



An empirical evaluation of some long-horizon macroeconomic forecasts

Kurt G. Lunsford ^{a,*}, Kenneth D. West^b

^a Federal Reserve Bank of Cleveland, P.O. Box 6387, Cleveland, OH, 44101, USA

^b Department of Economics, University of Wisconsin, Madison, WI, 53706, USA

ARTICLE INFO

JEL classification:

C22
C53
E17

Keywords:

Fractional integration
Forecast interval
Low frequency

ABSTRACT

We use long-run annual cross-country data for 10 macroeconomic variables to evaluate the long-horizon forecast distributions of six forecasting models. The variables we use range from ones having little serial correlation to ones having persistence consistent with unit roots. Our forecasting models include simple time series models and frequency domain models developed in Müller and Watson (2016). For plausibly stationary variables, an AR(1) model and a frequency domain model that does not require the user to take a stand on the order of integration appear reasonably well calibrated for forecast horizons of 10 and 25 years. For plausibly non-stationary variables, a random walk model appears reasonably well calibrated for forecast horizons of 10 and 25 years.

1. Introduction

Long-horizon forecasts are important for economic policy, professional forecasters, climate research, and financial markets. The US Social Security Administration makes projections for several macroeconomic variables out to 75 years, producing intermediate, high-cost, and low-cost scenarios.¹ The Survey of Professional Forecasters asks for 10-year forecasts for several macroeconomic variables.² Nordhaus (2018) uses per capita output forecasts out to the year 2100 and the uncertainties around these forecasts to study climate policies. Inflation caps and floors, studied in Kitsul and Wright (2013), have maturities out to 10 years, and inflation swaps, used in Haubrich et al. (2012), have maturities out to 30 years.

In this paper, we study the forecast distributions of six univariate forecasting models for horizons of up to 50 years. We use three simple time series models and three frequency domain models. Our time series models consist of an independent and identically distributed (iid) model, an autoregressive model of order one (AR(1)), and a random walk model. We choose these models because they are commonly used and cover a range of potential data persistence. Our frequency domain models, which are taken from Müller and Watson (2016), produce forecast distributions using low-frequency information in the data. We choose these models because they were specifically developed for measuring uncertainty around long-horizon forecasts and rely in part on an approach that has been shown to have certain optimality properties within a class of frequency domain models (Lazarus et al., 2018). By focusing on low

* Corresponding author.

E-mail addresses: kurt.lunsford@clev.frb.org (K.G. Lunsford), kdwest@wisc.edu (K.D. West).

¹ The 2025 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds is available at <https://www.ssa.gov/oact/TR/2025/index.html>, accessed on February 11, 2026. In an earlier working paper, Lunsford and West (2021) compare the high- and low-cost scenarios to long-horizon forecast intervals.

² See page 29 of the Survey of Professional Forecasters' documentation at <https://www.philadelphiafed.org/-/media/frbp/assets/surveys-and-data/survey-of-professional-forecasters/spf-documentation.pdf>, accessed on April 17, 2024.

frequencies, these models use the information in the data that is most relevant for long-horizon forecasts without having to model higher-frequency fluctuations.³ The models are an integrated of order 0 model (MW0), an integrated of order 1 model (MW1), and a fractionally integrated model (MWd). For this last model, the order of integration, d , can be between -0.4 and 1 , and the user does not have to take a stand on this parameter.

We evaluate forecasts from these models using a pseudo out-of-sample analysis with long-run annual cross-country data for 10 macroeconomic variables that cover a range of data types and degrees of persistence. The variables include ones that are real (e.g., per capita real GDP growth), nominal (e.g., CPI inflation), and financial (e.g., nominal interest rates). In total, we use 136 series (10 variables with 7 to 18 countries per variable) with each series covering roughly 150 years.

Let x_t denote one of the series. We use the six models to construct forecasts and forecast distributions for the long-horizon average, $(x_{t+h} + \dots + x_{t+h})/h$, for $h = 10, 25$, and 50 years. As in [Zhou et al. \(2010\)](#) and [Müller and Watson \(2016\)](#), forecasting the long-horizon average allows us to appeal to central limit theorems to compute forecast distributions. We evaluate the models along a number of dimensions, as noted below, but our main focus is to evaluate the calibration of each model's forecast distribution for each h . That is, we assess if the forecast distributions can be used to make probability statements that are roughly correct. We make this our focus not only for concision, but also because our central interest is to assess the models' abilities to quantify the uncertainty inherent in long-horizon forecasts.

Specifically, we use the forecast distributions to construct 68 percent nominal forecast intervals and, in our pseudo out-of-sample analysis, we compute coverage rates by calculating how often the realized long-horizon averages fall in the intervals. Because 68 percent nominal forecast intervals are intended to cover 68 percent of data realizations, a correctly calibrated model has a coverage rate of 0.68. We also compute probability integral transforms (PITs) for the forecast distributions and check the uniformity of these PITs.

For forecast intervals and PITs, we evaluate model performance both substantively and statistically. For forecast intervals, we consider whether, across our many data series, the 68 percent intervals generally deliver actual coverage close to 68 percent (substantive) and whether we can reject the null that mean confidence interval coverage is 68 percent (statistical). We apply similar logic to PITs, considering substantive and statistical measures of uniformity. With occasional exceptions that we note, models that performed well substantively also performed well statistically.

Results differ for plausibly stationary (finite variance) and plausibly non-stationary variables. For plausibly stationary variables such as per capita real GDP growth, the AR(1), MW0, and MWd models appear reasonably well calibrated for $h = 10$. The AR(1) and MWd models also work well for $h = 25$. Overall, looking across our many samples of forecasts of stationary series, half of the AR(1) model's coverage rates are between 0.61 and 0.80 at $h = 10$ and between 0.53 and 0.78 at $h = 25$. We also find that the PITs for the AR(1) model are reasonably uniform for $h = 10$ and 25 . Results for the MWd model are similar. For $h = 50$, no model appears to be well calibrated in the substantive sense that their coverage rates are close to 0.68 or that their PITs are uniform. However, we cannot reject the null hypothesis that the AR(1) model's coverage rates or PITs are well calibrated at this horizon.

For plausibly non-stationary variables such as short-term nominal interest rates, the random walk and MW1 models appear reasonably well calibrated for $h = 10$. Coverage rates are typically between 0.58 and 0.77 for the random walk model and between 0.51 and 0.71 for the MW1 model. The PITs from both of these models are roughly uniform. For $h = 25$, the random walk model appears substantively well calibrated with PITs that appear reasonably uniform. However, our formal statistical tests reject the null hypothesis of correct calibration for all models. For $h = 50$, no model appears well calibrated.

One may be concerned that the forecast intervals from well-calibrated models have relatively long lengths. To assess the trade-off of calibration against interval length, we use [Winkler \(1972\)](#) interval scores.⁴ We find that the well-calibrated models also have relatively good Winkler scores.

In an online appendix, we report results on continuous ranked probability scores for whole forecast distributions, which are consistent with the Winkler score results. In the online appendix, we also report the bias, root mean square prediction errors, and mean absolute prediction errors of point forecasts, and we briefly discuss these point forecast results in the body of the paper.

In addition to these main results, we discuss three extensions in [Section 6](#). First, we use data from [Schmelzing \(2020\)](#) to construct 8 samples of long-run interest rates spanning 195 to 709 years. These series appear to be non-stationary. Forecast interval performance generally improves, perhaps because of longer sample sizes. The AR(1), MWd and MW1 models appear to be well calibrated for $h = 10$, as do the MWd and MW1 models for $h = 25$ and $h = 50$. The random walk model is an exception to this general improvement. Second, we experiment with incorporating stochastic volatility into one of our models. We do not find meaningful improvement in calibration. Third, we discuss and experiment with model combination. We compute an equally-weighted combination of the simple time series models, finding that this combination does not improve upon the underlying models.

To our knowledge, we are the first to evaluate long-horizon forecasting models on a variety of macroeconomic variables that cover a wide degree of persistence. In doing so, we show the value of long-run cross-country databases ([Jordà et al., 2017, 2019](#); [Bergeaud et al., 2016](#)) for forecasting research, and we complement the literature that uses long-run data to study in-sample properties of economic and financial variables (e.g. [Chambers et al., 2024](#)). We highlight two further implications of our results. First, reasonably well-calibrated forecasting models are available for horizons between 10 and 25 years for plausibly stationary variables and for

³ In this spirit, [Farmer et al. \(2024\)](#) noted, "the quarter-to-quarter dynamics of the variable may be quite uninformative about its longer-run properties."

⁴ Some forecasters judge that "[a] good prediction interval will be as small as possible while maintaining the specified coverage" and a scaled version of the Winkler score was used in the M4 forecasting competition ([Makridakis et al., 2020](#), p.19).

horizons of 10 years for plausibly non-stationary variables. Hence, models are available for policymakers, professional forecasters, climate researchers, and financial market participants to compute long-run forecast distributions of at least 10 years. Further, the random walk model's reasonably uniform PITs and relatively good Winkler scores at $h = 25$ make it a decent benchmark model for longer horizon forecasts of plausibly non-stationary data. Second, the forecasting models used for stationary variables, such as labor productivity growth, should be different than the models used for non-stationary variables, such as nominal interest rates.

The vast majority of theoretical and empirical work on forecasting considers horizons far shorter than ours. In addition to Müller and Watson (2016), exceptions include the following. Granger and Jeon (2007) proposed that long-horizon forecasts be based on very simple parametric models. They graphically analyze how well their simple models do in terms of forecasting the log level of GDP 10 and 15 years ahead, finding that realized GDP almost always falls within 90 percent confidence intervals. We, too, construct some of our forecasts with very simple parametric models.⁵ Chudý et al. (2020) evaluate the long-horizon forecast intervals of Pascual et al. (2004), Zhou et al. (2010), and Müller and Watson (2016). Their data are daily, so “long-horizon” means many days ahead. They find mixed results for 90 percent forecast intervals, with more accurate coverage at shorter rather than longer horizons. Karmakar et al. (2022) evaluate a long-horizon forecasting model with many predictors. They evaluate their model with high-frequency electricity prices and find that the model we call MWd has good coverage up to 17 weeks ahead. Pesaran et al. (2006) use a univariate model with regime changes to forecast interest rates 60 months ahead, finding that Bayesian procedures work well. Christensen et al. (2018) use Müller and Watson (2016) models and a survey of experts to estimate uncertainty around long-run GDP growth forecasts but do not evaluate these measures of uncertainty. Giannone et al. (2019) use vector autoregressions to predict macroeconomic variables out to 10 years but only evaluate point forecasts. Müller et al. (2022) construct and evaluate a low-frequency model of cross-country GDP growth but do not study the 9 other variables in our data set.

Several of the papers cited above use multivariate models and models of structural change, while our models are univariate and have constant parameters. By showing that these univariate and constant parameter models can be reasonably well calibrated across many variables out to 25 years, we document readily available models for practitioners and also highlight models that can be used as comparative benchmarks in future research. We leave for future research evaluation of more complex forecasting models. Clark (2011) and D'Agostino et al. (2013) find that models of structural change are useful at forecast horizons out to 2 or 3 years, and such models could be interesting to study at longer horizons.

Section 2 of the paper describes our models, Section 3 our data, and Section 4 the mechanics of our forecasting analysis. Section 5 presents the results of our analysis. Section 6 provides extensions, and Section 7 concludes. An online-only appendix provides robustness results and technical details.

2. Forecasting models

For a given variable, x_t , the forecasts that we construct and evaluate are long-horizon averages, $\bar{x}_{\tau,h} = (x_{\tau+1} + \dots + x_{\tau+h})/h$. Anticipating our forecast evaluation below, we begin with a discussion of timing, using notation from West (2006). The total sample size of observed data is T . We use the first R observations to estimate an initial set of model parameters and make an initial forecast. We use $P = T + 1 - R - h$ observations to evaluate the forecasts. We estimate the forecasting models with both rolling and recursive sample schemes. For $\tau = R, \dots, T - h$, the rolling estimation sample is $\{x_{\tau-R+1}, \dots, x_\tau\}$ and the recursive estimation sample is $\{x_1, \dots, x_\tau\}$. For simplicity, we only show notation for the recursive sample going forward.

