argument is analogous to the proof of Theorem 3. □

Table 1: Size of tests under DGP 1.1

	$K = 5$		$K = 10$		$K = 5$		$K = 10$	
T	$N = 50$	$N = 100$	$N = 50$	$N = 100$	$N = 50$	$N = 100$	$N = 50$	$N = 100$
	Panel A: Time-block t-test				**Panel B: Time-block permutation test**			
50	0.049	0.05	0.056	0.058	0.068	0.063	0.052	0.062
100	0.056	0.065	0.052	0.065	0.07	0.085	0.055	0.061
500	0.052	0.053	0.062	0.056	0.06	0.068	0.061	0.054
1000	0.047	0.04	0.048	0.05	0.069	0.062	0.048	0.048
	Panel C: Group t-test				**Panel D: Group permutation test**			
50	0.053	0.049	0.056	0.05	0.055	0.062	0.057	0.053
100	0.057	0.063	0.049	0.049	0.06	0.07	0.048	0.05
500	0.055	0.056	0.042	0.058	0.065	0.07	0.042	0.06
1000	0.036	0.054	0.05	0.042	0.045	0.067	0.051	0.045
	Panel E: Decorrelated group t-test				**Panel F: Decorrelated group permutation test**			
50	0.05	0.051	0.056	0.042	0.057	0.062	0.053	0.043
100	0.056	0.056	0.051	0.054	0.06	0.07	0.053	0.058
500	0.052	0.057	0.045	0.059	0.066	0.069	0.047	0.063
1000	0.043	0.052	0.054	0.046	0.049	0.074	0.057	0.049

	$N = 50$	$N = 100$
	Panel G: Single cross-section test	
1	0.050	0.057

	$N = 50$	$N = 100$	$N = 50$	$N = 100$
	Panel H: S1		**Panel I: S3**	
50	0.088	0.108	0.098	0.116
100	0.1	0.094	0.104	0.102
500	0.065	0.081	0.071	0.08
1000	0.07	0.061	0.068	0.061

	$K = 5$		$K = 10$		$K = 5$		$K = 10$	
T	$N = 50$	$N = 100$	$N = 50$	$N = 100$	$N = 50$	$N = 100$	$N = 50$	$N = 100$
	Panel J: C1				**Panel K: C3**			
50	0.151	0.177	0.216	0.228	0.257	0.305	0.549	0.568
100	0.148	0.154	0.213	0.189	0.203	0.193	0.372	0.349
500	0.101	0.094	0.124	0.123	0.112	0.111	0.178	0.167
1000	0.097	0.075	0.086	0.089	0.102	0.08	0.108	0.119

The size of tests is estimated with p-value rejection threshold, $\alpha = 0.05$.

Table 2: Size of tests under DGP 1.2

	$K=5$		$K=10$		$K=5$		$K=10$	
T	$N=50$	$N=100$	$N=50$	$N=100$	$N=50$	$N=100$	$N=50$	$N=100$
	Panel A: Time-block t-test				**Panel B: Time-block permutation test**			
50	0.055	0.071	0.106	0.118	0.084	0.079	0.109	0.12
100	0.056	0.045	0.082	0.068	0.066	0.063	0.083	0.068
500	0.061	0.042	0.051	0.045	0.067	0.063	0.056	0.046
1000	0.056	0.034	0.056	0.055	0.064	0.056	0.06	0.058
	Panel C: Group t-test				**Panel D: Group permutation test**			
50	0.043	0.056	0.049	0.047	0.052	0.066	0.051	0.049
100	0.05	0.052	0.058	0.048	0.068	0.062	0.058	0.051
500	0.042	0.04	0.061	0.045	0.06	0.051	0.055	0.048
1000	0.06	0.048	0.045	0.048	0.069	0.071	0.044	0.046
	Panel E: Decorrelated group t-test				**Panel F: Decorrelated group permutation test**			
50	0.046	0.055	0.056	0.048	0.064	0.068	0.056	0.049
100	0.051	0.052	0.053	0.049	0.065	0.055	0.055	0.051
500	0.041	0.04	0.06	0.047	0.061	0.049	0.06	0.051
1000	0.06	0.046	0.039	0.048	0.068	0.062	0.043	0.047

	$N=50$	$N=100$
	Panel G: Single cross-section test	
1	0.066	0.048

	$N=50$	$N=100$	$N=50$	$N=100$
	Panel H: S1		**Panel I: S3**	
50	0.237	0.27	0.254	0.276
100	0.244	0.235	0.259	0.239
500	0.171	0.145	0.14	0.15
1000	0.114	0.124	0.136	0.131

	$K=5$		$K=10$		$K=5$		$K=10$	
T	$N=50$	$N=100$	$N=50$	$N=100$	$N=50$	$N=100$	$N=50$	$N=100$
	Panel J: C1				**Panel K: C3**			
50	0.592	0.598	0.801	0.822	0.714	0.721	0.961	0.968
100	0.56	0.545	0.773	0.789	0.627	0.64	0.883	0.897
500	0.289	0.283	0.446	0.476	0.315	0.308	0.546	0.549
1000	0.243	0.25	0.334	0.329	0.263	0.259	0.381	0.369

The size of tests is estimated with p-value rejection threshold, $\alpha = 0.05$.

