Demetrescu, Matei; Hanck, Christoph; Kruse-Becher, Robinson

Robust inference under time-varying volatility: A real-time evaluation of professional forecasters

Journal of Applied Econometrics

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Demetrescu, Matei; Hanck, Christoph; Kruse-Becher, Robinson (2022) : Robust inference under time-varying volatility: A real-time evaluation of professional forecasters, Journal of Applied Econometrics, ISSN 1099-1255, Wiley, Hoboken, NJ, Vol. 37, Iss. 5, pp. 1010-1030, https://doi.org/10.1002/jae.2906

Received: 10 June 2016 | Revised: 1 December 2021 | Accepted: 13 December 2021

DOI: 10.1002/jae.2906

WILEY Journal of APPLIED ECONOMETRICS

RESEARCH ARTICLE

Robust inference under time-varying volatility: A real-time evaluation of professional forecasters

Matei Demetrescu[1] | **Christoph Hanck**[2] | **Robinson Kruse-Becher**[3,4]

[1]Department of Statistics, TU Dortmund, Vogelpothsweg 78, Dortmund, Germany

[2]Faculty of Economics and Business Administration, University of Duisburg-Essen, Universitätsstraße 12, Essen, Germany

[3]Faculty of Economics, University of Hagen, Universitätsstraße 41, Hagen, Germany

[4]CREATES, Aarhus University, School of Economics and Management, Fuglesangs Allé 4, Aarhus V, Denmark

Correspondence
Robinson Kruse-Becher, University of Hagen, Faculty of Economics, Universitätsstr. 41, 58097 Hagen, Germany.
Email: robinson.kruse-becher@fernuni-hagen.de

Summary

In many forecast evaluation applications, standard tests as well as tests allowing for time-variation in relative forecast ability build on heteroskedasticity-and-autocorrelation consistent (HAC) covariance estimators. Yet, the finite-sample performance of these asymptotics is often poor. "Fixed-*b*" asymptotics, used to account for long-run variance estimation, improve finite-sample performance under homoskedasticity, but lose asymptotic pivotality under time-varying volatility. Moreover, loss of pivotality due to time-varying volatility is found in the standard HAC framework in certain cases as well. We prove a wild bootstrap implementation to restore asymptotically pivotal inference for the above and new CUSUM- and Cramér-von Mises-based tests in a fairly general setup, allowing for estimation uncertainty from either a rolling window or a recursive approach when fixed-*b* asymptotics are adopted to achieve good finite-sample performance. We then investigate the (time-varying) performance of professional forecasters relative to naive no-change and model-based predictions in real-time. We exploit the Survey of Professional Forecasters (SPF) database and analyze nowcasts and forecasts at different horizons for output and inflation. We find that not accounting for time-varying volatility seriously affects outcomes of tests for equal forecast ability: wild bootstrap inference typically yields convincing evidence for advantages of the SPF, while tests using non-robust critical values provide remarkably less. Moreover, we find significant evidence for time-variation of relative forecast ability, the advantages of the SPF weakening considerably after the "Great Moderation."

KEYWORDS

bootstrap, forecast evaluation, HAC estimation, hypothesis testing, structural breaks

1 | INTRODUCTION

Forecasting plays a crucial role in economics, finance, and many other disciplines. Policy makers, firms, investors, and households have various needs for macroeconomic predictions. Many of those are available, for example, from the IMF and OECD, governmental forecasts like "Teal Book" forecasts from the Federal Reserve, or commercial forecasters (e.g., Blue Chip Economic Indicators, Data Resources Inc., or the Survey of Professional Forecasters [SPF]). The SPF is

the most comprehensive database available to assess the performance of professional forecasters. A fundamental question is then whether SPF forecasts outperform simple (model-based) alternatives, that is, have significantly smaller forecast error loss differentials on average. For example, Zarnowitz and Braun (1993) reveal that SPF forecasts perform well in comparison with standard time series models (see also Croushore, 1993; Stark, 2010). With data from 1969 to 2017, we re-evaluate SPF forecasts for US output growth and GDP deflator inflation using robust inference methods.

This long evaluation period contains subsamples with structural changes mainly due to the "Great Moderation," but also during and after the "Great Financial Crisis." The "Great Moderation" is a period of considerable reduction in macroeconomic volatility as well as of sharp decline in predictability (Campbell, 2007). The "Great Financial Crisis" changed volatility, although to a lesser extent than the "Great Moderation," and yet less is known about its consequences on predictability. Changing macroeconomic volatility and changing predictability have important implications for forecast evaluation tests. While the first feature typically leads to time-varying volatility (in the sense of possibly unconditional heteroskedasticity over time) in forecast error loss differentials, the second might imply an instability of their mean. Ignoring these features may lead to significant size distortions and power losses; see the rich literature on forecasting in unstable environments (e.g., Coroneo & Iacone, 2020; Giacomini & Rossi, 2010; Rossi, 2013).

Here, we discuss the Diebold and Mariano (DM, 1995), fluctuation (Giacomini & Rossi, 2010), and new CUSUM and Cramér-von Mises tests from the perspective of time-variation, in particular time-varying volatility. While the DM test focuses on comparisons in stable environments, the latter three statistics capture time-varying relative forecast performance explicitly. The fluctuation, CUSUM, and Cramér-von Mises statistics are however generally not robust to time-varying volatility, as their limiting null distributions depend on limit processes for partial sums, which do not converge to standard Wiener processes under time-varying volatility (cf. Section 2.2).

Moreover, we conduct the discussion in the "fixed-b" paradigm as pioneered by Kiefer and Vogelsang (2005). This paradigm goes beyond the standard heteroskedasticity and autocorrelation consistent (HAC) framework (see the seminal contributions of Andrews, 1991; Newey & West, 1987), in which, for example, Diebold and Mariano (1995) and Giacomini and Rossi (2010) also derive their limiting distributions for the cited test statistics. HAC permits to use critical values from standard distributions, like the χ^2 or standard normal. These asymptotic distributions, however, turn out to be rather poor approximations to actual finite-sample distributions. Hence, substantial size distortions arise in practice. In particular, test results turn out to be sensitive to the choice of bandwidth B and kernel k employed for long-run variance estimation. The poor performance of HAC's asymptotic approximation can be explained by the "small-b" requirement that a vanishing fraction $b := B/P \to 0$ of the number of observations P be used for estimating autocovariances, while of course $b > 0$ in finite-samples. To tackle this issue, Kiefer et al. (2000) and Kiefer and Vogelsang (2002a, 2002b, 2005) propose "fixed-b" asymptotics, which do not assume that $b \to 0$. This leads to nonstandard distributions (reviewed in Section 2). Conveniently and unlike in the standard small-b HAC framework, the new distributions reflect the choice of B and k even in the limit. The above papers convincingly demonstrate that the new distributions provide, in the absence of time-varying variances, substantially better approximations to actual finite-sample distributions. For these reasons, Choi and Kiefer (2010) advocate the use of Diebold and Mariano (1995) tests with fixed-b critical values; see also Li and Patton (2018). However, fixed-b critical values rely too on asymptotics for partial sums, which are affected by time-varying volatility, such that the fixed-b based Diebold and Mariano (1995) test then lacks pivotality, too.

Our main theoretical contribution is then to develop time-varying volatility-robust wild bootstrap versions of DM, fluctuation (Giacomini & Rossi, 2010), and the new CUSUM and Cramér-von Mises statistics under the fixed-b paradigm. We allow for parameter estimation error (West, 1996) in estimated non-nested forecast models (such that one may, as we shall occasionally do, also refer to the DM statistic as a Diebold-Mariano-West statistic) and cover both rolling window and recursive estimation for a fairly general nonlinear GMM setup.

In more detail, Section 2 rigorously shows time-varying variances to affect fixed-b limiting distributions of all the above four statistics (discussed in more detail in Section 2.1) and thus to lead to a loss of asymptotic pivotality (see also Müller, 2014, p. 314). This actually emphasizes a strength of the fixed-b approach, as it implies that the variability of the variances—influencing finite-sample behavior—is reflected in the limiting distribution. It does, however, come at the cost of yet different critical values. Such time-varying variances are pervasive in applied work in general and in our empirical application in Section 3 specifically.[1]

[1]Indeed, Groen et al. (2013) find variance changes to be important for inflation forecasting. More generally, time-varying volatility is present in many macroeconomic (e.g., Clark & Ravazzolo, 2015; Justiniano & Primiceri, 2008; Sensier & van Dijk, 2004; Stock & Watson, 2002) and financial (e.g., Amado & Teräsvirta, 2013; Guidolin & Timmermann, 2006; Rapach & Strauss, 2008) series such as economic growth, inflation, and excess returns.

Adopting the parameter estimation framework of West (1996) (see Section 2.2), we characterize the resulting additional terms affecting the fixed-b distribution of the discussed tests for a class of generic nonlinear GMM estimators. We then develop a wild bootstrap correction (Section 2.3) replicating these features of the asymptotic distribution and establish its asymptotic validity. An appendix provides numerical results indicating considerable size distortions, due to time-varying volatility, resulting from using the non-bootstrapped conventional asymptotic critical values even in the limit. At the same time, the proposed bootstrap is shown to work well.

Section 3 compares the predictive ability of SPF forecasts for output and inflation to no-change and model-based approaches based on rolling window and recursive estimation. We focus on nowcasts and one-quarter and 1-year ahead forecasts and evaluate these by considering the first and the final release of data. Overall, we find forecast error loss differentials to exhibit substantial heteroskedasticity. This has a direct impact on test decisions when comparing outcomes of traditional and our new robust tests: While the bootstrap provides strong evidence for the superiority of SPF forecasts (especially for nowcasts), there are notably fewer and weaker rejections when using asymptotic critical values. Our findings strongly suggest that SPF forecasts perform better early in the sample, but also that this advantage shrank considerably in the 1980s, leading to equal predictive ability starting in the mid-1980s. There are some signs of recoveries of forecast superiority around 2000 for GDP deflator inflation. We discuss our findings in relation to the literature on SPF accuracy, in general as well as with emphasis on the loss in relative predictability related to the "Great Moderation."

In recent related work, Coroneo and Iacone (2020) study the use of the full-sample Diebold and Mariano (1995) statistic $\mathcal{T}^{DM}$ for unconditional predictive ability testing. They adopt the framework of Giacomini and White (2006); that is, they work with observed loss differentials—estimated from rolling forecasts—directly and hence do not explicitly model effects of parameter estimation in the limiting distributions as we do in our nonlinear GMM setup. Next to an application of fixed-b inference using the Bartlett kernel, Coroneo and Iacone (2020) use an alternative weighted periodogram estimate of the long-run variance with associated "fixed-m" asymptotics to improve the finite-sample performance of $\mathcal{T}^{DM}$. Additionally, they compare the effectiveness of these testing approaches to a stationary block bootstrap (Politis & Romano, 1994). Their fixed-b and fixed-m approaches rule out time-varying volatility.[2] Under time-varying volatility, as is also present in, for example, their empirical applications to the SPF, Coroneo and Iacone (2020) suggest to split the sample into subsamples for which an assumption of constant variance is more credible and hence would allow for the use of standard fixed-b or fixed-m asymptotics. Sometimes, economic considerations (e.g., the "Great Moderation") may provide useful guidance about suitable splits of the whole sample. However, there are several problems with ad hoc choices regarding selected sample splits. These issues touch upon the unknown existence, number and locations of break points see, for example, Rossi and Sekhposyan (2016). Our proposed tests do not require the researcher to possess such knowledge. Section 4 concludes. A series of appendices collects proofs (unless indicated otherwise in the main text), other derivations, simulation results and further empirical results.

2 | FIXED-b INFERENCE UNDER TIME-VARYING VOLATILITY

2.1 | Hypotheses and tests

We test the null of equal predictive ability of two competing forecasts for a target series z_t, either generated by models or obtained from surveys. We shall not assume a specific loss function but work with generic loss differentials directly (Diebold & Mariano, 1995),

$$y_t = \mathcal{L}_t\left(z_{t+h}, f_{1,t}\right) - \mathcal{L}_t\left(z_{t+h}, f_{2,t}\right). \tag{1}$$

Here, $f_{i,t}$, $i = 1, 2$, denote the competing h-step ahead forecasts for time $t + h$ and $\mathcal{L}_t(u_1, u_2)$ the loss function relevant at time t for horizon h. Typically, one focuses on one horizon h at a time, and we, therefore avoid any explicit dependence of $f_{i,t}$ and $\mathcal{L}_t$ on h in the following.

The forecasts $f_{i,t}$ depend on various predictors (including, e.g., z_t and lags of z_t) in the model-based case, gathered in the vector $\boldsymbol{x}_{i,t}$, and on parameters of a model, say $\boldsymbol{\theta}_i \in \mathbb{R}^{M_i}$. Sometimes, $\boldsymbol{\theta}_i$ is known, and we write $f_{i,t} = f_i\left(\boldsymbol{x}_{i,t}, \boldsymbol{\theta}_i\right)$ as "ideal forecasts."[3] In practice, however, parameters of forecast models are typically unknown, and one uses $\hat{f}^{\mathrm{r}}_{i,t} = f_i\left(\boldsymbol{x}_{i,t}, \hat{\boldsymbol{\theta}}^{\mathrm{r}}_{i,t}\right)$.

[2]The sampling properties of the periodogram also depend on time-varying volatility (see, e.g., Demetrescu & Sibbertsen, 2016).

[3]This includes cases such as driftless random-walk forecasts that do not require parameter estimation.