We use six models to make our forecasts. These include three simple times series models: an iid model, an AR(1) model, and a random walk model. The remaining three models are the MW0, MWd, and MW1 frequency domain models. All of the models are univariate, using only a sample of the variable of interest, $\{x_1, \dots, x_\tau\}$, to forecast $\bar{x}_{\tau,h}$. In addition, we do not model deterministic time trends. In practice, this will mean forecasting growth rates – not levels – of trending variables, such as per capita GDP and CPI. We will forecast the levels rather than growth rates of variables without an obvious time trend – even variables, such as nominal interest rates, whose persistence is consistent with the presence of unit roots.

We first discuss the simple time series models and then discuss the frequency domain models.

2.1. Simple forecasting models

For the iid, AR(1), and random walk models, we appeal to the central limit theorem and treat the future average realization, $\bar{x}_{\tau,h}$, as normally distributed with a conditional mean and variance, $f_{\tau,h}$ and $V_{\tau,h}$. That is, we use $\bar{x}_{\tau,h} \sim N(f_{\tau,h}, V_{\tau,h})$. The adjective “conditional” means dependent on data on x_t realized up to time τ . The three different models have different forms for $f_{\tau,h}$ and $V_{\tau,h}$ that we now show.

The iid Model. The iid model assumes that the data generating process (DGP) for x_t is $x_t = \mu + u_t$, in which u_t is iid with a mean of zero and variance of σ^2 . We estimate μ and σ^2 with $\hat{\mu}_\tau = \tau^{-1} \sum_{t=1}^\tau x_t$ and $\hat{\sigma}_\tau^2 = (\tau - 1)^{-1} \sum_{t=1}^\tau (x_t - \hat{\mu}_\tau)^2$. The mean and variance of the forecast distribution are $\hat{f}_{\tau,h} = \hat{\mu}_\tau$ and $\hat{V}_{\tau,h} = [(1/h) + (1/\tau)]\hat{\sigma}_\tau^2$.

The Random Walk Model. The model assumes that the DGP is $x_t = x_{t-1} + u_t$, in which u_t is iid with a mean of zero and variance of σ^2 . We estimate σ^2 with $\hat{\sigma}_\tau^2 = (\tau - 1)^{-1} \sum_{t=2}^\tau (x_t - x_{t-1})^2$. The mean and variance of the forecast distribution are $\hat{f}_{\tau,h} = x_\tau$ and $\hat{V}_{\tau,h} = (h + 1)(2h + 1)\hat{\sigma}_\tau^2/(6h)$.

⁵ Granger and Jeon (2007) propose forecasting the level of log GDP with a random walk with drift. This is equivalent to our use of what we call an iid model to forecast GDP growth.

The AR(1) Model. The model assumes that the DGP is $x_t = \rho_0 + \rho_1 x_{t-1} + u_t$, in which u_t is iid with a mean of zero and variance of σ^2 . We first estimate ρ_0 and ρ_1 with ordinary least squares and then bias-adjust following Yamamoto and Kunitomo (1984), denoting the bias-adjusted estimates with $\tilde{\rho}_0$ and $\tilde{\rho}_1$ (we suppress dependence on τ for notational simplicity). If $\tilde{\rho}_1 \geq 1$, then we forecast with the random walk model. If $\tilde{\rho}_1 < 1$, then we proceed with the AR(1) model. We compute $\tilde{u}_t = x_t - \tilde{\rho}_0 - \tilde{\rho}_1 x_{t-1}$ and $\tilde{\sigma}_\tau^2 = (\tau - 3)^{-1} \sum_{t=2}^\tau \tilde{u}_t^2$. The mean of the forecast distribution is $\hat{f}_{\tau,h} = \frac{\tilde{\rho}_0}{1-\tilde{\rho}_1} + \frac{1}{h}(\tilde{\rho}_1 + \tilde{\rho}_1^2 + \dots + \tilde{\rho}_1^h)(x_\tau - \frac{\tilde{\rho}_0}{1-\tilde{\rho}_1})$ and the variance is

$$\hat{V}_{\tau,h} = [1 + (1 + \tilde{\rho}_1)^2 + \dots + (1 + \tilde{\rho}_1 + \dots + \tilde{\rho}_1^{h-1})^2] \tilde{\sigma}_\tau^2 / h^2 \\ + [1 + (1 + \tilde{\rho}_1) + \dots + (1 + \tilde{\rho}_1 + \dots + \tilde{\rho}_1^{h-1})] \tilde{\sigma}_\tau^2 / (h^2(\tau - 1)).$$

We provide derivations of the above means and variances in the online appendix. We make three comments here. First, the variance of the iid model has two components. The $(1/\tau)\hat{\sigma}_\tau^2$ component reflects the in-sample uncertainty around $\hat{\mu}_\tau$, which is the mean of the forecast distribution. The $(1/h)\hat{\sigma}_\tau^2$ component reflects uncertainty about future realizations of the data. Because we are forecasting averages, this second component shrinks as the forecast horizon grows. Second, unlike the iid model, the variance of the random walk model increases with the forecast horizon, reflecting that uncertainty about future realizations of the data accumulates with a random walk DGP. Third, the AR(1) model acts as a bridge between the iid and random walk models. If $\tilde{\rho}_1 = 0$, the AR(1) model variance reduces to a form that matches the iid model variance aside from some degree of freedom adjustments. As we note above, if $\tilde{\rho}_1 \geq 1$, then we forecast with the random walk model instead of an explosive AR(1) model.

For these simple models, which are normally distributed, we compute 68 percent equal-tailed forecast intervals with $f_{\tau,h} \pm \sqrt{\hat{V}_{\tau,h}}$, plugging in the mean and variance for each respective model. We use the value of $f_{\tau,h}$ for each respective model as the point forecast.

2.2. Frequency domain forecasting models

The frequency domain forecasting models are from Müller and Watson (2016) and were specifically developed for measuring uncertainty around long-horizon forecasts. These models are based on extracting low-frequency patterns from the sample $\{x_1, \dots, x_\tau\}$ by using a small number, q , of slowly cycling cosine waves. By focusing on low-frequency patterns, these models use the information in the data that is most relevant for long-horizon forecasts without having to model higher-frequency fluctuations. The t th observation of the j th cosine wave is given by $\psi_{j,t} = \sqrt{2} \cos(\pi j(t - 1/2)/\tau)$, which completes one cycle in $2\tau/j$ periods.⁶ To extract the low-frequency patterns, we compute the simple average $\hat{\mu}_\tau = \tau^{-1} \sum_{t=1}^\tau x_t$ and the cosine-weighted averages, $\hat{\beta}_{\tau,j} = \tau^{-1} \sum_{t=1}^\tau \psi_{j,t} x_t$ for $j = 1, \dots, q$. Müller and Watson (2016) prove a central limit theorem, allowing us to treat the forecast distribution for $\tilde{x}_{\tau,h}$ conditional on $\hat{\mu}_\tau$ and $\hat{\beta}_{\tau,1}, \dots, \hat{\beta}_{\tau,q}$ as a generalized Student- t distribution with q degrees of freedom.

We now discuss each frequency domain model with additional details in the online appendix.

The MW0 Model. The model assumes that the DGP is $x_t = \mu + u_t$, in which u_t is a mean zero and integrated of order zero or $I(0)$ process. The forecast distribution has a location parameter of $\hat{\mu}_\tau$ and a scale parameter of $\hat{s}_{\tau,h} = \sqrt{[(1/h) + (1/\tau)](\tau \hat{\beta}'_{\tau,1:q} \hat{\beta}_{\tau,1:q} / q)}$, in which $\hat{\beta}_{\tau,1:q} = [\hat{\beta}_{\tau,1}, \dots, \hat{\beta}_{\tau,q}]'$. The 68 percent equal-tailed forecast interval is $\hat{\mu}_\tau \pm \hat{s}_{\tau,h} t_{0.84}^q$, in which $t_{0.84}^q$ is the 84th percentile of the Student- t distribution with q degrees of freedom. $\hat{\mu}_\tau$ is the point forecast.

In the formula for $\hat{s}_{\tau,h}$, $(\tau \hat{\beta}'_{\tau,1:q} \hat{\beta}_{\tau,1:q} / q)$ corresponds to textbook averaging of periodogram ordinates (e.g. Brockwell and Davis (1991, ch. 10)), but relying on a cosine rather than a Fourier transform. This term is an estimate of the long-run variance of x_t . In contrast to standard treatments, which aim at a consistent estimate of the long-run variance and let the number of ordinates averaged grow with sample size, Müller and Watson (2016) asymptotics hold q fixed as $h, \tau \rightarrow \infty$ with h/τ approaching a nonzero constant and deliver the t distribution. With such asymptotics, the MW0 model adapts both the long-run variance estimator and the fixed- b critical values of the equally weighted cosine test in Lazarus et al. (2018) for out-of-sample forecasting.⁷

We make three remarks, highlighting that the MW0 model generalizes the iid model to an $I(0)$ DGP. First, the location parameter for both models is the sample average. Second, $\hat{s}_{\tau,h}^2$ for the $I(0)$ model and $\hat{V}_{\tau,h}$ for the iid model are both an estimate of the long-run variance of x_t scaled by $(1/h) + (1/\tau)$. For the iid model, the long-run variance estimate of x_t is simply $\hat{\sigma}_\tau^2$. Third, the MW0 model has a Student- t^q distribution, while the iid model is normal. As discussed in the next section, we often use $q = 8$, generating a t^8 distribution that adjusts for distortions that can occur when using estimates of long-run variances to measure uncertainty (Lazarus et al., 2018).

The MW1 Model. The model assumes that the DGP is $x_t = \mu + u_t$, in which u_t is a mean zero and integrated of order one or $I(1)$ process. The formulas for the location and scale parameters are less intuitive than those for the MW0 model and so are in the online appendix. We note here that the location parameter is not simply $\hat{\mu}_\tau$. The 68 percent forecast interval is the location parameter plus or minus the scale parameter times $t_{0.84}^q$. The location parameter is the point forecast.

The MWd Model. The model assumes that the DGP is $x_t = \mu + u_t$, in which u_t is a mean zero and fractionally integrated or $I(d)$ process with fractional parameter $d \in (-0.5, 1)$. This model acts as a bridge between the MW0 ($d = 0$) and MW1 ($d = 1$) models. We treat d as unknown and use a Bayesian approach to construct the forecast distribution. We set a grid of potential values of d ,

⁶ For example, with $q = 8$ cosine waves and an estimation sample size of 48 years, the cosine waves have periods from 12 to 96 years.

⁷ Lazarus et al. (2018) also use Newey and West (1987) to estimate long-run variances with inference based on fixed- b critical values. This method could similarly be adapted for out-of-sample forecasting. However, the fixed- b critical values are nonstandard when using Newey and West (1987), and we do not study them here.

$\{-0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8, 1\}$, and use a uniform prior with equal mass on each grid point. The resulting Bayes predictive density is a weighted average of generalized Student- t densities with q degrees of freedom (Müller and Watson, 2016). Unlike our other forecast distributions, this distribution will not generally be symmetric. We compute the 16th and 84th percentiles of this distribution numerically to form our 68 percent equal-tailed forecast interval. For point forecasts, we use the mean of this distribution when computing bias and root mean squared prediction errors and the median when computing mean absolute errors.⁸

3. Data

We evaluate the forecasting models on 10 annually observed variables: per capita real GDP growth, labor productivity growth, CPI inflation, broad money growth, population growth, total equity returns, short-term nominal interest rates, long-term nominal interest rates, real exchange rates, and investment to GDP ratios. We choose these variables to cover a range of macroeconomic data types and degrees of persistence.

The source for all variables except labor productivity growth is the Macrohistory Database (Jordà et al., 2017, 2019), which has data beginning as early as 1870 and running through 2020. Productivity data, running from 1871 to 2022, are from the Long-Term Productivity Database (Bergeaud et al., 2016).⁹ We convert data with obvious upward trends, such as per capita real GDP, to percentage growth rates as $100 \times \log$ differences. We construct the real exchange rate from the Macrohistory Database's bilateral nominal US dollar exchange rates using its CPI data for the US and the relevant other country.

