Do *Any* Economists Have Superior Forecasting Skills?

Ritong Qu* Allan Timmermann† Yinchu Zhu‡

October 30, 2019

To answer this question, we develop new testing methods for identifying superior forecasting skills in settings with arbitrarily many forecasters, outcome variables, and time periods. Our methods allow us to address if *any* economists had superior forecasting skills for *any* variables or at *any* point in time while carefully controlling for the role of "luck" which can give rise to false discoveries when large numbers of forecasts are evaluated. We propose new hypotheses and test statistics that can be used to identify specialist, generalist, and event-specific skills in forecasting performance. We apply our new methods to a large set of Bloomberg survey forecasts of US economic data show that, overall, there is very little evidence that any individual forecasters can beat a simple equal-weighted average of peer forecasts.

Key words: Economic forecasting; superior predictive skills; panel data; multiple hypothesis testing; bootstrap; Bloomberg survey.

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

†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

News media and the popular press frequently report stories about forecasters who correctly predicted major economic and financial calamities such as the 2007-2009 mortgage market meltdown and the ensuing global financial crisis or "Black Monday" when the British pound depreciated sharply. Academic research has argued for the existence of "super forecasters" who possess extraordinary judgment and an innate ability to produce forecasts more accurate than their peers, see e.g., Tetlock and Gardner (2016). Often lost in this discussion is the fact that the pursuit of spectacularly accurate forecasts introduces a multiple hypothesis testing problem as they may be the result of an extensive search conducted over a potentially very large set of outcome variables, forecasters and time periods. If not properly controlled for, this process is likely to spuriously identify many cases with seemingly superior forecasting skills, wrongly attributing luck to predictive skill.

This paper develops new methods for conducting inference on the existence of forecasters with superior predictive skills in a panel data setting in which we observe forecasts of multiple variables reported by many forecasters for a large number of time periods. We pose our economic hypotheses as "Sup" tests which state that the benchmark forecasts are at least as accurate as all the forecasts from some alternative set. Our analysis exploits the panel data structure to conduct tests of the existence of superior forecasting skills for *any* economic forecaster, for *any* outcome variable, or at *any* point in time. The existence of both a cross-sectional and a time-series dimension for an arbitrarily large set of individual forecasters introduces a high-dimensional multiple hypothesis testing problem–as potentially many performance measures are being compared–which our methods carefully handle so as to control the false discovery rate.

To more accurately identify evidence of superior forecasting performance, we develop new economic hypotheses and associated tests that provide information about the nature and source of any superior skills that forecasters may possess. First, comparing forecasting performance across individual variables or subsets of variables with similar features, we can gain a sense of whether forecasters are *specialists*, with the ability to produce highly accurate forecasts for a particular variable or subset of variables, such as a certain type of economies (e.g., developing markets) or a certain type of firms (e.g., technology firms). Economic forecasters with superior ability to predict the outcome of a single or a few variables with common characteristics would be consistent with domain-specific expertise and is suggestive of forecasters with superior private information or analytical skills pertaining to a narrow set of variables. Conversely, common access to a large amount of public information about a particular variable (unit) creates a more level playing field which would make it more difficult for individual forecasters to produce forecasts with superior accuracy relative to

their peers.

Second, comparing individual forecasters' mean performance–averaged across many different variables–we can test whether some forecasters have superior *generalist* skills, indicating all-round forecasting abilities. This type of skill could arise as a result of some forecasters' ability to produce superior forecasts of common factors influencing the majority of variables. This type of skill would allow forecasters to produce accurate forecasts not only for a narrow set of variables but across the board.

Third, comparing individual forecasters' average performance across multiple variables over a short event window, or even in a single time period, we can test whether any forecaster possessed superior predictive ability during a particular period. For example, we might be interested in testing for superior predictive ability during the Global Financial Crisis to see whether any individual forecasters correctly anticipated this event. Because estimation of predictive skills in a single period cannot rely on time-series averages, event-specific skills can be detected only if this information shows up in better forecasting performance when averaged across multiple variables in that period.

The distinction between specialist and generalist skills is important for understanding the source of forecasting skills. Since private information is less likely to be available for (or even relevant to) a large set of heterogeneous variables, generalist skills are indicative of a forecaster's ability to process and analyze public information, suggesting that the forecaster may have superior modeling skills. Agents' forecasting skills reflect not only the economic signals they observe but also their efforts in distilling and processing information from such signals. Forecasters' effort level and, hence, the expected precision of their forecasts, can be viewed as the outcome of a constrained optimization problem: Forecasters have limited resources for processing information and can choose to focus predominantly on variable-specific information (specialists) or, conversely, on general information (generalists) based on the marginal cost and benefit of information acquisition and processing which can vary across investors and across time (economic states). Our economic hypotheses can, therefore, be used to test some of the implications of recent theoretical work on endogenous information acquisition and rational inattention. As a first example, Van Nieuwerburgh and Veldkamp (2010) show that investors should collect a disproportionate amount of information on those assets that they expect to overweight in their portfolio. Conversely, they should hold a balanced portfolio of the remaining assets–for which they do not possess an informational advantage–for diversification purposes. This type of strategy would be consistent with some investors possessing an informational advantage for a limited set of stocks (or industries) which they are overexposed to. As a second example, Mackowiak and Wiederholt (2009) develop a model in which firms with limited attention rationally choose to pay most attention to the more volatile firm-specific shocks and disregard aggregate shocks. This suggests

that firms within a specific industry may have an advantage at predicting outcomes in their industry and, thus, possess specialist skills but not generalist skills.[1] Our hypotheses allows us to specifically test whether specialist forecasting skills (predictive ability limited to variables with similar features) are more common than generalist skills (predictive ability across very different variables).

Outcomes in any individual period are likely to be strongly cross-sectionally correlated, so good forecasting performance across many variables in a particular period could simply be due to luck, i.e., the result of a forecaster essentially making one good judgment about the realization of a highly influential (global) factor. Provided that the cross-sectional dimension of the data set is large enough–or, equivalently, that the forecasts of sufficiently many variables are being compared–we show that we can perform valid comparisons of predictive accuracy in a *single* cross-section, i.e., for a *single* time period. This result requires us to apply a cross-sectional central limit theorem. We establish conditions under which this can be justified in the context of a model that decomposes the forecast errors of the individual variables into a correlated component that is driven by exposures to a set of common factors and an uncorrelated idiosyncratic component.

Again, time- or state-dependence in predictive skills–ideally suited for being identified through our new cross-sectional tests–can be used to test recent economic theories. For example, Kacperczyk et al. (2014) and Kacperczyk et al. (2016) propose a model in which information about the general performance of the stock market (as opposed to stock-specific information) is most valuable in recessions for fund managers attempting to time the market. The idea is that the common (market) component in stock returns is more important in recessions which are periods with high aggregate market volatility and macroeconomic uncertainty. As a second example, in short-term bond markets investors and fund managers may choose only to acquire information that allows them to generate more accurate forecasts of asset and redemption values in certain "information sensitive" states, while they prefer not to acquire costly information in more normal states.[2] Because large, sophisticated investors may have an advantage at collecting and processing information compared with smaller investors, we would expect to find greater evidence of heterogeneity in predictive skills in the information sensitive states. Conducting separate tests for superior predictive skills in normal versus information-sensitive states would be one way to address these implications

[1] Afrouzi (2019) analyses a setting in which firms operating in more competitive industries monitor prices more accurately because of the larger benefits to such firms from being able to rapidly change prices following unanticipated aggregate shocks. This makes it optimal for firms in competitive industries to acquire more information than firms in non-competitive industries with more stale prices. In turn, this should enable firms in competitive industries to more accurately predict price dynamics in their industry and so suggests that we should expect to find the strongest evidence of specialist forecasting skills in highly competitive industries.

[2] See Gorton and Ordonez (2014) and Gallagher et al. (2019) for further discussion of this point.

from theory.

Our Sup tests apply and extend the bootstrap methods recently proposed by Chernozhukov et al. (2018). We use the bootstrap to test the null hypothesis of equal predictive accuracy for a potentially large set of forecast comparisons (multiple moment conditions) against the alternative that the null gets rejected in at least one case. The bootstrap procedure is easy to implement and, in addition to testing the null that no forecast is more accurate than a given benchmark, allows us to identify the variables, forecasters, or time periods for which the benchmark is beaten. To the best of our knowledge, our approach provides the first tests of equal predictive accuracy conducted over multiple units in a panel setting.

Although our analysis builds on Chernozhukov et al. (2018), there are also some important technical differences between our proposed method and theirs. In particular, we develop studentized test statistics that apply to dependent data, whereas Chernozhukov et al. (2018) study a non-studentized test statistic (see their Appendix B.2). In many empirical applications, the scale of forecast errors can differ drastically across units and/or time and so normalizing test statistics to have unit variance under the null hypotheses typically improves the power of the test (Hansen (2005)). In practice, this really matters as we demonstrate both through Monte Carlo simulations and through empirical work.

Our paper makes several contributions to the existing literature on multiple comparisons of predictive performance.[3] The seminal paper of White (2000) and subsequent work by Hansen (2005), Romano and Wolf (2005), Hansen et al. (2011) address the multiple hypothesis problem in settings with a large-dimensional model space but a single dependent variable. These studies, therefore, rely on test statistics based on time-series averages. In contrast, we consider inference in a number of different settings. First, we consider a setting with a low-dimensional model (forecaster) space but multiple variables. This allows us to address questions such as whether a particular forecaster (model) is more accurate than the benchmark and, if so, for which variables. Second, we consider the case where the dimensions of both the forecaster and the cross-section are large, retaining the ability to identify which units and which models (forecasters) produce superior predictive performance. Third, we generalize results in the extant literature to a setting that uses a single cross-section to conduct inference on the average forecasting performance across a large number of (cross-sectional) units in a single time period.

We use our new methods to address whether economic forecasters possess superior fore-

[3] An earlier literature develops methods for conducting inference on the relative accuracy of a pair of forecasting models. For example, Chong and Hendry (1986) develop tests of forecast encompassing, while Diebold and Mariano (1995) and West (1996) develop distribution theory and propose statistics for testing the null of equal predictive accuracy for the non-nested case. Clark and McCracken (2001) and Giacomini and White (2006) develop methods for testing equal predictive accuracy for forecasts generated by nested models.

casting skills in an empirical application that considers a large set of monthly forecasts of 14 economic variables from a survey conducted by Bloomberg. The data set spans the period from 1997 to 2019 and covers hundreds of individual forecasters and firms, leading to more than one thousand forecast comparisons in some of our tests. We find significant evidence that the best forecasters can beat a simple autoregressive benchmark. While pairwise forecast comparisons suggest that some of the individual forecasters have the ability to outperform a simple equal-weighted average of their peers, once we account for the multiple hypothesis testing problem, we fail to find meaningful empirical evidence of any type of superior predictive skills either for individual variables or "on average", across variables.

We emphasize that our results should not be interpreted as suggesting that economic forecasters have no forecasting skills. Comparing the predictive accuracy of individual forecasters to the precision of forecasts from a simple autoregressive model, in fact we find numerous cases in which individual forecasters are significantly more accurate than this statistical model. Our empirical implementation interprets superior forecasting skills as the ability of individual forecasters to beat a simple equal-weighted average of their peers. We use this as our benchmark because it is often an available option for forecast users. However, we also acknowledge that this is not an easy benchmark to beat given the evidence on the overall good performance of simple equal-weighted forecast combinations, see, e.g., Timmermann (2006); Genre et al. (2013).[4]

Our second application uses our methodology to analyze the "term structure" of forecast errors and sheds light on the timing of the information flow that allows forecasts to become more accurate as the time to the predicted event becomes shorter. Specifically, we analyze country-level forecasts of inflation and GDP growth from the IMF's World Economic Outlook (WEO) publication recorded over the 27-year period from 1990 to 2016. Forecasts for next year and the current year are reported in the Spring and Fall WEO issues, giving us four different points on the term structure of forecasting performance. As the forecast horizon gets shorter and more information becomes available, we would expect forecasts to become more accurate. Comparing the forecast accuracy across long and short(er) horizons, we test if individual country-level forecasts are indeed becoming significantly more accurate over time and, if so, between which revision points we observe the largest improvements. Empirically, we find little evidence of systematic improvements in the accuracy of the next-year forecasts

[4] An alternative strategy for identifying superior forecasting skills would be to use the model confidence set of Hansen et al. (2011). This approach can be used to identify forecasters with better performance than others without benchmarking the performance against a specific alternative such as the equal-weighted peer-group average. As such, this approach only ranks relative forecasting performance and risks identifying a group of forecasters with "less bad" forecasting performance than other forecasters. For example, in a sample that includes some very bad forecasters and some forecasters with somewhat better but still poor performance, the approach simply eliminates the worst forecasters. It is not clear that the remaining forecasters can be viewed as being superior in a meaningful way.

between the Spring and Fall WEO issues, whereas there is strong evidence that current-year forecasts improve between the spring and fall. Because we observe forecasts for 180 different countries spanning different regions and types of economies, the WEO data is well suited for identifying domain-specific skills. We find that forecast improvements are strongest for advanced economies and generally far weaker for emerging market and developing economies, consistent with data quality being a challenge for the latter economies.

The outline of the paper is as follows. Section 2 introduces our economic hypotheses, while Section 3 describes our test methodology. Section 4 explains our methodology for conducting comparisons of forecasting performance on a single cross-section. Section 5 conducts an empirical analysis of the Bloomberg survey of forecasters, while Section 6 uses our methods to study the term structure of forecasting performance for the IMF's WEO forecasts. Section 7 concludes. Monte Carlo simulation results and technical proofs are contained in a set of Appendices.

2 Identifying Superior Predictive Skills in Panel Data

This section first introduces our framework for evaluating and comparing forecasting performance in a panel setting. Next, we develop a set of hypotheses that can be used to identify different types of predictive skills reflecting forecasters with domain-specific (specialist) or more generalist abilities. Throughout the analysis, we emphasize the importance of accounting for the multiple hypothesis testing problem that arises when many test statistics are being compared.

2.1 Setup

Consider a set of h-period-ahead forecasts of a panel of variables (units) $i = 1, ..., N$ observed over T time periods $t + h = 1, ..., T$. We refer to the outcome of variable i at time $t + h$ as $y_{i,t+h}$ and the associated forecast as $\hat{y}_{i,t+h|t,m}$ and assume that there are $m = 1, ..., M$ different forecasters or models. Our notation reflects that h-step-ahead forecasts of $y_{i,t+h}$ are generated at time t. In all cases, the forecast horizon, $h \geq 0$, is a non-negative integer.

This type of panel data gives us three dimensions over which to compute averages and conduct inference: variables $(i = 1, ..., N)$, forecasters $(i = 1, ..., M)$, and time periods $(t + h = 1, ..., T)$.[5] This is important not only because distributional results for time series require different assumptions than results for cross-sectional averages, but also because averages taken over different dimensions of the data can be used to test very different economic

[5] The forecast horizon, h, can be considered as a fourth dimension. However, we ignore this in our analysis because many data sets only cover a single forecast horizon and inference across multiple horizons is not very common.

hypotheses as we explain in this section and in Section 4.

To evaluate forecasting performance, we need a loss function. Following standard practice (e.g., Granger (1999)), we assume that the loss function takes as its inputs the outcome and the forecast, $L(y_{i,t+h}, \hat{y}_{i,t+h|t,m})$, and maps these values to the real line. By far the most common loss function used in applied work is squared error loss which takes the form

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

where $e_{i,t+h,m} = y_{i,t+h} - \hat{y}_{i,t+h|t,m}$ is the forecast error.

Forecasting performance is usually measured relative to some benchmark, m_0. For example, in tests of the efficient market hypothesis for financial markets, the benchmark might be a random walk forecast which assumes that asset prices fully incorporate all publicly available information at each point in time. Alternatively, the benchmark might be an incumbent model while the M alternative forecasts represent possible alternatives that could replace the benchmark.

The resulting *loss differential* of forecast m, measured relative to the benchmark m_0, is given by[6]

$$\Delta L_{i,t+h,m} = L(y_{i,t+h}, \hat{y}_{i,t+h|t,m_0}) - L(y_{i,t+h}, \hat{y}_{i,t+h|t,m}). \tag{2}$$

Under squared error loss, $\Delta L_{i,t+h,m} = e^2_{i,t+h,m_0} - e^2_{i,t+h,m}$. Positive values of $\Delta L_{i,t+h,m}$ show that forecast m generated a lower loss than the benchmark, m_0, in period $t+h$, while negative values show the reverse.

We next develop a set of hypotheses that exploit the assumed panel structure of the data. Previous work in the literature on multiple hypothesis testing has focused on the case with a single variable ($N = 1$), see, e.g., White (2000), Hansen (2005), and Romano and Wolf (2005). Conversely, papers that use panel data to evaluate predictive accuracy have not accounted for the multiple hypothesis testing issue, see, e.g., Keane and Runkle (1990) and Davies et al. (1995). As we shall see, *combining* panel data with insights from the literature on multiple hypothesis testing allows us to formulate and test a rich set of economic hypotheses.

2.2 Single Pairwise Comparison of Forecast Accuracy

We begin with a simple setting for comparing the predictive accuracy of two forecasts for a single variable ($N = 1$). Suppose we are interested in testing the hypothesis that two forecasts are equally accurate, on average, for a particular variable (unit), i. The key assumption is that the two forecasts used in the comparison (m_0 and m) as well as the identity of the

[6] For simplicity, we drop the reference to t and m_0 in the subscripts of ΔL.

variable (i) used in the forecast comparison are fixed ex-ante (predetermined). Under this assumption, the comparison does *not* involve a multiple hypothesis testing problem and so inference can be conducted using the test statistic proposed by Diebold and Mariano (1995):

$$H_0^{DM} : E[\Delta L_{i,t+h,m}] = 0. \tag{3}$$

Assuming that a time series of outcomes and forecasts $\{y_{i,t+h}, \hat{y}_{i,t+h|t,m}\}$ $t+h = 1, ..., T$ is observed, the Diebold-Mariano null in (3) can readily be tested using a t-test for the time-series average of the loss differential $\overline{\Delta L}_{i,m} = T^{-1} \sum_{t+h=1}^{T} \Delta L_{i,t+h,m}$.[7] Provided that parameter estimation error ("learning") can be ignored or is incorporated as part of the null hypothesis, the test statistic will asymptotically be normally distributed.[8]

Whenever we do not fix the variable, i, or the forecast, m, and instead consider tests conducted either across multiple variables or across many forecasts, a multiple hypothesis testing problem arises and we cannot rely on the conventional distribution results that underpin tests of H_0^{DM}. We next analyze a variety of alternative economic hypotheses and discuss which economic insights they can provide and the challenges they pose from a testing perspective.

2.3 Comparing Multiple Forecasts of a Single Variable

White (2000), Hansen (2005), and Romano and Wolf (2005) consider the null that, for a given variable, i, no forecast $m = 1, ..., M$ can beat the benchmark m_0:

$$H_0^{RC} : \max_{m \in \{1,...,M\}} E[\Delta L_{i,t+h,m}] \leq 0. \tag{4}$$

This "reality check" (RC) null is relevant if there is only a single outcome variable ($N = 1$) or, alternatively, if *ex-ante* we are interested in studying forecasting performance for a specific unit such as United States in a large cross-country analysis. The RC null in (4) is concerned with whether at least one forecast is better than some benchmark for a specific variable.[9] This could be relevant, for example, in tests of the efficient market hypothesis (EMH), with

[7] It is common to use heteroskedasticity and autocorrelation consistent standard errors when conducting this test, see Diebold and Mariano (1995).

[8] Giacomini and White (2006) discuss conditions under which this type of test is valid even for forecasts generated by nested models while Clark and West (2007) derive distributional results that account for parameter estimation error and nested models. Clark and McCracken (2001) consider the case with recursive updating in the parameters of nested forecasting models and show that this gives rise to a non-standard distribution of the resulting test statistics.

[9] White (2000) proposes a test based on the maximum value of the vector of time-series averages of loss differentials and develops a bootstrap methodology for characterizing the distribution of the maximum test statistic. Hansen (2005) proposes various refinements to White's approach, including using a studentized (pivotal) test statistic. Romano and Wolf (2005) develop a step-wise approach that can identify the set of forecasts that are superior to the benchmark.

strict versions of the EMH ruling out that *any* forecast can predict time-series variation in (risk-adjusted) returns. This approach is, however, limited to evaluating forecasting performance for a single variable and so tests of (4) cannot be used to draw conclusions about how a particular forecaster or forecasting method performs more broadly across multiple variables.

Suppose, however, that we are interested in testing whether a particular forecaster is more accurate than the benchmark, and that this forecaster reports predictions for many variables, $i = 1, ..., N$. Failing to find that this forecaster outperforms the benchmark for a particular variable, i, does not imply that the forecaster has no superior skills. For example, the forecaster could perform well for a subset of variables or even *on average*, across all variables. We next discuss hypotheses designed specifically to identify such predictive skills.

2.4 Performance Averaged Across Many Variables: Generalist Skills

Suppose we are interested in testing whether any forecaster (or model) has superior skills across multiple units. One interpretation of this case is that the forecaster is a skilled "generalist" who many not beat the benchmark for every single variable but is expected to perform well on average, i.e., in comparisons across a broad set of variables. A natural way to identify this type of skill is by testing whether any of the forecasters' predictive performance, averaged across multiple units, i, is better than the benchmark. We can test this by modifying the null in (4) to

$$H_0^G : \max_{m \in \{1,...,M\}} E[\frac{1}{N} \sum_{i=1}^{N} \Delta L_{i,t+h,m}] \leq 0. \tag{5}$$

Under this null, none of the M forecasters has a smaller expected loss than the benchmark. As stated, the generalist null in (5) can allow the individual forecasts, $m = 1, ..., M$, to outperform the benchmark for some variables, i, as long as the average forecasting performance is worse than that of the benchmark. Conversely, generalist skills that lead to a rejection of the null in (5) may arise from a forecaster's ability to predict the values of a set of common factors affecting multiple variables.

Tests of the null in (5) can be conducted in the same manner in which we would test the RC null in (4). The only difference is that we first compute the average cross-sectional loss differential in period $t + h$, $\overline{\Delta L}_{t+h,m} = \frac{1}{N} \sum_{i=1}^{N} \Delta L_{i,t+h,m}$, then use this measure as opposed to the loss differential only for variable i, $\Delta L_{i,t+h,m}$, as the basis for the test. As a result, the multiple hypothesis testing problem again only needs to account for the number of forecasts (M) entering the comparisons.

Since the null in H_0^G is concerned with individual forecasters' average performance, tests

of this null can help identify forecasters with *generalist* skills, i.e., the ability to outperform some benchmark on average. However, average forecasting performance could also be dominated by a single variable, i, whose performance is sufficiently superior that it overrides weaker evidence of inferior performance for other variables. When interpreting the results, it is, therefore, important to inspect both the cross-sectional average forecasting performance along with the performance recorded for individual units, $\overline{\Delta L}_{i,t+h,m}$.[10]

2.5 Performance for Specific Variables: Specialist Skills

Rather than testing superior predictive ability for a single pre-specified variable (as in (4)) or "on average" (as in (5)), we may be interested in comparing a single pair of forecasts (m_0 versus m) across multiple variables ($i = 1, ..., N$) and testing whether a particular forecast, m, is better than the benchmark for *any* of the variables:

$$H_0^S : \max_{i \in \{1,...,N\}} E[\Delta L_{i,t+h,m}] \leq 0. \tag{6}$$

Here the null is that a particular forecast, m, does not improve on the benchmark, m_0, for *any* of the variables $i = 1, ..., N$. This null focuses on a single forecast (m) and instead searches across the set of variables $i = 1, ..., N$ and so the dimension of the joint hypothesis test is now N. Failing to reject the "specialist" null in (6) would suggest that a given forecaster is not significantly better than the benchmark for any variable. Conversely, rejections indicate that the forecaster outperforms the benchmark for at least one variable.

If a subset of variables with common features can be identified ex-ante, an alternative way to test for specialist skills is to compare the average predictive accuracy for units within this subset, e.g., developing versus advanced economies. Specialist skills would now be defined as the ability to improve on the predictive accuracy of the benchmark for variables within this particular subset. To this end, consider a subset (cluster) C_k comprising $N_k < N$ of the variables. We can then test for predictive skills for this subset of variables for any of the M forecasters by means of the hypothesis

$$H_0^{S'} : \max_{m=1,...,M} E[\frac{1}{N_k} \sum_{i \in C_k} \Delta L_{i,t+h,m}] \leq 0, \tag{7}$$

where we assume that the clusters C_k (and N_k) are determined prior to the analysis. These clusters can be designed to identify domain expertise of particular economic interest, i.e., financial versus real variables or inflation versus GDP growth.

[10] An alternative approach towards testing for superior generalist skills is to base the test statistic on the proportion of variables, i, for which $\overline{\Delta L}_{i,t+h,m} < 0$. This approach is likely to work best if N is large but could have weak power because it disregards information about the magnitude of the loss differentials.

Alternatively, we can test for domain-specific predictive skills within a particular cluster by restricting the hypothesis in (6) to variables within a cluster, C_k :

$$H_0^{SC_k} : \max_{i \in C_k} E[\Delta L_{i,t+h,m}] \leq 0. \tag{8}$$

While this hypothesis is a special case of (6), in finite samples it is possible that tests conducted for a more narrow set of variables will have stronger power and, thus, be able to better identify cases of superior predictive skill.

