DISCUSSION PAPER SERIES

DP13746

COMPARING FORECASTING PERFORMANCE WITH PANEL DATA

Allan Timmermann and Yinchu Zhu

FINANCIAL ECONOMICS

ISSN 0265-8003

COMPARING FORECASTING PERFORMANCE WITH PANEL DATA

Allan Timmermann and Yinchu Zhu

Discussion Paper DP13746
Published 21 May 2019
Submitted 13 May 2019

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

COMPARING FORECASTING PERFORMANCE WITH PANEL DATA

Abstract

Abstract This paper develops new methods for testing equal predictive accuracy in panels of forecasts that exploit information in the time series and cross-sectional dimensions of the data. Using a common factor setup, we establish conditions on cross-sectional dependencies in forecast errors which allow us to conduct inference and compare performance on a single cross-section of forecasts. We consider both unconditional tests of equal predictive accuracy as well as tests that condition on the realization of common factors and show how to decompose forecast errors into exposures to common factors and an idiosyncratic variance component. 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

JEL Classification: N/A

Keywords: Economic forecasting, panel data, GDP growth, Inflation forecasts

Allan Timmermann - atimmerm@ucsd.edu
UCSD and CEPR

Yinchu Zhu - yzhu6@uoregon.edu
University of Oregon

Comparing Forecasting Performance with Panel Data*

Allan Timmermann† Yinchu Zhu‡

April 29, 2019

Abstract

This paper develops new methods for testing equal predictive accuracy in panels of forecasts that exploit information in the time series and cross-sectional dimensions of the data. Using a common factor setup, we establish conditions on cross-sectional dependencies in forecast errors which allow us to conduct inference and compare performance on a single cross-section of forecasts. We consider both unconditional tests of equal predictive accuracy as well as tests that condition on the realization of common factors and show how to decompose forecast errors into exposures to common factors and an idiosyncratic variance component. 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; GDP growth; Inflation Forecasts

*We thank Ritong Qu for excellent research assistance with the empirical analysis.

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

‡Lundquist College of Business, University of Oregon, 1208 University St, Eugene, OR 97403, U.S.A.; yzhu6@uoregon.edu

1 Introduction

Panels of forecasts of outcomes recorded over multiple periods of time for many different variables are now ubiquitous in economics and finance. For example, the IMF produces annual forecasts of a variety of economic indicators such as real GDP growth and inflation for around 180 countries. Financial analysts predict company earnings and revenue for hundreds of firms spanning multiple industries. Credit card companies predict in real time whether a charge is fraudulent for millions of transactions on a daily basis.[1]

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

To address this point, we can instead compare two forecasts' accuracy either by averaging along the time-series dimension (e.g., years) for individual variables or clusters of variable or, alternatively, by averaging along the cross-sectional dimension for a single time period or a cluster of periods, in both cases testing whether a set of forecasts are equally accurate within each cluster. 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 the Global Financial Crisis (GFC) in 2007-2009, while the two sets of forecasts are equally accurate during more normal times. Such a finding would indicate that the model-based forecasts adapted too slowly to the unusual economic conditions during the GFC, while conversely survey participants used important forward-looking information to improve their forecasts during this period. Alternatively, we could use a longer time-series record to separately compare the predictive accuracy of two competing approaches to predict company earnings clustered

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

by industry, region, or country–in all cases clustering forecasts and outcomes by pre-specified groups so as to gain insights into the forecasts' relative accuracy for different types of firms.

Considerations such as these lead us to first study tests of equal predictive accuracy that average over time for pre-specified cross-sectional clusters of units as well as tests that average cross-sectionally using time-series observations. 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–a point we address subsequently.

Panels of economic forecasts often have a small time-series dimension but a large cross-sectional dimension. This situation arises, for example, if the outcome is measured either at a quarterly or at an annual frequency, limiting the number of time-series observations, or in the case of surveys that do not go far back in time. A paucity of time-series observations means that conventional tests of equal predictive accuracy conducted by comparing time series of forecasts for individual outcome variables lack power, particularly if performed out-of-sample.[2] By relying on time-series averages, conventional tests of equal predictive accuracy are also heavily dependent on time-series stationarity assumptions which are often questionable. An attractive alternative to such tests is to exploit the cross-sectional dimension to test whether, for a short sub-sample or even a single time period, the cross-sectional average predictive accuracy is identical across alternative forecasts. Focusing on individual time periods also facilitates faster real-time evaluation of forecasting performance than conventional methods that require calculating often lengthy time-series averages of forecasting performance.

The key challenge that arises in cross-sectional comparisons of forecasting performance–measured, e.g., by mean squared forecast errors–is the possible presence of a common component in squared error loss differentials. Common components, if sufficiently strong, can invalidate the use of a cross-sectional CLT to derive distributional results for test statistics based on cross-sectional averages computed for

[2]See Inoue and Kilian (2005) and Hansen and Timmermann (2015) for a discussion of the reduced power of out-of-sample tests.

a single time period. To address this challenge, we next develop a common factor framework that allows differences in squared forecast errors to be correlated both over time and cross-sectionally. We cover both the case with homogeneous factor loadings as well as the general case in which factor loadings are heterogeneous across units. The case with homogeneous factor loadings gives rise to tests of equal unconditionally expected predictive accuracy, while the case with heterogeneous factor loadings is better covered by a test that conditions on factor realizations.

The magnitude of any common components in squared forecast error differences turns out to contain important economic information. To the extent that large common shocks to the outcome were unanticipated by all forecasters, they will cancel out from comparisons of *relative* squared-error forecasting performance. Conversely, idiosyncratic shocks that are specific to the individual forecasters should not cancel out from squared error loss differentials. To get a better sense of the commonality and predictability of economic shocks, we therefore propose a new decomposition of the squared forecast error differential into a squared bias component, which tracks differences in two forecasts' exposures to common factors, and an idiosyncratic variance component. Since only the total squared forecast error differential is observed, we develop three approaches to estimate the common factors in forecast errors, namely (i) a cluster method that imposes homogeneity restrictions on factor loadings within clusters; (ii) a common correlated effects estimator based on Pesaran (2006); and (iii) a principal components approach. These approaches work under a variety of assumptions about the number of factors and possible structure in factor loadings and so should cover most situations encountered by applied researchers.

We illustrate our 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 sample of 180 countries covering 27 annual observations and four forecast horizons over the period from 1990 to 2016. 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. Interestingly, during our sample the idiosyncratic variance component of the IMF GDP growth forecasts has been systematically de-

clining relative to that of the Consensus Economics forecasts.

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. Inspecting the underlying forecast error components, we find that the relatively accurate IMF inflation forecasts arises from a systematically lower squared bias component compared with both the Consensus Economics and autoregressive forecasts.

Looking at the term structure of squared forecast errors across different forecast horizons, our tests allow us to identify the points in time in 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.

Our paper contributes to a large literature that compares the predictive accuracy of time-series forecasts of a single outcome. In an early contribution, Chong and Hendry (1986) propose tests of forecast encompassing. More recently, Diebold and Mariano (1995) and West (1996) develop tests for comparing the null of equal predictive accuracy. Clark and McCracken (2001) and McCracken (2007) focus on comparisons of predictive accuracy for forecasts that are generated by nested models, while accounting for the effect of recursive updating in the parameter estimates used to generate forecasts. Giacomini and White (2006) propose a test of equal predictive accuracy that accounts for the presence of non-vanishing parameter estimation error and develop methods for conditional forecast comparisons. We build on these earlier contributions, but show how the presence of a cross-sectional dimension can (i) enrich the set of economic hypotheses that can be tested; (ii) dispense with the need for restrictive assumptions on time-series stationarity for the underlying data generating process; and (iii) increase the power of existing time-series tests generalized to a panel setting.

There is also a literature on evaluating efficiency of forecasts with panel data; see, e.g., Keane and Runkle (1990), Davies and Lahiri (1995), and Patton and Timmer-

mann (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.

The outline of the paper is as follows. Section 2 introduces tests of equal predictive accuracy for panels of forecasts conducted for the pooled average (pooling both cross-sectionally and across time) or pooled separately across time clusters or cross-sectional clusters. Using these test, Section 3 conducts an empirical analysis that compares the IMF forecasts of GDP growth and inflation to the equivalent Consensus Economics and autoregressive forecasts. Section 4 introduces a common factor decomposition of forecast errors and proposes ways to compare predictive accuracy for individual cross-sections. Section 5 proposes a decomposition of the mean squared forecast errors into a squared bias and an idiosyncratic error variance component and derives statistics for testing the null that these two components are of the same magnitude across different forecasts. Section 6 concludes. Technical proofs are in an Appendix.

2 Panel Tests of Equal Predictive Accuracy

Consider a panel of data with y_{it+h} denoting the realized value of unit i at time $t+h$, where $i = 1, ..., n$ refers to the cross-sectional dimension and $t + h = 1, ..., T$ refers to the time-series dimension.[3] Further, suppose we observe a series of h-step-ahead forecasts of the outcome, y_{it+h}, generated conditional on information available to the forecaster at time t. We denote these by $\hat{y}_{it+h|t,m}$, 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 $L_{it+h|t,m} = L(y_{it+h}, \hat{y}_{it+h|t,m})$. Consistent with most empirical work, we assume that the loss is a quadratic function

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

of the forecast error, $e_{it+h,m} = y_{it+h} - \hat{y}_{it+h|t,m}$, and thus takes the form[4]

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

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 $L_{it+h|t,m}$. In particular, we do not consider the effect of estimation error on the distribution of the test statistics which we derive.[5]

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:

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

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, $t + h$. 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

[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).

m_2 for unit i at time $t+h$ as

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

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

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

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

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

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

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:

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

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

Theorem 1. *Suppose that* $\max_{1 \leq t \leq T} E|R_{t+h}|^r$ *is bounded with* $r > 2$ *and that* $\{R_{t+h}\}_{t+h=1}^{T}$ *is* α*-mixing of size* $-r/(r-2)$. *Also assume that* $\hat{\sigma}(\Delta L_{t+h|t}) = \bar{\sigma}_{n,T} + o_P(1)$ *and* $\bar{\sigma}_{n,T} > 0$ *is bounded away from zero, where* $\bar{\sigma}^2_{n,T} =$

$\text{Var}\left((nT)^{-1/2}\sum_{t=1}^{T}\sum_{i=1}^{n}\Delta L_{i,t+h|t}\right)$. *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. However, it does not require restrictions on the degree of cross-sectional dependence and in fact allows for arbitrary cross-sectional dependence in the loss differentials.

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 detect differences across clusters of time-series or cross-sectional data.

2.2 Testing Equal Predictive Accuracy for Clusters

In many situations, the relative accuracy of a set of economic forecasts can be expected to differ across time or across units 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 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 are equally accurate either across certain periods of time or for different cross-sectional groups or clusters. This subsection develops a framework for conducting such tests.

2.2.1 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 $\cup_{j=1}^{K} t_j = [h+1 : T+h]$. Denote the associated test statistics by $\{R_{t_1}, R_{t_2}, ..., R_{t_K}\}$. For example, if each cluster has equal length, q, the test statistic for the jth cluster can be computed as $R_{t_j} = q^{-1} \sum_{t=(j-1)q+h}^{jq+h-1} R_t$.[6] 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-series 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.[7] 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:

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

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 inde-

[6] 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.

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

pendent across the K clusters. 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.[8] The assumption of asymptotic normality rules out that the cross-sectional dependence in the loss differentials is so strong that it invalidates the use of a cross-sectional CLT–a point we shall be more specific about in Section 4.

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 not, 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 $H_0^{Tcluster}$ *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 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 recently proposed by Canay et al. (2017). Define the randomization p-value:

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

where $\xi_1, ..., \xi_K \in \{-1, 1\}$ are all possible combinations of the K variables $\xi_1, ..., \xi_K$,

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

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*

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

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 5 clusters, inference at the 5% significance level only rejects when the p-value is exactly zero. On the other hand, for larger values of K, Theorem 3 implies a similarity property of the randomization test.[9] However, 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 $\alpha - 2^{-K}$.

2.2.2 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

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

elements in the jth cluster, i.e., the cardinality of H_j, with $\sum_{j=1}^{K} |H_j| = n$ and define

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

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

$$H_0^{Ccluster} : \; ED_1 = ED_2 = \cdots = ED_K = 0. \tag{11}$$

This setup is equivalent to that in Section 2.2.1. 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

$$J_n^D = \frac{\sqrt{K}\bar{D}}{\sqrt{(K-1)^{-1} \sum_{j=1}^{K} (D_j - \bar{D})^2}}. \tag{12}$$

Analogous to the result for the time-series clusters, we make the following assumption:[10]

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. We discuss alternative approaches that do not require this assumption in Section 4 and, for now, focus on situations where the assumption holds. 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 $\alpha \leq 0.2$.

[10] 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 $\Delta L_{i,t+h|t}$.

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:

$$\hat{p}_D = 2^{-K} \sum_{\xi_1,\ldots,\xi_K\in\{-1,1\}} \mathbf{1}\left\{\left|\sum_{j=1}^{K} D_j\xi_j\right| > \left|\sum_{j=1}^{K} D_j\right|\right\}. \tag{13}$$

Using Assumption 2, we have

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

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

Provided that the conditions in Assumption 2 hold, this result means that we can apply the randomization test in (13) for the cross-sectional clusters.

3 Empirical Results

To illustrate the economic insights that can be gained from the test statistics introduced in section 2, 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.[11] We compare the WEO forecasts to forecasts from a private-sector organization (Consensus Economics) as well as forecasts from a simple autoregressive model.

[11] 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).

3.1 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 1, this produces a set of four forecast horizons, listed in decreasing order: $\{h = 1, S;\ h = 1, F;\ h = 0, S;\ h = 0, F\}$.[12] 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 somewhat shorter data sample. For all countries, the last outcome is recorded for 2016.

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 smaller than that covered by the WEO forecasts, restricting the cross-sectional dimension of 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. In total, we can compare the

[12] 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.

WEO country-level forecasts to 180 or 181 forecasts generated by the AR(1) model.

3.2 Comparisons of GDP Growth Forecasts

The top row in Table 1 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.[13]

First, consider the forecasts of real GDP growth (Panel A). The pooled average t-test in equation (5) is positive for the three shortest forecast horizons and negative only for the shortest horizon ($h = 0, F$). 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 (two right-most columns) show that the WEO Fall forecasts (though not the Spring forecasts) are significantly more accurate, on average, than the AR forecasts with a t-statistic of 2.25 and a p-value of 0.02.

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 clusters so that the current-year forecasts ($h = 0$) are based on $K = 27$ one-year time clusters, while next-year forecasts ($h = 1$) are based on $K = 26$ one-year clusters. Rows 3 and 4 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 0.10. The randomization test (10) reported in the fifth row leads to identical conclusions with p-values ranging from 0.38 to 0.92, 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 2010-2016.[14] The results, listed in line six for the randomization p-value,

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

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

show that we cannot reject the null that the CE and WEO forecasts of GDP growth were equally accurate both during the GFC and in more normal times. Conversely, with p-values below 0.01, there is very strong evidence that the Spring and Fall one-year-ahead 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 so shows that additional power can be gained from grouping time periods based on economic characteristics–in this case the unfolding of a major global crisis.

Finally, we cluster the country observations along a set of IMF classifications which consider geographical regions and economic development stages. Specifically, we use a partition of $K = 7$ clusters of countries, namely (i) Advanced Economies (36 countries in 2016), (ii) Emerging and Developing Europe (9), (iii) Emerging and Developing Asia (27), (iv) Latin America and the Caribbean (32), (v) Middle East, North Africa, Afghanistan, and Pakistan (21), (vi) Commonwealth of Independent States (12), and (vii) Sub-Sahara Africa (45), with 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 East, North Africa, Afghanistan, and Pakistan, and Sub-Sahara Africa clusters into one cluster labeled DMS.[15] This leaves us with five clusters for the WEO vs. CE comparisons.

Row seven in Table 2 reports the t-statistic from equation (12) with p-values in rows eight and nine, the latter for the randomization test (13). Again we fail to find evidence of significant differences in the accuracy of the WEO versus CE forecasts. Conversely, the one-year Fall WEO forecast is significantly more accurate than the autoregressive forecast with a p-value of 0.02 or 0.00, depending on the choice of test statistic.