The notation $\hat{\theta}^{\mathrm{r}}_{i,t}$ emphasizes that one can update the estimators over time, either in a rolling ($\mathrm{r} = rol$) or a recursive ($\mathrm{r} = rec$) fashion.

Time-variation in the loss differentials (1) may arise for a variety of reasons. The most obvious are time-varying features in the series z_{t+h} and the forecasts $f_{i,t}$, but changes in the loss function (such as different weights attached to forecast errors at different times) may also play a role. Less apparent but potentially no less important is the effect of parameter estimation, $\hat{f}^{\mathrm{r}}_{i,t} - f_{i,t}$; see Section 2.2.

We focus on tests of unconditional (cf. Remark 6 for alternative cases) equal predictive accuracy for all t. Hence, the null of interest is that of a zero loss differential at each point in time (Giacomini & Rossi, 2010)

$$H_0 : E(y_t) \equiv \mu_t = 0 \forall t,$$

extending the pair of hypotheses of "average" equal versus unequal predictive ability as pioneered by Diebold and Mariano (1995). One may also consider one-sided alternatives (cf., e.g., Remark 3). Imposing constancy of μ_t has important consequences: As pointed out by Giacomini and Rossi (2010), one can expect some loss of power and reduced interpretability of rejections by tests based on falsely assuming a (time-)homogenous alternative. We follow the seminal work of Giacomini and Rossi (2010) and allow for time-variation in μ_t under the alternative (e.g., as a consequence of forecast breakdowns or other forms of structural instabilities in the relative predictive performance).

To accommodate parameter estimation, we follow closely the setup pioneered by West (1996). There are R preliminary observations used to obtain estimates $\hat{\theta}_{1,R}$ and $\hat{\theta}_{2,R}$. These are used to set up the forecasts $\hat{f}_{1,R}$ and $\hat{f}_{2,R}$, which are compared with z_{R+h}. Then, for the rolling window approach, one estimates the parameters using observations $t = 2, \ldots, R{+}1$ (resulting in $\hat{\theta}^{rol}_{i,R+1}$), while the estimation sample is expanded by one observation for the recursive approach (resulting in $\hat{\theta}^{rec}_{i,R+1}$). The forecast comparison is then conducted for $t = R$, until $t = R + P - 1$. Here, P denotes the number of out-of-sample observations, $z_{R+h}, \ldots, z_{R+P-1+h}$, which are available for forecast comparison together with $\hat{f}_{i,R}, \ldots, \hat{f}_{i,R+P-1}$. According to West (1996), R and P should go to infinity jointly, with $P/R \to \pi > 0$ to ensure that the estimation effect is reflected in the asymptotics.[4] To fix ideas, we focus on the class of (possibly overidentified) GMM estimators with at least as many moment conditions N_i as parameters M_i. Like in West (1996), pseudo-true values $\boldsymbol{\theta}_i$ are taken to exist, such that, as the sample size grows, one may write $\hat{\theta}^{\mathrm{r}}_{i,t} \xrightarrow{p} \boldsymbol{\theta}_i \; \forall t \geq R$, for $\mathrm{r} \in \{rol, rec\}$. Section 2.2 states precise assumptions on the estimators. The observed forecast losses are then given by $\mathcal{L}_t\left(z_{t+h}, \hat{f}^{\mathrm{r}}_{i,t}\right) \equiv \mathcal{L}_t\left(z_{t+h}, f_i\left(\boldsymbol{x}_{i,t}, \hat{\boldsymbol{\theta}}^{\mathrm{r}}_{i,t}\right)\right)$; so one uses

$$\hat{y}^{\mathrm{r}}_t = \mathcal{L}_t\left(z_{t+h}, \hat{f}^{\mathrm{r}}_{1,t}\right) - \mathcal{L}_t\left(z_{t+h}, \hat{f}^{\mathrm{r}}_{2,t}\right), \qquad t = R, \ldots, R + P - 1, \tag{2}$$

for testing rather than the unobserved y_t.

Testing the null restriction $E(y_t) = 0$ under the assumption of (time-)homogeneity may be done via a Diebold-Mariano-West Wald-type statistic building on $\hat{y}^{\mathrm{r}}_t$ (Diebold & Mariano, 1995; West, 1996). Concretely, let

$$\mathcal{T}^{DM} = \frac{1}{P} \frac{\left(\sum_{t=R}^{R+P-1} \hat{y}^{\mathrm{r}}_t\right)^2}{\hat{\Omega}}, \tag{3}$$

where $\hat{\Omega}$ is a suitable estimator of the relevant long-run variance. Estimation of $\hat{\Omega}$ is discussed in more detail below. Considering heterogeneity, the first method used here to test $\mu_t = 0$ against $\mu_t \neq 0$ without imposing constant expectations is the fluctuations test of Giacomini and Rossi (2010). With $\hat{\Omega}$ based on all P pseudo out-of-sample observations available,[5] consider

[4] By considering the contribution of estimation uncertainty, our framework therefore focuses on (adopting the taxonomy of Giacomini & Rossi, 2010) comparing forecasting *models* (and, in so doing, on non-nested models) rather than comparing forecasting *methods*, as in, e.g., Giacomini and White (2006), where the losses depend on parameters estimated in sample using so-called limited-memory estimators.

[5] We hence follow Giacomini and Rossi (2010) and focus on a full-sample estimate of the long-run variance. In a time-varying framework like the present one, it is, following a suggestion of a referee, natural to also study time-varying estimates $\hat{\Omega}_t$ of the long-run variance. We investigate this option in our Monte-Carlo study, but find full-sample estimates to typically perform better, at least in the experiments considered there.

$$\mathcal{T}^F = \max_{t\in\{\lfloor S/2\rfloor+R,\dots,P+R-\lfloor S/2\rfloor\}} \left| \frac{1}{\sqrt{S\hat{\Omega}}} \sum_{j=t-\lfloor S/2\rfloor}^{t+\lfloor S/2\rfloor-1} \hat{y}_j^{\mathrm{r}} \right|, \; S = \lfloor \nu P \rfloor \;\text{ with }\; \nu \in (0,1). \tag{4}$$

We consider two additional statistics to deal with time-varying relative predictive ability, namely, a CUSUM-type and a Cramér-von Mises functional.[6] The CUSUM-type statistic is directly based on the partial sums of $\hat{y}_t^{\mathrm{r}}$,[7]

$$\mathcal{T}^Q = \max_{R\le t\le R+P-1} \sqrt{\frac{S_t^2}{\hat{\Omega}P}} \qquad \text{with} \qquad S_t = \sum_{j=R}^{t} \hat{y}_j^{\mathrm{r}}. \tag{5}$$

The Cramér-von Mises statistic is given by

$$\mathcal{T}^C = \frac{1}{P^2} \sum_{t=R}^{R+P-1} \frac{S_t^2}{\hat{\Omega}}. \tag{6}$$

Standard regularity conditions assumed, the small-b limiting distribution of $\mathcal{T}^x$, $x \in \{DM, F, Q, C\}$ are known under unconditional homoskedasticity, and can be obtained as particular cases of Proposition 2.2, which deals with the encompassing case of time-varying volatility.

Let us now take a closer look at the long-run variance estimator. Given suitable choices for the kernel k and the bandwidth $B = \lfloor bP \rfloor$(see Andrews, 1991; Newey & West, 1987),

$$\hat{\Omega} = \hat{\gamma}_0 + 2\sum_{j=1}^{P-1} k\left(j/B\right)\hat{\gamma}_j \tag{7}$$

is a long-run variance estimator with $\hat{\gamma}_j = P^{-1}\sum_{t=|j|+R}^{R+P-1}\left(y_t-\bar{y}\right)\left(y_{t-|j|}-\bar{y}\right)$. Regularity conditions assumed, $\hat{\Omega}$ is consistent for the long-run variance of y_t. Whenever y_t is unobserved, one computes $\hat{\Omega}$ based on $\hat{y}_t^{\mathrm{r}}$. However, West (1996) shows that, when parameters need to be estimated, the resulting long-run variance estimator does not standardize the partial sums of $\hat{y}_t^{\mathrm{r}}$ correctly in general. See theorem 4.1 of West (1996), which also indicates how to explicitly correct the long-run variance estimator. Yet, we shall not require West's *explicit* correction here, since the wild bootstrap we use to deal with time-varying volatility in the fixed-b framework (see Section 2.3, and in particular Step 4 of Algorithm 1) *implicitly* correctly replicates the behavior of the test statistics in the limit by constructing bootstrap samples in such a way that they do capture the effect of estimation error.

Although (cf. Remark 1) the small-b asymptotic distributions of the above statistics do not depend on k and b,[8] Kiefer and Vogelsang (2005) argue for $\mathcal{T}^{DM}$ (and this extends to $\mathcal{T}^x, x \in \{F, Q, C\}$) that finite-sample dependence on tuning parameters translates into poor finite-sample behavior. To alleviate this, Choi and Kiefer (2010) resort to fixed-b asymptotics for $\mathcal{T}^{DM}$.

However, fixed-b based limiting distributions are affected by time-varying variances, such that one solution immediately prompts the next problem. Proposition 2.2 contains a formal treatment; see also Demetrescu et al. (2019) and the references therein. To illustrate the main issues with such time-varying variances, consider the case of known parameters and tests based on $\mathcal{T}^{DM}$. To make the dependence of the distribution of $\mathcal{T}^{DM}$ on k and b explicit, Kiefer and Vogelsang (2005) let $b \in (0, 1]$ in the limit. Under homoskedasticity, the resulting limiting distribution is free of nuisance parameters (any scale matrix cancelling out), but is nonstandard. Concretely, Choi and Kiefer (2010) show that

[6]These appear to be more popular in the statistical literature, with prominent econometric exceptions such as the KPSS test for stationarity.

[7]The (perhaps more familiar) CUSUM statistic for a break in mean involves $S_t/t - S_P/P$. This effectively demeans the series, and such a test is rather for a break in relative predictive power. We however test for departures from the null $\mu_t = 0$ rather than μ_t being a constant unknown mean, so centering S_t at zero is the natural choice here.

[8]Since $B = \lfloor bP \rfloor$, we may switch freely between the use of the bandwidth B and the fraction b; however, since b appears in the limit distributions, we use it from now on.

$$\mathcal{T}^{DM} \xrightarrow{d} \mathcal{B}_{k,b} \quad \text{with} \quad \mathcal{B}_{k,b} = W^2(1)/\Lambda_{k,b}(W) \text{ and}$$
$$\Lambda_{k,b}(W) \equiv \begin{cases} -\int_0^1 \int_0^1 \frac{1}{b^2} k'' \left(\frac{r-s}{b}\right) \bar{W}(r)\bar{W}(s)\mathrm{d}r\mathrm{d}s & \text{for } k \text{ differentiable twice} \\ \frac{2}{b}\left(\int_0^1 \bar{W}(r)^2\mathrm{d}r - \int_0^{1-b} \bar{W}(r+b)\bar{W}(r)\mathrm{d}r\right) & \text{for the Bartlett kernel,} \end{cases} \tag{8}$$

where $\bar{W}(s) \equiv W(s) - sW(1)$ with $W(s)$ a standard Wiener process. The distinct feature of fixed-b asymptotics is that $\mathcal{B}_{k,b}$ depends on the *entire* path of the Wiener process $W(s)$ obtained as the limit process of the partial sums of y_t—and not only on $W(1)$, like for small-b. Since time-varying volatility implies a *different* limit for partial-sums processes (see, e.g., Cavaliere, 2004), this has important consequences for fixed-b when the volatility of y_t varies over time. Such dependence of the limiting distributions on the variance pattern extends to the case of estimated parameters and forecast instabilities; see Proposition 2.2.

Remark 1. For $b \to 0$, $\Lambda_{k,b}(W) \xrightarrow{d} 1$ and $\mathcal{B}_{k,b} \xrightarrow{d} \chi^2_1$ (Kiefer & Vogelsang, 2005). In this sense, small-b asymptotics are a particular case of fixed-b asymptotics. Interestingly, $\mathcal{T}^{DM}$ is asymptotically robust under the null to time-varying volatility under small-b asymptotics.[9] Yet, as mentioned above, the finite-sample quality of the HAC-based χ^2-approximation is poor, so the two extant options presented above effectively force practitioners to choose for $\mathcal{T}^{DM}$ between two problems under possible time-varying volatility: either non-pivotal fixed-b distributions, or asymptotically robust small-b distributions with poor finite-sample quality.

2.2 | Assumptions and limiting behavior

This subsection states our maintained assumptions on the DGP and GMM estimation with $N_i \geq M_i$ moment conditions, and provides relevant asymptotic theory.