2.6 Performance Across Multiple Variables and Multiple Forecasts

The broadest type of forecast comparison involves testing whether there exist any variables, i, for which any of the forecasts, m, beats the benchmark. Testing this broad "no superior skill" hypothesis requires that we model the distribution of the test statistic obtained by maximizing both over i and over m:

$$H_0^{NSS} : \max_{i \in \{1,..,N\}} \max_{m \in \{1,..,M\}} E[\Delta L_{i,t+h,m}] \leq 0. \tag{9}$$

This null reflects the very real possibility that apparently superior forecasting performance may simply be the result of looking at a very large set of pairwise comparisons. When the number of variables, N, and the number of forecasts, M, are large, tests of this null involve a high-dimensional modeling problem. This poses special challenges to the test procedure, as we discuss in the next section.

Because tests of the null H_0^{NSS} take account of all possible forecast comparisons, it could well lead to conservative inference and loss in power in situations where few forecasts are genuinely superior to the benchmark and these are mixed with forecasts that either do not improve on the benchmark or are inferior. In such cases, it can be beneficial to constrain the set of comparisons to subset of variables and/or forecasters based on ex-ante economic reasoning on which variables and forecasters are most likely to be associated with superior predictive skills. Alternatively, statistical procedures for moment selection can be used as we discuss further below.

We next develop statistics for testing the hypotheses introduced above.

3 Test statistics

To handle cases with a large number of variables (N) relative to the time-series dimension, T, we use the approach developed by Chernozhukov et al. (2018). In turn, this approach implements a version of the high-dimensional bootstrap from Chernozhukov et al. (2013,

2017) which accounts for serial dependence using a blocking technique.

Suppose we only compare the performance of a single forecast (m) to that of the benchmark (m_0) so that, without any risk of confusion, we can drop the forecast subscript, m, from (2) and define $\Delta L_{i,t+h} = L(y_{i,t+h}, \hat{y}_{i,t+h|t,m_0}) - L(y_{i,t+h}, \hat{y}_{i,t+h|t,m_1})$ and $\hat{\mu}_i = T^{-1}\sum_{t+h=1}^{T} \Delta L_{i,t+h}$. Appendix B.1 of Chernozhukov et al. (2018) considers the test statistic $J_T = \max_{1\leq i\leq N} \sqrt{T}\hat{\mu}_i$. We depart from the analysis in their paper by introducing a studentized test statistic. As suggested by Hansen (2005), studentization can improve the power in tests of predictive performance in many empirical applications where $\hat{\mu}_i$ displays strong forms of heteroskedasticity. Such heteroskedasticity may arise due to differences in sample lengths used to compute the test statistics or differences in the degree of variability in the loss differentials across different variables.

Let B_T be an integer that measures the average block length used in the bootstrap and define the number of blocks $K := K_T = \lfloor T/B_T \rfloor$. For $j \in \{1, ..., K-1\}$, let $H_j = \{(j-1)B_T + 1, ..., jB_T\}$ and $H_K = \{(K-1)B_T + 1, ..., T\}$ denote the jth and kth time-series blocks. Consider the following test statistic for the maximum value of the average loss differential, computed across the $i = 1, ..., N$ cross-sectional units:

$$R_T = \max_{1\leq i\leq N} \frac{T^{-1/2}\sum_{t+h=1}^{T} I_{i,t+h}\Delta L_{i,t+h}}{\hat{a}_i}, \tag{10}$$

where $I_{i,t+h} = \mathbf{1}\{\Delta L_{i,t+h} \text{ is observed}\}$ and $\hat{a}_i > 0$ is a normalizing quantity that is either deterministic or estimated from the data. Ideally, we observe all the loss differentials, $I_{i,t+h} = 1$ for all i and all $t+h$. In practice, it is common that not all of the $\Delta L_{i,t+h}$'s are available.

Example 3.1. We consider a variety of possible normalizations of the test statistic:

- No normalization: $\hat{a}_i = 1$ for $1 \leq i \leq N$. This choice does not attempt to balance differences in $Var(T^{-1/2}\sum_{t=1}^{T} I_{i,t+h}\Delta L_{i,t+h})$ across i. Hence, the behavior of R_T will depend mostly on indices i corresponding to the largest values of $Var(T^{-1/2}\sum_{t=1}^{T} I_{i,t+h}\Delta L_{i,t+h})$.

- Full normalization: $\hat{a}_i = \sqrt{K^{-1}\sum_{j=1}^{K}\left(B_T^{-1/2}\sum_{t+h\in H_j}(I_{i,t+h}\Delta L_{i,t+h} - \hat{\mu}_i)\right)^2}$. This normalization is an estimate of the long-run variance and hence can correct the cross-sectional differences in scale of $T^{-1/2}\sum_{t+h=1}^{T} I_{i,t+h}\Delta L_{i,t+h}$. However, this could be a rather noisy estimate as it is essentially computed from K observations, each observation being the sum of data in a block. In small samples, the noise in this estimate could create substantial size distortions.

- Partial normalization: $\hat{a}_i = \sqrt{T^{-1}\sum_{t+h=1}^{T}(I_{i,t+h}\Delta L_{i,t+h} - \hat{\mu}_i)^2}$ with $\hat{\mu}_i = T^{-1}\sum_{t+h=1}^{T} I_{i,t+h}\Delta L_{i,t+h}$. This choice of normalization corrects for different scales

in the unconditional variance of $Var(I_{i,t+h}\Delta L_{i,t+h})$. This is a sensible choice when the variability of $I_{i,t+h}\Delta L_{i,t+h}$ differs significantly across i but does not guarantee that the variance of $T^{-1/2}\sum_{t+h=1}^{T} I_{i,t+h}\Delta L_{i,t+h}/\hat{a}_i$ stays approximately constant across i.[11]

- Sample-sized normalization: $\hat{a}_i = \sqrt{T_i/T}$, where $T_i = \sum_{t+h=1}^{T} I_{i,t+h}$. This choice is sensible when T_i/T varies significantly across i and the variance of $T^{-1/2}\sum_{t+h=1}^{T} I_{i,t+h}\Delta L_{i,t+h}$ is driven by the number of observations in each series.

- Double normalization: $\hat{a}_i = \sqrt{T_i/T}\times\sqrt{T_i^{-1}\sum_{t+h=1}^{T} I_{i,t+h}(\Delta L_{i,t+h} - \hat{\mu}_{(i)})^2}$ with $\hat{\mu}_{(i)} = T_i^{-1}\sum_{t+h=1}^{T} I_{i,t+h}\Delta L_{i,t+h}$. This choice normalizes both by the number of observations T_i and the unconditional variance of the observed $\Delta L_{i,t+h}$.

Critical values for R_T in (10) can be based on the following multiplier bootstrap procedure. Let $\{\xi_j\}_{j=1}^{K}$ be a set of i.i.d $N(0,1)$ variables used to construct the statistic

$$R_T^* = \max_{1\leq i\leq N} R_{i,T}^*, \tag{11}$$

where

$$R_{i,T}^* = \frac{K^{-1/2}\sum_{j=1}^{K}\xi_j\left(B_T^{-1/2}\sum_{t\in H_j} I_{i,t+h}\Delta L_{i,t+h}\right)}{\hat{a}_i}.$$

To cover the different hypotheses and test statistics used in our empirical analysis, we consider a general setting in which the number of forecasts of $y_{i,t+h}$ is also large. Suppose that for each $1 \leq i \leq N$, we have $|\mathcal{D}_i| + 1$ forecasts for all $1 \leq t+h \leq T$, say $\hat{y}_{i,t+h|t,m}$ for $m = m_0$ and $m \in \mathcal{D}_i$. Hence, we can allow the number of forecasters to vary across variables, although for simplicity we assume that this number does not depend on time.[12]

The following general setup covers as special cases the earlier null hypotheses considered in Section 2:[13]

$$H_0 : \max_{1\leq i\leq N}\max_{m\in\mathcal{D}_i} E\left[\Delta L_{i,t+h,m}\right] \leq 0. \tag{12}$$

To test this null, define

$$U_{t+h} = \left(\{\Delta L_{1,t+h,m}\}_{m\in\mathcal{D}_1}, \{\Delta L_{2,t+h,m}\}_{m\in\mathcal{D}_2}, ..., \{\Delta L_{N,t+h,m}\}_{m\in\mathcal{D}_N}\right),$$

and suppose that U_{t+h} is a column vector of dimension $\mathcal{N} = \sum_{i=1}^{N}|\mathcal{D}_i|$ with kth component

[11] The reason is that the unconditional variance is not the same as the long-run variance of the partial sum $T^{-1/2}\sum_{t+h=1}^{T} I_{i,t+h}\Delta L_{i,t+h}$ since the latter also depends on any serial correlation.

[12] Extension to the case where $\mathcal{D}_i$ is time-varying is conceptually trivial but makes the notation more cumbersome without offering additional insights.

[13] The null hypothesis in (6) can be stated as a special case of (12) with $\mathcal{D}_i = \{m\}$. Similarly, we can accommodate the null hypothesis in (8) by replacing $\max_{1\leq i\leq N}$ with $\max_{i\in C_k}$ in (12). Next, writing $U_{t+h} = \{\Delta L_{i,t+h,m}\}_{i\in C_k,\ m\in\mathcal{D}_i}$, the rest of the procedure follows directly from this.

denoted by $U_{k,t+h}$. Consider the test statistic

$$\tilde{R}_T = \max_{1\leq k\leq \mathcal{N}} \frac{T^{-1/2}\sum_{t+h=1}^{T} U_{k,t+h}}{\hat{a}_k}, \tag{13}$$

where $\hat{a}_k$ is computed using any of the schemes described in Example 3.1.

Bootstrap critical value are constructed analogously

$$\tilde{R}^*_T = \max_{1\leq k\leq \mathcal{N}} \tilde{R}^*_{k,T}, \tag{14}$$

where

$$\tilde{R}^*_{k,T} = \frac{K^{-1/2}\sum_{j=1}^{K}\xi_j\left(B_T^{-1/2}\sum_{t+h\in H_j} U_{k,t+h}\right)}{\hat{a}_k}.$$

To establish the distributional properties of the test statistic in (13), we require a set of regularity conditions. To this end, let $W_{k,t+h} = U_{k,t+h} - E(U_{k,t+h})$, while $W_{t+h} = (W_{1t+h}, ..., W_{\mathcal{N}t+h})$. We summarize our assumptions as follows:

Assumption 1. *Suppose that the following conditions hold:*
(1) The distribution of W_{t+h} does not depend on t.
(2) $P(\max_{1\leq t+h\leq T}\|W_{t+h}\|_\infty \leq D_T) = 1$ for some $D_T \geq 1$.
(3) $\{W_{t+h}\}_{t+h=1}^{T}$ is β-mixing with mixing coefficient $\beta_{\text{mixing}}(\cdot)$.
(4) $c_1 \leq E\left(k^{-1/2}\sum_{t+h=s+1}^{s+k} W_{j,t+h}\right)^2$, $E\left(k^{-1/2}\sum_{t+h=s+1}^{s+k} W_{j,t+h}\right)^2 \leq C_1$ for any j, s and k.
(5) $T^{1/2+b}D_T\log^{5/2}(\mathcal{N}T) \lesssim B_T \lesssim T^{1-b}/(\log\mathcal{N})^2$ and $\beta_{\text{mixing}}(s) \lesssim \exp(-b_1 s^{b_2})$ for some constant $b, b_1, b_2 > 0$.
(6) There exist a nonrandom vector $a = (a_1, ..., a_{\mathcal{N}})' \in \mathbb{R}^{\mathcal{N}}$ and constants $\kappa_1, \kappa_2 > 0$ such that $\kappa_1 \leq a_j \leq \kappa_2$ for all $1 \leq j \leq \mathcal{N}$ and $\max_{1\leq j\leq\mathcal{N}}|\hat{a}_j - a_j| = o_P(1/\log\mathcal{N})$.

Part (1) of Assumption 1 requires strict stationarity and can be relaxed at the expense of more technicalities in the proof. Part (2) imposes a bound on the tail behavior of the loss difference. When the loss difference is bounded, we can choose D_T to be a constant; when the loss difference is sub-Gaussian, we can choose $D_T \asymp \sqrt{\log(\mathcal{N}T)}$ and adapt the proof to handle $P(\max_{1\leq t\leq T}\|W_{t+h}\|_\infty \leq D_T) \to 1$. This bound on the variables is needed for the high-dimensional bootstrap and Gaussian approximation even in the i.i.d case; see Chernozhukov et al. (2013, 2017, 2018).[14] The β-mixing condition in part (3) is routinely imposed in the literature and holds for many stochastic processes. Part (4) requires the loss differences for all variables to be of roughly the same order of magnitude. Part (5) imposes

[14] One way to relax part (2) of Assumption 1 is to use the union bound together with moderate deviation inequalities for self-normalized sums, but this might lead to more conservative procedures; see Chernozhukov et al. (2018).

rate conditions; notice that we allow $\mathcal{N} \gg T$. Finally, part (6) states that $\hat{a}_j$ needs to be uniformly consistent for some non-random quantity that is bounded away from zero and infinity.

Note that we can verify that part (6) of Assumption 1 holds for the normalization schemes listed above as we next formalize:

Lemma 3.1. *Let Assumption 1(1)-(5) hold. Then all the normalizations in Example 3.1 satisfy part (6) of Assumption 1.*

Using Assumption 1, we have the following result:

Theorem 3.1. *Suppose Assumption 1 holds. Under H_0 in (12), we have*

$$\limsup_{T\to\infty} P\left(\tilde{R}_T > \tilde{Q}^*_{T,1-\alpha}\right) \leq \alpha,$$

where $\tilde{Q}^*_{T,1-\alpha}$ is the $(1-\alpha)$ quantile of $\tilde{R}^*_T$ conditional on the data. Moreover, if $E(\Delta L_{i,t+h,m}) = 0$ for all $1 \leq i \leq N$ and $m \in \mathcal{D}_i$, then

$$\limsup_{T\to\infty} P\left(\tilde{R}_T > \tilde{Q}^*_{T,1-\alpha}\right) = \alpha.$$

Theorem 3.1 establishes the asymptotic validity of the proposed procedure. Under the null of equal expected loss for all variables, the multiplier bootstrap test is asymptotically exact and, hence, not conservative.

The studentization used for $\tilde{R}_T$ serves a similar role as the self-normalization in Chernozhukov et al. (2018) for the independent case and typically improves on power properties. By arguments similar to those in Chernozhukov et al. (2018), we expect the test to have non-trivial power against alternatives of order $\max_{1\leq i\leq N} \max_{m\in\mathcal{D}_i} E(\Delta L_{i,t+h,m}) = O(\sqrt{T^{-1}\log\mathcal{N}})$ with a rate that is minimax optimal. Since the number of hypotheses tested only enters through a logarithmic factor, the proposed test has consistency against fixed alternatives even if this number grows exponentially with T.

It is important to note that the dimension $\mathcal{N}$ only has a very small impact on the requirements that guarantee the validity of the procedure. This is because in the regularity conditions (Assumption 1), only the rate $\log(\mathcal{N})$ matters, which means that $\mathcal{N}$ can increase at the rate T^c for any constant $c > 0$.

3.1 Family-wise error rate

We next use Theorem 3.1 to construct confidence sets for under- and overperforming units. For notational simplicity, we consider $|\mathcal{D}_i| = 1$ so $\mathcal{N} = N$. Define $A = \{i : \mu_i > 0\}$, where

$\mu_i = T^{-1}\sum_{t+h=1}^{T} E\Delta L_{i,t+h}$, so A is the set of units, i, for which an alternative forecast, m, beats the benchmark, m_0.

To estimate this set, consider

$$\hat{A} = \left\{ i : \ \frac{T^{-1/2}\sum_{t+h=1}^{T} \Delta L_{i,t+h}}{\hat{a}_i} > Q^*_{T,1-\alpha} \right\}.$$

If $\hat{A}$ contains a unit that is not in A, i.e., $\hat{A}\backslash A \neq \emptyset$, $\hat{A}$ makes a false discovery since it includes units for which the alternative forecast performs no better than m_0.

A consequence of Theorem 3.1 is that the probability of a false discovery is asymptotically at most α. To see this, notice that

$$\begin{aligned}
&P\left(\hat{A}\backslash A \neq \emptyset\right) \\
&= P\left(\frac{T^{-1/2}\sum_{t=1}^{T} \Delta L_{i_0,t+h}}{\hat{a}_{i_0}} > Q^*_{T,1-\alpha} \quad \text{for some } i_0 \in \hat{A}\backslash A \right) \\
&\leq P\left(\max_{i \in A^c} \frac{T^{-1/2}\sum_{t=1}^{T} \Delta L_{i,t+h}}{\hat{a}_i} > Q^*_{T,1-\alpha} \right) \\
&\leq \alpha + o(1),
\end{aligned}$$

where the last inequality follows by Theorem 3.1 applied to A^c (instead of $\{1, ..., N\}$). By construction, $\max_{i \in A^c} E\mu_i \leq 0$. We summarize this result as follows:

Corollary 3.1. *Suppose Assumption 1 holds. Consider A and $\hat{A}$ defined above. Then*

$$\limsup_{T\to\infty} P\left(\hat{A} \subseteq A\right) \geq 1 - \alpha.$$

Hence, with probability at least $1-\alpha+o(1)$, $\hat{A}$ only selects cases in which the alternative forecast outperforms the benchmark.

Our approach to bootstrapping the distribution of the maximum value chosen from a large set of test statistics is related to the reality check methodology pioneered by White (2000), but there are also important differences. Most notably, White (2000) tests hypotheses about the population parameter value.[15] Moreover, he assumes that the forecasts are generated by parametric models and thus take the form $f_{t+h|t} = f(Z_t, \hat{\beta}_h)$ using the parameter updating scheme discussed in West (1996).[16] Finally, White (2000) assumes that the number of forecasts each time period is fixed, whereas we allow it to be expanding with the sample size, T. As pointed out by White (2000) (page 1111) and Chernozhukov et al. (2018)

[15] See, e.g., the discussion on page 1099 in White (2000).
[16] See Assumption A.2 in the Appendix to West (1996).

(Comment 4.7), assuming a fixed number of forecasts, models or moment conditions is an important limitation in many empirical applications. Here we allow the number of forecasts to be much larger than T which can be quite important for panel forecasts with large N.

3.2 Improving power by moment selection

The literature on testing moment inequalities suggests that test power can be improved by reducing the number of inequalities, p, via moment selection; see e.g., Hansen (2005); Andrews and Soares (2010); Romano et al. (2014). To see how this works, we start with the goal of testing moment inequalities in $A = \{1, ..., N\}$.[17] We would like to use the data to find a set A_0 such that with high probability, say $1-\beta$, the moment inequalities contained in $A \backslash A_0$ are satisfied. Provided that this holds, we only need to test the moment inequalities in A_0. When $|A_0| \ll |A|$, excluding the moment inequalities in $A \backslash A_0$ can be expected to improve the power of the test, although we need to adjust the size of the test to be $\alpha - \beta$ when testing the moment inequalities in A_0.

Most of the literature on testing moment inequalities focuses on the case where $|A|$ is fixed.[18] Here, we follow the spirit of Romano et al. (2014) and use a bootstrapped threshold. We summarize the details in Algorithm 1.

Algorithm 1. *Implement the following steps:*

1. *Choose $\beta \in (0, \alpha)$ to be either a constant or a sequence tending to zero.*

2. *Compute*
$$R_{i,T} = \frac{T^{-1/2} \sum_{t=1}^{T} \Delta L_{i,t+h}}{\hat{a}_i} \qquad \forall 1 \leq i \leq N.$$

3. *Compute the bootstrapped threshold C_β, which is the $1-\beta$ quantile of $\|R_T^*\|_\infty$ conditional on the data, where R_T^* is defined in (11). In other words, $P(\|R_T^*\|_\infty > C_\beta \mid data) = \beta$.*

4. *Select $A_0 = \{i : \ R_{i,T} > -C_\beta\}$.*

5. *Compute the test statistic $\max_{i \in A_0} R_{i,T}$.*

6. *Compute the bootstrap critical value $C_{\alpha-\beta,A_0}$ satisfying $P(\max_{i \in A_0} R_{i,T}^* > C_{\alpha-\beta,A_0} \mid data) = \alpha - \beta$, where $R_{i,T}^*$ is defined in (11).*

[17] This can be generalized to $A = \{1, ..., \mathcal{N}\}$, where $\mathcal{N}$ varies depending on which null hypothesis is being tested. For example, $\mathcal{N} = N$ in H_0^S, whereas $\mathcal{N} = N \times M$ in H_0^{NSS}. Again, for simplicity, we focus on the case of $|\mathcal{D}_i| = 1$ (so $\mathcal{N} = N$).

[18] Hansen (2005) proposes a threshold of $\sqrt{\log \log N}$ based on the law of iterated logarithm so that A_0 contains moments whose sample counterpart is larger than $-\sqrt{T^{-1} \log \log N}$.

Although we need to decrease the size of the test from α to $\alpha - \beta$ for small β, the test statistic and the bootstrap critical value are computed as the maximum over indices in A_0 rather than over the original set $\{1, ..., N\}$. When $|A_0|$ is much smaller than N, the price we pay for using a reduced nominal size is small and the procedure can result in improved power.[19]

3.3 Monte Carlo Simulations

Appendix A reports the results from a set of Monte Carlo simulations which we use to study the finite sample properties of our new test statistics. We draw the following conclusions from these simulations. Both the studentized and non-studentized test statistics have reasonable size properties when N and M are small, regardless of the time-series dimension, T. However, as N and M grow bigger, the test statistics tend to become under-sized. This holds particularly for the studentized test statistic when $\alpha = 0.05$. Interestingly, the undersizing is less of a concern for $\alpha = 0.10$ and using a critical level of $\alpha = 0.10$ for the studentized test statistic in many cases gets us close to a size of 5%-10%.

The Monte Carlo simulations also show that the power of the studentized test statistic is far better than that of the non-studentized test statistic, even when size-adjusted critical values are used in the power calculations. This is an important consideration because accounting for the multiple hypothesis testing problem easily leads to procedures with weak power and, hence, conservative inference. For this reason, we use studentized test statistics with a size of $\alpha = 0.10$ throughout our empirical applications.

4 Forecasting Performance in a Single Period

In situations with a large number of variables, N, we might be able to exploit the cross-sectional dimension of the data to address whether the performance of individual forecasters, averaged cross-sectionally, is better than the benchmark in a *single* period or over a short time span. For example, we might be interested in testing if some forecasters were able to generate more accurate predictions of inflation or GDP than the benchmark during the Global Financial crisis but not be interested in predictive accuracy outside of this period.

Alternatively, we might be interested in testing whether agents possess forecasting skills in different states of the economy. Kacperczyk et al. (2016) develop a theoretical framework in which the informational advantage of skilled fund managers (and, hence, the relative accuracy of their forecasts) increases during periods of heightened risk since the payoff from more extensive information acquisition is higher in such states. Moreover, in empirical work Kacperczyk et al. (2014) find that the type of predictive ability possessed by skilled fund

[19] The high-dimensional testing problem is further discussed by Chernozhukov et al. (2018).

managers varies over the economic cycle, shifting from stock picking skills in expansions towards market timing skills during recessions.

Tests for superior predictive skills in a single cross-section can be based on the distribution of the average cross-sectional loss differentials, $\hat{\mu}_{m,t+h} = N^{-1}\sum_{i=1}^{N} \Delta L_{i,t+h,m}$. For inference to be valid, we require the use of a cross-sectional central limit theorem for the resulting test statistic which means that the cross-sectional dependency in the loss differentials cannot be too strong. To establish conditions under which this holds, consider the following factor structure for the forecasting errors

$$e_{i,t+h,m} = \lambda'_{i,m} f_{t+h} + u_{i,t+h,m}, \tag{15}$$

for $1 \leq i \leq N$ and $1 \leq t+h \leq T$, where $f_{t+h} \in \mathbb{R}^k$ is a set of latent factors common to the forecast errors. Many outcome variables contain a common component that none of the forecasters anticipated which can make forecast errors highly correlated. The factor structure assumed in (15) is a natural representation of this situation.