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 Fall forecasts of GDP growth are

[15] In addition, our comparison of the WEO forecasts and the autoregressive 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.

significantly more accurate than the autoregressive forecasts for at least one period (time cluster) and one cross-sectional group (region cluster).

To assist in interpreting the aggregate test statistics reported in Table 1, in Table 2 we break down the comparisons of predictive accuracy by country clusters and, thus, compute 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, we find that 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), though not for any of the other clusters.

3.3 Comparisons of Inflation Forecasts

Turning to the inflation forecasts, the top row of Panel B in Table 1 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 0.02, 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, 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. Conversely, the WEO and CE one-year-ahead inflation forecasts appear to be of broadly similar accuracy with most of the test statistics computed for this horizon failing to reject the null of equal predictive accuracy at this horizon.[16]

Our comparison of the average predictive accuracy of the WEO and autoregressive forecasts of inflation unequivocally leads 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

[16] At the one-year horizon, only the randomization p-value method based on the GFC time blocks suggest that the WEO Fall inflation forecasts are significantly more accurate than the CE forecasts.

are obtained from the time-series cluster and regional cluster tests. This shows that there are both time periods and regions for which the WEO inflation forecasts are significantly more accurate than the autoregressive forecasts of inflation.

Turning again to the sources of the outcomes of these aggregate test results, Panel B in Table 2 reports results for the individual country clusters. We find that current-year WEO forecasts for the Latin American and Caribbean countries, the Commonwealth of Independent States, and the DMS countries are significantly more accurate than the CE forecasts. Conversely, for advanced economies and developing Europe, there is no evidence to suggest that the WEO inflation forecasts are more accurate than the CE forecasts at any of the horizons. 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 all clusters included in our analysis.

3.4 Comparisons Across Forecast Horizons

Because we observe WEO forecasts of the same outcome (real GDP growth or inflation for country i in year t) produced 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

$$H_0^{horizon} : E[e^2_{h=0,F}] \leq E[e^2_{h=0,S}] \leq E[e^2_{h=1,F}] \leq E[e^2_{h=1,S}]. \tag{14}$$

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

$$\begin{aligned}
\Delta L_{i,t+h}(h = 1, S; h = 1, F) &= e^2_{i,t+1,S} - e^2_{i,t+1,F} \\
\Delta L_{i,t+h}(h = 1, F; h = 0, S) &= e^2_{i,t+1,F} - e^2_{i,t+0,S}, \\
\Delta L_{i,t+h}(h = 0, S; h = 0, F) &= e^2_{i,t+0,S} - e^2_{i,t+0,F}, \\
\Delta L_{i,t+h}(h = 1, S; h = 0, F) &= e^2_{i,t+1,S} - e^2_{i,t+0,F}
\end{aligned} \tag{15}$$

The last difference in (15) 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 cumulates 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 may not even be 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 as the "event date" draws closer.

Table 3 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 shortest horizon is more accurate than the forecasts computed at the longer horizon. 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 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 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 the accuracy of their one-year-ahead inflation forecasts between the spring and fall WEO issues during the Global Financial Crisis.

Both the pooled average and region-cluster tests 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 $h = 0, S$), 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 groups.

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 0.08. This is unsurprising since a considerable amount of information relevant for forecasting current-year GDP growth gets released between April and October, the two dates at which these WEO forecasts are reported. All tests also very strongly reject the null of no improvement from the longest ($h = 1, S$) to the shortest ($h = 0, F$) forecast horizon showing that, on a cumulative basis, the predictive accuracy of the WEO forecasts improves both in specific years, and for some economic groups.

Table 4 mirrors Table 2 and presents results broken down by the 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, and Latin America and the Caribbean. Conversely, we only see significant improvements in predictive accuracy on a cumulative basis for the CIS and DMS economies.

Turning to the the predictive accuracy of the inflation forecasts (Table 3, 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 highly 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 results disaggregated by regional cluster (Panel B in Table 4) 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.

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. In particular, the strong improvements in predictive accuracy observed in 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 actively use this information to improve their forecasts.

4 Comparing Predictive Accuracy for Individual Cross-sections

An important limitation of formal comparisons of the (relative) accuracy of different economic forecasts is that sample sizes tend to be quite short and so the statistical power of tests based on time-series averages can be quite low. This is particularly relevant for microeconomic applications that often rely on short surveys, see, e.g., Giacomini et al. (2019) and Liu et al. (2019). In addition, inference that relies on averaging test statistics across time can be affected by non-stationarities in the underlying data generating process which may adversely affect our ability to analyze time-series averages of loss differentials. Finally, new time-series observations arrive only slowly when outcomes are measured at a monthly, quarterly, or annual frequency, reducing our ability to conduct real-time comparisons of predictive accuracy. These points highlight shortcomings of inference on predictive accuracy that is based on time-series averages.

In contrast, individual forecasting models can often be used to generate hundreds or even thousands of cross-sectional forecasts each period, as in the case of forecasts for individual customers, transactions, product categories, or firms. The presence of such data with small T and large n suggests the feasibility of comparing the accuracy of a pair of forecasts in a particular time period or over a short period of time. Conducting such tests requires, however, an understanding of the assumptions under which we can establish the distribution of cross-sectional averages underlying the test statistics. Most obviously, the loss differentials cannot be too strongly cross-sectionally dependent–otherwise a CLT will not apply to the cross-sectional test statistics.

To more accurately characterize the sources of cross-sectional dependencies in forecast errors, suppose we can decompose the forecast error of model m, $e_{i,t+h,m} = y_{i,t+h} - \hat{y}_{i,t+h|t,m}$, into a common component, f_{t+h}, with factor loadings λ_{im}, and an idiosyncratic component, $u_{it+h,m}$, so that, for $m = 1, 2$,

$$e_{i,t+h,m} = \lambda'_{im} f_{t+h} + u_{i,t+h,m}. \tag{16}$$

Under this setup, forecast errors are allowed to be affected by the same common factors, f_{t+h}, but we allow for differences in the factor loadings (λ_{im}) across units, i, and across forecasts, m. Factor loadings, λ_{im}, can be either random or fixed; we are clear on which of these assumptions we make in the analysis below.

4.1 Null Hypotheses

Under the assumed common factor structure in (16), it is not, in general, possible to use test statistics that rely on a cross-sectional CLT because of the presence of a common component that does not disappear asymptotically even as $n \to \infty$. To address this issue, we consider two approaches for testing the null of equal predictive accuracy in a single cross-section. First, we can conduct a test of an unconditional null of equal predictive accuracy. As we show below, this type of test requires that the common factor component that introduces dependence in forecast errors cancels out in the loss differentials. Second, we can condition on the factor realization and conduct a conditional test of equal predictive accuracy. This approach will be valid provided that, conditional on the realized factor, a cross-sectional CLT applies to the idiosyncratic error components. Without these assumptions, the cross-sectional averages of loss differentials will not, in general, be asymptotically normal.

First, consider testing the null that the cross-sectional average loss differential at time $t + h$, $\overline{\Delta L}_{t+h} = n^{-1} \sum_{i=1}^{n} \Delta L_{i,t+h|t}$, equals zero in expectation:

$$H_{0,t+h}^{unc} : \; E(\overline{\Delta L}_{t+h}) = 0. \tag{17}$$

While the forecasts are only expected to be equally accurate at a single point in time, $t + h$, differences in predictive accuracy at that time are hypothesized to balance out across units, $i = 1, ..., n$. This could happen, for example, because the errors of one of the forecasts have small loadings on the common factor, f_{t+h}, but a high

idiosyncratic error component, $u_{i,t+h,m}$, relative to a competing model–a point we return to in Section 5.

We can also consider testing whether two forecasts are expected to be equally accurate, at time $t+h$, *conditional* on a particular outcome of the factor realizations, f_{t+h}, as well as the factor loadings $\{\lambda_{i1}, \lambda_{i2}\}_{i=1}^{n}$:

$$H_{0,t+h}^{cond}: \; E(\overline{\Delta L}_{t+h} \mid \mathcal{F}) = 0, \tag{18}$$

where $\mathcal{F} = \sigma(f_{t+h}, \{\lambda_{i1}, \lambda_{i2}\}_{i=1}^{n})$.

The interpretation of this conditional null is clearly different from that of the unconditional null in (17) but can be of separate economic interest. For example, we can use (18) to test whether, conditional on the unusual realizations of the factors that occurred during the Global Financial Crisis, the predictive accuracy of a set of alternative forecasts was the same. Or, as the complement to this, we can test whether the forecasts were equally accurate during more "normal" years.

If, in fact, factor realizations were the main driver of differences in the predictive accuracy of a pair of forecasts, we can imagine situations in which we reject the null in (17) without rejecting (18). Conversely, two forecasts could be equally accurate "on average" in a given period because one forecast is more strongly affected by shocks to the common factors and less affected by idiosyncratic error shocks, while the reverse holds for the other forecast and the effects exactly balance out. In this case, we do not reject the null in (17), whereas the conditional null in (18) is rejected.

For now we do not discuss how we can test if factor loadings are homogeneous or heterogeneous. However, we note that the test for equal squared bias developed in the next section can be used to address this question.

4.2 Homogeneous Factor Loadings

Suppose the loadings on the common factors affecting the individual forecast errors in (16) are the same across the two forecasts so that $\lambda_{i1} = \lambda_{i2} = \lambda_i$. Assuming quadratic error loss, we then have

$$\Delta L_{i,t+h|t} = (u_{i,t+h,1}^2 - u_{i,t+h,2}^2) + 2(u_{i,t+h,1} - u_{i,t+h,2})\lambda_i' f_{t+h}. \tag{19}$$

Common unpredictable shocks that are not picked up by any of the forecasts can

be thought of as satisfying the assumption of homogeneous factor loadings since they can have a different effect on different units, but will affect the forecasts in the same way, i.e., $\lambda'_{i1} = \lambda'_{i2}$. These shocks will, therefore, cancel out from the forecast error differentials. For example, if the effects of a major event such as the Global Financial Crisis were unanticipated by both forecasts and affected them in the same amount, they cancel out from the loss differential.

Under homogeneous factor loadings, the cross-sectional dependence arising from the forecasts' exposure to the common factors, f_{t+h}, does not play an important role in deriving the asymptotics of tests of the null in (17) since $\lambda'_i f_{t+h}$ is multiplied by $(u_{i,t+h,1} - u_{i,t+h,2})$. This is assured under the following assumption which requires (conditionally) independent idiosyncratic error components as well as a Lyapounov condition:

Assumption 3. *Suppose that the loadings are homogeneous, i.e.,* $\lambda_{i1} = \lambda_{i2} = \lambda_i$ *for* $i = 1, ..., n$. *Conditional on* $\mathcal{F} = \sigma(f_{t+h}, \{\lambda_{i1}, \lambda_{i2}\}_{i=1}^n)$, $\{(u_{i,t+h,1}, u_{i,t+h,2})\}_{i=1}^n$ *is independent across* i *with mean zero and with bounded* $(4+\delta)$ *moments for some* $\delta > 0$. *Moreover,* $\min_{1\leq i\leq n} \mathrm{Var}[(u_{i,t+h,1} - u_{i,t+h,2}) \mid \mathcal{F}] \geq c$ *for some constant* $c > 0$ *and*

$$\frac{\left(\sum_{i=1}^n |\lambda'_i f_{t+h}|^{2+\delta}\right)^{1/(2+\delta)}}{\left(\sum_{i=1}^n |\lambda'_i f_{t+h}|^2\right)^{1/2}} = o_P(1).$$

To test the null of equal expected loss for the pooled average in (17), consider the test statistic

$$Q_{t+h} = \frac{n^{1/2}\overline{\Delta L}_{t+h|t}}{\sqrt{n^{-1}\sum_{i=1}^n \left(\Delta L_{i,t+h|t}\right)^2}}. \tag{20}$$

Under the assumption of pair-wise homogeneous factor loadings, (19) shows that testing the null of equal predictive accuracy in period $t+h$ amounts to testing that $E(u^2_{i,t+h,1} - u^2_{i,t+h,2}) = 0$. This is easily accomplished under Assumption 3 which ensures independence across i for $(u_{i,t+h,1}, u_{i,t+h,2})$ so that asymptotic normality can be established for Q_{t+h} in (20).[17]

Using Assumption 3, we can show the following result:

[17] Alternatively, we can test this null under assumptions of stationarity which allows us to exploit time-series variation in the factors. Assumptions of stationarity and weak serial dependence gets us back to using the test statistic in (5). Of course, assuming stationarity and calculating time-series averages in this manner means that we cannot meaningfully use (20) to address how the forecasts performed during specific periods such as the global financial crisis.

Theorem 6. *Suppose Assumption 3 holds. Then under the null of equal expected cross-sectional predictive accuracy, $H_{0,t+h}^{unc}$ in (17), we have*

$$\limsup_{n\to\infty} P\left(|Q_{t+h}| > z_{1-\alpha/2}\right) \leq \alpha,$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of a $N(0,1)$ variable.

Theorem 6 shows that homogeneous factor loadings lead to a simple test of the null of equal expected loss for the pooled average using data only on a single cross-section. Moreover, the test statistic follows a Gaussian distribution in large cross-sections.

For now, we do not go into details of how the assumption of homogeneous loadings can be tested. However, as we show in Section 5, our approach for testing the null in (17) remains valid as long as $n^{-1}\sum_{i=1}^{n}[(\lambda_{i,1}'f_{t+h})^2 - (\lambda_{i,2}'f_{t+h})^2] = 0$. Moreover, as we show below, we can test that this condition holds using the procedure from Theorem 9, Theorem 11 or the test statistic in (44). depending on how one estimates the dependence structure in the forecast errors.

4.3 Heterogeneous Factor Loadings

Next, consider the case with heterogeneous factor loadings for the forecast errors, i.e., $\lambda_{i,1} \neq \lambda_{i,2}$. For this case, the expression in (19) is generalized to

$$\begin{aligned}\Delta L_{i,t+h|t} = &\left[(\lambda_{i,1}'f_{t+h})^2 - (\lambda_{i,2}'f_{t+h})^2\right] \\ &+ \left[u_{i,t+h,1}^2 - u_{i,t+h,2}^2 + 2(\lambda_{i,1}'f_{t+h}u_{i,t+h,1} - \lambda_{i,2}'f_{t+h}u_{i,t+h,2})\right].\end{aligned} \tag{21}$$

When the factor loadings differ for the forecasts, equation (21) shows that the relative predictive accuracy in period $t+h$ contains a systematic risk component, $E\left[(\lambda_{i,1}'f_{t+h})^2 - (\lambda_{i,2}'f_{t+h})^2\right]$. Even if f_{t+h} is independent of the factor loadings, $\{(\lambda_{i,1},\lambda_{i,2})\}_{i=1}^{n}$, and these loadings are independent across i, $n^{-1/2}\sum_{i=1}^{n}\left[(\lambda_{i,1}'f_{t+h})^2 - (\lambda_{i,2}'f_{t+h})^2\right]$ is asymptotically normal only conditional on f_{t+h}. This suggests conducting a test of equal expected predictive accuracy conditional on the factor realization as is done in (18).