Table 3: Size of tests under DGP 2.1

	$K = 5$		$K = 10$		$K = 5$		$K = 10$	
T	$N = 50$	$N = 100$	$N = 50$	$N = 100$	$N = 50$	$N = 100$	$N = 50$	$N = 100$
	Panel A: Time-block t-test				**Panel B: Time-block permutation test**			
50	0.049	0.054	0.064	0.052	0.072	0.068	0.061	0.055
100	0.049	0.047	0.048	0.052	0.062	0.062	0.048	0.057
500	0.044	0.047	0.055	0.052	0.057	0.058	0.053	0.056
1000	0.048	0.052	0.048	0.053	0.054	0.066	0.049	0.055
	Panel C: Group t-test				**Panel D: Group permutation test**			
50	0.057	0.049	0.043	0.041	0.069	0.072	0.047	0.045
100	0.055	0.045	0.047	0.055	0.067	0.053	0.046	0.053
500	0.055	0.044	0.049	0.043	0.074	0.05	0.05	0.045
1000	0.047	0.058	0.049	0.034	0.06	0.065	0.047	0.034
	Panel E: Decorrelated group t-test				**Panel F: Decorrelated group permutation test**			
50	0.062	0.051	0.06	0.042	0.066	0.066	0.059	0.042
100	0.059	0.047	0.045	0.056	0.069	0.055	0.046	0.055
500	0.057	0.043	0.05	0.044	0.078	0.051	0.051	0.045
1000	0.046	0.052	0.049	0.037	0.063	0.066	0.049	0.036
	Panel G: S1				**Panel H: S3**			
50	0.491	0.74	0.346	0.638	0.115	0.1	0.105	0.105
100	0.494	0.734	0.329	0.611	0.095	0.099	0.089	0.088
500	0.426	0.728	0.325	0.582	0.078	0.07	0.071	0.075
1000	0.415	0.697	0.315	0.601	0.06	0.062	0.068	0.062
	Panel I: C1				**Panel J: C3**			
50	0.926	0.997	0.948	1	0.249	0.234	0.514	0.537
100	0.911	0.995	0.924	1	0.173	0.204	0.331	0.334
500	0.895	0.995	0.9	1	0.084	0.105	0.161	0.146
1000	0.876	0.996	0.888	1	0.089	0.079	0.114	0.104

The size of tests is estimated with p-value rejection threshold, $\alpha = 0.05$.

Table 4: Size of tests under DGP 2.2

	$K = 5$		$K = 10$		$K = 5$		$K = 10$	
T	$N = 50$	$N = 100$	$N = 50$	$N = 100$	$N = 50$	$N = 100$	$N = 50$	$N = 100$
	Panel A: Time-block t-test				**Panel B: Time-block permutation test**			
50	0.044	0.051	0.048	0.044	0.056	0.074	0.054	0.057
100	0.049	0.051	0.055	0.048	0.058	0.074	0.061	0.054
500	0.042	0.052	0.057	0.033	0.063	0.071	0.054	0.04
1000	0.039	0.037	0.054	0.046	0.058	0.056	0.058	0.046
	Panel C: Group t-test				**Panel D: Group permutation test**			
50	0.143	0.163	0.271	0.29	0.172	0.185	0.266	0.288
100	0.151	0.161	0.286	0.298	0.177	0.178	0.292	0.298
500	0.163	0.168	0.276	0.292	0.173	0.176	0.28	0.286
1000	0.156	0.178	0.285	0.288	0.169	0.189	0.283	0.29
	Panel E: Decorrelated group t-test				**Panel F: Decorrelated group permutation test**			
50	0.052	0.06	0.053	0.048	0.065	0.062	0.049	0.047
100	0.041	0.048	0.059	0.064	0.061	0.068	0.06	0.064
500	0.052	0.046	0.042	0.049	0.061	0.061	0.05	0.054
1000	0.057	0.058	0.042	0.04	0.071	0.071	0.041	0.039
	Panel G: S1				**Panel H: S3**			
50	0.693	0.854	0.661	0.847	0.1	0.103	0.095	0.101
100	0.698	0.841	0.649	0.828	0.079	0.11	0.094	0.104
500	0.672	0.848	0.639	0.84	0.075	0.073	0.071	0.06
1000	0.673	0.845	0.66	0.812	0.07	0.079	0.069	0.065
	Panel I: C1				**Panel J: C3**			
50	0.968	0.999	0.984	1	0.228	0.239	0.505	0.502
100	0.963	0.999	0.989	1	0.178	0.188	0.324	0.337
500	0.959	0.998	0.979	1	0.091	0.1	0.135	0.14
1000	0.953	1	0.96	1	0.098	0.072	0.127	0.102

The size of tests is estimated with p-value rejection threshold, $\alpha = 0.05$.

Table 5: Size of tests under DGP 2.3

	$K = 5$		$K = 10$		$K = 5$		$K = 10$	
T	$N = 50$	$N = 100$	$N = 50$	$N = 100$	$N = 50$	$N = 100$	$N = 50$	$N = 100$
	Panel A: Time-block t-test				**Panel B: Time-block permutation test**			
50	0.055	0.045	0.065	0.057	0.063	0.062	0.068	0.052
100	0.051	0.054	0.068	0.056	0.062	0.065	0.069	0.051
500	0.046	0.055	0.064	0.05	0.065	0.068	0.066	0.048
1000	0.042	0.053	0.048	0.048	0.055	0.064	0.05	0.053
	Panel C: Group t-test				**Panel D: Group permutation test**			
50	0.048	0.042	0.055	0.047	0.06	0.06	0.054	0.043
100	0.045	0.052	0.067	0.046	0.061	0.065	0.07	0.048
500	0.037	0.046	0.067	0.052	0.05	0.054	0.071	0.052
1000	0.055	0.054	0.054	0.045	0.079	0.066	0.05	0.046
	Panel E: Decorrelated group t-test				**Panel F: Decorrelated group permutation test**			
50	0.049	0.049	0.046	0.043	0.063	0.067	0.05	0.044
100	0.046	0.054	0.065	0.046	0.063	0.068	0.066	0.052
500	0.039	0.051	0.067	0.051	0.057	0.058	0.073	0.052
1000	0.052	0.05	0.05	0.053	0.073	0.069	0.054	0.055
	Panel G: Single cross-section test							
1	0.040	0.049	0.052	0.042				
	Panel H: S1				**Panel I: S3**			
50	0.104	0.104	0.101	0.084	0.107	0.109	0.099	0.098
100	0.095	0.108	0.106	0.089	0.103	0.111	0.114	0.1
500	0.061	0.07	0.089	0.052	0.053	0.072	0.095	0.06
1000	0.056	0.065	0.072	0.059	0.057	0.068	0.072	0.059
	Panel J: C1				**Panel K: C3**			
50	0.158	0.156	0.232	0.195	0.264	0.258	0.509	0.52
100	0.163	0.154	0.194	0.184	0.208	0.198	0.342	0.336
500	0.09	0.098	0.085	0.105	0.105	0.109	0.129	0.148
1000	0.057	0.073	0.099	0.086	0.071	0.08	0.121	0.107

The size of tests is estimated with p-value rejection threshold, $\alpha = 0.05$.