Under the factor structure in (15), the squared error loss differential takes the form

$$\begin{aligned}\Delta L_{i,t+h,m} &= \left(\lambda'_{i,m_0} f_{t+h} + u_{i,t+h,m_0}\right)^2 - \left(\lambda'_{i,m} f_{t+h} + u_{i,t+h,m}\right)^2 \\ &= f'_{t+h}(\lambda_{i,m_0}\lambda'_{i,m_0} - \lambda_{i,m}\lambda'_{i,m}) f_{t+h} + u^2_{i,t+h,m_0} - u^2_{i,t+h,m} \\ &\quad + 2 f'_{t+h}(\lambda_{i,m_0} u_{i,t+h,m_0} - \lambda_{i,m} u_{i,t+h,m}).\end{aligned} \tag{16}$$

To rule out that the cross-sectional dependencies are so strong as to prevent us from establishing distributional results for comparisons of the cross-sectional average loss differentials, we assume that the idiosyncratic terms are independent conditional on the factor structure, as we next make clear:

Assumption 2. *Let $\mathcal{F}$ be the σ-algebra generated by $\{f_{t+h}\}_{1\leq t+h\leq T}$ and $\{\lambda_{i,m}\}_{1\leq i\leq N,\ 0\leq m\leq M}$. Conditional on $\mathcal{F}$, $\{u_i\}_{i=1}^N$ is independent across i and $E(u_i \mid \mathcal{F}) = 0$, where $u_i = \{u_{i,t+h,m}\}_{1\leq t+h\leq T, 1\leq m\leq M} \in \mathbb{R}^{T\times M}$.*

Using Assumption 2, we have

$$\frac{1}{N}\sum_{i=1}^{N} \Delta L_{i,t+h,m} - E\left(\frac{1}{N}\sum_{i=1}^{N} \Delta L_{i,t+h,m} \mid \mathcal{F}\right) = \frac{1}{N}\sum_{i=1}^{N} \xi_{i,t+h,m},$$

where $\xi_{i,t+h,m} = 2f'_{t+h}(\lambda_{i,m_0} u_{i,t+h,m_0} - \lambda_{i,m} u_{i,t+h,m}) + \left(u^2_{i,t+h,m_0} - u^2_{i,t+h,m}\right) - E\left(u^2_{i,t+h,m_0} - u^2_{i,t+h,m} \mid \mathcal{F}\right)$. Under Assumption 2, $\{\xi_{i,t+h,m}\}_{i=1}^N$ has mean zero and is independent across i conditional on $\mathcal{F}$. Therefore, we can use a central limit theorem to show that $\frac{1}{N}\sum_{i=1}^{N} \Delta L_{i,t+h,m}$ is an asymptotically normal estimator for $E\left(\frac{1}{N}\sum_{i=1}^{N} \Delta L_{i,t+h,m} \mid \mathcal{F}\right)$. By

virtue of a high-dimensional Gaussian approximation, we can extend this intuition to a simultaneous test across many periods, $t+h$, and/or forecasts, m.

The conditional null that, given $\mathcal{F}$, a particular forecast, m, is not expected to be more accurate, on average across all units, than the benchmark in a particular time period, $t+h$, can be tested by cross-sectionally averaging the loss differentials in period $t+h$:

$$H_0^{ES} : \max_{(t+h,m)\in\mathcal{A}} E\left(\overline{\Delta L}_{t+h,m} \mid \mathcal{F}\right) \leq 0, \tag{17}$$

where $\overline{\Delta L}_{t+h,m} = \frac{1}{N}\sum_{i=1}^{N} \Delta L_{i,t+h,m}$ is the cross-sectional average loss differential for forecast m and $\mathcal{A}$ is the set defined by $\mathcal{A} = \{t+h\} \times \{m = 1, ..., M\}$. Note that the hypothesis in (17) is strictly about performance in period $t+h$ and so we refer to this null hypothesis as being about "event skills" (ES). Put slightly differently, the null in (17) is concerned with whether the average predictive accuracy in period $t+h$ of any of the forecasters is better than that of the benchmark.

We can also test whether, across all periods $t+h = 1, ..., T$ and all forecasts, $m = 1, ..., M$, *any* of the forecasts were more accurate, on average across all units, than the benchmark in *any* time period (given $\mathcal{F}$):

$$H_0^{ES'} : \max_{t+h\in\{1,...,T\}} \max_{m\in\{1,...,M\}} E\left(\overline{\Delta L}_{t+h,m} \mid \mathcal{F}\right) \leq 0, \tag{18}$$

where now $\mathcal{A} = \{t+h = 1, ..., T\} \times \{m = 1, ..., M\}$ in (18). This null can be used to test whether any forecaster's cross-sectional average performance beats the benchmark during any period in the sample.

The test statistic we propose for testing (17) or (18) is given by

$$Z = \max_{(t+h,m)\in\mathcal{A}} \frac{\sqrt{N}\overline{\Delta L}_{t+h,m}}{\sqrt{N^{-1}\sum_{i=1}^{N} \widetilde{\Delta L}_{i,t+h,m}^2}}, \tag{19}$$

where $\widetilde{\Delta L}_{i,t+h,m} = \Delta L_{i,t+h,m} - \overline{\Delta L}_{t+h,m}$ is the demeaned loss differential of variable i for forecast m. Critical values for this test statistic can be obtained from a bootstrap

$$Z_* = \max_{(t+h,m)\in\mathcal{A}} \frac{N^{-1/2}\sum_{i=1}^{N} \varepsilon_i \widetilde{\Delta L}_{i,t+h,m}}{\sqrt{N^{-1}\sum_{i=1}^{N} \widetilde{\Delta L}_{i,t+h,m}^2}}, \tag{20}$$

where the multipliers $\varepsilon_i \sim N(0,1)$ are generated independently of the data. Note that we assume cross-sectional conditional independence for the idiosyncratic terms. Moreover, we assume that the multipliers ε_i are i.i.d, rather than having the block structure needed to handle serial dependence in the earlier test statistics which use data from multiple time

periods.

We can now establish the validity of the above procedure:

Theorem 4.1. *Let Assumption 2 hold. Suppose that* $(\kappa_{N,3}^3 \vee \kappa_{N,4}^2 \vee B_N)^2 \log^{7/2}(TMN) \lesssim N^{1/2-c}$ *for some* $c \in (0, 1/2)$, *where* $B_N = (E \max_{t,m,i} |\xi_{i,t+h,m}|^4)^{1/4}$, $\kappa_{N,3} = (\max_{i,t,m} E|\xi_{i,t+h,m}|^3)^{1/3}$ *and* $\kappa_{N,4} = (\max_{i,t,m} E|\xi_{i,t+h,m}|^4)^{1/4}$. *Then under* H_0 *in (17) we have*

$$\limsup_{N\to\infty} P\left(Z > Q^*_{N,1-\alpha,Z}\right) \leq \alpha,$$

where $Q^*_{N,1-\alpha,Z}$ *is the* $(1-\alpha)$ *quantile of* Z_* *conditional on the data. Moreover, if* $E\left(\frac{1}{N}\sum_{i=1}^N \Delta L_{i,t+h,m} \mid \mathcal{F}\right) = 0$ *for all* $(t+h, m) \in \mathcal{A}$, *then*

$$\limsup_{N\to\infty} P\left(Z > Q^*_{N,1-\alpha,Z}\right) = \alpha.$$

Here, B_N, $\kappa_{N,3}$ and $\kappa_{N,4}$ measure the tail of $\xi_{i,t+h,m}$, which is the deviation of the loss differential $\Delta L_{i,t+h,m}$ from its conditional mean. When deviations are bounded, $B_N, \kappa_{N,3}$ and $\kappa_{N,4}$ are positive constants. If $\xi_{i,t+h,m}$ has sub-Gaussian tails, then $B_N = O(\log(TMN))$ and $\kappa_{N,3}$ and $M_{N,4}$ are constants. The proof of Theorem 4.1 follows almost exactly the same lines as the proof of Theorem 4.3 of Chernozhukov et al. (2018) with two exceptions: (1) the independence assumption is replaced by conditional independence given $\mathcal{F}$ and (2) the assumption of identical distributions is changed and can be handled by slight changes to the definition of B_N, $\kappa_{N,3}$ and $\kappa_{N,4}$. We omit the details of the proof for this reason.

While Theorem 4.1 is stated for the null in (17), the null hypothesis in (18) can equally be tested in the same way by replacing $\max_{(t+h,m)\in\mathcal{A}}$ with $\max_{t+h\in\{1,\ldots,T\}} \max_{m\in\{1,\ldots,M\}}$ in (19) and (20).

5 Bloomberg Survey of Economic Forecasters

This section applies our methods to the Bloomberg survey of economic forecasters which reports forecasts for a range of economic variables. The Bloomberg data offer a setting with many forecasters, multiple outcome variables and a long sample, i.e., a setting where N, M and T are reasonably large. This means, first, that we can use our methods to explore the full range of economic hypothesis developed in Sections 2 and 4 and, second, that the multiple hypothesis testing issue becomes important and so requires the use of our new methods.

5.1 Bloomberg Survey

Bloomberg conducts monthly surveys of several economic variables. We focus on the forecasts of outcomes (or preliminary estimates) of a set of 14 U.S. variables with reasonably large sample coverage: The Fed funds rate (FDTR), GDP growth (GDP), growth in personal consumption (GDPC), growth in industrial production (IP), change in nonfarm payrolls (NFP), new home sales (NHS), building permits (NHSPA), housing starts (NHSPS), percentage changes in the core price index (PCEC), percentage changes in the price index (PCE), the unemployment rate (UN), average hourly earnings (AHE), consumer price inflation (CPI), and existing home sales (ETSL).

Table 1 lists the 14 variables along with a few summary statistics. Data samples vary across the individual variables, beginning as early as August 1997 (nonfarm payrolls) and as late as March 2010 (average hourly earnings), with all series ending at some point after May 2019. The number of monthly time-series observations varies from 111 to 254. Forecasts are reported both for individual forecasters and for individual firms, the difference being that some forecasters belong to the same firm so the number of individual forecasters is slightly greater than the number of firms. Data on the individual firms generally offer greater time-series coverage, so we use firms as the unit of observation in most of our analysis. Many of the survey participants report very few forecasts, so we require a minimum of five observations for each participant to conduct a meaningful comparison of predictive accuracy. After imposing this requirement, the number of firms reporting valid forecasts over our sample (M) varies from 38 for average hourly earnings to 153 for nonfarm payrolls.

Bloomberg refers to a *release date* which is the date when the official data source publishes the actual value of a variable, and an earlier *observation date* which is the end of the period covered by the survey. For example the observation date could be 4/30/2019 with the release date being 5/10/2019. To avoid stale forecasts, we only include forecasts recorded within 7-10 days of the release date.

For 12 of the 14 variables, the data frequency is monthly and so each release date is easily paired with a single observation date. The remaining two variables, GDP growth and growth in personal consumption, are quarterly. For these variables the BEA releases three different estimates of the "actual" value for a given quarter, namely a preliminary estimate followed by a second and a third estimate, respectively. These estimates get released in separate months and so we treat them as three monthly observations of the same variable.[20]

[20] For example, for Q3, 2018 the GDP series has an observation date of 9/29/2018, along with three release dates: 10/26/2018, 11/28/2018, and 12/21/2018.

5.2 Pairwise Forecast Comparisons

For each of the 14 variables covered by the survey, we first present the outcome of pairwise comparisons of firm-level forecasts relative to two benchmarks, namely (i) forecasts from an AR(1) model with an intercept whose parameters are estimated recursively, using the first 24 months of the sample as a warm-up period; and (ii) the equal-weighted (average) forecast computed across all forecasters included in the survey in a given month. By design, the simple AR(1) model is restricted to incorporate very little information–essentially the historical persistence of the predicted variable. Even so, simple, parsimonious models have often proven difficult to beat in empirical analyses of out-of-sample forecasting performance, see, e.g., Faust and Wright (2013).

Our second benchmark, the equal-weighted average, provides a natural reference point since it allows us to address if any forecaster is significantly better than the simple strategy of just using the peer-group average forecast. This would seem to be a minimal requirement in order to make it worthwhile for a decision maker to elect to rely on a single forecaster as opposed to using the consensus average.

Figure 1 shows histograms depicting the distribution of the ratio of the root mean squared forecast errors (RMSE) for individual forecasters relative to forecasts from the AR(1) model (left column) and equal-weighted mean (right column) benchmarks. We focus on five variables, namely ETSL (existing home sales), GDP, IP (industrial production) NFP (nonfarm payrolls), and UN (unemployment rate).[21] Ratios below unity indicate that the firms are producing more accurate forecasts than the benchmark, while ratios above unity suggest that the benchmark forecasts are best. Relative to the AR(1) forecasts, the vast majority of firms generate forecasts with lower RMSE values for all five variables. The opposite holds when we compare firm forecasts to the equal-weighted mean. Notice also the very large spreads in the individual forecasters' performance relative to the benchmarks.

Figure 1 provides insights into the heterogeneity in individual forecasters' performance relative to our two benchmarks, but does not give a sense of the statistical significance of (relative) differences in forecasting performance. As a first way to conduct more formal test results, Table 2 reports pairwise Diebold-Mariano (DM) t-tests computed as

$$t_{i,m}^{DM} = T^{-1/2} \frac{\sum_{t=1}^{T} \Delta L_{i,t+h,m}}{\hat{\sigma}(\Delta L_{i,t+h,m})}. \tag{21}$$

We categorize the test statistics according to whether $t_{i,m}^{DM}$ falls in one of four intervals whose bounds are defined using a 95% critical value for a one-sided test: $t_{i,m}^{DM} < -1.645$, $-1.645 \leq t_{i,m}^{DM} < 0$, $0 \leq t_{i,m}^{DM} < 1.645$, and $t_{i,m}^{DM} \geq 1.645$. The DM tests are set up

[21] To preserve a legible scale, for some of the variables a few outliers in the right tail have been omitted from these plots.

using $\Delta L_{it+h,m} = e^2_{i,t+h,m_0} - e^2_{i,t+h,m}$ with $m_0 = \{AR(1), average\}$ as benchmarks and m representing the individual firm-level forecasts. Positive values of the loss differential therefore indicate that individual forecasts are more accurate than either the AR(1) forecasts or consensus forecasts, while negative values suggest the reverse. We list the number of test statistics that fall in each bin, with the total number of pairwise comparisons listed in the bottom row.

First consider the comparison of the individual forecasters' precision against that of the forecasts generated by an AR(1) model (Panel A). For most variables, there is strong evidence that a majority of forecasters are more accurate than the forecasts from the AR(1) model–in many cases significantly so, as evidenced by the many positive and statistically significant DM t-statistics. In fact, for the AHE, CPI, GDP, GDPC, IP, NFP, NHSPS, and PCEC variables, more than half of the forecasters produce significantly more accurate forecasts than the AR(1) model. Conversely, there is only weak evidence that individual forecasters produce significantly less accurate forecasts than the AR(1) model: for all but one of the variables (UN) we observe a DM test statistic below -1.645 for three or fewer forecasters and for the unemployment rate only four out of 149 forecasters significantly underperform the AR(1) model.

Next, consider the comparison of the individual forecasts against the mean forecast (Panel B of Table 2). For all variables, we find few instances – always less than a handful – in which individual forecasters are significantly more accurate than the consensus mean (t-statistic above 1.645). Conversely, there are multiple instances in which the reverse holds and individual forecasters are significantly *less* accurate than the mean, with numbers varying from a minimum of nine cases (FDTR and NHSPA, corresponding to 8-10% of all cases) to a maximum of 63 cases (UN), representing just over 40% of the forecasters.

5.3 Comparisons Across Many Forecasters

The results in Table 2 are difficult to interpret because they do not account for the fact that we are considering so many pairwise comparisons — in some cases well over one hundred. Some of the individual forecasters that beat the average might have done so due to luck. We therefore next present results that account for the multiple hypothesis testing problem.

We first conduct tests of the null that no individual forecaster is able to outperform a given benchmark, i.e., the Reality Check null in (4). Results from such tests are presented in Table 3. In each panel, the top row lists the p-value of the null hypothesis, followed by the number of rejections. As a reminder, in this and all subsequent tables, we use a test size of $\alpha = 0.10$. Our Monte Carlo simulations suggest that this choice yields a nominal finite-sample size closer to 5% for the studentized test statistic which tends to be undersized

in finite samples.[22]

First, consider the comparison of the individual forecasters' precision against that of the forecasts generated by an AR(1) model (Panel A). For all variables but one (NHSPA), we find strong evidence (with p-values at or below 0.02) against the null hypothesis that none of the forecasters can beat the forecasts from the AR(1) model and, thus, conclusively reject the Reality Check null in (4). However, after accounting for the multiple hypothesis testing problem the number of forecasters deemed to be significantly better than the AR(1) model drops to a much smaller number than suggested by the pair-wise test statistics in Table 2.

Next, consider the results when we reverse the null hypothesis in (4) and assign the individual forecasters to m_0, thus testing the null that none of the individual forecasters performs significantly *worse* than the AR(1) model. Panel B of Table 3 shows that, for all 14 variables, we fail to find a single rejection of the null that all the individual forecasters are at least as accurate as the AR(1) forecasts. This is quite a strong finding that stands in marked contrast to our pair-wise comparisons for variables such as the unemployment rate (UN) for which we found that four of the forecasters generated significantly negative Diebold-Mariano test statistics when compared against the AR(1) benchmark. Apparently, this evidence does not stand up to closer scrutiny which makes sense accounting for the fact that these four cases were selected from a set of 149 pairwise comparisons.

Comparing the individual forecasters' performance to the simple mean forecast, computed as the cross-sectional average across forecasters, creates a markedly higher hurdle than the AR(1) forecasts. Panel C tests the null that none of the individual forecasters can outperform the simple average. Across the 14 variables, we find only two instances (one, each, for GDPC and NFP) in which this null is rejected.[23] Testing the reverse null hypothesis – that none of the professional forecasters are *less* accurate than the mean forecast – leads to many more rejections, however, particularly for the GDP, IP and UN variables (Panel D). While there is very weak evidence that individual forecasters can beat the average forecast, there is, thus, plenty of evidence that individual forecasters can be *worse* than the average forecast for some variables.

Figure 2 illustrates the relation between the pairwise DM p-values (listed along the horizontal axis) and the p-values computed from bootstrapping the distribution of test statistics associated with testing the Reality Check null in (4) (listed on the vertical axis). These p-values are computed using forecasts of the unemployment rate variable. The DM p-values are conducted on a pair-wise basis and thus ignore multiple hypothesis testing while the

[22] By making this choice, we implicitly err on the side of finding more rejections than if a more traditional level of 5% were used, so as to give forecasters the benefit of the doubt and better be able to detect if any superior skills are present.

[23] The two firms for which we reject the null are Sim Kee Boon Institute for Financial Economics (8 observations) and the Canada Pension Plan Investment Board (5 observations).

Reality Check p-values account for this. The DM p-values are therefore smaller and so the points should appear above the 45-degree line in the figure. Points in the bottom left corner indicate cases in which the DM and Sup tests both reject, whereas points in the top left corner show instances in which the DM test rejects but the Sup test does not. The distance to the 45-degree line can therefore be viewed as a measure of the importance of the multiple hypothesis testing problem.

The top right panel in Figure 2 compares the DM p-values to the p-values for the Reality Check null that the AR(1) forecasts are more accurate than all of the individual forecasts. We fail to find any p-values below 0.5 on the vertical axis (so no rejections of the RC null), although there are many instances with low DM p-values (small values on the horizontal axis). Clearly, the effect of multiple hypothesis testing is important for this case. The largest discrepancy between the two sets of p-values appears in the comparison of the Reality check test of the null that none of the individual forecasters can beat the simple equal-weighted average and the pairwise DM test (bottom left panel). The smallest Reality Check p-values exceed 0.8 while there are four DM p-values below 0.10.

An alternative to studying forecasting performance separately for each outcome variable is to compare individual forecasters' performance averaged across all variables, i.e., to test the null of no generalist skills in (5).[24] Results from such tests are reported in the right-most column in Table 3.[25] We reject the null that the AR(1) forecasts are as accurate, on average, as those produced by all the individual forecasters, finding 49 forecasters for which the null gets rejected (Panel A). Conversely, we fail to reject the reverse null, i.e., we find no significant evidence that the average predictive accuracy of the individual forecasters is significantly worse than that of the AR(1) model (Panel B).

Panel C reports results from comparisons of individual forecasters' average performance to the consensus average. We fail to identify a single forecaster with significantly more accurate average performance than the consensus forecast. Conversely, we identify 36 forecasters whose average performance is significantly worse than the equal-weighted average (Panel D).

Rather than basing tests on individual variables or on the grand average across all variables, we can form groups of "similar" variables and test for domain-specific skills. To this end, we first form five clusters labeled inflation (consisting of AHE, CPI, PCEC, and PCE), housing market (ETSL, NHS, NHSPA, and NHSPS), economic growth (GDP, GDPC, IP), labor market (NFP, UN) and the funds rate (FDTR). Using (7), we can test whether the average forecasting performance within a particular cluster is always at least as good

[24] In these comparisons, we omit the two quarterly GDP series (GDP and GDPC) and the Fed Funds rate series which only has eight annual data points.

[25] We require the individual forecasters to report results for at least five variables to be included in the comparison and compute the test statistic from the loss differential averaged across these variables.

for the benchmark forecast as it is for all forecasts contained in the alternative set.

Table 4 reports the outcome of our tests. We find a large number of rejections of the null that the AR(1) forecasts are at least as accurate as all of the individual firm forecasts, with rejections concentrated in the inflation, growth and labor categories and far fewer rejections appearing for the housing market and fed funds rate. Conversely, we do not find a single rejection for any of the categories of the null that the individual forecasts are at least as accurate as the AR(1) forecasts.

The opposite picture emerges from the comparison of the firm forecasts to the equal-weighted average. Here we only find a single rejection (for the labor market category) of the null that the mean forecast is at least as accurate as all of the individual forecasts. In contrast, we find a large number of rejections of the null that the individual firm forecasts are at least as accurate as the equal-weighted average with the highest number of rejections emerging for the inflation, growth, and labor categories and only one and zero rejections coming from the housing market and federal funds rate categories, respectively.

We also implemented the moment selection procedure for our Sup tests of equal predictive accuracy. In general, for the Bloomberg data we do not find that this procedure leads to more rejections of the Reality Check null in (5). Finally, we explored the sensitivity of our results with respect to requirements on the minimum number of observations for each forecaster. We find that our results are robust to requiring each forecaster to have a certain number of minimum observations (e.g., 18 months).

Overall, this evidence suggests that there is little-to-no evidence of superior domain-specific skills among individual forecasters once we compare the forecasts to the equal-weighted average. Conversely, many individual forecasters are significantly less accurate than the peer average for variables in the inflation, growth and labor categories.

5.4 Tests for Any Superior Forecasting Skills

We next consider whether *any* of the individual forecasters can beat the benchmark for *any* of the variables, i.e., the null of "no superior skill" in (9).

Table 5 reports the outcome of our tests. First consider the results that require each forecaster to have reported a minimum of five forecasts (Panels A-D). Using the AR(1) forecasts as the baseline and the firm-level forecasts as the alternative (Panel A), across 1,001 pairwise comparisons, we obtain a p-value of 0.00 and identify 49 individual forecasters who are significantly more accurate than the AR(1) model for at least one variable. In contrast, the reverse null – that all forecasters are at least as accurate for all variables as the AR(1) forecasts – fails to be rejected, with a p-value of 0.65. Undertaking the same test for the individual forecasters expands the set of pairwise comparisons from 1,001 to 1,207. Using these forecasts, we now find 47 rejections of the null that the AR(1) forecasts are at least as

accurate as those produced by the individual forecasters. Hence, we continue to arrive at a similar conclusion regardless of whether we use the firm-level or individual forecaster-level data.[26]

Next, consider the comparison of the predictive accuracy of the individual forecasters to the consensus mean. Results are very different for this comparison. With a p-value of 0.03, Panel C shows that we identify only a single instance in which an individual forecaster beats the equal-weighted average[27]. In contrast, there are six cases for which the equal weighted average is significantly better than individual forecasters.[28] Similar results hold regardless of whether we use firm-level or forecaster-level data (Panel D).

Recent research indicates that the size of bootstrapping methods can be distorted when a large fraction of the units (in our case the individual forecasters) have small sample sizes.[29] To address the importance of this point, Panels E-H in Table 5 show a similar set of results when we require the individual forecasters to report at least 25 forecasts. This restriction limits both the number of forecasters (M) and the number of short track records used in the comparisons. Panels E and F show that the results comparing individual forecasters to predictions from the AR(1) model do not change: we continue to find 49 cases in which individual firms produce significantly more accurate forecasts than the AR(1) benchmark. Turning to the comparison between the individual forecasters and the equal-weighted mean, the results become a little stronger as we no longer find even a single rejection of the null that the equal-weighted forecasts are at least as accurate, across all forecasters and all variables, as the forecasts reported by individual firms. Moreover, the number of rejections of the reverse null–that the individual firm-level forecasts are at least as accurate as the equal-weighted mean–increases from 6 cases to 11, indicating that the power of the bootstrap test has increased a bit as a result of omitting forecasters with short records.

5.5 Identifying Periods with Superior Forecasting Performance

We next provide empirical tests of the null of no superior predictive skills at a given point in time using cross-sectional average performance, i.e., the null in (17). To make the results easier to interpret, we compute our test statistics using non-overlapping 12-month blocks, averaging the squared forecast errors within individual calendar years. The top panels in

[26] In sharp contrast, if instead we use the non-studentized test statistic, we obtain p-values ranging between 0.19 and 0.99 and fail to reject the null in a single case. This clearly demonstrates the weaker power of the non-studentized test statistic.

[27] The rejection is recorded for the NFP forecasts produced by the Canada Pension Plan Investment Board. It is worth noting that this forecaster only reports forecasts for five variables during 2005.

[28] These rejections are for Manulife Asset Management Limited (CPI), UniCredit Bank (CPI), Nord/LB (PCE), Canadian Imperial Bank of Commerce (UN), Desjardins Financial Group (UN), and Westpac Banking Corp (UN).

[29] See, e.g., Andrikogiannopoulou and Papakonstantinou (2019).

Figure 3 plot results from these tests with blue asterisks in the upper row indicating periods where at least one forecaster is significantly better than the AR(1) forecasts (left panel) or the equal-weighted average (right panel) during a particular year. We find that at least one individual forecaster produced significantly more accurate forecasts than the AR(1) benchmark in 16 of the 17 years. Conversely, we fail to find a single year in which at least one forecaster is *less* accurate than the AR(1) benchmark.

Turning to the comparison of the individual forecasters to the equal-weighted average, the top right panel in Figure 3 shows that there are three years (2015, 2018, and 2019) during which at least one forecaster generates significantly more accurate predictions than the equal-weighted average, while the reverse holds every single year in our sample. In other words, conducting the cross-sectional tests separately on the individual years, we find at least one forecaster with significantly worse performance than the equal-weighted mean in every single year during our sample.