To see how the conditional null in (18) can be tested, again let $\mathcal{F} = \sigma(f_{t+h}, \{\lambda_{i1},\lambda_{i2}\}_{i=1}^{n})$ and assume that $E(u_{i,t+h,1} \mid \mathcal{F}) = E(u_{i,t+h,2} \mid \mathcal{F}) = 0$. De-

fine

$$\xi_{i,t+h} = \left(u^2_{i,t+h,1} - u^2_{i,t+h,2}\right) - E\left(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}\right) + 2(\lambda'_{i,1} f_{t+h} u_{i,t+h,1} - \lambda'_{i,2} f_{t+h} u_{i,+ht,2}).$$

Using equation (21), we have

$$\overline{\Delta L}_{t+h} - E(\overline{\Delta L}_{t+h} \mid \mathcal{F}) = n^{-1} \sum_{i=1}^{n} \xi_{i,t+h}. \tag{22}$$

The ideal variance estimate for the object in (22) is $n^{-1}\sum_{i=1}^n \xi^2_{i,t+h}$. However, at the unit level, we only have data on $e_{i,t+h,m}$ and hence are limited to computing $n^{-1}\sum_{i=1}^n (\Delta L_{i,t+h|t} - \overline{\Delta L}_{t+h})^2$. Consider the following test statistic

$$Q_{t+h} = \frac{\sqrt{n}\overline{\Delta L}_{t+h}}{\sqrt{n^{-1}\sum_{i=1}^n (\Delta L_{i,t+h|t} - \overline{\Delta L}_{t+h})^2}}. \tag{23}$$

To establish properties of the test statistic in (23), we need a set of regularity conditions which we summarize in the following assumption:

Assumption 4. *Conditional on* $\mathcal{F} = (f_{t+h}, \{\lambda_{i1}, \lambda_{i2}\}_{i=1}^n)$, $\{(u_{i,t+h,1}, u_{i,t+h,2})\}_{i=1}^n$ *is independent across* i *with mean zero and bounded* $(4+\delta)$ *moments for some* $\delta > 0$. *Moreover,* $\min_{1\le i\le n} \mathrm{Var}[\xi_{i,t+h} \mid \mathcal{F}] \ge c$ *for some constant* $c > 0$.

Using this assumption, we can now test the null $n^{-1}\sum_{i=1}^n E\left(\Delta L_{i,t+h|t} \mid \mathcal{F}\right) = 0$ or, equivalently, establish a confidence interval for $E(\overline{\Delta L}_{t+h} \mid \mathcal{F})$:

Theorem 7. *Suppose Assumption 4 holds. Then, under the null hypothesis in (18), the following result holds for the test statistic in (20)*

$$\limsup_{n\to\infty} P\left(|Q_{t+h}| > z_{1-\alpha/2}\right) \le \alpha,$$

where $z_{1-\alpha/2}$ *is the* $(1-\alpha/2)$ *quantile of a* $N(0,1)$ *variable.*

Results based on the test statistic in (23) can be interpreted in two ways. First, as explained above, they can be viewed as tests of the null hypothesis $H_0: E(\overline{\Delta L}_{t+h} \mid \mathcal{F}) = 0$. Second, if we assume that the factor loadings $\{(\lambda_{i,1}, \lambda_{i,2})\}_{i=1}^n$ are random, independent across i and independent of f_{t+h}, we can use the test statistic in (23) to test $H_0: E(\overline{\Delta L}_{t+h} \mid f_{t+h}) = 0$ instead of testing $E(\overline{\Delta L}_{t+h} \mid \mathcal{F}) = 0$, with

the latter also conditioning on the factor loadings. Testing the former hypothesis ($E(\overline{\Delta L}_{t+h} \mid f_{t+h}) = 0$) rather than the latter–and, hence not conditioning on the factor loadings ($\lambda_{i,1}$ and $\lambda_{i,2}$)–introduces an additional term in the numerator of (23)

$$E(\overline{\Delta L}_{t+h} \mid \mathcal{F}) - E(\overline{\Delta L}_{t+h|t} \mid f_{t+h}) \\ = f'_{t+h} \left(n^{-1} \sum_{i=1}^{n} \left[\lambda_{i,1}\lambda'_{i,1} - \lambda_{i,2}\lambda'_{i,2} - E(\lambda_{i,1}\lambda'_{i,1} - \lambda_{i,2}\lambda'_{i,2}) \right] \right) f_{t+h}.$$

However, the denominator in (23) still overestimates the variance of the numerator of the test statistic under the null. As a result, Theorem 7 remains valid even for testing the null $E(\overline{\Delta L}_{t+h} \mid f_{t+h}) = 0$ and the critical values remain the same.

Under either interpretation, it follows from (21) that the variance estimate in (23) is conservative. Under the first interpretation, this follows because the variance estimate takes into account variation in the factor structure and in $E(u^2_{i,t+h,1} - u^2_{i,t+h,2})$. Under the second interpretation, the variance estimate still includes cross-sectional variations in $E(u^2_{i,t+h,1} - u^2_{i,t+h,2})$. This seems unavoidable without introducing additional modeling assumptions that impose structure on this variation.[18]

4.4 Empirical Results for the Cross-sectional Tests

We next illustrate empirically how our new cross-sectional tests of equal predictive accuracy can be used to compare the WEO and CE forecasts introduced earlier. To this end, Figure 2 plots time-series of the cross-sectional average test statistics for the null of equal predictive accuracy (17) for individual years. We show two lines corresponding to the test statistics assuming either homogeneous factor loadings, (20) for testing (17), or heterogeneous loadings, (23) for testing (18). The results use the current-year spring forecasts, $h = 0, S$ in the WEO versus CE comparison and next-year Fall WEO forecasts in the comparison with the autoregressive forecasts. All empirical results assume the presence of two common factors.

The first point to notice is that, in most years, the two sets of test statistics in (20) and (23) are very similar even though they test potentially very different hypotheses and deal with factor-related shocks in very different ways. The reason for

[18] Essentially, we have a CLT for independent but non-identically distributed variables, $\Delta L_{i,t+h|t} - E\left[\Delta L_{i,t+h|t} | f_{t+h}\right]$, but the exact variance is difficult to estimate because $E\left[\Delta L_{i,t+h|t} | f_{t+h}\right]$ cannot be estimated from the observed data.

this similarity is that the tests only differ with respect to which terms they include in the denominator and this turns out to be of little importance. Turning to the individual test statistics, the top left panel shows that the WEO forecasts of real GDP growth were significantly more accurate than the corresponding CE forecasts only in one year during our sample which happens to be the peak of the global financial crisis. Conversely, the WEO forecasts were never significantly less accurate than the CE forecasts..

Compared to the AR(1) forecasts (top right panel), the Fall next-year WEO forecasts of real GDP growth were significantly more accurate during the Global Financial Crisis. Reassuringly, the WEO forecasts of GDP growth are not less accurate than the autoregressive forecasts in any years during our sample.

Figure 3 plots disaggregate t-statistics for comparing the null of equal expected squared error loss for WEO versus CE (rows 1 and 3) or WEO versus AR forecasts (rows 2 and 4). The test statistics are computed for three time clusters, namely 1995-2006, 2007-2009 (Global Financial Crisis), and 2010-16 and for five to seven groups of countries.

First consider the GDP growth forecasts (top two rows). We cannot reject the null of equal predictive accuracy of the WEO and CE forecasts for any of the economic groups during the early sample (1995-2006). Conversely, the WEO forecasts were significantly more accurate than the CE forecasts during the GFC (2007-2009) and for the Latin America and Caribbean and dms groups in the last part of the sample (2010-16). Measured against the AR forecasts, there is little to suggest that the WEO forecasts were significantly more accurate for any of the clusters in the early sample (1995-2006). However, the WEO forecasts were significantly more accurate than the AR forecasts during the financial crisis for three of seven regions. In the last part of the sample (2010-2016), the WEO forecasts of GDP growth for advanced economies and emerging and developing Europe were also significantly more accurate than their autoregressive counterparts.

Turning to the inflation forecasts (bottom two panels in Figure 2), the WEO forecasts are significantly more accurate than the CE forecasts in 2000, 2007-09 and, again, in 2013 and 2016. We find no years in which the CE inflation forecasts were significantly more accurate than their WEO counterparts. Moreover, the WEO inflation forecasts performed significantly better than the AR forecasts in most years, a notable exception being 2008 for which the reverse holds.

Next, consider the disaggregate inflation forecasts presented in the bottom two rows of Figure 3. The key finding here is that while the WEO inflation forecasts for advanced economies were not significantly more accurate than the CE forecasts prior to and after the financial crisis, they were far more accurate (with a t-statistic of 10) during the crisis. WEO inflation forecasts also fared well against the CE forecasts for the other groups during the crisis. Compared to the autoregressive forecasts, the WEO forecasts were significantly more accurate for all economic groups both in the early and late sample (1995-2006, 2010-16), but only so for advanced economies during the crisis.

These plots demonstrate a number of points. First, the results are very robust to whether we assume homogeneous or heterogeneous loadings on the common factors and test the null of equal cross-sectional average predictive accuracy unconditionally or conditional on the factor loadings. Second, there is sufficient year-on-year variation in the cross-sectional tests to allow us to pinpoint those years in which we can reject that two forecasts are equally accurate. For example, the sometimes very different performance of the test statistics computed for the Global Financial Crisis shows that the WEO forecasts did not uniformly dominate (or get dominated by) the CE or AR forecasts. Third, disaggregating by economic region or development stage and crisis and non-crisis years, we can gain important insights into which forecasts perform best for which types of economies across different economic states.

4.4.1 Predictive Accuracy at Different Horizons

Figure 4 plots time series of the single-year cross-sectional test statistics, now comparing the WEO forecasts at long and short horizons. Because the test statistics in (20) and (23), are again extremely similar regardless of which methodology we use, we only present results for the conditional test that allows for heterogeneous factor loadings and show the outcomes for GDP growth and inflation forecasts in the same panels. First consider the one-year-ahead spring versus fall forecasts of GDP growth (top left corner). As we move forward in time from the spring to the fall WEO forecasts, we find no years in which the GDP growth forecasts become significantly less accurate. Conversely, we see significant improvements in predictive accuracy at shorter horizons in 2002 and, again, during and after the GFC. Larger improvements in the accuracy of the WEO GDP growth forecasts emerge when moving from prior-year fall predictions to current-year spring forecasts (top right panel) and, even more so,

from current-year spring to current-year fall forecasts (bottom left panel), with many of the individual years generating t-statistics above two. Finally, we see evidence of significant improvements in predictive accuracy when moving from next-year spring to current-year fall forecasts of GDP growth (bottom right panel) for almost all years with exception of 1997-1998.

Next, consider the corresponding WEO inflation forecasts recorded at the four different horizons. With one exception (the 2009 comparison of the one-year-ahead fall and spring WEO inflation forecasts), we do not find any years with significant deterioration in predictive accuracy as the forecast horizon is reduced. Conversely, the WEO inflation forecasts tend to become significantly more accurate in individual years (as well as overall) as we move from the one-year Fall to the current-year Spring or from the current-year Spring to the current-year Fall forecasts and more information relevant for predicting inflation becomes available.

5 Decomposing Differences in Forecasting Performance

Under heterogeneous factor loadings, the conditionally expected cross-sectional average loss differential depends on differences in the two forecasts' factor loadings times the factor realizations, $(\lambda_{i,1}' f_{t+h})^2 - (\lambda_{i,2}' f_{t+h})^2$, as well as differences in the squared idiosyncratic error terms, $u_{i,t+h,1}^2 - u_{i,t+h,2}^2$, both averaged cross-sectionally.

To help interpret the economic sources of differences in forecasting performance, we can quantify the magnitude of these components. For example, we might be interested in testing whether differences in two forecasts' accuracy during the Global Financial crisis (2007-2008) was due to differences in their exposure (loadings) to a set of common factors that took on unusually large values. For example, one forecast may have been particularly exposed to financial market performance. Alternatively, differences in predictive accuracy could be due to differences in the variance of the idiosyncratic errors.

This section discusses how to separately conduct inference on the squared conditional bias and idiosyncratic variance components.

5.1 Decomposing the Conditional Squared Error Loss

Using equation (21), we can express the (cross-sectional) average conditional squared error loss difference as the sum of the average difference in squared conditional bias and the average difference in the conditional idiosyncratic error variance:

$$\underbrace{n^{-1}\sum_{i=1}^{n} E\left(\Delta L_{i,t+h|t} \mid \mathcal{F}\right)}_{E(\overline{\Delta L}_{t+h} \mid \mathcal{F})} = \underbrace{n^{-1}\sum_{i=1}^{n}\left[(\lambda_{i,1}' f_{t+h})^2 - (\lambda_{i,2}' f_{t+h})^2\right]}_{bias^2_{t+h}} + \underbrace{n^{-1}\sum_{i=1}^{n} E\left(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}\right)}_{E\left(\Delta^2_{ut+h} \mid \mathcal{F}\right)}. \tag{24}$$

The terms on the right hand side of the decomposition in (24) are unobserved, so we need to impose more structure on the data generating process to be able to separately identify these components.

Suppose the factor realizations, f_{t+h}, are independent of the factor loadings, $(\lambda_{i,1}, \lambda_{i,2})$ and, moreover, that f_{t+h} is stationary. In this case, we can compute the difference in the relative predictive accuracy of the forecasts at each point in time and relate this to changes in the relative accuracy of the idiosyncratic part, $E(u^2_{i,t+h,1} - u^2_{i,t+h,2})$. Moreover, provided that $n^{-1}\sum_{i=1}^{n} E[(\lambda_{i,1}' f_{t+h})^2 - (\lambda_{i,2}' f_{t+h})^2]$ is time invariant, we can use time series data to consistently estimate this quantity and characterize the asymptotics of the estimate.

Conversely, if f_{t+h} is nonstationary and $E(u^2_{i,t+h,1} - u^2_{i,t+h,2})$ is allowed to change with $t+h$, we cannot identify $n^{-1}\sum_{i=1}^{n} E[(\lambda_{i,1}' f_{t+h})^2 - (\lambda_{i,2}' f_{t+h})^2]$. However, since we can estimate $(\lambda_{i,1}' f_{t+h})^2 - (\lambda_{i,2}' f_{t+h})^2$ from time-series data, we can still perform tests on $n^{-1}\sum_{i=1}^{n} E(u^2_{i,t+h,1} - u^2_{i,t+h,2})$.[19]

Building on these observations, note that

$$\overline{\Delta L}_{t+h} - bias^2_{t+h} = \overline{\Delta}^2_{ut+h} + \frac{2}{n}\sum_{i=1}^{n}\left[\lambda_{i,1}' f_{t+h} u_{i,t+h,1} - \lambda_{i,2}' f_{t+h} u_{i,t+h,2}\right], \tag{25}$$

where $\overline{\Delta}^2_{ut+h} = n^{-1}\sum_{i=1}^{n}\left(u^2_{i,t+h,1} - u^2_{i,t+h,2}\right)$.

Provided that n is relatively large so the last term on the right side of equation (25) is small, the bias-adjusted average loss differential on the left hand side of (25) can be expected to be a good estimate of the difference in the two forecasts' idiosyncratic

[19] Notice that even if we know the random variables $\{(\lambda_{i,1}' f_{t+h})^2 - (\lambda_{i,2}' f_{t+h})^2\}_{i=1}^{n}$, this does not mean that we can estimate $n^{-1}\sum_{i=1}^{n} E[(\lambda_{i,1}' f_{t+h})^2 - (\lambda_{i,2}' f_{t+h})^2]$ if f_{t+h} is non-stationary.

variance at time t, $E\left(\Delta_{ut+h}^2 \mid \mathcal{F}\right)$.[20]

We next discuss three strategies for computing $(\lambda_{i,1}' f_{t+h})^2 - (\lambda_{i,2}' f_{t+h})^2$. The first exploits group patterns in factor loadings and so is applicable when factor loadings are homogeneous within certain groups of units. This approach poses no limit on the number of factors affecting the forecast errors but requires that clusters can be identified within which there is little or no heterogeneity in the factor loadings. The second approach uses the common correlated effects (CCE) method of Pesaran (2006). While this approach does not impose tight restrictions on factor loadings, in practice it puts limits on the number of common factors driving the forecast errors. The third approach, principal components (PCA), is similar to the CCE approach but does not impose tight bounds on the number of common factors in the forecast error differentials.

5.2 Clustering in Factor Loadings

It is common in empirical applications to have data on units that share certain observable characteristics or features which make them more similar than randomly selected units. For example, advanced economies may react in a broadly similar way to certain supply shocks which, in turn, affect emerging or developing economies in a very different manner. Or, the effect of an interest rate increase on the default probability of credit card holders may be quite different across high, medium, and low income households, yet be broadly similar within these three categories.