Assumption 1. Let $\bar{\mathbf{C}}^b_{i,a} \equiv \sum_{j=a}^b \mathbf{C}_{i,j,\boldsymbol{\theta}_i}$. For $t = R, \ldots, R+P-1$ and $\mathrm{r} \in \{rol, rec\}$, let the following decompositions hold:

$$\hat{\boldsymbol{\theta}}^{\mathrm{r}}_{i,t} = \boldsymbol{\theta}_i + \left(\bar{\mathbf{C}}^{t,\prime}_{i,\mathcal{R}} \mathbf{W}_{i,\boldsymbol{\theta}_i} \bar{\mathbf{C}}^{t}_{i,\mathcal{R}}\right)^{-1} \bar{\mathbf{C}}^{t,\prime}_{i,\mathcal{R}} \mathbf{W}_{i,\boldsymbol{\theta}_i} \sum_{j=\mathcal{R}}^{t} \boldsymbol{a}_{i,j,\boldsymbol{\theta}_i} + \boldsymbol{r}^{\mathrm{r}}_{i,t}$$

where $\mathcal{R} = t - R + 1$ for $\mathrm{r} = rol$ and $\mathcal{R} = 1$ for $\mathrm{r} = rec$. Furthermore,

(i) $\sup_{R<t\leq R+P} \left\|\boldsymbol{r}^{\mathrm{r}}_{i,t}\right\| = o_p\left(R^{-1/2}\right)$ as $R, P \to \infty$ with $P/R \to \pi$,
(ii) $\mathbf{W}_{i,\boldsymbol{\theta}_i} > 0$ are deterministic, symmetric full-rank matrices,
(iii) $E(\boldsymbol{a}_{i,t,\boldsymbol{\theta}_i}) = \mathbf{0}$ and
(iv) $\bar{\mathbf{C}}^t_{i,\mathcal{R}}$ are full-rank with probability approaching unity as specified in Assumption 4.

This assumption gives the usual linearized representation of a standard nonlinear GMM estimator which minimizes the suitably weighted quadratic form of sample moment conditions. The condition that $E\left(\boldsymbol{a}_{i,t,\boldsymbol{\theta}_i}\right) = \mathbf{0}$ at the true $\boldsymbol{\theta}_i$ follows from specifying moment conditions for estimating $\boldsymbol{\theta}_i$. The $\mathbf{C}_{i,j,\boldsymbol{\theta}_i}$ are the Jacobians of the moment conditions and the $\mathbf{W}_{i,\boldsymbol{\theta}_i}$ are the limiting weighting matrices (note that the formulation allows for estimated optimal weights). The dependence on $\boldsymbol{\theta}_i$ arises from having possibly nonlinear moment conditions which are linearized for the asymptotics.

In the linear GMM case, the $\mathbf{C}_{i,j,\boldsymbol{\theta}_i}$ are simply the cross-products of instruments and regressors, while the $\boldsymbol{a}_{i,t,\boldsymbol{\theta}_i}$ are the products of instruments and regression errors, say, $\epsilon_{i,t}$. Moreover, $\boldsymbol{r}^{\mathrm{r}}_{i,t} = \mathbf{0}$ in the linear setup. For OLS, of course, regressors serve as instruments and weight matrices cancel out. We thus simply have that $\hat{\boldsymbol{\theta}}^{rol}_{i,t} = \boldsymbol{\theta}_i +$

[9] The explanation is that the full-sample sum in the numerator of the $\mathcal{T}^{DM}$ converges upon normalization to a normal distribution even under time-varying volatility, while the long-run variance estimator converges under small-b to the average long-run variance of the loss differentials as required for robustness (see Cavaliere, 2004).

$\left(\sum_{j=t-R+1}^{t} \boldsymbol{x}_{j,t}\boldsymbol{x}'_{j,t}\right)^{-1} \sum_{j=t-R+1}^{t} \boldsymbol{x}_{i,t}\epsilon_{i,t}$ (and analogously for $\hat{\boldsymbol{\theta}}_{i,t}^{rec}$). Appendix S1 provides further details for the important special case of a linear regression.

In line with the literature (again, see West, 1996), we assume the loss and forecast functions to be smooth enough to allow for an evaluation of the impact of the estimation noise. The assumption covers leading loss functions such as squared error loss as well as generic forecast functions, (cf. again Appendix S1 for a specific example). The gradient characterizing the impact of changes in the parameters on the loss is

$$\boldsymbol{d}_i(f, \boldsymbol{t}) = \left.\frac{\partial \mathcal{L}_t}{\partial u_2}\right|_{\substack{u_1=z_{t+h} \\ u_2=f}} \left.\frac{\partial f_i}{\partial \boldsymbol{\theta}}\right|_{\substack{\boldsymbol{x}_{i,t} \\ \boldsymbol{\theta}_{=t}}}, \tag{9}$$

and we assume it to be uniformly continuous in the following sense.

Assumption 2. There exists $0 < \epsilon < 1/2$ such that, for the neighborhood $\Phi_P = \times_{i=1,2}\left\{\tilde{\boldsymbol{\theta}}_i : \left\|\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right\| < CP^{-1/2+\epsilon}, C > 0\right\}$ of $\left(\boldsymbol{\theta}'_1; \boldsymbol{\theta}'_2\right)'$, it holds as $R, P \to \infty$ with $P/R \to \pi$ that

$$\sup_{(\tilde{\boldsymbol{\theta}}'_1, \tilde{\boldsymbol{\theta}}'_2)' \in \Phi_P; t=R, \ldots, P+R-1} \left\|\boldsymbol{d}_i(\tilde{f}_{i,t}, \tilde{\boldsymbol{\theta}}_i) - \boldsymbol{d}_i(f_{i,t}, \boldsymbol{\theta}_i)\right\| \xrightarrow{p} 0$$

where $\tilde{f}_{i,t} = f_i\left(\boldsymbol{x}_{i,t}, \tilde{\boldsymbol{\theta}}_i\right), i = 1, 2.$

As a consequence, we may write

$$\hat{y}_t^{\mathrm{r}} = y_t + \sum_{i=1}^{2} (-1)^{i+1} \boldsymbol{d}'_i(f_{i,t}, \boldsymbol{\theta}_i) \cdot \left(\hat{\boldsymbol{\theta}}_{i,t}^{\mathrm{r}} - \boldsymbol{\theta}_i\right) + o_p(1), \qquad t = R, \ldots, R+P-1, \tag{10}$$

where the $o_p(1)$ term is negligible uniformly in t (see the proof of Lemma 2.2) and (the transpose of) $\boldsymbol{d}'_i(f_{i,t}, \boldsymbol{\theta}_i)$ is defined in (9). Assumption 2 serves the same purpose as the corresponding Assumption 1(b) of West (1996) requiring a certain boundedness of second derivative of the $f_{i,t}$. The conditions are useful in this form for dealing with the bootstrap later on; see in particular the proof of consistency of our proposed bootstrap approach (Proposition 2) below. It is fulfilled, for example, when the Jacobians of $\boldsymbol{d}_i$ are bounded on Φ_P. To describe the effect of the "estimation noise" terms $\boldsymbol{d}'_i(f_{i,t}, \boldsymbol{\theta}_i) \cdot \left(\hat{\boldsymbol{\theta}}_{i,t}^{\mathrm{r}} - \boldsymbol{\theta}_i\right)$, we make the following mild high-level assumption serving to guarantee a law of large numbers for the average of the derivatives to hold.[10]

Assumption 3. As $P, R \to \infty$ with $P/R \to \pi$, the weak convergence $P^{-1} \sum_{t=R}^{R+[sP]-1} \boldsymbol{d}_i(f_{i,t}, \boldsymbol{\theta}_i) \Rightarrow \boldsymbol{h}_i(s)$, $i = 1, 2$ holds on $s \in [0, 1]$, where $\boldsymbol{h}_i$ are Lipschitz-continuous deterministic vector functions.

To quantify the departures from the standard small-b limits, we specify the behavior of the moment conditions *jointly* with that of y_t (and also characterize the limit behavior of the Jacobians of the moment conditions $\mathbf{C}_{i,j,\boldsymbol{\theta}_i}$):

Assumption 4. Let $\boldsymbol{\xi}_t = \left(\boldsymbol{a}'_{1,t,\boldsymbol{\theta}_1}, \boldsymbol{a}'_{2,t,\boldsymbol{\theta}_2}, y_t - \mu_t\right)' \in \mathbb{R}^{N_1+N_2+1}$ s.t. $\boldsymbol{\xi}_t = \mathbf{G}(t/R)\tilde{\boldsymbol{\nu}}_t$. Assume that

(i) $\mathbf{G}(u)$ is a matrix of piecewise Lipschitz functions, full-rank at all $u \in [0, 1+\pi]$,
(ii) $\tilde{\boldsymbol{\nu}}_t$ has zero mean and unit long-run covariance, and is $L_{2+\delta}$-bounded for some $\delta > 0$, strictly stationary and strong mixing with mixing coefficients $\alpha(j)$ satisfying the summability condition $\sum_{j\geq 0} \alpha(j)^{1/p-1/(2+\delta)} < \infty$ for some $2 < p < 2+\delta$, and
(iii) there exist matrices $\mathbf{C}_i(u)$ of deterministic Lipschitz functions, full-rank for all $u > 0$, such that the weak convergence $R^{-1} \sum_{t=1}^{[uR]} \mathbf{C}_{i,t,\boldsymbol{\theta}_i} \Rightarrow \mathbf{C}_i(u)$ holds on $[0, 1+\pi]$.

[10] One could alternatively state slightly more low-level assumptions on average of $\boldsymbol{d}_i(f_{i,t}, \boldsymbol{\theta}_i)$ for the full sample $t = 1, \ldots, R+P-1$. However, as can be seen in (10), one only needs observations at times $R, \ldots, R+P-1$, so that we state our assumption on $P^{-1} \sum_{t=R}^{R+[sP]-1} \boldsymbol{d}_i(f_{i,t}, \boldsymbol{\theta}_i)$ directly.

The structure of $\mathbf{G}$ is not restricted, since its role is to generate time-varying, symmetric, positive definite (local) long-run covariance matrices $\mathbf{G}(t/R)\mathbf{G}'(t/R)$ for $\boldsymbol{\xi}_t$. Assumption 4 allows for a wide range of patterns of time-varying volatility, including (possibly multiple) abrupt or smooth changes, as well as periodic patterns of heteroskedasticity. The assumption of a non-stochastic variance function $\mathbf{G}(u)$ can moreover be relaxed, for example, under independence conditions between $\mathbf{G}(u)$ and $\tilde{\boldsymbol{\nu}}_t$. The strong mixing condition is fairly mild, too; it is a typical requirement for CLTs and invariance principles for dependent sequences and allows, under suitable restrictions, for various forms of, for example, Markov switching or GARCH models (the surveys of Bradley, 2005, and Lindner, 2009, provide more technical discussions).

Partitioning $\mathbf{G}$ conformably with the components of $\boldsymbol{\xi}_t$, we note that the off-diagonal blocks induce (long-run) correlation of the moment conditions and the loss differentials, which may therefore be time-varying. Correspondingly, block diagonality of $\mathbf{G}$ implies asymptotic independence of the average moment conditions and the loss differentials, case in which the time-variation is rather in their marginal covariance matrices. Clearly, the mixing requirement on $\boldsymbol{\xi}_t$ and the deterministic limit of the sample averages of the Jacobians of the moment conditions imply short memory, so we do not allow for unit root behavior of regressors or instruments in the GMM estimation procedure. We obtain from, for example, Smeekes and Urbain (2014, Lemma 1) the following partial sum behavior:

Lemma 1. *Under Assumption 4 with $\boldsymbol{W}$ a $N_1 + N_2 + 1$ vector of independent Wiener processes, $R^{-1/2}\sum_{t=1}^{[uR]}\boldsymbol{\xi}_t \Rightarrow \int_0^u \mathbf{G}(s)\mathrm{d}\boldsymbol{W}(s) \equiv (\boldsymbol{A}_1'(u), \boldsymbol{A}_2'(u), A_y(u))'$ on $[0, 1+\pi]$.*

The process $\int_0^u \mathbf{G}(s)\mathrm{d}\boldsymbol{W}(s)$ is Gaussian with independent, zero-mean increments, but not a Brownian motion as its quadratic variation $\int_0^s \mathbf{G}(r)\mathbf{G}'(r)\mathrm{d}r$ is nonlinear whenever $\mathbf{G}(\cdot) \neq const$. In particular, this can occur due to breaks or smooth transitions in variances or covariances of $\boldsymbol{\xi}_t$. Its components $\boldsymbol{A}_i$ and A_y are simply the limit processes for the partial sums of the GMM moment conditions and the loss differentials, respectively. We then have the following behavior of the partial sums of $\hat{y}_t^{\mathrm{r}}$, $\mathrm{r} \in \{rol, rec\}$, in the evaluation period $t = R, \ldots, R+P-1$.

Lemma 2. *Let $\mathcal{A}(s) \equiv \left(A_y(1+s\pi) - A_y(1)\right)/\sqrt{\pi}$, and, for $\mathrm{r} \in \{rol, rec\}$, $\tilde{\mathbf{C}}_i^{rol}(s) \equiv \mathbf{C}_i(1+\pi s) - \mathbf{C}_i(\pi s)$, $\tilde{\mathbf{C}}_i^{rec}(s) \equiv \mathbf{C}_i(1+\pi s)$, $\tilde{\boldsymbol{A}}_i^{rol}(s) \equiv \boldsymbol{A}_i(1+\pi s) - \boldsymbol{A}_i(\pi s)$ and $\tilde{\boldsymbol{A}}_i^{rec}(s) \equiv \boldsymbol{A}_i(1+\pi s)$. Under Assumptions 1–4 and the null $\mu_t = 0\ \forall t$, we have, for $s \in [0,1]$,*

$$\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[sP]-1}\hat{y}_t^{\mathrm{r}} \Rightarrow \mathcal{A}(s) + \sqrt{\pi}\sum_{i=1}^{2}(-1)^{i+1}\int_0^s \boldsymbol{N}_i^{\mathrm{r}\prime}(r)(\mathbf{M}_i^{\mathrm{r}})^{-1}(r)\mathrm{d}\boldsymbol{h}_i(r) \equiv B_{\mathbf{G},\pi}^{\mathrm{r}}(s),$$

where $\mathbf{M}_i^{\mathrm{r}}(s) \equiv \tilde{\mathbf{C}}_i^{\mathrm{r}\prime}(s)\mathbf{W}_{i,\boldsymbol{\theta}_i}\tilde{\mathbf{C}}_i^{\mathrm{r}}(s)$ and $\boldsymbol{N}_i^{\mathrm{r}}(s) \equiv \tilde{\mathbf{C}}_i^{\mathrm{r}\prime}(s)\mathbf{W}_{i,\boldsymbol{\theta}_i}\tilde{\boldsymbol{A}}_i^{\mathrm{r}}(s)$.

