An Empirical Evaluation of Some Long-Horizon Macroeconomic Forecasts[*]

Kurt G. Lunsford[†] Kenneth D. West[‡]

November 2021

Abstract

We use long-run annual cross-country data to evaluate pseudo out-of-sample forecasts of five variables for horizons up to 50 years. The variables we forecast are real per capita GDP growth, CPI inflation, labor productivity growth, and long- and short-term nominal interest rates. Our models for forecasting include simple time series models and frequency domain methods recently developed in Müller and Watson (2016). We focus on coverage of 68% forecast intervals (that is, coverage of 68% confidence intervals for forecasts). For GDP growth, CPI inflation and labor productivity growth, median coverage across countries is roughly 68% for several models, but with considerable dispersion around that median. For these three series, a reasonable model choice is a frequency domain model that does not require the user to take a stand on the order of integration of the data. For interest rates, forecast intervals for most models and samples include markedly fewer than 68% of realized values. For interest rates, a reasonable model choice is a driftless random walk. For real per capita GDP and labor productivity growth, we find that forecasts and forecast intervals from the best-performing models are very similar to the Social Security Administration's (SSA's) long-run projections. In contrast, for CPI inflation and long-term interest rates, we find that forecasts from the best-performing models have wider forecast intervals than intervals implied by the SSA's projections for their low- and high-cost scenarios.

Keywords: Fractional Integration, Forecast Interval, Low Frequency, Social Security

JEL Codes: C22, C53, E17, H55

[*] We thank Tu Cao for excellent research assistance. The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Cleveland or the Federal Reserve System. The research reported herein was performed pursuant to grant RDR18000003 from the US Social Security Administration (SSA) funded as part of the Retirement and Disability Research Consortium. The opinions and conclusions expressed are solely those of the authors and do not represent the opinions or policy of SSA, any agency of the Federal Government, or NBER. Neither the United States Government nor any agency thereof, nor any of their 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.

[†] Federal Reserve Bank of Cleveland, P.O. Box 6387, Cleveland, OH 44101; kurt.lunsford@clev.frb.org

[‡] Department of Economics, University of Wisconsin, Madison, WI 53706 and National Bureau of Economic Research; kdwest@wisc.edu

1 Introduction

Many public policy issues require judgement about long-run economic outcomes. The long-run feasibility of current policies relating to government spending and taxation clearly depends on future prospects for variables such as growth, productivity and interest rates on government debt. So, too, does the desirability of alternative short-term policies, some of which might treat current generations in a way that is not feasible to sustain, or alternatively, might short-sightedly sacrifice the welfare of current generations in favor of future generations.

Recognizing the centrality of long-run economic outcomes, the U.S. Social Security Administration (henceforth SSA) makes projections up to 75 years ahead.[1] These projections are not the product of a formal econometric model, and, for the most part, are not accompanied by formal measures of uncertainty such as confidence intervals around the projections. While that approach has considerable appeal, it is not the only possible way to quantify future outcomes.

In this paper, we lay the groundwork to complementing the approach of the SSA. Using data from 23 mostly developed countries (Bergeaud, Cette, and Lecat, 2016; Jordà, Schularick, and Taylor, 2017), we construct and evaluate forecasts and forecast intervals up to 50 years ahead for five variables. The variables are annual per capita GDP growth, CPI inflation, labor productivity growth, and long- and short-term interest rates. Depending on the series and the country, the start date is usually in the 1870s but sometimes is as late as 1891. We start our forecasting exercise between 1918 and 1939, leaving us, for each variable in each country, about a century's worth of data to use to evaluate predictions. We construct forecasts and forecast intervals with both simple time series models and recently developed frequency domain models (Müller and Watson (2016)–henceforth MW). All of our models are univariate. We construct pseudo out-of-sample forecasts and forecast intervals for the average values of each of our variables in each of our countries over the next 10, 25 and 50 years.

We focus on intervals with 68% nominal coverage. Across samples–that is, across variables and across countries–the ideal outcome of course would be to find that about 68% of realized values fall within the 68% forecast intervals in sample after sample. For convenience of exposition, let us interpret "about 68%" as 68% ± 10%. Unfortunately, it is only in a minority of samples that "about 68%" of the realized values fall within our forecast intervals. Depending on the variable and the model, sometimes distinctly more than 68% of the realized average values fall in the 68%

[1] According to Dev (2015), Social Security has been making projections for horizons well beyond 5 years since 1956; the 75 year horizon has been in place since 1992. The *The 2021 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds*, also called the OASDI Trustees Report, which contains the most recent projections at the time of this writing, is available at `https://www.ssa.gov/OACT/TR/2021/`.

intervals, and sometimes distinctly fewer.

For the three variables generally modeled as stationary (per capita GDP growth, inflation, productivity growth), median coverage is about 68%, with as many samples including too many realized values as samples including too few realized values. For these three variables, a reasonable choice of a model is a frequency domain model that does not require one to specify whether the data are stationary or have a unit root or fall somewhere in between. For the two interest rate series, our models generally deliver forecast intervals that are too narrow, with many more samples having empirical coverage less than 68% than greater than 68%. An exception to the previous sentence is the random walk model (one of our simple time series models) for short-term interest rates, which deliver roughly as many samples with too many as with too few realized values in its 68% intervals.

We use the results to interpret 25-year-ahead projections for real per capita GDP growth, labor productivity growth, CPI inflation and long-term interest rates from the 2021 Report from the Trustees of the Social Security Administration.[2] The Report's "intermediate" projections are accompanied by alternative projections under what are called "low-cost" and "high-cost" scenarios. We ask the question: if our forecasting models are used to inform the SSA's projections, what would they tell us? In answering this question, we compare the projections of low- and high-cost scenarios to the ends of 68% forecast intervals produced by our models.

For GDP and productivity growth, SSA projections are basically in agreement with our models. For CPI inflation, our point forecasts tend to be slightly higher than the SSA projection and our 68% forecast intervals are notably wider than the difference between the low- and high-cost projections. For long-term interest rates, our point forecasts are distinctly below the SSA projection and, as is the case for CPI inflation, our 68% forecast intervals are wider than that implied by low- and high-cost projections.

Taking our forecasts and 68% forecast intervals at face value, this suggests that the U.S. economy is likelier to breach the bounds of the low-cost and high-cost projections for CPI than for GDP or productivity growth, and still more likely to breach the bounds for interest rates. In addition, there is substantial probability that long-term interest rates will be lower than projected by SSA.

To our knowledge, the vast majority of theoretical and empirical work on forecasting considers horizons shorter than even our shortest horizon of 10 years. 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

[2] Formally, this is *The 2021 Annual Report* referenced in the previous footnote. We do not compare our results for short-term interest rates because this document does not seem to include projections of short-term rates.

realized GDP almost always falls within 90% confidence intervals. We, too, construct some of our forecasts with very simple parametric models.[3] Chudý, Karmakar, and Wu (2020) evaluate the long-horizon forecast intervals of Pascual, Romo, and Ruiz (2004), Zhou, Xu, and Wu (2010) and Müller and Watson (2016). Their data are daily, so "long-horizon" means many days ahead. They find mixed results for 90% forecast intervals, with more accurate coverage at shorter rather than longer horizons. To the best of our knowledge, we are the first evaluate long-horizon forecasting with annual macroeconomic data.

Some studies have analyzed the ex-post performance of long horizon projections of the SSA (Dev, 2015; Kashin, King, and Soneji, 2015). Like those studies, we use the difference between low- and high-cost projections as a measure of SSA uncertainty. Unlike those studies, our ex-post evaluation is applied to our models rather than to SSA projections, and we contrast up to date forecasts of our models with up-to-date SSA projections. Stock (2019) also presents up-to-date forecasts of some variables of central interest to the SSA.

We stated above that we merely "lay the groundwork" for an alternative approach to long-horizon forecasts. This is partly because the performance of our forecast intervals was not ideal, as is often the case in simulations in the closely related literature on estimation of long run variances (e.g., Vogelsang (2018)). As well, we have only considered five variables, and we focus on forecast interval coverage. We only briefly study the accuracy of point forecasts, leaving more thorough analysis of point forecasts for future research.

Section 2 of the paper describes our models, section 3 our data, and section 4 the mechanics of our forecasting analysis. Section 5 presents forecast evaluation results. Section 6 compares our forecasts and forecast intervals to those implied by the projections of the SSA. Section 7 concludes. The appendix has some technical details.

2 Forecasting Models

The forecasts that we construct and evaluate in this paper are long-horizon averages of an economic variable, x_t. We impose two restrictions on all of the forecasting models used in this paper. First, all of the models will be univariate. That is, we will only use the available data sample of the variable of interest, $\{x_1, \dots, x_T\}$, to forecast the long-horizon average, $(x_{T+1} + \cdots + x_{T+h})/h$. Second, we will assume that x_t has no deterministic time trend. In practice, this will mean forecasting growth rates – not levels – of trending variables, including per capita gross domestic product (GDP), labor productivity, and the consumer price index (CPI). However, we will forecast nominal long- and

[3] 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.

short-term interest rates in levels.

In this paper, we compute point forecasts and use the notation $f_{T,h}$ to denote the forecast made at time T of horizon h. We also compute 68% forecast intervals, which are intended to give a measure of uncertainty around the point forecast. We will interpret these intervals as saying, "the probability that $(x_{T+1} + \cdots + x_{T+h})/h$ is contained in the forecast interval is 0.68."

We use six models to make our forecasts: an independent and identically distributed (iid) model, a random walk (RW) model with iid innovations, an autoregressive model of order one (AR(1)) with iid innovations, MW's integrated of order zero (MW0) model, MW's integrated of order one (MW1) model, and MW's fractionally integrated (MWd) model. For each variable we use either the iid or random walk model but not both. We use the iid model for data generally considered stationary and the random walk model for interest rates. Thus, each variable is forecast by five models: AR(1), MW0, MW1, MWd, and either iid or RW.

The iid, RW and AR(1) models are relatively simple and put a good amount of structure on the data. We describe these models in subsection 2.1. The MW0, MW1, and MWd models require more exposition, and we describe these models in 2.2.

2.1 Simple Forecasting Models

We now describe the iid, random walk, and AR(1) forecasting models. For these three models, we treat the future realization of the data, $(x_{T+1}+\cdots+x_{T+h})/h$, as normally distributed with a mean of the point forecast, $f_{T,h}$, and a variance, $V_{T,h}$. That is, we use $(x_{T+1}+\cdots+x_{T+h})/h \sim N(f_{T,h}, V_{T,h})$, and the three different models will give different forms for $f_{T,h}$ and $V_{T,h}$. We provide derivations in the appendix.

The iid Model. Our first model assumes that the data generating process (DGP) for x_t is

$$x_t = \mu + u_t, \tag{2.1}$$

in which u_t is iid with a mean of zero and variance of σ_u^2. We compute the estimates of μ and σ_u^2 with $\hat{\mu} = T^{-1}\sum_{t=1}^{T} x_t$ and $\hat{\sigma}_u^2 = T^{-1}\sum_{t=1}^{T}(x_t - \hat{\mu})^2$. We then use $\hat{\mu} = T^{-1}\sum_{t=1}^{T} x_t$, the sample average, as the point forecast

$$f_{T,h}^{iid} = \hat{\mu}. \tag{2.2}$$

The estimated variance of the forecast is given by

$$\hat{V}_{T,h}^{iid} = [(1/h) + (1/T)]\hat{\sigma}_u^2. \tag{2.3}$$

Then, using the normal distribution's ±1 standard deviation around the mean to compute 68% forecast intervals, we have

$$\hat{\mu} \pm \sqrt{[(1/h) + (1/T)]\hat{\sigma}_u^2} \tag{2.4}$$

as our 68% forecast interval.

The Random Walk Model. Our second model assumes that the DGP for x_t is a random walk with no drift

$$x_t = x_{t-1} + u_t, \tag{2.5}$$

in which u_t is iid with a mean of zero and variance of σ_u^2. We compute the estimate of σ_u^2 with $\hat{\sigma}_u^2 = (T-1)^{-1}\sum_{t=2}^{T}(x_t - x_{t-1})^2$. We then use the last observation in the sample, x_T, as the point forecast

$$f_{T,h}^{rw} = x_T \tag{2.6}$$

The estimated variance of the forecast is given by

$$\hat{V}_{T,h}^{rw} = (h+1)(2h+1)\hat{\sigma}_u^2/(6h). \tag{2.7}$$

Then, we use

$$x_T \pm \sqrt{(h+1)(2h+1)\hat{\sigma}_u^2/(6h)} \tag{2.8}$$

as our 68% forecast interval.

The AR(1) Model. Our third model assumes that the DGP for x_t is an AR(1)

$$x_t = \rho_0 + \rho_1 x_{t-1} + u_t, \tag{2.9}$$

in which u_t is iid with a mean of zero and variance of σ_u^2. We estimate ρ_0 and ρ_1 with ordinary least squares, denoting the estimates with $\hat{\rho}_0$ and $\hat{\rho}_1$. We only use the AR(1) model if $|\hat{\rho}_1| < 1$, so that the model implies that x_t is stationary. If we estimate $\hat{\rho}_1 \geq 1$, then we forecast with the random walk model.