The test statistics reported in the top panels of Figure 3 are conducted on an annual basis and so do not address the multiple hypothesis testing problem induced by looking at 17 separate test statistics. Indeed, we might well expect some of the rejections to be down to "luck". To address this issue and obtain an overall test that answers whether there was *any* year during our sample in which at least one forecaster performed significantly better (or worse) than the benchmark, we use our bootstrap methodology to test the null in (18).

The lower panels in Figure 3 show result from these tests. We would expect the number of rejections to decline compared to those found for the tests conducted for the individual years (upper panels) and this is exactly what we find, although the difference is not very large. Specifically, we find that the null that the AR(1) forecasts are at least as accurate every year in the sample as the predictions reported by the individual forecasters gets rejected for all but three years. We also fail to find any evidence that individual forecasters produced forecasts that were less accurate than those from the AR(1) model at any point during our sample.

The number of years in which at least one firm generates forecasts that are more accurate than the equal-weighted mean is reduced from three (upper panel) to two after accounting for the multiple hypothesis testing problem.[30] Although we no longer find that at least one firm is less accurate than the equal-weighted average for every single year in our sample, this holds for all but three years.

[30] The two firms for which we reject the null in the top line in panel D are Commerzbank AG (2018) and JP Morgan (2019).

6 Term Structure of Forecast Errors

Our second empirical application uses our new test methods to assess the evolution in the performance of the International Monetary Fund's (IMFs) World Economic Outlook (WEO) forecasts of real GDP growth and inflation across the world's economies.[31] The WEO publication contains the "flagship" forecasts published by the IMF which perhaps receives more attention than any other global forecasts and is widely covered by public media and in academic research. We consider the accuracy of next-year and current-year forecasts as recorded in the Spring and Fall WEO issues. This involves conducting pairwise comparisons ($M = 1$) of forecasts recorded at long and short horizons. However, the cross-sectional dimension (country-level) is quite large for this data as the WEO forecasts are reported for around 180 countries. This allows us to analyze forecasting performance for particular clusters or subsets of countries and better identify for which types of economies forecasts become more accurate as time progresses.

6.1 Predictive Accuracy Across Different Horizons

The IMF's World Economic Outlook (WEO) gets published twice each year, namely in April (labeled Spring, or S) and October (Fall, or F). The WEO publication contains forecasts for the current-year ($h = 0$) and next year ($h = 1$), producing a set of four forecast horizons, listed in decreasing order: $\{h = 1, S; h = 1, F; h = 0, S; h = 0, F\}$.[32] 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, providing us with a somewhat shorter data sample. In all cases, the last outcome for our data is recorded for 2016. Our analysis focuses on forecasts of real GDP growth and inflation.

We would expect the predictive accuracy to improve as the forecast horizon is reduced and more information about the outcome becomes available. Because we observe the WEO forecasts of the same outcome (real GDP growth or inflation in country i in year t) at four different horizons, we can test if this holds. Ordering the WEO forecasts from the longest ($h = 1, S$) to the shortest ($h = 0, F$) horizon, under the squared error loss in (1) we have

$$E[e^2_{h=0,F}] \leq E[e^2_{h=0,S}] \leq E[e^2_{h=1,F}] \leq E[e^2_{h=1,S}]. \tag{22}$$

Define the squared error loss differential for forecasts of the outcome in period t given

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

[32] 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.

forecasts generated at short and long horizons, $t - h_S$ and $t - h_L$ for $h_L > h_S$:

$$\Delta L_{i,t,h_L \to h_S} = (y_{i,t} - \hat{y}_{i,t|t-h_S})^2 - (y_{i,t} - \hat{y}_{i,t|t-h_L})^2. \tag{23}$$

Similarly, define the change in the squared forecast error from reversing the order and going from the short to the long horizon:

$$\Delta L_{i,t,h_S \to h_L} = (y_{i,t} - \hat{y}_{i,t|t-h_L})^2 - (y_{i,t} - \hat{y}_{i,t|t-h_S})^2. \tag{24}$$

Using the timing of the WEO data, this suggests four comparisons of changes in predictive accuracy across forecast horizons, namely (i) Spring versus fall next year ($h_L = 1, S; h_S = 1, F$); (ii) Fall next year versus spring current-year ($h_L = 1, F; h_S = 0, S$); (iii) Spring versus fall current-year ($h_L = 0, S; h_S = 0, F$); and (iv) Spring next year versus fall current-year ($h_L = 1, S; h_S = 0, F$).

We apply our approach to test for improvements in squared error loss accuracy across all countries for a given pair of forecast horizons. For example, to test the null that, for each country, i, the forecast is at least as accurate at the short horizon, h_S, as it is at the long horizon, $h_L > h_S$, we test the null

$$H_0 : \max_{i \in \{1,\ldots,N\}} (E[\Delta L_{i,t,h_L \to h_S}]) \leq 0. \tag{25}$$

Conversely, to test the null that, for each country, i, the forecast is at least as accurate at the long horizon as it is at the short horizon (and, thus, fails to improve with new information), we test the null

$$H_0 : \max_{i \in \{1,\ldots,N\}} (E[\Delta L_{i,t,h_S \to h_L}]) \leq 0. \tag{26}$$

Rejections of the null in (26) means that the forecasts are improving as the forecast horizon gets shorter. To get a sense of how the accuracy in the WEO forecasts evolves across different horizons, Figure 4 shows a heat diagram depicting how the accuracy of the WEO country-level inflation forecasts evolves as we move from $h = 1, S$ to $h = 1, F$ (top left panel), from $h = 1, F$ to $h = 0, S$ (top right panel), from $h = 0, S$ to $h = 0, F$ (bottom left panel) and on a cumulative basis ($h = 1, S$ versus $h = 0, F$). The last comparison measures whether the WEO current-year Fall forecasts ($h = 0, F$) are more accurate than the prior-year Spring forecasts ($h = 1, S$) and thus accumulates any gains in accuracy over the three preceding six-month intervals.

In each diagram, the color applied to each country is based on the p-values for testing the null in (26). Red colors correspond to small p-values, indicating that short horizon

forecasts are significantly more accurate than long-horizon forecasts. Green colors indicate the reverse, i.e., weak evidence against significant improvements in predictive accuracy as the forecast horizon gets shorter.

There is essentially no evidence that the forecasts at the longest horizon ($h = 1, S$) are significantly less accurate than the forecasts at the shorter next-year horizon ($h = 1, F$), indicating that little useful information arrives one year ahead of time–or at least that such information is not incorporated in the next-year fall forecast. Evidence of improvements in predictive accuracy starts showing up in the top right ($h = 1, F$ vs $h = 0, S$) and bottom left ($h = 0, S$ vs $h = 0, F$) panels, with notable improvements for many European countries, United States, and Australia. Finally, on a cumulative basis, we see clear evidence of improved accuracy of the inflation forecasts for the aforementioned countries in addition to countries such as Canada, Chile, and India.

Table 6 supplements Figure 4 by reporting the results of comparisons of the accuracy of the WEO forecasts across the four different forecast horizons. Panels A and C set up the test statistic so that rejections (small p-values) indicate significant improvements as the forecast horizon gets longer, thus testing the null in (25). We fail to find any instances for which the GDP growth forecasts (Panel A) or inflation forecasts (Panel C) are significantly *less* accurate at the shorter forecast horizons than at the longer horizons.

In contrast, testing the reverse null in (26) we find three instances–including large economies such as Brazil and Italy–for which the Sup test identifies significant improvements in the accuracy of the next-year GDP growth forecasts as we move from the spring to the fall WEO (Panel B). Evidence of significant improvements in the accuracy of the GDP growth forecasts becomes stronger for the current-year forecasts ($h = 0, S$ versus $h = 0, F$) for which the null is rejected for nine countries. On a cumulative basis (last column), we identify 18 countries for which we can reject the null that the long-horizon forecasts ($h = 1, S$) are at least as accurate as the short-horizon forecasts ($h = 0, F$).

For the inflation forecasts, we observe even more rejections of the null that the long-horizon forecast is at least as accurate as the short-horizon forecast (Panel D), with notable cases including France, Germany, Italy, Turkey, and United States. Improvements in predictive accuracy are concentrated in the revisions from next-year Fall to current-year Spring and from current-year Spring versus Fall forecasts as well as on a cumulative basis across horizons ($h = 1, S$ vs $h = 0, F$).

Interestingly, there is no evidence to suggest that the accuracy of the WEO inflation forecasts improves significantly for any country between the Spring and Fall of the previous year. Little useful information appears to get released one year out as improvements in the accuracy of the IMF's inflation forecasts mainly are concentrated between the fall of the previous year and the fall of the current year.

We can apply the hypothesis in (7) to test whether improvements in the accuracy of the WEO forecasts across the four horizons differ across geographical regions and types of economies. To this end, we next group the countries into nine categories adopted by the IMF, namely (i) advanced economies, (ii) emerging and developing economies, (iii) emerging and developing Europe, (iv) low income developing countries, (v) Latin America and Caribbean, (vi) Commonwealth of Independent States, (vii) Middle East, North Africa, Afghanistan and Pakistan, (viii) Emerging and developing Asia, and (ix) Sub-Sahara Africa.

The results, presented in Table 7, suggest large variation across types of economies and across the two variables (GDP growth versus inflation). Two observations stand out. First, we clearly see more countries with significant improvements for the inflation series than for the GDP growth series. For example, on a cumulative basis, we record significant improvements in the accuracy of the inflation forecasts for 28 out of 36 advanced economies. The corresponding number is 15 out of 36 countries for GDP growth. Second, there is clearly stronger evidence that the WEO forecasts are getting significantly more accurate as the forecast horizon shrinks for the advanced economies than for any of the other groups. Evidence of improvements in predictive accuracy as the forecast horizon is reduced is weakest for emerging market and developing economies.

To identify which advanced economies the forecasts improve significantly for, Table 8 applies the Sup test associated with (8) to the group of advanced economies. Once again, we find no case for which the short-horizon GDP growth is deemed to be less accurate than the corresponding forecasts generated at longer horizons (Panel A). Conversely, we now identify significant reductions in squared error losses after accounting for the multiple comparison problem not only for the current-year forecasts ($h = 0, S$ versus $h = 0, F$) but also for the one-year-ahead forecasts (Panel B). For example, for next-year GDP growth forecasts, we see significant improvements in predictive accuracy between the Spring and Fall WEO issues for Italy, Japan and Portugal. In addition, we see significant improvements in predictive accuracy for six countries between $h = 1, F$ and $h = 0, S$ and for 14 countries from $h = 0, S$ to $h = 0, F$ forecasts of GDP growth.

For the inflation forecasts, again we find no cases of significant deterioration in predictive accuracy as the forecast horizon shrinks (Panel C). Conversely, we find 12 countries for which the predictive accuracy of current-year Spring forecasts is significantly better than that of the next-year fall forecasts and 16 countries for which the current-year fall inflation forecasts are significantly more accurate than the current-year spring forecasts (Panel D).

6.2 Improvements in Predictive Accuracy in Individual Years

Figure 5 shows results from cross-sectional comparisons of average forecasting performance in individual years for the four different forecast horizons and thus tests the null hypothesis

in (17) (top row) or (18) (bottom row). Asterisks in this figure indicate when the null that forecasts at the longest horizon are at least as accurate as forecasts at the shorter horizon gets rejected at the 10% significance level. Years are shown on the horizontal axis, with outcomes for the four forecast horizons tracked in separate rows on the y-axis.

The upper panels test the null (17) that the longer-horizon forecast is at least as accurate as the shorter-horizon forecast in particular years. Comparing the two longest horizons ($h = 1, S$ versus $h = 1, F$), we fail to reject this null for almost half of the years in the sample. Conversely, we only fail to reject this null in two or three years for the $h = 1, F$ versus $h = 0, S$ comparison. For the two shortest horizons ($h = 0, S$ versus $h = 0, F$) and the cumulative revision ($h = 1, S$ versus $h = 0, F$), however, we reject the null in every single year. These results suggest that the forecasts improve significantly in almost all years at the current-year horizons, as well as on a cumulative basis. In contrast, there is weaker evidence of improvements in predictive accuracy every single year at the longer one-year-ahead horizons.

Turning to the tests of (18), we find only weak evidence against the null that the one-year-ahead spring forecasts ($h = 1, S$) are at least as accurate as the one-year-ahead fall forecasts ($h = 1, F$) every single year during the sample, regardless of whether we consider the GDP or inflation series. In fact, we identify only two years during our sample (2009 and 2010) in which the null gets rejected for GDP growth and a single year (2008) where the null gets rejected for the inflation rate (top row). Rejections of the null in (18) get stronger at the shorter horizons and, notably for the cumulative revisions, suggesting that short-horizon forecasts are significantly more accurate than long-horizon forecasts in a majority of the years during our sample after accounting for the multiple hypothesis testing problem.

These results are in line with what we would expect and illustrate that our tests possess the power to identify significant differences in predictive accuracy. They also illustrate that the forecast horizon is crucial to the predictive accuracy of macroeconomic forecasts and that improvements in accuracy are concentrated at relatively short horizons.

7 Conclusion

We develop new methods for evaluating the accuracy of panel forecasts and testing if individual forecasts are significantly more accurate than some benchmark forecast for at least one outcome variable, one forecaster (model), or one time-period after accounting for the multiple hypothesis testing problem associated with comparing forecasting performance across multiple tests. Building on Chernozhukov et al. (2018), we show that a bootstrap approach can be used to test economically interesting hypotheses that take the form of multiple moment conditions. Our approach allows us to test for different types of forecasting skills–specialist,

generalist, or event-specific–and can be used to identify both the forecasters, variables, and time periods for which forecasters may possess superior skills.

In an empirical application to survey forecasts of a range of economic variables, we find that pair-wise comparisons of individual forecasters' performance against a simple peer-group average suggest that some forecasters have superior skills. However, once we account for the multiple hypothesis testing problem, ("luck"), there is little evidence that individual forecasters possess any types of superior predictive skills.

The paucity of evidence in support of the existence of economic forecasters with superior skills does *not*, in any way, suggest that economic forecasters are "unskilled". In fact, we find plenty of evidence to suggest the reverse: for many variables, economic forecasters are able to beat simple, robust statistical forecasts. Moreover, there is clear evidence that organizations such as the IMF update their predictions as new information becomes available and that the precision of their forecasts improves as the forecast horizon shrinks.

Our empirical findings have important implications for how economic forecasts should be used. Often, decision makers have the option of either relying on the predictions of a single forecaster or using some combination of individual forecasts. Our weak evidence in support of the existence of individual forecasters with superior skills suggests that decision makers should be disinclined to focus on the predictions of individual forecasters–even those with a good historical track record. Whenever feasible, a less risky strategy with a higher chance of success is to rely on simple peer-group averages.

Our findings are suggestive of an environment where professional forecasters have access to similar data sets and use broadly similar models or approaches to generate forecasts, consistent with asymmetry of information not being very important for forecasts of broad economic outcomes.

References

Afrouzi, H. (2019). Strategic inattention, inflation dynamics and the non-neutrality of money.

Andrews, D. W. and Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, 78(1):119–157.

Andrikogiannopoulou, A. and Papakonstantinou, F. (2019). Reassessing false discoveries in mutual fund performance: Skill, luck, or lack of power? *The Journal of Finance*, forthcoming.

Chen, X., Shao, Q.-M., Wu, W. B., and Xu, L. (2016). Self-normalized cramér-type moderate deviations under dependence. *The Annals of Statistics*, 44(4):1593–1617.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of gaussian random vectors. *Probability Theory and Related Fields*, 162(1-2):47–70.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309–2352.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2018). Inference on causal and structural parameters using many moment inequalities. *Review of Economic Studies (Forthcoming), available at arXiv:1312.7614*.

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.

Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1):291–311.

Davies, A., Lahiri, K., et al. (1995). A new framework for analyzing survey forecasts using three-dimensional panel data. *Journal of Econometrics*, 68(1):205–228.

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

Faust, J. and Wright, J. H. (2013). Forecasting inflation. In *Handbook of economic forecasting*, volume 2, pages 2–56. Elsevier.

Gallagher, E., Schmidt, L., Timmermann, A., and Wermers, R. (2019). Investor information acquisition and money market fund risk rebalancing during the 2011-12 eurozone crisis. *Robert H. Smith School Research Paper No. RHS*, 2886171.

Genre, V., Kenny, G., Meyler, A., and Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, 29(1):108–121.

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

Gorton, G. and Ordonez, G. (2014). Collateral crises. *American Economic Review*, 104(2):343–78.

Granger, C. W. J. (1999). Outline of forecast theory using generalized cost functions. *Spanish Economic Review*, 1(2):161–173.

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

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

Kacperczyk, M., Nieuwerburgh, S. V., and Veldkamp, L. (2014). Time-varying fund manager skill. *The Journal of Finance*, 69(4):1455–1484.

Kacperczyk, M., Van Nieuwerburgh, S., and Veldkamp, L. (2016). A rational theory of mutual funds' attention allocation. *Econometrica*, 84(2):571–626.

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.

Mackowiak, B. and Wiederholt, M. (2009). Optimal sticky prices under rational inattention. *American Economic Review*, 99(3):769–803.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2014). A practical two-step method for testing moment inequalities. *Econometrica*, 82(5):1979–2002.

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

Tetlock, P. E. and Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.

Timmermann, A. (2006). Forecast combinations. In Elliott, G., Granger, C., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, pages 135–196. Elsevier.

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

Van Nieuwerburgh, S. and Veldkamp, L. (2010). Information acquisition and under-diversification. *The Review of Economic Studies*, 77(2):779–805.

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

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

Zhu, Y. and Bradic, J. (2018). Significance testing in non-sparse high-dimensional linear models.

Table 1: **Summary Statistics for Bloomberg economic survey variables**

Variable	Description	Type[1]	Frequency	Begins	Ends	Number of time periods	Number of forecasters	Number of firms	Number of firms>5 forecasts
AHE	Average hourly earnings	Δ% YoY	monthly	3/5/2010	5/3/2019	111	104	86	38
CPI	CPI	Δ% YoY	monthly	11/19/2002	5/10/2019	197	178	134	67
ETSL	Existing homes sales	Level	monthly	3/23/2005	5/21/2019	171	215	162	92
FDTR	Fed Funds rate	Rate Level	8 times/year[2]	12/22/1998	12/11/2019	169	544	395	88
GDP	GDP	Δ% QoQ	monthly[4]	4/30/1997	5/30/2019	254	309	221	134
GDPC	GDP Personal Consumption	Δ% QoQ	monthly[4]	1/30/2003	4/26/2019	193	167	130	50
IP	Industrial Production	Δ% MoM	monthly	6/16/1998	5/15/2019	252	288	204	121
NFP	Nonfarm payrolls	Δ MoM	monthly	8/1/1997	6/7/2019	254	324	234	153
NHS	New home sales	Level	monthly	6/2/1998	5/23/2019	251	273	196	103
NHSPA	Building permits	Level	monthly	8/16/2002	5/16/2019	202	205	150	69
NHSPS	Housing starts	Level	monthly	6/16/1998	5/16/2019	252	278	198	99
PCEC	PCE Core	Δ% YoY	monthly[3]	6/25/2001	5/31/2019	180	164	121	45
PCE	PCE	Δ% YoY	monthly[3]	5/28/2004	5/31/2019	181	154	118	65
UN	Unemployment	Rate Level	monthly	1/10/1997	6/7/2019	253	308	224	149

Notes: The table summarizes our data on 14 variables collected from the Bloomberg survey of economists, including variable name, description, frequency, sample period, number of time-series observations, and the number of different forecasters for each variable. We combine the forecasts of economists serving in the same company/institution by computing a simple average for each variable each period. To be included in the analysis, the reported forecasts must have a forecast horizon of 7-10 days, i.e., we only include forecasts reported between 7 and 10 days prior to the release date. Values published on the release dates are used as the 'actual value' in the forecast evaluation. We convert the level forecasts (ETSL, NHS, NHSPA, NGSPS) to percentage changes prior to evaluating the forecasting performance for these variables.

[1] Δ% stands for percentage change; QoQ, MoM, and YoY refer to quarter-on-quarter, month-over-month and year-over-year, respectively.

[2] Release Date for FDTR is 8 times per year.

[3] PCEC and PCE get released monthly.

[4] Although GDP and GDPC are quarterly variables, the BEA releases three estimates for every quarterly value, one each month, labeled preliminary, second and third estimates respectively. Most economists made forecasts for all three estimates before each release date so we concatenate the three separate releases into one monthly time series.

Table 2: Distribution of pairwise Diebold-Mariano test statistics. Firm-level forecasters vs. AR(1) or the equal-weighted mean

Panel A: Individual firm-level forecasters vs AR(1)

	AHE	CPI	ETSL	FDTR	GDP	GDPC	IP	NFP	NHS	NHSPA	NHSPS	PCEC	PCE	UN
tstat<-1.645	0	0	0	0	2	0	0	1	3	3	1	0	1	4
-1.645<tstat<0	2	0	6	3	0	0	9	13	16	38	3	0	5	15
0<tstat<1.645	9	17	48	61	14	8	32	42	57	21	21	10	34	78
tstat>1.645	27	50	38	24	118	42	80	97	27	7	74	35	25	52

Panel B: Individual firm-level forecasters vs equal-weighted mean

	AHE	CPI	ETSL	FDTR	GDP	GDPC	IP	NFP	NHS	NHSPA	NHSPS	PCEC	PCE	UN
tstat<-1.645	14	28	30	9	56	15	45	58	17	9	20	19	27	63
-1.645<tstat<0	17	28	39	41	51	23	57	72	57	39	57	17	30	59
0<tstat<1.645	6	11	19	34	24	11	18	19	27	20	20	7	6	26
tstat>1.645	1	0	4	4	3	1	1	4	2	1	2	2	2	1
total	38	67	92	88	134	50	121	153	103	69	99	45	65	149

Notes: This table reports the distribution of Diebold-Mariano test statistics for comparing the accuracy of individual (firm-level) forecasters to predictions from an AR(1) model (Panel A) or simple equal-weighted mean forecasts (Panel B). We estimate the parameters of the AR(1) model using an initial warm-up period of 24 months. Positive t-statistics suggest that the individual forecasters were more accurate than the benchmarks, while negative values suggest the opposite. We discard individual firms with less than five forecasts.

Table 3: Sup tests for superior predictive accuracy: Individual firm-level forecasters versus AR(1) model or the equal-weighted mean

	AHE	CPI	ETSL	FDTR	GDP	GDPC	IP	NFP	NHS	NHSPA	NHSPS	PCEC	PCE	UN	Average
Panel A: m_0 = AR(1), m_1 = firm forecasters															
p-value	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.02	0.00	0.00
no. rejections	18	18	4	2	51	12	15	27	2	0	29	15	3	9	49
Panel B: m_0 = firm forecasters, m_1 = AR(1)															
p-value	1.00	1.00	1.00	0.99	0.94	1.00	1.00	0.79	0.29	0.37	0.90	1.00	0.61	0.53	1.00
no. rejections	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Panel C: m_0 = mean, m_1 = firm forecasters															
p-value	0.94	0.98	0.20	0.53	1.00	0.07	0.97	0.01	0.84	0.93	0.82	0.24	0.72	0.84	1.00
no. rejections	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
Panel D: m_0 = firm forecasters, m_1 = mean															
p-value	0.03	0.00	0.02	0.53	0.00	0.05	0.01	0.03	0.16	0.10	0.15	0.01	0.00	0.00	0.00
no. rejections	2	5	4	0	7	1	6	5	0	1	0	3	4	9	36
no. forecasters	38	67	92	88	134	50	121	153	103	69	99	45	65	149	121

Notes: The first 14 columns of this table report the outcome of a set of Sup tests for superior predictive accuracy. The null hypothesis in these tests is that the forecasts in the m_0 set are at least as accurate as the forecasts in the m_1 set. Rejections of this null indicate that forecasts in m_0 perform worse than at least one forecast from m_1. The tests are conducted separately for each variable and so tests the hypothesis in equation (4) in the paper. The last column in the table reports test statistics for the null applied to the cross-sectional average forecasting performance, i.e., the predictive accuracy averaged across all the monthly variables in our sample (AHE, CPI, ETSL, IP, NFP, NHS, NHSPA, NHSPS, PCEC, PCE and UN). This is equation (5) in the paper. The mean forecast is calculated as the cross-sectional average across firm-level forecasts. We estimate the parameters of the AR(1) model using a warm-up period of 24 months. Panel A and B compare AR(1) forecasts to individual (firm-level) forecasts. Panel C and D compare mean forecasts to the individual forecasts. We discard individual firms with less than five reported forecasts. Moreover, the comparison of average performance in the last column excludes firms with fewer than five variables, each of which has a minimum of five reported forecasts. In addition, we normalize the forecast error of each variable by the sample RMSE of the mean forecasts to put the forecast errors of the different variables on a comparable scale. The last row lists the number of firms in the sample. The row labeled "p-value" in each panel is the p-values of the Sup test. The second row in each panel is the number of rejections at the 10% significance level using a studentized bootstrapped test statistic.