Details on sample periods and country coverage are in the online appendix. Here is a summary. Most series run 1870s–2020. The number of countries for a given series runs from a low of 7 (investment to GDP) to a high of 18 (labor productivity growth). Table 1 gives the exact number for each variable. To give an idea of country coverage, we list the 17 countries for per capita real GDP growth, CPI inflation, and population growth: Australia, Belgium, Canada, Switzerland, Germany, Denmark, Spain, Finland, France, Great Britain, Italy, Japan, Netherlands, Norway, Portugal, Sweden, and the US. With the exception of labor productivity growth, country coverage for all other series is a subset of these 17. The labor productivity data, which cover 18 countries, omits Australia and Portugal but adds Chile, Greece, and Ireland.

4. Pseudo out-of-sample analysis

4.1. Samples and measures of forecast performance

We evaluate the forecasting models with the following pseudo out-of-sample analysis. $R = 48$ is our initial sample size. For example, for all countries for per capita GDP growth, the initial rolling and recursive sample is 1871–1918. We use this sample to estimate model parameters and compute the initial forecast distribution from each model. We make forecasts over horizons $h = 10, 25$, and 50 years ahead so that the initial forecast distributions are for averages during 1919–1928 ($h = 10$), 1919–1943 ($h = 25$), and 1919–1968 ($h = 50$). We then move one year forward, re-estimate the model parameters, and compute forecast distributions for 1920–1929 ($h = 10$), 1920–1944 ($h = 25$), and 1920–1969 ($h = 50$). We repeat this process until we get to the end of our sample, with the final forecast distributions being for averages during 2011–2020 ($h = 10$), 1996–2020 ($h = 25$), and 1971–2020 ($h = 50$). This routine gives $P = 93$ for $h = 10$, $P = 78$ for $h = 25$, and $P = 53$ for $h = 50$ for all countries for per capita GDP growth. We use this routine with $R = 48$ for all other variables, and the values for P will vary slightly depending on the available sample.

For the Müller and Watson (2016) models, we use $q = 8$ cosine waves for the rolling estimation sample. With $R = 48$, the 8 cosine waves have periods from 12 to 96 years. For the recursive estimation samples, we compute q by rounding $\tau/6$ to the nearest integer for $\tau = R, \dots, T - h$, so that the ratio of the estimation sample size to q roughly matches that of the rolling samples.

For each estimation sample in our pseudo out-of-sample analysis, we estimate the AR(1) model's bias-adjusted slope coefficient and the maximum likelihood value of d from the MWd model.¹⁰ Hence, we have a large number of estimated persistence parameters for each variable: P forecasts times 2 sample schemes (rolling and recursive) for each country. Table 1 summarizes these persistence estimates by showing the median estimated value and the interquartile range (IQR) of estimated values for each of our 10 variables.

Table 1 displays the variables in plausibly stationary (lines (2) to (7)) or plausibly non-stationary (lines (9) to (12)) groups. Plausibly non-stationary variables are those whose estimates of the fractional integration parameter d tend to be above 0.5, plus the real exchange rate, whose median d was 0.4 but which was very highly serially correlated.¹¹ We divide the variables into plausibly stationary and plausibly non-stationary because our discussion below finds this useful for presentational purposes: reducing our hundreds of forecasts to these two groups allows us to highlight central patterns.

Table 1 shows that our variables have a wide range of persistence estimates, ranging from near zero serial correlation in productivity growth to apparent unit roots in long-term nominal interest rates. Hence, our pseudo out-of-sample analysis should be informative about how the forecasting models perform over the range of persistence values encountered in macroeconomic variables.

⁸ See Gneiting (2011) for a discussion of matching point forecasts with loss functions.

⁹ Each data set has only the most recent data vintage. Thus, our forecasting analysis abstracts from the effects of data revisions.

¹⁰ We give estimation details in the online appendix.

¹¹ Instead of categorizing whole variable types into stationary and non-stationary categories, we could do this categorization on a series by series basis. For example, instead of categorizing all population growth variables as stationary, we could categorize population growth for each country depending on, for example, the estimate of d for each country. We show this type of categorization in the online appendix and it has no material effect on our coverage rate results. We also show coverage rate results on a variable by variable basis in the online appendix.

Table 1
Persistence statistics estimated in the $h = 10$ pseudo out-of-sample analysis.

(1)	(2)	(3)	(4)	(5)
variable	no. of countries	no. of estimates of $\hat{\rho}_1$ and $\hat{d}$	$\hat{\rho}_1$	$\hat{d}$
A. Plausibly stationary variables:				
(2) GDP growth	17	3162	0.24 (0.04, 0.33)	-0.2 (-0.4, 0.0)
(3) productivity growth	18	3410	0.10 (-0.01, 0.25)	0.0 (-0.2, 0.2)
(4) CPI inflation	17	3162	0.62 (0.50, 0.70)	0.2 (0.0, 0.4)
(5) money growth	12	2208	0.64 (0.48, 0.77)	0.4 (0.0, 0.6)
(6) population growth	17	3162	0.81 (0.63, 0.91)	0.4 (0.2, 0.6)
(7) equity returns	11	2030	0.18 (0.02, 0.34)	0.0 (-0.4, 0.2)
(8) No. of stationary series	92			
B. Plausibly non-stationary variables:				
(9) short-term interest	9	1662	0.90 (0.80, 0.96)	0.8 (0.8, 1.0)
(10) long-term interest	12	2236	1.00 (0.96, 1.03)	1.0 (0.8, 1.0)
(11) real exchange rate	16	3002	0.88 (0.80, 0.93)	0.4 (0.2, 0.6)
(12) I/Y ratio	7	1314	0.96 (0.91, 1.00)	0.6 (0.4, 0.8)
(13) No. of non-stationary series	44			

Notes:

1. The “no. of estimates” column gives the total number of estimates for the $h = 10$ horizon. For example, for $h = 10$, GDP growth has 17 countries with 2 sample schemes per country and $P = 93$ forecasts per sample scheme. This gives $17 \times 2 \times 93 = 3162$ total forecasts and estimates.
2. The “ $\hat{\rho}_1$ ” column shows the median and IQR (in parentheses) of the estimates of the bias-adjusted AR(1) slope coefficients.
3. The “ $\hat{d}$ ” column shows the median and IQR (in parentheses) of the maximum likelihood estimates of d . The maximum likelihood value is chosen from the grid $\{-0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8, 1\}$.

The main focus of our pseudo out-of-sample analysis will be to evaluate the calibration of each model's forecast distributions. To do this, we construct 68 percent equal-tailed forecast intervals from the 16th and 84th percentiles of the forecast distributions, with formulas provided above.¹² For a given variable, country, and sampling scheme, let $FI_{\tau,h}^{68}$ denote the 68 percent forecast interval made with sample $\{x_t\}_{t=1}^{\tau}$ for horizon h . We then compute the coverage rate, $P^{-1} \sum_{\tau=R}^{T-h} \mathbf{1}(\bar{x}_{\tau,h} \in FI_{\tau,h}^{68})$ with $\mathbf{1}(\cdot)$ being the indicator function, for each variable, country, and sampling scheme. Because 68 percent forecast intervals are intended to cover 68 percent of data realizations, a well-calibrated forecasting model will have a coverage rate of 0.68. Essentially, we are checking if the forecast distributions can be used to make probability statements that are roughly correct.

In addition to checking the calibration of the 68 percent forecast intervals, we compute the probability integral transforms (PITs) to check the calibration of the whole forecast distribution. For a given variable, country, and sampling scheme, let $F_{\tau,h}(\cdot)$ be the forecast cumulative distribution function made with sample $\{x_t\}_{t=1}^{\tau}$ for horizon h . Then, $F_{\tau,h}(\bar{x}_{\tau,h})$ is the PIT. Well-calibrated models have PITs that are uniform on $[0, 1]$ (Dawid, 1984; Diebold et al., 1998).

Using methods described in Subsection 4.2 below, we test the null hypothesis of correct calibration of coverage rates and of PITs. In the results we report below, we indicate with a “+” and a “+” when these tests fail to reject at the 0.05 level for coverage rates and PITs, respectively.

In our pseudo out-of-sample analysis, we also compute average forecast interval lengths and average Winkler (1972) interval scores, with formulas provided below. The Winkler score is a proper and consistent scoring (or loss) function for equal-tailed forecast intervals (Gneiting and Raftery, 2007; Brehmer and Gneiting, 2021), and we use it to assess the trade-off between the calibration and the interval length of a forecasting model.¹³

We close this subsection with four comments. First, we compute coverage rates and the other statistics for 136 series (10 variables with 7 to 18 countries per variable) times 2 sampling schemes, giving 272 values of each statistic. Doubling the number of series in

¹² These equal-tailed forecast intervals are the same as the highest predictive density forecast intervals for the iid, AR(1), random walk, MW0, and MW1 models. Only the MWd model will have different equal-tailed and highest predictive density intervals.

¹³ In the context of our application, a proper scoring rule is one in which use of the true density results in no higher an expected loss than use of any other density. A consistent scoring rule is one in which a researcher who wishes to minimize expected loss will truthfully report the quantiles of the density presumed by the researcher.

lines (8) and (13) in Table 1 to reflect the 2 sampling schemes, we will have 184 values of each statistic for stationary series and 88 for non-stationary series. To summarize this large number of statistics, we report medians and IQRs over the 184 and 88 observations in each of these two groups.

Second, our pseudo out-of-sample forecasts have a heavily overlapping nature. Using per capita GDP growth as an example, the number of forecasts that we have for each country and sampling scheme is $P = 93$ for $h = 10$, $P = 78$ for $h = 25$, and $P = 53$ for $h = 50$ as noted above. Thus, we only have at most 10 non-overlapping forecasts for $h = 10$, 4 non-overlapping forecasts for $h = 25$, and 2 non-overlapping forecasts for $h = 50$. These small samples lead us to follow Müller and Watson (2018) and our own work (Lunsford and West, 2019) in using 68 percent forecast intervals. Because realizations in the tails of a distribution are infrequent, observing behavior in the tails, as is required for evaluation of 90 or 95 percent intervals, would likely require longer samples than we have available.

Third, we use both our judgment and results from hypothesis tests when discussing the calibration of the forecasting models. Judgment is important in light of the huge literature documenting the difficulties of inference when overlapping data are used (see, e.g., Lazarus et al., 2018). For our hypothesis tests, discussed in Subsection 4.2 below, we make adjustments to account for our heavily overlapping data and also to make use of the cross section of variables in our data.

Fourth, we also compute the continuous ranked probability scores of the forecast distributions and the bias, root mean square prediction errors, and mean absolute prediction errors of the point forecasts. We provide some discussion of these statistics below. For concision, we report them in detail in the online appendix.

4.2. Inference about forecast performance

We test the null hypothesis of correct calibration of coverage rates and of PITs for a given model and horizon, against the alternative of incorrect calibration. We make use of our panel of data and account for potential cross-sectional correlations and for our highly overlapping forecasts. For each of our six models and for each measure of forecast performance (coverage rates and PITs), we construct two tests of correct calibration at each horizon. One test examines calibration of forecast intervals for the set of plausibly stationary series, the other for the set of plausibly non-stationary series. The relevant t - and Wald statistics are shown in the online appendix.

We first present some specifics on testing coverage rates and then extend that approach to testings the PITs. For a given model and given horizon h , let $d_{i,\tau} = \mathbf{1}(\bar{x}_{i,\tau} \in FI_{i,\tau}^{68})$ indicate when the realization for series i averaged over $t = \tau + 1, \dots, \tau + h$ falls into the 68 percent forecast interval made in period τ . (In this subsection, we have changed notation from above. In $d_{i,\tau}$ and $FI_{i,\tau}^{68}$, we use i to indicate a series and omit for simplicity subscripts for the model and horizon, h .) We test the null hypothesis of correct calibration, $\mathbb{E}(d_{i,\tau}) = 0.68$, against the alternative, $\mathbb{E}(d_{i,\tau}) \neq 0.68$, by pooling the two sampling schemes for all series in the set and computing the fraction of realizations that fall in the 68 percent forecast intervals. Let S_τ be the set and n_τ be the number of series available at date τ . We compute the fraction via the pooled regression of $d_{i,\tau}$ on a constant. Let $\bar{d} = P^{-1} \sum_{\tau=R}^{T-h} \left(n_\tau^{-1} \sum_{i \in S_\tau} d_{i,\tau} \right)$ be the resulting fraction. We compute a t -statistic as $\bar{d} - 0.68$ divided by a standard error. Computation of the standard error follows Driscoll and Kraay (1998), which permits both cross-sectional and temporal dependence. To compute the long-run variance for the standard error, we cross-sectionally average the estimated residuals, $\hat{e}_{i,\tau} = d_{i,\tau} - \bar{d}$, to get $\tilde{e}_\tau = n_\tau^{-1} \sum_{i \in S_\tau} \hat{e}_{i,\tau}$ and then apply the Bartlett kernel (Newey and West, 1987) with a bandwidth of h to autocovariances of $\tilde{e}_\tau$ for $\tau = R, \dots, T - h$.