If $|\hat{\rho}_1| < 1$, we compute $\hat{u}_t = x_t - \hat{\rho}_0 - \hat{\rho}_1 x_{t-1}$ and $\hat{\sigma}_u^2 = (T-1)^{-1}\sum_{t=2}^{T}\hat{u}_t^2$. Using these estimates, the estimated unconditional mean of x_t is $\hat{\rho}_0/(1-\hat{\rho}_1)$, and we use

$$f_{T,h}^{ar1} = \frac{\hat{\rho}_0}{1-\hat{\rho}_1} + \frac{1}{h}(\hat{\rho}_1 + \hat{\rho}_1^2 + \cdots + \hat{\rho}_1^h)\left(x_T - \frac{\hat{\rho}_0}{1-\hat{\rho}_1}\right) \tag{2.10}$$

as the point forecast. The estimated variance of the forecast is given by

$$\hat{V}_{T,h}^{ar1} = \frac{1}{h^2}[1 + (1 + \hat{\rho}_1)^2 + \cdots + (1 + \hat{\rho}_1 + \cdots + \hat{\rho}_1^h)^2]\hat{\sigma}_u^2. \tag{2.11}$$

Then, we use

$$f_{T,h}^{ar1} \pm \sqrt{\hat{V}_{T,h}^{ar1}}, \tag{2.12}$$

as our 68% forecast interval.

2.2 Müller and Watson's Forecasting Models

The MW forecasting models are all based on extracting long-run patterns from the sample $\{x_1, \ldots, x_T\}$ by using a small number, $q << T$, of slowly cycling cosine waves. Hence, we may also refer to MW's methodology as a "low-frequency" or "frequency domain" methodology.

The tth observation of the jth cosine wave is given by $\psi_{j,t} = \sqrt{2}\cos(\pi j(t - 1/2)/T)$ for $t = 1, \ldots, T$ and $j = 1, \ldots, q$. We show the first four of these cosine waves in Figure 2.1. The first wave completes one cycle in $2T$ periods, the second wave completes one cycle in T periods, the third wave completes on cycle in $2T/3$ periods, and the fourth wave complete one cycle in $T/2$ periods. The general pattern is that the jth wave completes one cycle in $2T/j$ periods. For example, if the sample size is $T = 48$ and the number of cosine waves is $q = 8$, then the first cosine wave completes one cycle in 96 years and the eighth cosine wave completes one cycle in 12 years. These eight cosine waves can then be used to extract long-run patterns in the data that occur between 12 and 96 years.

We extract the long-run patterns in the data with linear projection,

$$x_t = \beta_0 + \beta_1\psi_{1,t} + \cdots + \beta_q\psi_{q,t} + e_t. \tag{2.13}$$

We estimate $\beta_0, \beta_1, \ldots, \beta_q$ with ordinary least squares. Before providing equations for the estimates of $\beta_0, \beta_1, \ldots, \beta_q$, we note three features of the cosine waves. First, they sum to zero over time, $\sum_{t=1}^T \psi_{j,t} = 0$ for $j = 1, \ldots, q$. Second, their squares sum to T over time, $\sum_{t=1}^T \psi_{j,t}^2 = T$ for $j = 1, \ldots, q$. Third, their cross products sum to zero over time, $\sum_{t=1}^T \psi_{j,t}\psi_{k,t} = 0$ for $j \neq k$. With these three features, we have

$$\hat{\beta}_0 = T^{-1}\sum_{t=1}^T x_t, \tag{2.14}$$

which is simply the sample average of x_t. We note that $\hat{\beta}_0$ is equivalent to $\hat{\mu}$ for the iid model.

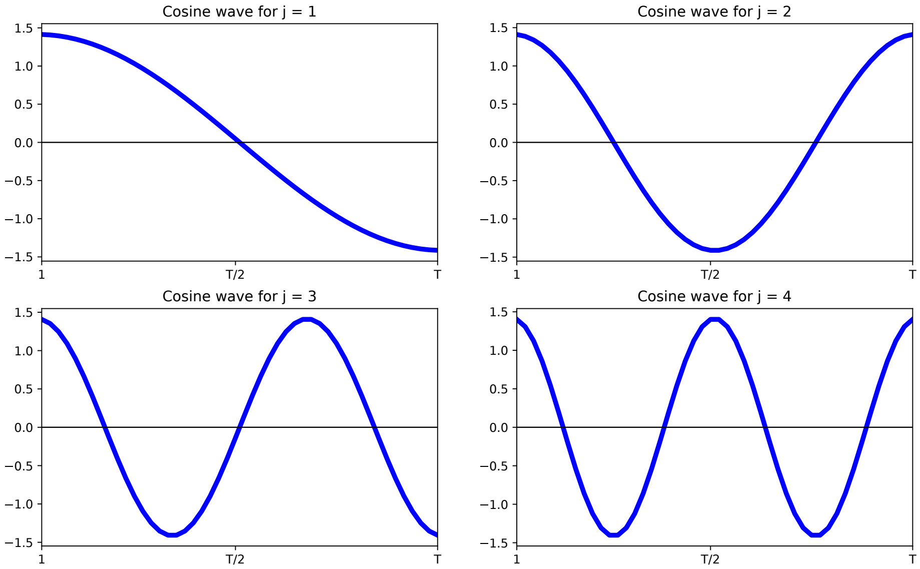


Figure 2.1: The cosine waves for $j = 1, \dots, 4$ for extracting long-run patterns.

Next, we have

$$\hat{\beta}_j = T^{-1} \sum_{t=1}^{T} \sqrt{2} \cos(\pi j (t - 1/2)/T) x_t, \quad \text{for } j = 1, \dots, q. \tag{2.15}$$

Hence, $\hat{\beta}_j$ is a weighted average of x_t in which the jth cosine wave provides the weights.

To illustrate what these long-run patterns look like, we can compute the long-run trend, using the q cosine waves and the estimates in (2.14) and (2.15)

$$\hat{x}_t^{trend} = \hat{\beta}_0 + \hat{\beta}_1 \psi_{1,t} + \cdots + \hat{\beta}_q \psi_{q,t}. \tag{2.16}$$

Figure 2.2 shows this trend for labor productivity. Figure 2.2 has 48 years of annual productivity growth, from 1973 to 2020, and we use $q = 8$ to compute the trend.[4] The trend follows the data but smooths through the year-to-year fluctuations.[5] The trend in Figure 2.2 highlights that labor productivity persistently exceeded its sample average from the mid-1990s to the mid-2000s and that labor productivity was below its sample average from 2010 to 2019.

The estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_q$ summarize the information in the long-run trend and will be used to compute the MW forecasts. For forecasting, we use the notation $\hat{\beta}_{1:q} = [\hat{\beta}_1, \dots, \hat{\beta}_q]'$ to denote the

[4] Labor productivity data are from the supplemental single-year tables to the 2021 SSA Trustees Report. See Section 3.2.

[5] This method of computing a long-run trend is sometimes referred to as “low-frequency filtering” or “band-pass filtering.” As described above, with $T = 48$ and $q = 8$, we are filtering out the frequencies that do not correspond to periods between 12 ad 96 years.

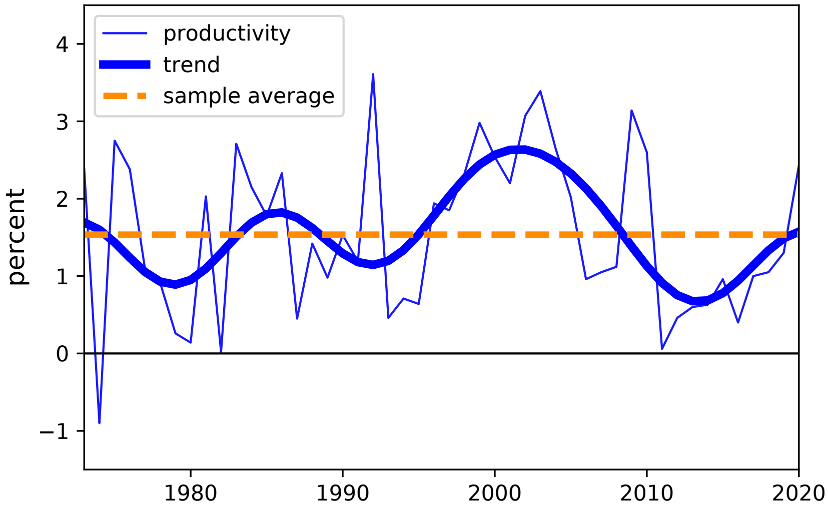


Figure 2.2: Labor productivity growth with its sample average and long-run trend, 1973-2020.

$(q \times 1)$ vector of slope estimates computed with (2.15). We also use $y_{T,h} = (x_{T+1} + \cdots + x_{T+h})/h - \hat{\beta}_0$. We assume that $\lim_{T\to\infty}(h/T) = r > 0$ so that the forecast horizon does not become too small as the sample size increases. Then, MW prove a central limit theorem (CLT)

$$T^{1-\kappa}\begin{bmatrix}\hat{\beta}_{1:q}\\ y_{T,h}\end{bmatrix} \Rightarrow \begin{bmatrix}\tilde{\beta}\\ y\end{bmatrix} \sim N(0,\Sigma), \quad \Sigma = \begin{bmatrix}\Sigma_{\beta\beta} & \Sigma_{\beta y}\\ \Sigma_{y\beta} & \Sigma_{yy}\end{bmatrix}, \tag{2.17}$$

in which κ is a scaling factor that depends on the relevant model for x_t. For the MW0 model, $\kappa = 1/2$. For the MW1 model, $\kappa = 3/2$. For the MWd model, $\kappa = 1/2 + d$. The usefulness of this CLT is that, when T is large, we can treat $T^{1-\kappa}\hat{\beta}_{1:q}$ and $T^{1-\kappa}y_{T,h} = T^{1-\kappa}(x_{T+1} + \cdots + x_{T+h})/h - T^{1-\kappa}\hat{\beta}_0$ as jointly normally distributed, allowing us to compute a distribution for $(x_{T+1} + \cdots + x_{T+h})/h$ given $\hat{\beta}_0$ and $\hat{\beta}_{1:q}$.

In addition to different scaling factors, the form of the covariance matrix, Σ, will be different for each of the MW models. We now discuss forecasting with each of the models.

The MW0 Model. The DGP for x_t is

$$x_t = \mu + u_t, \tag{2.18}$$

in which u_t is a mean zero and $I(0)$ process. By $I(0)$ process, we mean that the dependence between distant observations of u_t is limited so that u_t reverts to 0 relatively quickly.[6] In general, we intend

[6] More formally, we assume that the jth autocovariance of u_t, denoted by $E(u_t u_{t-j})$, has two properties. First, the jth autocovariance only depends on j but not on t. Second, the jth autocovariance goes to zero quickly as j

for the assumption that u_t is $I(0)$ to be flexible, covering cases where u_t is an ARMA process with potentially infinite orders and with potentially non-iid innovations. In other words, we can think of the MW0 model as being similar to but more general than the iid or the AR(1) models.

When u_t is $I(0)$, MW can compute the covariance matrix, Σ, in (2.17) analytically. Let σ^2_{lrv} be the long-run variance of u_t.[7] Then, $\Sigma_{\beta\beta} = \sigma^2_{lrv} I_q$ with I_q being the $(q \times q)$ identity matrix. It is also the case that $\Sigma_{y\beta} = \Sigma'_{\beta y}$ is a $(1 \times q)$ matrix of zeros. Finally, $\Sigma_{yy} = [(1/h) + (1/T)](T\sigma^2_{lrv})$.

Based on this covariance matrix, MW show that $(x_{T+1} + \cdots + x_{T+h})/h$ is a Student-t random variable with q degrees of freedom – not a normally distributed random variable as with the simple forecasting models in Section 2.1. The 68% forecast interval is

$$\hat{\beta}_0 \pm \sqrt{[(1/h) + (1/T)](T\hat{\beta}'_{1:q}\hat{\beta}_{1:q}/q)}\ t^q_{0.84}, \tag{2.19}$$

in which $t^q_{0.84}$ denotes the 0.84 quantile of the Student-t distribution with q degrees of freedom.

When computing forecasts, the values of $t^q_{0.84}$ will change with different values of q. Using the example of $q = 8$, $t^8_{0.84} = 1.06$, which is 6 percent larger than the ±1 standard deviation used in the simple models. Larger values of q will cause $t^q_{0.84}$ to move closer to 1.

We make two observations about the MW0 model. First, the point forecast, $\hat{\beta}_0$, is just the sample average based on (2.14) and is equivalent to $\hat{\mu}$ in the iid model. Thus, the MW0 and iid model give the same point forecast. Second, the only practical difference between the MW0 and iid models is the width of the forecast interval. Comparing the forecast intervals in (2.19) and (2.4), we see that both forecast intervals are scaled by $\sqrt{[(1/h) + (1/T)]}$. Hence, the only difference is that the MW0 model uses $\sqrt{T\hat{\beta}'_{1:q}\hat{\beta}_{1:q}/q}\ t^q_{1-\alpha/2}$ while the iid model uses $\sqrt{\hat{\sigma}^2_u}$.[8]

The MW1 Model. The DGP for x_t is

$$x_t = \mu + u_t, \tag{2.20}$$

in which u_t is a mean zero and $I(1)$ process. By $I(1)$ process, we mean that differences in u_t, $\Delta u_t = u_t - u_{t-1}$, are $I(0)$. We may also refer to u_t as being a "unit root" process. Hence, this model is similar to but generalizes the random walk model.

increases. See Section 2.1 of Stock (1994) for technical assumptions.