Table 6: Size of tests under DGP 3

	$K = 5$					$K = 10$					$K = 5$					$K = 10$				
T	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$
	Panel A: Time-block t-test										**Panel B: Time-block permutation test**									
50	0.042	0.053	0.036	0.056	0.042	0.042	0.042	0.045	0.038	0.046	0.054	0.062	0.052	0.069	0.057	0.046	0.043	0.046	0.039	0.04
100	0.048	0.055	0.045	0.045	0.055	0.059	0.046	0.043	0.044	0.056	0.067	0.07	0.064	0.058	0.072	0.058	0.048	0.04	0.045	0.061
500	0.049	0.054	0.045	0.056	0.053	0.05	0.048	0.056	0.048	0.056	0.06	0.061	0.057	0.061	0.075	0.049	0.049	0.049	0.049	0.056
1000	0.041	0.037	0.045	0.046	0.054	0.042	0.041	0.055	0.041	0.053	0.049	0.049	0.059	0.071	0.069	0.043	0.041	0.058	0.041	0.054
	Panel C: Group t-test										**Panel D: Group permutation test**									
50	0.067	0.084	0.073	0.07	0.045	0.111	0.077	0.074	0.065	0.061	0.076	0.092	0.08	0.079	0.068	0.107	0.074	0.074	0.068	0.059
100	0.086	0.076	0.062	0.077	0.065	0.143	0.09	0.067	0.075	0.083	0.102	0.097	0.08	0.096	0.088	0.141	0.094	0.063	0.076	0.086
500	0.082	0.074	0.072	0.076	0.071	0.124	0.081	0.095	0.072	0.077	0.108	0.087	0.094	0.084	0.082	0.126	0.084	0.097	0.076	0.08
1000	0.068	0.072	0.073	0.072	0.061	0.109	0.072	0.092	0.073	0.076	0.083	0.093	0.094	0.09	0.068	0.108	0.081	0.093	0.073	0.076
	Panel E: Decorrelated group t-test										**Panel F: Decorrelated group permutation test**									
50	0.041	0.059	0.052	0.051	0.034	0.047	0.049	0.055	0.053	0.028	0.053	0.069	0.066	0.061	0.05	0.048	0.049	0.054	0.058	0.029
100	0.054	0.048	0.044	0.059	0.056	0.066	0.055	0.043	0.047	0.055	0.058	0.071	0.059	0.07	0.074	0.065	0.056	0.044	0.054	0.058
500	0.052	0.051	0.054	0.057	0.049	0.05	0.061	0.067	0.049	0.056	0.067	0.06	0.068	0.064	0.063	0.055	0.061	0.064	0.047	0.058
1000	0.037	0.048	0.059	0.052	0.039	0.041	0.043	0.057	0.039	0.053	0.054	0.069	0.066	0.064	0.052	0.044	0.045	0.054	0.046	0.051
T	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$						$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$					
	Panel G: S1										**Panel H: S3**									
50	0.121	0.112	0.097	0.104	0.114						0.052	0.064	0.057	0.063	0.05					
100	0.144	0.124	0.102	0.096	0.129						0.078	0.062	0.045	0.056	0.063					
500	0.119	0.101	0.119	0.1	0.119						0.05	0.055	0.063	0.059	0.055					
1000	0.117	0.11	0.11	0.096	0.119						0.045	0.05	0.056	0.047	0.06					
	$K = 5$					$K = 10$					$K = 5$					$K = 10$				
T	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$
	Panel I: C1										**Panel J: C3**									
50	0.106	0.11	0.111	0.127	0.13	0.085	0.132	0.13	0.133	0.165	0.102	0.123	0.135	0.129	0.12	0.295	0.263	0.295	0.274	0.266
100	0.119	0.12	0.117	0.112	0.143	0.077	0.127	0.129	0.128	0.182	0.086	0.077	0.085	0.089	0.082	0.136	0.152	0.141	0.152	0.155
500	0.101	0.101	0.091	0.105	0.139	0.051	0.109	0.114	0.132	0.165	0.059	0.057	0.051	0.065	0.065	0.080	0.08	0.071	0.077	0.094
1000	0.117	0.089	0.106	0.1	0.118	0.053	0.1	0.105	0.128	0.15	0.064	0.046	0.048	0.047	0.044	0.071	0.065	0.057	0.062	0.06

The size of tests is estimated with p-value rejection threshold, $\alpha = 0.05$.