In this subsection we develop a class of estimators using the identifying assumption that clusters of cross-sectional units share the same factor loadings, while allowing factor loadings to differ across clusters. More formally, suppose that a set of K clusters $\bigcup_{k=1}^{K} H_k = \{1, ..., n\}$ form a partition of all n units so that each unit belongs to one unique cluster, H_k, i.e., $H_j \cap H_l = \emptyset$ with $n_k = |H_k|$ elements in the kth cluster. We assume that the cluster membership for each unit is known ex ante and so is not determined endogenously from the data. Moreover, suppose that the factor loadings $(\lambda_{i,1}, \lambda_{i,2})$ can differ across clusters $(\lambda_{i,1}, \lambda_{i,2}) \neq (\lambda_{j1}, \lambda_{j2})$ for $i \in H_k$ and $j \in H_l$, but are homogeneous within clusters

$$(\lambda_{i,1}, \lambda_{i,2}) = (\lambda_{1,(k)}, \lambda_{2,(k)}) \text{ for all } i \in H_k. \tag{26}$$

[20] Of course, we do not directly observe the idiosyncratic errors and factors. However, since $\overline{\Delta L}_{t+h}$ is observed, from (25) we only need to estimate the factor-induced squared bias term, $bias_{t+h}^2$.

5.2.1 Testing Equal Idiosyncratic Error Variances

We first discuss testing the conditional null given $\mathcal{F}$of equal average idiosyncratic error variance for the two forecasts for all units in cluster k:

$$H_0^{idio}: \; n_k^{-1} \sum_{i \in H_k} E(u_{i,t+h,1}^2 - u_{i,t+h,2}^2 \mid \mathcal{F}) = 0. \tag{27}$$

To test this null, we need to construct an estimate of the idiosyncratic variance within each cluster. To see how group patterns in factor loadings allow us to identify the idiosyncratic variance component, $\overline{\Delta}_{ut+h}^2$, define the errors from the two forecasts, averaged within each cluster, as

$$\overline{e}_{1,k,t+h} \equiv n_k^{-1} \sum_{i \in H_k} \left(y_{i,t+h} - \hat{y}_{i,t+h|t,1}\right) = \lambda_{1,(k)}' f_{t+h} + n_k^{-1} \sum_{i \in H_k} u_{i,t+h,1},$$

and

$$\overline{e}_{2,k,t+h} \equiv n_k^{-1} \sum_{i \in H_k} \left(y_{i,t+h} - \hat{y}_{i,t+h|t,2}\right) = \lambda_{2,(k)}' f_{t+h} + n_k^{-1} \sum_{i \in H_k} u_{i,t+h,2}.$$

Squaring these within-cluster average forecast errors and computing their difference, we have

$$\begin{aligned} \overline{e}_{1,k,t+h}^2 - \overline{e}_{2,k,t+h}^2 = (\lambda_{1,(k)}' f_{t+h})^2 - (\lambda_{2,(k)}' f_{t+h})^2 + \left(n_k^{-1} \sum_{i \in H_k} u_{i,t+h,1}\right)^2 - \left(n_k^{-1} \sum_{i \in H_k} u_{i,t+h,2}\right)^2 \\ + 2\lambda_{1,(k)}' f_{t+h} n_k^{-1} \sum_{i \in H_k} u_{i,t+h,1} - 2\lambda_{2,(k)}' f_{t+h} n_k^{-1} \sum_{i \in H_k} u_{i,t+h,2}. \end{aligned} \tag{28}$$

Let $\overline{\Delta u}_{t+h,k}^2$ be the average loss differential for cluster k, i.e., $\overline{\Delta L}_{t+h,k} \equiv n_k^{-1} \sum_{i \in H_k} \Delta L_{i,t+h}$, adjusted for the difference $(\overline{e}_{1,k,t+h}^2 - \overline{e}_{1,k,t+h}^2)$:

$$\overline{\Delta u}_{t+h,k}^2 = \overline{\Delta L}_{t+h,k} - (\overline{e}_{1,k,t+h}^2 - \overline{e}_{1,k,t+h}^2). \tag{29}$$

Using this result, we can use the following test statistic to test H_0^{idio} in (27):

$$S_k = \frac{\sqrt{n_k} \overline{\Delta u}_{t+h,k}^2}{\sqrt{n_k^{-1} \sum_{i \in H_k} (\Delta L_{i,t+h} - \overline{\Delta L}_{t+h,k})^2}}. \tag{30}$$

Theorem 8. *Suppose Assumption 4 holds. Then under the null hypothesis* H_0^{idio} : $n_k^{-1}\sum_{i\in H_k} E(u_{i,t+h,1}^2 - u_{i,t+h,2}^2 \mid \mathcal{F}) = 0$, *we have*

$$\limsup_{n_k\to\infty} P\left(|S_k| > z_{1-\alpha/2}\right) \leq \alpha,$$

where $z_{1-\alpha/2}$ *is the* $(1-\alpha/2)$ *quantile of a* $N(0,1)$ *variable.*

Alternatively, we can test that the null of equal expected squared idiosyncratic forecast errors holds on average, i.e., across all units though not necessarily within each cluster:

$$H_0^{idio-av} : \; n^{-1}\sum_{i=1}^{n} E(u_{i,t+h,1}^2 - u_{i,t+h,2}^2 \mid \mathcal{F}) = 0. \tag{31}$$

To this end, let $\overline{\Delta u}_{t+h}^2 = \sum_{k=1}^K \frac{n_k}{n}\overline{\Delta u}_{t+h,k}^2$ be the cluster-weighted average difference in squared idiosyncratic forecast errors, and consider the test statistic

$$S_c = \frac{\sqrt{n}\overline{\Delta u}_{t+h}^2}{\sqrt{n^{-1}\sum_{k=1}^K\sum_{i\in H_k}(\Delta L_{i,t+h} - \overline{\Delta L}_{t+h,k})^2}}. \tag{32}$$

We use S_c to test the null in (31) of equal average idiosyncratic forecast error variance:

Corollary 1. *Suppose Assumption 4 holds and assume that* $\lim_{n\to\infty} n_k/n > 0$ *for all* $1 \leq k \leq K$. *Then under the null in (31), we have*

$$\limsup_{n\to\infty} P\left(|S_c| > z_{1-\alpha/2}\right) \leq \alpha,$$

where $z_{1-\alpha/2}$ *is the* $(1-\alpha/2)$ *quantile of a* $N(0,1)$ *variable.*

For example, using Corollary 1, we can compute a $1-\alpha$ confidence interval for the squared idiosyncratic forecast errors $\overline{\Delta u}_{t+h}^2$ as

$$\overline{\Delta u}_{t+h}^2 \pm \frac{z_{1-\alpha}}{\sqrt{n}}\sqrt{n^{-1}\sum_{k=1}^K\sum_{i\in H_k}(\Delta L_{i,t+h} - \overline{\Delta L}_{t+h,k})^2}. \tag{33}$$

5.2.2 Testing Equal Squared Biases

Next, consider the squared bias component of the expected loss differential in (24). Under the assumed homogeneous factor loadings within clusters in (26), we have

$$n^{-1}\sum_{i=1}^{n}\left[(\lambda_{i,1}'f_{t+h})^2-(\lambda_{i,2}'f_{t+h})^2\right]=\sum_{k=1}^{K}\frac{n_k}{n}\left((\lambda_{1,(k)}'f_{t+h})^2-(\lambda_{2,(k)}'f_{t+h})^2\right).$$

We can estimate $(\lambda_{1,(k)}'f_{t+h})^2-(\lambda_{2,(k)}'f_{t+h})^2$ by $\overline{e}_{1,k,t+h}^2-\overline{e}_{1,k,t+h}^2$. By (28), we have

$$\begin{aligned}\overline{e}_{1,k,t+h}^2-\overline{e}_{1,k,t+h}^2&=(\lambda_{1,(k)}'f_{t+h})^2-(\lambda_{2,(k)}'f_{t+h})^2\\&\quad+2\lambda_{1,(k)}'f_{t+h}n_k^{-1}\sum_{i\in H_k}u_{i,t+h,1}-2\lambda_{2,(k)}'f_{t+h}n_k^{-1}\sum_{i\in H_k}u_{i,t+h,2}+O_P(n_k^{-1}).\end{aligned}$$

To test the null of equal squared bias, we use the following test statistic:

$$B_{n,1}=\frac{\sqrt{n}\sum_{k=1}^{K}\frac{n_k}{n}(\overline{e}_{1,k,t+h}^2-\overline{e}_{2,k,t+h}^2)}{2\sqrt{n^{-1}\sum_{k=1}^{K}\sum_{i\in H_k}(\overline{e}_{1,k,t+h}\hat{u}_{i,t+h,1}-\overline{e}_{2,k,t+h}\hat{u}_{i,t+h,2})^2}},\tag{34}$$

where, for $i\in H_k$, $\hat{u}_{i,t+h,1}=y_{i,t+h}-\hat{y}_{i,t+h|t,m_1}-\overline{e}_{1,k,t+h}$ and $\hat{u}_{i,t+h,2}=y_{i,t+h}-\hat{y}_{i,t+h|t,m_2}-\overline{e}_{2,k,t+h}$. We can show that $B_{n,1}\left(n^{-1}\sum_{i=1}^{n}\left[(\lambda_{i,1}'f_{t+h})^2-(\lambda_{i,2}'f_{t+h})^2\right]\right)\to^d N(0,1)$, and so:

Theorem 9. *Suppose Assumption 3 holds and assume that* $\lim_{n\to\infty}n_k/n>0$ *for all* $1\le k\le K$. *Then under* $H_0:\ n^{-1}\sum_{i=1}^{n}\left[(\lambda_{i,1}'f_{t+h})^2-(\lambda_{i,2}'f_{t+h})^2\right]=0$, *we have*

$$\limsup_{n\to\infty}P\left(|B_{n,1}|>z_{1-\alpha/2}\right)\le\alpha.$$

Note that the null of equal squared bias relates to our earlier discussion of homogeneous versus heterogeneous factor loadings: If factor loadings are the same across two sets of forecasts, their squared bias differential should also be close to zero.

Using Theorem 9, a $1-\alpha$ confidence interval for $n^{-1}\sum_{i=1}^{n}\left[(\lambda_{i,1}'f_{t+h})^2-(\lambda_{i,2}'f_{t+h})^2\right]$ is

$$\sum_{k=1}^{K}\frac{n_k}{n}(\overline{e}_{1,k,t+h}^2-\overline{e}_{2,k,t+h}^2)\pm\frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{n^{-1}\sum_{k=1}^{K}\sum_{i\in H_k}(\overline{e}_{1,k,t+h}\hat{u}_{i,t+h,1}-\overline{e}_{2,k,t+h}\hat{u}_{i,t+h,2})^2}.\tag{35}$$

Note that because $B_{n,1}\left(n^{-1}\sum_{i=1}^{n}\left[(\lambda_{i,1}'f_{t+h})^2-(\lambda_{i,2}'f_{t+h})^2\right]\right)\to^d N(0,1)$, the confidence interval is asymptotically exact.

5.3 Factor Structure Estimated by CCE

In many empirical applications, a cluster structure may not be suitable either because units are not easily assigned to individual clusters or because factor loadings are not homogeneous within clusters. For such applications, a more traditional factor setting may be more appropriate. To this end, suppose we observe a panel of forecast errors $\{e_{i,s+h,m}\}_{1\leq i\leq n,\ 1\leq s\leq T}$ generated according to the factor model in (16), $e_{i,s+h,m}=\lambda_{i,m}'f_{s+h}+u_{i,s+h,m}$, where $m=1,2$, $\lambda_{i,m}\in\mathbb{R}^{r\times v}$ and $f_{s+h}\in\mathbb{R}^r$ with $v\geq r$, so the number of observables, v, is at least equal to the number of factors, r. The requirement that $v\geq r$ implies that if we do not include observables other than the two sets of forecast errors, we can allow for at most two factors. Conversely, including more observable variables that are driven by the same factors lets us relax this restriction and allow for additional factors.

5.3.1 Difference in Idiosyncratic Error Variances

Let $e_{i,s+h}=(e_{i,s+h,1},e_{i,s+h,2})'\in\mathbb{R}^2$ and $u_{i,s+h}=(u_{i,s+h,1},u_{i,s+h,2})'\in\mathbb{R}^2$ be 2×1 vectors of forecast errors and idiosyncratic residuals and define the cross-sectional averages $\bar{e}_{s+h}=n^{-1}\sum_{i=1}^{n}e_{i,s+h}$, $\bar{u}_{s+h}=n^{-1}\sum_{i=1}^{n}u_{i,s+h}$ and $\bar{\lambda}=n^{-1}\sum_{i=1}^{n}\lambda_i$ with $\lambda_i=(\lambda_{i,1},\lambda_{i,2})\in\mathbb{R}^{r\times 2}$. Assuming that we can invoke a CLT for the cross-sectional average of the idiosyncratic shocks, $\bar{u}_{s+h}$ will be small and $\bar{e}_{s+h}\approx\bar{\lambda}'f_{s+h}$ can be used as a proxy for the unobserved factors. This is the common correlated effects (CCE) idea proposed in Pesaran (2006). In turn, we can estimate the individual factor loadings, λ_{im}, from a time-series regression

$$\hat{\lambda}_i'=\left(\sum_{s=1}^{T}e_{i,s+h}\bar{e}_{s+h}'\right)\left(\sum_{s=1}^{T}\bar{e}_{s+h}\bar{e}_{s+h}'\right)^{-1}.$$

Let $\lambda_{i,1}$ denote the first column of λ_i, with similar notations used for $\hat{\lambda}_{i,1}$ and $\hat{\lambda}_{i,2}$. Consider the following regularity conditions:

Assumption 5. *The following conditions hold for $m=1,2$:*
(1) the smallest eigenvalue of $\bar{\lambda}\bar{\lambda}'$ is bounded away from zero.

(2) $\{u_{i,t+h,m}\}_{i=1}^{n}$ *has mean zero and is independent across* i*.*

The first part of Assumption 5 implies that the number of factors cannot exceed the dimension of $e_{i,s+h}$–otherwise the smallest eigenvalue of $\bar{\lambda}\bar{\lambda}'$ is zero.

Using Assumption 5, we can characterize the difference between the average squared forecast errors and the average squared factor values, both weighted by the factor loadings, λ_i' :

Lemma 1. *Under Assumption 5, we have*

$$n^{-1/2}\sum_{i=1}^{n}[(\hat{\lambda}_{i,1}'\bar{e}_{t+h})^2-(\lambda_{i,1}'f_{t+h})^2]=2n^{-1/2}\bar{u}_{t+h}'\bar{\lambda}'(\bar{\lambda}\bar{\lambda}')^{-1}\left(\sum_{i=1}^{n}\lambda_{i,1}\lambda_{i,1}'\right)f_{t+h}+o_P(1).$$

Next, consider the null that the difference in the squared idiosyncratic variance component of the forecast errors equals zero:

$$H_0:\ n^{-1}\sum_{i=1}^{n}E(u_{i,t+h,1}^2-u_{i,t+h,2}^2\mid\mathcal{F})=0. \tag{36}$$

To test this hypothesis, we use the following test statistic

$$S_{cce}=\frac{\sqrt{n}\overline{\Delta\hat{u}}_{t+h}^2}{\sqrt{n^{-1}\sum_{i=1}^{n}(\Delta L_{i,t+h|t}-[(\hat{\lambda}_{i,1}'\bar{e}_{t+h})^2-(\lambda_{i,2}'\bar{e}_{t+h})^2]-\hat{c}_{t+h}+\hat{u}_{i,t+h}'\hat{D}_{t+h})^2}}, \tag{37}$$

where $\hat{c}_{t+h}=n^{-1}\sum_{i=1}^{n}(\Delta L_{i,t+h|t}-[(\hat{\lambda}_{i,1}'\bar{e}_{t+h})^2-(\lambda_{i,2}'\bar{e}_{t+h})^2]+\hat{u}_{i,t+h}'\hat{D}_{t+h})$,

$$\overline{\Delta\hat{u}}_{t+h}^2=n^{-1}\sum_{i=1}^{n}\Delta L_{i,t+h}-n^{-1}\sum_{i=1}^{n}[(\hat{\lambda}_{i,1}'\bar{e}_{t+h})^2-(\hat{\lambda}_{i,2}'\bar{e}_{t+h})^2] \tag{38}$$

and

$$\hat{D}_{t+h}=n^{-1}\sum_{i=1}^{n}(\hat{\lambda}_{i,1}\hat{\lambda}_{i,1}'-\hat{\lambda}_{i,2}\hat{\lambda}_{i,2}')\bar{e}_{t+h}. \tag{39}$$

Using these definitions, we now have the following result:

Theorem 10. *Suppose Assumption 5 holds. Then under the null hypothesis* $n^{-1}\sum_{i=1}^{n}E(u_{i,t+h,1}^2-u_{i,t+h,2}^2\mid\mathcal{F})=0$, $S_{cce}\to^d N(0,1)$.