[7] The long-run variance is $\sigma^2_{lrv} = E(u_t^2) + 2\sum_{j=1}^{\infty} E(u_t u_{t-j})$. Following the assumptions in Stock (1994), the long-run variance of an $I(0)$ process is finite but non-zero.

[8] These differences have statistical interpretations. $\hat{\sigma}^2_u$ is the estimated variance of x_t in the iid model. When using the iid assumption, $\hat{\sigma}^2_u$ is equivalent to the long-run variance of x_t. In contrast, when u_t, and thus x_t, is an $I(0)$ process, $T\hat{\beta}'_{1:q}\hat{\beta}_{1:q}/q$ is MW's estimate of the long-run variance of x_t. Further, because MW rely only on the estimates $\hat{\beta}_1, \ldots, \hat{\beta}_q$ for computing forecasting intervals, they effectively shrink their sample size to q. Hence, they use Student-t^q quantiles rather than standard normal quantiles.

Unlike when u_t is $I(0)$, MW do not provide an analytical form for every element of Σ when u_t is $I(1)$. Because of this, we use an approximation of Σ, proving formulas in the appendix. As with the MW0 model, MW show that $(x_{T+1} + \cdots + x_{T+h})/h$ is a Student-t random variable with q degrees of freedom – not a normally distributed random variable as with the simple forecasting models in Section 2.1. The forecast interval is

$$\hat{\beta}_0 + \Sigma_{y\beta}\Sigma_{\beta\beta}^{-1}\hat{\beta}_{1:q} \pm \sqrt{(\Sigma_{yy} - \Sigma_{y\beta}\Sigma_{\beta\beta}^{-1}\Sigma_{\beta y})(\hat{\beta}'_{1:q}\Sigma_{\beta\beta}^{-1}\hat{\beta}_{1:q}/q)}\ t^q_{0.84}, \tag{2.21}$$

in which $t^q_{0.84}$ denotes the 0.84 quantile of the Student-t distribution with q degrees of freedom. Using the example of $q = 8$, $t^8_{0.84} = 1.06$.

Comparing (2.19) and (2.21), we see that the point forecasts of the MW0 and MW1 models are different: $f^{MW0}_{T,h} = \hat{\beta}_0$ and $f^{MW1}_{T,h} = \hat{\beta}_0 + \Sigma_{y\beta}\Sigma_{\beta\beta}^{-1}\hat{\beta}_{1:q}$. The additional term for the MW1 model moves the forecast away from the sample average and toward the last value of the long-run trend. Using Figure 2.2 as an example for productivity growth, the MW0 point forecast will be the sample average while the MW1 point forecast will approximately be the value of the trend for 2020 (which happen to be very similar in Figure 2.2).

In addition, while the forecast intervals for the MW0 and MW1 models both use Student-t^q quantiles, they have different scalings that precede these quantiles. As we will see below, it is generally the case that the MW1 model's forecast intervals are wider than the MW0 model's forecast intervals.

The MWd Model. The DGP for x_t is

$$x_t = \mu + u_t, \tag{2.22}$$

in which u_t is a mean zero and $I(d)$ process with $-0.5 < d < 1.5$. This MWd model functions as a more general model than either the MW0 or MW1 models. That is, if we set $d = 0$, then we recover the MW0. If we set $d = 1$, then we recover the MW1 model. However, we may also choose any value of d such that $-0.5 < d < 1.5$.[9]

If d is known, then we can approximate Σ in (2.17) and compute forecasts in manner that parallels the MW1 model. However, in practice, d is not known. Instead, following MW, we use a Bayesian approach. Our prior is that d has an equal probability of being one of the values in $\{-0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8, 1\}$. We then numerically compute the Bayes predictive density as in Section 3.2 of MW. Using $\bar{x}_{T+h} = (x_{T+1} + \cdots x_{T+h})/h$, we denote the Bayes predictive density

[9] This model may also be called the "fractionally integrated" model, referring to the fact that d can be a fraction. See Baillie (1996) for an overview.

with $f^{bayes}(\bar{x}_{T+h}|\hat{\beta}_0, \hat{\beta}_{1:q})$.

With $f^{bayes}(\bar{x}_{T+h}|\hat{\beta}_0, \hat{\beta}_{1:q})$ in hand, we use the expectation or mean as our point forecast

$$f_{T,h}^{MWd} = \int_{-\infty}^{\infty} \bar{x}_{T+h} f^{bayes}(\bar{x}_{T+h}|\hat{\beta}_0, \hat{\beta}_{1:q}) d\bar{x}_{T+h}. \tag{2.23}$$

To compute the 68% forecast intervals, we then find the values of $\underline{b}$ and $\bar{b}$ such that $0.16 = \int_{-\infty}^{\underline{b}} f^{bayes}(\bar{x}_{T+h}|\hat{\beta}_0, \hat{\beta}_{1:q}) d\bar{x}_{T+h}$ and $0.84 = \int_{-\infty}^{\bar{b}} f^{bayes}(\bar{x}_{T+h}|\hat{\beta}_0, \hat{\beta}_{1:q}) d\bar{x}_{T+h}$. The MWd forecast interval is then $\{\underline{b}, \bar{b}\}$.

3 Data

In this section, we describe the data used in this paper. Section 3.1 covers the data for our pseudo out-of-sample forecasting exercises, which are described in Section 4 with results shown in Section 5. Section 3.2 covers the data used to compare the forecast intervals from the forecasting models with the low- and high-cost scenarios in OASDI Trustees Report in Section 6.

3.1 Data for Pseudo Out-of-Sample Analysis

We use two datasets for our pseudo out-of-sample analysis. First is the Macrohistory Database (Jordà, Schularick, and Taylor, 2017). Second is the Long-Term Productivity Database (Bergeaud, Cette, and Lecat, 2016).[10] We compute pseudo out-of-sample forecasts for five variables: per capita real GDP growth, CPI inflation, labor productivity growth, the level of long-term interest rates, and the level of short-term interest rates. We observe all data at an annual frequency. The two databases allow us to observe each variable for many different countries. For each variable, we include a country if it has a long sample with no missing observations. We also extend the CPI and interest rate data to 2020 when possible using data from the OECD and from individual country central bank websites. In the following bullets, we list which countries we use for each variable and their respective samples.

- **per capita real GDP:** We use a panel of 17 countries: AUS, BEL, CAN, CHE, DEU, DNK, ESP, FIN, FRA, GBR, ITA, JPN, NLD, NOR, PRT, SWE, and USA. All countries have data samples from 1871 to 2017.
- **CPI inflation:** We use a panel of 17 countries: AUS, BEL, CAN, CHE, DEU, DNK, ESP, FIN, FRA, GBR, ITA, JPN, NLD, NOR, PRT, SWE, and USA. All countries have data

[10] Each dataset presents only the latest vintage. Thus, our forecasting analysis abstracts from the effects of data revisions, which in any case would likely be small for long-horizon averages.

samples from 1871 to 2020.

- **Labor productivity growth:** We use a panel of 23 countries: AUS, AUT, BEL, CAN, CHE, CHL, DEU, DNK, ESP, FIN, FRA, GBR, GRC, IRL, ITA, JPN, MEX, NLD, NOR, NZL, PRT, SWE, and USA. AUS, AUT, and BEL have data samples from 1891 to 2018. All other countries have data samples from 1891 to 2019.

- **Long-term interest rate level:** We use an unbalanced panel of 15 countries. Due to the nature of the unbalanced panel, we list the data samples for each country: AUS (1870-2020), BEL (1920-2020), CAN (1870-2020), CHE (1919-2020), DNK (1870-2020), FRA (1870-2020), GBR (1870-2020), IRL (1922-2020), ITA (1870-2020), JPN (1870-2020), NLD (1870-2020), NOR (1870-2020), PRT (1870-2020), SWE (1870-2020), USA (1870-2020).

- **Short-term interest rate level:** We use an unbalanced panel of 12 countries. Due to the nature of the unbalanced panel, we list the data samples for each country: BEL (1920-2020), CHE (1870-2020), DNK (1875-2020), FIN (1870-2020), FRA (1922-2020), GBR (1870-2020), IRL (1920-2020), ITA (1922-2020), NLD (1870-2020), PRT (1880-2020), SWE (1870-2020), USA (1870-2020).

3.2 Data for Social Security Comparison

For comparing point and interval forecasts from our models to the SSA's long-run projections, we use data from the supplemental single-year tables to the SSA's 2021 Trustees Report, downloaded from https://www.ssa.gov/OACT/TR/2021/. For per capita real GDP growth, we use total population from Table V.A3 to compute population growth. We then subtract that population growth from real GDP growth in Table V.B2. We pull labor productivity growth and CPI inflation from Table V.B1. We use the nominal average annual interest rate from Table V.B2 as a long-term interest rate.[11] We do not have a short-term interest rate for this analysis.

4 Pseudo Out-of-Sample Analysis

We make forecasts over horizons $h = 10$, 25 and 50 years ahead, with 50 year forecasts made only for samples longer than 125 years. To illustrate the mechanics, suppose that the first observation in the data series is 1871, as is the case for much of our cross-country data. For concreteness, let us

[11] This interest rate is for newly issued Social Security trust fund securities and is the average average of the nominal interest rates for special U.S. Government obligations issuable to the trust fund in each of the 12 months of the year. See page 116 of the 2021 Trustees Report. We find that this interest rate is very close to the annual average of the 10-year Treasury constant maturity rate from the FRED database, https://fred.stlouisfed.org/series/GS10, and so treat it as a long-term interest rate.

suppose that GDP growth is the variable under study and that we are using 68% forecast intervals. As a reminder, for a normal distribution a 68% forecast interval is ±1 standard deviation around the mean. Our cross-country GDP data end in 2017.

Let us use "estimate our models" as shorthand for "estimate parameters that our models need to make forecasts and construct forecast intervals." (As distinguished from, estimating the forecasting performance of our models.) We proceed as follows:

1. Using a 48-year sample running from 1871-1918, estimate each of our models (iid, AR(1), three MW models).

2. Construct forecasts and 68% forecast intervals for average GDP growth for horizons of 10, 25 and 50 years: 1919-1928 ($h = 10$), 1919-1943 ($h = 25$), 1919-1968 ($h = 50$).

3. Using the data for the indicated 10, 25 and 50 year intervals, compute the ex-post forecast error and note whether the ex-post average growth rate is within the 68% forecast interval.

4. Add one year to the end of the sample used to estimate our models. Repeat steps 1-3. Repeat over and over, until the available data are exhausted. For example, for $h = 10$ year ahead forecasts, we forecast and construct forecast intervals for average GDP growth from 1919-1928, then 1920-1929, and so on, up to 2008-2017.

In a sample running 1871-2017, we end up with 90 forecasts for $h = 10$, 75 forecasts for $h = 25$ and 50 forecasts for $h = 50$. As a result, we have 90, 75 or 50 observations on the ex-post average growth rate. To explain some of the statistics we compute from the forecasts and forecast errors, consider $h = 10$. We compute a time series of 90 forecasts errors, running from $T + 10 = 1928$ to $T + 10 = 2017$. Let

$$\bar{x}_{T+10} = (x_{T+1} + \cdots + x_{T+10})/10 = \text{realized 10 year average}, \tag{4.1a}$$

$$\eta_{T+10} = \text{forecast error} = \bar{x}_{T+10} - (\text{forecast made using a sample ending at } T), \tag{4.1b}$$

$$FI_{T+10} = 68\% \text{ forecast interval for } T+10, \text{made using a sample ending at } T. \tag{4.1c}$$

For $T + 10$ running from 1928 to 2017, with a sample size of predictions and realizations $P = 90$, we compute among other statistics:

$$\text{absolute value of bias (|bias|):} \quad \sum\nolimits_T \eta_{T+10}/P \tag{4.2}$$

$$\text{root mean squared forecast error (RMSFE):} \sqrt{\sum\nolimits_T \eta^2_{T+10}/P} \tag{4.3}$$

$$\text{fraction of realized averages that fall within nominal 68\% forecast intervals:} \tag{4.4}$$
$$\left[\sum\nolimits_T \mathbf{1}(\bar{x}_{T+10} \in FI_{T+10})\right]/P.$$

In (4.4), $\mathbf{1}(\bar{x}_{T+10} \in FI_{T+10})$ is an indicator function that takes the value of 1 if realized average growth $\bar{x}_{T+10}$ falls within the forecast interval, and is 0 otherwise. An ideal procedure to construct 68% forecast intervals would yield a value of 0.68 for this statistic.

4.1 Samples and Sampling Schemes

There remains a question of whether one should use distant data in making forecasts and constructing forecast intervals: in estimating models to forecasting average GDP growth 2011-2020, should one use data on GDP growth all the way back to 1871? Distant data allows forecasts to reflect behavior in the distant past that may recur in the future. On the other hand, one might view what happened to GDP growth in (say) World War II as uninformative or perhaps even misleading about prospects for GDP growth in the 21st century.