Table 4: Sup tests of superior average predictive accuracy across subsets of similar variables: Individual firm-level forecasters versus an AR(1) model and an equal-weighted mean

	Inflation	Housing market	Growth	Labor	Funds rate
Panel A: m_0 = AR(1), m_1 = firm forecasters					
p-value	0.00	0.00	0.00	0.00	0.00
no. rejections	33	17	66	36	7
Panel B: m_0 = firm forecasters, m_1 = AR(1)					
p-value	0.95	1.00	0.99	0.97	0.99
no. rejections	0	0	0	0	0
Panel C: m_0 = mean, m_1 = firm forecasters					
p-value	0.98	1.00	0.72	0.04	0.52
no. rejections	0	0	0	1	0
Panel D: m_0 = firm forecasters, m_1 = mean					
p-value	0.00	0.02	0.00	0.00	0.53
no. rejections	12	1	16	17	0
no. forecasters	87	123	147	155	88

Notes: This table reports test statistics for Sup tests of the null hypothesis that none of the forecasts in the m_0 set is less accurate (has a strictly higher average MSE value computed than the alternative forecasts in the m_1 set. These tests are restricted to variables belonging to a set of clusters of variables and so test the hypothesis in equation (7) in the paper. We use the following clusters: Inflation (AHE, CPI, PCEC, PCE); housing market (ETSL, NHS, NHSPA, NHSPS); economic growth (GDP, GDPC, IP); labor market (NFP, UN); and Fed funds rate (FDTR). We estimate the parameters of the AR(1) model using a warm-up period of 24 months. Panels A and B compare AR(1) forecasts to individual (firm-level) forecasters. Panels C and D compare mean forecasts to the individual forecasters. We discard individual firms with less than five reported forecasts. The comparison of average performance within each cluster excludes firms with fewer than two variables (except for the Fed funds cluster which contains a single variable), each of which has a minimum of five reported forecasts. In addition, we normalize the forecast error of each variable by the sample RMSE of the mean forecasts to put the forecast errors of the different variables on a comparable scale. The last row lists the number of firms in the sample. The row labeled "p-value" in each panel is the p-values of the Sup test. The second row in each panel is the number of firms rejected at a 10% significance level using a studentized test statistic.

Table 5: Sup tests comparing predictive ability across multiple outcome variables and multiple forecasters

	Benchmark vs. firm forecasters		Benchmark vs. individual forecasters	
5 observations minimum				
Panel A			**Panel B**	
	m_0 =AR(1)	Reverse	m_0 =AR(1)	Reverse
p-value	0.00	0.65	0.00	0.71
n rejections	49	0	47	0
Panel C			**Panel D**	
	m_0 =mean	Reverse	m_0 =mean	Reverse
p-value	0.03	0.01	0.02	0.00
n rejections	1	6	1	7
n variables × forecasters	1001	1001	1207	1207
25 observations minimum				
Panel E			**Panel F**	
	m_0 =AR(1)	Reverse	m_0 =AR(1)	Reverse
p-value	0.00	0.99	0.00	0.99
n rejections	49	0	47	0
Panel G			**Panel H**	
	m_0 =mean	Reverse	m_0 =mean	Reverse
p-value	0.65	0.01	0.66	0.00
n rejections	0	11	0	13
n variables × forecasters	443	443	452	452

Notes: This table reports Sup test statistics that compare the predictive accuracy of firm-level (left panel) and individual forecasters (right panel) to forecasts from an AR(1) model (Panels A, B, E and F) or the equal-weighted mean forecast (Panels C, D, G and H). The null hypothesis is that the accuracy of forecasts in m_0 are at least as high as that of forecasts in the set of alternatives, m_1. The analysis includes 11 variables: AHE, CPI, ETSL, IP, NFP, NHS, NHSPA, NHSPS, PCEC, PCE and UN. For the top 4 panels, we exclude forecasters with less than five reported values. For the bottom 4 panels, we exclude forecasters with less than twenty-five reported values. The AR(1) model is estimated with an expanding window, using an initial 24-month warm-up sample. The first row in each panel reports the p-value of the Sup test while the second row reports the number of variable-firm rejections. The tests are studentized and use a significance level of $\alpha = 0.10$.

Table 6: Sup tests comparing predictive accuracy across different horizons

Panel A: GDP, m_0 = short horizon, m_1 = long horizon			
H_0: Short-horizon forecasts at least as accurate as long-horizon forecasts			
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
0.992	0.999	0.925	1.000
Panel B: GDP, m_0 = long horizon, m_1 = short horizon			
H_0: Long-horizon forecasts at least as accurate as short-horizon forecasts			
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
0.091	0.002	0.008	0.005
Brazil	Switzerland	Chile	Argentina
Italy	Venezuela	Israel	Brazil
Portugal		Italy	Comoros
		Japan	Congo, Democratic
		Spain	Guyana
		St. Kitts Nevis	Haiti
		Switzerland	Israel
		Ukraine	Italy
		United Kingdom	Kenya
			Lebanon
			Panama
			Peru
			Portugal
			Switzerland
			Tunisia
			United States
			Venezuela
			Zimbabwe
Panel C: Inflation, m_0 = short horizon, m_1 = long horizon			
H_0: Short-horizon forecasts at least as accurate as long-horizon forecasts			
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
0.316	0.944	1.000	0.998
Panel D: Inflation, m_0 = long horizon, m_1 = short horizon			
H_0: Long-horizon forecasts at least as accurate as short-horizon forecasts			
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
0.127	0.000	0.000	0.000
	Angola	Belgium	Angola
	Australia	Dominican Republic	Austria
	Cyprus	Finland	Bangladesh
	Egypt	France	Belarus
	Finland	Indonesia	Belgium
	France	Italy	Canada
	Germany	Japan	Cyprus
	Hungary	Lithuania	Denmark
	Luxembourg	Nepal	Dominican Republic
	Madagascar	Peru	Egypt
	New Zealand	Poland	Estonia
	Slovak Republic	Portugal	Ethiopia
	Slovenia	Singapore	Finland
	Spain	United States	France
	Switzerland		Germany
	Zimbabwe		Ghana
			Guatemala
			India
			Indonesia
			Italy
			Kenya
			Lithuania
			Luxembourg
			Malaysia
			Mongolia
			Mozambique
			New Zealand
			Norway
			Portugal
			Romania
			Spain
			Sweden
			Switzerland
			Thailand
			United States
			Zambia
			Zimbabwe

Notes: The first row in each panel reports the p-value of the Sup test for the null that the benchmark forecasts m_0 are at least as accurate as the alternative forecasts m_1 for all countries included in the comparison. Small p-values indicate rejections of the null. For cases where the null is rejected, we list the countries for which the alternative forecast is significantly more accurate than the benchmark forecast, i.e., countries whose t-statistics are higher than the 90% quantile of the maximum value of the bootstrapped t-statistic. Panels A and B examine GDP forecasts while Panels C and D examine inflation forecasts. Columns 1-3 compare WEO forecasts at consecutive revision points, while column 4 evaluates the cumulative revision from the one-year-ahead spring forecasts ($h = 1, S$) to the current-year fall forecasts ($h = 0, F$).

Table 7: **Sup tests comparing predictive accuracy across clusters of economies: Long- versus short-horizon WEO forecasts**

Panel A: GDP Growth

	world	ae	emde	lics	eeur	dasia	lac	menap	cis	ssa
h=1,S vs. h=1,F	0.01	0.01	0.01	0.07	0.14	0.42	0.00	0.14	0.04	0.05
no. rejections	3	3	1	1	0	0	2	0	1	1
h=1,F vs. h=0,S	0.00	0.01	0.00	0.07	0.20	0.03	0.00	0.12	0.03	0.05
no. rejections	2	6	1	1	0	1	3	0	4	2
h=0,S vs. h=0,F	0.01	0.00	0.04	0.10	0.02	0.03	0.01	0.04	0.00	0.07
no. rejections	9	14	3	0	1	2	5	3	2	1
h=1,S vs. h=0,F	0.00	0.00	0.00	0.00	0.03	0.03	0.00	0.01	0.01	0.00
no. rejections	20	15	15	9	2	4	12	5	3	10
no. countries	186	36	150	58	12	28	32	23	12	43

Panel B: Inflation

	world	ae	emde	lics	eeur	dasia	lac	menap	cis	ssa
h=1,S vs. h=1,F	0.13	0.16	0.11	0.06	0.15	0.13	0.43	0.33	0.43	0.04
no. rejections	0	0	0	2	0	0	0	0	0	2
h=1,F vs. h=0,S	0.00	0.00	0.03	0.03	0.00	0.05	0.03	0.01	0.04	0.01
no. rejections	17	13	7	3	6	1	3	4	2	4
h=0,S vs. h=0,F	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.06	0.02	0.14
no. rejections	15	16	5	2	3	5	5	1	2	0
h=1,S vs. h=0,F	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00
no. rejections	38	28	20	9	5	9	7	7	6	8
no. countries	185	36	149	58	12	28	31	23	12	43

Notes: This table reports p-values for Sup tests comparing long-horizon to short-horizon WEO forecasts. The null hypothesis is that none of the long-horizon WEO forecasts are less accurate than the corresponding short-horizon forecasts for each of the countries within a particular group. Small p-values indicate that the null is rejected and some short-horizon WEO forecasts are significantly more accurate than their long-horizon counterparts. Each panel also shows the number of countries for which the null hypothesis is rejected using a nominal size of $\alpha = 0.1$. 'ae' refers to advanced economies, 'emde' is emerging and developing economies, 'eeur' is emerging and developing Europe, 'lics' is low income developing countries, '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.

Table 8: Sup tests comparing predictive accuracy across different forecast horizons for advanced economies

Panel A: GDP, m_0 = short horizon, m_1 = long horizon			
H_0: Short-horizon forecasts at least as accurate as long-horizon forecasts			
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
0.975	0.755	1.000	1.000

Panel B: GDP, m_0 = long horizon, m_1 = short horizon			
H_0: Long-horizon forecasts at least as accurate as short-horizon forecasts			
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
0.007	0.007	0.003	0.002
Italy	Canada	Belgium	Belgium
Japan	Hong Kong SAR	Canada	Canada
Portugal	Luxembourg	Cyprus	Cyprus
	Portugal	Estonia	Finland
	Switzerland	France	France
	United States	Israel	Germany
		Italy	Greece
		Japan	Hong Kong SAR
		Latvia	Ireland
		New Zealand	Israel
		Portugal	Italy
		Spain	Japan
		Switzerland	Luxembourg
		United Kingdom	Malta
			Portugal
			Switzerland
			United States

Panel C: Inflation, m_0 = short horizon, m_1 = long horizon			
H_0: Short-horizon forecasts at least as accurate as long-horizon forecasts			
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
0.888	1.000	1.000	1.000

Panel D: Inflation, m_0 = long horizon, m_1 = short horizon				
H_0: Long-horizon forecasts at least as accurate as short-horizon forecasts				
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	
0.151	0.000	0.001	0.000	
	Australia	Belgium	Austria	Netherlands
	Cyprus	Canada	Belgium	New Zealand
	Finland	Denmark	Canada	Norway
	France	Finland	Cyprus	Portugal
	Germany	France	Czech Republic	Singapore
	Italy	Germany	Denmark	Slovak Republic
	Luxembourg	Italy	Estonia	Slovenia
	New Zealand	Japan	Finland	Spain
	Slovak Republic	Lithuania	France	Sweden
	Slovenia	New Zealand	Germany	Switzerland
	Spain	Norway	Ireland	United Kingdom
	Switzerland	Portugal	Italy	United States
		Singapore	Japan	
		Slovak Republic	Korea	
		United Kingdom	Lithuania	
		United States	Luxembourg	

Notes: The first row in each panel reports the p-value of the Sup test for the null that the benchmark forecasts m_0 are at least as accurate as the forecasts in the alternative set m_1 for all countries included in the comparison. Small p-values indicate rejections of the null. For cases where the null is rejected, we list the countries for which the alternative forecast is significantly more accurate than the benchmark forecast, i.e., countries whose t-statistics are higher than the 90% quantile of the maximum value of the bootstrapped t-statistic. Panels A and B examine GDP forecasts while Panels C and D examine inflation forecasts. Columns 1-3 compare WEO forecasts at consecutive revision points, while column 4 evaluates the cumulative revision from the one-year-ahead spring forecasts ($h = 1, S$) to the current-year fall forecasts ($h = 0, F$).

Figure 1: Distribution of RMSE ratios for individual forecasters relative to benchmarks.
This figure shows histograms depicting the distribution of the ratio of the root mean squared forecast errors of individual forecasters relative to forecasts from AR(1) model (left column) and equal-weighted mean forecasts (right column). We present results for ETSL (existing home sales), GDP, IP (industrial production) NFP (nonfarm payrolls), and UN (unemployment rate).

(a) Firms vs AR(1) (b) Firms vs mean

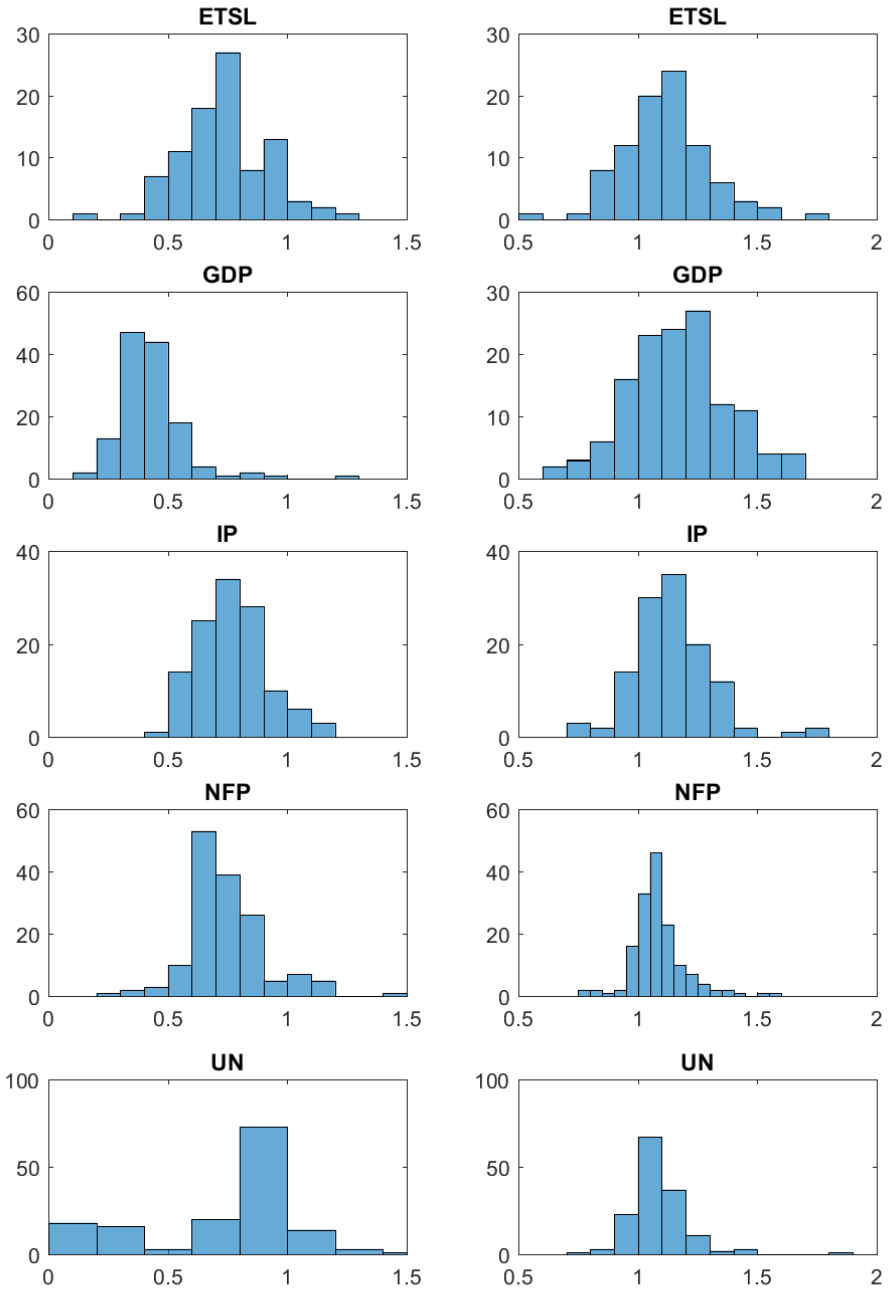
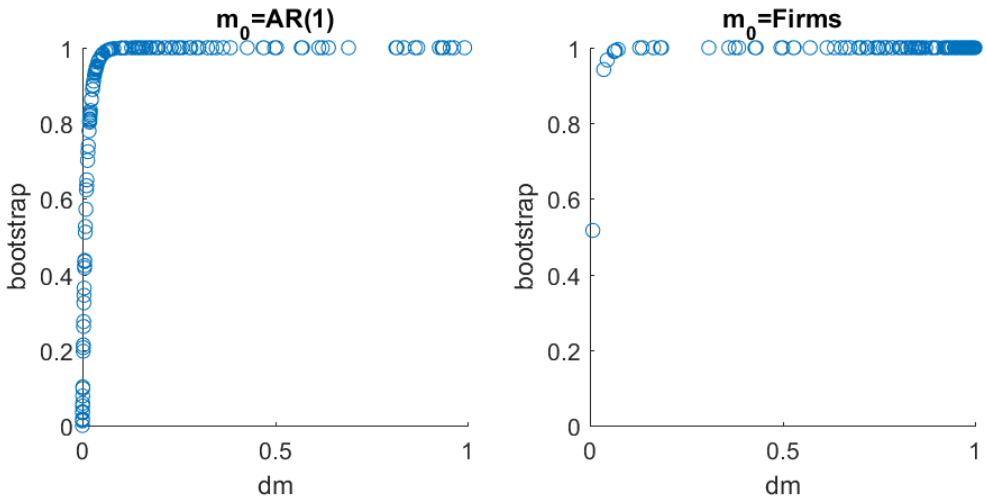


Figure 2: Comparisons of p-values from pairwise Diebold-Mariano and Sup tests for unemployment forecasts.

This figure plots p-values from the pairwise Diebold-Mariano tests on the horizontal axis against bootstrapped p-values from the Sup tests (vertical axis) for the null that the forecasts in benchmark set m_0 are at least as accurate as the forecasts in the alternative set, m_1. The top left window uses AR(1) forecasts as m_0, while individual forecasters are in the alternative set. The top right window reverses this, assigning individual forecasters to the set m_0 and the AR(1) forecasts to the alternative set. Similarly, the bottom left window uses the equal-weighted mean forecasts as set m_0, while individual forecasters are in the alternative set and the bottom right window assigns individual forecasters to set m_0 and the equal-weighted mean forecasts to the alternative set.

(a) Firms vs AR(1)



(b) Firms vs mean

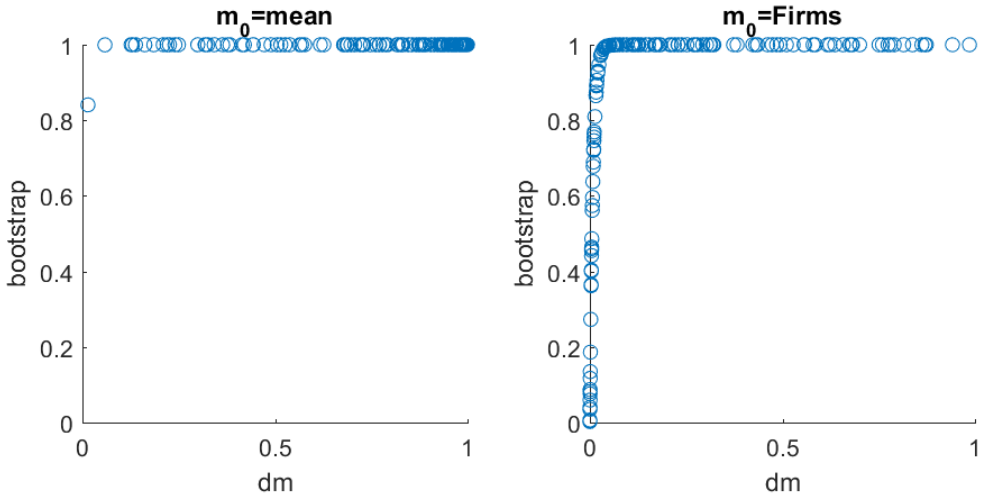
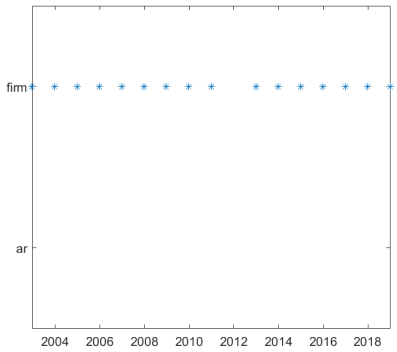
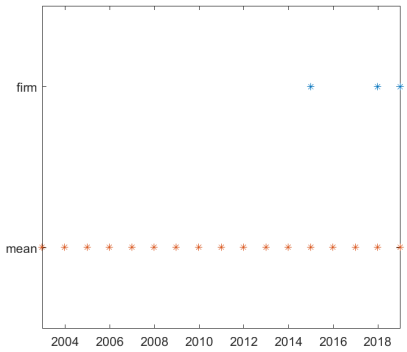


Figure 3: Sup test comparing (firm-level) forecasters' average performance in individual calendar years to forecasts from AR(1) and equal-weighted mean.
The top panels test whether the benchmark forecasts (m_0) are at least as accurate as all forecasts in the alternative set (m_1) in individual years. The bottom panels test whether the benchmark forecasts are at least as accurate as all forecasts in the alternative set during every single year in the sample. Blue asterisks listed in the top row of each panel indicate periods in which the null is rejected at the 10% significance level, suggesting that at least one firm forecaster is more accurate than the AR(1) (left two panels) or equal-weighted mean forecast (right two panels). Red asterisks in the bottom rows indicate periods in which the opposite holds and at least one forecaster is significantly less accurate than the AR(1) (left two panels) or equal-weighted mean forecast (right two panels).

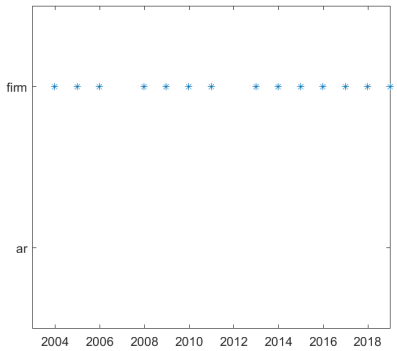
(a) Firms vs AR(1)



(b) Firms vs mean



(c) Firms vs AR(1)



(d) Firms vs mean

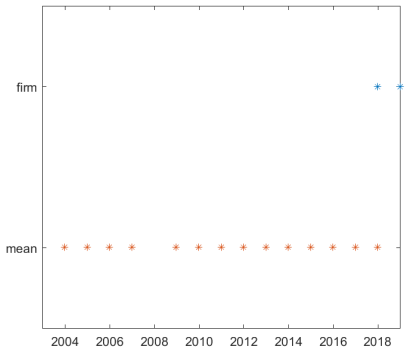
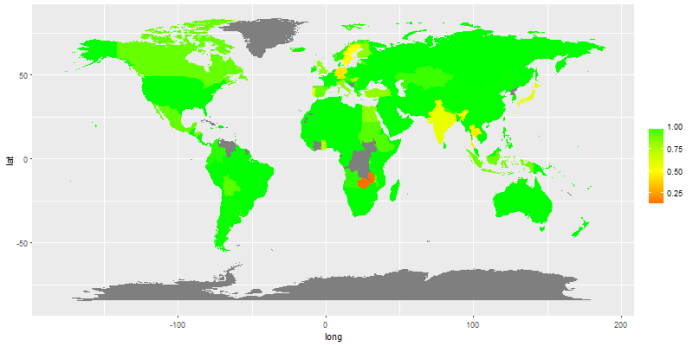
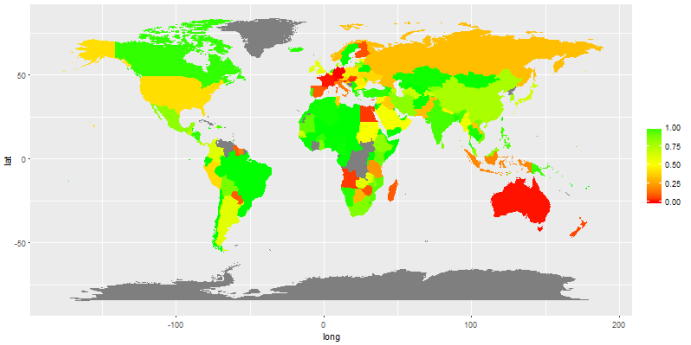


Figure 4: Sup tests comparing predictive accuracy across different horizons for country-level WEO inflation rates.
The figures report p-values from one-sided Sup tests. Red color (a small p-value) indicates that long-horizon forecasts are significantly less accurate for a particular country than the short-horizon forecasts. Green color (large p-values) indicates weak evidence against long-horizon forecasts being at least as accurate as the short-horizon forecasts.