Using that S_{cce} follows a standard Gaussian distribution asymptotically, we can

compute a $1-\alpha$ confidence interval for $n^{-1}\sum_{i=1}^n E(u_{i,t,1}^2 - u_{i,t,2}^2 \mid \mathcal{F})$ as

$$\overline{\Delta\hat{u}}_{t+h}^2 \pm \frac{z_{1-\alpha}}{\sqrt{n}} \sqrt{n^{-1}\sum_{i=1}^n (\Delta L_{i,t+h|t} - [(\hat{\lambda}_{i,1}'\bar{e}_{t+h})^2 - (\lambda_{i,2}'\bar{e}_{t+h})^2] - \hat{c} + \hat{u}_{i,t+h}'\hat{D}_{t+h})^2} \quad (40)$$

5.3.2 Squared Bias Differences

Next, consider the squared bias component of the MSE loss differential. Define

$$D_{t+h} = \bar{\lambda}'(\bar{\lambda}\bar{\lambda}')^{-1}\left(n^{-1}\sum_{i=1}^n [\lambda_{i,1}\lambda_{i,1}' - \lambda_{i,2}\lambda_{i,2}']\right) f_{t+h}.$$

Using

$$\sqrt{n}\left(n^{-1}\sum_{i=1}^n \left[(\hat{\lambda}_{i,1}'\bar{e}_{t+h})^2 - (\hat{\lambda}_{i,2}'\bar{e}_{t+h})^2\right] - n^{-1}\sum_{i=1}^n \left[(\lambda_{i,1}'f_{t+h})^2 - (\lambda_{i,2}'f_{t+h})^2\right]\right) = 2n^{1/2}\bar{u}_{t+h}'D_{t+h} + o_P(1),$$

it follows that $n^{-1}\sum_{i=1}^n \left[(\hat{\lambda}_{i,1}'\bar{e}_{t+h})^2 - (\hat{\lambda}_{i,2}'\bar{e}_{t+h})^2\right]$ is a $\sqrt{n}$-consistent estimator for the average difference in the squared bias differential, $n^{-1}\sum_{i=1}^n \left[(\lambda_{i,1}'f_{t+h})^2 - (\lambda_{i,2}'f_{t+h})^2\right]$, where the estimation error is asymptotically $2\bar{u}_{t+h}'D_{t+h}$. To construct tests for the squared bias difference, consider the following test statistic

$$B_{n,2} = \frac{n^{-1/2}\sum_{i=1}^n \left[(\hat{\lambda}_{i,1}'\bar{e}_{t+h})^2 - (\hat{\lambda}_{i,2}'\bar{e}_{t+h})^2\right]}{2\sqrt{n^{-1}\sum_{i=1}^n (\hat{u}_{i,t+h}'\hat{D}_{t+h})^2}}, \quad (41)$$

where, again, $\hat{u}_{i,t+h} = e_{i,t+h} - \hat{\lambda}_i'\bar{e}_{t+h}$. The following result characterizes the distribution of this statistic:

Theorem 11. *Suppose that Assumption 5 holds. Then under the null hypothesis* $n^{-1}\sum_{i=1}^n \left[(\lambda_{i,1}'f_{t+h})^2 - (\lambda_{i,2}'f_{t+h})^2\right] = 0$, *we have* $B_{n,2} \to^d N(0,1)$.

Using Theorem 11, we can construct a confidence interval for the average squared bias differential, $n^{-1}\sum_{i=1}^n \left[(\lambda_{i,1}'f_{t+h})^2 - (\lambda_{i,2}'f_{t+h})^2\right]$ as

$$n^{-1}\sum_{i=1}^n \left[(\hat{\lambda}_{i,1}'\bar{e}_{t+h})^2 - (\hat{\lambda}_{i,2}'\bar{e}_{t+h})^2\right] \pm \frac{z_{1-\alpha/2}}{\sqrt{n}} \sqrt{n^{-1}\sum_{i=1}^n (\hat{u}_{i,t+h}'\hat{D}_{t+h})^2}. \quad (42)$$

Again, this confidence interval is asymptotically exact.

5.4 Factor Structure Estimated by PCA

We next describe an alternative to the CCE approach in Section 5.3 which uses principal components analysis (PCA) to extract the common factors. A notable advantage of the PCA approach is that, unlike the CCE approach, the number of observed forecast errors does not pose an upper bound on the number of factors. In practice, this means that we can allow for more factors under the PCA approach.

Define the difference in the idiosyncratic forecast error variance

$$\overline{\Delta\hat{u}}^2_{t+h} = n^{-1}\sum_{i=1}^{n} \Delta L_{i,t+h} - n^{-1}\sum_{i=1}^{n}\left[(\hat{\lambda}'_{i,1}\hat{f}_{t+h})^2 - (\hat{\lambda}'_{i,2}\hat{f}_{t+h})^2\right]. \tag{43}$$

As before, let $\hat{f}_{t+h}$ and $\hat{\lambda}_i$ be the estimated factors and factor loadings obtained using PCA estimation. Then we have the following results on $\overline{\Delta\hat{u}}^2_{t+h}$:

Lemma 2. *Under Assumptions A-F in Bai (2003), we have*

$$\begin{aligned}
&\sqrt{n}\left[\overline{\Delta\hat{u}}^2_{t+h} - n^{-1}\sum_{i=1}^{n} E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F})\right] \\
&= n^{-1/2}\sum_{i=1}^{n}\left[(u^2_{i,t+h,1} - u^2_{i,t+h,2}) - E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}) + 2(\lambda'_{i,1}f_{t+h}u_{i,t+h,1} - \lambda'_{i,2}f_{t+h}u_{i,t+h,2})\right] \\
&\quad + o_P(1).
\end{aligned}$$

Notice that we no longer have a term involving D_{t+h}. Depending on the distribution of the idiosyncratic term, the PCA approach might yield a more efficient estimator than the CCE approach since it does not require us to estimate this term.

From this point, all steps in the inference procedure are exactly the same as those in Section 5.3, except that $(\hat{\lambda}'_{i,1}\bar{e}_{t+h}, \hat{\lambda}'_{i,2}\bar{e}_{t+h})$ is replaced by the PCA estimate $(\hat{\lambda}'_{i,1}\hat{f}_{t+h}, \hat{\lambda}'_{i,2}\hat{f}_{t+h})$ and we set $\hat{D}_{t+h} = 0$. Specifically, in Equations (37), (38) and (40), we replace $(\hat{\lambda}'_{i,1}\bar{e}_{t+h}, \hat{\lambda}'_{i,2}\bar{e}_{t+h})$ with the PCA estimate $(\hat{\lambda}'_{i,1}\hat{f}_{t+h}, \hat{\lambda}'_{i,2}\hat{f}_{t+h})$ and set $\hat{D}_{t+h} = 0$. We also replace $B_{n,2}$ defined in (41) with the following

$$\tilde{B}_{n,2} = \frac{n^{-1/2}\sum_{i=1}^{n}\left[(\hat{\lambda}'_{i,1}\hat{f}_{t+h})^2 - (\hat{\lambda}'_{i,2}\hat{f}_{t+h})^2\right]}{2\sqrt{n^{-1}\sum_{i=1}^{n}(\hat{\lambda}'_{i,1}\hat{f}_{t+h}\hat{u}_{i,t+h,1} - \hat{\lambda}'_{i,2}\hat{f}_{t+h}\hat{u}_{i,t+h,2})^2}}, \tag{44}$$

where $\hat{u}_{i,t+h,m} = e_{i,t+h,m} - \lambda'_{i,m} f_{t+h}$.

5.5 Empirical Results

Figure 5 uses a set of heat diagrams to show the annual values of the cross-sectional tests statistics used to test the null of equal idiosyncratic variances (36) for the WEO versus CE forecasts or the WEO versus autoregressive forecasts. Each panel corresponds to a separate pair-wise comparison. The heat diagrams use blue to indicate years where the WEO forecasts have a smaller idiosyncratic variance component than the CE or autoregressive forecasts with red color indicating the reverse. Diamonds indicate years in which the test statistic is significant at the 5% level, using a two-sided test. Each diagram contains three rows showing results based on the cluster, CCE, and PCA approaches, respectively.

First consider the comparison of the WEO versus CE forecasts of real GDP growth (top panel). Although the test statistics in (32) and (37) fluctuate around zero in most years without being statistically significant, there is a clear trend from mostly negative values in the early part of the sample up to 2002 towards positive values later on in the sample. This suggests that the WEO forecasts of real GDP growth went from initially having a larger idiosyncratic error variance to subsequently having a smaller idiosyncratic variance than the CE forecasts in each and every year since 2008. Comparing the WEO and AR(1) forecasts (second panel), the idiosyncratic error variance component is again seen to be smaller for the WEO forecasts for the majority of years in our sample and this differential is statistically significant in some years during the Global Financial Crisis and its aftermath, 2008-2011.

Turning to the inflation forecasts (lower panels in Figure 5), idiosyncratic error variances are quite similar for the WEO and CE forecasts and we continue to find in most years that we cannot reject the null that the magnitude of the idiosyncratic variance of the WEO and CE forecasts is identical. We find strong evidence, however, that the WEO inflation forecasts have a significantly smaller idiosyncratic error variance than the AR forecasts in most years–particularly based on the test statistic (32) that uses the cluster approach–and there is no single year where the AR forecasts produce a significantly smaller idiosyncratic error variance than the WEO inflation forecasts.

Figure 6 shows the outcome of cross-sectional comparisons of the squared bias

component in the errors of the WEO, CE and autoregressive forecasts using the test statistics in (34), (41) and (44). For the GDP growth series (top two panels), we continue to find that the squared bias difference between the WEO and CE forecasts is insignificant in most years although again with a trend towards the squared bias of the WEO forecasts becoming smaller relative to the CE forecasts in the second half of the sample. Interestingly, the squared bias of the WEO GDP growth forecasts was significantly smaller than for the CE forecasts in 2008 and 2009, i.e., during the GFC. Comparing the squared bias components of the WEO and AR(1) GDP growth forecasts (second panel), we find that the squared bias is bigger for the AR(1) forecasts for most years in our sample, though not significantly so.

For the inflation series (bottom two panels), there is systematic evidence that the WEO forecasts had a smaller squared bias than the CE forecasts in most years. In fact, the WEO forecasts have a significantly smaller squared bias component than the CE forecasts in seven years under the test statistic that relies on the PCA factor model, although the evidence is weaker when we use the two other approaches to extract the squared bias component. Finally, there is strong evidence that the WEO inflation forecasts have a significantly smaller squared bias than the autoregressive inflation forecasts during most years in our sample (a notable exception being 2008), regardless of which method is used to extract the bias.

To summarize these findings, Table 5 reports the mean and median values of the idiosyncratic error variance and squared bias components for the CE versus WEO forecasts, with positive values indicating that the variance (or bias) of the CE forecasts errors exceeds that of the WEO forecast errors. To reduce the effect of outliers, we have standardized the forecast errors before computing cross-sectional averages. For GDP growth (Panel A), the difference in the average idiosyncratic variance is quite small, ranging between 5% and 16%, although this value grows far higher during the GFC. Differences in squared biases are also, on average, quite small, ranging from -2% to 9%, although again these values are substantially higher (and positive) during the GFC. For inflation (Panel B) we find substantially larger (positive) differences in the idiosyncratic variance and squared bias components, both on average and during the GFC period.

These results confirm our original finding that the accuracy of the WEO and CE forecasts of real GDP growth is comparable during our sample and show that this can be explained by idiosyncratic error variance and squared bias terms that were of

similar magnitude. Conversely, the significantly lower mean squared forecast errors for the WEO inflation forecasts compared to the CE inflation forecasts results from the WEO forecasts having both a substantially lower idiosyncratic error variance and a lower squared bias.

6 Conclusion

This paper develops new methods for testing the null of equal predictive accuracy in the context of panel data in which we observe time series of forecasts and outcomes of multiple units. We show that 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. In situations where forecast errors are cross-sectionally correlated as captured by a set of common factors, we show that inference about equal predictive accuracy can be conducted under a set of assumptions about exposures to the common factors. In particular, we show that the null of equal predictive accuracy can be conducted in settings with a small time-series dimension–even just a single period–and a large cross-sectional dimension if either (i) factor loadings are homogeneous across units so that the effect of common factors on forecast errors cancels out in loss differentials; or (ii) we condition on factor realizations and conduct a test of equal predictive accuracy, given these factors.

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. They also demonstrate the feasibility of identifying shifts over time in the relative factor exposure ("sensitivity") across different forecasts.

An important advantage of our new tests is that they can be computed using very few time-series observations–in fact only a single cross-section-provided that cross-sectional dependencies are properly accounted for. This makes the tests particularly useful in microeconomic forecast applications which often have short time-series dimensions due to infrequently conducted surveys or the attrition of individual house-

holds that enter and exit.[21] Panel data with longer time-series dimensions may, in turn, be subject to non-stationary dynamics as agents learn and modify their behavior. Approaches such as ours that do not depend on conducting inference on long time-series averages offer an advantage also in these settings.

References

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, pages 135–171.

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.

Chen, X., Shao, Q.-M., and Wu, W. B. (2016). Supplement to "self-normalized cramér-type moderate deviations under dependence". *The Annals of Statistics*.

Chong, Y. Y. and Hendry, D. F. (1986). Econometric evaluation of linear macro-economic models. *The Review of Economic Studies*, 53(4):671–690.

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.

[21] Giacomini et al. (2019) discuss micro forecasting approaches for annual PSID panels while Liu et al. (2018) and Liu et al. (2019) develop ways to forecast in panels with very short time-series dimensions.

Giacomini, R., Lee, S., and Sarpietro, S. (2019). Microforecasting with individual forecast selection. *Unpublished working paper, UCL*.

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.

Inoue, A. and Kilian, L. (2005). In-sample or out-of-sample tests of predictability: Which one should we use? *Econometric Reviews*, 23(4):371–402.

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.

Liu, L., Moon, H. R., and Schorfheide, F. (2018). Forecasting with dynamic panel data models. *Unpublished working paper, University of Pennsylvania*.

Liu, L., Moon, H. R., and Schorfheide, F. (2019). Forecasting with a panel tobit model. *Unpublished working paper, University of Pennsylvania*.

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.

Peña, V. H., Lai, T. L., and Shao, Q.-M. (2008). *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012.

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

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.

A Proofs

This section presents proofs of the theoretical results in the main body of our paper.

A.1 Theorem 1

Proof. By Theorem 5.20 of White (2014), we have

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

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

A.2 Theorems 2 and 4

Theorems 2 and 4 follow from Theorem 1 of Ibragimov and Müller (2010) and the continuous mapping theorem.

A.3 Theorems 3 and 5

Theorems 3 and 5 follow from Theorem 3.1 of Canay et al. (2017).