We allow for both of these positions and in two ways. First, we use two sampling schemes, one of which does use all available data (the *recursive* scheme) and one of which drops the observation that was formerly at the beginning of the sample when another observation is added on at the end (the *rolling* scheme). For both schemes, the sample size used for estimation of our models starts at 48. To illustrate, suppose data run from 1871-2017 and consider the $h = 10$ horizon. For the recursive scheme, the sample size used for estimation grows year by year until it reaches 137. At that point, data for 1871-2007 is used to forecast 2008-2017. For the rolling scheme, the sample size stays fixed at 48, with the initial observation dropped when an year is added at the end. Thus, a sample running 1960-2007 is used to forecast 2008-2017. The final rolling samples for $h = 25$ and $h = 50$ are of course shifted back 15 and 40 years earlier than the final 1960-2007 sample for $h = 10$.

Second, although all our data go back to the 19th century, we repeat all of our estimation using samples that start in 1919 (or early 1920s, depending on data availability). With a sample start of 1919, the first 10-year forecast is for 1967-1976. The choice of 1919 is not entirely arbitrary. In

Table 4.1: Samples for different horizons

	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)
	$h = 10$		$h = 25$		$h = 50$	
data sample	first predicted observation	no. of predictions P	first predicted observation	no. of predictions P	first predicted observation	no. of predictions P
1871-2020	1919-1928	93	1919-1943	78	1919-1968	53
1919-2020	1967-1976	45	1967-1991	30	n.a.	n.a.

Notes: Some series start or end on dates slightly different than 1871, 1919 or 2020. See text for exact dates. The first forecast is always for an h-year period that begins 48 years after the beginning of the sample. For example, for series whose coverage begins in 1870, the first predicted 10 year average is for 1918-1927, and there would be 94, 79 and 54 predictions for the three horizons.

addition to giving us about 100 years of data, such a starting point eliminates some potentially anomalous periods. It omits World War I entirely and makes forecasts start well after the Great Depression and World War II. Because the sample is shorter, we only predict at 10- and 25- but not 50-year horizons.

For rolling samples, forecasts for a given observation (say, average GDP growth 1967-1976) are identical in the complete sample and the sample that ignores data prior to 1919. That is, the forecasts (and forecast errors) in the sample that ignores data prior to 1919 are a subset of those in the complete sample. For recursive samples, forecasts and forecast errors for a given observation are, in general, different in the complete and 1919- samples.

Table 4.1 summarizes information on sample sizes. Per the description of data in the previous section, the start date for some series is not exactly 1871 or 1919 and some series end slightly prior to 2020. Whatever the start and end dates, for a given series in a given country, the number of predictions is the same for the rolling and recursive schemes.

In our view, the number of predictions given in Table 4.1 should be understood in light of the heavily overlapping nature of our forecasts. Despite the sample sizes of 93 / 78 / 53 in the 1871-2020 line in Table 4.1, there are only 9 non-overlapping 10-year periods, 3 non-overlapping 25-year periods and 1 non-overlapping 50-year period. Among other implications, this leads us to follow Müller and Watson (2018) and our own work (Lunsford and West, 2019) in using 68% 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% forecast intervals, requires more data than evaluation of behavior that includes realizations towards the more-frequently-observed center of a distribution.

Of course, even for $h = 50$ we have 77 different samples with one non-overlapping set of 50

Table 4.2: Actual coverage of 68% forecast intervals, US GDP growth, rolling samples

		(1) $h = 10$ (90 forecasts)	(2) $h = 25$ (75 forecasts)	(3) $h = 50$ (50 forecasts)	(4) median across horizons
(1)	iid	72%	80%	82%	80%
(2)	AR(1)	71%	78%	63%	71%
(3)	MW0	71%	80%	74%	74%
(4)	MW1	77%	88%	94%	88%
(5)	MWd	70%	77%	74%	74%
(6)	median across models	71%	80%	74%	77%

Notes:
1. The sample period runs 1871-2017.
2. See the text for definitions of the models.
3. The value of 72% in the iid row, $h = 10$ column indicates that 65 of the 90 realizations of 10 year average GDP growth fall within the iid model's nominal 68% forecast interval. Other entries are defined similarly.

observations.[12] Our hope is that these 77 samples, though highly correlated, will provide enough variation for us to meaningfully evaluate the forecasting performance of our models.

4.2 Reporting of Results

We report results both for quality of forecast intervals for predictions, and of the quality of the predictions relative to one another. These results are aggregated over all countries.

To illustrate how such aggregations are constructed, Table 4.2 reports how our 68% forecast interval coverage works for a single country and sampling scheme: U.S. GDP growth, for rolling samples. To illustrate, consider the 72% figure for iid, $h = 10$ in row (1), column (1). This indicates that 72%, or 65 of the 90 realized values for 10 year average growth, fall within the 68% forecast interval that was constructed using the iid model. (The figures in the table are the fractions in equation (4.4), multiplied by 100 to convert to percentage.) Of course, an ideal figure would be 68%. So slightly more realizations than are ideal fall into the forecast intervals. The 80% and 82% figures for $h = 25$ and $h = 50$ indicate that 60 of 75 ($h = 25$) or 41 of 50 ($h = 50$) realized values fall within the intervals for those horizons.[13]

[12] 77 = GDP growth for 17 countries + inflation for 17 countries + productivity growth for 23 countries + long interest rates for 12 countries + short interest rates for 8 countries. Since this same data is used for other horizons, for $h = 10$, the comparable figure is 9×77 sets of non-overlapping 10-year observations and for $h = 25$ the figure is 3×77.

[13] We noted above that the iid and MW0 models use identical point forecasts. But as can be seen in the table for $h = 50$, the fact that they use different procedures to construct forecast intervals means coverage can be quite different.

In our tables below, we report median coverage over a set of forecasts as one summary statistic of behavior. Table 4.2 involves few enough entries that perhaps no summary statistics are needed. But for illustration we present medians over horizons, over models, and overall over both horizons and models. In columns (1)-(3), the bottom row in the table presents the median across the five models for a given horizon; in rows (1)-(5), the final column presents the median across the three horizons; row (6), column (4) presents the median across the 15 entries in the table.

In the results about to be presented, we aggregate in a way analogous to the aggregation in the final row and final column of Table 4.2. For each of our data series, the basic unit of observation is a set of forecasts for a given country, horizon and sampling scheme. For example, we shall summarize (among other statistics) the behavior of the each model aggregated across horizons. For GDP growth, this means across 102 observations. Here, $102 = 17$ countries × 3 horizons × 2 sampling schemes. (For the sample starting in 1919, our sample includes 68 observations, where $68 = 17$ countries × 2 horizons × 2 sampling schemes.) Thus, in our tables below, for GDP growth, using all available data, and for a given model such as AR(1), median coverage is computed using 102 observations rather than (as in the rightmost column of Table 4.2) using 3 observations.

We also summarize accuracy of coverage via a histogram of coverage. We use four bins. For nominal 68% coverage, we report the percentage of observations in which: coverage is less than 38%; between 38% and 58%; between 58% and 78% (labeled in our graphs as 68% ± 10%); and greater than 78%. A well-performing model will have a pile-up of observations in the 68% ± 10% bin–that is, will generally have coverage that is close to nominal size of 68%. Our use of "within 10%" as defining "close to 68%" is arbitrary, but we think suffices to distinguish well- and poorly-performing models.

We also report histograms after aggregating over models. Figure 4.1 illustrates this when we aggregate over models and report a histogram for the 15 entries in Table 4.2. Since 9 of the 15 entries in the table are within 68% ± 10%, the graph reports 0.6 (=9/15) for the 68% ± 10% entry.

For point estimates, such as |bias| or RMSFE (see equations (4.2) and (4.3)), we look at medians across aggregates, and express the results relative to an arbitrarily chosen baseline model. We also report the percentage of sets of forecasts in which a given model produces the lowest RMSFE.

5 Pseudo Out-of-Sample Results

We present results for coverage of 68% forecast intervals in Section 5.1 and then for |bias| and RMSFE in Section 5.2. In each case, we begin by discussing results for GDP growth, inflation and productivity growth, for the longest possible samples and for samples that begin around 1919. We then discuss the results for interest rates.

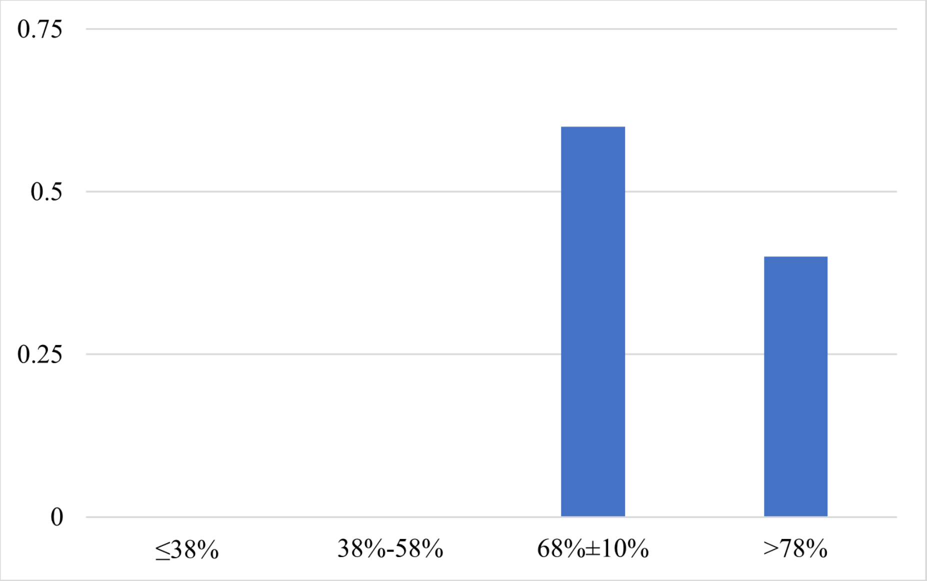


Figure 4.1: US GDP growth: 68% coverage

This is a histogram of the figures in Table 4.2. The vertical axis is the fraction of the 15 entries in Table 4.2 that fall into the indicated bin.

For simplicity, we will refer to GDP growth, inflation and productivity growth as "stationary" variables rather than use the longer but more accurate description "variables often presumed to be stationary." We distinguish these from our interest rate variables, which are often interpreted as displaying behavior suggestive of unit roots.

5.1 Results for Coverage of 68% Forecast Intervals

Table 5.1 has summary information on performance of 68% confidence intervals. The top panel relies on the longest time series available. The bottom panel relies on data starting in 1919 or slightly later. In each panel, for our stationary variables (columns (3)-(5)), the baseline simple time series model is the iid model. We do not compute random walk forecasts for the stationary variables, and show "n.a." in the random walk row. In columns (6) and (7), the baseline simple time series model for interest rates is a random walk. We do not compute iid model forecasts for these variables, and show "n.a." in the iid row.

Overall, the models perform better for stationary variables than for interest rate variables. We discuss stationary and interest rate variables in turn.

For the stationary variables, performance is better in the complete sample going back to the 19th century (panel A). Here, the "median coverage" columns in panel A indicate to us that coverage

Table 5.1: Actual 68% forecast interval coverage: medians and fraction of samples within 10% of nominal coverage

A. Samples starting 1870s or 1891

	(1)	(2)	(3a)	(3b)	(3c)	(4a)	(4b)	(4c)	(5a)	(5b)	(5c)	(6a)	(6b)	(6c)	(7a)	(7b)	(7c)
(1)				GDP growth			CPI inflation			Productivity growth			Long-term interest rates			Short-term interest rates	
(2)				17 countries			17 countries			23 countries			12 countries			8 countries	
(3)	horizon	model	no. of samples	median coverage	fraction 68%±10%	no. of samples	median coverage	fraction 68%±10%	no. of samples	median coverage	fraction 68%±10%	no. of samples	median coverage	fraction 68%±10%	no. of samples	median coverage	fraction 68%±10%
(4)	*all*	*all*	510	70%	0.36	510	69%	0.25	690	71%	0.26	360	31%	0.07	240	34%	0.09
(5)	*all*	RW	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	72	34%	0.13	48	66%	0.35
(6)	*all*	iid	102	61%	0.35	102	36%	0.13	138	61%	0.29	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
(7)	*all*	MW0	102	67%	0.36	102	62%	0.31	138	65%	0.30	72	27%	0.00	48	20%	0.00
(8)	*all*	MW1	102	86%	0.26	102	88%	0.16	138	90%	0.07	72	38%	0.13	48	44%	0.08
(9)	*all*	MWd	102	64%	0.46	102	76%	0.38	138	70%	0.33	72	34%	0.08	48	35%	0.02
(10)	*all*	AR1	102	60%	0.36	102	57%	0.29	138	57%	0.29	72	20%	0.01	48	20%	0.00
(11)	10	*all*	170	72%	0.58	170	75%	0.36	230	75%	0.35	120	43%	0.13	80	43%	0.16
(12)	25	*all*	170	74%	0.31	170	65%	0.29	230	67%	0.25	120	25%	0.02	80	27%	0.04
(13)	50	*all*	170	57%	0.19	170	58%	0.12	230	69%	0.17	120	24%	0.07	80	29%	0.08