Table 7: Size of tests under DGP 4

	$K = 5$					$K = 10$					$K = 5$					$K = 10$				
T	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$
	Panel A: Time-block t-test										**Panel B: Time-block permutation test**									
50	0.052	0.066	0.044	0.048	0.044	0.045	0.058	0.043	0.049	0.05	0.065	0.076	0.055	0.061	0.053	0.05	0.061	0.046	0.053	0.048
100	0.058	0.043	0.044	0.043	0.053	0.049	0.048	0.046	0.036	0.054	0.073	0.064	0.065	0.059	0.069	0.051	0.049	0.049	0.037	0.055
500	0.051	0.049	0.053	0.049	0.055	0.063	0.055	0.057	0.043	0.05	0.073	0.059	0.061	0.058	0.073	0.067	0.055	0.058	0.048	0.053
1000	0.053	0.048	0.05	0.049	0.06	0.062	0.048	0.054	0.049	0.047	0.055	0.064	0.063	0.062	0.077	0.062	0.052	0.059	0.048	0.046
	Panel C: Group t-test										**Panel D: Group permutation test**									
50	0.54	0.637	0.678	0.732	0.794	0.646	0.704	0.719	0.777	0.822	0.554	0.652	0.681	0.747	0.808	0.636	0.703	0.72	0.776	0.817
100	0.544	0.617	0.683	0.758	0.788	0.644	0.689	0.733	0.787	0.814	0.544	0.616	0.676	0.752	0.795	0.648	0.69	0.732	0.789	0.817
500	0.541	0.61	0.678	0.765	0.803	0.622	0.672	0.712	0.791	0.82	0.544	0.622	0.676	0.775	0.802	0.626	0.672	0.716	0.793	0.82
1000	0.526	0.64	0.67	0.741	0.805	0.651	0.699	0.737	0.771	0.828	0.543	0.641	0.682	0.739	0.798	0.648	0.701	0.734	0.777	0.828
	Panel E: Decorrelated group t-test										**Panel F: Decorrelated group permutation test**									
50	0.048	0.048	0.054	0.06	0.05	0.046	0.05	0.047	0.058	0.058	0.063	0.064	0.071	0.073	0.064	0.049	0.052	0.05	0.059	0.062
100	0.033	0.04	0.046	0.045	0.052	0.048	0.041	0.043	0.043	0.058	0.05	0.053	0.061	0.064	0.064	0.051	0.047	0.043	0.043	0.056
500	0.043	0.055	0.05	0.058	0.051	0.044	0.045	0.047	0.069	0.052	0.055	0.067	0.062	0.063	0.056	0.043	0.05	0.046	0.069	0.054
1000	0.051	0.037	0.054	0.053	0.046	0.055	0.034	0.049	0.044	0.043	0.066	0.043	0.072	0.053	0.052	0.059	0.033	0.049	0.046	0.047

T	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$
	Panel G: S1					**Panel H: S3**				
50	0.466	0.589	0.653	0.728	0.801	0.07	0.069	0.059	0.061	0.068
100	0.47	0.583	0.659	0.742	0.791	0.066	0.046	0.039	0.057	0.06
500	0.45	0.569	0.636	0.744	0.794	0.065	0.057	0.049	0.045	0.058
1000	0.464	0.594	0.648	0.713	0.81	0.054	0.049	0.058	0.046	0.049

	$K = 5$					$K = 10$					$K = 5$					$K = 10$				
T	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$	$N = 10$	$N = 20$	$N = 30$	$N = 50$	$N = 100$
	Panel I: C1										**Panel J: C3**									
50	0.282	0.409	0.481	0.6	0.712	0.168	0.322	0.399	0.519	0.656	0.122	0.126	0.129	0.141	0.117	0.262	0.275	0.266	0.293	0.274
100	0.262	0.401	0.512	0.614	0.681	0.155	0.303	0.411	0.526	0.614	0.101	0.082	0.091	0.073	0.095	0.149	0.153	0.161	0.144	0.156
500	0.252	0.376	0.475	0.606	0.706	0.156	0.273	0.4	0.525	0.651	0.056	0.06	0.053	0.065	0.048	0.067	0.074	0.068	0.089	0.08
1000	0.249	0.418	0.506	0.582	0.722	0.148	0.297	0.421	0.507	0.665	0.054	0.062	0.067	0.059	0.054	0.067	0.068	0.069	0.072	0.057

The size of tests is estimated with p-value rejection threshold, $\alpha = 0.05$.

Table 8: Tests of Equal Predictive Accuracy

Panel A: GDP Growth	CE				AR	
	h=1, S	h=1, F	h=0, S	h=0, F	h=1, S	h=1, F
t-stat pooled average	0.98	0.49	0.93	0.00	1.45	2.45
p-value	(0.33)	(0.62)	(0.35)	(1.00)	(0.15)	(0.01)
t-stat time clusters	0.70	0.50	1.00	0.00	1.40	2.11
p-value	(0.49)	(0.62)	(0.33)	(1.00)	(0.18)	(0.05)
Randomization p-value, 1-year clusters	(0.50)	(0.71)	(0.34)	(1.00)	(0.20)	(0.01)
Randomization p-value, GFC clusters	(0.25)	(0.50)	(0.25)	(0.75)	(0.00)	(0.00)
t-stat group clusters	1.14	0.57	1.64	0.26	2.37	2.65
p-value	(0.32)	(0.60)	(0.18)	(0.81)	(0.06)	(0.04)
Randomization p-value group clusters	(0.31)	(0.63)	(0.12)	(0.81)	(0.00)	(0.00)
p-value decorrelated group clusters	(0.84)	(0.54)	(0.48)	(0.88)	(0.16)	(0.04)
Randomization p-value decorrelated group clusters	(0.87)	(0.44)	(0.50)	(0.81)	(0.19)	(0.05)
Sup-test, null: WEO is worse than benchmarks	(0.40)	(0.26)	(0.10)	(0.01)	(0.00)	(0.00)
Sup-test, null: WEO is better than benchmarks	(0.47)	(0.83)	(0.15)	(0.21)	(0.09)	(0.64)
Panel B: Inflation						
t-stat pooled average	2.08	1.83	3.66	2.79	6.56	6.63
p-value	(0.04)	(0.07)	(0.00)	(0.01)	(0.00)	(0.00)
t-stat time clusters	2.21	1.83	3.69	3.19	8.16	8.24
p-value	(0.04)	(0.08)	(0.00)	(0.00)	(0.00)	(0.00)
Randomization p-value, 1-year clusters	(0.03)	(0.07)	(0.00)	(0.00)	(0.00)	(0.00)
Randomization p-value, GFC clusters	(0.50)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
t-stat group clusters	1.05	0.90	2.28	1.40	1.98	1.98
p-value	(0.35)	(0.42)	(0.08)	(0.23)	(0.10)	(0.10)
Randomization p-value group clusters	(0.31)	(0.56)	(0.00)	(0.25)	(0.03)	(0.03)
p-value decorrelated group cluster	(0.70)	(0.66)	(0.32)	(0.18)	(0.47)	(0.44)
Randomization p-value decorrelated group clusters	(0.88)	(0.69)	(0.31)	(0.19)	(0.47)	(0.40)
Sup-test, null: WEO is worse than benchmarks	(0.32)	(0.52)	(0.09)	(0.00)	(0.00)	(0.00)
Sup-test, null: WEO is better than benchmarks	(0.40)	(0.13)	(0.69)	(0.08)	(0.23)	(0.97)