Following Lazarus et al. (2018), we use fixed- b critical values (Kiefer and Vogelsang, 2002, 2005) for our hypothesis tests. For $h = 10$, we use $d_{i,\tau}$ from 1928 through 2020, giving us 93 values for $\tilde{e}_\tau$. This gives us a ratio of Bartlett kernel bandwidth to sample size, or b parameter, of $10/93$. We then have $b = 25/78$ for $h = 25$ and $b = 50/53$ for $h = 50$. For $h = 10$ and 25, b is less than the Sun (2014) recommended upper bound of 0.4, and we compute critical values as in Sun (2014).¹⁴ For $h = 50$, b is close to 1, and we take our critical value from Kiefer and Vogelsang (2002). The 95 percent critical values are 2.26, 2.99, and 4.77 for $h = 10, 25$ and 50, respectively.

For testing correct calibration with the PITs, we follow the spirit of Knüppel (2015) and test if the moments of the PITs match the moments of the uniform distribution on $[0, 1]$. This approach allows us to closely follow our coverage rate tests. Let $z_{i,\tau}$ denote the PIT for series i with a forecast distribution made in period τ for a given model and horizon h . We test the null of correct calibration by jointly testing the null of $\mathbb{E}(z_{i,\tau}) = 0.5$ and $\mathbb{E}((z_{i,\tau} - 0.5)^2) = 1/12$ against the alternative of $\mathbb{E}(z_{i,\tau}) \neq 0.5$ or $\mathbb{E}((z_{i,\tau} - 0.5)^2) \neq 1/12$. We estimate pooled regressions of $z_{i,\tau}$ and $(z_{i,\tau} - 0.5)^2$ on a constant with $\bar{m}_1 = P^{-1} \sum_{\tau=R}^{T-h} \left(n_\tau^{-1} \sum_{i \in S_\tau} z_{i,\tau} \right)$ and $\bar{m}_2 = P^{-1} \sum_{\tau=R}^{T-h} \left(n_\tau^{-1} \sum_{i \in S_\tau} (z_{i,\tau} - 0.5)^2 \right)$, respectively. We then compute Wald statistics with $[\bar{m}_1, \bar{m}_2]' - [1/2, 1/12]'$. For the long-run variance-covariance matrix in the Wald statistic, we compute regression residuals, cross-sectionally average the regression residuals, and then apply the Bartlett kernel with a bandwidth of h jointly to those cross-sectional averages. We use fixed- b critical values with $b = 10/93, 25/78$, and $50/53$, for $h = 10, 25$ and 50. For $h = 10$ and 25, critical values follow Sun (2014).¹⁵ For $h = 50$, the critical value is from Table II of Kiefer et al. (2000). The 95 percent critical values are 9.04, 22.50, 51.41 for $h = 10, 25$ and 50, respectively.

¹⁴ Let F_{df_1, df_2}^α be the $1 - \alpha$ quantile of the F distribution with degrees of freedom given by df_1 and df_2 . The critical value is the square root of $\kappa F_{1,K}^\alpha$ with $\kappa = (\exp(b) + 1 + b)/2$ and $K = \max(\text{ceil}(3/(2b)), 1)$.

¹⁵ The critical value is $2\kappa F_{2,K}^\alpha$ with $\kappa = (\exp(5b/3) + 1 + 5b/3)/2$ and $K = \max(\text{ceil}(3/(2b)), 2) - 1$.

Table 2
Coverage rates of nominal 68 percent forecast intervals: medians and IQRs.

(1) (2)	(1) (2)	(2)	(3a) Stationary variables 184 samples		(4b) Non-stationary variables 88 samples	
			median coverage	IQR	median coverage	IQR
(3)	horizon	model				
(4)	10	iid	0.50	(0.39, 0.66)	0.13	(0.10, 0.20)
(5)	10	MW0	0.64 [‡]	(0.55, 0.73)	0.35	(0.26, 0.43)
(6)	10	AR(1)	0.70 [‡]	(0.61, 0.80)	0.55	(0.45, 0.65)
(7)	10	MWd	0.73 [‡]	(0.66, 0.80)	0.56	(0.45, 0.66)
(8)	10	RW	0.95	(0.90, 0.97)	0.70 [‡]	(0.58, 0.77)
(9)	10	MW1	0.77 [‡]	(0.72, 0.83)	0.63	(0.51, 0.71)
(10)	25	iid	0.42	(0.27, 0.63)	0.09	(0.06, 0.12)
(11)	25	MW0	0.58	(0.44, 0.69)	0.22	(0.15, 0.30)
(12)	25	AR(1)	0.66 [‡]	(0.53, 0.78)	0.40	(0.27, 0.58)
(13)	25	MWd	0.67 [‡]	(0.55, 0.77)	0.39	(0.27, 0.56)
(14)	25	RW	0.98	(0.95, 0.99)	0.72	(0.39, 0.84)
(15)	25	MW1	0.86	(0.81, 0.91)	0.61	(0.31, 0.77)
(16)	50	iid	0.27	(0.13, 0.59)	0.04	(0.00, 0.12)
(17)	50	MW0	0.49	(0.22, 0.70)	0.12	(0.00, 0.20)
(18)	50	AR(1)	0.54 [‡]	(0.28, 0.80)	0.31	(0.24, 0.50)
(19)	50	MWd	0.58 [‡]	(0.32, 0.81)	0.33	(0.26, 0.52)
(20)	50	RW	1.00	(0.98, 1.00)	0.78 [‡]	(0.44, 0.89)
(21)	50	MW1	0.92	(0.89, 0.95)	0.64 [‡]	(0.36, 0.83)

Notes:

1. ‡ indicates that the model failed to reject the null hypothesis of correct calibration at the 0.05 level for the relevant set of variables (either stationary or non-stationary). See [Subsection 4.2](#) for details.
2. See [Table 1](#) for categorization of variables as plausibly stationary or plausibly non-stationary.
3. In the list of models in column (2), we use the shorthand “RW” for the random walk model.
4. The number of samples for each group in row (2) is the sum of the number of variables in that group times the two sampling schemes (rolling and recursive).
5. “Median coverage” refers to the median across the number of samples in row (2). Consider the figure of 0.50 in row (4), column (3a). Per columns (1) and (2), the iid model is used to produce 184 sets (92 series $\times$ 2 sampling schemes) of pseudo out-of-sample forecasts for $h = 10$. Actual coverage rates of forecast intervals with nominal 68 percent coverage are computed for each set of pseudo out-of-sample forecasts. 0.50 is the median value across these 184 coverage rates.
6. IQR is interquartile range. It shows the 25th and 75th percentiles of the coverage rates across the number of samples in row (2). These percentiles are computed analogously to the median in the previous note.

5. Pseudo out-of-sample results

5.1. Coverage rates of 68 percent forecast intervals

For each forecasting model and forecast horizon, we compute 184 coverage rates for the plausibly stationary variables and 88 coverage rates for the plausibly non-stationary variables. [Table 2](#) summarizes these coverage rates by reporting the median coverage rate and IQR of the coverage rates for each forecasting model and forecast horizon across the two data categories. ‡s in [Table 2](#), placed next to the median IQR for convenience, reflect not a test about the median but which models fail to reject the null hypothesis of correct calibration at the 5 percent level, based on the hypothesis tests that we discussed in [Subsection 4.2](#).

For arguably stationary variables, we see in columns (3a) and (3b) of [Table 2](#) that the MW0, AR(1), and MWd models are reasonably well calibrated for $h = 10$. We show in the online appendix that the iid model performs comparably to these three models for the subset of arguably stationary variables that have very little serial correlation (i.e., for GDP growth, productivity growth, and equity returns). We say “reasonably well” calibrated to indicate models that have coverage rates near the target of 0.68 and that also do not reject the null hypothesis of correct calibration. We view actual coverage rates of 0.64 to 0.73 (the range of median values for the MW0, AR(1), and MWd models for $h = 10$ in column (3a) of [Table 2](#)) as good outcomes. The IQRs in column (3b) indicate that half or more of the samples yielded coverage rates tolerably close to 0.68. By contrast, the random walk model performs poorly. While the MW1 model does not reject the null hypothesis of correct calibration, it does not appear to be substantively well calibrated. Its IQR in row (9) and column (3b) indicates that its coverage rates are above 0.68 in over 75 percent of the samples. Further, as discussed below, the MW1 model produces forecast intervals that are relatively long and have high Winkler scores (see row (9) and columns

(3a) and (3b) of Tables 3 and 4). Hence, we do not recommend that practitioners use the MW1 model for forecasting stationary variables.

We see in columns (3a) and (3b) that as h increases from 10 to 25 to 50, performance degrades for all models. For the models that performed well for $h = 10$, actual size falls as the horizon increases. The median coverage rates in column (3a) for the AR(1) and MWd models stay quite close to 0.68 for $h = 25$; however, the lengths of their IQRs in column (3b) have gone up, indicating wider dispersion away from 0.68. For $h = 50$, no model appears to be well calibrated in the substantive sense that their coverage rates are often close to 0.68. While we cannot statistically reject that the AR(1) and MWd models are correctly calibrated, their IQRs are quite wide. For MWd, for example, we see from the median and upper end of the IQR that only 25 percent of the samples had actual coverage between 0.58 and 0.81.

Fig. 1 depicts the results for the PITs. We bin the realized PITs separately for the stationary and non-stationary variables into the intervals $[0, 0.2)$, $[0.2, 0.4)$, $[0.4, 0.6)$, $[0.6, 0.8)$, and $[0.8, 1]$. Fig. 1 then shows the fraction of the PITs that fall into each bin for each model and forecast horizon. Because well-calibrated models have PITs that are uniform on $[0, 1]$, an ideal model would have a value of 0.2 in each bin. $\dagger$ s in the legends indicate which models fail to reject the null hypothesis of correct calibration, based on the hypothesis tests that we discuss in Subsection 4.2. We provide numerical values for Fig. 1 in the online appendix.