B. Samples starting 1919 or 1920s

	(1)	(2)	(3a)	(3b)	(3c)	(4a)	(4b)	(4c)	(5a)	(5b)	(5c)	(6a)	(6b)	(6c)	(7a)	(7b)	(7c)
(1)				GDP growth			CPI inflation			Productivity growth			Long-term interest rates			Short-term interest rates	
(2)				17 countries			17 countries			23 countries			15 countries			12 countries	
(3)	horizon	model	no. of samples	median coverage	fraction 68%±10%	no. of samples	median coverage	fraction 68%±10%	no. of samples	median coverage	fraction 68%±10%	no. of samples	median coverage	fraction 68%±10%	no. of samples	median coverage	fraction 68%±10%
(4)	*all*	*all*	340	85%	0.24	340	73%	0.25	460	80%	0.25	300	30%	0.03	240	34%	0.11
(5)	*all*	RW	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	60	29%	0.10	48	57%	0.35
(6)	*all*	iid	68	74%	0.35	68	30%	0.10	92	66%	0.28	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
(7)	*all*	MW0	68	80%	0.29	68	64%	0.32	92	72%	0.25	60	30%	0.00	48	28%	0.00
(8)	*all*	MW1	68	100%	0.10	68	91%	0.19	92	100%	0.15	60	40%	0.03	48	42%	0.15
(9)	*all*	MWd	68	86%	0.19	68	87%	0.31	92	86%	0.28	60	30%	0.00	48	35%	0.04
(10)	*all*	AR1	68	78%	0.28	68	67%	0.32	92	66%	0.28	60	21%	0.00	48	26%	0.02
(11)	10	*all*	170	79%	0.29	170	75%	0.32	230	78%	0.30	150	38%	0.03	120	41%	0.14
(12)	25	*all*	170	96%	0.19	170	67%	0.18	230	82%	0.20	150	21%	0.03	120	27%	0.08

Notes:

1. For all series, the sample end date is between 2017 and 2020. See text for list of countries, explanation of models and exact sample periods.
2. In the lists of horizons and models in columns (1) and (2), "all" means numbers aggregated over all five models or over all three (panel A) or two (panel B) horizons.
3. In the list of models in column (2), we use the shorthand "RW" and "AR1" for the random walk and AR(1) models.
4. To explain "median coverage," consider the figure of 67% in row (7), column (3b) in panel A. Per column (3a), the MW0 model is used to produce 102 sets of pseudo out-of-sample forecasts of average GDP growth; here, 102 = 17 countries × 3 horizons × 2 sampling schemes. Confidence intervals with nominal 68% coverage are computed for each forecast in each of the 102 sets of forecasts. We show the median value across these 102 sets of forecasts.
5. To explain "fraction 68% ± 10%," continue the example in the previous note by considering the figure of 0.36 in column (3c) in the MW0 row in panel A. Across the 102 sets of forecasts, actual coverage is between 58% and 78% in 37 of the 102 sets, giving the fraction 37/102=0.36.

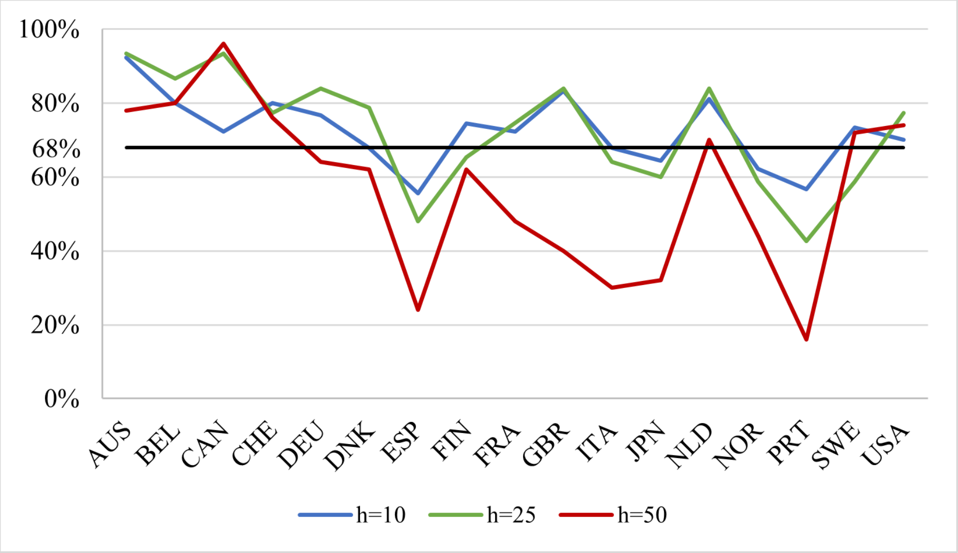


Figure 5.1: 68% coverage for GDP growth, MWd model, rolling scheme

Note: The results plotted here are a subset of those reported in column (3) of panel A of Table 5.1.

tends to be centered tolerably close to the ideal value of 68%. Of the 27 entries for "median coverage" in panel A for the three stationary variables, all but 4 fall between 57% and 76%, and three of the four exceptions are for the MW1 model that is not intended for stationary data.

These results indicate to us that coverage is centered reasonably close to the ideal value of 68%. However, and even putting aside MW1, there is considerable dispersion around these medians. Columns (3c), (4c) and (5c) in panel A indicate that typical values for the fraction of actual coverage that is 68% ± 10% is about 0.25 to 0.35; there are occasional higher or lower values. Put differently, in two thirds or more of the sets of forecasts, actual coverage is not within 10% of nominal 68% coverage. Perhaps unsurprisingly, coverage is better at shorter horizons ($h = 10$ [row (11), panel A]) than at longer horizons ($h = 25$ or $h = 50$). This is true whether on measures quality of coverage by closeness of the median to 68% or by fraction 68% ± 10%.

Figure 5.1 gives some insight into behavior across horizons, as well as variability across countries. For GDP growth, rolling samples, it plots coverage of 68% intervals for each of our 17 countries. Country mnemonics appear in alphabetical order on the horizontal axis. For the USA, the three values plotted are the same as those presented in the MWd row of Table 4.2.

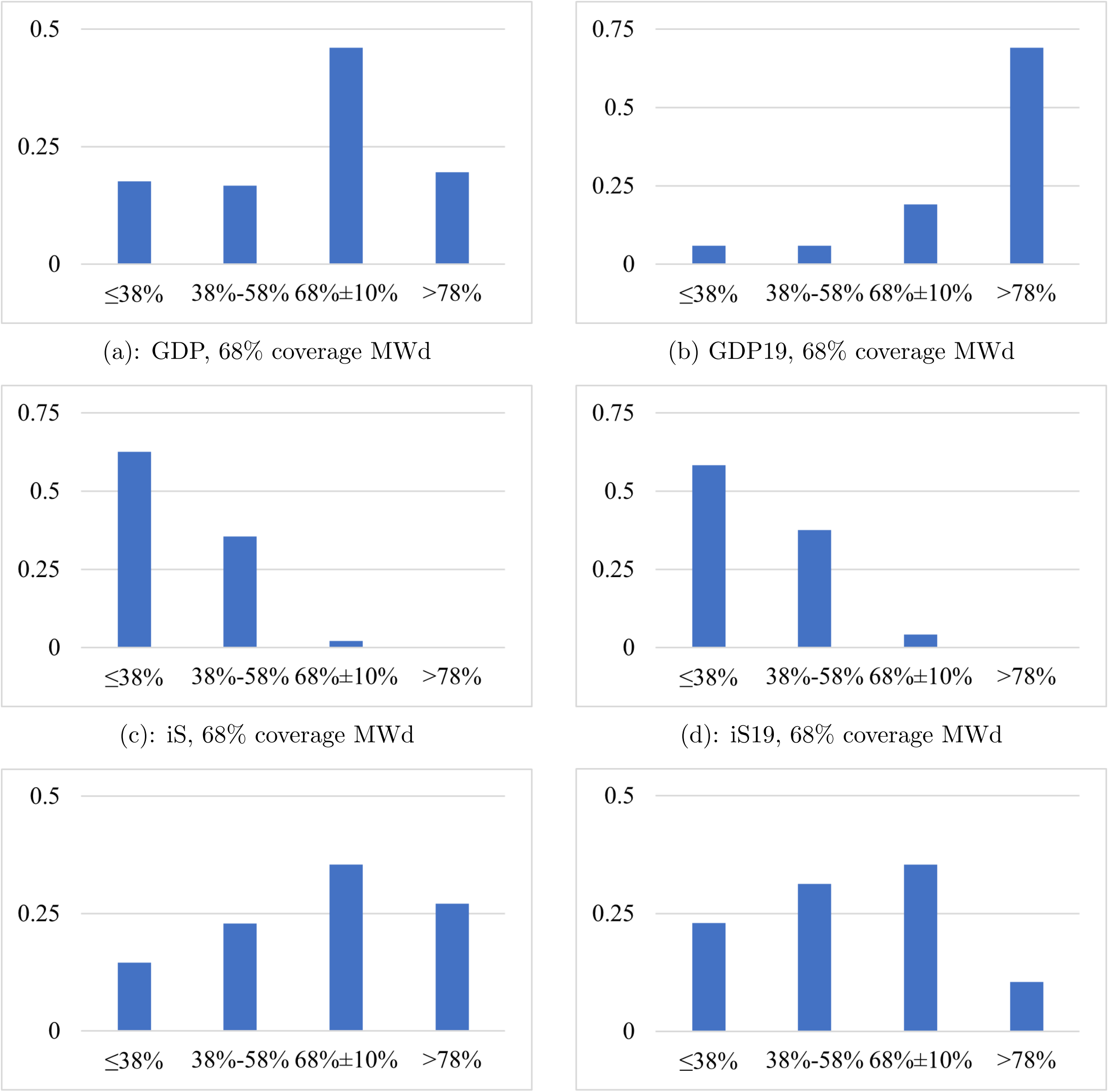
One can see notable degradation of performance in the $h = 50$ horizon relative to $h = 25$ and $h = 10$, along with smaller degradation for $h = 25$ relative to $h = 10$. As well, one can see that $h = 10$ and $h = 25$ track one another, with over- or under-coverage in one tending to be associated with the same in the other – unsurprising, given that the forecasts and realizations come from the same data. Finally, one can see that there is considerable variation. Even limiting ourselves to $h = 10$ and $h = 25$, actual coverage runs from 43% to 93%. Teasing out circumstances that lead to over- or under-coverage is an interesting task, but one that we leave for future research.

Let us return to Table 5.1. In panel A, across models (rows (6) to (10)), MWd is best according to "fraction 68% ± 10%" for the stationary variables, and it split the honors with MW0 for median closest to 68%. Figure 5.2(a) presents a histogram of coverage for MWd. The spike at 68% ± 10% is pronounced, which is good; what is bad is that many (indeed, slightly over half) of the sets of forecasts yield coverage that is outside of 68% ± 10%. Also, one can see that there are (slightly) more sets of forecasts with coverage below 58% (leftmost two bars) than above 78%. That is, median coverage is below 68%. Something new to the figure and not presented in the table is the fact that coverage for some sets of forecasts is far from 68%, with about one sixth of the forecasts having coverage below 38%.

We will discuss the other graphs in Figure 5.2 shortly. For the present, let us turn to results for shorter samples, in panel B of Table 5.1. For the stationary variables, country coverage is the same as in panel A. So any differences between panel A and panel B are attributable to the sample periods considered. For almost each and every permutation of model and horizon, median coverage in panel B is higher than in panel A. The increase is particularly notable for GDP growth. Since, in our view, coverage is well centered in panel A, the implication is that median coverage is generally above 68% in panel B. Unless one would rather have coverage that is above 68% than below 68%–which of course is entirely possible–performance is less satisfactory in panel B. As well, "fraction 68% ± 10%" falls in virtually every case, sometimes dramatically so (e.g., for $h = 10$, GDP growth, it falls from 0.58 to 0.29).

The performance of MWd for the stationary variables degrades notably in the 1919- sample in panel B. The histogram in Figure 5.2(b) shows that nearly three-fourths of the sets of forecasts have coverage greater than 78%. That is, the MWd forecast intervals are almost always too broad, with realized growth rates landing in the 68% forecast intervals far more often than 68% of the time. Indeed, the vertical scale of the graph is different than Figure 5.2(a), to accommodate the high fraction falling in the ">78%" bin.

To study how sampling schemes affect our differing results for the different samples, Table 5.2 breaks down the "fraction 68% ± 10%" results for both the rolling and recursive schemes. The "all" line repeats the "fraction 68% ± 10%" that appears in row (4) of panel A and of panel B in Table 5.1. The rolling and recursive lines report this statistic for forecasts constructed according to the indicated scheme. Since exactly half of the *all* samples used each scheme, the *all* figure is an average of the rolling and recursive figures (apart from rounding).



(a): GDP, 68% coverage MWd

(b) GDP19, 68% coverage MWd

(c): iS, 68% coverage MWd

(d): iS19, 68% coverage MWd

(e): iS, 68% coverage RW

(f): iS19, 68% coverage RW

Figure 5.2

These are histograms of the fraction of samples in which actual forecast interval coverage fall in the indicated range. "GDP" refers to GDP growth and "iS" refers to short-term interest rates. The left column relies on the entire sample, the right column on samples starting around 1919. In (a)-(f) above, the values for the 68% ± 10% bins repeat values from panel A of Table 5.1 (left column above) or panel B of Table 5.1 (right column above), as follows: (a) and (b): row (9), column (3c); (c) and (d): row (9), column (7c); (e) and (f): row (5), column (7c).