Panel A uses real GDP growth forecasts, while Panel B uses inflation data. Positive values of the t-stats indicate that the WEO forecasts are more accurate than the counterparts, while negative values suggest the reverse. All p-values in the empirical exercise are based on two-sided tests (with the exception of Sup tests whose null hypotheses are one-sided), and are reported in brackets. In each panel, the first row reports the t-stat for the null of equal predictive accuracy for the pooled average, averaging both cross-sectionally and across time. The second row reports the associated p-value for this test. Rows three through five report the outcomes of tests of equal predictive accuracy during each year in our sample using either 26 ($h = 1$) or 27 ($h = 0$) time clusters. Row six uses three time clusters centered around the time of the Global Financial Crisis (2007-2009), namely 1995-2006, 2007-2009, and 2010-2019. These tests are all based on cross-sectional average forecasting performance. Rows three and four use the Ibragimov-Muller (2010) cluster test, while rows five and six are based on the randomization test of Canay, Romano and Shaikh (2017). Similarly, Rows seven through nine report the outcomes of tests of equal predictive accuracy for clusters of countries with similar characteristics and thus average both across time and across the countries within each cluster. Rows seven and eight use the Ibragimov-Muller (2010) cluster test, while row nine uses the randomization test of Canay, Romano and Shaikh (2017). Rows ten through eleven of each panel reports similar tests statistics for clusters of countries as in rows eight through nine, while decorrelating country-cluster wise average loss differentials. Row twelve and thirteen respectively reports p-values of the null that all WEO forecasts are no better than benchmarks for all country series and its reverse null hypothesis.

Table 9: Tests of Equal Predictive Accuracy Across Economic Groupings

Panel A: GDP Growth

WEO vs. CE

	ae	eeur	lac	cis	dms
h=1, S	0.80	0.21	1.82	-0.49	-0.87
h=1, F	1.17	0.10	0.76	-0.99	-0.50
h=0, S	1.70	-0.78	1.83	0.47	-0.04
h=0, F	0.34	-0.92	-0.14	0.17	0.10

WEO vs. AR

	ae	eeur	dasia	lac	menap	cis	ssa
h=1, S	1.62	2.64	0.36	0.15	0.70	1.43	1.00
h=1, F	1.73	3.27	1.45	2.25	1.13	1.91	1.19

Panel B: Inflation

WEO vs. CE

	ae	eeur	lac	cis	dms
h=1, S	0.06	0.45	-2.21	2.00	0.62
h=1, F	-0.48	1.94	-1.47	2.26	0.22
h=0, S	-0.10	-0.51	1.70	3.50	2.56
h=0, F	-1.13	-1.38	3.26	2.86	2.63

WEO vs. AR

	ae	eeur	dasia	lac	cis	ms
h=1, S	6.49	4..65	3.60	3.50	4.06	3.66
h=1, F	7.90	4.66	3.63	3.50	4.06	2.71

The table reports the outcome of pooled-average tests of equal squared error predictive accuracy comparing the IMF World Economic Outlook (WEO) forecasts to Consensus Economics (CE, first four rows) and autoregressive (AR, rows five and six) forecasts within each country group. Positive values of the t-tests indicate that the WEO forecasts are more accurate than the CE or AR forecasts, while negative values suggest the reverse. In each panel, each row reports a t-statistic for the null of equal predictive accuracy for the pooled average within economic groupings, averaging both cross-sectionally and across time. 'ae' refers to advanced economies, 'eeur' is emerging and developing Europe, 'lac' is Latin America and Caribbean, 'cis' is Commonwealth of Independent States, 'menap' is Middle East, North Africa, Afghanistan, and Pakistan, 'dasia' is emerging and developing Asia, and 'ssa' is Sub-Sahara Africa. Finally, 'dms' combines dasia, menap, ssa while 'ms' refers to menap and ssa combined.

Table 10: Tests of Equal Predictive Accuracy Across Different Forecast Horizons

Panel A: GDP Growth				

	h = 1,S vs
	h = 1,F
		h = 1,F vs
		h = 0,S
			h = 0,S vs
			h = 0,F
				h = 1,S vs
				h = 0,F

t-stat pooled average	0.56	1.88	2.17	5.92
p-value	(0.58)	(0.06)	(0.03)	(0.00)
t-stat time clusters	0.58	1.41	2.28	5.28
p-value	(0.57)	(0.17)	(0.03)	(0.00)
Randomization p-value, 1-year clusters	(0.63)	(0.18)	(0.00)	(0.00)
Randomization p-value, GFC cluster	(0.00)	(0.00)	(0.00)	(0.00)
t-stat group clusters	−0.03	2.69	2.16	2.87
p-value	(0.98)	(0.05)	(0.10)	(0.05)
Randomization p-value group clusters	(1.00)	(0.00)	(0.00)	(0.00)
p-value decorrelated group cluster	(0.12)	(0.04)	(0.03)	(0.01)
Randomization p-value decorrelated group clusters	(0.06)	(0.00)	(0.00)	(0.00)
Sup-test, null: short-horizon forecasts are worse	(0.01)	(0.03)	(0.01)	(0.00)
Sup-test, null: long-horizon forecasts are worse	(1.00)	(1.00)	(1.00)	(1.00)
Panel B: Inflation				
t-stat pool	2.92	5.41	6.41	6.64
p-value	(0.00)	(0.00)	(0.00)	(0.00)
t-stat time cluster	2.68	6.15	7.17	7.45
p-value	(0.01)	(0.00)	(0.00)	(0.00)
Randomization p-value, 1-year cluster	(0.00)	(0.00)	(0.00)	(0.00)
Randomization p-value, GFC cluster	(0.20)	(0.07)	(0.03)	(0.07)
t-stat region cluster	2.75	3.26	2.30	2.96
p-value	(0.05)	(0.03)	(0.08)	(0.04)
Randomization p-value region cluster	(0.00)	(0.00)	(0.06)	(0.00)
p-value decorrelated group cluster	(0.32)	(0.04)	(0.01)	(0.27)
Randomization p-value decorrelated group clusters	(0.25)	(0.06)	(0.00)	(0.12)
Sup-test, null: short-horizon forecasts are worse	(0.12)	(0.00)	(0.00)	(0.00)
Sup-test, null: long-horizon forecasts are worse	(0.35)	(1.00)	(1.00)	(1.00)