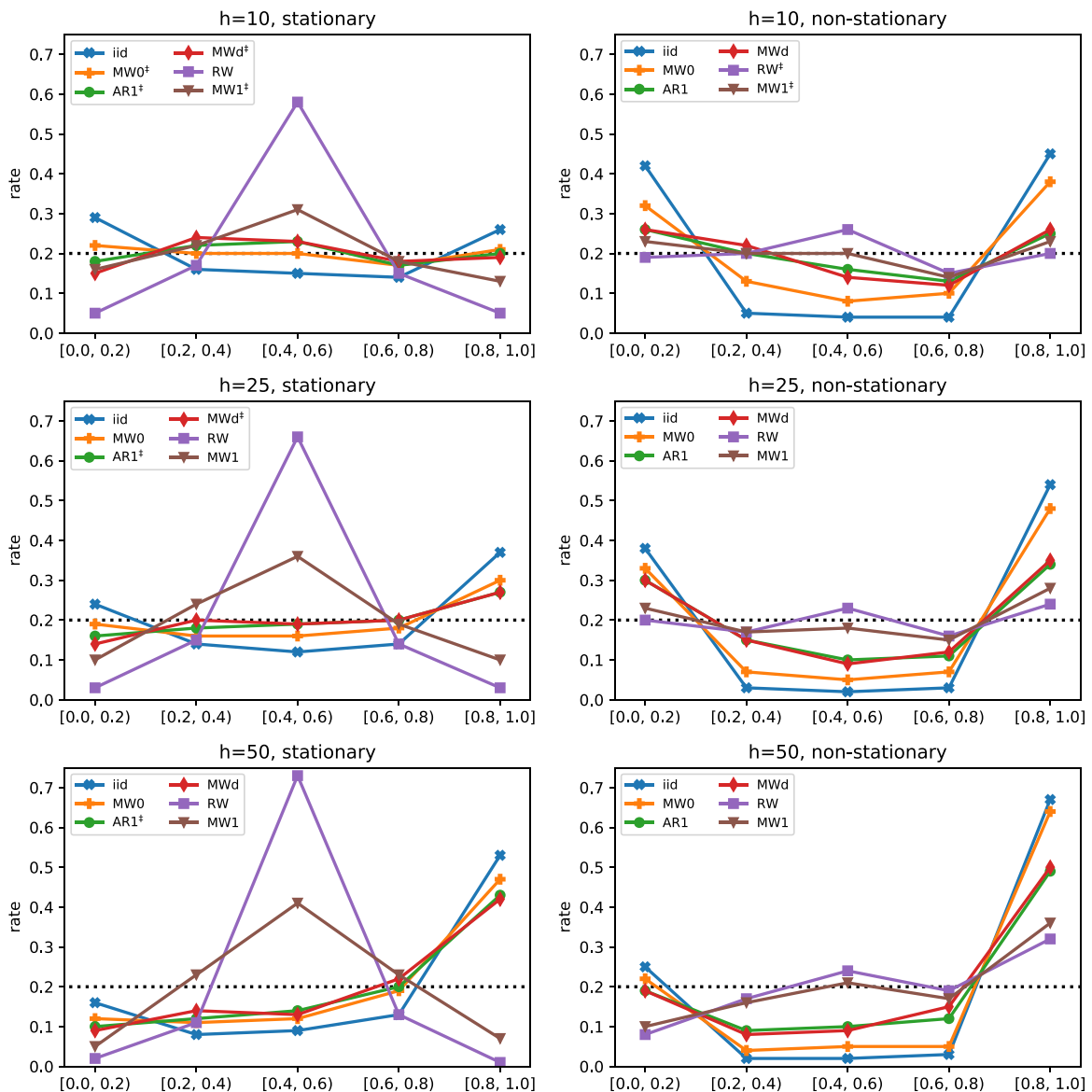


Fig. 1. Rates at which PITs fall into different bins.

Notes: Dotted black lines show the target rate of 0.2. $\dagger$ s in the legends indicate models which fail to reject the null hypothesis of correct calibration at the 0.05 level.

We find that the PITs are generally consistent with our coverage rate results. For the plausibly stationary variables in the left-hand column of Fig. 1, the PITs are close to uniform for the MW0, AR(1), and MWd models for $h = 10$, and we do not reject correct calibration for these models. We also do not reject correct calibration for the MW1 model; however, its PITs appear to be less uniform. For $h = 25$, the AR(1) and MWd models again appear close to uniform, and we do not reject correct calibration. For $h = 50$, no model appears to be well calibrated in the substantive sense that its PITs are approximately uniform, but we cannot statistically reject correct calibration for the AR(1) model.

We return to Table 2 to discuss plausibly non-stationary variables. For $h = 10$, we see in columns (4a) and (4b) of Table 2 that the random walk performs well, with a median coverage rate of 0.70. We reject the null hypothesis of correct calibration for the MW1 model, but its median coverage rate of 0.63 is tolerably close to 0.68. Unsurprisingly, the iid and MW0 models perform quite poorly for this set of highly persistent variables. The AR(1) model (perhaps surprisingly) performs comparably to the MWd model. Indeed, the AR(1) and MWd models often perform similarly across data types, forecast horizons, and evaluation statistics, and we typically discuss these models together. For longer horizons, performance degrades for all models. For $h = 25$ and 50, the median coverage rates are tolerable for the random walk and MW1 models (e.g., 0.72 for random walk for $h = 25$ and 0.64 for MW1 for $h = 50$), but the IQRs are quite long. All models reject the null of correct calibration for $h = 25$ but the random walk and MW1 model do not reject that null for $h = 50$.

The PITs in the right-hand column of Fig. 1 reinforce our coverage rate results for $h = 10$. The PITs for the random walk and MW1 models appear reasonably uniform and we do not reject correct calibration. For $h = 25$, the random walk model appears well calibrated in the substantive sense that its PITs appear reasonably uniform. As reported in detail in our online-only appendix, the distance to 0.2 varies little across the five bins, and the random walk model has the same average distance to 0.2 at $h = 25$ as the AR(1) and MWd models have for the stationary variables. However, despite these substantive results, we statistically reject correct calibration for the random walk model. For $h = 50$, no model appears to have uniform PITs for the plausibly non-stationary variables.

In sum, then, for arguably stationary variables, the AR(1) and MWd models are well calibrated for $h = 10$ and $h = 25$. For $h = 50$, no model appears to be substantively well calibrated, but we cannot statistically reject that the AR(1) and MWd models are correctly calibrated. For arguably non-stationary variables, the random walk and MW1 models are well calibrated for $h = 10$. For $h = 25$, the random walk model appears substantively well calibrated with a median coverage rate and PITs that appear close to target, but the hypothesis tests reject correct calibration. For $h = 50$, the PITs indicate that no model is well calibrated.¹⁶

5.2. Interval lengths

We compute average interval lengths for each model and series by averaging the lengths of the forecast intervals over the P forecasts. We have 184 average interval lengths for plausibly stationary variables and 88 average interval lengths for plausibly non-stationary variables for each forecast horizon. These lengths will scale with the potentially different units of the different series, and we use two approaches to summarize the results. First, we cancel out the units by dividing each average interval length by that from the iid model for the same series and sampling scheme. We then report median and IQR values for relative (to the iid model) average interval length. Second, we report the fraction of the samples in which a given model produces the lowest average interval length among our six models. These two approaches will give a sense of how much model choice affects interval length and how often a particular model has the smallest average interval length.

Table 3 summarizes our results. For both the plausibly stationary and plausibly non-stationary variables, the models can generally be ordered from smallest to largest relative length (with a few exceptions) as follows: iid, MW0, AR(1), MWd, MW1, and random walk. The iid model almost always has the smallest average interval length.

To relate the results in Table 3 to those in Table 2, consider first the iid and MW0 models. These two models have identical point forecasts by construction. In Table 2, judging by either median or IQR, MW0 is better calibrated. Thus, this is a case where, unambiguously, better calibration comes from generally larger interval lengths. Next consider the MW0, AR(1), and MWd models. For $h = 10$, the three models are similarly well calibrated for the stationary variables while having similar interval lengths. For $h = 25$, the AR(1) and MWd models remain well calibrated, while the MW0 model becomes undersized. Table 3 shows that the AR(1) and MWd models' relative interval lengths increase slightly from $h = 10$ to $h = 25$, while the MW0 model's relative lengths do not. This example again shows that calibration is related to interval length.

We defer further discussion of interval lengths to the next subsection on Winkler scores.

5.3. Winkler interval scores

We have just seen that the random walk model has relatively long intervals and the iid model has relatively short intervals. In our discussion of calibration in Subsection 5.1, the random walk model received some support (for non-stationary variables), while the iid model received little support for either the stationary or the non-stationary series. These two models highlight that forecasters may face a trade-off between (1) models that are reasonably well calibrated but have relatively long forecast intervals, and (2) models that

¹⁶ A reader asked whether our negative results for forecasts 50 years ahead have implications for policymakers, such as Social Security, who make even longer-term projections. Our negative result may only apply to the six models that we study. For example, to our understanding, the way Social Security constructs projections (see "The Long-range Economic Assumptions for the 2025 Trustees Report" at https://www.ssa.gov/oact/TR/2025/2025_Long-Range_Economic_Assumptions.pdf, accessed on February 11, 2026) is so different from the way we construct forecasts that we do not feel we can extrapolate from our forecasts to Social Security's projections.

Table 3

Average lengths: medians and IQRs of relative values and fraction with minimum value.

	(1)	(2)	(3a)	(3b)	(3c)	(4a)	(4b)	(4c)
(1)			Stationary variables			Non-stationary variables		
(2)			184 samples			88 samples		
			median			median		
	horizon	model	relative	IQR	fraction	relative	IQR	fraction
(3)			length		with min	length		with min
					length			length
(4)	10	iid	1.00	(1.00, 1.00)	0.88	1.00	(1.00, 1.00)	1.00
(5)	10	MW0	1.39	(1.15, 1.82)	0.08	2.30	(2.18, 2.42)	0.00
(6)	10	AR(1)	1.42	(1.15, 1.96)	0.03	2.18	(1.88, 2.34)	0.00
(7)	10	MWd	1.56	(1.24, 2.15)	0.01	2.38	(2.17, 2.62)	0.00
(8)	10	RW	6.64	(5.00, 7.64)	0.00	2.91	(2.12, 3.58)	0.00
(9)	10	MW1	2.73	(2.23, 3.08)	0.00	2.49	(2.21, 3.15)	0.00
(10)	25	iid	1.00	(1.00, 1.00)	0.83	1.00	(1.00, 1.00)	1.00
(11)	25	MW0	1.38	(1.14, 1.81)	0.07	2.28	(2.17, 2.40)	0.00
(12)	25	AR(1)	1.51	(1.17, 2.25)	0.04	3.70	(3.37, 4.27)	0.00
(13)	25	MWd	1.73	(1.25, 2.67)	0.07	3.80	(3.50, 4.04)	0.00
(14)	25	RW	14.25	(10.83, 16.66)	0.00	6.46	(4.71, 7.83)	0.00
(15)	25	MW1	5.75	(4.62, 6.53)	0.00	5.38	(4.53, 6.58)	0.00
(16)	50	iid	1.00	(1.00, 1.00)	0.78	1.00	(1.00, 1.00)	1.00
(17)	50	MW0	1.37	(1.12, 1.81)	0.08	2.22	(2.11, 2.38)	0.00
(18)	50	AR(1)	1.62	(1.18, 2.38)	0.05	5.28	(4.11, 7.01)	0.00
(19)	50	MWd	1.82	(1.20, 2.91)	0.10	5.16	(4.39, 6.07)	0.00
(20)	50	RW	24.12	(18.99, 28.50)	0.00	12.25	(9.11, 13.83)	0.00
(21)	50	MW1	9.98	(7.79, 11.18)	0.00	10.47	(8.80, 11.46)	0.00

Notes:

1. See Table 1 for categorization of variables as plausibly stationary or plausibly non-stationary.
2. In the list of models in column (2), we use the shorthand “RW” for the random walk model.
3. The number of samples for each group in row (2) is the sum of the number of variables in that group times the two sampling schemes (rolling and recursive).
4. For each model in each sample, average interval length is expressed relative to average interval length for the iid model in that sample. Medians and IQRs (interquartile ranges) of the resulting relative average interval lengths were constructed as described in the notes to Table 2.
5. “fraction with min length” reports the fraction of the samples for a given horizon in which the corresponding model has the lowest average length among the six models.

are poorly calibrated but have relatively short forecast intervals.¹⁷ To quantify this trade-off, we use Winkler (1972) interval scores to assign a loss to each model (for each variable, forecast horizon, and sampling scheme). We compute Winkler scores as follows. Let $u_{\tau,h}$ and $\ell_{\tau,h}$ be the upper (84th percentile) and lower (16th percentile) bounds of the 68 percent forecast interval made using $\{x_t\}_{t=1}^{\tau}$ for horizon h . Then, the Winkler score is

$$u_{\tau,h} - \ell_{\tau,h} + 2(\ell_{\tau,h} - \bar{x}_{\tau,h})\mathbf{I}(\ell_{\tau,h} > \bar{x}_{\tau,h})/(1 - 0.68) + 2(\bar{x}_{\tau,h} - u_{\tau,h})\mathbf{I}(u_{\tau,h} < \bar{x}_{\tau,h})/(1 - 0.68), \quad (5.1)$$

in which $\mathbf{I}(\cdot)$ is the indicator function. We average these Winkler scores over $\tau = R, \dots, T - h$ and all of our discussion is in terms of these averages. This loss penalizes both interval length and inaccurate values for the 16th and 84th percentiles. Hence, it can be used to adjudicate the calibration-length trade-off. As we noted above, this measure is both consistent and proper (Gneiting and Raftery, 2007; Brehmer and Gneiting, 2021). Lower Winkler scores are better.