Table 5.2: Fraction of samples with coverage 68% ± 10%

		(1a)	(1b)	(2a)	(2b)	(3a)	(3b)
		GDP growth		Inflation		Prod. growth	
		—Sample—		—Sample—		—Sample—	
		Full	1919-	Full	1919-	Full	1919-
(1)	*all*	0.36	0.24	0.25	0.25	0.26	0.25
(2)	rolling	0.41	0.33	0.26	0.19	0.29	0.34
(3)	recursive	0.30	0.10	0.25	0.31	0.23	0.16

Notes:
1. The *all* line repeats "fraction 68% ± 10%" presented in line (4) of panels A and B of Table 5.1.
2. The rolling and recursive lines split the results underling the *all* line into the halves associate with the rolling and with the recursive scheme. See text for definition.
3. See notes to Table 5.1.

Coverage for "all" in Table 5.2 is worse in the 1919- sample for GDP growth and similar across samples for inflation and productivity growth. Consistent with these overall results, each scheme's performance is worse in the 1919- sample for GDP growth. For inflation, the rolling scheme's performance is worse in the 1919- sample while the recursive scheme improved. Productivity has the opposite results. Hence, we cannot conclude that changes in performance across samples is explained by the use of distant data by the recursive scheme. Rather, we leave the changes in performance as unexplained at present.

We return now to panel B in Table 5.1. Across the two horizons, coverage is better for $h = 10$ than for $h = 25$. Across the five models, the AR(1) model is best for the stationary variables, as measured by closeness of median to 68% or "fraction 68% ± 10%." The iid model performs similarly to the AR(1) model for GDP and productivity growth, but is distinctly worse for inflation. The MW0 is similar to the AR(1) model for "fraction 68% ± 10%" for the stationary variables.

Now let us consider results for interest rates, noting that country coverage is broader in panel B than in panel A. For long-term rates, performance in each of the samples (both panels A and B) is roughly similar, and, in general, is poor relative to that for GDP growth, inflation or productivity growth. Median coverage is generally well below 68% (column (6b)). The fraction of sets of forecasts with coverage of 68% ± 10% is tiny, generally less than 0.10 (column (6c)). Of this seemingly poor collection of models, the random walk and MW1 models perform best.

For short-term rates, the random walk model is distinctly the best. It produces median coverage and "fraction 68% ± 10%" that is comparable to that for stationary variables in panel A: median coverage is 66% (panel A) and 57% (panel B); "fraction 68% ± 10%" is 0.35 in both panels. To our eye, therefore, the RW model's performance is tolerable.

While the coverage of the MWd model is intended to be robust to very persistent variables, such as interest rates, it performs slightly worse than the MW1 model for both long- and short-term rates and for both samples based on both median coverage and "fraction 68% ± 10%." It also performs worse than the random walk model based on "fraction 68% ± 10%," materially so for short-term rates. However, the MWd model generally performs better than the MW0 and AR(1) models for both long- and short-rates. Overall, we view these results as indicating that the coverage of the MWd model has some robustness to high degrees of persistence.

For interest rates, across horizons, performance is better for $h = 10$ than for longer horizons.

Figures 5.2(c) and 5.2(d) depict the behavior of MWd for short term interest rates. For both the complete (Fig. 5.2(c)) and 1919- samples (Fig. 5.2(d)), over half the sets of forecasts have actual coverage below 38%. For interest rates, MWd is neither the best nor worst performing model, so these histograms are broadly representative.

Figures 5.2(e) and 5.2(f) present coverage histograms for the random walk model. These have the same flavor as the histogram for MWd/GDP growth in Figure 5.2(a), with a nice peak in the 68% ± 10% bin.

5.2 Results for |bias| and RMSFE

Table 5.3 has results for median absolute value of bias (|bias|) and root mean squared forecast error (RMSFE). We present these with values normalized relative to the median value of a baseline model. With this convention, the median values for the baseline model are 1. The choice of baseline model is arbitrary. We normalize to allow a clean presentation of relative sizes of median |bias| and median RMSFE.[14]

Recall that the iid and MW0 models both use the sample mean to forecast and, hence, for each data set the two models yield identical |bias| and RMSFE. For the three series generally modeled as stationary (GDP growth, inflation, productivity growth), the baseline model is iid/MW0. For the two interest rates series, the baseline model is RW.

To understand the column "% lowest RMSFE," consider the 40% figure for iid/MW0 for GDP growth in the left corner of panel A. Of the 102 samples (102 = 17 countries × 3 horizons × 2 sampling schemes), the iid/MW0 forecast produces the lowest RMSFE in 41, or 40%, of these samples. Apart from rounding, the percentages in the "% lowest RMSFE" column add to 100 (because in any given sample, one of the models produces the lowest RMSFE).

We begin with the six sets of results in panels A and B for our stationary variables: GDP growth, inflation and productivity growth. A striking result is that MW1 has exceptionally small

[14] We found the raw numerical values unrevealing, and hence express them in relative fashion.

Table 5.3: Absolute value of forecast bias (\|bias\|) and root mean squared forecast error (RMSFE)

A. Samples starting 1870s or 1891

	(1)	(2a)	(2b)	(2c)	(3a)	(3b)	(3c)	(4a)	(4b)	(4c)	(5a)	(5b)	(5c)	(6a)	(6b)	(6c)
(1)		GDP growth			CPI inflation			Productivity growth			Long term interest rates			Short term interest rates		
(2)		17 countries, 102 samples			17 countries, 102 samples			23 countries, 138 samples			12 countries, 72 samples			8 countries, 48 samples		
(3)	model	median |bias|	median RMSFE	% lowest RMSFE	median |bias|	median RMSFE	% lowest RMSFE	median |bias|	median RMSFE	% lowest RMSFE	median |bias|	median RMSFE	% lowest RMSFE	median |bias|	median RMSFE	% lowest RMSFE
(4)	RW	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1	1	51%	1	1	23%
(5)	iid/MW0	1	1	40%	1	1	50%	1	1	36%	6.32	1.21	29%	2.28	1.03	44%
(6)	MW1	0.16	2.54	0%	0.59	2.07	0%	0.27	2.15	0%	0.84	1.08	3%	0.99	1.03	6%
(7)	MWd	0.95	1.03	25%	0.60	1.12	20%	0.93	1.08	43%	0.93	1.08	11%	0.79	0.98	4%
(8)	AR1	0.96	0.97	34%	0.81	0.96	30%	0.99	1.01	22%	1.07	1.08	6%	0.93	0.97	23%

B. Samples starting 1919 or 1920s

	(1)	(2a)	(2b)	(2c)	(3a)	(3b)	(3c)	(4a)	(4b)	(4c)	(5a)	(5b)	(5c)	(6a)	(6b)	(6c)
(1)		GDP growth			CPI inflation			Productivity growth			Long term interest rates			Short term interest rates		
(2)		17 countries, 68 samples			17 countries, 68 samples			23 countries, 92 samples			15 countries, 60 samples			12 countries, 48 samples		
(3)	model	median |bias|	median RMSFE	% lowest RMSFE	median |bias|	median RMSFE	% lowest RMSFE	median |bias|	median RMSFE	% lowest RMSFE	median |bias|	median RMSFE	% lowest RMSFE	median |bias|	median RMSFE	% lowest RMSFE
(4)	RW	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1	1	52%	1	1	31%
(5)	iid/MW0	1	1	28%	1	1	6%	1	1	20%	1.89	1.32	33%	0.76	0.99	40%
(6)	MW1	0.69	1.62	1%	1.28	1.07	24%	0.34	1.55	17%	1.32	1.06	3%	1.11	1.01	15%
(7)	MWd	1.24	1.01	37%	0.91	0.90	43%	0.86	1.17	36%	1.15	1.06	5%	0.91	1.00	2%
(8)	AR1	1.17	0.96	34%	1.28	0.90	28%	0.92	1.19	27%	1.34	1.02	7%	0.65	1.10	13%

Notes:
1. For each model and data series, the median value of |bias| and RMSFE is computed for the number of samples given in row (2). The median values are then expressed relative to the median value for a baseline model. The baseline model is iid/MW0 in columns (2)-(4), RW in columns (5) and (6).
2. "% lowest RMSFE" gives the percentage of sets of forecasts in which the indicated model had the lowest RMSFE. For example, in row (5), column (2c) the value of 40% reflects the fact that in 41 of the 102 sets of forecasts, the iid/MW0 model had the lowest RMSFE.
3. In row (5) in each panel, results for the iid and MW0 models are identical.
4. See notes to Table 5.1.

|bias| and exceptionally large RMSFE. The most extreme example is for GDP growth in panel A. For MW1, |bias| is about one-sixth of the next best model ($0.16 \approx 0.95/6$) while RMSFE is over twice that of the next worst model ($2.54 > 2 \times 1.03$). Clearly this means that for such data MW1 has huge variance along with small bias. That MW1 has large variance is unsurprising, since MW1 is intended for unit root series and the data under discussion are generally considered stationary. That it has small |bias| is surprising. We have done a deep dive into a couple of the underlying series and learned little other that to confirm that, indeed, forecasts from MW1 tend to be scattered far away, but symmetrically, around realizations. Hence the forecast errors average to near zero but have large variance. At this point, we do not have a satisfying explanation for the performance of MW1, which possibly is a fluke. At any rate, we henceforth put aside MW1 in our discussion of the three stationary series.

Each performance metric (for example, median |bias|) appears six times across the three stationary series, three times in panel A and three times in panel B. MWd has smallest median |bias| in five of the six cases (the exception being GDP growth in panel B). The AR(1) model has smallest median RMSFE in four of the six cases (the exceptions being productivity growth in both panels). MWd's "% lowest RMSFE" is the largest in four of the six cases (the exceptions being GDP growth and inflation in panel A.) That is, in four of the six cases, MWd produces the lowest RMSFE more often than any other model.

Finally, consider the two interest rate series. The random walk model is clearly the best model for long-term interest rates, having top performance for all three measure in panel B and top for all but |bias| in panel A. For short-term interest rates, the full sample yields a mixed picture while the 1919- sample indicates that iid/MW0 performs best.

5.3 Summary

The bottom line is that the MWd model is a good choice for stationary variables, in terms of both forecast interval coverage and accuracy of predictions as measured by |bias| or RMSFE. The iid, MW0 and AR(1) models perform tolerably well, though they are a step behind MWd. The MW1 model is not a good choice for stationary variables. For the interest rate variables, the random walk model is probably the preferred choice, though its advantages relative to our alternatives are not as broadly based. MW1 and MWd come next, with MW0 and perhaps AR(1) not good choices for interest rates.

The reader may reasonably look at the MWd and random walk models as merely the best of a bad lot. For forecast interval coverage, even the best performing model on the most accommodating data–probably MWd for GDP growth, in the complete sample in panel A of Table 5.1–has flaws.

It by no means delivers an unambiguously attractive outcome in which actual coverage is clustered close to 68% in the vast majority of samples. For what it is worth, the performance summarized in Table 5.1 compares favorably to performance of some well established procedures for long-horizon inference about impulse response functions for VARs. The object of study in that literature is not the same as ours, but it is related. Simulations in that literature find, as did we using actual data, that there can be dramatic under- or over-coverage. See, for example, Pesavento and Rossi (2007). As well, the closely related literature on estimation of long-run variances has long struggled to devise procedures that work well for inference for a wide range of data series. For a relatively recent study whose simulations yield both distinctly worse and distinctly better coverage than shown here for actual data, see Vogelsang (2018).

6 Social Security Projection and Model Comparisons

In this section, we compare forecasts and forecast intervals from our models to projections of the Social Security Administration (SSA).[15] We note that the SSA projections are not forecasts from models such as ours; rather, they reflect the judgement of the SSA Trustees. Our goal then is to answer the question: if our forecasting models are used to inform the judgement of the SSA Trustees and their projections, what would our models tell us?

Our source for the SSA projections is the 2021 Trustee's Report. The Report includes 75-year projections for many variables relevant to the balance sheet of the Social Security system. These include real per capita GDP growth, CPI inflation, labor productivity growth and a long-term interest rate series. These series align well but not perfectly with the U.S. data used in previous sections. We were unable to locate a good analogue to the short-term interest rate series used in previous sections, and hence do not consider short rates in the present section. See Section 3.2 for some additional details.

For each of our variables, there are three SSA projections associated with what are called intermediate, low- and high-cost scenarios. We interpret the intermediate projection as a central tendency, and the projections associated with high- and low-cost scenarios as bracketing a range of plausible outcomes. For brevity, we use the "low-cost projection" as the shorthand for "the projection associated with the low-cost scenario," and similarly for intermediate and high-cost projections.

To estimate our models, we use 48 years of data, from 1973 to 2020, which matches the rolling

[15] The SSA projections for some variables, such as productivity growth and inflation, are also viewed as economic assumptions to help evaluate the financial operations of the Old-Age and Survivors Insurance and Federal Disability Insurance trust funds.

scheme sample size in our pseudo out-of-sample analysis. Further, we note that the 2021 Social Security Trustees Report focuses on the years 1969 to 2019 when discussing its economic projections, which has a high degree of overlap with the sample we use.[16] We consider forecasts and forecast intervals for a horizon of $h = 25$ years, running 2021-2045. We choose that horizon because our pseudo out-of-sample results were not particularly reassuring about $h = 50$ year horizons, and $h = 10$ is on the low end of horizons of interest.