Positive values of the t-stats indicate that the shorter-horizon forecasts are more accurate than longer-horizon forecasts, while negative values suggest the reverse. All p-values in the empirical exercise are based on two-sided tests (with the exception of Sup tests whose null hypotheses are one-sided), and are reported in brackets. In each panel, the first row reports the t-stat for the null of equal predictive accuracy for the pooled average, averaging both cross-sectionally and across time. The second row reports the associated p-value for this test. Rows three through five report the outcomes of tests of equal predictive accuracy during each year in our sample using either 26 ($h = 1$) or 27 ($h = 0$) time clusters. Row six uses three time clusters centered around the time of the Global Financial Crisis (2007-2009), namely 1995-2006, 2007-2009, and 2010-2019. These tests are all based on cross-sectional average forecasting performance. Rows three and four use the Ibragimov-Muller (2010) cluster test, while rows five and six are based on the randomization test of Canay, Romano and Shaikh (2017). Similarly, Rows seven through nine report the outcomes of tests of equal predictive accuracy for clusters of countries with similar characteristics and thus average both across time and across the countries within each cluster. Rows seven and eight use the Ibragimov-Muller (2010) cluster test, while row nine uses the randomization test of Canay, Romano and Shaikh (2017). Rows ten through eleven of each panel reports similar tests statistics for clusters of countries as in rows eight through nine, while decorrelating country-cluster wise average loss differentials. Row twelve and thirteen respectively reports p-values of the null that shorter-horizon forecasts are no better than longer-horizon for all country series and its reverse null hypothesis.

Table 11: Tests of Equal Predictive Accuracy Across Different Forecast Horizons: Results by economic groupings

Panel A: GDP Growth

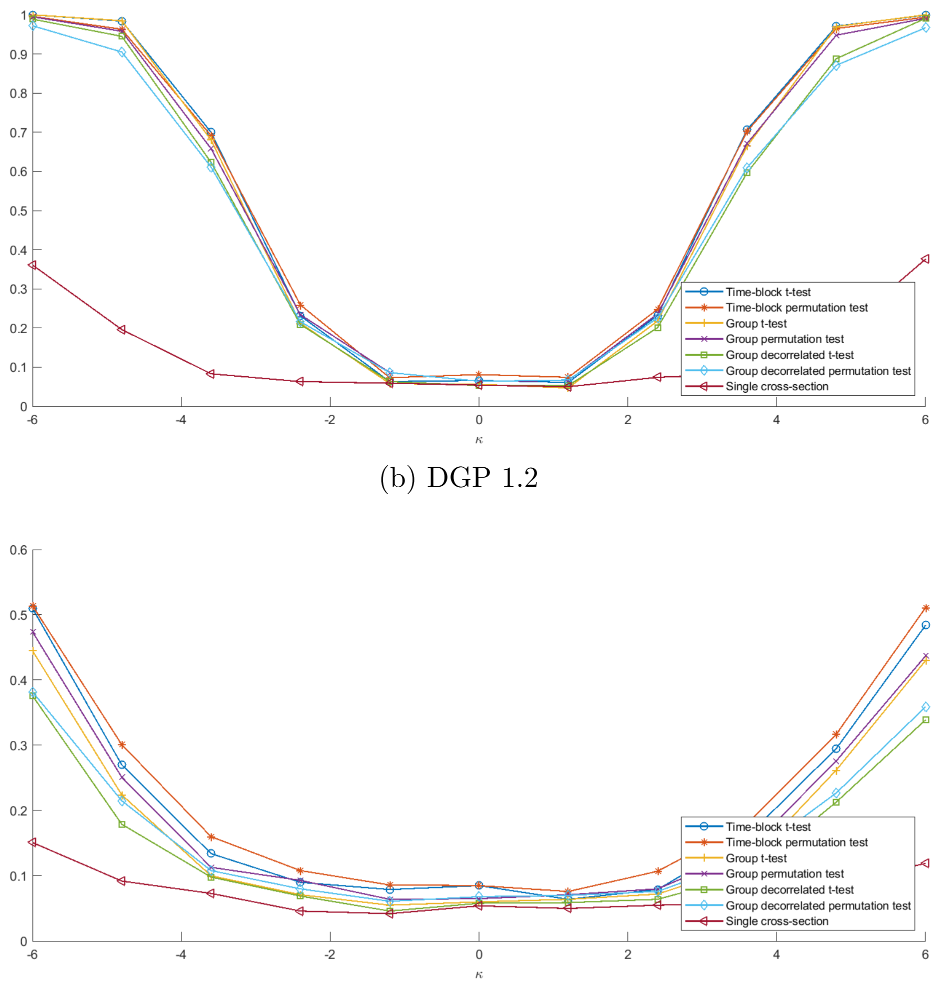
	ae	eeur	dasia	lac	menap	cis	ssa
h=1,S vs h=1,F	1.86	2.30	2.39	4.08	0.41	-0.90	1.09
h=1,F vs h=0,S	2.36	2.24	3.43	3.59	0.38	1.42	1.54
h=0,S vs h=0,F	5.49	2.02	5.61	8.49	1.35	1.87	4.16
h=1,S vs h=0,F	2.68	2.44	4.51	5.10	2.87	2.42	3.02

Panel B: Inflation

	ae	eeur	dasia	lac	cis	ms
h=1,S vs h=1,F	2.56	1.55	-0.30	3.07	1.01	2.52
h=1,F vs h=0,S	4.86	4.80	3.25	4.26	3.86	5.11
h=0,S vs h=0,F	4.95	2.07	4.08	5.34	5.38	5.52
h=1,S vs h=0,F	4.87	3.39	5.01	6.28	4.19	5.81

The table reports the outcome of pooled-average tests of equal squared error predictive accuracy comparing WEO forecasts of different horizons within each country group. Positive values of the t-tests indicate that the shorter-horizon forecasts are more accurate than the longer-horizon forecasts, while negative values suggest the reverse. In each panel, each row reports a t-statistic for the null of equal predictive accuracy for the pooled average within economic groupings, averaging both cross-sectionally and across time. The definition of country clusters are identical to the table notes of Table 9.