Proof. See Online Appendix. □

Remark 2. As already discussed by West (1996, Sec. 4), there are situations in which the effect of estimation error is negligible. Lemma 2.2 shows that it is sufficient that $\boldsymbol{h}_i(s) = \boldsymbol{0}$ for all s, as the weak limit of $P^{-1/2}\sum_{t=R+1}^{R+[sP]}\hat{y}_t^{\mathrm{r}}$ then only depends on the limit process for the loss differential, A_y. Verifying whether the condition $\boldsymbol{h}_i(s) = \boldsymbol{0}$ holds or not in a particular application requires information beyond the observed forecast errors. A sufficient condition for this to hold is that $\frac{\partial \hat{\mathcal{L}}_t}{\partial u_2}$ has zero expectation and is uncorrelated with $\frac{\partial f_i}{\partial \boldsymbol{\theta}}$ for both $i = 1, 2$. The first condition (unbiasedness) is quite mild. The second, however, implies both $f_{1,t}$ and $f_{2,t}$ to be rational forecasts. The statistics under study test for equal predictive accuracy only, so rationality may be quite restrictive. It will, however, at least approximately be met in an interesting situation: under stationarity and estimation under the relevant loss, their product $\boldsymbol{d}_i\left(f_{i,t}, \boldsymbol{\theta}_i\right)$ may be close to zero because it represents a f.o.c. for the estimators (following from minimizing the observed loss, $\sum \mathcal{L}\left(z_{t+h}; f_{i,t}(\boldsymbol{\theta})\right)$ w.r.t. $\boldsymbol{\theta}$). See, for example, Appendix S1 for a leading example. The bottom line is that, for all tests considered here, the estimation effect depends in general on the examined forecasting procedures via $\frac{\partial f_i}{\partial \boldsymbol{\theta}}$. In order to compare forecasts, one therefore requires information regarding their construction, that is, information in addition to the point forecasts and the actual realizations, see West (1996) again.

Lemma 2.2 confirms that one also recovers the case without estimation error for $\pi \to 0$ (i.e., when "many" preliminary observations R are available relative to the forecasting periods P), where, again $P^{-1/2}\sum_{t=R}^{R+[sP]-1}\hat{y}_t \Rightarrow \mathcal{A}(s)$. At the same time, for $\pi \to \infty$, the estimation effect dominates.

When the researcher knows that she is in a situation like one of those discussed in this remark, she may simply set $\boldsymbol{d}_i = \boldsymbol{0}$ in Step 4 of the bootstrap algorithm 1 introduced in the following subsection.

Since the processes $B^{\mathrm{r}}_{\mathbf{G},\pi}$ are not Brownian motions is general, Lemma 2.2 implies non-pivotal null distributions for the statistics of interest. With $\Lambda_{k,b}$ from (8), we have the following

Proposition 1. *Under the assumptions of Lemma 2.2 and the null $\mu_t = 0\ \forall t$, we have for $B^{\mathrm{r}}_{\mathbf{G},\pi}$, $\mathrm{r} \in \{rol, rec\}$,*

$$\mathcal{T}^{DM} \xrightarrow{d} (B^{\mathrm{r}}_{\mathbf{G},\pi}(1))^2 / \Lambda_{k,b}\left(B^{\mathrm{r}}_{\mathbf{G},\pi}\right), \qquad \mathcal{T}^{F} \Rightarrow \sup_{s\in[\nu/2,1-\nu/2]} \frac{1}{\nu} \frac{\left|B^{\mathrm{r}}_{\mathbf{G},\pi}\left(s+\frac{\nu}{2}\right) - B^{\mathrm{r}}_{\mathbf{G},\pi}\left(s-\frac{\nu}{2}\right)\right|}{\sqrt{\Lambda_{k,b}\left(B^{\mathrm{r}}_{\mathbf{G},\pi}\right)}}$$

$$\mathcal{T}^{Q} \Rightarrow \sup_{s\in[0,1]} \frac{\left|B^{\mathrm{r}}_{\mathbf{G},\pi}(s)\right|}{\sqrt{\Lambda_{k,b}\left(B^{\mathrm{r}}_{\mathbf{G},\pi}\right)}}, \qquad \mathcal{T}^{C} \Rightarrow \frac{1}{\Lambda_{k,b}\left(B^{\mathrm{r}}_{\mathbf{G},\pi}\right)} \int_0^1 (B^{\mathrm{r}}_{\mathbf{G},\pi}(s))^2 \mathrm{d}s.$$

Proof. See Online Appendix. □

Remark 3. Evidently, the limiting random variables presented in Proposition 2.2 may, together with suitable critical values (see Section 2.3), also be adopted for one-sided testing whenever the researcher has specific alternatives in mind. For example, a signed version of (5), $\max_{R\le t\le R+P-1} S_t/\sqrt{\hat{\Omega}P}$, together with large quantiles of $\sup_{s\in[0,1]} B^{\mathrm{r}}_{\mathbf{G},\pi}(s)/\sqrt{\Lambda_{k,b}\left(B^{\mathrm{r}}_{\mathbf{G},\pi}\right)}$ may be used for right-tailed CUSUM-type tests. See Section 3 for an illustration of one-sided testing. □

Remark 4. Notwithstanding Remark 1, the limiting distributions of $\mathcal{T}^F$, $\mathcal{T}^Q$ and $\mathcal{T}^C$ depend on the entire path of the processes $B^{\mathrm{r}}_{\mathbf{G},\pi}$ via their numerator even when $b \to 0$. Therefore, small-b robustness to time-varying volatility is only given for $\mathcal{T}^{DM}$ in general. □

Given the dependence on time-varying variances in this particular form, a wild bootstrap is a natural candidate to restore asymptotically valid inference. See, for example, Hansen (2000, p. 106) for an early application of the wild bootstrap to replicate sampling distributions affected by unconditional heteroskedasticity. We provide implementation details in the next subsection.

Remark 5. There are alternative ways to deal with time-varying (co)variances, some of which we explore in related work (Demetrescu et al., 2019). These build (i) on estimating $\mathbf{G}$ and using the estimate to time-transform the series so as to restore homoskedasticity and hence apply standard fixed-b inference or (ii) on using a pretesting approach where, depending on the outcome of a test of no unconditional heteroskedasticity, either standard or heteroskedasticity robust fixed-b methods are used. We provide evidence that the wild bootstrap's performance is superior in terms of both size and power. We therefore focus in a wild bootstrap implementation here.

Remark 6. Tests of equal *conditional* predictive ability are obtained by leveraging the loss differentials with a vector $\boldsymbol{w}_t$ of K suitable test functions (Giacomini & White, 2006). To cover this case, one may set $\boldsymbol{y}_t = \boldsymbol{w}_t\left(\mathcal{L}_t\left(z_{t+h}, f_{1,t}\right) - \mathcal{L}_t\left(z_{t+h}, f_{2,t}\right)\right)$ and correspondingly test the null $H_0 : E\left(\boldsymbol{y}_t\right) = \boldsymbol{0}$. Appendix S3 contains the details of a multivariate implementation of tests of equal predictive accuracy. Of course, $w_t = 1$ recovers the unconditional approach on which we focus here. In any case, conditional tests are of course equally affected by time-varying volatility.

2.3 | A wild bootstrap correction

To correct for inherent non-pivotality via the wild bootstrap, the bootstrap scheme must replicate the properties of $B^{\mathrm{r}}_{\mathbf{G},\pi}$, $\mathrm{r} \in \{rol, rec\}$, in the limit. In particular, the wild bootstrap algorithm we propose focuses at replicating the *volatility-related* time-varying properties of all involved series. These properties depend, among others, on $\boldsymbol{h}_i(\cdot)$, $\mathbf{C}_i(\cdot)$, and the joint behavior of $A_y(\cdot)$ and $\boldsymbol{A}_i(\cdot)$. Since $\mathbf{C}_i(\cdot)$, $\mathbf{W}_i$ and $\boldsymbol{h}_i(\cdot)$ are deterministic, this can be achieved by jointly bootstrapping y_t and $\boldsymbol{a}_{i,t}$. To do so, one must however resort to estimated quantities, since y_t and especially $\boldsymbol{a}_{i,t}$ are not observed directly (unless

there is no estimation error, such that y_t is observed and the other quantities do not enter the test statistics at all). While $\hat{y}_t^{\mathrm{r}}$, $\mathrm{r} = \{rec, rol\}$, is a natural estimator for y_t, estimates of $\boldsymbol{a}_{i,t,\boldsymbol{\theta}_i}$, $\mathbf{W}_{i,\boldsymbol{\theta}_i}$ and $\mathbf{C}_{i,t,\boldsymbol{\theta}_{i,t}}$ require plugging in estimates of $\boldsymbol{\theta}_i$, leading to $\hat{\mathbf{C}}_{i,t}^{\mathrm{r}}$, $\hat{\mathbf{W}}_{i,t}^{\mathrm{r}}$ and $\hat{\boldsymbol{a}}_{i,t}^{\mathrm{r}}$:

Algorithm 1

1. Compute $\hat{y}_t^{\mathrm{r}}$ from (2) and $\hat{\mathbf{C}}_{i,t}^{\mathrm{r}}$, $\hat{\mathbf{W}}_{i,t}^{\mathrm{r}}$ and $\hat{\boldsymbol{a}}_{i,t}^{\mathrm{r}}$, $\mathrm{r} = \{rec, rol\}$ as follows:
   - For rolling window estimation:

$$\hat{\mathbf{C}}_{i,t}^{rol} = \mathbf{C}_{i,t,\hat{\boldsymbol{\theta}}_{i,R}^{rol}}, \quad \hat{\mathbf{W}}_{i,t}^{rol} = \mathbf{W}_{i,\hat{\boldsymbol{\theta}}_{i,R}^{rol}}, \hat{\boldsymbol{a}}_{i,t}^{rol} = \boldsymbol{a}_{i,t,\hat{\boldsymbol{\theta}}_{i,R}^{rol}}, \text{ for } t = 1, \ldots, R$$
$$\hat{\mathbf{C}}_{i,t}^{rol} = \mathbf{C}_{i,t,\hat{\boldsymbol{\theta}}_{i,t}^{rol}}, \quad \hat{\mathbf{W}}_{i,t}^{rol} = \mathbf{W}_{i,\hat{\boldsymbol{\theta}}_{i,t}^{rol}}, \hat{\boldsymbol{a}}_{i,t}^{rol} = \boldsymbol{a}_{i,t,\hat{\boldsymbol{\theta}}_{i,t}^{rol}}, \text{ for } t = R+1, \ldots, R+P-1.$$

   - For recursive estimation: set $\hat{\boldsymbol{\theta}}_{i,t}^{rec} = \mathbf{0}$ for $t < N_i$ and compute

$$\hat{\mathbf{C}}_{i,t}^{rec} = \mathbf{C}_{i,t,\hat{\boldsymbol{\theta}}_{i,t}^{rec}}, \hat{\mathbf{W}}_{i,t}^{rec} = \mathbf{W}_{i,\hat{\boldsymbol{\theta}}_{i,t}^{rec}}, \hat{\boldsymbol{a}}_{i,t}^{rec} = \boldsymbol{a}_{i,t,\hat{\boldsymbol{\theta}}_{i,t}^{rec}}, \ t = 1, \ldots, R+P-1.$$