Table 6.1 has results. Row (1) has SSA projections. The intermediate projection is the number in the top half of the row. The numbers in the parentheses in the lower half of the row come from the low- and high-cost projections, with the numerically smaller value put on the left end of the range. We will refer to the object in parentheses as the "projection interval." Subsequent lines in the table present the model point forecasts along with 68% forecast intervals. Of course, the point forecasts for the iid and MW0 models are identical in columns (1)-(3), while the forecast intervals are different.

For the three stationary variables in columns (1)-(3), MW1 is an outlier. For example, the point estimate of -0.0% for GDP growth falls starkly far from the 1.5% to 1.7% values of all other entries in the column. Now, in both the overall set of results and in the U.S. results presented in the pseudo out-of-sample analysis of the previous section of this paper, the MW1 model had by far the worst performance for stationary variables. This was true whether one measured performance by either RMSFE or forecast interval coverage. Hence, we put little weight on the MW1 forecast for these three variables, and will put MW1 aside until we discuss the interest rate results.

Focus first on real per capita GDP growth in column (1). Point estimates are very similar in all rows (again, putting aside MW1), and are close the intermediate SSA projection. Our 68% forecast intervals tend to be a little broader than the SSA low-high interval. This is notably so for MWd, which is the model that tended to perform best overall for stationary variables. However, the discrepancies are relatively small.

Next, jump to productivity growth in column (3). As with GDP growth, there is little disagreement amongst the models' point forecasts and the SSA intermediate projection. Further, the agreement between SSA low-high interval and the models' forecast intervals is even stronger than for GDP growth.

For CPI inflation in column (2), we see some larger discrepancies. The SSA projection of 2.4% is at the low end of our point forecasts, which range from 2.4% to 3.8% (putting aside MW1). In the previous section's pseudo out-of-sample results for inflation, while MWd's performance in terms of bias and RMSFE overall was probably best, both the iid/MW0 and AR(1) were pretty close. On balance, these three models that worked relatively well suggest that the SSA intermediate

[16] See pages 105 to 116 of the 2021 OASDI Trustees Report.

Table 6.1: SSA Projections and Model Forecasts, 2021-2045 ($h = 25$)

		(1) GDP growth	(2) CPI inflation	(3) Productivity growth	(4) Long-term interest rates
(1)	SSA projections	1.6% (1.2%, 2.1%)	2.4% (1.8%, 3.0%)	1.6% (1.2%, 1.9%)	4.1% (3.3%, 5.1%)
(2)	random walk	n.a.	n.a.	n.a.	1.0% (-1.7%, 3.7%)
(3)	iid	1.7% (1.2%, 2.2%)	3.8% (3.1%, 4.6%)	1.5% (1.3%, 1.8%)	n.a.
(4)	MW0	1.7% (1.2%, 2.3%)	3.8% (2.2%, 5.5%)	1.5% (1.2%, 1.9%)	6.1% (4.2%, 8.1%)
(5)	MW1	-0.0% (-2.3%, 2.2%)	1.5% (-2.3%, 5.3%)	1.6% (0.3%, 2.9%)	2.3% (-0.9%, 5.5%)
(6)	MWd	1.5% (0.7%, 2.3%)	2.4% (-0.4%, 5.1%)	1.5% (1.0%, 2.0%)	2.9% (-0.0%, 5.8%)
(7)	AR(1)	1.5% (1.0%, 2.1%)	3.0% (1.4%, 4.6%)	1.5% (1.3%, 1.8%)	0.5% (-1.9%, 2.9%)

Notes:

1. The top half of row (1) shows the SSA's intermediate projection. The bottom half of row (1) shows the SSA's low- and high-cost projections in parentheses, with the numerically smaller value put on the left end of the range.

2. The top half of rows (2)-(7) show point forecasts from the respective model. The bottom half of rows (2)-(7) show the 68% forecast intervals from the respective model.

projection of 2.4% is on the low side.[17]

In terms of interval width, those from all of our models are broader than that for the SSA's projection interval. In particular, the width of the MW0 model at 3.3% is nearly triple the width of the projection interval at 1.2%, and the widths of the AR(1) and MWd models are even wider than that of the MW0 model. In contrast to GDP and productivity growth, which have SSA projection intervals that are similar to the 68% forecast intervals from our forecasting models, CPI inflation has SSA projection interval width that is narrower than is consistent with a 68% forecast interval. Put differently, if one wishes that projection intervals for each variable reflect equal probability

[17] In a private communication, an economist from Social Security commenting on this section of the paper noted that the SSA Trustees view the high inflation of the 1970s and 80s as unlikely to be repeated in the future.

of occurrence, either the GDP and productivity projection intervals are too broad or the inflation projection interval is too narrow.[18]

For long-term interest rates in column (4), our point forecasts are generally lower than the SSA's intermediate projection. Further, after we downweight the MW0 and AR(1) forecasts because of their poor performance with long-term interest rates in our pseudo out-of-sample analysis, our remaining three forecasts (random walk, MW1 and MWd) are notably below those of the SSA projections. Indeed, the upper bound of our forecast interval for the random walk model (3.7%) is below the intermediate SSA projection of 4.1%.

In addition to disagreement between the models' point forecasts and SSA's intermediate projection, the models also have forecast intervals for the interest rate that are wider than SSA's projection interval. The random walk, MW1 and MWd models have forecast interval widths that range from about 5.5% to 6.5%. In contrast, the SSA projection interval has a width of 1.8%. Further, the fact that there tended to be undercoverage in our pseudo out-of-sample forecast intervals for interest rates suggests that the models' intervals should be even broader. On the other hand, the effective lower bound on nominal interest rates may cut the other way. The lower end of the forecast interval for the random walk model is -1.7%. One may reasonably argue that the odds of long term nominal rates averaging -1.7% over a 25 year period are quite slim, in which case a -1.7% value for the lower end of the forecast interval seems doubtful. Perhaps some additional structure is necessary for a variable such as interest rates with a bound. Nonetheless, we do think our results indicate that the SSA projection is on the high side, and that the projection interval seems to reflect less uncertainty than do the intervals for GDP or labor productivity growth.

In sum, our models and the SSA projections line up well for per capita GDP growth and productivity growth. Hence, if SSA views 68% forecast intervals as covering a reasonable range of projection scenarios, then our models do not indicate the need for any material change in the construction of per capita GDP or productivity growth projections. By contrast, for inflation and long-term interest rates, there is some disagreement between SSA's intermediate projections and the models' point forecasts. Further, the projection intervals for inflation and long-term interest rates are narrow compared to our 68% forecast intervals. It is our understanding that projections for all four variables may be linked to one another. So even if one agrees that our models raise concerns about inflation and interest rate projections, it might not be feasible to adjust one set of

[18] In a private communication, an economist from Social Security commenting on this section of the paper stated that one should not "judge" the projection intervals with confidence intervals from our models. To be clear: we are not judging the projection intervals. We are contrasting them with forecast intervals from our models. One can endorse the projection intervals as accomplishing exactly what SSA aims to accomplish in their use of high- and low-cost scenarios and simultaneously observe that in a probabilistic sense some intervals are likelier to be breached than are other intervals.

projections while holding the other set fixed. In that case, we interpret the models as suggesting that the chain of steps that leads from baseline assumptions to interest rate and inflation projections may warrant re-consideration.

7 Conclusions

In this paper we have used pseudo out-of-sample analysis to evaluate some models for long-horizon forecasts and forecast intervals. Our data for the out-of-sample analysis stretches for a century or more for up to 23 countries. The variables, which are forecast for horizons up to 50 years, are real per capita GDP growth, CPI inflation, labor productivity growth, and long- and short-term nominal interest rates. We find that a frequency domain model that does not require one to take a stand on the order of integration is a good choice for forecasting GDP growth, CPI inflation and labor productivity growth (Müller and Watson (2016)). A driftless random walk model is probably the best choice for forecasting nominal interest rates.

We then compare the point and interval forecasts from our models, after excluding the poorest performing models in our pseudo out-of-sample analysis, to the Social Security Administration's (SSA's) projections. We find that the SSA's projections for real per capita GDP and productivity growth are similar to point forecasts and 68% forecast intervals from our models. In contrast, we find that the distance between the SSA's low- and high-cost scenario projections for CPI inflation and nominal interest rates is materially smaller than the 68% forecast intervals from our models. In other words, our models indicate that the probability that inflation or interest rates breach the projections of either the low- or high-cost scenarios is higher than the comparable probability for GDP and labor productivity growth.

Tasks for future research include consideration of a wider set of models for forecasting and for inference. Possibilities include models with bias adjustments for highly serially correlated data, methods such as in Chudý, Karmakar, and Wu (2020) for construction of forecast intervals for stationary data, model averaging, and multivariate models.

A Forecast Distributions of the Simple Models

The iid Model. The model is $x_t = \mu + u_t$, with u_t being iid with mean zero and variance σ_u^2. We use the estimates $\hat{\mu} = T^{-1}\sum_{t=1}^{T} x_t$ and $\hat{\sigma}_u^2 = T^{-1}\sum_{t=1}^{T}(x_t - \hat{\mu})^2$. When h and T are each sufficiently large, we treat $h^{1/2}[(x_{T+1} + \cdots + x_{T+h})/h - \mu]$ and $T^{1/2}(\hat{\mu} - \mu)$ each as normally distributed with $h^{1/2}[(x_{T+1} + \cdots + x_{T+h})/h - \mu] \sim N(0, \sigma_u^2)$ and $T^{1/2}(\hat{\mu} - \mu) \sim N(0, \sigma_u^2)$. With u_t iid, we also have that $h^{1/2}[(x_{T+1} + \cdots + x_{T+h})/h]$ and $T^{1/2}\hat{\mu}$ are independent so that $(x_{T+1} + \cdots + x_{T+h})/h - \hat{\mu} \sim N(0, [(1/h) + (1/T)]\sigma_u^2)$ or

$$(x_{T+1} + \cdots + x_{T+h})/h = \hat{\mu} + \sqrt{[(1/h) + (1/T)]\sigma_u^2}\ \xi,$$

in which ξ is a standard normal random variable. For the forecast intervals, we use $\hat{\sigma}_u^2$ in the place of σ_u^2.

The Random Walk Model. The model is $x_t = x_{t-1} + u_t$, with u_t being iid with mean zero and variance σ_u^2. We use the estimate $\hat{\sigma}_u^2 = (T-1)^{-1}\sum_{t=2}^{T}(x_t - x_{t-1})^2$. It is then the case that

$$\begin{aligned}(x_{T+1} + \cdots + x_{T+h})/h - x_T &= [(x_{T+1} - x_T) + \cdots + (x_{T+h} - x_T)]/h \\ &= [u_{T+1} + (u_{T+1} + u_{T+2}) + \cdots + (u_{T+1} + \cdots u_{T+h})]/h \\ &= [hu_{T+1}/h + (h-1)u_{T+2}/h + \cdots u_{T+h}/h].\end{aligned}$$

Hence, $h^{-1/2}[(x_{T+1} + \cdots + x_{T+h})/h - x_T] = h^{-1/2}[hu_{T+1}/h + (h-1)u_{T+2}/h + \cdots u_{T+h}/h]$, which we treat as normally distributed when h is large. Using $\sum_{j=1}^{h} j^2 = h(h+1)(2h+1)/6$ from Equation 16.1.10 in Hamilton (1994), we compute $h^{-1/2}[(x_{T+1} + \cdots + x_{T+h})/h - x_T] \sim N(0, (h+1)(2h+1)\sigma_u^2/(6h^2))$. Hence, we have

$$(x_{T+1} + \cdots + x_{T+h})/h = x_T + \sqrt{(h+1)(2h+1)\sigma_u^2/(6h)}\ \xi,$$

in which ξ is a standard normal random variable. For the forecast intervals, we use $\hat{\sigma}_u^2$ in the place of σ_u^2.