Figure 1: **Power curve: DGP 1 ($\alpha = 0.05$)**

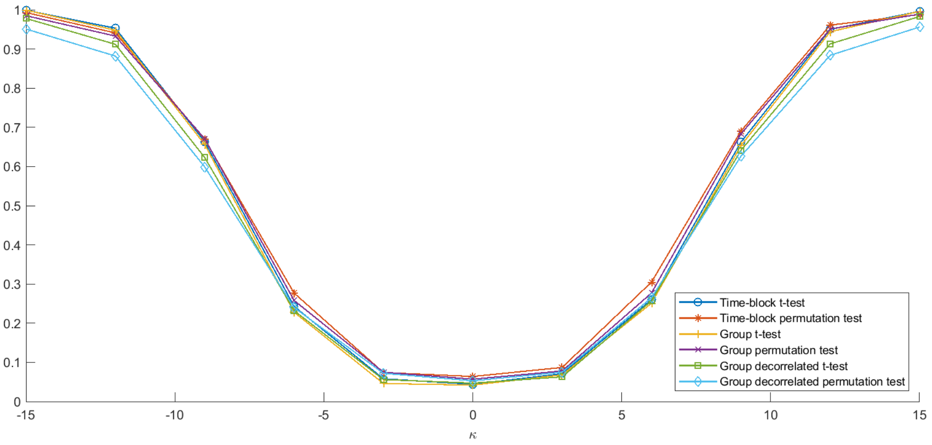


(a) DGP 1.1

(b) DGP 1.2

Figure 2: **Power curve: DGP 2 ($\alpha = 0.05$)**

(a) DGP 2.1



(b) DGP 2.2

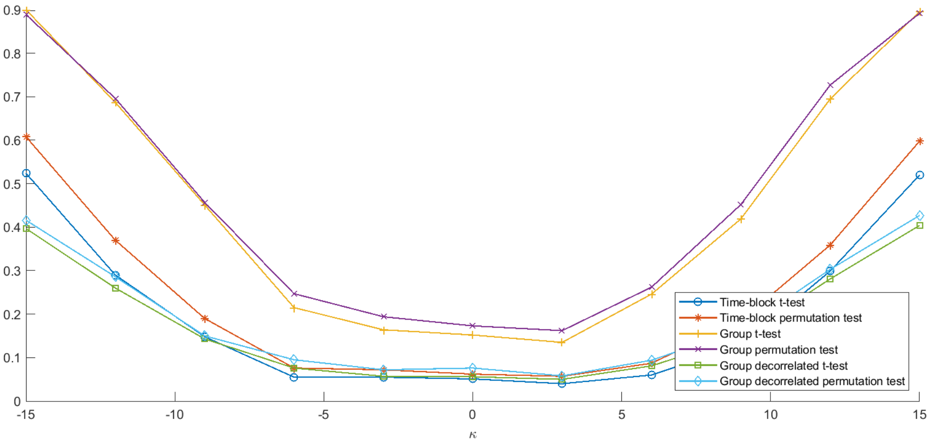


Figure 3: Forecast Horizon Used in the WEO Forecasts.

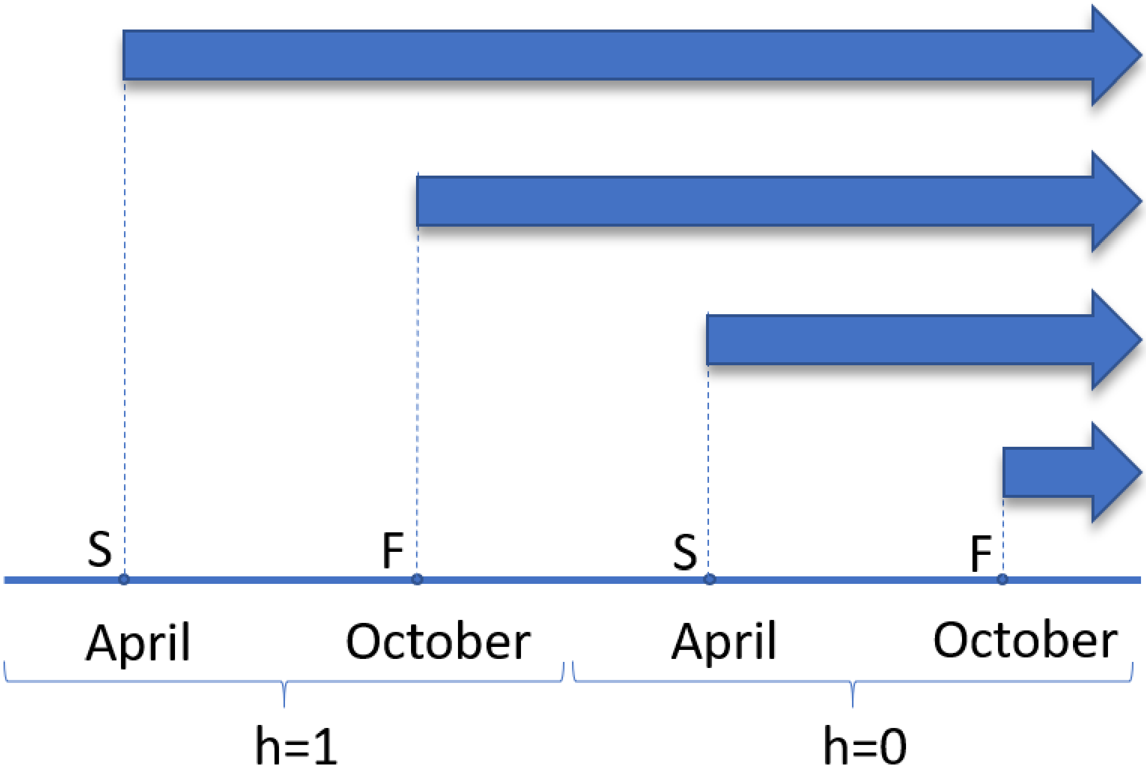


Figure 4: Single period cross-sectional comparison of WEO inflation forecasts against Consensus Economists