A.4 Theorem 6

Proof. Using (19), we have

$$\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t} \mid \mathcal{F}) = (u^2_{i,t+h,1} - u^2_{i,t+h,2}) - E[u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}] + 2(u_{i,t+h,1} - u_{i,t+h,2})\lambda_i' f_{t+h}.$$

Hence, conditional on $\mathcal{F}$, $\{\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t} \mid \mathcal{F})\}_{i=1}^n$ is independent across i with mean zero. By Assumption 3, the sequence $\{\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t} \mid \mathcal{F})\}_{i=1}^{n_t}$ conditional on $\mathcal{F}$ satisfies the Lyapunov condition. Hence, a standard argument yields

$$\frac{n^{-1/2}\sum_{i=1}^n [\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t} \mid \mathcal{F})]}{\sqrt{n^{-1}\sum_{i=1}^n [\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t} \mid \mathcal{F})]^2}} \xrightarrow{d} N(0,1).$$

Under the null that $n^{-1}\sum_{i=1}^n E(\Delta L_{i,t+h|t} \mid \mathcal{F}) = 0$, we have

$$\frac{n^{-1/2}\sum_{i=1}^n \Delta L_{i,t+h|t}}{\sqrt{n^{-1}\sum_{i=1}^n [\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t} \mid \mathcal{F})]^2}} \xrightarrow{d} N(0,1).$$

The result now follows by noticing that $n^{-1}\sum_{i=1}^n [\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t} \mid \mathcal{F})]^2 \leq n^{-1}\sum_{i=1}^n (\Delta L_{i,t+h|t})^2$. □

A.5 Theorem 8

Proof. Start by noticing that

$$\sqrt{n_k}\left(\overline{\Delta u}^2_{t+h,k} - n_k^{-1}\sum_{i\in H_k} E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F})\right)$$
$$= n_k^{-1/2}\sum_{i\in H_k}\left[(u^2_{i,t+h,1} - u^2_{i,t+h,2}) - E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}) + 2(\lambda_1' f_{t+h} u_{i,t+h,1} - \lambda_2' f_{t+h} u_{i,t+h,2})\right]$$

$$+ n_k^{-1/2} \left(n_k^{-1} \sum_{i \in H_k} u_{i,t+h,2} \right)^2 - n_k^{-1/2} \left(n_k^{-1} \sum_{i \in H_k} u_{i,t+h,1} \right)^2 .$$

By a CLT,

$$n_k^{-1/2} \left(n_k^{-1} \sum_{i \in H_k} u_{i,t+h,2} \right)^2 - n_k^{-1/2} \left(n_k^{-1} \sum_{i \in H_k} u_{i,t+h,1} \right)^2 = O_P(n_k^{-3/2}) = o_P(1).$$

Therefore, $\overline{\Delta u}^2_{t+h,k}$ is a $\sqrt{n_k}$-consistent estimator for $n_k^{-1} \sum_{i \in H_k} E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F})$. By the same CLT argument, it is also asymptotically normal. To estimate the variance of $\overline{\Delta u}^2_{t+h,k}$, we use $n_k^{-1} \sum_{i \in H_k} (\Delta L_{i,t+h} - \overline{\Delta L}_{t+h,k})^2$, where $\overline{\Delta L}_{t+h,k} = n_k^{-1} \sum_{i \in H_k} \Delta L_{i,t+h}$. □

A.6 Corollary 1

Proof. The result follows once we notice that

$$\begin{aligned}
&\sqrt{n} \left(\sum_{k=1}^{K} \frac{n_k}{n} \overline{\Delta u}^2_{t+h,k} - n^{-1} \sum_{i=1}^{n} E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}) \right) \\
&= \sqrt{n} \sum_{k=1}^{K} \frac{n_k}{n} \left(\overline{\Delta u}^2_{t+h,k} - n_k^{-1} \sum_{i \in H_k} E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}) \right) \\
&= \sum_{k=1}^{K} \frac{n_k}{\sqrt{n}} \left\{ n_k^{-1} \sum_{i \in H_k} \left[(u^2_{i,t+h,1} - u^2_{i,t+h,2}) - E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}) \right] + O_P(n_k^{-1}) \right\} \\
&= n^{-1/2} \sum_{i=1}^{n} \left[(u^2_{i,t+h,1} - u^2_{i,t+h,2}) - E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}) \right] + O_P(n^{-1/2}).
\end{aligned}$$

□

A.7 Theorem 9

Proof. By equation (28), we have

$$\sqrt{n} \left[\sum_{k=1}^{K} \frac{n_k}{n} (\overline{e}^2_{1,k,t+h} - \overline{e}^2_{2,k,t+h}) - \sum_{k=1}^{K} \frac{n_k}{n} \left((\lambda'_{1,(k)} f_{t+h})^2 - (\lambda'_{2,(k)} f_{t+h})^2 \right) \right]$$

$$= n^{-1/2} \sum_{k=1}^{K} n_k \left\{ \left(n_k^{-1} \sum_{i \in H_k} u_{i,t+h,1} \right)^2 - \left(n_k^{-1} \sum_{i \in H_k} u_{i,t+h,2} \right)^2 \right\}$$
$$+ 2n^{-1/2} \sum_{i=1}^{n} \left(\lambda_{i,1}' f_{t+h} u_{i,t+h,1} - \lambda_{i,2}' f_{t+h} u_{i,t+h,2} \right).$$

Again as in the proof of Theorem 8, we can show that $n^{-1/2} \sum_{k=1}^{K} n_k \left\{ \left(n_k^{-1} \sum_{i \in H_k} u_{i,t+h,1} \right)^2 - \left(n_k^{-1} \sum_{i \in H_k} u_{i,t+h,2} \right)^2 \right\} = o_P(1)$, and so

$$\sqrt{n} \left[\sum_{k=1}^{K} \frac{n_k}{n} (\overline{e}_{1,k,t+h}^2 - \overline{e}_{2,k,t+h}^2) - n^{-1} \sum_{i=1}^{n} \left[(\lambda_{i,1}' f_{t+h})^2 - (\lambda_{i,2}' f_{t+h})^2 \right] \right]$$
$$= 2n^{-1/2} \sum_{i=1}^{n} \left(\lambda_{i,1}' f_{t+h} u_{i,+ht,1} - \lambda_{i,2}' f_{t+h} u_{i,t+h,2} \right) + o_P(1).$$

The rest of the proof follows by a CLT as in the proof of Theorem 8. □

A.8 Lemma 1

Proof. Since we can write $f_{s+h} = (\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}(\bar{e}_{s+h} - \bar{u}_{s+h})$, we have $e_{i,s+h} = \lambda_i'(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}\bar{e}_{s+h} + u_{i,s+h} - \lambda_i'(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}\bar{u}_{s+h}$. It is not difficult to see that

$$\hat{\lambda}_i' = \left(\sum_{s+h=1}^{T} [\lambda_i'(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}\bar{e}_{s+h} + u_{i,s+h} - \lambda_i'(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}\bar{u}_{s+h}]\bar{e}_{s+h}' \right) \left(\sum_{s+h=1}^{T} \bar{e}_{s+h}\bar{e}_{s+h}' \right)^{-1}$$
$$= \lambda_i'(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda} + \left(\sum_{s+h=1}^{T} u_{i,s+h}\bar{e}_{s+h}' \right) \left(\sum_{s+h=1}^{T} \bar{e}_{s+h}\bar{e}_{s+h}' \right)^{-1}$$
$$- \lambda_i'(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda} \left(\sum_{s+h=1}^{T} \bar{u}_{s+h}\bar{e}_{s+h}' \right) \left(\sum_{s+h=1}^{T} \bar{e}_{s+h}\bar{e}_{s+h}' \right)^{-1},$$

and thus

$$\hat{\lambda}_i'\bar{e}_{t+h} = \lambda_i' f_{t+h} + \xi_{i,t+h} + \varepsilon_{i,t+h} + \zeta_{i,t+h}.$$

where $\xi_{i,t+h} = \lambda_i'(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}\bar{u}_{t+h}$, $\varepsilon_{i,t+h} = \left(\sum_{s+h=1}^{T} u_{i,s+h}\bar{e}_{s+h}' \right) \left(\sum_{s+h=1}^{T} \bar{e}_{s+h}\bar{e}_{s+h}' \right)^{-1} \bar{e}_{t+h}$ and $\zeta_{i,t+h} = -\lambda_i'(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda} \left(\sum_{s+h=1}^{T} \bar{u}_{s+h}\bar{e}_{s+h}' \right) \left(\sum_{s+h=1}^{T} \bar{e}_{s+h}\bar{e}_{s+h}' \right)^{-1} \bar{e}_{t+h}$.

Next, observe that

$$n^{-1/2}\sum_{i=1}^{n}[(\hat{\lambda}_{i,1}'\bar{e}_{t+h})^2-(\lambda_{i,1}'f_{t+h})^2]$$
$$=n^{-1/2}\sum_{i=1}^{n}\left(\xi_{i,t+h,1}+\varepsilon_{i,t+h,1}+\zeta_{i,t+h,1}\right)^2+2n^{-1/2}\sum_{i=1}^{n}\left(\xi_{i,t+h,1}+\varepsilon_{i,t+h,1}+\zeta_{i,t+h,1}\right)\lambda_{i,1}'f_{t+h}.$$

Moreover,

$$n^{-1/2}\sum_{i=1}^{n}\xi_{i,t+h,1}\lambda_{i,1}'f_{t+h}=n^{-1/2}\bar{u}_{t+h}'\bar{\lambda}'(\bar{\lambda}\bar{\lambda}')^{-1}\left(\sum_{i=1}^{n}\lambda_{i,1}\lambda_{i,1}'\right)f_{t+h}$$

and

$$n^{-1/2}\sum_{i=1}^{n}\zeta_{i,t+h,1}\lambda_{i,1}'f_{t+h}$$
$$=-n^{-1/2}\bar{e}_{t+h}'\left(\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{e}_{s+h}'\right)^{-1}\left(\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{u}_{s+h}'\right)\bar{\lambda}'(\bar{\lambda}\bar{\lambda}')^{-1}\left(\sum_{i=1}^{n}\lambda_{i,1}\lambda_{i,1}'\right)f_{t+h}.$$

Finally, we have

$$n^{-1/2}\sum_{i=1}^{n}\varepsilon_{i,t+h,1}\lambda_{i,1}'f_{t+h}=n^{-1/2}f_{t+h}'\left(\sum_{s+h=1}^{T}\left(\sum_{i=1}^{n}\lambda_{i,1}u_{i,s+h}\right)\bar{e}_{s+h}'\right)\left(\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{e}_{s+h}'\right)^{-1}\bar{e}_{t+h}.$$

One can show that $n^{-1/2}\sum_{i=1}^{n}\varepsilon_{i,t+h,1}\lambda_{i,1}'f_{t+h}=o_P(1)$ and $n^{-1/2}\sum_{i=1}^{n}\zeta_{i,t+h,1}\lambda_{i,1}'f_{t+h}=o_P(1)$. Therefore, we have

$$n^{-1/2}\sum_{i=1}^{n}[(\hat{\lambda}_{i,1}'\bar{e}_{t+h})^2-(\lambda_{i,1}'f_{t+h})^2]=2n^{-1/2}\bar{u}_{t+h}'\bar{\lambda}'(\bar{\lambda}\bar{\lambda}')^{-1}\left(\sum_{i=1}^{n}\lambda_{i,1}\lambda_{i,1}'\right)f_{t+h}+o_P(1).$$

□

A.9 Theorem 10

Proof. We notice that

$$\begin{aligned}
&\sqrt{n}\left[\overline{\Delta\hat{u}}^2_{t+h} - n^{-1}\sum_{i=1}^{n} E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F})\right] \\
&= n^{-1/2}\sum_{i=1}^{n}\Big[(u^2_{i,t+h,1} - u^2_{i,t+h,2}) - E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}) \\
&\quad + 2(\lambda'_{i,1}f_{t+h}u_{i,t+h,1} - \lambda'_{i,2}f_{t+h}u_{i,t+h,2}) + u'_{i,t+h}D_{t+h}\Big], \qquad (45)
\end{aligned}$$

where $D_{t+h} = \bar{\lambda}'(\bar{\lambda}\bar{\lambda}')^{-1}\left(n^{-1}\sum_{i=1}^{n}[\lambda_{i,1}\lambda'_{i,1} - \lambda_{i,2}\lambda'_{i,2}]\right)f_{t+h}$. Since $\hat{\lambda}'_i - \lambda'_i(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda} = o_P(1)$, $\bar{e}_{t+h} = \bar{\lambda}'f_{t+h} + o_P(1)$ and $(\bar{\lambda}\bar{\lambda}')^{-1}$ exists asymptotically, we have $\hat{D}_{D_{t+h}} = D_{t+h} + o_P(1)$. Since $\{u_{i,t+h,m}\}_{i=1}^n$ is independent across i, the result then follows by the classical CLT and a self-normalized CLT; see e.g., Theorem 4.1 of Chen et al. (2016); Peña et al. (2008). □

A.10 Theorem 11

Proof. By Lemma 1, we have that under the null hypothesis,

$$n^{-1/2}\sum_{i=1}^{n}\left[(\hat{\lambda}'_{i,1}\bar{e}_{t+h})^2 - (\hat{\lambda}'_{i,2}\bar{e}_{t+h})^2\right] = 2n^{-1/2}\bar{u}'_{t+h}\bar{\lambda}'(\bar{\lambda}\bar{\lambda}')^{-1}\sum_{i=1}^{n}\left(\lambda_{i,1}\lambda'_{i,1} - \lambda_{i,2}\lambda'_{i,2}\right)f_{t+h} + o_P(1).$$

Since $\{u_{i,t+h,m}\}_{i=1}^n$ is independent across i, the result then follows by the classical CLT and a self-normalized CLT; see e.g., Theorem 4.1 of Chen et al. (2016); Peña et al. (2008). □

A.11 Lemma 2

Proof. Under Assumptions A-F and Theorem 3 in Bai (2003), recall that the following result holds:

$$\hat{\lambda}'_i\hat{f}_{t+h} - \lambda'_i f_{t+h} = n^{-1}\lambda'_i\left(n^{-1}\sum_{j=1}^{n}\lambda_j\lambda'_j\right)^{-1}\sum_{j=1}^{n}\lambda_j u_{jt+h}$$

$$+ T^{-1} f'_{t+h} \left(T^{-1} \sum_{s+h=1}^{T} f_{s+h} f'_{s+h} \right)^{-1} \sum_{s+h=1}^{T} f_{s+h} u_{is+h} + O_P(1/\min\{n,T\}).$$

Using this result, we have $\hat{\lambda}'_{i,m} f_{t+h} = \lambda'_{i,m} f_{t+h} + \xi_{i,t+h,m}$ for $m \in \{1,2\}$, where

$$\begin{aligned}\xi_{i,t+h,m} = {} & n^{-1} \lambda'_i \left(n^{-1} \sum_{j=1}^{n} \lambda_j \lambda'_j \right)^{-1} \sum_{j=1}^{n} \lambda_{j,m} u_{jt+h,m} \\ & + T^{-1} f'_{t+h} \left(T^{-1} \sum_{s+h=1}^{T} f_{s+h} f'_{s+h} \right)^{-1} \sum_{s+h=1}^{T} f_{s+h} u_{is+h,m} + O_P(1/\min\{n,T\}).\end{aligned}$$

It follows that

$$\begin{aligned}n^{-1} \sum_{i=1}^{n} \left[(\hat{\lambda}'_{i,1} \hat{f}_{t+h})^2 - (\hat{\lambda}'_{i,2} \hat{f}_{t+h})^2 \right] = {} & n^{-1} \sum_{i=1}^{n} [(\lambda'_{i,1} f_{t+h})^2 - (\lambda'_{i,2} f_{t+h})^2] \\ & + 2n^{-1} \sum_{i=1}^{n} [\lambda'_{i,1} f_{t+h} \xi_{i,t+h,1} - \lambda'_{i,2} f_{t+h} \xi_{i,t+h,2}] \\ & + n^{-1} \sum_{i=1}^{n} [\xi^2_{i,t+h,1} - \xi^2_{i,t+h,2}].\end{aligned}$$

The last term is of order $1/\min\{n,T\}$, which is negligible if $\sqrt{n}/T = o(1)$. Under assumptions of weak (cross-sectional and serial) dependence in $u_{i,t+h,m}$ (e.g., Assumptions E and F in Bai (2003)), we can show that

$$n^{-1} \sum_{i=1}^{n} \lambda'_{i,m} f_{t+h} \xi_{i,t+h,m} = n^{-1} \sum_{i=1}^{n} \lambda'_{i,m} f_{t+h} u_{i,t+h,m} + o_P(n^{-1/2}).$$

Using this, it follows that

$$\begin{aligned}n^{-1} \sum_{i=1}^{n} \left[(\hat{\lambda}'_{i,1} \hat{f}_{t+h})^2 - (\hat{\lambda}'_{i,2} \hat{f}_{t+h})^2 \right] = {} & n^{-1} \sum_{i=1}^{n} [(\lambda'_{i,1} f_{t+h})^2 - (\lambda'_{i,2} f_{t+h})^2] \\ & + 2n^{-1} \sum_{i=1}^{n} [\lambda'_{i,1} f_{t+h} u_{i,t+h,1} - \lambda'_{i,2} f_{t+h} u_{i,t+h,2}] + o_P(n^{-1/2}).\end{aligned}$$