     To save computing time, one may evaluate $\hat{\mathbf{C}}_{i,t}^{\mathrm{r}}$, $\hat{\mathbf{W}}_{i,t}^{\mathrm{r}}$ and $\boldsymbol{a}_{i,t,\cdot}$ at $\hat{\boldsymbol{\theta}}_{i,R+P-1}^{\mathrm{r}}$.
2. For $t = 1, \ldots, R+P-1$, construct wild bootstrap variates $\left(\boldsymbol{a}_{1,t}^{*,\prime}, \boldsymbol{a}_{2,t}^{*,\prime}, y_t^*\right)'$ as $\left(\hat{\boldsymbol{a}}_{1,t}^{\mathrm{r},\prime}, \hat{\boldsymbol{a}}_{2,t}^{\mathrm{r},\prime}, \hat{y}_t^{\mathrm{r}}\right)' r_t^*$, where the multipliers r_t^* are an i.i.d.(0,1) sequence, independent of the data, with $E\left(|r_t^*|^w\right) < \infty \ \forall w \in \mathbb{N}$. Note that, for $t < R$, one may use any values for y_t and $\hat{y}_t^{\mathrm{r}}$ since these do not enter the test statistics $\mathcal{T}^x$, $x \in \{DM, F, Q, C\}$.
3. Construct the bootstrap analogues

$$\hat{\boldsymbol{\theta}}_{i,t}^{*,\mathrm{r}} = \left(\sum_{j=\mathcal{R}}^{t} \hat{\mathbf{C}}_{i,j}^{\mathrm{r},\prime} \hat{\mathbf{W}}_{i,t}^{\mathrm{r}} \sum_{j=\mathcal{R}}^{t} \hat{\mathbf{C}}_{i,j}^{\mathrm{r}}\right)^{-1} \sum_{j=\mathcal{R}}^{t} \hat{\mathbf{C}}_{i,j}^{\mathrm{r},\prime} \hat{\mathbf{W}}_{i,t}^{\mathrm{r}} \sum_{j=\mathcal{R}}^{t} \boldsymbol{a}_{i,j}^* + \hat{\boldsymbol{\theta}}_{i,R+P}^{\mathrm{r}}$$

   for $t = R, \ldots, R+P-1$, where $\mathcal{R} = t - R + 1$ for $\mathrm{r} = rol$ and $\mathcal{R} = 1$ for $\mathrm{r} = rec$.
4. Letting $\hat{f}_{i,t}^{\mathrm{r},*} = f_i\left(\boldsymbol{x}_{i,t}, \hat{\boldsymbol{\theta}}_{i,t}^{\mathrm{r},*}\right)$, $\mathrm{r} \in \{rol, rec\}$, construct the bootstrap sample

$$\hat{y}_t^{\mathrm{r},*} = y_t^* + \boldsymbol{d}_1'(\hat{f}_{1,t}^{\mathrm{r},*}, \hat{\boldsymbol{\theta}}_{1,t}^{*,\mathrm{r}}) \cdot \left(\hat{\boldsymbol{\theta}}_{1,t}^{*,\mathrm{r}} - \hat{\boldsymbol{\theta}}_{1,R+P}^{\mathrm{r}}\right) - \boldsymbol{d}_2'(\hat{f}_{2,t}^{\mathrm{r},*}, \hat{\boldsymbol{\theta}}_{2,t}^{*,\mathrm{r}}) \cdot \left(\hat{\boldsymbol{\theta}}_{2,t}^{*,\mathrm{r}} - \hat{\boldsymbol{\theta}}_{2,R+P}^{\mathrm{r}}\right)$$

   for $t = R, \ldots, R+P-1$.
5. Using the bootstrap sample $\hat{y}_t^{\mathrm{r},*}$, $t = R, \ldots, R+P-1$, compute the bootstrap analogues $\mathcal{T}^{x,*}$, $x \in \{DM, F, Q, C\}$, of the test statistics (3)-(6).
6. Obtain the quantile(s) $q_{1-\alpha}^{x,*}$, $x \in \{DM, F, Q, C\}$, of the respective bootstrap distributions.

In practice, the distribution functions of the bootstrap statistics $\mathcal{T}^{x,*}$ are not known, but can be simulated in the usual way by repeating Steps 2–5 M times for a reasonably large M to obtain consistent empirical analogues via Monte Carlo simulation. Typical choices for the distribution of r_t^* are the Gaussian, Rademacher, or Mammen (1993) distributions.

Some additional conditions are required for establishing the validity of this bootstrap.

Assumption 5.

(i) $\mathbf{W}_{i,\boldsymbol{\theta}_i}$ is continuous in $\boldsymbol{\theta}_i$,

(ii) for $\max\{N_1, N_2\} \leq t \leq R + P - 1$, $\sup_t \left\| \hat{\mathbf{C}}^{\mathrm{r}}_{i,t} - \mathbf{C}_{i,t,\boldsymbol{\theta}_i} \right\| \xrightarrow{p} 0$,

(iii) $\exists \gamma > 0$ such that $\sup_t \left\| \boldsymbol{d}_i(f_{i,t}, \boldsymbol{\theta}_i) \right\| = O_p\left(P^{1/2-\gamma}\right)$ and $\sup_t \left\| \hat{\boldsymbol{a}}^{\mathrm{r}}_{i,t} - \boldsymbol{a}_{i,t,\boldsymbol{\theta}_i} \right\| = O_p\left(P^{-\gamma}\right)$,

(iv) $E\left(\tilde{\boldsymbol{\nu}}_t \tilde{\boldsymbol{\nu}}_t'\right) = c \cdot \mathbf{I}_{N_1+N_2+1}$ with $c > 0$.

Proposition 2. *Under Assumptions 1–5, it holds under the null $\mu_t = 0\ \forall t$ that*

$$P\left(\mathcal{T}^x \geq q^{x,*}_{1-\alpha}\right) \to \alpha,\ x \in \{DM, F, Q, C\}, \quad as \quad R, P \to \infty\ with\ P/R \to \pi.$$

Proof. See Online Appendix. □

Remark 7. The additional Assumption 5(i)-(iii) refers essentially to required smoothness of $\hat{\mathbf{C}}^{\mathrm{r}}_{i,t}$ and $\hat{\boldsymbol{a}}^{\mathrm{r}}_{i,t}$ as functions of the estimators, and is fulfilled in, for example, the linear GMM case; see, for example, Appendix S1. In a nutshell, it transfers the smoothness requirements from Assumption 2 to the bootstrap world. Assumption 5(iv) implies the proposed bootstrap scheme to asymptotically work under the additional condition that $E\left(\tilde{\boldsymbol{\nu}}_t \tilde{\boldsymbol{\nu}}_t'\right) = c \cdot \mathbf{I}_{N_1+N_2+1}$, namely that the covariance and long-run covariance matrices of $\tilde{\boldsymbol{\nu}}_t$ are proportional. This is trivially fulfilled in the case without estimation error, and may for example also be side-stepped when there is one factor driving the volatility changes having the same impact on all components, that is, when $\mathbf{G}(s) = g(s) \cdot \mathbf{G}_0$ for some constant full-rank matrix $\mathbf{G}_0$ and $g(s)$ a piecewise Lipschitz scalar function. A further slightly more restrictive example of this condition being fulfilled is given in case of common dynamics. That we require this condition is a consequence of using a plain-vanilla wild bootstrap in step 2 of the above algorithm, which imposes no serial correlation in the bootstrap error replicates, therefore producing equal covariance and long-run covariance matrices (conditional on the data). The condition would be violated when, for example, the researcher overdifferences the involved series to obtain a reduced-rank long-run covariance matrix. In such cases, one could for example resort to a sieve wild bootstrap (see, e.g., Cavaliere et al., 2010, for an implementation in co-integrated models with time-varying volatility) or, in a less parametric vein, to a block wild bootstrap (see, e.g., Smeekes & Urbain, 2014, who explicitly permit singular long-run covariance matrices) both of which allow to capture the relevant long-run covariance matrix.

Remark 8. As argued in the proof of Proposition 2, $q^{x,*}_{1-\alpha}$ remains unaffected under local alternatives $\mu_t = R^{-1/2}\mu(t/R)$ with μ a non-zero deterministic Lipschitz function $\mu(\cdot)$; see the discussion following Equation (S16) in Appendix S2. At the same time, the limiting behavior of $\mathcal{T}^x$, $x \in \{DM, F, Q, C\}$ can easily be seen to change, so that the bootstrap tests have nontrivial local power.

Remark 9. The algorithm is easily modified to account for the case where only one of the forecasts involves estimated parameters, or when the two forecasts resort to different estimation schemes, one rolling and the other recursive.

Remark 10. While the bootstrap from Algorithm 1 is feasible when a researcher possesses all the necessary information regarding the construction of the forecast, some external sources (cf. Section 3) only publish point forecasts and actual realizations. Such information is not sufficient to assess the relative strengths of privately constructed forecast *models*. Among others, the covariance of $\boldsymbol{A}_i$ and A_y is often not known to "outsiders," making it impossible to apply a suitable bootstrap.

Remark 11. Appendix S4 presents the results of extensive Monte Carlo simulations confirming good finite-sample performance of the bootstrap versions of all statistics considered in this section.

Remark 12. Multiple forecast comparisons, for example, of the kind used for model confidence sets (Hansen et al., 2011), may also be implemented using the proposed bootstrap procedure.

3 | EMPIRICAL RESULTS

3.1 | The Survey of Professional Forecasters data—summary statistics

The survey started in 1968 (conducted by the American Statistical Association and the National Bureau for Economic Research) and is administered by the Federal Reserve Bank of Philadelphia since 1990. Participants are asked to predict main US macroeconomic variables in the middle of each quarter for the current and the following four quarters. We consider two key variables: output growth (RGDP, "Real Gross National Product/Gross Domestic Product") and inflation (PGDP, "Price Index for Gross National Product/Gross Domestic Product").[11]

Our sample includes the 1970s with its severe oil price shocks, leading to increases in macroeconomic volatility and conversely, the "Great Moderation," lasting until the mid-1980s, which exhibited a sharp decline in volatility and predictability (see Campbell, 2007). It is well documented that the "Great Moderation" led to enhanced macroeconomic stability which eased forecasting in general, but also made it more difficult to beat simple time series models (see, e.g., Stock & Watson, 2007). Similarly, Groen et al. (2013) find that regime changes in the variance play an important role for real-time (inflation) forecasting. The sample also covers the "Great Financial Crisis" in 2007/2008. Such a long sample is interesting as it may be possible to identify different episodes in relative forecast performance.

We consider three horizons, namely, nowcasting ($h = 0$), one-quarter ahead ($h = 1$) and 1-year ahead ($h = 4$) forecasts, and two vintages (the first and final releases). Macroeconomic data are often revised significantly, see Croushore and Stark (2001). Faust and Wright (2013) and Stark (2010) discuss and demonstrate the importance of the vintage structure when evaluating SPF (inflation) forecasts. We compare the SPF to model-based forecasts generated in real-time to enable a fair comparison with regard to the available information; see also Stark (2010), D'Agostino et al. (2006), and Coroneo and Iacone (2020).

The dynamic forecast models are economically motivated and include a predictor x_t and an autoregressive term: $z_t = \theta_0 + \theta_1 x_{t-1} + \theta_2 z_{t-1} + e_t$. For output, we use the term spread (in short: TMS), that is, the difference between long-term bond rates and short-term yields, as a predictor. Important references include Estrella and Hardouvelis (1991) for the term spread being an important predictor of real output and Giacomini and Rossi (2006) for the instability of its forecasting performance after the "Great Moderation." For inflation, we use a Phillips curve-based model (in short: PC), see, for example, Stock and Watson (1999). Here, x_t is the unemployment rate. By using the unemployment rate and an intercept rather than the unemployment gap, this specification is in line with the assumption of a constant NAIRU. The forecasting performance of the model and its empirical instability are investigated in, for example, Giacomini and Rossi (2009) and recently in Perron and Yamamoto (2021).

Real-time data from the Federal Reserve Bank of Philadelphia[12] is used to construct rolling window and recursive forecasts with $R = 60$. Interest rate data are taken from the updated data set of Welch and Goyal (2008).[13] In the following, we present evaluation results for the first release and rolling window estimation and discuss differences and similarities for the final release and recursive estimation towards the end of this section.

Figure 1 displays representative mean squared error loss differentials for $h = 0$ for the full sample, which covers 191 quarterly observations from 1969Q4 to 2017Q2.[14] The series reveal that (i) loss differentials are mostly, but not always, positive, indicating advantages of SPF forecasts, (ii) there is potentially some time-variation in the mean, (iii) there are striking volatility changes and (iv) there is some mild to intermediate autocorrelation. Appendices S8 and S9 contain further Figures S33–S37 (S49–S51) for other horizons and releases with similar patterns.

Table 1 provides summary statistics. We report root mean squared error ratios of competing forecasts relative to the SPF, such that values >1 indicate a better performance of the SPF. In all cases, the SPF appears to outperform its competitors. However, there is some notable heterogeneity. The SPF is particularly successful at nowcasting (most strongly so for output). The advantages typically shrink with an increasing forecast horizon. However, the term spread model (TMS) is a strong competitor at $h = 4$, while Phillips curve-based (PC) forecasts are less competitive.

[11]The data files are located at https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/data-files/error-statistics. Appendix S10 presents some results indicating robustness of our findings when investigating unemployment and housing starts, which are also available from the SPF.

[12]The data files are located at https://www.philadelphiafed.org/research-and-data/real-time-center/real-time-data/data-files.

[13]See Amit Goyal's website http://www.hec.unil.ch/agoyal/. In the notation of Welch and Goyal (2008 p. 1459), the ten-year long-term government bond yield and the three-month Treasury bill secondary market rate are labeled as "lty" and "tbl," respectively.

[14]Some series contain a few missing values. Details on imputation are provided in Appendix S7. As there are relatively many missing values in the first year of the survey, we decided to start in 1969Q4.