As with average interval lengths, we compute relative (to the iid model) average Winkler scores for each variable, country, and sampling scheme in order to divide out units. We report medians and IQRs of these relative values.¹⁸ We also compute the fraction of samples in which a given model has the smallest average Winkler score.

Table 4 shows these results. For the plausibly stationary variables, at $h = 10$, we see in columns (3a) and (3b) of Table 4 that the relative Winkler scores for the iid, MW0, AR(1), and MWd models are quite similar. We also see in column (3c) that the AR(1) and MWd models produce the lowest Winkler scores at rates similar to those of the iid model. Taken together with the coverage rate results in Table 2, we find that the MW0, AR(1), and MWd models are better calibrated than the iid model but without a noticeable increase in average Winkler scores. That Winkler performance for iid is comparable to MW0, AR(1) and MWd, despite the poor results for coverage, thus reflects the contribution of interval length to a Winkler score. Or, to state the flip side of the same point, the longer intervals of the MW0, AR(1), and MWd models do not outweigh their better calibration according to the metric of Winkler scores.

For the stationary variables, at $h = 25$, Table 4 shows that the iid, MW0, AR(1), and MWd models again have similar relative Winkler scores and similar rates of producing the lowest Winkler score. We saw above that the AR(1) and MWd models are the best

¹⁷ This is a well-known trade-off that may also be referred to as being between calibration and “sharpness” (Gneiting et al., 2007; Gneiting and Katzfuss, 2014).

¹⁸ We have also computed relative average Winkler scores using the MW0, AR(1), and MWd models in the denominator, but found little change in ranking of median and IQR values.

Table 4

Winkler scores: medians and IQRs of relative values and fraction with minimum value.

(1) (2)	(1)	(2)	(3a) (3b) Stationary variables 184 samples		(3c)	(4a) (4b) Non-stationary variables 88 samples		(4c)
			median relative Winkler	IQR		median relative Winkler	IQR	
(3)	horizon	model			fraction with min Winkler			fraction with min Winkler
(4)	10	iid	1.00	(1.00, 1.00)	0.32	1.00	(1.00, 1.00)	0.03
(5)	10	MW0	0.98	(0.91, 1.03)	0.11	0.82	(0.80, 0.85)	0.08
(6)	10	AR(1)	0.99	(0.88, 1.07)	0.29	0.54	(0.46, 0.65)	0.28
(7)	10	MWd	0.98	(0.87, 1.08)	0.26	0.57	(0.50, 0.68)	0.02
(8)	10	RW	3.15	(1.62, 4.42)	0.02	0.52	(0.45, 0.68)	0.48
(9)	10	MW1	1.63	(1.21, 2.13)	0.01	0.56	(0.46, 0.73)	0.10
(10)	25	iid	1.00	(1.00, 1.00)	0.26	1.00	(1.00, 1.00)	0.00
(11)	25	MW0	0.97	(0.87, 1.03)	0.24	0.88	(0.83, 0.90)	0.18
(12)	25	AR(1)	0.99	(0.89, 1.08)	0.27	0.79	(0.64, 0.88)	0.15
(13)	25	MWd	1.02	(0.86, 1.13)	0.21	0.75	(0.61, 0.87)	0.23
(14)	25	RW	4.99	(2.67, 10.01)	0.01	0.83	(0.68, 1.00)	0.27
(15)	25	MW1	2.26	(1.51, 3.38)	0.02	0.85	(0.72, 1.02)	0.17
(16)	50	iid	1.00	(1.00, 1.00)	0.29	1.00	(1.00, 1.00)	0.01
(17)	50	MW0	0.96	(0.86, 1.04)	0.25	0.89	(0.84, 0.94)	0.16
(18)	50	AR(1)	1.01	(0.86, 1.14)	0.22	0.81	(0.69, 0.89)	0.17
(19)	50	MWd	1.04	(0.88, 1.19)	0.21	0.78	(0.68, 0.86)	0.23
(20)	50	RW	8.31	(3.83, 16.39)	0.01	0.91	(0.65, 1.29)	0.25
(21)	50	MW1	3.41	(1.91, 5.68)	0.02	0.86	(0.65, 1.31)	0.18

Notes:

1. See [Table 1](#) for categorization of variables as plausibly stationary or plausibly non-stationary.
2. In the list of models in column (2), we use the shorthand “RW” for the random walk model.
3. The number of samples for each group in row (2) is the sum of the number of variables in that group times the two sampling schemes (rolling and recursive).
4. For each model in each sample, the average Winkler score is expressed relative to the average Winkler score for the iid model in that sample. Medians and IQRs (interquartile ranges) of the resulting relative average Winkler scores were constructed as described in the notes to [Table 2](#).
5. “fraction with min Winkler” reports the fraction of the samples for a given horizon in which the corresponding model has the lowest average Winkler score among the six models.

calibrated at this horizon. Hence, forecasters can use the AR(1) and MWd models to achieve reasonably good calibration with no degradation in Winkler loss relative to any of our other models.

For the plausibly stationary variables, the random walk and MW1 models usually have the highest Winkler scores at all horizons. These high values result from wide forecast intervals and poor calibration. For example, for $h = 10$, column (3a) in [Table 3](#) shows that the median relative interval length for the random walk is 6.64 (as opposed to, for example, 1.39 for MW0). Large interval lengths are consistent with the random walk and MW1 models generally having interval coverage well above 68 percent in [Table 2](#).

For the plausibly non-stationary variables for $h = 10$, we see in columns (4a) and (4b) of [Table 4](#) that the relative Winkler scores for the AR(1), MWd, random walk, and MW1 models are similar. The relative scores for the iid and MW0 models are noticeably higher. Taken together with the coverage rate results in [Table 2](#), we find that the random walk and MW1 models are better calibrated than the AR(1) and MWd models but without a noticeable increase in relative Winkler scores. Column (4c) in [Table 4](#) shows that the random walk model often produces the lowest average Winkler scores.

For the plausibly non-stationary variables, at $h = 25$, the relative Winkler scores for the random walk model appear slightly higher than those for the AR(1) and MWd models. However, the random walk model produces the lowest Winkler scores in the highest fraction of samples. Overall, the random walk model’s scores do not appear materially worse than those for the AR(1) and MWd models despite the random walk model’s materially larger interval lengths.

As noted above, Winkler scores assess the trade-off between interval length and the joint accuracy of the 16th and 84th forecast percentiles. We use these interval scores for consistency with our coverage rate and interval length results, all of which are based on 68 percent forecast intervals. However, we also compute a continuous ranked probability score, which is a proper scoring function for the whole forecast distribution ([Gneiting and Raftery, 2007](#)), for each forecast distribution that we compute. The continuous ranked probability scores show essentially the same results as the Winkler scores, and we show the continuous ranked probability scores in the online appendix.

5.4. Point forecasts

While long-horizon forecast intervals and distributions are our focus, researchers and practitioners may also be interested in the quality of long-horizon point forecasts. In this subsection, we summarize results on the root mean square prediction errors (RMSPEs) and mean absolute prediction errors (MAPEs) of our forecasting models. Detailed results are in our online appendix. Briefly, we find that our point forecast results generally parallel our Winkler score results.

For the stationary variables at all horizons, the AR(1) and MWd models have similar RMSPEs and MAPEs relative to the iid and MW0 models (which have the same point forecasts). The random walk and MW1 models have noticeably higher RMSPEs and MAPEs relative to the iid and MW0 models.¹⁹ Hence, practitioners can use the AR(1) and MWd models to achieve good calibration (as documented in Subsection 5.1) without sacrificing point forecast accuracy or Winkler loss.

For the non-stationary variables for $h = 10$, the AR(1), MWd, random walk, and MW1 models have low and similar RMSPEs and MAPEs relative to the iid and MW0 models, again paralleling our Winkler score results. Further, the random walk model produces the lowest RMSPEs and MAPEs more often than the other models, again indicating that the random walk model generally performs well for non-stationary variables for $h = 10$. The point forecast results for non-stationary variables for $h = 25$ parallel the Winkler score results less closely. The AR(1) and MWd models have similar RMSPEs and MAPEs relative to the iid and MW0 models, but the random walk and MW1 models have slightly higher relative values. This result indicates that while the random walk model appears reasonably well calibrated for non-stationary variables for $h = 25$, its point forecasts may be less accurate than those from several of our other models.

6. Extensions

In this section, we provide four extensions of our analysis. First, we study a different collection of long-run nominal interest rates from Schmelzing (2020). Second, we extend the iid model to include stochastic volatility, and illustrate the effects on the calibration of forecast intervals. Third, we discuss combining forecasting models and experiment with an equally-weighted combination of our simple time series models. Fourth, we briefly summarize an extension studying the use of real-time data in forecasting US GDP growth. Detailed results for this fourth extension are in our appendix because we found the use of such data made virtually no difference in the performance of forecast intervals or PITs.

6.1. Very long samples of long-term nominal interest rates

For our main analysis discussed above, our arguably non-stationary variables, including nominal long-term interest rates, typically had samples near 150 years. We now study the long-term nominal interest rate data used in Schmelzing (2020), from which we get samples of 195 to 709 years that allow many more non-overlapping forecasts.²⁰ Specifically, we use the following 8 samples: Italy (1310–1699), Italy (1807–2018), United Kingdom (1310–2018), Holland (1400–1700), Holland (1750–1944), Germany (1326–1943), France (1387–2018), and USA (1786–2018).²¹

We do a pseudo out-of-sample analysis on these 8 samples that parallels our analysis above, including estimation of model parameters with both rolling and recursive samples. This analysis gives us 16 coverage rates and collections of PITs for studying calibration. This analysis also gives us 5668 bias-adjusted estimates of ρ_1 and maximum likelihood estimates of d . The IQR of bias-adjusted estimates of ρ_1 is (0.87, 1.00), and the IQR of the maximum likelihood estimates of d is (0.6, 1.0). This range of estimates closely aligns with the estimates for our baseline collection of plausibly non-stationary variables in Table 1.

Table 5 presents results in a fashion parallel to Table 2, with the addition of $\ddagger$ in column (3) to flag results in which the uniformity of PITs is not rejected. Compared to the results for non-stationary variables (column (4)) in Table 2, forecast performance improves for several models. The AR(1), MWd and MW1 models appear to be well calibrated for $h = 10$, and the MW1 model fails to reject correct calibration for both coverage rates and PITs. For $h = 25$ and 50, the MWd model and especially the MW1 model continue to perform well. In contrast to the entries in column (4) of Table 2, the random walk model's forecast intervals no longer perform well: for all three horizons the lower end of the IQRs for RW is at or above 0.68 (though uniformity of PITs is not rejected for $h = 25$).

Good calibration of some models even for the $h = 50$ horizon may reflect the longer span of the time series (up to 700 observations here, as opposed to up to 150 observations in the data set in Section 5). But such an interpretation clearly is speculative. Analysis with additional time series of comparable span is required to evaluate the role of longer samples. The message that we take from Table 5 is that it, too, indicates that it is possible to have well-calibrated forecast intervals over long horizons.

6.2. Stochastic volatility

Some research studying forecast horizons shorter than ours finds benefits of allowing time-varying volatility when producing forecast distributions (e.g. Clark, 2011). The MW0, MW1 and MWd models account for time-varying variances nonparametrically as long as that time variation is stationary and short memory. Our other models do not incorporate time varying variances.

In this subsection, we extend our iid model to incorporate non-stationary stochastic volatility (SV). This SV model assumes that the DGP is $x_t = \mu + u_t$ with $u_t = \sigma_t \epsilon_t$, $\log(\sigma_t^2) = \log(\sigma_{t-1}^2) + e_t$, $e_t \stackrel{iid}{\sim} N(0, 1)$, $\epsilon_t \stackrel{iid}{\sim} N(0, \sigma_e^2)$, and ϵ_t and e_t are independent. As in the iid model, u_t is serially uncorrelated, but now the log of its variance follows a random walk. We adapt code from Exercise 19.1 of Chan et al.

¹⁹ If we measure the quality of point forecasts by bias rather than RMSPE, the random walk and MW1 models do quite well for reasons explained and illustrated in Lunsford and West (forthcoming).

²⁰ We downloaded these data from <https://www.pfsmelzing.me/all-research-and-data> on July 18, 2025. These data are also available at <https://www.bankofengland.co.uk/working-paper/2020/eight-centuries-of-global-real-interest-rates-r-g-and-the-suprasecular-decline-1311-2018>.