The stated result follows from this. □

Table 1: Tests of Equal Predictive Accuracy

	CE				AR	
	h=1, S	h=1, F	h=0, S	h=0, F	h=1, S	h=1, F
Panel A: GDP Growth						
t-stat pooled average	0.72	0.36	0.84	-0.07	1.05	2.25
p-value	(0.46)	(0.71)	(0.39)	(0.93)	(0.29)	(0.02)
t-stat time clusters	0.51	0.37	0.90	-0.09	0.98	1.93
p-value	(0.61)	(0.71)	(0.37)	(0.92)	(0.33)	(0.06)
Randomization p-value, 1-year clusters	(0.61)	(0.79)	(0.38)	(0.92)	(0.48)	(0.02)
Randomization p-value, GFC clusters	(0.25)	(0.49)	(0.24)	(0.75)	(0.00)	(0.00)
t-stat group clusters	1.09	0.40	1.58	-0.01	1.53	2.94
p-value	(0.33)	(0.70)	(0.18)	(0.98)	(0.17)	(0.02)
Randomization p-value group clusters	(0.31)	(0.56)	(0.12)	(0.87)	(0.04)	(0.00)
Panel B: Inflation						
t-stat pooled average	1.42	0.73	2.86	2.25	4.81	7.24
p-value	(0.15)	(0.46)	(0.00)	(0.02)	(0.00)	(0.00)
t-stat time clusters	1.59	0.83	3.17	2.59	4.46	7.72
p-value	(0.12)	(0.41)	(0.00)	(0.01)	(0.00)	(0.00)
Randomization p-value, 1-year clusters	(0.12)	(0.41)	(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	1.01	0.72	2.27	1.55	6.14	5.56
p-value	(0.36)	(0.50)	(0.08)	(0.19)	(0.00)	(0.00)
Randomization p-value group clusters	(0.37)	(0.56)	(0.00)	(0.18)	(0.00)	(0.00)

Notes:: This table reports the outcome of tests of equal predictive accuracy, comparing the IMF World Economic Outlook (WEO) forecasts to Consensus Economics (CE, first four columns) and autoregressive (AR, columns five and six) forecasts. 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. Similarly, small p-values indicate better performance of the WEO forecasts relative to the CE or AR forecasts. 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 2010-2016. 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). All p-values are based on two-sided tests. h is the forecast horizon and refers to either current-year forecasts ($h = 0$) or next-year forecasts ($h = 1$). WEO and CE forecasts are recorded for the spring (S) and fall (F) of the current and previous year. Panel A uses real GDP growth forecasts, while Panel B uses inflation data.

Table 2: **Tests of Equal Predictive Accuracy Across Economic Groupings**

Panel A: GDP Growth

WEO vs. CE

	ae	eeur	lac	cis	dms
h=1, S	0.82	0.20	1.31	-0.47	-0.98
h=1, F	1.16	0.10	0.34	-0.96	-0.57
h=0, S	1.70	-0.79	1.60	0.45	-0.06
h=0, F	0.79	-0.92	-0.32	0.16	0.12

WEO vs. AR

	ae	eeur	dasia	lac	menap	cis	ssa
h=1, S	1.39	2.65	0.48	0.27	0.79	1.57	-0.16
h=1, F	1.74	3.20	1.52	2.45	0.94	1.85	1.12

Panel B: Inflation

WEO vs. CE

	ae	eeur	lac	cis	dms
h=1, S	0.06	0.26	-1.65	1.51	0.41
h=1, F	-0.42	1.38	-1.47	2.13	0.01
h=0, S	-0.08	-0.53	2.29	2.60	2.46
h=0, F	-1.14	-1.27	3.60	1.98	2.57

WEO vs. AR

	ae	eeur	dasia	lac	cis	ms
h=1, S	4.37	2.50	4.87	4.33	2.46	3.30
h=1, F	8.78	2.66	5.00	6.05	2.94	4.76

Notes:: This table reports the outcome of 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. 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 3: 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	0.37	1.65	2.18	5.62
p-value	(0.70)	(0.09)	(0.02)	(0.00)
t-stat time clusters	0.38	1.23	2.28	4.97
p-value	(0.70)	(0.22)	(0.03)	(0.00)
Randomization p-value, 1-year clusters	(0.75)	(0.24)	(0.00)	(0.00)
Randomization p-value, GFC cluster	(0.00)	(0.00)	(0.00)	(0.00)
t-stat group clusters	0.01	2.65	2.05	3.57
p-value	(0.99)	(0.03)	(0.08)	(0.01)
Randomization p-value group clusters	(0.98)	(0.00)	(0.00)	(0.00)
Panel B: Inflation				
t-stat pool	2.68	5.71	6.57	6.90
pval	(0.00)	(0.00)	(0.00)	(0.00)
t-stat time cluster	2.50	6.31	7.15	7.54
pval	(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	(0.00)	(0.00)	(0.00)	(0.00)
t-stat region cluster	2.75	5.70	3.63	5.06
pval	(0.03)	(0.00)	(0.01)	(0.00)
Randomization p-value region cluster	(0.01)	(0.00)	(0.00)	(0.00)

Notes: This table reports the outcome of tests of equal squared error predictive accuracy comparing the IMF World Economic Outlook (WEO) forecasts across different forecast horizons. h is the forecast horizon and refers to either next-year ($h = 1$) or current-year ($h = 0$) Spring (S) or Fall (F) forecasts. Positive test statistics (small p-values) indicate that the forecasts computed at the short horizon are more accurate than the forecasts computed at the longer horizon, i.e., that predictive accuracy improves as the forecast horizon is reduced. 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 2010-2016. 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). All p-values are based on two-sided tests. Panel A uses real GDP growth forecasts, while Panel B uses inflation data.

Table 4: **Tests of Equal Predictive Accuracy Across Different Forecast Horizons: Results by economic groupings**

Panel A: GDP Growth					
	ae	eeur	lac	cis	dms
h=1,S vs h=1,F	1.87	2.29	3.60	-0.90	0.88
h=1,F vs h=0,S	2.29	2.28	3.57	1.42	0.88
h=0,S vs h=0,F	5.22	1.99	8.21	1.87	1.83
h=1,S vs h=0,F	2.60	2.45	4.83	2.45	4.22
Panel B: Inflation					
	ae	eeur	lac	cis	dms
h=1,S vs h=1,F	2.64	1.49	2.73	1.30	2.06
h=1,F vs h=0,S	5.29	4.25	3.63	3.79	5.56
h=0,S vs h=0,F	5.63	1.68	5.26	4.66	5.47
h=1,S vs h=0,F	5.40	3.09	5.81	4.11	6.51

Notes:: This table reports the outcome of tests of equal squared error predictive accuracy comparing the IMF World Economic Outlook (WEO) forecasts across different forecast horizons. h is the forecast horizon and refers to either next-year ($h = 1$) or current-year ($h = 0$) Spring (S) or Fall (F) forecasts. Positive test statistics (small p-values) indicate that the predictive accuracy improves as the forecast horizon is reduced. In each panel, each row reports the 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. 'dms' combines three IMF groups, namely Middle East, North Africa, Afghanistan, and Pakistan, emerging and developing Asia, and Sub-Sahara Africa.

Table 5: Differences in idiosyncratic error variance and squared bias components in the squared error loss.

Panel A: GDP growth

	Difference in idiosyncratic variance				Difference in squared bias			
	1990-2016		2007-2009		1995-2016		2007-2009	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Cluster	0.16	0.09	0.59	1.00	-0.02	0.05	1.41	1.04
CCE	0.05	0.12	0.16	0.12	0.08	0.00	1.85	0.69
PCA	0.05	0.12	1.09	1.01	0.09	0.01	0.91	1.03

Panel B: Inflation

	Difference in idiosyncratic variance				Difference in squared bias			
	1990-2016		2007-2009		1995-2016		2007-2009	
Cluster	0.74	0.30	2.35	2.60	0.56	0.25	1.10	0.67
CCE	0.67	0.16	1.74	0.85	0.63	0.29	1.71	0.69
PCA	0.35	0.05	0.86	0.28	0.96	0.40	2.59	1.56

Notes:: This table reports the mean and median values of the difference in the idiosyncratic error variance (left panel) and squared bias components (right panel) of the mean squared forecast errors computed for the Consensus Economics (CE) versus IMF WEO forecasts. Positive values indicate that the idiosyncratic error variance or squared bias is larger for the CE forecasts than for the WEO forecasts, while negative values suggest the reverse. Forecast errors have been normalized before computing the cross-sectional average and the results refer to current-year spring forecasts. Panel A uses errors from forecasting GDP growth, while Panel B refers to inflation forecast errors. The first row in each column assumes a cluster structure on factor loadings while the second and third rows use the CCE and PCA estimation methods to compute factors.

Figure 1: Forecast Horizon Used in the WEO Forecasts.

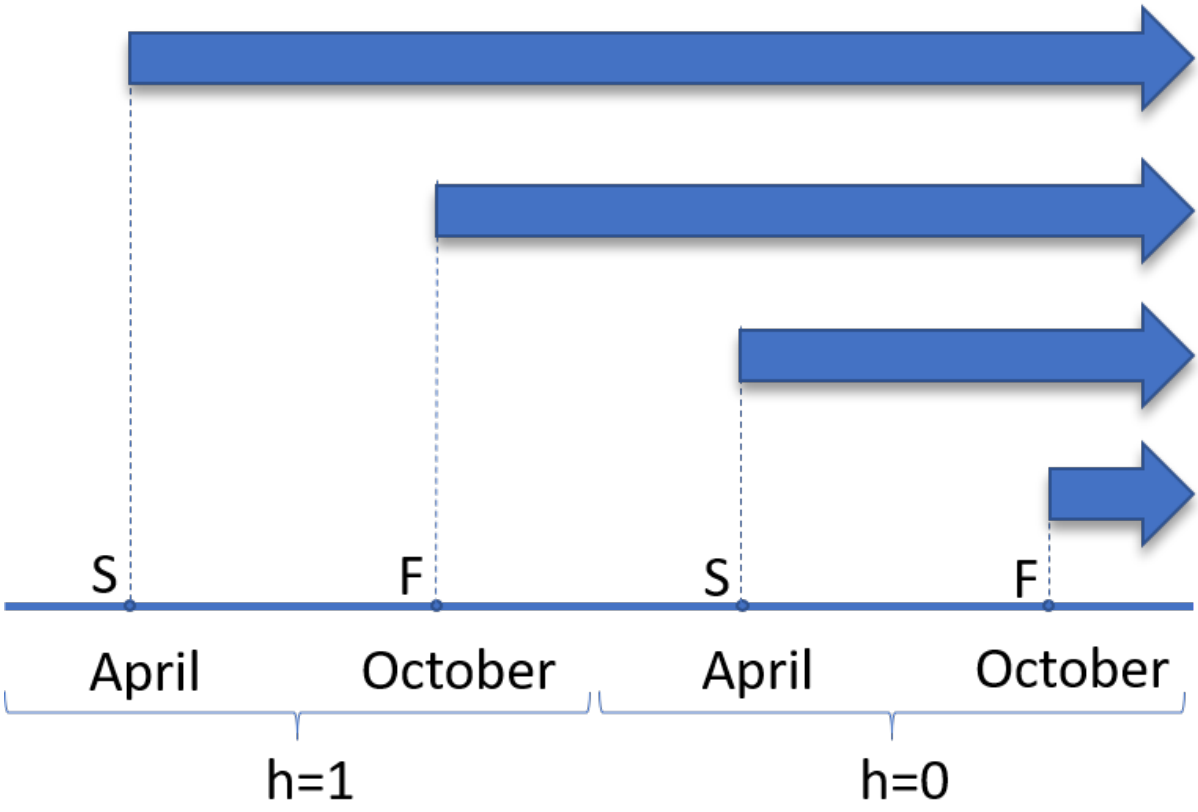
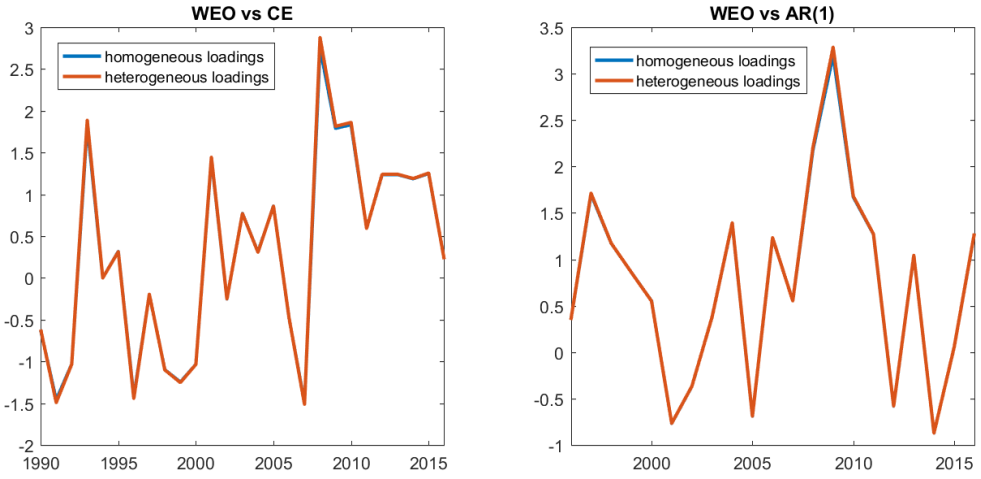


Figure 2: Cross-sectional test-statistics comparing predictive accuracy in individual years in the sample. Positive values of the test statistics indicate that the WEO forecasts are more accurate than either the Consensus Economics (CE) or autoregressive (AR) forecasts. Negative values suggest the reverse. The figures use current year spring forecasts in the comparison of WEO and CE forecasts and one-year ahead fall forecast when comparing WEO and AR(1) forecasts.