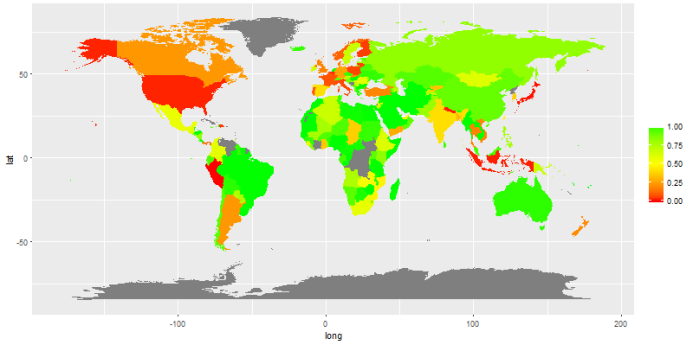
(a) h=1, S vs. h=1, F



(b) h=1, F vs. h=0, S



(c) h=0, S vs. h=0, F



(d) h=1, S vs. h=0, F

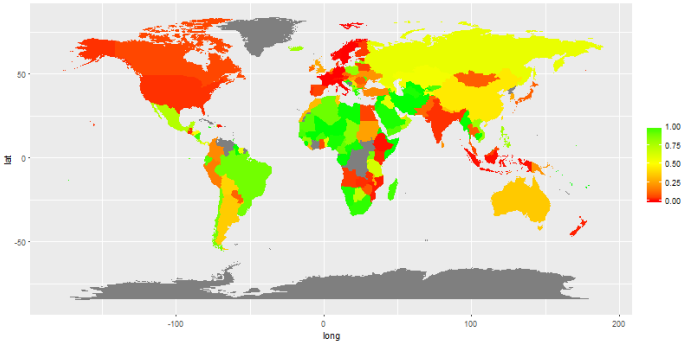
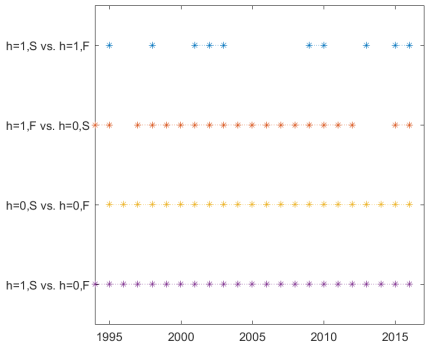


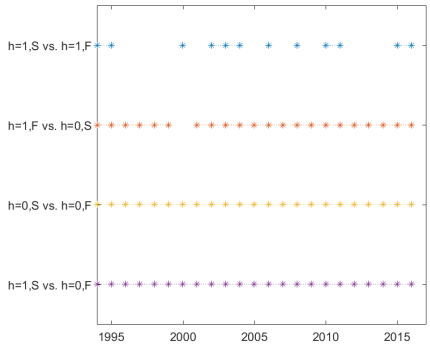
Figure 5: Sup test comparing average performance of long- and short-horizon forecasts in individual calendar years.

The top panels test whether the benchmark forecasts (m_0) are at least as accurate as all forecasts in the alternative set (m_1) in individual years. The bottom panels test whether the benchmark forecasts are at least as accurate as all forecasts in the alternative set during every single year in the sample. Asterisks indicate periods in which the null is rejected at the 10% significance level, suggesting that the short-horizon forecast is significantly more accurate than the long-horizon forecast for at least one country.

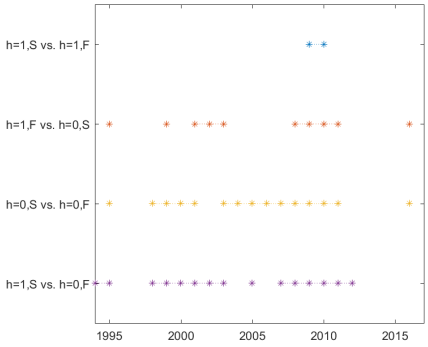
(a) GDP



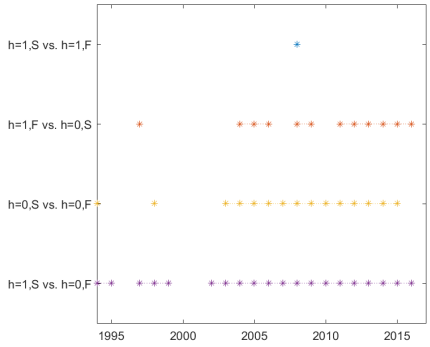
(b) CPI



(c) GDP



(d) CPI



Web Appendices

A Monte Carlo Simulations

To explore the finite-sample performance of our bootstrap procedure for identifying superior forecasting skills, this section conducts a series of Monte Carlo simulations addressing both the size and the power of our bootstrap. Consider the following setup: for forecasters $m = 1, ..., M$, variables $i = 1, ..., N$ and time periods $t + h = 1, ..., T$, the forecast errors are assumed obey the factor structure

$$e_{i,t+h,m} = \lambda_{i,m} f_{t+h} + u_{i,t+h,m},$$

where f_{t+h} is a mean-zero Gaussian AR(1) process with autoregressive coefficient ρ and variance σ_f^2. We generate $\lambda_{i,m}$ as i.i.d random variables from a $N(0, \sigma_\lambda^2)$ distribution truncated such that $\lambda_{i,m}^2 \sigma_f^2 \leq 0.9$; we then set $u_{i,t+h,m}$ as a mean-zero Gaussian AR(1) process with AR coefficient ρ and variance $1 - \lambda_{i,m}^2 \sigma_f^2$. Here, $\{f_{t+h}\}_{t+h=1}^T$, $\{\lambda_{i,m}\}_{1 \leq i \leq N, 1 \leq m \leq M}$ and $\{u_{i,t+h,m}\}_{1 \leq i \leq N, 1 \leq m \leq M, 1 \leq t+h \leq T}$ are mutually independent. We set $(\sigma_f, \sigma_\lambda) = (2, 1.2)$. When $T > 30$, we use $\rho = 0.5$ and a block size $B_T = T^{0.6}$; otherwise, we use $\rho = 0$ and $B_T = 1$. We consider both a no normalization and a partial normalization scheme, both of which are described in Example 3.1. Under these schemes, all forecast errors have MSE values equal to one and thus the null hypothesis that no forecasts underperform the baseline model, m_0, holds.

Table A1 reports size results from 1,200 Monte Carlo simulations using a variety of combinations for the sample size, $T = \{25, 50, 100, 200\}$, the number of forecasters, $M = \{2, 10, 100\}$ and the number of outcome variables $N = \{1, 10, 25, 50, 100\}$.[33] We report results both with and without studentizing the Sup test statistic and use critical values of $\alpha = 0.05, 0.10$.

In general, the size of the non-studentized test statistic is closely aligned with the true size although it tends to be undersized for large values of N and T, particularly when M is also large. The size properties of the studentized test statistic are quite good for small-to-modest values of N,M, and T, but this test statistic tends to be severely undersized when N, T, M are large. The undersizing is particularly pronounced for $\alpha = 0.05$. Interestingly, when the time-series dimension is small ($T = 25$), the studentized test statistic is actually over-sized and the rejection rate increases in the number of variables, N. This pattern reverses in the tests that use larger sample sizes, i.e., $T = 50, 100, 200$.

The size simulations can be used to compute size-adjusted critical values that deliver more accurate finite-sample performance. Although using α as the critical value for the

[33] Each MC simulation uses 250 bootstraps.

p-value leads to asymptotically exact tests, this might not be the case in finite samples. In particular, for each value of (N, M, T), we can compute size-adjusted critical values for the p-value such that the rejection probability for this sample size under the null hypothesis is made to be exactly α. Note that whenever the rejection rate in Table A1 exceeds α, the corresponding size-adjusted p-value, displayed in Table A2, will be adjusted downwards (below α), whereas the reverse holds when the rejection rate in Table A1 falls below α. An interesting observation from Table A2 is that using a critical level of $\alpha = 0.10$ for the studentized test statistic in many cases gets us close to a size of 5%. This is the chief reason why we use a 10% size throughout the empirical analysis.

To explore the power properties of the Sup test statistics with and without studentization, consider the following setup. For each of the N outcome variables, we use one forecast as the benchmark while the remaining $(M-1)$ forecasts are competitors. In other words, we split the NM forecasts into N benchmarks and $(N-1)M$ competing forecasts. Next, we randomly select 20% of these competing forecasts and add $\left(2T^{-1}\log(MN)\right)^{1/8}$ to the selected forecast errors, which then have larger MSE than the baseline forecasts.

This design for the power experiments is in line with that in Chernozhukov et al. (2018). In their simulations, 5% of the moments violate the null hypothesis while this figure is 20% in ours. We choose a larger percentage of moments that violate the null hypothesis due to the smaller sample size in our experiments: their sample size is always 400 and our sample size ranges from 25 to 200. Because of our smaller sample sizes, it is necessary to let the magnitude of departures from the null depend on the sample size in order to obtain meaningful comparisons; if we change $\left(2T^{-1}\log(MN)\right)^{1/8}$ to a fixed value, we would likely find that the methods either have power close to one or close to the nominal size.

Table A3 reports the power of the Sup test statistics with and without studentization. To facilitate comparisons across the two test statistics, we use size-adjusted critical values. With exception of a few instances when $N = 1$, across the board, the power of the Sup test statistic is much higher with studentization than without, e.g., the power can be 10-20% for the non-studentized test statistic but 70-80% for the studentized test statistic. The general conclusion from is, thus, that using the studentized rather than the non-studentized test statistic yields far better power.

B Proofs

This appendix provides proofs for the theoretical results in our paper.

B.1 Preliminary results

Before proving our theorem, we start by recalling some results from Chernozhukov et al. (2018), re-stated here using our notation so that these results can be readily used in our analysis. Let $q_T > r_T$ with $q_T + r_T \leq T/2$. Further, let $B_T = q_T + r_T$ and $K = K_T = \lfloor T/(q_T + r_T) \rfloor$ (the integer part of $T/(q_T + r_T)$). For $1 \leq k \leq K$, define $A_k = \{t : (k-1)(q_T + r_T) + 1 \leq t \leq (k-1)(q_T + r_T) + q_T\}$. Let $\{\varepsilon_k\}_{k=1}^K$ be i.i.d $N(0,1)$ random variables that are independent of the data. In the proofs, we use W_t instead of W_{t+h} for notational simplicity; changing $t + h$ to t does not affect the theoretical arguments.

Theorem B.1 (Theorem B.1 of Chernozhukov et al. (2018))**.** *Let Assumption 1 hold. Then there exist constants $C, c > 0$ depending only on c_1, c_2 and C_1 such that*

$$E \sup_{x \in \mathbb{R}} \left| P\left(\max_{1 \leq j \leq \mathcal{N}} T^{-1/2} \sum_{t=1}^{T} W_{jt} \leq x \right) - P\left(\max_{1 \leq j \leq \mathcal{N}} \frac{1}{\sqrt{K q_T}} \sum_{k=1}^{K} \sum_{t \in A_k} (W_{j,t} - \hat{\mu}_j)\varepsilon_k \leq x \mid \{W_s\}_{s=1}^T \right) \right| \leq C T^{-c},$$

where $\hat{\mu}_j = T^{-1} \sum_{t=1}^T W_{jt}$. Moreover,

$$\sup_{x \in \mathbb{R}} \left| P\left(\max_{1 \leq j \leq \mathcal{N}} T^{-1/2} \sum_{t=1}^{T} W_{jt} \leq x \right) - P\left(\max_{1 \leq j \leq \mathcal{N}} Z_j \leq x \right) \right| \leq C T^{-c},$$

where $Z = (Z_1, ..., Z_{\mathcal{N}})' \in \mathbb{R}^{\mathcal{N}}$ is a centered Gaussian vector with variance matrix $EZZ' = (K q_T)^{-1} \sum_{k=1}^K E\left[\left(\sum_{t \in A_k} W_t \right) \left(\sum_{t \in A_k} W_t \right)' \right]$.

Proof. The first claim follows from the statement of Theorem B.1 of Chernozhukov et al. (2018). The second statement is from the proof of Theorem B.1 of Chernozhukov et al. (2018); see Equation (94) therein. □

B.1.1 Bootstrap approximation of normalized test statistic

The following theorem is a general result on the bootstrap approximation of the normalized test statistic, assuming a good approximation of the normalization. In Appendix B.1.2, we provide further results on the approximation of the normalization. We first state the following Theorem B.2, present its proof, and then prove the auxiliary lemmas used.

Theorem B.2. *Let Assumption 1 hold. Suppose that $T^{-1}\sqrt{K} r_T (\log \mathcal{N})^{3/2} = o(1)$, $T^{-1} q_T (\log \mathcal{N})^{3/2} = o(1)$, $T^{-1} K r_T \log^2 \mathcal{N} = o(1)$ and $T^{-1} r_T^2 D_T^2 \log^3 \mathcal{N} = o(1)$. Let*

$a = (a_1, ..., a_{\mathcal{N}})' \in \mathbb{R}^{\mathcal{N}}$ be nonrandom with $\kappa_1 \leq a_j \leq \kappa_2$ for all $1 \leq j \leq \mathcal{N}$ and $\kappa_1, \kappa_2 > 0$. Suppose that $\hat{a} = (\hat{a}_1, ..., \hat{a}_{\mathcal{N}})$ satisfies $\min_{1\leq j\leq \mathcal{N}} \hat{a}_j > 0$ and $\|\hat{a} - a\|_\infty = o_P(1/\log \mathcal{N})$. Then

$$\sup_{x\in\mathbb{R}} \left| P\left(\max_{1\leq j\leq \mathcal{N}} \frac{T^{-1/2}\sum_{t=1}^T W_{jt}}{\hat{a}_j} \leq x \right) - P\left(\max_{1\leq j\leq \mathcal{N}} \frac{T^{-1/2}\sum_{k=1}^K \sum_{t\in \bar{A}_k} (W_{j,t} - \hat{\mu}_j)\varepsilon_k}{\hat{a}_j} \leq x \mid \{W_s\}_{s=1}^T \right) \right| = o_P(1),$$

where $\bar{A}_k = \{t : \ (k-1)(q_T + r_T) + 1 \leq t \leq k(q_T + r_T)\}$.

Proof. We first apply Theorem B.1 to $\{(a_1^{-1}W_{1t}, ..., a_{\mathcal{N}}^{-1}W_{\mathcal{N}t})\}_{t=1}^T$, obtaining

$$E\sup_{x\in\mathbb{R}} \left| P\left(\max_{1\leq j\leq \mathcal{N}} a_j^{-1}T^{-1/2}\sum_{t=1}^T W_{jt} \leq x \right) - P\left(\max_{1\leq j\leq \mathcal{N}} a_j^{-1}(Kq_T)^{-1/2}\sum_{k=1}^K \sum_{t\in A_k} (W_{j,t} - \hat{\mu}_j)\varepsilon_k \leq x \mid \{W_s\}_{s=1}^T \right) \right| \leq CT^{-c},$$

where $C, c > 0$ are constants that only depend on c_1, c_2, C_1, κ_1 and κ_2. By Lemma B.5,

$$\sup_{x\in\mathbb{R}} \left| P\left(\max_{1\leq j\leq \mathcal{N}} a_j^{-1}(Kq_T)^{-1/2}\sum_{k=1}^K \sum_{t\in A_k} (W_{j,t} - \hat{\mu}_j)\varepsilon_k \leq x \mid \{W_s\}_{s=1}^T \right) - P\left(\max_{1\leq j\leq \mathcal{N}} \hat{a}_j^{-1}T^{-1/2}\sum_{k=1}^K \sum_{t\in \bar{A}_k} (W_{j,t} - \hat{\mu}_j)\varepsilon_k \leq x \mid \{W_s\}_{s=1}^T \right) \right| = o_P(1).$$

Therefore, we have

$$\sup_{x\in\mathbb{R}} \left| P\left(\max_{1\leq j\leq \mathcal{N}} a_j^{-1}T^{-1/2}\sum_{t=1}^T W_{jt} \leq x \right) - P\left(\max_{1\leq j\leq \mathcal{N}} \hat{a}_j^{-1}T^{-1/2}\sum_{k=1}^K \sum_{t\in \bar{A}_k} (W_{j,t} - \hat{\mu}_j)\varepsilon_k \leq x \mid \{W_s\}_{s=1}^T \right) \right| = o_P(1).$$

It follows from Lemma B.3 that

$$\sup_{x\in\mathbb{R}} \left| P\left(\max_{1\leq j\leq \mathcal{N}} a_j^{-1}T^{-1/2}\sum_{t=1}^T W_{jt} \leq x \right) - P\left(\max_{1\leq j\leq \mathcal{N}} \hat{a}_j^{-1}T^{-1/2}\sum_{t=1}^T W_{jt} \leq x \right) \right| = o_P(1).$$

The desired result follows from this. □

Lemma B.1. *Let* $R = (R_1, ..., R_{\mathcal{N}})'$, $\hat{R} = (\hat{R}_1, ..., \hat{R}_{\mathcal{N}})'$, $\zeta = (\zeta_1, ..., \zeta_{\mathcal{N}})'$, $\hat{\zeta} = (\hat{\zeta}_1, ..., \hat{\zeta}_{\mathcal{N}})'$ *and* $Z = (Z_1, ..., Z_{\mathcal{N}})'$ *be random vectors in* $\mathbb{R}^{\mathcal{N}}$. *Suppose that* ζ *and* $\hat{\zeta}$ *are* $\mathcal{F}$*-measurable for some* σ*-algebra* $\mathcal{F}$. *Also assume that* Z *is a centered Gaussian vector with* $\min_{1\leq j\leq \mathcal{N}} E(Z_j^2) \geq b$ *almost surely for some constant* $b > 0$. *If* $\max_{1\leq j\leq \mathcal{N}} |\hat{R}_j - R_j| = o_P(1/\sqrt{\log \mathcal{N}})$ *as* $N \to \infty$ *(or other dimensions tend to infinity), then the following holds:*

$$E \sup_{x\in\mathbb{R}} \left| P\left(\max_{1\leq j\leq \mathcal{N}} Z_j \leq x \right) - P(\max_{1\leq j\leq \mathcal{N}} \hat{R}_j \leq x \mid \mathcal{F}) \right| \\ \leq 3E \sup_{x\in\mathbb{R}} \left| P\left(\max_{1\leq j\leq \mathcal{N}} Z_j \leq x \right) - P\left(\max_{1\leq j\leq \mathcal{N}} R_j \leq x \mid \mathcal{F} \right) \right| + o(1).$$

Moreover, if $\|\hat{\zeta} - \zeta\|_\infty = o_P(1/\sqrt{\log \mathcal{N}})$, *the following holds:*

$$\sup_{x\in\mathbb{R}} \left| P\left(\max_{1\leq j\leq \mathcal{N}} Z_j \leq x \right) - P\left(\max_{1\leq j\leq \mathcal{N}} \hat{\zeta}_j \leq x \right) \right| \\ \leq 3 \sup_{x\in\mathbb{R}} \left| P\left(\max_{1\leq j\leq \mathcal{N}} Z_j \leq x \right) - P\left(\max_{1\leq j\leq \mathcal{N}} \zeta_j \leq x \right) \right| + o(1).$$

Proof. **Step 1:** show the first claim.

For an arbitrary $\eta > 0$, let $c = \eta/\sqrt{\log \mathcal{N}}$. Define the event $\mathcal{M} = \{\max_{1\leq j\leq \mathcal{N}} |\hat{R}_j - R_j| \leq c\}$ and variables $\xi = \max_{1\leq j\leq \mathcal{N}} R_j$ and $\hat{\xi} = \max_{1\leq j\leq \mathcal{N}} \hat{R}_j$. Let $a_N = \sup_{x\in\mathbb{R}} |P(\max_{1\leq j\leq \mathcal{N}} Z_j \leq x) - P(\xi \leq x \mid \mathcal{F})|$.

We first notice that, given the event $\mathcal{M}$, $|\hat{\xi} - \xi| \leq c$, and thus

$$\begin{aligned} &\left| \mathbf{1}\{\xi \leq x\} - \mathbf{1}\left\{\hat{\xi} \leq x\right\} \right| \\ &= \mathbf{1}\left\{\hat{\xi} \leq x \text{ and } \xi > x\right\} + \mathbf{1}\left\{\hat{\xi} > x \text{ and } \xi \leq x\right\} \\ &= \mathbf{1}\left\{\xi - \hat{\xi} \geq \xi - x \text{ and } \xi - x > 0\right\} + \mathbf{1}\left\{\xi - \hat{\xi} < \xi - x \text{ and } \xi - x \leq 0\right\} \\ &\leq \mathbf{1}\left\{|\xi - x| \leq |\hat{\xi} - \xi|\right\} \leq \mathbf{1}\{|\xi - x| \leq c\}. \end{aligned} \tag{27}$$

Hence,

$$\begin{aligned} &\left| P(\xi \leq x \mid \mathcal{F}) - P(\hat{\xi} \leq x \mid \mathcal{F}) \right| \\ &\leq E\left[\left| \mathbf{1}\{\xi \leq x\} - \mathbf{1}\left\{\hat{\xi} \leq x\right\} \right| \mid \mathcal{F} \right] \\ &\leq P(|\xi - x| \leq c \mid \mathcal{F}) + P(\mathcal{M}^c \mid \mathcal{F}) \\ &\leq P(\xi \leq x + c \mid \mathcal{F}) - P(\xi \leq x - 2c) + P(\mathcal{M}^c \mid \mathcal{F}) \end{aligned}$$

$$\leq P\left(\max_{1\leq j\leq \mathcal{N}} Z_j \leq x+c \mid \mathcal{F}\right) - P\left(\max_{1\leq j\leq \mathcal{N}} Z_j \leq x-2c\right) + P(\mathcal{M}^c \mid \mathcal{F}) + 2a_N. \tag{28}$$

Let $\iota = (1, ..., 1)' \in \mathbb{R}^{\mathcal{N}}$. Then, by Lemma A.1 of Chernozhukov et al. (2017), it follows that almost surely, for any $x \in \mathbb{R}$,

$$P\left(Z \leq (x+c)\iota \mid \mathcal{F}\right) - P\left(Z \leq (x-2c)\iota \mid \mathcal{F}\right) \leq 3cC_b\sqrt{\log \mathcal{N}},$$

where $C_b > 0$ is a constant that only depends on b. Here, $Z \leq (x+c)\iota$ means $Z_j \leq (x+c)$ for all $1 \leq j \leq \mathcal{N}$; similarly, $Z \leq (x-2c)\iota$ means that $Z_j \leq (x-2c)$ for all $1 \leq j \leq \mathcal{N}$. Hence, for any $z \in \mathbb{R}$, $P(Z \leq z\iota) = P(\max_{1\leq j\leq \mathcal{N}} Z_j \leq z)$. Therefore, the above display implies that for any $x \in \mathbb{R}$,

$$P\left(\max_{1\leq j\leq \mathcal{N}} Z_j \leq x+c \mid \mathcal{F}\right) - P\left(\max_{1\leq j\leq \mathcal{N}} Z_j \leq x-2c\right) \leq 3cC_b\sqrt{\log \mathcal{N}}. \tag{29}$$

By (28), we have

$$\left|P(\xi \leq x \mid \mathcal{F}) - P(\hat{\xi} \leq x \mid \mathcal{F})\right| \leq 3cC_b\sqrt{\log \mathcal{N}} + 2a_N + P(\mathcal{M}^c \mid \mathcal{F}).$$

Since the above display holds for any $x \in \mathbb{R}$, we have

$$\sup_{x\in\mathbb{R}}\left|P(\xi \leq x \mid \mathcal{F}) - P(\hat{\xi} \leq x \mid \mathcal{F})\right| \leq 3cC_b\sqrt{\log \mathcal{N}} + 2a_N + P(\mathcal{M}^c \mid \mathcal{F}),$$

and, thus,

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\leq j\leq \mathcal{N}} Z_j \leq x\right) - P(\hat{\xi} \leq x \mid \mathcal{F})\right| \leq 3a_N + 3cC_b\sqrt{\log \mathcal{N}} + P(\mathcal{M}^c \mid \mathcal{F}).$$

Taking expectations on both sides, we obtain

$$E\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\leq j\leq \mathcal{N}} Z_j \leq x\right) - P(\hat{\xi} \leq x \mid \mathcal{F})\right| \leq 3Ea_N + 3cC_b\sqrt{\log \mathcal{N}} + P(\mathcal{M}^c) = 3Ea_N + 3\eta C_b + P(\mathcal{M}^c).$$

Since $\max_{1\leq j\leq \mathcal{N}} |\hat{R}_j - R_j| = o_P(1/\sqrt{\log \mathcal{N}})$, $P(\mathcal{M}^c) = o(1)$, and so

$$E\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\leq j\leq \mathcal{N}} Z_j \leq x\right) - P(\hat{\xi} \leq x \mid \mathcal{F})\right| \leq 3Ea_N + 3\eta C_b + o(1).$$

Because $\eta > 0$ is arbitrary, it follows that

$$E\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\leq j\leq \mathcal{N}} Z_j \leq x\right) - P(\hat{\xi} \leq x \mid \mathcal{F})\right| = Ea_N + o(1).$$

Step 2: show the second claim.

The argument is similar to Step 1, but we include the details for completeness.