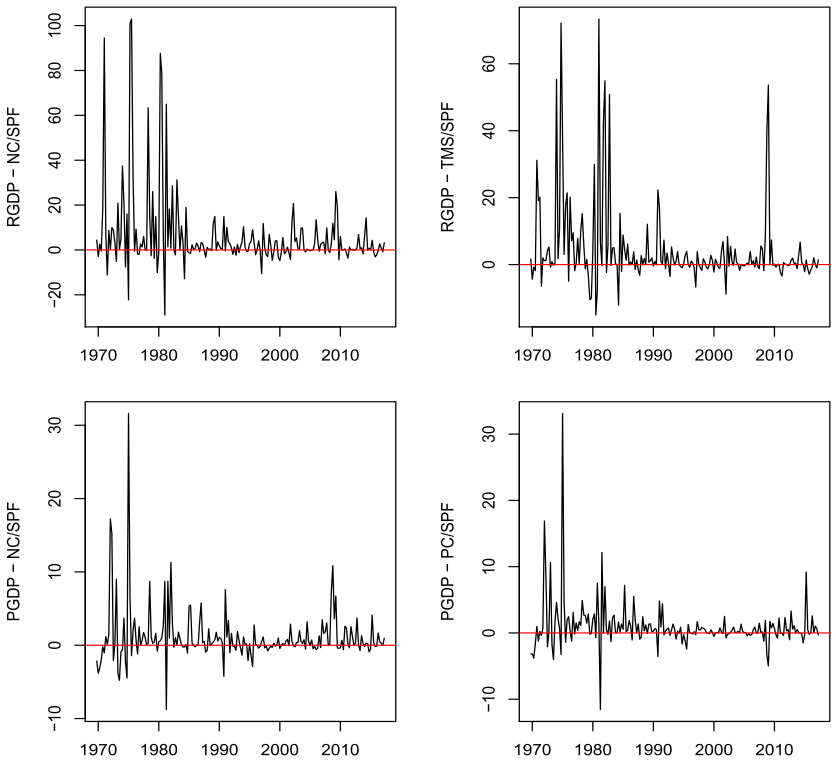

FIGURE 1 Loss differential series for output growth (RGDP) and GDP deflator inflation (PGDP) of competing forecasts against the SPF (NC: no-change; TMS: term spread; PC: Phillips curve). Nowcasts are evaluated against the first release for mean squared error loss

TABLE 1 Summary statistics for output growth (RGDP) and GDP deflator inflation (PGDP) using the first data release

Statistic Sample		RelLoss 1969–2017	SD(I) 1969–1984	SD(II) 1985–2006	SD(III) 2007–2017	AC(1) 1969–2017
RGDP - NC/SPF	$h = 0$	1.69	28.67	4.95	6.02	0.24
	$h = 1$	1.51	60.61	5.49	14.02	0.14
	$h = 4$	1.40	55.26	8.17	15.76	0.44
RGDP - TMS/SPF	$h = 0$	1.52	19.33	4.10	10.39	0.21
	$h = 1$	1.16	16.49	5.42	8.21	0.22
	$h = 4$	1.06	20.27	3.99	1.99	0.04
PGDP - NC/SPF	$h = 0$	1.38	5.88	1.68	2.41	0.08
	$h = 1$	1.23	9.82	2.01	1.91	0.26
	$h = 4$	1.12	16.62	2.33	2.57	0.29
PGDP - PC/SPF	$h = 0$	1.32	5.75	1.43	1.99	-0.02
	$h = 1$	1.26	11.00	1.65	1.83	0.25
	$h = 4$	1.29	22.55	2.41	2.58	0.41

Note: RelLoss denotes the relative root mean squared error loss of the competing forecasts against the SPF (NC: no-change; TMS: term spread; PC: Phillips curve); SD(·) labels the standard deviation of the loss differentials in the subsample I (1969–1984), II (1985–2006), or III (2007–2017). AC(1) denotes the empirical first-order autocorrelation coefficient of the loss differential series.

Unconditional standard deviations for the subsamples I (1969Q4–1984Q4, 61 observations), II (1985Q1–2006Q4, 88 observations) and III (2007Q1–2017Q2, 42 observations) indicate strong overall changes in volatility. This underlines the need for suitable inferential procedures. Structural changes associated with the "Great Moderation" are strongest for real GDP growth (with many break factors being even smaller than 1/5). For output, volatility of loss differentials increased a bit during the "Great Financial Crisis" (relative to the "Great Moderation"), while it stays fairly constant for inflation. Finally, the empirical first-order autocorrelation coefficient indicates a mild to intermediate degree of serial correlation in the loss differentials.

3.2 | Tests for equal predictive ability and time-variation

For all statistics $\mathcal{T}^x$, $x \in \{DM, F, Q, C\}$, we consider $b \in \{0, 0.1, \ldots, 1\}$ for the fixed-b bandwidth parameter. We thus include a classic Newey-West type statistic ($b = 0$, see also Appendix S3, fn. 24) and also the fixed-b versions proposed by Choi and Kiefer (2010). We focus on the Bartlett kernel (i.e., $k(x) = 1 - |x|$ for $|x| < 1$ and $k(x) = 0$ otherwise) due to its higher power relative to the Quadratic Spectral kernel, where both have similar size (cf. Appendix S4). Test decisions and their strengths based on asymptotic, non-robust ("asy") and wild bootstrap ("bs") critical values are compared.

No-change forecasts do not involve parameter estimation while model-based forecasts generally do. For the SPF, the estimation error is not available and therefore, no correction of estimation error is applied, see the discussion in Giacomini and Rossi (2010) and Rossi and Sekhposyan (2016). Therefore, we employ the bootstrap algorithm given in Algorithm 1 with the additional restrictions from Remarks 2 and 9 using $M = 5000$ replications, see also Appendix S1 for further details.

First, we test for equal predictive ability using the full-sample statistic $\mathcal{T}^{DM}$. Table 2 reports rejections at significance levels of 1%, 5%, and 10%. These are labeled as "***", "**", and "*" to ease the presentation of the many results and to conserve space by not reporting six different critical values for each statistic. We consider one-sided tests against the alternative that the SPF outperforms the benchmark.

Starting with output growth (RGDP) and no-change (NC) forecasts, the bootstrap version (subscript "bs") rejects equal predictive ability across the full sample in all cases—at least at the nominal ten percent level, but mostly at the 5% level or lower. This finding holds for all horizons h and all values of the bandwidth-parameter b. It thus clearly suggests that the

TABLE 2 Test decisions for the full-sample $\mathcal{T}^{DM}$-statistic for equal predictive ability of competing forecasts against the SPF (NC: no-change; TMS: term spread; PC: Phillips curve)—either based on wild bootstrap ("bs") or asymptotic critical values ("asy")

	RGDP - NC/SPF						RGDP - TMS/SPF					
	$h=0$		$h=1$		$h=4$		$h=0$		$h=1$		$h=4$	
b	$\mathcal{T}^{DM}_{bs}$	$\mathcal{T}^{DM}_{asy}$	$\mathcal{T}^{DM}_{bs}$	$\mathcal{T}^{DM}_{asy}$	$\mathcal{T}^{DM}_{bs}$	$\mathcal{T}^{DM}_{asy}$	$\mathcal{T}^{DM}_{bs}$	$\mathcal{T}^{DM}_{asy}$	$\mathcal{T}^{DM}_{bs}$	$\mathcal{T}^{DM}_{asy}$	$\mathcal{T}^{DM}_{bs}$	$\mathcal{T}^{DM}_{asy}$
0	***	***	***	***	***	***	***	***	**	**	*	*
0.1	***	***	***	***	***	**	***	***	**	*	*	*
0.2	***	**	***	**	***	**	***	**	**	**	**	*
0.3	***	*	***	*	**	*	**	**	**	**	**	*
0.4	***	*	***	*	**	*	**	**	**	**	**	
0.5	**	*	***	*	**	*	**	*	**	**	**	
0.6	**	*	***	*	**	*	**	*	**	**	**	
0.7	**	*	***	*	*		**	*	**	**	*	
0.8	**	*	***	*	*		**	*	**	**	*	
0.9	**	*	***	*	*		**	*	**	**	*	
1	**	*	***	*	*		**	*	**	**	*	
	PGDP - NC/SPF						PGDP - PC/SPF					
	$h=0$		$h=1$		$h=4$		$h=0$		$h=1$		$h=4$	
b	$\mathcal{T}^{DM}_{bs}$	$\mathcal{T}^{DM}_{asy}$	$\mathcal{T}^{DM}_{bs}$	$\mathcal{T}^{DM}_{asy}$	$\mathcal{T}^{DM}_{bs}$	$\mathcal{T}^{DM}_{asy}$	$\mathcal{T}^{DM}_{bs}$	$\mathcal{T}^{DM}_{asy}$	$\mathcal{T}^{DM}_{bs}$	$\mathcal{T}^{DM}_{asy}$	$\mathcal{T}^{DM}_{bs}$	$\mathcal{T}^{DM}_{asy}$
0	***	***	***	**			***	***	***	***	***	**
0.1	***	***	***	***	**	*	***	***	***	***	***	*
0.2	***	***	***	***	**	*	***	**	***	**	***	*
0.3	***	**	***	***	**	*	***	**	***	**	***	*
0.4	***	**	***	**	**	*	**	*	***	**	***	*
0.5	***	**	***	**	**		**	*	***	**	***	
0.6	***	**	***	**	**		**	*	***	**	***	
0.7	***	**	***	**	**		**	*	***	*	***	
0.8	***	**	***	**	**		**	*	***	**	***	
0.9	***	**	***	**	**		**	*	***	**	***	
1	***	**	***	**	**		**	*	***	**	***	

Note: Nowcasts ($h = 0$), one-quarter ($h = 1$) and 1-year ahead forecasts ($h = 4$) are evaluated against the first data release. Evaluation sample runs from 1969Q4 to 2017Q2.

*Significance level at 10%.
**Significance level at 5%.
***Significance level at 1%.

SPF significantly outperforms its competitors over the full sample. On the contrary, asymptotic critical values produce far weaker and fewer rejections. Results for the term spread model (TMS) are quite similar.

For GDP deflator inflation (PGDP), bootstrap inference leads to rejections at the 1% level in all cases for the shortest horizons $h = 0$ and $h = 1$. Relying on asymptotic critical values mainly produces rejections at the 5% level. We find a clear difference in test decisions for 1-year ahead forecasts ($h = 4$): While the bootstrap detects significant differences, asymptotic inference hardly indicates any significant deviation from equal predictive ability. The differences between the outcomes for testing the superiority of the SPF over no-change or Phillips-curve based model forecasts are quite small.

In sum, the volatility-robust full sample results convincingly indicate the usefulness of the SPF for both variables, especially at short horizons. We next consider tests suitable for detecting time-variation in the relative forecast performance. To this end, we proceed in two steps. First, we apply the $\mathcal{T}^F$ (with $\nu = 0.3$ as suggested in Giacomini & Rossi, 2010), $\mathcal{T}^Q$ and $\mathcal{T}^C$ statistics presented in Section 2 as two-sided versions to test for time-variation in both directions and to ensure that we do not overlook potential periods in which the SPF is outperformed by the benchmarks. It may occur that the SPF is outperformed in some periods and that this feature is reversed in another part of the sample. Second, we investigate the time-varying nature of relative predictive ability of the SPF further by studying the time-varying components of the fluctuation and the CUSUM statistic and consider signed versions of the aforementioned test statistics with one-sided (in favor of the SPF) critical values, see Remark 3. The time-varying components are in particular the (i) rolling standardized mean squared error difference and (ii) scaled partial sum of the loss differential to identify different episodes of relative predictability, if present.

TABLE 3 Test decisions for the time-variation $\mathcal{T}^{\{Q,C,F\}}$-statistics for time-variation in the predictive ability of competing forecasts against the SPF (NC: no-change; TMS: term spread; PC: Phillips curve)—either based on wild bootstrap ("bs") or asymptotic critical values ("asy")

	RGDP - NC/SPF																	
	$h = 0$						$h = 1$						$h = 4$					
b	$\mathcal{T}^Q_{bs}$	$\mathcal{T}^Q_{asy}$	$\mathcal{T}^C_{bs}$	$\mathcal{T}^C_{asy}$	$\mathcal{T}^F_{bs}$	$\mathcal{T}^F_{asy}$	$\mathcal{T}^Q_{bs}$	$\mathcal{T}^Q_{asy}$	$\mathcal{T}^C_{bs}$	$\mathcal{T}^C_{asy}$	$\mathcal{T}^F_{bs}$	$\mathcal{T}^F_{asy}$	$\mathcal{T}^Q_{bs}$	$\mathcal{T}^Q_{asy}$	$\mathcal{T}^C_{bs}$	$\mathcal{T}^C_{asy}$	$\mathcal{T}^F_{bs}$	$\mathcal{T}^F_{asy}$
0	***	***	***	***	***	***	***	***	***	***	**	***	***	**	***	***	**	***
0.1	***	**	***	***	***	***	***	**	***	***	***	***	**	*	***	**	**	**
0.2	***	*	***	**	**	**	***	*	***	**	*	*	**		**	**		*
0.3	**		**	**	*		**		**	*			*		**	*		
0.4	**		**	*			**		**	*					**	*		
0.5	*		**	*			*		**	*					*			
0.6	*		**				*		**						*			
0.7	*		**				*		**						*			
0.8			**				*		**						*			
0.9			**				*		**						*			
1			**				*		*									
	RGDP - TMS/SPF																	
	$h = 0$						$h = 1$						$h = 4$					
b	$\mathcal{T}^Q_{bs}$	$\mathcal{T}^Q_{asy}$	$\mathcal{T}^C_{bs}$	$\mathcal{T}^C_{asy}$	$\mathcal{T}^F_{bs}$	$\mathcal{T}^F_{asy}$	$\mathcal{T}^Q_{bs}$	$\mathcal{T}^Q_{asy}$	$\mathcal{T}^C_{bs}$	$\mathcal{T}^C_{asy}$	$\mathcal{T}^F_{bs}$	$\mathcal{T}^F_{asy}$	$\mathcal{T}^Q_{bs}$	$\mathcal{T}^Q_{asy}$	$\mathcal{T}^C_{bs}$	$\mathcal{T}^C_{asy}$	$\mathcal{T}^F_{bs}$	$\mathcal{T}^F_{asy}$
0	***	***	***	***	***	***	**	**	*	**						*		*
0.1	***	**	***	***	***	***				*						*		**
0.2	**	*	***	**	**	**	*		*	*			**		***	**	**	**
0.3	*		**	**	*		**	*	**	**			*		**	*	*	*
0.4	*		**	*			**	*	**	**					**			
0.5			*	*			*	*	**	**					**			
0.6			*	*			*		*	*					*			
0.7			*	*			**		**	**					*			
0.8			*				*		**	*					*			
0.9			*				*		**	*					*			
1			*				*		**	*					*			

Note: Nowcasts ($h = 0$), one-quarter ($h = 1$) and 1-year ahead forecasts ($h = 4$) are evaluated against the first data release. Evaluation sample runs from 1969Q4 to 2017Q2.