²¹ We note that the raw data have long stretches in which interest rates do not fluctuate, such as Italy (1700–1806) and Holland (1701–1749). We remove such stretches that are at least 48 years long and separately study other samples for those countries. Also, we do not use Schmelzing's data for Spain (1800–1936) and Japan (1870–2018), which have smaller sample sizes than in our baseline analysis.

Table 5

Coverage rates of forecast intervals: Schmelzing's (2020) interest rate data.

	(1)	(2)	(3)	(4)
(1)	horizon	model	median coverage	IQR
(2)	10	iid	0.15	(0.12, 0.24)
(3)	10	MW0	0.49	(0.41, 0.64)
(4)	10	AR(1)	0.68 [†]	(0.61, 0.75)
(5)	10	MWd	0.69 [†]	(0.65, 0.78)
(6)	10	RW	0.78	(0.71, 0.89)
(7)	10	MW1	0.71 ^{‡‡}	(0.64, 0.79)
(8)	25	iid	0.09	(0.07, 0.15)
(9)	25	MW0	0.29	(0.25, 0.37)
(10)	25	AR(1)	0.52	(0.46, 0.68)
(11)	25	MWd	0.62 [†]	(0.54, 0.73)
(12)	25	RW	0.75 [‡]	(0.68, 0.90)
(13)	25	MW1	0.69 ^{‡‡}	(0.58, 0.82)
(14)	50	iid	0.06	(0.04, 0.11)
(15)	50	MW0	0.19	(0.11, 0.25)
(16)	50	AR(1)	0.54	(0.34, 0.66)
(17)	50	MWd	0.57 [†]	(0.51, 0.74)
(18)	50	RW	0.79	(0.70, 0.94)
(19)	50	MW1	0.74 ^{‡‡}	(0.66, 0.89)

Notes:

1. In the list of models in column (2), we use the shorthand “RW” for the random walk model.

2. The total number of samples is 16 (8 series times 2 sampling schemes). “Median coverage” refers to the median coverage rates across the 16 samples. IQR is interquartile range, showing the 25th and 75th percentiles of the coverage rates across the 16 samples.

3. [†] indicates models with coverage rates that fail to reject the null hypothesis of correct calibration at the 0.05 level.4. [‡] indicates models with PITs that fail to reject the null hypothesis of correct calibration at the 0.05 level.**Table 6**Forecast interval coverage with stochastic volatility, $h = 10$.

(1)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)
	GDP Growth		Productivity Growth		Equity Returns	
model	median coverage	IQR	median coverage	IQR	median coverage	IQR
(1) iid	0.65	(0.55, 0.71)	0.56	(0.47, 0.72)	0.62	(0.57, 0.73)
(2) SV	0.52	(0.43, 0.59)	0.43	(0.38, 0.47)	0.71	(0.64, 0.80)

Notes:

1. Line (1) presents results for a subset of the forecasts—those for the iid model—that were summarized in row (4), columns (3a) and (3b) in Table 2.

2. The model in line (2) assumes that the data are serially uncorrelated around an unknown mean as in line (1), but forecast intervals were constructed using a stochastic volatility model, as explained in the text.

(2020) to produce 68 percent forecast intervals from this SV model.²² We forecast the distribution of $(x_{\tau+1} + \dots + x_{\tau+h})/h$ conditional on $\{x_1, \dots, x_\tau\}$ by sampling from the posterior of $\{\sigma_1, \dots, \sigma_\tau, \mu, \sigma_e^2\}$ conditional on $\{x_1, \dots, x_\tau\}$ and making use of $e_{\tau+j}$ being iid for $j = 1, \dots, h$ conditional on σ_e^2 and $x_{\tau+j}$ being normal conditional on $\{\mu, \sigma_\tau, e_{\tau+1}, \dots, e_{\tau+j}\}$ for $j = 1, \dots, h$.

To evaluate the change from the iid to the SV model, we use the variables in Table 1 that show little serial correlation, consistent with both the iid and SV models: GDP growth, productivity growth, and equity returns. In forecasts for the iid model that are a subset of those summarized in Table 2, the iid model performed relatively well for these variables.

We show results for the iid and SV models for GDP growth, productivity growth, and equity returns for $h = 10$ in Table 6.²³ Line (1) reflects forecasts that are a subset of those included in the iid model in column (3) of Table 2. Line (2) adds in stochastic volatility. For GDP growth and productivity growth, results degrade when SV is incorporated. For equity returns results perhaps improve slightly:

²² We downloaded the code from https://web.ics.purdue.edu/~jltobias/second_edition/Chapter19/chapt19.html on October 16, 2025. We take all modeling and estimation choices, priors, number of posterior draws, and burn in period from that code.

²³ Results for $h = 25$ are qualitatively similar.

Table 7
Coverage rates of nominal 68 percent forecast intervals: medians and IQRs.

	(1)	(2)	(3a)	(3b)	(4a)	(4b)
(1)			Stationary variables		Non-stationary variables	
(2)			184 samples		88 samples	
(3)	horizon	model	median coverage	IQR	median coverage	IQR
(4)	10	mixture	0.85	(0.81, 0.90)	0.65 [†]	(0.56, 0.72)
(5)	25	mixture	0.87	(0.78, 0.94)	0.53	(0.39, 0.66)
(6)	50	mixture	0.89	(0.74, 0.96)	0.37	(0.30, 0.64)

Notes:

1. The “mixture” model is an equally-weighted average of the normally-distributed forecast distributions from the iid, AR(1), and random walk models.
2. See notes to [Table 2](#).

the median is closer to 0.68, but the IQRs are of equal width. Overall, allowing for SV in the iid model for these three variables does not deliver substantive benefits.

One possible reason for this is that our use of rolling samples already allows variances (and other moments) to change relatively quickly across estimation samples.²⁴ A second possible reason why incorporating SV does not improve results is that our horizons are long averages. It is well known that for financial variables, time variation in volatility tends to be less apparent when data are aggregated over longer and longer time intervals (e.g., [Baillie and Bollerslev, 1989](#), [Poon and Taylor, 1992](#)). This is perhaps because under suitable assumptions about financial or other variables, a central limit theorem will deliver normality for long enough horizons ([Diebold, 1988](#) cited in [Drost and Nijman \(1993\)](#)). Third, it is possible that other models of persistence and time-varying volatility applied to other variables would yield different results.²⁵

6.3. Combining models

We now discuss and experiment with combining the forecast distributions of our forecasting models. As described in [Subsection 2.2](#), the MWd model is a combination of some underlying distributions. Two of the underlying distributions are those for the MW0 and MW1 models. The procedure to combine is Bayesian, and the combination weights on each underlying model depend on a prior of equal weights and the in-sample likelihood of each underlying model. We have seen that the MWd model appears better calibrated than the MW0 or MW1 models for the stationary variables and worse calibrated than the MW1 model for non-stationary variables. Hence, the MWd model is an example of a model combination that helps for some but not all variables.

Because the MWd model already combines frequency domain models, we now experiment with combining our simple time series models. Specifically, we place equal weights on the forecast distributions of the iid, AR(1), and random walk models. [Rossi and Sekhposyan \(2014\)](#) also evaluate the calibration of equally-weighted combinations. Our combination produces a forecast distribution that is a mixture of normal distributions, which we use to compute 68 percent forecast intervals.

[Table 7](#) shows the coverage rates for the mixture model. For arguably stationary variables, the mixture model does not appear well calibrated at any horizon. For arguably non-stationary variables, it appears well calibrated at $h = 10$ but not at any other horizon.

A recent survey on forecast combination, [Wang et al. \(2023\)](#), points out that combining distributions might not lead to forecast improvements in the same way that combining point forecasts often does. Combining distributions with different means can cause the variance of the mixture to increase. This effect seems to plausibly explain the poor calibration for stationary variables: forecast intervals for these variables have coverage rates that are above target at all horizons (columns (3a) and (3b) in [Table 7](#)). This logic also implies that combining densities with low coverage rates, such as the iid and AR(1) models for non-stationary variables, can lead to a better calibrated mixture. For the non-stationary variables, we indeed see that the mixture is well calibrated for $h = 10$. However, this good calibration does not persist at longer horizons.

6.4. Data revisions

We use real-time data on US GDP growth to consider the possible effects of data revisions on forecast interval coverage and on PITs. We find that using real-time has no meaningful effect on our US GDP results. Hence, we provide a brief summary of results here. We describe data sources and give detailed results in our online appendix.²⁶

²⁴ Like all of our results, results in [Table 6](#) reflect both rolling and recursive samples.

²⁵ A variety of time-varying volatility models exist. [Clark and Ravazzolo \(2015\)](#) and [Chan and Grant \(2016\)](#) study different stochastic volatility and autoregressive conditional heteroskedasticity models when producing forecasts. Random walk stochastic volatility was also favored for macroeconomic variables by [Clark and Ravazzolo \(2015\)](#).

²⁶ More formally, our real-time data use both US GNP and GDP; however, we simply refer to GDP here for ease of exposition. See our online appendix for details.

Let us use “JST” to denote the [Jordà et al. \(2017\)](#) database used in [Section 5](#). We use “real-time” to denote data that combines real-time data on US GDP, 1947–2011, with pre-1947 data from JST. The first vintage for our real-time data was from 1965, and spanned 1947–1965; the last, from 2011, spanned 1947–2011. We made $P = 47$ real-time forecasts at a horizon of $h = 10$. In evaluating forecast accuracy, we took JST data to be the final version of GDP growth. We compared coverage rates and PITs using real-time data to coverage rates and PITs using JST.²⁷ There were 12 comparisons, covering 6 models and two sampling schemes (rolling and recursive).

Across the 12 comparisons, the maximum discrepancy in coverage rates between real-time and JST data was 0.02. Almost all coverage rates were identical to two digits (e.g., for the iid model, coverage was 0.77 both in the real-time and JST data). Across the 12 comparisons, the correlation between PITs for real-time and JST forecasts was near perfect, ranging from 0.96 to 1.0.

7. Conclusions

Long-horizon forecasts are important for economic policy, professional forecasters, climate research, and financial markets. We study the forecast distributions of six univariate forecasting models – three simple time series models and three frequency domain models – for horizons out to 50 years. We find that it is possible to have well-calibrated forecasts for horizons of up to 25 but not 50 years.

We use long-run annual data for 10 macroeconomic variables with up to 18 countries per variable, allowing us to construct pseudo out-of-sample forecasts for 136 different series with widely varying degrees of persistence. For plausibly stationary variables, an AR(1) model and the [Müller and Watson \(2016\)](#) MWd model appear reasonably well calibrated for forecast horizons of both 10 and 25 years. For plausibly non-stationary variables, a random walk model appears reasonably well calibrated for forecast horizons of both 10 and 25 years. Very long samples of nominal interest rate data from [Schmelzing \(2020\)](#) indicate that the [Müller and Watson \(2016\)](#) MW1 model may also be well calibrated for plausibly non-stationary data.

A priority for future research is considering multivariate models and models that incorporate structural change.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Kenneth D. West reports financial support was provided by US Social Security Administration. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are grateful to two anonymous referees, an Associate Editor, conference and seminar participants at the 2024 NBER-NSF Time Series Conference, the Federal Reserve Banks of Cleveland and New York, the Bank of England, the 2023 International Symposium on Forecasting, and the University of East Anglia. West thanks the US Social Security Administration (SSA) for financial support. The views, opinions, and conclusions expressed are solely those of the authors and do not represent the views, opinions, or policy of the SSA, any agency of the federal government, the Federal Reserve Bank of Cleveland, the Federal Reserve System, or NBER. Neither the United States government nor any agency thereof, nor any of its employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of the contents of this report. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply endorsement, recommendation, or favoring by the United States government or any agency thereof.