(a) GDP



(b) Inflation

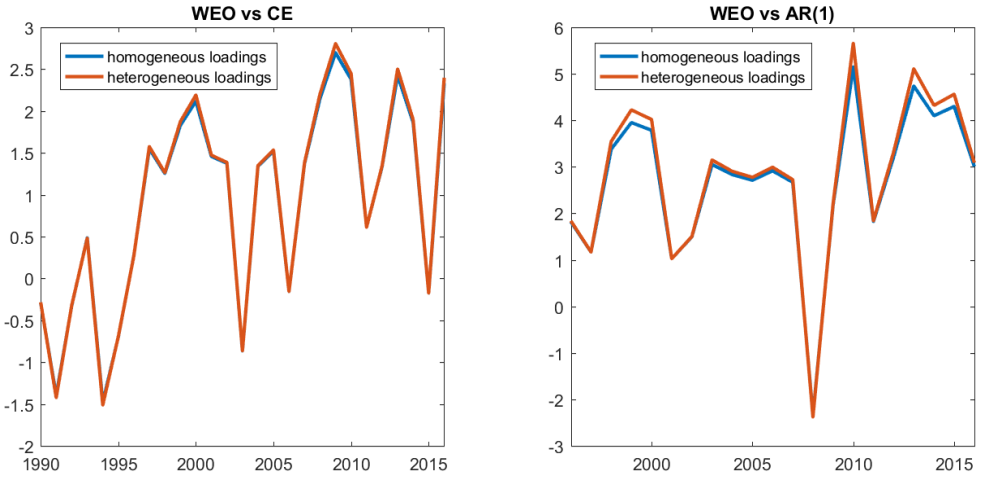
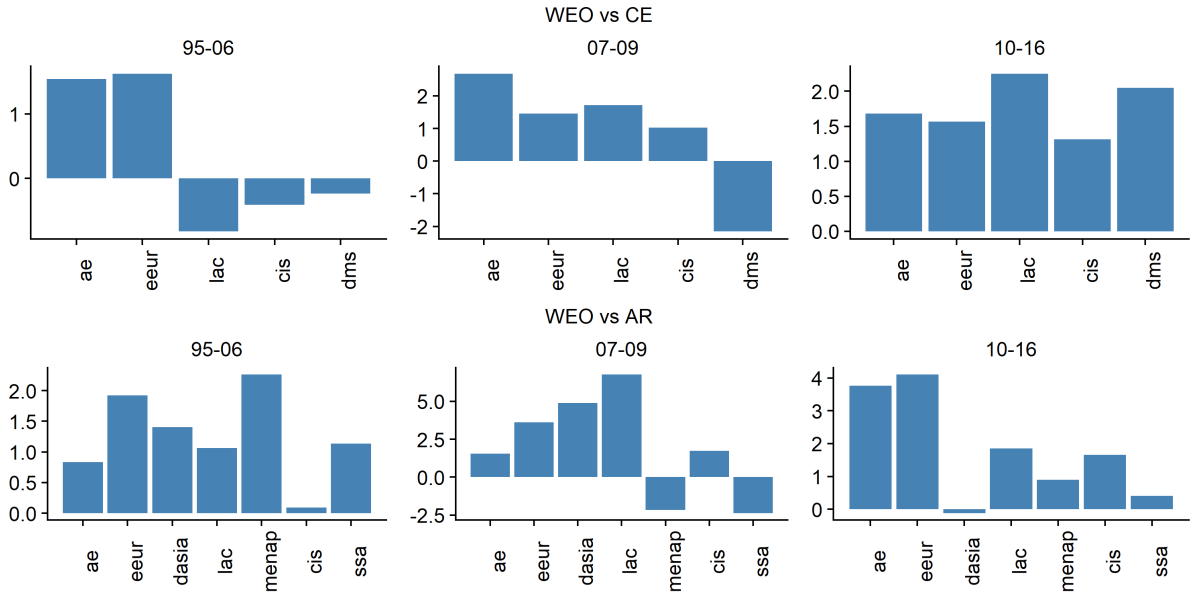


Figure 3: This figure presents t-statistics for comparing the null of equal expected squared error loss for WEO versus CE (panels in rows 1 and 3) current-year spring (h = 0, S) forecasts or WEO versus AR (panels in rows 2 and 4) next-year fall (h = 1, F) forecasts of GDP growth (top two rows) and inflation (bottom two rows). The test statistics are computed for three time clusters, namely 1995-2006, 2007-2009 (Global Financial Crisis), and 2010-16 and for five groups of countries, namely advanced economies (ae), emerging and developing Europe (eeur), Latin America and Caribbean (lac), Commonwealth of Independent State (cis) and a fifth group containing the remaining countries (dms). Positive test statistics suggest that WEO forecasts are more accurate than the alternative forecasts (CE or AR).

(a) GDP



(b) Inflation

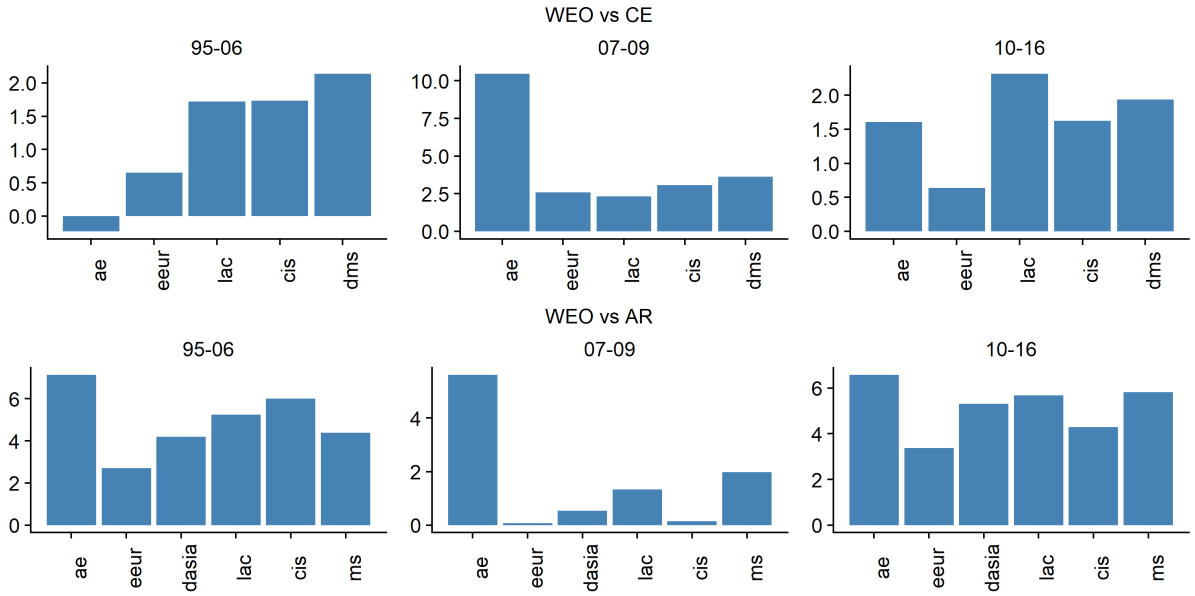


Figure 4: Cross-sectional test-statistics comparing the accuracy in individual years in the sample of WEO forecasts of GDP growth and inflation produced at different forecast horizons, h. Positive values of the test statistics indicate that the WEO forecasts produced at the shorter forecast horizon are more accurate than the forecasts generated at a longer horizon. Negative values suggest the opposite.

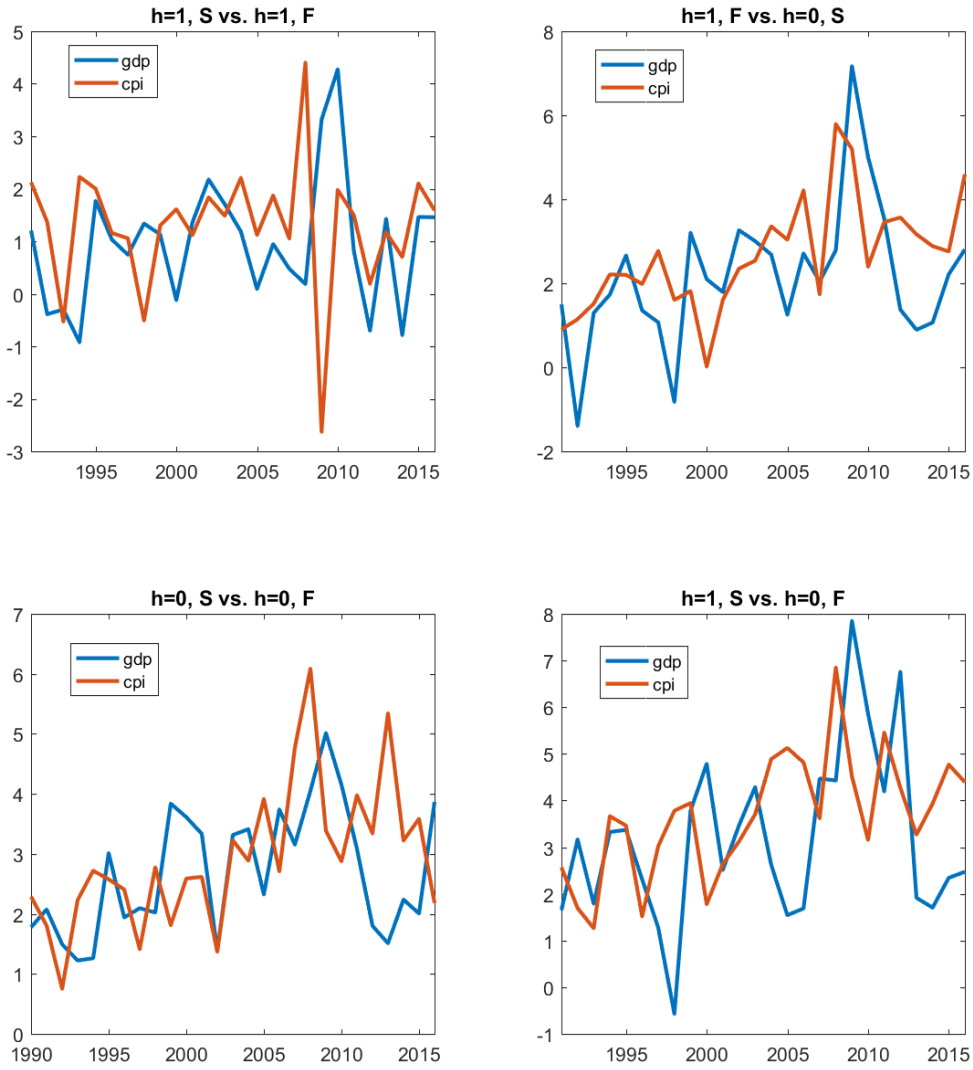
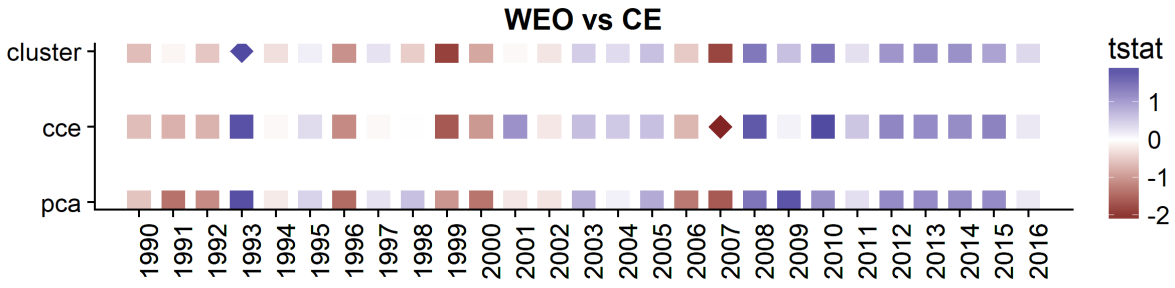
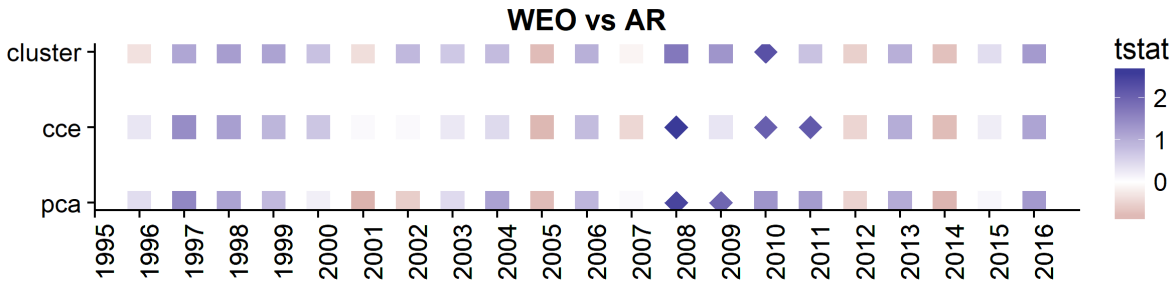


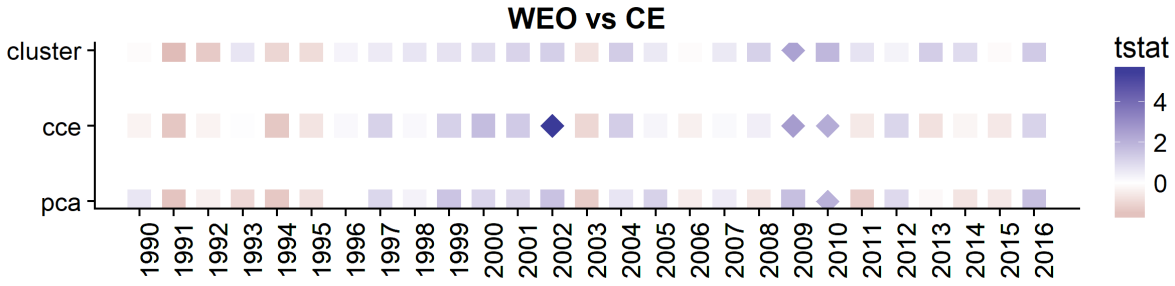
Figure 5: Cross-sectional comparisons of the magnitude of idiosyncratic variances in individual years in the sample. Each panel shows the outcome of a cross-sectional test of the null that two forecasts have the same idiosyncratic error variance in a given year. The first and third rows of panels compare WEO and Consensus Economics forecasts while rows two and four compare WEO forecasts to predictions from a simple autoregressive model. Blue color indicates that the idiosyncratic error variance component of the WEO forecasts is smaller than the corresponding value for the alternative forecast. Red color indicates the reverse. The first line in each panel is calculated by assuming identical factor loadings within each cluster. The second and third lines of each panel estimate the factors by CCE or PCA, respectively. The figures use current year spring forecasts to compare the idiosyncratic error variance component of the WEO and CE forecasts and one-year ahead fall forecast when comparing WEO and AR(1) forecasts. Diamond shapes indicate that the null of equal idiosyncratic error variance is rejected at the 5% level.

(a) GDP





(b) Inflation



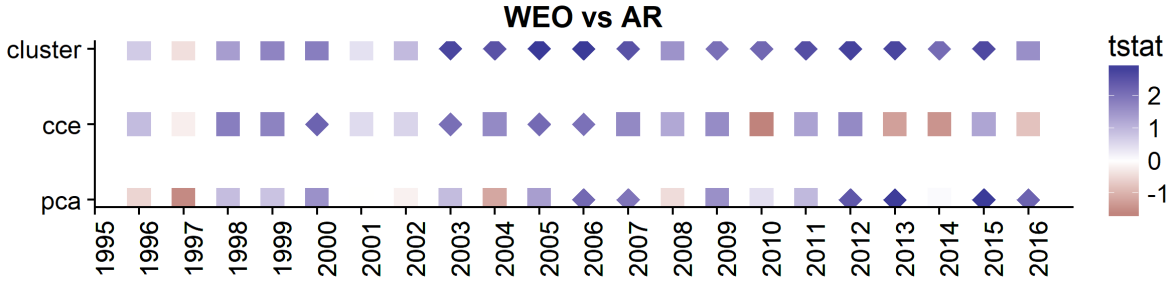
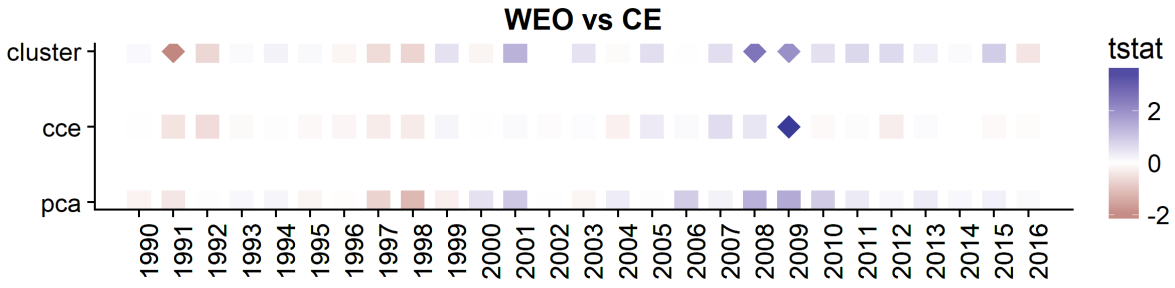


Figure 6: Cross-sectional comparisons of the magnitude of squared biases in individual years in the sample. Each panel shows the outcome of a cross-sectional test of the null that two forecasts have the same squared bias in a given year. The first and third rows of panels compare WEO and Consensus Economics forecasts while rows two and four compare WEO forecasts to predictions from a simple autoregressive model. Blue color indicates that the squared bias component of the WEO forecasts is smaller than the corresponding value for the alternative forecast. Red color indicates the reverse. The first line in each panel is calculated by assuming identical factor loadings within each cluster. The second and third lines of each panel estimate the factors by CCE or PCA, respectively. The figures use current year spring forecasts to compare the squared bias component of the WEO and CE forecasts and one-year ahead fall forecast when comparing WEO and AR(1) forecasts. Diamond shapes indicate that the null of equal squared bias is rejected at the 5% level.

(a) GDP



(b) Inflation