[*] Significance level at 10%.
[**] Significance level at 5%.
[***] Significance level at 1%.

TABLE 4 Continued from Table 3

	PGDP - NC/SPF																	
	$h = 0$						$h = 1$						$h = 4$					
b	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$
0	***	***	***	***	***	***	**	*	*	**		*						
0.1	***	***	***	***	***	***	***	***	***	***	***	***						
0.2	***	**	***	***	***	***	***	**	***	***	***	***						
0.3	***	**	***	**	**	**	***	**	***	***	***	***						
0.4	***	**	***	**	**	*	***	**	***	**	**	**			*			
0.5	***	**	***	**	**	*	***	**	***	**	**	**			*			
0.6	***	**	***	**	**	*	***	**	***	**	**	**			*			
0.7	***	**	***	**	**	*	***	*	***	**	**	**			**			
0.8	**	**	***	**	*	*	***	*	***	**	**	**			*			
0.9	***	**	***	**	**	*	***	*	***	**	**	**			*			
1	***	**	***	**	**	*	***	*	***	**	**	**			*			
	PGDP - PC/SPF																	
	$h = 0$						$h = 1$						$h = 4$					
b	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$	$\mathcal{T}^Q_{\text{bs}}$	$\mathcal{T}^Q_{\text{asy}}$	$\mathcal{T}^C_{\text{bs}}$	$\mathcal{T}^C_{\text{asy}}$	$\mathcal{T}^F_{\text{bs}}$	$\mathcal{T}^F_{\text{asy}}$
0	***	***	***	***	***	***	***	**	**	***	**	**	**		**	**		
0.1	***	**	***	***	***	***	***	**	***	***	**	***	**		**	*		
0.2	**	*	***	**	**	**	***	**	***	**	**	**	*		**	*		
0.3	**		**	**	*		***	*	***	**	*	**			**	*		
0.4	*		**	*			**		***	**					**			
0.5	*		**	*			**		***	*					**			
0.6	*		**				**		***	*					**			
0.7	*		**				**		***	*					**			
0.8			**				**		***	*					**			
0.9	*		**				**		***	*					**			
1	*		**				**		***	*					**			

*Significance level at 10%.
**Significance level at 5%.
***Significance level at 1%.

Tables 3 and 4 report results. Once more, bootstrapped versions of the test statistics provide stronger rejections than their asymptotic counterparts. Since both $\mathcal{T}^{DM}$ (aiming at testing against a constant alternative) and $\mathcal{T}^x$, $x \in \{F, Q, C\}$ (tests allowing for time-varying alternatives) yield quite similar rejections overall, the current results are not clear-cut as to the nature of the alternative.

Figure 2 reveals a sizable and significant deterioration in nowcast predictability in the early 1980s associated with the "Great Moderation." This breakdown is significant and permanent, while the mild recoveries observed for inflation (versus no-change forecasts only) in the early 2000s are too weak for a rejection. For output growth, the results suggest that there is no comeback in relative predictive ability of the SPF. Interestingly, relative forecast performance did not change a lot during the "Great Financial Crisis" even though volatility changed somewhat, but to a much lesser extent when compared with the "Great Moderation." Appendix S8 shows that these results also hold for other horizons, see Figures S38–S39. Figures S43–S45 show the unscaled rolling window mean squared error difference between the SPF and its competitors. They generally support the previous interpretation and reveal that, at least, no-change forecasts never significantly outperform the SPF. The CUSUM statistic indicates a breakdown in relative forecast performance as it also turns significant in the early 1980s, implying that the accumulated changes are large enough for a rejection.[15]

Our interpretation is that the full sample results are mainly driven by the first part of the sample (until the mid-1980s) in which the SPF clearly performed better. As the statistics for time-variation further indicate clearly and robustly, the advantages in relative predictability largely disappear in the mid-1980s. Most of the evidence for time-variation, however, would not have been detected by a traditional analysis using asymptotic critical values.

[15]Its behavior at the beginning and end of the sample provides additional information which $\mathcal{T}^F$ cannot provide due to trimming. Before 1976, there are signs for time-variation in all series. GDP deflator inflation and output growth apparently exhibit some further time-variation after 2010.

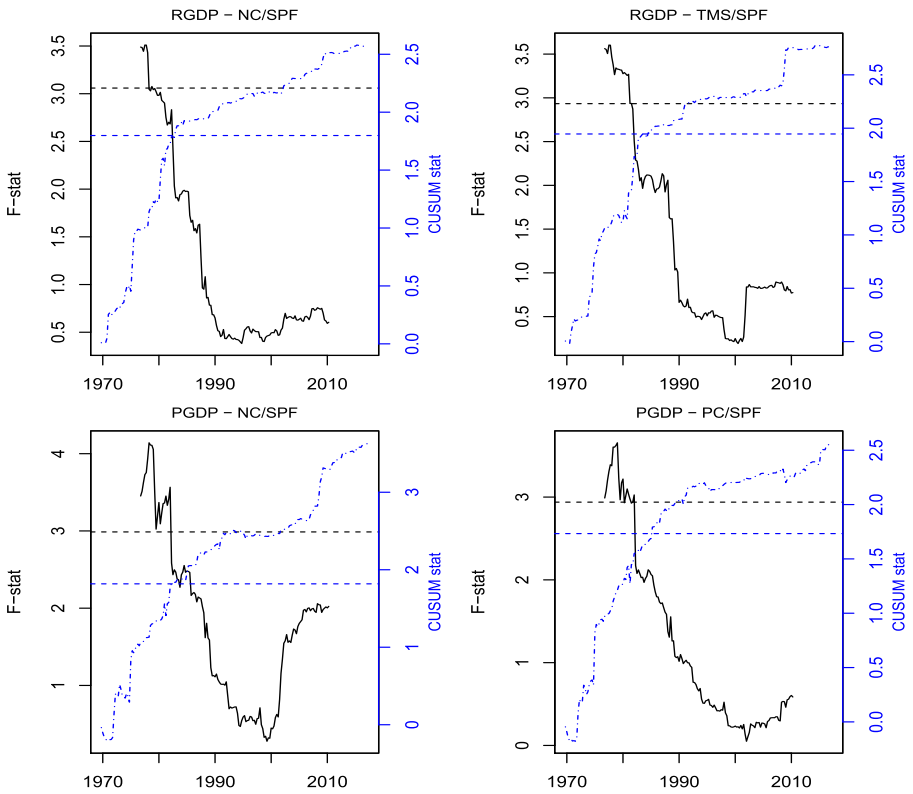

FIGURE 2 The plots show the signed time-varying components of the fluctuation statistic (left axis, solid black line) and the CUSUM statistic (right axis, dashed-dotted blue line), see Equations (4) and (5) and Remark 3. Horizontal dashed lines are the corresponding one-sided five percent critical values for the *maximum* of the displayed statistics. Nowcasts are evaluated against the first release; $b = 0.2$, $\nu = 0.3$

Our results are fairly robust with respect to the vintage (first and final release) and the employed estimation scheme (rolling and recursive). Starting with the descriptive statistics reported in Tables S8 and S12 (see Appendices S8 and S9), we observe very similar patterns to the baseline case in Table 1. The loss differentials in Figures S33–S37 and S49–S51 generally reveal strong heteroskedasticity. Regarding the full sample results (Tables S9 and S13), our main conclusions continue to hold. A notable difference is the case of real output growth when forecasts are evaluated against the final rather than the first release. Here, we find no more evidence for the superiority of the SPF over the term spread model, except when looking at nowcasts. For inflation, on the contrary, results are quite robust throughout various settings. These findings are not affected by the estimation scheme. When looking are tests for time-variation, we obtain very similar conclusions, see Tables S10–S11 and S14–S15. Figures S46–S48 show rolling averages of loss differentials[16] (analogous to Figure 1); see Figures S40–S42 for the components of the statistics designed to detect time-variation (analogous to Figure 2).[17] In nearly all cases, we find the same pattern of advantages for the SPF in the early part of the sample (prior to the "Great Moderation") with a significant decline in the mid-eighties and limited recovery in the 2000s (if at all). An exception (for one- and four-quarter ahead forecasts) is the recursively estimated term spread model for which the relative SPF performance improves since the 2000s.

3.3 | Discussion of our results in light of the related literature

We now provide a comparison of our findings with those of previous studies on the performance of the SPF. Most of these use the Diebold and Mariano (1995) test for differences in mean squared error. One strand of the literature deals with the accuracy of the SPF in general, while a second smaller one focusses on the decline of predictability in connection to the "Great Moderation." A comparison is generally complicated by the fact that studies obviously use different variables (and definitions), benchmarks, vintages, horizons, samples, and so forth. However, two articles, namely, D'Agostino et al. (2006) and Coroneo and Iacone (2020), are particularly close to the scope of our work.

There is some consensus that the SPF provides accurate forecasts, especially nowcasts, for real output growth and inflation. Zarnowitz and Braun (1993) and Croushore (1993) (see also references therein) provide early evidence on the

[16]See Figures S55–S57 for the case of recursively estimated models.

[17]See Figures S52–S54 for the case of recursively estimated models.

good performance of SPF forecasts for real GDP and inflation. Ang et al. (2007) find that surveys (including the SPF) forecast inflation better than macro variables, time series models (including no-change forecasts as advocated by Atkeson & Ohanian, 2001) and asset markets. They also find that when allowing for time-variation, the SPF dominates throughout the whole sample. Croushore (2010) finds confirmatory evidence using real-time data.

The advantages of SPF nowcasts has been documented in several influential studies, for example, Giannone et al. (2008). Liebermann (2014) considers real-time nowcasting for output growth and compares professional forecasters and a dynamic factor model to simple autoregressive and no-change forecasts. The author finds that gains in forecasting accuracy are pronounced for $h = 0$ and decrease in h. For a sample from 1985Q1 to 2007Q4, Stark (2010) similarly finds that the accuracy of the SPF declines significantly for $h > 1$ and that the SPF outperforms no-change forecasts.

We now turn to the discussion of D'Agostino et al. (2006) and Coroneo and Iacone (2020). Both use a naive benchmark (without estimation) under mean squared error loss and deal with time-variation by running tests on subsamples. In contrast to our tests, theirs are not robust to time-varying volatility and do not exploit the full sample to formally and endogenously test for time-variation.

Coroneo and Iacone (2020) propose a Diebold and Mariano (1995) statistic with fixed-m asymptotics (cf. the introduction). Their full-sample test has good size under homoskedasticity even in samples of only 40 observations, while tests using standard small-b-type asymptotics are oversized. Another advantage is the ensured positivity of the estimated long-run variance which is particularly important in small samples and with relatively long forecast horizons, see e.g. Harvey et al. (2017). In addition, Coroneo and Iacone (2020) consider a stationary block-bootstrap version of the test and find it to yield better size than standard asymptotics, again under homoskedasticity, and to be equally powerful as the fixed-m approach. In a sample ranging from 1987Q1 to 2016Q4, the SPF significantly outperforms a naive random walk in some cases for real output growth and inflation (as well as unemployment and interest rates). For output growth and inflation, there is evidence against the null at all horizons except three-quarters ahead. Generally, the evidence is stronger for shorter horizons.

In a subsample analysis with three blocks of ten years of data, the authors investigate time-variation and find: (i) for output growth, the SPF provides constantly superior nowcasts in all three subsamples, while the results for other horizons and subsamples are mixed—overall, the evidence is declining over the subsamples and for horizons beyond one-quarter; (ii) for inflation, relative advantages of the SPF are mainly observed for their last subsample period from 2007 to 2016 at all horizons (except three-quarters ahead). Thus, our findings only partly corroborate those of Coroneo and Iacone ((2020), tabs. 1 and 2) for these two variables. Unlike Stark (2010) and Coroneo and Iacone (2020), we do not find that the SPF easily outperforms naive output and inflation forecasts after the "Great Moderation." In order to further investigate whether the use of different testing environments may serve as an explanation for these differences, we provide an additional analysis reported in Appendix S11. First, we run the $\mathcal{T}^{DM}$-statistic on each of the three subsamples (for SPF vs. no-change nowcasts and one-quarter and one-year ahead forecasts). The different tests mostly agree and give the same answers. Such an outcome is in line with the theory in Section 2 since the volatility varies much more across the individual subsamples rather than within. Second, we run our $\mathcal{T}^{x}, x \in \{F, Q\}$-tests on their subsample to identify periods of instability in relative forecasting performance and thereby, we are able to further compare the test results in light of the applied testing environments. Actually, we find differences as the results do not match very closely. This leads us to conclude that the observed differences in our main analysis may indeed be attributed to the different tests in use. As a by-product, we further provide evidence for instability within the subsamples studied in Coroneo and Iacone (2020) and thus recommend the usage of fluctuation and related tests in general.