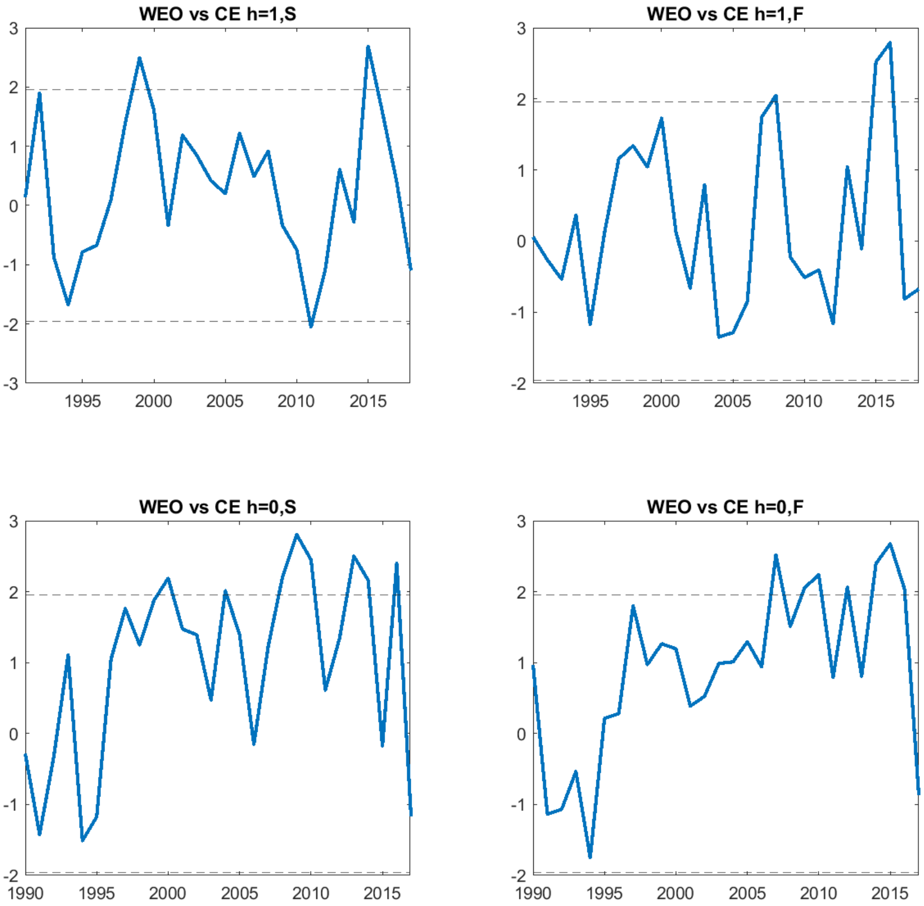
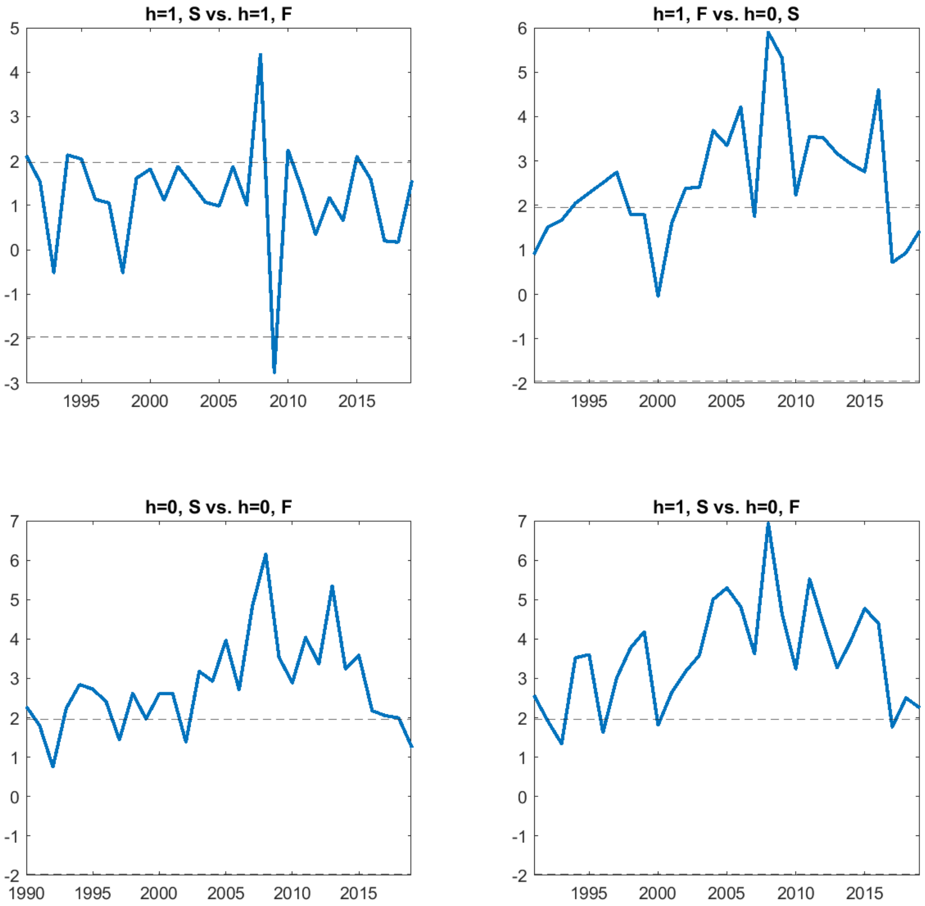


Figure 5: Single period cross-sectional comparison of WEO inflation forecasts of different horizons


h=1, S vs. h=1, F
h=1, F vs. h=0, S
h=0, S vs. h=0, F
h=1, S vs. h=0, F