The AR(1) Model. The model is $x_t = \rho_0 + \rho_1 x_{t-1} + u_t$, with u_t being iid with mean zero and variance σ_u^2. We compute $\hat{\rho}_0$ and $\hat{\rho}_1$ with ordinary least squares. As noted in the paper, we only forecast with the AR(1) model if $|\hat{\rho}_1| < 1$. If $\hat{\rho}_1 >= 1$, we forecast with the random walk model. If $|\hat{\rho}_1| < 1$, we compute $\hat{u}_t = x_t - \hat{\rho}_0 - \hat{\rho}_1 x_{t-1}$ and $\hat{\sigma}_u^2 = (T-1)^{-1}\sum_{t=2}^{T}\hat{u}_t^2$. We compute the period-by-period forecasts recursively

$$\hat{x}_{T+1} = \hat{\rho}_0 + \hat{\rho}_1 x_T,$$
$$\hat{x}_{T+j} = \hat{\rho}_0 + \hat{\rho}_1 \hat{x}_{T+j-1}, \quad j = 2, \ldots, h.$$

These equations imply $\hat{x}_{T+j} = (1 + \hat{\rho}_1 + \cdots + \hat{\rho}_1^{j-1})\hat{\rho}_0 + \hat{\rho}_1^j x_T$. Using $(1 - \hat{\rho}_1^j)/(1 - \hat{\rho}_1) = 1 + \hat{\rho}_1 + \cdots + \hat{\rho}_1^{j-1}$, we then have

$$\hat{x}_{T+j} = \frac{\hat{\rho}_0}{1 - \hat{\rho}_1} + \hat{\rho}_1^j \left(x_T - \frac{\hat{\rho}_0}{1 - \hat{\rho}_1} \right)$$

and then

$$\frac{1}{h}(\hat{x}_{T+1} + \cdots + \hat{x}_{T+h}) = \frac{\hat{\rho}_0}{1 - \hat{\rho}_1} + \frac{1}{h}(\hat{\rho}_1 + \hat{\rho}_1^2 + \cdots + \hat{\rho}_1^h) \left(x_T - \frac{\hat{\rho}_0}{1 - \hat{\rho}_1} \right),$$

which is the point forecast used in the paper. Fuller and Hasza (1980) show that these forecasts are unbiased when u_t is drawn from a normal distribution. However, Magnus and Pesaran (1991) point out that this unbiasedness result depends on assumptions about the initial observation x_1 and that unbiasedness holds if $E(x_1) = \rho_0/(1 - \rho_1)$.

To simplify the analysis, we assume that $\hat{\rho}_0$ and $\hat{\rho}_1$ equal the population values ρ_0 and ρ_1 with certainty. With these assumptions, we have $h^{1/2}[(x_{T+1} + \cdots + x_{T+h})/h - (\hat{x}_{T+1} + \cdots + \hat{x}_{T+h})/h] = h^{1/2}[(1 + \rho_1 + \cdots + \rho_1^{h-1})u_{T+1}/h + (1 + \rho_1 + \cdots + \rho_1^{h-2})u_{T+2}/h + \cdots + u_{T+h}/h]$, which we treat as normally distributed when h is large. We compute the distribution $h^{1/2}[(x_{T+1} + \cdots + x_{T+h})/h - (\hat{x}_{T+1} + \cdots + \hat{x}_{T+h})/h] \sim N(0, [1 + (1 + \rho_1)^2 + \cdots + (1 + \rho_1 + \cdots + \rho_1^{h-1})^2]\sigma_u^2/h)$. Hence, we have

$$\begin{aligned}(x_{T+1} + \cdots + x_{T+h})/h = {} & \frac{\hat{\rho}_0}{1 - \hat{\rho}_1} + \frac{1}{h}(\hat{\rho}_1 + \hat{\rho}_1^2 + \cdots + \hat{\rho}_1^h) \left(x_T - \frac{\hat{\rho}_0}{1 - \hat{\rho}_1} \right) \\ & + \sqrt{[1 + (1 + \rho_1)^2 + \cdots + (1 + \rho_1 + \cdots + \rho_1^{h-1})^2]\sigma_u^2/h^2}\ \xi,\end{aligned}$$

in which ξ is a standard normal random variable. For the forecast intervals, we use $\hat{\sigma}_u^2$ in the place of σ_u^2.

B Covariance Approximations for Müller and Watson

Given a data sample, $\{x_1, \ldots, x_T\}$, the forecasting approach in Müller and Watson (2016) uses $\hat{\beta}_0 = T^{-1} \sum_{t=1}^{T} x_t$ and $\hat{\beta}_j = T^{-1} \sum_{t=1}^{T} \sqrt{2} \cos(\pi j(t-1/2)/T) x_t$ for $j = 1, \ldots, q$, in which q is much smaller than T. Write $\hat{\beta}_{1:q} = [\hat{\beta}_1, \ldots, \hat{\beta}_q]'$ as a $(q \times 1)$ vector and $y_{T,h} = (x_{T+1} + \cdots + x_{T+h})/h - \hat{\beta}_0$ as a scalar. Then, Müller and Watson (2016) prove a central limit theorem

$$T^{1-\kappa} \begin{bmatrix} \hat{\beta}_{1:q} \\ y_{T,h} \end{bmatrix} \Rightarrow \begin{bmatrix} \tilde{\beta} \\ y \end{bmatrix} \sim N(0, \Sigma), \quad \Sigma = \begin{bmatrix} \Sigma_{\beta\beta} & \Sigma_{\beta y} \\ \Sigma_{y\beta} & \Sigma_{yy} \end{bmatrix},$$

in which κ is a scaling factor that depends on the relevant model for x_t. For the MW0 model, $\kappa = 1/2$. For the MW1 model, $\kappa = 3/2$. For the MWd model, $\kappa = 1/2 + d$.

The forecasting approach in Müller and Watson (2016) relies on knowing the form of Σ. For the MW0 model, Müller and Watson (2016) provide analytical values for every element of Σ. However, for the MW1 and MWd models, we use the numerical approximations from Section 3.2 of Müller and Watson (2020).

To start our numerical approximation, let $r = h/T$ be the ratio of the forecast horizon to the sample size. Then, we use $N = 1000$ and compute the integer, $H = \text{round}(rN)$. Using the notation $\psi_{j,t} = \sqrt{2} \cos(\pi j(t-1/2)/N)$, we write the $(N \times q)$ matrix

$$\Psi = \begin{bmatrix} \psi_{1,1} & \psi_{2,1} & \cdots & \psi_{q,1} \\ \psi_{1,2} & \psi_{2,2} & \cdots & \psi_{q,2} \\ \vdots & \vdots & & \vdots \\ \psi_{1,N} & \psi_{2,N} & \cdots & \psi_{q,N} \end{bmatrix}.$$

Then, we write the $((N+H) \times (q+1))$ matrix

$$\Xi = \begin{bmatrix} \Psi & -\mathbf{1}_{N \times 1} \\ \mathbf{0}_{H \times q} & (N/H)\mathbf{1}_{H \times 1}, \end{bmatrix}$$

in which $\mathbf{1}_{m \times n}$ denotes an $(m \times n)$ matrix of ones and $\mathbf{0}_{m \times n}$ denotes an $(m \times n)$ matrix of zeros. Next, let L be a lower-triangular $((N+H) \times (N+H))$ matrix with ones on and below the main diagonal. Then, we approximate Σ for the MW1 model with

$$\Sigma = \sigma_{lrv}^2 (\Xi' L L' \Xi)/N^3,$$

in which σ_{lrv}^2 denotes the long-run variance of Δu_t. The value of σ_{lrv}^2 is unknown, but the form of the forecast interval is such that any value of $\sigma^2_{lrv} > 0$ yields the same forecast interval. We can see this by noting that σ^2_{lrv} scales every element of Σ equally, causing it to be divided out of the terms $\Sigma_{y\beta}\Sigma^{-1}_{\beta\beta}$ and $(\Sigma_{yy} - \Sigma_{y\beta}\Sigma^{-1}_{\beta\beta}\Sigma_{\beta y})(\hat{\beta}'_{1:q}\Sigma^{-1}_{\beta\beta}\hat{\beta}_{1:q}/q)$ in the forecast interval. Hence, we set $\sigma^2_{lrv} = 1$ and compute

$$\Sigma = (\Xi' L L' \Xi)/N^3, \tag{B.1}$$

for the MW1 model.

For the MWd model, if the value of d is such that $-0.5 < d < 0.5$, define a $((N+H) \times (N+H))$ matrix Λ in which the (i,j) element is given by

$$\lambda_{i,j} = \frac{\Gamma(k+d)\Gamma(1-2d)}{\Gamma(k+1-d)\Gamma(1-d)\Gamma(d)},$$

in which $k = |i-j|$ and $\Gamma(\cdot)$ denotes the gamma function. Then, we set $\sigma^2_{lrv} = 1$ as in the MW1 model[19] and compute

$$\Sigma = (\Xi' \Lambda \Xi)/N^{1+2d}. \tag{B.2}$$

If the value of d is such that $0.5 < d < 1.5$, compute $\tilde{d} = d - 1$ and define a $((N+H) \times (N+H))$ matrix Λ in which the (i,j) element is given by

$$\lambda_{i,j} = \frac{\Gamma(k+\tilde{d})\Gamma(1-2\tilde{d})}{\Gamma(k+1-\tilde{d})\Gamma(1-\tilde{d})\Gamma(\tilde{d})},$$

in which $k = |i-j|$ and $\Gamma(\cdot)$ denotes the gamma function. Then, we set $\sigma^2_{lrv} = 1$ and compute

$$\Sigma = (\Xi' L \Lambda L' \Xi)/N^{1+2d}. \tag{B.3}$$

[19] For the MWd model, σ^2_{lrv} denotes the long-run variance of $(1-B)^d u_t$ with B being the backshift or lag operator.

References

Baillie, Richard T. 1996. "Long Memory Processes and Fractional Integration in Econometrics." *Journal of Econometrics* 73 (1):5–59. URL https://doi.org/10.1016/0304-4076(95)01732-1.

Bergeaud, Antonin, Gilbert Cette, and Rémy Lecat. 2016. "Productivity Trends in Advanced Countries between 1890 and 2012." *Review of Income and Wealth* 62 (3):420–444. URL https://doi.org/10.1111/roiw.12185.

Chudý, Marek, Sayar Karmakar, and Wei Biao Wu. 2020. "Long-Term Prediction Intervals of Economic Time Series." *Empirical Economics* 58 (1):191–222. URL https://doi.org/10.1007/s00181-019-01689-2.

Dev, Abhishek. 2015. "Forecasting Error in the US Social Security Administration's Economic Assumptions." Bard College *Senior Projects Fall 2015*, Paper 26. URL https://digitalcommons.bard.edu/senproj_f2015/26/.

Fuller, Wayne A. and David P. Hasza. 1980. "Predictors for the First-Order Autoregressive Process." *Journal of Econometrics* 13 (2):139–157. URL https://doi.org/10.1016/0304-4076(80)90012-3.

Granger, Clive W.J. and Yongil Jeon. 2007. "Long-Term Forecasting and Evaluation." *International Journal of Forecasting* 23 (4):539–551. URL https://doi.org/10.1016/j.ijforecast.2007.07.002.

Hamilton, James D. 1994. *Time Series Analysis*. Princeton University Press.

Jordà, Òscar, Moritz Schularick, and Alan M. Taylor. 2017. "Macrofinancial History and the New Business Cycle Facts." In *NBER Macroeconomics Annual 2016*, edited by Martin Eichenbaum and Jonathan A. Parker. University of Chicago Press, 213–263. URL https://doi.org/10.1086/690241.

Kashin, Konstantin, Gary King, and Samir Soneji. 2015. "Systematic Bias and Nontransparency in US Social Security Administration Forecasts." *Journal of Economic Perspectives* 29 (2):239–258. URL https://doi.org/10.1257/jep.29.2.239.

Lunsford, Kurt G. and Kenneth D. West. 2019. "Some Evidence on Secular Drivers of US Safe Real Rates." *American Economic Journal: Macroeconomics* 11 (4):113–139. URL https://doi.org/10.1257/mac.20180005.

Magnus, Jan R. and Bahram Pesaran. 1991. "The Bias of Forecasts from a First-Order Autoregression." *Econometric Theory* 7 (2):222–235. URL https://doi.org/10.1017/S0266466600004424.

Müller, Ulrich K. and Mark W. Watson. 2016. "Measuring Uncertainty about Long-Run Predictions." *Review of Economic Studies* 83 (4):1711–1740. URL https://doi.org/10.1093/restud/rdw003.

———. 2018. "Long-Run Covariability." *Econometrica* 86 (3):775–804. URL https://doi.org/10.3982/ECTA15047.

———. 2020. "Low-Frequency Analysis of Economic Time Series." Working Paper, Princeton University.

Pascual, Lorenzo, Juan Romo, and Esther Ruiz. 2004. "Bootstrap Predictive Inference for ARIMA Processes." *Journal of Time Series Analysis* 25 (4):449–465. URL https://doi.org/10.1111/j.1467-9892.2004.01713.x.

Pesavento, Elena and Barbara Rossi. 2007. "Impulse Response Confidence Intervals for Persistent Data: What Have We Learned?" *Journal of Economic Dynamics and Control* 31 (7):2398–2412. URL https://doi.org/10.1016/j.jedc.2006.07.006.

Stock, James H. 1994. "Unit Roots, Structural Breaks and Trends." In *Handbook of Econometrics, Volume 4*, edited by Robert F. Engle and Daniel L. McFadden, chap. 46. Elsevier Science B.V., 247–284. URL https://doi.org/10.1016/S1573-4412(05)80015-7.

———. 2019. "Long Term Growth Rates: Estimation and Uncertainty." Presentation for Social Security Technical Panel on Assumptions and Methods, March 29, 2019. URL https://www.ssab.gov/wp-content/uploads/2021/03/7-SLIDES-Stock-Long-Term-Growth-Rates.pdf.

Vogelsang, Timothy J. 2018. "Comment on "HAR Inference: Recommendations for Practice"." *Journal of Business & Economic Statistics* 36 (4):569–573. URL https://doi.org/10.1080/07350015.2018.1497503.

Zhou, Zhou, Zhiwei Xu, and Wei Biao Wu. 2010. "Long-Term Prediction Intervals of Time Series." *IEEE Transactions in Information Theory* 56 (3):1436–1446. URL https://doi.org/10.1109/TIT.2009.2039158.