D'Agostino et al. (2006) find a significant decline in relative predictive accuracy of the SPF for inflation and output growth for $h = 1$ to $h = 4$. Their full-sample (1975Q1 to 1999Q4) results indicate that the advantages of the SPF appear to be driven by the period prior to 1985 in which the SPF outperformed the naive benchmark, with no significant advantage thereafter. This points strongly to instabilities in the relative forecast performance. Our findings corroborate their results and sharpen them in showing that this phenomenon also holds for nowcasting. In addition, Campbell (2007), D'Agostino and Whelan (2008) and Gamber and Smith (2009) find, through analyses of various subsamples and consistent with our results, declining predictability of the SPF after the "Great Moderation" for output growth and inflation. Explanations regarding the causes of the forecast breakdown differ across these studies and remain an open issue.

By applying robust tests to a fairly long sample of more than 40 years, we obtain results which support several previous findings. Among these are (i) the advantages of the SPF for shortest horizons, but smaller advantages for one-year ahead forecasts; (ii) a significant decline in relative predictability during the 1980s; (iii) the robustness of the relative performance of the SPF to data revisions. Our results yield the following new insights: (i) advantages of the SPF forecasts are minimal in the 1990s, with weak signs of recoveries for GDP deflator inflation later on; (ii) relative forecast performance did not

change during the "Great Financial Crisis," even though volatility increased (although relatively less than during the "Great Moderation") and (iii) the time-variation in the relative performance of the SPF is robust to the evaluation against simple no-change forecasts and dynamic models based on the term spread or the Phillips curve.

The observed recoveries possibly turn into a significant comeback of SPF forecasts in the future. In this case, the exact timing would very likely be unknown (Inoue & Rossi, 2005), rendering a subsample analysis inappropriate. In general, the ad hoc choice of break points may easily lead to biases. Moreover, it is not always possible to invoke economic reasons like the well-studied "Great Moderation." In contrast, the methods proposed here are suitable for data containing possibly multiple unknown breakpoints in forecast performance alongside changes in volatility.

4 | CONCLUDING REMARKS

This paper proposes wild bootstrap tests for equal predictive ability that can be applied when volatility and relative forecast performance may be time-varying, and proves their validity. Both features are present in many macroeconomic and financial forecast comparisons. The tests account for, when needed, rolling and recursive estimation of parameters of forecast models (West, 1996). The considered tests are either full sample tests (Diebold & Mariano, 1995) or CUSUM, Cramér-von Mises and fluctuation statistics when testing for time-variation. All employ fixed-b asymptotics which deliver more accurately sized tests in finite-samples.

Our empirical application investigates the (time-varying) forecast performance of professional forecasters obtained from the SPF relative to simple no-change and model-based forecasts in real-time. The analysis suggests that ignoring time-varying variance seriously affects conclusions regarding the null of equal predictive ability. Traditional tests provide considerably weaker evidence against the null than the wild bootstrap versions. Tests allowing for time-variation indicate that the SPF had significant advantages until the mid-1980s, but not thereafter. Further research might address to what extent the time-varying relative forecast performance can be explained (e.g., Campbell, 2007). Another interesting avenue is to investigate the Fed's popular "Teal Book" forecasts (e.g., D'Agostino & Whelan, 2008; Romer & Romer, 2000; Rossi & Sekhposyan, 2016).

ACKNOWLEDGEMENTS

The authors would like to thank three anonymous referees and the editor Barbara Rossi for their very constructive comments. Seminar and conference participants in Aarhus, Amsterdam, Bath, Bonn, Cologne, Graz, Hannover, Konstanz, Liverpool, Maastricht, Montreal, Münster, Rostock, and Vienna, in particular Ulrich Müller provided very useful comments. The first two authors gratefully acknowledge the support of the German Research Foundation (DFG) through the projects DE 1617/4-2 and SFB 823, TP4. Kruse-Becher gratefully acknowledges financial support from CREATES funded by the Danish National Research Foundation (DNRF78).

OPEN RESEARCH BADGES

This article has been awarded Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. Data is available at http://qed.econ.queensu.ca/jae/datasets/demetrescu002/.

DATA AVAILABILITY STATEMENT

The raw newspaper text data are commercial and cannot be shared but are available from Dow Jones via their newspaper text API. However, the replication packet includes the query used to retrieve data from the Dow Jones API (this may be found in DataSources/DowJones/770cf8c9-5da1-4a02-a57a-2d101b114b57_query.json). Intermediate data include computed text metrics and term-frequency matrices. They are not included in the replication packet because they contain article-level data (such as counts of words by article) that is commercial. The results data are available in the replication packet: these may be found in 'data/results/' and are used to create all of the charts and tables for the paper. The code is also included in the replication packet. The replication packet is available at http://qed.econ.queensu.ca/jae/ under the relevant volume and issue numbers.

REFERENCES

Amado, C., & Teräsvirta, T. (2013). Modelling volatility by variance decomposition. *Journal of Econometrics*, *175*(2), 142–153.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, *59*(3), 817–858.

Ang, A., Bekaert, G., & Wei, M. (2007). Do macro variables, asset markets, or surveys forecast inflation better? *Journal of Monetary Economics*, *54*(4), 1163–1212.

Atkeson, A., & Ohanian, L. E. (2001). Are phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis Quarterly Review*, *25*, 2–11.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, *2*, 107–144.

Campbell, S. D. (2007). Macroeconomic volatility, predictability, and uncertainty in the great moderation: Evidence from the survey of professional forecasters. *Journal of Business & Economic Statistics*, *25*(2), 191–200.

Cavaliere, G. (2004). Unit root tests under time-varying variances. *Econometric Reviews*, *23*(3), 259–292.

Cavaliere, G., Rahbek, A., & Taylor, A. M. R. (2010). Testing for co-integration in vector autoregressions with non-stationary volatility. *Journal of Econometrics*, *158*(1), 7–24.

Choi, H. S., & Kiefer, N. M. (2010). Improving robust model selection tests for dynamic models. *The Econometrics Journal*, *13*(2), 177–204.

Clark, T. E., & Ravazzolo, F. (2015). Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *Journal of Applied Econometrics*, *30*(4), 551–575.

Coroneo, L., & Iacone, F. (2020). Comparing predictive accuracy in small samples using fixed-smoothing asymptotics. *Journal of Applied Econometrics*, *35*(4), 391–409.

Croushore, D. (1993). Introducing: the survey of professional forecasters. *Business Review-Federal Reserve Bank of Philadelphia*, *6*, 3.

Croushore, D. (2010). An evaluation of inflation forecasts from surveys using real-time data. *The BE Journal of Macroeconomics*, *10*(1), 1–32.

Croushore, D., & Stark, T. (2001). A real-time data set for macroeconomists. *Journal of Econometrics*, *105*(1), 111–130.

D'Agostino, A., Giannone, D., & Surico, P. (2006). (Un)Predictability and macroeconomic stability. Working Paper Series 605, European Central Bank.

D'Agostino, A., & Whelan, K. (2008). Federal reserve information during the great moderation. *Journal of the European Economic Association*, *6*(2-3), 609–620.

Davidson, J. (1994). *Stochastic limit theory*. Oxford University Press.

Demetrescu, M., Hanck, C., & Kruse, R. (2019). Robust fixed-*b* inference in the presence of time-varying volatility. Mimeo.

Demetrescu, M., & Sibbertsen, P. (2016). Inference on the long-memory properties of time series with non-stationary volatility. *Economics Letters*, *144*, 80–84.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263.

Estrella, A., & Hardouvelis, G. A. (1991). The term structure as a predictor of real economic activity. *The Journal of Finance*, *46*(2), 555–576.

Faust, J., & Wright, J. H. (2013). Forecasting inflation. In Elliott, G., & Timmermann, A. (Eds.), *Handbook of economic forecasting*, Vol. *2*. Elsevier, pp. 2–56.

Gamber, E. N., & Smith, J. K. (2009). Are the fed's inflation forecasts still superior to the private sector's? *Journal of Macroeconomics*, *31*(2), 240–251.

Giacomini, R., & Rossi, B. (2006). How stable is the forecasting performance of the yield curve for output growth? *Oxford Bulletin of Economics and Statistics*, *68*(s1), 783–795.

Giacomini, R., & Rossi, B. (2009). Detecting and predicting forecast breakdowns. *Review of Economic Studies*, *76*(2), 669–705.

Giacomini, R., & Rossi, B. (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics*, *25*(4), 595–620.

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

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, *55*(4), 665–676.

Groen, J. J. J., Paap, R., & Ravazzolo, F. (2013). Real-time inflation forecasting in a changing world. *Journal of Business & Economic Statistics*, *31*(1), 29–44.

Guidolin, M., & Timmermann, A. (2006). An econometric model of nonlinear dynamics in the joint distribution of stock and bond returns. *Journal of Applied Econometrics*, *21*(1), 1–22.

Hansen, B. E. (2000). Testing for structural change in conditional models. *Journal of Econometrics*, *97*(1), 93–115.

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

Harvey, D. I., Leybourne, S. J., & Whitehouse, E. J. (2017). Forecast evaluation tests and negative long-run variance estimates in small samples. *International Journal of Forecasting*, *33*(4), 833–847.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, *45*(7), 1–47.

Horowitz, J. L., & Savin, N. E. (2000). Empirically relevant critical values for hypothesis tests: A bootstrap approach. *Journal of Econometrics*, *95*(2), 375–389.

Inoue, A., & Rossi, B. (2005). Recursive predictability tests for real-time data. *Journal of Business & Economic Statistics*, *23*, 336–345.

Justiniano, A., & Primiceri, G. E. (2008). The time-varying volatility of macroeconomic fluctuations. *American Economic Review*, *98*(3), 604–641.

Kiefer, N. M., & Vogelsang, T. J. (2002a). Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation. *Econometrica*, *70*(5), 2093–2095.

Kiefer, N. M., & Vogelsang, T. J. (2002b). Heteroskedasticity-autocorrelation robust testing using bandwidth equal to sample size. *Econometric Theory*, *18*(6), 1350–1366.

Kiefer, N. M., & Vogelsang, T. J. (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory*, *21*(6), 1130–1164.

Kiefer, N. M., Vogelsang, T. J., & Bunzel, H. (2000). Simple robust testing of regression hypotheses. *Econometrica*, *68*(3), 695–714.

Li, J., & Patton, A. J. (2018). Asymptotic inference about predictive accuracy using high frequency data. *Journal of Econometrics*, *203*(2), 223–240.

Liebermann, J. (2014). Real-time nowcasting of GDP: A factor model vs. professional forecasters. *Oxford Bulletin of Economics and Statistics*, *76*(6), 783–811.

Lindner, A. M. (2009). Stationarity, mixing, distributional properties and moments of GARCH(p,q)–processes. In T. G. Andersen, R. A. Davis,, J.-P. Kreiss, & T. Mikosch, (Eds.), *Handbook of financial time series* (pp. 43–69). Springer.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics*, *21*, 255–285.

Müller, U. K. (2014). HAC corrections for strongly autocorrelated time series. *Journal of Business & Economic Statistics*, *32*(3), 311–322.

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

Perron, P., & Yamamoto, Y. (2021). Testing for changes in forecasting performance. *Journal of Business & Economic Statistics*, *39*(1), 148–165.

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

Rapach, D. E., & Strauss, J. K. (2008). Structural breaks and GARCH models of exchange rate volatility. *Journal of Applied Econometrics*, *23*(1), 65–90.

Romer, C. D., & Romer, D. H. (2000). Federal reserve information and the behavior of interest rates. *American Economic Review*, *90*(3), 429–457.

Rossi, B. (2013). Advances in forecasting under instability. In Elliott, G., & Timmermann, A. (Eds.), *Handbook of economic forecasting* (Vol. *2*, pp. 1203–1324). Elsevie.

Rossi, B., & Sekhposyan, T. (2016). Forecast rationality tests in the presence of instabilities, with applications to federal reserve and survey forecasts. *Journal of Applied Econometrics*, *31*(3), 507–532.

Sensier, M., & van Dijk, D. (2004). Testing for volatility changes in U.S. macroeconomic time series. *The Review of Economics and Statistics*, *86*(3), 833–839.

Smeekes, S., & Urbain, J.-P. (2014). A multivariate invariance principle for modified wild bootstrap methods with an application to unit root testing. Maastricht University GSBE Research Memoranda, RM/14/008.

Stark, T. (2010). Realistic evaluation of real-time forecasts in the survey of professional forecasters. (*Special Report*): Federal Reserve Bank of Philadelphia, Research Department.

Stock, J., & Watson, M. (1999). Forecasting inflation. *Journal of Monetary Economics*, *44*(2), 293–335.

Stock, J. H., & Watson, M. W. (2002). Has the business cycle changed and why? *NBER Macroeconomics Annual*, *17*(1), 159–218.

Stock, J. H., & Watson, M. W. (2007). Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking*, *39*(S1), 3–33.

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *21*(4), 1455–1508.

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

Zarnowitz, V., & Braun, P. (1993). Twenty-two years of the NBER-ASA quarterly economic outlook surveys: Aspects and comparisons of forecasting performance. In *Business cycles, indicators and forecasting* (pp. 11–94). University of Chicago Press.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Demetrescu, M., Hanck, C., & Kruse-Becher, R. (2022) Robust inference under time-varying volatility: A real-time evaluation of professional forecasters. *Journal of Applied Econometrics*, *37*(5), 1010–1030. https://doi.org/10.1002/jae.2906