Fix an arbitrary $\eta > 0$. Let $c_1 = \eta/\sqrt{\log \mathcal{N}}$, $\psi = \max_{1\le j\le \mathcal{N}} \zeta_j$ and $\hat{\psi} = \max_{1\le j\le \mathcal{N}} \hat{\zeta}_j$. Define $d_N = \sup_{x\in\mathbb{R}} |P(\psi \le x) - P(\max_{1\le j\le \mathcal{N}} Z_j \le x)|$. Define the event $\mathcal{M}_1 = \{\|\hat{\zeta} - \zeta\|_\infty \le c_1\}$. As in (27), we notice that, given the event $\mathcal{M}_1$,

$$\left|\mathbf{1}\{\psi \le x\} - \mathbf{1}\left\{\hat{\psi} \le x\right\}\right| \le \mathbf{1}\{|\psi - x| \le c_1\}.$$

Thus,

$$\begin{aligned}
&\left|P(\psi \le x) - P\left(\hat{\psi} \le x\right)\right| \\
&\le P(|\psi - x| \le c_1) + P(\mathcal{M}_1^c) \\
&= P(x - c_1 \le \psi \le x + c_1) + P(\mathcal{M}_1^c) \\
&\le P(\psi \le x + c_1) - P(\psi \le x - 2c_1) + P(\mathcal{M}_1^c) \\
&\le P\left(\max_{1\le j\le \mathcal{N}} Z_j \le x + c_1\right) - P\left(\max_{1\le j\le \mathcal{N}} Z_j \le x - 2c_1\right) + 2d_N + P(\mathcal{M}_1^c) \\
&\overset{\text{(i)}}{\le} 3c_1 C_b \sqrt{\log \mathcal{N}} + 2d_N + P(\mathcal{M}_1^c),
\end{aligned}$$

where (i) follows by (29) (with c replaced by c_1). Since the above bound holds for any $x \in \mathbb{R}$, we have that, given the event $\mathcal{M}_1$,

$$\sup_{x\in\mathbb{R}} \left|P(\psi \le x) - P\left(\hat{\psi} \le x\right)\right| \le 3c_1 C_b \sqrt{\log \mathcal{N}} + 2d_N = 6c_1 C_b \eta + 2d_N + P(\mathcal{M}_1^c).$$

Therefore,

$$\sup_{x\in\mathbb{R}} \left|P\left(\max_{1\le j\le \mathcal{N}} Z_j \le x\right) - P\left(\hat{\psi} \le x\right)\right| \le 6c_1 C_b \eta + 3d_N + P(\mathcal{M}_1^c).$$

Notice that $P(\mathcal{M}_1^c) = o(1)$ due to $\|\hat{\zeta} - \zeta\|_\infty = o_P(1/\sqrt{\log \mathcal{N}})$. Since $\eta > 0$ is arbitrary, we have

$$\sup_{x\in\mathbb{R}} \left|P(\psi \le x) - P\left(\hat{\psi} \le x\right)\right| \le 3d_N + o(1).$$

This completes the proof. □

Lemma B.2. *Suppose that $Z = (Z_1, ..., Z_{\mathcal{N}})'$ is a centered Gaussian vector with $\max_{1\le j\le \mathcal{N}} E(Z_j^2) \le b$ for some constant $b > 0$. Then for any $z \in (0, \mathcal{N}/5)$, $P\left(\|Z\|_\infty \ge \sqrt{2b\log(\mathcal{N}/z)}\right) \le 2z$.*

Proof. Clearly, $Z_j/\sqrt{EZ_j^2} \sim N(0,1)$. Thus, $P\left(|Z_j|/\sqrt{EZ_j^2} > x\right) = 2\Phi(-x) = 2 - 2\Phi(x)$,

where $\Phi(\cdot)$ denotes the cdf of a $N(0,1)$ variable. Since $EZ_j^2 \leq b$, we have that for any $x > 0$, $P\left(|Z_j| > \sqrt{b}x\right) \leq 2(1-\Phi(x))$. By the union bound, it follows that for any $x > 0$,

$$P\left(\max_{1\leq j\leq \mathcal{N}} |Z_j| > \sqrt{b}x\right) \leq 2\mathcal{N}(1-\Phi(x)).$$

Taking $x = \Phi^{-1}(1-a)$ for $a \in (0,1)$, we have $P(\|Z\|_\infty > \sqrt{b}\Phi^{-1}(1-a)) \leq 2\mathcal{N}a$. By Lemma 1 of Zhu and Bradic (2018), for $a \leq 1/5$, $\Phi^{-1}(1-a) \leq \sqrt{2\log(1/a)}$. This means that $P(\|Z\|_\infty > \sqrt{2b\log(1/a)}) \leq 2\mathcal{N}a$. The desired result follows by setting $a = z/\mathcal{N}$. □

Lemma B.3. *Let the assumptions of Theorem B.2 hold. Then*

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\leq j\leq \mathcal{N}} a_j^{-1}T^{-1/2}\sum_{t=1}^{T} W_{jt} \leq x\right) - P\left(\max_{1\leq j\leq \mathcal{N}} \hat{a}_j^{-1}T^{-1/2}\sum_{t=1}^{T} W_{jt} \leq x\right)\right| = o_P(1).$$

Proof. Define $\zeta = (\zeta_1, ..., \zeta_{\mathcal{N}})'$ and $\hat{\zeta} = (\hat{\zeta}_1, ..., \hat{\zeta}_{\mathcal{N}})'$, where $\zeta_j = a_j^{-1}T^{-1/2}\sum_{t=1}^{T} W_{jt}$ and $\hat{\zeta}_j = \hat{a}_j^{-1}T^{-1/2}\sum_{t=1}^{T} W_{jt}$. Also, let $Z = (Z_1, ..., Z_{\mathcal{N}})' \in \mathbb{R}^{\mathcal{N}}$ be a centered Gaussian vector with variance matrix $EZZ' = (Kq_T)^{-1}\sum_{k=1}^{K} E\left[\left(\sum_{t\in A_k} W_t\right)\left(\sum_{t\in A_k} W_t\right)'\right]$.

By Theorem B.1, we have

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\leq j\leq \mathcal{N}} Z_j \leq x\right) - P\left(\max_{1\leq j\leq \mathcal{N}} \zeta_j \leq x\right)\right| = o(1). \tag{30}$$

Applying the same argument with (Z_j, ζ_j) replaced by $(-Z_j, -\zeta_j)$, we obtain

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\leq j\leq \mathcal{N}} (-Z_j) \leq x\right) - P\left(\max_{1\leq j\leq \mathcal{N}} (-\zeta_j) \leq x\right)\right| = o(1).$$

By assumption, $\max_{1\leq j\leq \mathcal{N}} EZ_j^2 \leq C_1$. Hence, by Lemma B.2, we have that, for any $\eta \in (0, 1/5)$,

$$P\left(\|Z\|_\infty \geq \sqrt{2C_1\log(\eta\mathcal{N})}\right) \leq 2\eta.$$

It follows that

$$P\left(\max_{1\leq j\leq \mathcal{N}} \zeta_j > \sqrt{2C_1\log(\eta\mathcal{N})}\right) \leq P\left(\max_{1\leq j\leq \mathcal{N}} Z_j > \sqrt{2C_1\log(\eta\mathcal{N})}\right) + o(1) \leq 2\eta + o(1)$$

and

$$P\left(\max_{1\le j\le\mathcal{N}}(-\zeta_j) > \sqrt{2C_1\log(\eta\mathcal{N})}\right) \le P\left(\max_{1\le j\le\mathcal{N}}(-Z_j) > \sqrt{2C_1\log(\eta\mathcal{N})}\right)+o(1) \le 2\eta+o(1).$$

Therefore,

$$\begin{aligned}&P\left(\|\zeta\|_\infty > \sqrt{2C_1\log(\eta\mathcal{N})}\right)\\ &\le P\left(\max_{1\le j\le\mathcal{N}}\zeta_j > \sqrt{2C_1\log(\eta\mathcal{N})}\right) + P\left(\max_{1\le j\le\mathcal{N}}(-\zeta_j) > \sqrt{2C_1\log(\eta\mathcal{N})}\right) \le 4\eta + o(1).\end{aligned}$$

Since η is arbitrary, we have

$$\|\zeta\|_\infty = O_P\left(\sqrt{\log\mathcal{N}}\right). \tag{31}$$

By Assumption 1, $\min_{1\le j\le\mathcal{N}} E(Z_j^2) \ge c_1$. Next, we verify

$$\|\hat{\zeta}-\zeta\|_\infty = o_P(1/\sqrt{\log\mathcal{N}}). \tag{32}$$

Notice that $\hat{\zeta}_j = \hat{a}_j^{-1}a_j\zeta_j$. Thus,

$$\|\hat{\zeta}-\zeta\|_\infty = \max_{1\le j\le\mathcal{N}}|\hat{\zeta}_j-\zeta_j| \le \|\zeta\|_\infty \max_{1\le j\le\mathcal{N}}\left|\hat{a}_j^{-1}a_j-1\right|.$$

Since $\min_{1\le j\le\mathcal{N}} a_j \ge \kappa_1$ and $\|\hat{a}-a\|_\infty = o_P(1)$, we have $\max_{1\le j\le\mathcal{N}}\left|\hat{a}_j^{-1}a_j-1\right| = O_P(\|\hat{a}-a\|_\infty)$. Since $\|\zeta\|_\infty = O_P(\sqrt{\log\mathcal{N}})$, we have (32) by the assumption on $\|\hat{a}-a\|_\infty$.

Therefore, from Lemma B.1, we have

$$\begin{aligned}&\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le\mathcal{N}}Z_j\le x\right)-P\left(\max_{1\le j\le\mathcal{N}}\hat{\zeta}_j\le x\right)\right|\\ &\qquad\le 3\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le\mathcal{N}}Z_j\le x\right)-P\left(\max_{1\le j\le\mathcal{N}}\zeta_j\le x\right)\right| + o_P(1) \overset{\text{(i)}}{=} o_P(1),\end{aligned}$$

where (i) follows by (30). Now by the triangular inequality, (30) implies

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le\mathcal{N}}\zeta_j\le x\right)-P\left(\max_{1\le j\le\mathcal{N}}\hat{\zeta}_j\le x\right)\right| = o_P(1).$$

This completes the proof. □

Lemma B.4. *Let the assumptions of Theorem B.2 hold. Then*

$$\max_{1\leq j\leq \mathcal{N}}\left|\sum_{k=1}^{K}\sum_{t\in\bar{A}_k\setminus A_k}(W_{j,t}-\hat{\mu}_j)\varepsilon_k\right| = O_P\left(\sqrt{Kr_T\log\mathcal{N}}+r_TD_T\log\mathcal{N}+r_T\sqrt{T^{-1}K\log\mathcal{N}}\right),$$

where $\bar{A}_k=\{t:\ (k-1)(q_T+r_T)+1\leq t\leq k(q_T+r_T)\}$ *for* $1\leq k\leq K-1$ *and* $\bar{A}_K=\{(K-1)(q_T+r_T)+q_T+1,...,T\}$.

Proof. For $1\leq k\leq K$, let $u_{j,k}=\sum_{t\in\bar{A}_k\setminus A_k}W_{j,t}\varepsilon_k$. By Berbee's coupling (e.g., Lemma 7.1 of Chen et al. (2016)), there exist variables $\{v_k\}_{k=1}^K$ with $v_k=(v_{1,k},...,v_{\mathcal{N},k})'$ such that (1) $\{v_k\}_{k=1}^K$ is independent across k and independent of $\{\varepsilon_k\}_{k=1}^K$; (2) v_k has the same distribution as $\sum_{t\in\bar{A}_k\setminus A_k}W_t$ and (3) $P(\bigcap_{k=1}^K\{v_k=\sum_{t\in\bar{A}_k\setminus A_k}W_t\})\geq 1-K\beta_{\text{mixing}}(q_T)$.

Since $\varepsilon_k\sim N(0,1)$, Assumption 1 implies that $\max_{1\leq j\leq\mathcal{N}}\sum_{k=1}^K E(v_{j,k}\varepsilon_k)^2\leq Kr_TC_1$. Also notice that $\|\sum_{t\in\bar{A}_k\setminus A_k}W_t\|_\infty\leq r_TD_T$. It follows by Lemma D.3 of Chernozhukov et al. (2018) that

$$E\left(\max_{1\leq j\leq\mathcal{N}}\left|\sum_{k=1}^K v_{j,k}\varepsilon_k\right|\right)\leq M\left(\sqrt{Kr_TC_1\log\mathcal{N}}+r_TD_T\log\mathcal{N}\right),$$

where $M>0$ is a universal constant. Since $\max_{1\leq j\leq\mathcal{N}}\left|\sum_{k=1}^{K-1}v_{j,k}\varepsilon_k\right| = \max_{1\leq j\leq\mathcal{N}}|\sum_{k=1}^{K-1}u_{j,k}|$ with a probability of at least $1-K\beta_{\text{mixing}}(q_T)$ and $K\beta_{\text{mixing}}(q_T)\leq T\beta_{\text{mixing}}(r_T)=o(1)$, we have that

$$\max_{1\leq j\leq\mathcal{N}}\left|\sum_{k=1}^K u_{j,k}\right| = O_P\left(\sqrt{Kr_T\log\mathcal{N}}+r_TD_T\log\mathcal{N}\right). \tag{33}$$

In the proof of Lemma B.3, we showed that $\max_{1\leq j\leq\mathcal{N}}\hat{\mu}_j=O_P(\sqrt{T^{-1}\log\mathcal{N}})$; see (31). By a similar argument, we have $\max_{1\leq j\leq\mathcal{N}}(-\hat{\mu}_j)=O_P(\sqrt{T^{-1}\log\mathcal{N}})$. Hence,

$$\|\hat{\mu}\|_\infty=O_P(\sqrt{T^{-1}\log\mathcal{N}}).$$

It follows that

$$\begin{aligned}\max_{1\leq j\leq\mathcal{N}}\left|\sum_{k=1}^K\sum_{t\in\bar{A}_k\setminus A_k}\hat{\mu}_j\varepsilon_k\right| &= r_T\|\hat{\mu}\|_\infty\left|\sum_{k=1}^{K-1}\varepsilon_k\right| \\ &= O_P(r_T\sqrt{T^{-1}\log\mathcal{N}}\times\sqrt{K-1})=O_P(r_T\sqrt{T^{-1}K\log\mathcal{N}}).\end{aligned}$$

The above display and (33) imply that

$$\max_{1\le j\le \mathcal{N}}\left|\sum_{k=1}^{K}\sum_{t\in \bar{A}_k\setminus A_k}(W_{j,t}-\hat{\mu}_j)\varepsilon_k\right| = O_P\left(\sqrt{Kr_T\log\mathcal{N}}+r_T D_T\log\mathcal{N}+r_T\sqrt{T^{-1}K\log\mathcal{N}}\right).$$

This completes the proof. □

Lemma B.5. *Let the assumptions of Theorem B.2 hold. Then*

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le\mathcal{N}}a_j^{-1}(Kq_T)^{-1/2}\sum_{k=1}^{K}\sum_{t\in A_k}(W_{j,t}-\hat{\mu}_j)\varepsilon_k\le x\mid\{W_s\}_{s=1}^{T}\right)\right.$$
$$\left.-P\left(\max_{1\le j\le\mathcal{N}}\hat{a}_j^{-1}T^{-1/2}\sum_{k=1}^{K}\sum_{t\in \bar{A}_k}(W_{j,t}-\hat{\mu}_j)\varepsilon_k\le x\mid\{W_s\}_{s=1}^{T}\right)\right| = o_P(1),$$

where $\{\bar{A}_k\}_{k=1}^{K}$ *is defined in the statement of Lemma B.4.*

Proof. Define $R=(R_1,...,R_{\mathcal{N}})'$ and $\hat{R}=(\hat{R}_1,...,\hat{R}_{\mathcal{N}})'$, where $R_j = a_j^{-1}(Kq_T)^{-1/2}\sum_{k=1}^{K}\sum_{t\in A_k}(W_{j,t}-\hat{\mu}_j)\varepsilon_k$ and $\hat{R}_j=\hat{a}_j^{-1}T^{-1/2}\sum_{k=1}^{K}\sum_{t\in\bar{A}_k}(W_{j,t}-\hat{\mu}_j)\varepsilon_k$. Let $\mathcal{F}$ denote the σ-algebra generated by $\{W_s\}_{s=1}^{T}$.

Also, let $Z=(Z_1,...,Z_{\mathcal{N}})'\in\mathbb{R}^{\mathcal{N}}$ be a centered Gaussian vector with variance matrix $EZZ'=(Kq_T)^{-1}\sum_{k=1}^{K}E\left[\left(\sum_{t\in A_k}W_t\right)\left(\sum_{t\in A_k}W_t\right)'\right]$.

By Theorem B.1, we have

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le\mathcal{N}}Z_j\le x\right)-P\left(\max_{1\le j\le\mathcal{N}}R_j\le x\mid\mathcal{F}\right)\right| = o_P(1). \tag{34}$$

Applying the same argument with (Z_j,R_j) replaced by $(-Z_j,-R_j)$, we obtain

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le\mathcal{N}}(-Z_j)\le x\right)-P\left(\max_{1\le j\le\mathcal{N}}(-R_j)\le x\mid\mathcal{F}\right)\right| = o_P(1).$$

By assumption, $\max_{1\le j\le\mathcal{N}}EZ_j^2\le C_1$. Hence, by Lemma B.2, we have that, for any $\eta\in(0,1/5)$,

$$P\left(\|Z\|_\infty\ge\sqrt{2C_1\log(\eta\mathcal{N})}\right)\le 2\eta.$$

It follows that

$$P\left(\max_{1\le j\le\mathcal{N}}R_j>\sqrt{2C_1\log(\eta\mathcal{N})}\mid\mathcal{F}\right)\le P\left(\max_{1\le j\le\mathcal{N}}Z_j>\sqrt{2C_1\log(\eta\mathcal{N})}\right)+o_P(1)\le 2\eta+o_P(1)$$

and

$$P\left(\max_{1\le j\le\mathcal{N}}(-R_j)>\sqrt{2C_1\log(\eta\mathcal{N})}\mid\mathcal{F}\right)\le P\left(\max_{1\le j\le\mathcal{N}}(-Z_j)>\sqrt{2C_1\log(\eta\mathcal{N})}\right)+o_P(1)\le 2\eta+o_P(1).$$

In turn, we have

$$\begin{aligned}&P\left(\|R\|_\infty>\sqrt{2C_1\log(\eta\mathcal{N})}\mid\mathcal{F}\right)\\&\le P\left(\max_{1\le j\le\mathcal{N}}R_j>\sqrt{2C_1\log(\eta\mathcal{N})}\mid\mathcal{F}\right)+P\left(\max_{1\le j\le\mathcal{N}}(-R_j)>\sqrt{2C_1\log(\eta\mathcal{N})}\mid\mathcal{F}\right)\le 4\eta+o_P(1).\end{aligned}$$

Since η is arbitrary, we have

$$\|R\|_\infty=O_P(\sqrt{\log\mathcal{N}}).\tag{35}$$

By Assumption 1, $\min_{1\le j\le\mathcal{N}}EZ_j^2\ge c_1$. Hence, by Lemma B.1, it suffices to verify that

$$\max_{1\le j\le\mathcal{N}}|\hat{R}_j-R_j|=o_P(1/\sqrt{\log\mathcal{N}}).\tag{36}$$

We notice that

$$\hat{a}_j\sqrt{T}\hat{R}_j-a_j\sqrt{Kq_T}R_j=\sum_{k=1}^{K}\sum_{t\in\bar{A}_k\setminus A_k}(W_{j,t}-\hat{\mu}_j)\varepsilon_k.$$

Hence, by Lemma B.4, we have

$$\max_{1\le j\le\mathcal{N}}\left|\hat{a}_j\sqrt{T}\hat{R}_j-a_j\sqrt{Kq_T}R_j\right|=O_P\left(\sqrt{Kr_T\log\mathcal{N}}+r_TD_T\log\mathcal{N}+r_T\sqrt{T^{-1}K\log\mathcal{N}}\right).$$

Since $\|\hat{a}-a\|_\infty=o_P(1)$ and $\min_{1\le j\le\mathcal{N}}a_j\ge\kappa_1>0$, we have

$$\max_{1\le j\le\mathcal{N}}\left|\hat{R}_j-a_j\hat{a}_j^{-1}\sqrt{Kq_T/T}R_j\right|=O_P\left(\sqrt{T^{-1}Kr_T\log\mathcal{N}}+T^{-1/2}r_TD_T\log\mathcal{N}+r_TT^{-1}\sqrt{K\log\mathcal{N}}\right).$$

By (35), it follows that

$$\begin{aligned}&\max_{1\le j\le\mathcal{N}}\left|\hat{R}_j-R_j\right|\\&\le\max_{1\le j\le\mathcal{N}}\left|\hat{R}_j-a_j\hat{a}_j^{-1}\sqrt{Kq_T/T}R_j\right|+\max_{1\le j\le\mathcal{N}}\left|a_j\hat{a}_j^{-1}\sqrt{Kq_T/T}-1\right|\times\max_{1\le j\le\mathcal{N}}|R_j|\\&=O_P\left(\sqrt{T^{-1}Kr_T\log\mathcal{N}}+T^{-1/2}r_TD_T\log\mathcal{N}+r_TT^{-1}\sqrt{K\log\mathcal{N}}\right)\\&\quad+\max_{1\le j\le\mathcal{N}}\left|a_j\hat{a}_j^{-1}\sqrt{Kq_T/T}-1\right|O_P(\sqrt{\log\mathcal{N}})\end{aligned}$$

$$\leq O_P\left(\sqrt{T^{-1}Kr_T\log\mathcal{N}}+T^{-1/2}r_TD_T\log\mathcal{N}+r_TT^{-1}\sqrt{K\log\mathcal{N}}\right)$$
$$+\max_{1\leq j\leq\mathcal{N}}\left|a_j\hat{a}_j^{-1}-1\right|\sqrt{Kq_T/T}O_P(\sqrt{\log\mathcal{N}})+\max_{1\leq j\leq\mathcal{N}}\left|\sqrt{Kq_T/T}-1\right|O_P(\sqrt{\log\mathcal{N}}). \tag{37}$$

By assumption, we have $\|\hat{a}-a\|_\infty=o_P(1/\log\mathcal{N})$ and $\min_{1\leq j\leq\mathcal{N}}a_j\geq\kappa_1>0$. Observe that

$$\max_{1\leq j\leq\mathcal{N}}\left|a_j\hat{a}_j^{-1}-1\right|\sqrt{Kq_T/T}\leq O_P\left(\max_{1\leq j\leq\mathcal{N}}|\hat{a}_j-a_j|\right)$$

and

$$\left|\sqrt{Kq_T/T}-1\right|=\frac{1-Kq_T/T}{1+\sqrt{Kq_T/T}}<\frac{T-Kq_T}{T}$$
$$<\frac{((K+1)(q_T+r_T)-Kq_T)}{T}=\frac{(K+1)r_T+q_T}{T}.$$

By the assumptions on the rates in the statement of Theorem B.2, we obtain (36) from (37). This completes the proof. □

We next show the following result.

B.1.2 Preliminary results on variance approximation

Lemma B.6. *Let the assumptions of Theorem B.2 hold. Moreover, suppose that* $\beta_{\text{mixing}}(i)\lesssim\exp(-b_1i^{b_2})$. *Then*

$$\max_{1\leq j\leq\mathcal{N}}|T^{-1}\sum_{t=1}^{T}(W_{j,t}-E(W_{j,t}))|=O_P\left(D_TT^{-1/2}\left(\log\mathcal{N}+(\log T)^{1/(2b_2)}\right)\right).$$

Proof. We apply Bernstein's blocking technique with the same block structure as A_k and $\bar{A}_k$, but the choice of K and q_T is only specific to the proof of this lemma. Let $R_{k,j}=\sum_{t\in A_k}(W_{j,t}-E(W_{j,t}))$. We choose r_T such that $K\beta_{\text{mixing}}(r_T)=o(1)$. This means that $Kq_T\exp(-b_1r_T^{b_2})=o(1)$.