Supplementary material

Supplementary material associated with this article can be found in the online version at [10.1016/j.jeconom.2026.106221](https://doi.org/10.1016/j.jeconom.2026.106221)

References

- Baillie, R.T., Bollerslev, T., 1989. The message in daily exchange rates: a conditional-variance tale. *J. Business Econ. Stat.* 7 (3), 297–305. <https://doi.org/10.1198/073500102753410390>
- Bergeaud, A., Cetté, G., Lecat, R., 2016. Productivity trends in advanced countries between 1890 and 2012. *Rev. Income Wealth* 62 (3), 420–444. <https://doi.org/10.1111/roiw.12185>
- Brehmer, J.R., Gneiting, T., 2021. Scoring interval forecasts: equal-tailed, shortest, and modal interval. *Bernoulli* 27 (3), 1993–2010. <https://doi.org/10.3150/20-BEJ1298>
- Brockwell, P.J., Davis, R.A., 1991. *Time Series: Theory and Methods*. Springer. <https://doi.org/10.1007/978-1-4419-0320-4>
- Chambers, D., Dimson, E., Ilmanen, A., Rintamäki, P., 2024. Long-run asset returns. *Annual Rev. Finan. Econ.* 16, 435–458. <https://doi.org/10.1146/annurev-financial-082123-105515>
- Chan, J., Koop, G., Poirier, D.J., Tobias, J.L., 2020. *Bayesian Econometric Methods, Second Edition*. Cambridge: Cambridge University Press.

²⁷ Since GDP growth is one of our stationary variables, results for JST considered here are a component of results reported for stationary variables, $h = 10$, in [Table 2](#) and [Fig. 1](#).

- Chan, J. C.C., Grant, A.L., 2016. Modeling energy price dynamics: GARCH versus stochastic volatility. *Energy Econ.* 54, 182–189. <https://doi.org/10.1016/j.eneco.2015.12.003>
- Christensen, P., Gillingham, K., Nordhaus, W., 2018. Uncertainty in forecasts of long-run economic growth. *Proc. National Acad. Sci.* 115 (21), 5409–5414. <https://doi.org/10.1073/pnas.1713628115>
- Chudý, M., Karmakar, S., Wu, W.B., 2020. Long-term prediction intervals of economic time series. *Empir. Econ.* 58 (1), 191–222. <https://doi.org/10.1007/s00181-019-01689-2>
- Clark, T.E., 2011. Real-time density forecasts from bayesian vector autoregressions with stochastic volatility. *J. Business Econ. Stat.* 29 (3), 327–341. <https://doi.org/10.1198/jbes.2010.09248>
- Clark, T.E., Ravazzolo, F., 2015. Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *J. Appl. Econom.* 30, 551–575. <https://doi.org/10.1002/jae.2379>
- D'Agostino, A., Gambetti, L., Giannone, D., 2013. Macroeconomic forecasting and structural change. *J. Appl. Econom.* 28, 82–101. <https://doi.org/10.1002/jae.1257>
- Dawid, A.P., 1984. Statistical theory: the prequential approach. *J. Royal Stat. Soc.: Series A (General)* 147 (2), 278–292. <https://doi.org/10.2307/2981683>
- Diebold, F.X., 1988. *Empirical Modeling of Exchange Rate Dynamics*. New York and Berlin: Springer-Verlag. [https://www.sas.upenn.edu/~fdiebold/papers/paper83/Diebold%20\(1987\)%20Lecture%20Notes.pdf](https://www.sas.upenn.edu/~fdiebold/papers/paper83/Diebold%20(1987)%20Lecture%20Notes.pdf)
- Diebold, F.X., Gunther, T.A., Tay, A.S., 1998. Evaluating density forecasts with applications to financial risk management. *Int. Econ. Rev.* 39 (4), 863–883. <https://doi.org/10.2307/2527342>
- Driscoll, J.C., Kraay, A.C., 1998. Consistent covariance matrix estimation with spatially dependent panel data. *Rev. Econ. Stat.* 80 (4), 549–560. <https://doi.org/10.1162/003465398557825>
- Drost, F.C., Nijman, T.E., 1993. Temporal aggregation of GARCH processes. *Econometrica* 61, 909–927. <https://doi.org/10.2307/2951767>
- Farmer, L.E., Nakamura, E., Steinsson, J., 2024. Learning about the long run. *J. Political Econ.* 132 (10), 3334–3377. <https://doi.org/10.1086/730207>
- Giannone, D., Lenza, M., Primiceri, G.E., 2019. Priors for the long run. *J. Am. Stat. Assoc.* 114 (526), 565–580. <https://doi.org/10.1080/01621459.2018.1483826>
- Gneiting, T., 2011. Making and evaluating point forecasts. *J. Am. Stat. Assoc.* 106 (494), 746–762. <https://doi.org/10.1198/jasa.2011.r10138>
- Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. *J. Royal Stat. Soc. Series B: Stat. Methodol.* 69 (2), 243–268. <https://doi.org/10.1111/j.1467-9868.2007.00587.x>
- Gneiting, T., Katzfuss, M., 2014. Probabilistic forecasting. *Annu. Rev. Stat. Appl.* 1, 125–151. <https://doi.org/10.1146/annurev-statistics-062713-085831>
- Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* 102 (477), 359–378. <https://doi.org/10.1198/016214506000001437>
- Granger, C. W.J., Jeon, Y., 2007. Long-term forecasting and evaluation. *Int. J. Forecast.* 23 (4), 539–551. <https://doi.org/10.1016/j.ijforecast.2007.07.002>
- Hau, B., Pennacchi, G., Ritchken, P., 2012. Inflation expectations, real rates, and risk premia: evidence from inflation swaps. *Rev. Finan. Stud.* 25 (5), 1588–1629. <https://doi.org/10.1093/rfs/hhs003>
- Jordà, O., Knoll, K., Kuvshinov, D., Schularick, M., Taylor, A.M., 2019. The rate of return on everything, 1870–2015. *Quart. J. Econ.* 134 (3), 1225–1298. <https://doi.org/10.1093/qje/qjz012>
- Jordà, O., Schularick, M., Taylor, A.M., 2017. Macrofinancial history and the new business cycle facts. In: Eichenbaum, M., Parker, J.A. (Eds.), *NBER Macroeconomics Annual 2016*. University of Chicago Press, pp. 213–263. <https://doi.org/10.1086/690241>
- Karmakar, S., Chudý, M., Wu, W.B., 2022. Long-term prediction intervals with many covariates. *J. Time Series Anal.* 43 (4), 587–609. <https://doi.org/10.1111/jtsa.12629>
- Kiefer, N.M., Vogelsang, T.J., 2002. Heteroskedasticity-autocorrelation robust standard errors using the bartlett kernel without truncation. *Econometrica* 70 (5), 2093–2095. <https://doi.org/10.1111/1468-0262.00366>
- Kiefer, N.M., Vogelsang, T.J., 2005. A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econ. Theory* 21 (6), 1130–1164. <https://doi.org/10.1017/S0266466605050565>
- Kiefer, N.M., Vogelsang, T.J., Bunzel, H., 2000. Simple robust testing of regression hypotheses. *Econometrica* 68 (3), 695–714. <https://doi.org/10.1111/1468-0262.00128>
- Kitsul, Y., Wright, J.H., 2013. The economics of options-implied inflation probability density functions. *J. Financ. Econ.* 110 (3), 696–711. <https://doi.org/10.1016/j.jfineco.2013.08.013>
- Knüppel, M., 2015. Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *J. Business Econ. Stat.* 33 (2), 270–281. <https://doi.org/10.1080/07350015.2014.948175>
- Lazarus, E., Lewis, D.J., Stock, J.H., Watson, M.W., 2018. Har inference: recommendations for practice. *J. Business Econ. Stat.* 36 (4), 541–559. <https://doi.org/10.1080/07350015.2018.1506926>
- Lunsford, K.G., West, K.D., 2019. Some evidence on secular drivers of US safe real rates. *Am. Econ. J.: Macroecon.* 11 (4), 113–139. <https://doi.org/10.1257/mac.20180005>
- Lunsford, K.G., West, K.D., 2021. An empirical evaluation of some long-horizon macroeconomic forecasts. NBER Retirement and Disability Research Center Paper No. NB21-18. URL <https://www.nber.org/programs-projects/projects-and-centers/retirement-and-disability-research-center/papers/nb21-18>.
- Lunsford, K.G., West, K.D., forthcoming. Random walk forecasts of stationary processes have low bias. *J. Business Econ. Stat.* URL <https://doi.org/10.1080/07350015.2025.2541724>.
- Makridakis, S., Hyndman, R.J., Petropoulos, F., 2020. Forecasting in social settings: the state of the art. *Int. J. Forecast.* 36 (1), 15–28. <https://doi.org/10.1016/j.ijforecast.2019.05.011>
- Müller, U.K., Stock, J.H., Watson, M.W., 2022. An econometric model of international growth dynamics for long-horizon forecasting. *Rev. Econ. Stat.* 104 (5), 857–876. https://doi.org/10.1162/rest_a_00997
- Müller, U.K., Watson, M.W., 2016. Measuring uncertainty about long-run predictions. *Rev. Econ. Stud.* 83 (4), 1711–1740. <https://doi.org/10.1093/restud/rdw003>
- Müller, U.K., Watson, M.W., 2018. Long-run covariability. *Econometrica* 86 (3), 775–804. <https://doi.org/10.3982/ECTA15047>
- Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55 (3), 703–708. <https://doi.org/10.2307/1913610>
- Nordhaus, W., 2018. Projections and uncertainties about climate change in an era of minimal climate policies. *Am. Econ. J.: Econ. Policy* 10 (3), 333–360. <https://doi.org/10.1257/pol.20170046>
- Pascual, L., Romo, J., Ruiz, E., 2004. Bootstrap predictive inference for ARIMA processes. *J. Time Series Anal.* 25 (4), 449–465. <https://doi.org/10.1111/j.1467-9892.2004.01713.x>
- Pesaran, M., Petteuzzo, D., Timmermann, A., 2006. Forecasting time series subject to multiple structural breaks. *Rev. Econ. Stud.* 73 (4), 1057–1084. <https://doi.org/10.1111/j.1467-937X.2006.00408.x>
- Poon, S.-H., Taylor, S.J., 1992. Stock returns and volatility: an empirical study of the UK stock market. *J. Bank. Finance* 16 (1), 37–59. [https://doi.org/10.1016/0378-4266\(92\)90077-D](https://doi.org/10.1016/0378-4266(92)90077-D)
- Rossi, B., Sekhposyan, T., 2014. Evaluating predictive densities of US output growth and inflation in a large macroeconomic data set. *Int. J. Forecast.* 30 (3), 662–682. <https://doi.org/10.1016/j.ijforecast.2013.03.005>
- Schmelzing, P., 2020. Eight centuries of global real interest rates, r-g, and the ‘suprasecular’ decline, 1311–2018. Bank of England Staff Working Paper No. 845. URL <https://www.bankofengland.co.uk/working-paper/2020/eight-centuries-of-global-real-interest-rates-r-g-and-the-suprasecular-decline-1311-2018>.
- Sun, Y., 2014. Let's fix it: fixed-b asymptotics versus small-b asymptotics in heteroskedasticity and autocorrelation robust inference. *J. Econom.* 178 (3), 659–677. <https://doi.org/10.1016/j.jeconom.2013.10.001>
- Wang, X., Hyndman, R.J., Li, F., Kang, Y., 2023. Forecast combinations: an over 50-year review. *Int. J. Forecast.* 39 (4), 1518–1547. <https://doi.org/10.1016/j.ijforecast.2022.11.005>

- West, K.D., 2006. Forecast evaluation. In: Elliott, G., Granger, C. W.J., Timmermann, A. (Eds.), *Handbook of Economic Forecasting*, Volume 1. Elsevier B.V. chapter 3, pp. 99–134. [https://doi.org/10.1016/S1574-0706\(05\)01003-7](https://doi.org/10.1016/S1574-0706(05)01003-7)
- Winkler, R.L., 1972. A decision-theoretic approach to interval estimation. *J. Am. Stat. Assoc.* 67 (337), 187–191. <https://doi.org/10.1080/01621459.1972.10481224>
- Yamamoto, T., Kunitomo, N., 1984. Asymptotic bias of the least squares estimator for multivariate autoregressive models. *Ann. Inst. Stat. Math.* 36, 419–430. <https://doi.org/10.1007/BF02481980>
- Zhou, Z., Xu, Z., Wu, W.B., 2010. Long-term prediction intervals of time series. *IEEE Trans. Inf. Theory* 56 (3), 1436–1446. <https://doi.org/10.1109/TIT.2009.2039158>