By Berbee's coupling and Lemma 8 of Chernozhukov et al. (2015), it follows that

$$\max_{1\leq j\leq\mathcal{N}}\left|\sum_{k=1}^{K}R_{k,j}\right|$$
$$=O_P\left(\max_{1\leq j\leq\mathcal{N},\ 1\leq k\leq K}|R_{k,j}|\sqrt{\log\mathcal{N}}+\sqrt{\max_{1\leq j\leq\mathcal{N}}\sum_{k=1}^{K}ER_{k,j}^2}\times\log\mathcal{N}\right)$$

$$\leq O_P\left(\max_{1\leq k\leq K}|A_k|\max_{1\leq j\leq\mathcal{N},\ 1\leq t\leq T}|W_{j,t}-E(W_{j,t})|\sqrt{\log\mathcal{N}}+\sqrt{\max_{1\leq j\leq\mathcal{N}}\sum_{k=1}^{K}ER_{k,j}^2}\times\log\mathcal{N}\right)$$
$$\overset{\text{(i)}}{=} O_P\left(D_Tq_T\sqrt{\log\mathcal{N}}+D_T\sqrt{Kq_T}\log\mathcal{N}\right),$$

where (i) follows from the definition of A_k and $ER_{k,j}^2\lesssim D_T^2q_T$ (due to Lemma 7.2 of Chen et al. (2016)). Moreover,

$$\max_{1\leq j\leq\mathcal{N}}\left|\sum_{t=1}^{T}(W_{j,t}-E(W_{j,t}))-\sum_{k=1}^{K}R_{k,j}\right|=\max_{1\leq j\leq\mathcal{N}}\left|\sum_{k=1}^{K}\sum_{t\in\bar{A}_k/A_k}(W_{j,t}-E(W_{j,t}))\right|$$
$$\leq 2D_T\sum_{k=1}^{K}|\bar{A}_k/A_k|\leq 2D_T(Kr_T+q_T).$$

Therefore,

$$\max_{1\leq j\leq\mathcal{N}}\left|\sum_{t=1}^{T}(W_{j,t}-E(W_{j,t}))\right|=O_P\left(D_Tq_T\sqrt{\log\mathcal{N}}+D_T\sqrt{Kq_T}\log\mathcal{N}+D_TKr_T\right).$$

Using $K\asymp T/(q_T+r_T)$, we choose $q_T\asymp\sqrt{Tr_T/\sqrt{\log\mathcal{N}}}$ to get

$$\max_{1\leq j\leq\mathcal{N}}\left|\sum_{t=1}^{T}(W_{j,t}-E(W_{j,t}))\right|=O_P\left(D_T\sqrt{T}\log\mathcal{N}+D_T\sqrt{Tr_T}\right).$$

Now the requirement of $Kq_T\exp(-b_1r_T^{b_2})=o(1)$ implies that we can choose $r_T\asymp(\log T)^{1/b_2}$, which means that

$$\max_{1\leq j\leq\mathcal{N}}\left|\sum_{t=1}^{T}(W_{j,t}-E(W_{j,t}))\right|=O_P\left(D_T\sqrt{T}\left(\log\mathcal{N}+(\log T)^{1/(2b_2)}\right)\right).$$

The desired result follows from this. □

Lemma B.7. *Let the assumptions of Theorem B.2 hold. Moreover, suppose that* $K\beta_{\text{mixing}}(q_T+r_T)=o(1)$. *Then*

$$\max_{1\leq j\leq\mathcal{N}}\left|T^{-1}\sum_{k=1}^{K}\left[\left(\sum_{t\in\bar{A}_k}(W_{j,t}-\mu_j)\right)^2-E\left(\sum_{t\in\bar{A}_k}(W_{j,t}-\mu_j)\right)^2\right]\right|$$
$$=O_P\left(\sqrt{T^{-1}(q_T+r_T)D_T^4\log\mathcal{N}}+T^{-1}(q_T+r_T)^2D_T^2\log\mathcal{N}\right).$$

Proof. For simplicity, we assume that K is an even number and denote $L_T = K/2$. Since $K\beta_{\text{mixing}}(q_T + r_T) = o(1)$, we can use Berbee's coupling and Lemma 8 of Chernozhukov et al. (2015), obtaining

$$\begin{aligned}
&\max_{1\le j\le \mathcal{N}} \left| \sum_{l=1}^{L_T} \left[\left(\sum_{t\in \bar{A}_{2l-1}} (W_{j,t} - \mu_j) \right)^2 - E\left(\sum_{t\in \bar{A}_{2l-1}} (W_{j,t} - \mu_j) \right)^2 \right] \right| \\
&= O_P\left(\sqrt{\sum_{l=1}^{L_T} E\left(\sum_{t\in \bar{A}_{2l-1}} (W_{j,t} - \mu_j) \right)^4 \log \mathcal{N}} + \max_{1\le l\le L_T,\ 1\le j\le \mathcal{N}} \left(\sum_{t\in \bar{A}_{2l-1}} (W_{j,t} - \mu_j) \right)^2 \times \log \mathcal{N} \right) \\
&\overset{\text{(i)}}{=} O_P\left(\sqrt{\sum_{l=1}^{L_T} |\bar{A}_{2l-1}|^2 (2D_T)^4 \times \log \mathcal{N}} + \max_{1\le l\le L_T} |\bar{A}_{2l-1}|^2 (2D_T)^2 \times \log \mathcal{N} \right) \\
&= O_P\left(\sqrt{T(q_T + r_T) D_T^4 \log \mathcal{N}} + (q_T + r_T)^2 D_T^2 \log \mathcal{N} \right),
\end{aligned}$$

where (i) follows by $E\left(\sum_{t\in \bar{A}_{2l-1}} (W_{j,t} - \mu_j)\right)^4 \lesssim |\bar{A}_{2l-1}|^2 (2D_T)^4$ (due to Lemma 7.2 of Chen et al. (2016)). Similarly, we can show

$$\begin{aligned}
&\max_{1\le j\le \mathcal{N}} \left| \sum_{l=1}^{L_T} \left[\left(\sum_{t\in \bar{A}_{2l}} (W_{j,t} - \mu_j) \right)^2 - E\left(\sum_{t\in \bar{A}_{2l}} (W_{j,t} - \mu_j) \right)^2 \right] \right| \\
&\qquad = O_P\left(\sqrt{T(q_T + r_T) D_T^4 \log \mathcal{N}} + (q_T + r_T)^2 D_T^2 \log \mathcal{N} \right).
\end{aligned}$$

The desired result follows from this. □

B.2 Proof of Theorem 3.1

We apply Theorem B.2. We separate $K_T = q_T + r_T$ with $q_T > r_T$. Specifically, we choose $r_T = \kappa_1 (\log T)^{1/b_2}$ with $\kappa_1 = (2/b_1)^{1/b_2}$ and $q_T = K_T - r_T$. It suffices to show that the conditions of Theorem B.2 can be satisfied by this choice of (q_T, r_T).

Since $K \asymp T/(q_T + r_T)$ and $(r_T/q_T)\log^2 \mathcal{N} = o(1)$, we have $T^{-1}\sqrt{K} r_T (\log \mathcal{N})^{3/2} = o(1)$ and $T^{-1} K r_T \log^2 \mathcal{N} = o(1)$. By $q_T D_T \log^{5/2}(\mathcal{N}T) \le C_1 T^{1/2 - c_2}$ and $r_T < q_T$, we have $T^{-1} q_T (\log \mathcal{N})^{3/2} = o(1)$ and $T^{-1} r_T^2 D_T^2 \log^3 \mathcal{N} = o(1)$. It remains to show that Assumption 1 is satisfied by this choice of (q_T, r_T). In particular, with $\mathcal{N} = N$, we need to verify the following

$$\max\{K\beta_{\text{mixing}}(r_T),\ (r_T/q_T)\log^2 \mathcal{N}\} \le C_1 T^{-c_2} \tag{38}$$

and

$$q_T D_T \log^{5/2}(\mathcal{N}T) \leq C_1 T^{1/2-c_2} \tag{39}$$

for some $0 < c_2 < 1/4$.

Since $\beta_{\text{mixing}}(r_T) \lesssim \exp(-b_1 r_T^{b_2})$, $K \leq T$, it follows that $K\beta_{\text{mixing}}(r_T) \lesssim T\exp(-b_1 r_T^{b_2})$. Hence, we have $K\beta_{\text{mixing}}(r_T) \lesssim T^{-1}$.

Since $K \asymp T/(q_T + r_T)$ and $q_T > r_T$, we have $q_T \asymp T/K$. Therefore,

$$(r_T/q_T)\log^2 \mathcal{N} \lesssim (\log T)^{1/b_2} K T^{-1} \log^2 \mathcal{N}.$$

By Assumption 1, $KT^{-1}\log^2 \mathcal{N} \lesssim T^{-b}$ for some $b \in (0, 1/4)$. Thus, we only need to choose $c_2 = b/2$ to obtain $(r_T/q_T)\log^2 \mathcal{N} \lesssim T^{-c_2}$. This proves (38).

By $q_T \asymp T/K$ and Assumption 1, we have

$$q_T D_T \log^{5/2}(\mathcal{N}T) \asymp K^{-1} T D_T \log^{5/2}(\mathcal{N}T) \lesssim T^{1/2-b} \lesssim T^{1/2-c_2},$$

which proves (39). Now we have verified all the conditions of Theorem B.2, which implies that

$$\sup_{x\in\mathbb{R}} \left| P\left(\max_{1\leq j\leq \mathcal{N}} \frac{T^{-1/2}\sum_{t=1}^T W_{jt}}{\hat{a}_j} \leq x \right) - P\left(\max_{1\leq j\leq \mathcal{N}} \frac{T^{-1/2}\sum_{k=1}^K \sum_{t\in \bar{A}_k} (W_{j,t} - \hat{\mu}_j)\varepsilon_k}{\hat{a}_j} \leq x \mid \{W_s\}_{s=1}^T \right) \right| = o_P(1).$$

This means that

$$\lim_{T\to\infty} P\left(\max_{1\leq j\leq \mathcal{N}} \frac{T^{-1/2}\sum_{t=1}^T (U_{j,t} - EU_{j,t})}{\hat{a}_j} > \tilde{Q}^*_{T,1-\alpha} \right) = \alpha.$$

Under the null hypothesis of $EU_{j,t} \leq 0$, we have that

$$\max_{1\leq j\leq \mathcal{N}} \frac{T^{-1/2}\sum_{t=1}^T (U_{j,t} - EU_{j,t})}{\hat{a}_j} \geq \max_{1\leq j\leq \mathcal{N}} \frac{T^{-1/2}\sum_{t=1}^T U_{j,t}}{\hat{a}_j} = \tilde{R}_T.$$

In turn, this means that

$$\limsup_{T\to\infty} P\left(\tilde{R}_T > \tilde{Q}^*_{T,1-\alpha} \right) \leq \alpha.$$

When $EU_{j,t} = 0$, the inequality in the above two equation displays hold with equality. This completes the proof.

B.3 Proof of Lemma 3.1

We consider two cases.

Case 1: $\hat{a}_j = 1$.

We only need to take $a_j = 1$. Then $\hat{a}_j - a_j = 1$ and the result clearly holds.

Case 2: $\hat{a}_j = \sqrt{K^{-1}\sum_{j=1}^{K}\left(B_T^{-1/2}\sum_{t\in H_j}(\Delta L_{i,t+h} - \hat{\mu}_i)\right)^2}$.

We inherit all the notations from before. Recall $\hat{\mu}_j = T^{-1}\sum_{t=1}^{T} W_{j,t}$. Let $a_j^2 = T^{-1}\sum_{k=1}^{K} E\left(\sum_{t\in\bar{A}_k}(W_{j,t} - \mu_j)\right)^2$ with $\mu_j = T^{-1}\sum_{t=1}^{T} E(W_{j,t})$. Let $\bar{W}_{j,t} = W_{j,t} - \mu_j$ and $\bar{a}_j^2 = T^{-1}\sum_{k=1}^{K}\left(\sum_{t\in\bar{A}_k}\bar{W}_{j,t}\right)^2$. Clearly, $\hat{\mu}_j - \mu_j = T^{-1}\sum_{t=1}^{T}\bar{W}_{j,t}$.

Notice that by triangular inequality for the Euclidean norm in $\mathbb{R}^K$, we have

$$\begin{aligned}
\max_{1\le j\le\mathcal{N}}|\hat{a}_j - \bar{a}_j| &= T^{-1/2}\max_{1\le j\le\mathcal{N}}\left|\sqrt{\sum_{k=1}^{K}\left(\sum_{t\in\bar{A}_k}\bar{W}_{j,t} - |\bar{A}_k|(\hat{\mu}_j - \mu_j)\right)^2} - \sqrt{\sum_{k=1}^{K}\left(\sum_{t\in\bar{A}_k}\bar{W}_{j,t}\right)^2}\right| \\
&\le T^{-1/2}\max_{1\le j\le\mathcal{N}}\sqrt{\sum_{k=1}^{K}|\bar{A}_k|^2(\hat{\mu}_j - \mu_j)^2} \\
&\le T^{-1/2}\|\hat{\mu} - \mu\|_\infty \max_{1\le k\le K}|\bar{A}_k|\sqrt{K} \\
&\overset{\text{(i)}}{=} O_P\left(D_T T^{-1}\left(\log\mathcal{N} + (\log T)^{1/(2b_2)}\right)\times(q_T + r_T)\sqrt{K}\right) \\
&= O_P\left(D_T K^{-1/2}\left(\log\mathcal{N} + (\log T)^{1/(2b_2)}\right)\right),
\end{aligned}$$

where (i) follows from Lemma B.6. On the other hand, Lemma B.7 implies that

$$\max_{1\le j\le\mathcal{N}}|\bar{a}_j^2 - a_j^2| = O_P\left(\sqrt{T^{-1}(q_T + r_T)D_T^4\log\mathcal{N}} + T^{-1}(q_T + r_T)^2 D_T^2\log\mathcal{N}\right).$$

Notice that the rate conditions in the assumption imply that the rate in the above two displays are $o_P(1/\log\mathcal{N})$. Since $\min_{1\le j\le\mathcal{N}} a_j$ is bounded away from zero, we have $\max_{1\le j\le\mathcal{N}}|\hat{a}_j - a_j| = o_P(1/\log\mathcal{N})$.

The proof for the other cases follows by similar arguments as for Case 2.

Table A1: Finite-sample size of Sup tests computed across multiple variables, forecasters, and time-periods

	$\alpha = 0.05$								$\alpha = 0.1$							
	Without studentization				With studentization				Without studentization				With studentization			
	$M = 2$				$M = 2$				$M = 2$				$M = 2$			
$N \setminus T$	25	50	100	200	25	50	100	200	$n \setminus T$	25	50	100	200	25	50	100	200
1	0.063	0.060	0.050	0.057	0.058	0.063	0.049	0.057	1	0.117	0.135	0.111	0.113	0.117	0.131	0.116	0.117
10	0.054	0.050	0.049	0.052	0.059	0.047	0.027	0.023	10	0.109	0.112	0.105	0.108	0.126	0.113	0.077	0.081
25	0.053	0.048	0.045	0.056	0.073	0.033	0.026	0.014	25	0.115	0.112	0.093	0.112	0.141	0.098	0.076	0.073
50	0.039	0.047	0.040	0.048	0.052	0.022	0.022	0.020	50	0.086	0.117	0.097	0.100	0.122	0.087	0.054	0.059
100	0.033	0.036	0.040	0.039	0.082	0.027	0.011	0.015	100	0.087	0.099	0.109	0.086	0.148	0.074	0.058	0.045
	$M = 10$				$M = 10$				$M = 10$				$M = 10$			
$N \setminus T$	25	50	100	200	25	50	100	200	$n \setminus T$	25	50	100	200	25	50	100	200
1	0.049	0.063	0.059	0.049	0.057	0.044	0.040	0.032	1	0.121	0.158	0.143	0.119	0.113	0.134	0.113	0.088
10	0.068	0.053	0.043	0.047	0.077	0.034	0.021	0.021	10	0.134	0.143	0.121	0.104	0.155	0.101	0.075	0.068
25	0.057	0.067	0.045	0.041	0.087	0.019	0.009	0.008	25	0.127	0.155	0.123	0.130	0.170	0.083	0.043	0.049
50	0.042	0.047	0.056	0.035	0.099	0.020	0.006	0.007	50	0.105	0.135	0.123	0.095	0.196	0.072	0.044	0.033
100	0.036	0.033	0.019	0.026	0.121	0.025	0.004	0.003	100	0.104	0.100	0.077	0.070	0.231	0.079	0.030	0.031
	$M = 100$				$M = 100$				$M = 100$				$M = 100$			
$N \setminus T$	25	50	100	200	25	50	100	200	$n \setminus T$	25	50	100	200	25	50	100	200
1	0.053	0.072	0.051	0.047	0.075	0.039	0.021	0.020	1	0.150	0.165	0.137	0.128	0.135	0.114	0.084	0.070
10	0.059	0.062	0.047	0.036	0.114	0.023	0.005	0.004	10	0.122	0.165	0.131	0.117	0.227	0.098	0.051	0.036
25	0.042	0.034	0.042	0.038	0.130	0.020	0.007	0.002	25	0.113	0.120	0.126	0.100	0.283	0.083	0.033	0.023
50	0.039	0.030	0.024	0.028	0.179	0.016	0.002	0.002	50	0.102	0.129	0.086	0.086	0.369	0.081	0.025	0.021
100	0.031	0.023	0.016	0.022	0.237	0.007	0.001	0.002	100	0.067	0.088	0.070	0.063	0.430	0.063	0.013	0.011

Notes: This table presents the size of Sup tests comparing the finite-sample accuracy of a set of benchmark forecasts m_0 to a set of alternative forecasts, m_1. All numbers are based on 2,000 Monte Carlo simulations conducted under the null of equal predictive accuracy of the forecasts in m_0 and m_1. N denotes the number of variables; M refers to the number of forecasters, while T denotes the number of time-series observations. The two panels on the left present results set the asymptotic size of the test to $\alpha = 0.05$ while the two panels on the right set the asymptotic size of the test to $\alpha = 0.10$. The Monte Carlo simulations generate the forecast errors as $e_{i,t+h,m} = \lambda_{i,m} f_{t+h} + u_{i,t+h,m}$, where f_t is a mean-zero Gaussian AR(1) process with autoregressive coefficient ρ and variance σ_f^2 . We generate $\lambda_{i,m}$ as i.i.d random variables from a $N(0, \sigma_\lambda^2)$ distribution and truncated such that $\lambda_{i,m}^2 \sigma_f^2 \leq 0.9$; we then set $u_{i,t+h,m}$ as a mean-zero Gaussian AR(1) process with AR coefficient ρ and variance $1 - \lambda_{i,m}^2 \sigma_f^2$. $\{f_{t+h}\}_{t+h=1}^{T}$, $\{\lambda_{i,m}\}_{1 \leq i \leq n, 1 \leq m \leq M}$ and $\{u_{i,t+h,m}\}_{1 \leq i \leq n, 1 \leq m \leq M, 1 \leq t+h \leq T}$ are assumed to be mutually independent. We choose $(\sigma_f, \sigma_\lambda) = (2, 1.2)$. When $T > 30$, we use $\rho = 0.5$ and a block size of $BT = T^{0.6}$; otherwise we use $\rho = 0$ and $B_T = 1$. We consider two studentization schemes: no studentization (the first and third panels) and (partial) studentization (the second and fourth panel), which are both described in Example 1. Under this scheme, all forecast errors have an MSE equal to one and thus the null hypothesis that no forecasts underperforms the baseline model holds.

Table A2: Size-adjusted critical values for the Sup test

	$\alpha = 0.05$								$\alpha = 0.1$							
	Without studentization				With studentization				Without studentization				With studentization			
	$M = 2$				$M = 2$				$M = 2$				$M = 2$			
$N \setminus T$	25	50	100	200	25	50	100	200	$n \setminus T$	25	50	100	200	25	50	100	200
1	0.040	0.048	0.052	0.044	0.048	0.040	0.052	0.044	1	0.084	0.076	0.084	0.088	0.076	0.080	0.080	0.088
10	0.048	0.052	0.052	0.048	0.040	0.052	0.080	0.080	10	0.096	0.088	0.096	0.092	0.076	0.092	0.124	0.120
25	0.048	0.056	0.056	0.044	0.032	0.064	0.080	0.084	25	0.096	0.092	0.108	0.092	0.068	0.108	0.132	0.128
50	0.068	0.056	0.064	0.052	0.048	0.068	0.092	0.092	50	0.112	0.092	0.104	0.104	0.088	0.112	0.140	0.152
100	0.064	0.060	0.060	0.072	0.032	0.076	0.092	0.108	100	0.112	0.104	0.096	0.112	0.064	0.124	0.136	0.168
	$M = 10$				$M = 10$				$M = 10$				$M = 10$			
$N \setminus T$	25	50	100	200	25	50	100	200	$n \setminus T$	25	50	100	200	25	50	100	200
1	0.052	0.040	0.048	0.052	0.044	0.056	0.056	0.068	1	0.084	0.072	0.076	0.084	0.092	0.080	0.092	0.112
10	0.036	0.048	0.056	0.052	0.036	0.068	0.080	0.080	10	0.076	0.080	0.088	0.096	0.060	0.100	0.120	0.128
25	0.044	0.040	0.056	0.052	0.024	0.080	0.108	0.104	25	0.088	0.072	0.088	0.088	0.060	0.112	0.148	0.156
50	0.056	0.056	0.048	0.064	0.020	0.084	0.108	0.132	50	0.096	0.084	0.088	0.104	0.052	0.116	0.156	0.180
100	0.064	0.068	0.076	0.084	0.016	0.076	0.128	0.140	100	0.100	0.104	0.116	0.124	0.040	0.116	0.172	0.188
	$M = 100$				$M = 100$				$M = 100$				$M = 100$			
$N \setminus T$	25	50	100	200	25	50	100	200	$n \setminus T$	25	50	100	200	25	50	100	200
1	0.048	0.036	0.048	0.052	0.036	0.060	0.076	0.084	1	0.072	0.064	0.080	0.084	0.076	0.092	0.116	0.128
10	0.044	0.044	0.052	0.056	0.024	0.076	0.100	0.120	10	0.088	0.068	0.080	0.092	0.044	0.104	0.140	0.168
25	0.056	0.060	0.056	0.064	0.016	0.076	0.112	0.132	25	0.096	0.092	0.084	0.104	0.040	0.112	0.160	0.184
50	0.060	0.064	0.072	0.068	0.012	0.084	0.128	0.148	50	0.100	0.088	0.112	0.112	0.028	0.116	0.172	0.196
100	0.080	0.076	0.084	0.084	0.008	0.096	0.140	0.176	100	0.128	0.108	0.128	0.132	0.016	0.124	0.176	0.224

Notes: This table presents size-adjusted critical values for Sup tests conducted on finite sample data generated in a Monte Carlo simulation. Here, N denotes the number of variables; M is the number of forecasters, and T is the number of time periods. The two panels on the left present results when the asymptotic size of the test is set to $\alpha = 0.05$ while the two panels on the right present result when the asymptotic size of the test is set to $\alpha = 0.10$. For each value of (N, M, T), we compute the critical value for the p-value such that the rejection probability for this sample size under the null hypothesis is set to equal α. We refer to these critical values as size-adjusted critical values. We consider two studentization schemes: no studentization (first and third panels) and (partial) studentization (second and fourth panels), both being described in Example 1. The forecast errors are generated in the same way as those in table A1.

Table A3: Power of Sup test using size-adjusted critical values

	$\alpha = 0.05$								$\alpha = 0.1$							
	Without studentization				With studentization				Without studentization				With studentization			
	$M = 2$				$M = 2$				$M = 2$				$M = 2$			
$N \setminus T$	25	50	100	200	25	50	100	200	$n \setminus T$	25	50	100	200	25	50	100	200
1	0.050	0.053	0.057	0.035	0.054	0.045	0.064	0.038	1	0.095	0.087	0.105	0.090	0.082	0.083	0.095	0.093
10	0.112	0.129	0.169	0.170	0.351	0.378	0.589	0.702	10	0.218	0.240	0.305	0.304	0.462	0.515	0.682	0.775
25	0.109	0.131	0.121	0.114	0.461	0.576	0.790	0.903	25	0.216	0.208	0.280	0.255	0.609	0.712	0.892	0.950
50	0.140	0.107	0.144	0.124	0.657	0.690	0.884	0.975	50	0.218	0.196	0.225	0.234	0.770	0.834	0.955	0.994
100	0.100	0.105	0.100	0.159	0.621	0.771	0.922	0.993	100	0.197	0.205	0.180	0.244	0.776	0.890	0.976	0.999
	$M = 10$				$M = 10$				$M = 10$				$M = 10$			
$N \setminus T$	25	50	100	200	25	50	100	200	$n \setminus T$	25	50	100	200	25	50	100	200
1	0.473	0.312	0.444	0.610	0.298	0.299	0.384	0.584	1	0.559	0.468	0.551	0.706	0.415	0.393	0.500	0.687
10	0.095	0.112	0.125	0.146	0.637	0.723	0.873	0.978	10	0.182	0.207	0.213	0.274	0.733	0.827	0.945	0.998
25	0.092	0.097	0.128	0.135	0.644	0.829	0.963	0.998	25	0.184	0.204	0.212	0.258	0.818	0.905	0.991	1.000
50	0.097	0.133	0.117	0.159	0.650	0.816	0.967	1.000	50	0.213	0.241	0.218	0.279	0.829	0.904	0.990	1.000
100	0.125	0.150	0.171	0.246	0.645	0.823	0.992	1.000	100	0.241	0.261	0.281	0.365	0.815	0.925	0.999	1.000
	$M = 100$				$M = 100$				$M = 100$				$M = 100$			
$N \setminus T$	25	50	100	200	25	50	100	200	$n \setminus T$	25	50	100	200	25	50	100	200
1	0.721	0.508	0.721	0.857	0.560	0.651	0.822	0.932	1	0.824	0.710	0.864	0.938	0.719	0.763	0.909	0.965
10	0.091	0.089	0.141	0.162	0.668	0.834	0.951	0.998	10	0.227	0.176	0.229	0.296	0.798	0.907	0.982	1.000
25	0.130	0.155	0.144	0.185	0.700	0.843	0.958	0.999	25	0.239	0.252	0.240	0.312	0.843	0.943	0.995	1.000
50	0.158	0.165	0.172	0.186	0.702	0.871	0.986	1.000	50	0.246	0.235	0.319	0.351	0.850	0.928	0.995	1.000
100	0.182	0.155	0.208	0.249	0.701	0.935	0.987	1.000	100	0.319	0.285	0.372	0.443	0.812	0.979	0.998	1.000

Notes: This table reports the finite-sample power of the Sup test conducted across multiple variables, forecasters and time-periods. N denotes the number of variables; M is the number of forecasters and T is the number of time periods. The two panels on the left present results using the 5% size-adjusted critical values from Table A2; the two panels on the right present result using the 10% size-adjusted critical values from Table A2. We consider two studentization schemes: no studentization (first and third panels) and (partial) studentization (second and fourth panels), which are both described in Example 1. To investigate the power properties, we consider the following. Of all the Mn forecasts, N forecasts are assigned to the baseline set m_0 (i.e., each of the N series has one baseline forecasts) while $(N-1)M$ forecasts are assigned to the set of alternatives, m_1. All forecast errors are generated in the same way as the simulation described in Table A1. Then, we randomly select 20% of the competing forecasts and add $(2T^{-1}\log(Mn))^{1/8}$ to their selected forecast errors, which then have larger MSE values than the baseline forecasts.