**Federal Reserve Bank of Cleveland Working Paper Series**

# An Empirical Evaluation of Some Long-Horizon Macroeconomic Forecasts

Kurt G. Lunsford and Kenneth D. West

Working Paper No. 24-20

September 2024

**Suggested citation:** Lunsford, Kurt G. and Kenneth D. West. 2024. "An Empirical Evaluation of Some Long-Horizon Macroeconomic Forecasts." Working Paper No. 24-20. Federal Reserve Bank of Cleveland. https://doi.org/10.26509/frbc-wp-202420.

**Federal Reserve Bank of Cleveland Working Paper Series**
ISSN: 2573-7953

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

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

September 4, 2024

**Abstract**

We use long-run annual cross-country data for 10 macroeconomic variables to evaluate the long-horizon forecast distributions of six forecasting models. The variables we use range from ones having little serial correlation to ones having persistence consistent with unit roots. Our forecasting models include simple time series models and frequency domain models developed in Müller and Watson (2016). For plausibly stationary variables, an AR(1) model and a frequency domain model that does not require the user to take a stand on the order of integration appear reasonably well calibrated for forecast horizons of 10 and 25 years. For plausibly non-stationary variables, a random walk model appears reasonably well calibrated for forecast horizons of 10 and 25 years. No model appears well calibrated for forecast horizons of 50 years.

**Keywords:** Fractional Integration, Forecast Interval, Low Frequency

**JEL Codes:** C22, C53, E17

[*] We are grateful to conference and seminar participants at the Federal Reserve Banks of Cleveland and New York, the Bank of England, the 2023 International Symposium on Forecasting, and the University of East Anglia. West acknowledges grant RDR18000003 from the US Social Security Administration (SSA) funded as part of the Retirement and Disability Research Consortium. The views, opinions, and conclusions expressed are solely those of the authors and do not represent the views, opinions or policy of the SSA, any agency of the federal government, the Federal Reserve Bank of Cleveland, the Federal Reserve System, or NBER. Neither the United States government nor any agency thereof, nor any of its employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of the contents of this report. Reference herein to any specific commercial product, process or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply endorsement, recommendation or favoring by the United States government or any agency thereof.

[†] 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; kdwest@wisc.edu

# 1 Introduction

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

In this paper, we study the forecast distributions of six univariate forecasting models for horizons of up to 50 years. We use three simple time series models and three frequency domain models. Our time series models consist of an independent and identically distributed (iid) model, an autoregressive model of order one (AR(1)), and a random walk model. We choose these models because they are commonly used and cover a range of potential data persistence. Our frequency domain models are from Müller and Watson (2016). These models produce forecast distributions using low-frequency information in the data and were specifically developed for measuring uncertainty around long-horizon forecasts. They are an integrated of order 0 model (MW0), an integrated of order 1 model (MW1), and a fractionally integrated model (MWd). For this last model, the order of integration, $d$, can be between $-0.4$ and 1, and the user does not have to take a stand on this parameter.

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

[1] The 2024 report of the Board of the Trustees of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds is available at `https://www.ssa.gov/oact/TR/2024/index.html`, accessed on May 7, 2024. In an earlier working paper, Lunsford and West (2021) compare the high- and low-cost scenarios to long-horizon forecast intervals.

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

Let $x_t$ denote one of the series. We use the six models to construct forecast distributions for the long-horizon average, $(x_{\tau+1} + \cdots + x_{\tau+h})/h$, for $h = 10$, 25, and 50 years. Then, our main focus is to evaluate the calibration of each model's forecast distribution for each $h$. We use the forecast distributions to construct 68 percent nominal forecast intervals and, in our pseudo out-of-sample analysis, we compute coverage rates by calculating how often the realized long-horizon averages fall in the interval. A correctly calibrated model has a coverage rate of 0.68.

For plausibly stationary variables such as per capita real GDP growth, the AR(1), MW0, and MWd models appear reasonably well calibrated for $h = 10$. The AR(1) and MWd models also work well for $h = 25$. Overall, looking across our many samples of forecasts of stationary series, half of the AR(1) model's coverage rates are between 0.61 and 0.80 at $h = 10$ and between 0.53 and 0.78 at $h = 25$. Results for the MWd model are similar. For $h = 50$, no model appears to be well calibrated.

For plausibly non-stationary variables such as short-term nominal interest rates, the random walk and MW1 models appear reasonably well calibrated for $h = 10$. Coverage rates are typically between 0.58 and 0.77 for the random walk model and between 0.51 and 0.71 for the MW1 model. For $h = 25$, the random walk model's coverage rates are quite spread out, but its probability integral transforms appear reasonably uniform, suggesting okay calibration for $h = 25$. For $h = 50$, it is again the case that no model appears well calibrated.

One may be concerned that the well-calibrated models have relatively long interval lengths. Investigation of the length of forecast intervals reveals that the best calibrated models sometimes have larger interval lengths than more poorly calibrated models. Thus, the spread of the forecast distribution is important for calibration. To assess the trade-off of calibration against interval length, we use Winkler (1972) interval scores.[3] We find that the well-calibrated models also have relatively good Winkler scores.

While our interest is in whole forecast distributions, we have found that focusing one component of a distribution, the 68 percent forecast interval, facilitates communicating the results for our 6 models and 136 data series. In an online appendix, we report probability integral transforms and continuous ranked probability scores for the forecast distributions. Aside from the probability

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

integral transforms informing our assessment of the random walk model for $h = 25$ for plausibly non-stationary variables as noted above, these statistics are consistent with our forecast interval results. In the online appendix, we also report the bias, root mean square prediction errors, and mean absolute prediction errors of point forecasts.

To our knowledge, we are the first to evaluate long-horizon forecasting models on a variety of macroeconomic variables that cover a wide degree of persistence. In doing so, we show the value of long-run cross-country databases (Jordà, Schularick, and Taylor, 2017; Jordà et al., 2019; Bergeaud, Cette, and Lecat, 2016) for forecasting research. We highlight two further implications of our results. First, reasonably well-calibrated forecasting models are available for horizons between 10 and 25 years. Second, we do not find a one-size-fits-all model for all variables. Neither of the models that are best calibrated for plausibly stationary variables for $h = 10$ and 25 (AR(1) and MWd) is the model that is best calibrated for plausibly non-stationary variables (random walk).

The vast majority of theoretical and empirical work on forecasting considers horizons far shorter than ours. In addition to Müller and Watson (2016), exceptions include the following. Granger and Jeon (2007) proposed that long-horizon forecasts be based on very simple parametric models. They graphically analyze how well their simple models do in terms of forecasting the log level of GDP 10 and 15 years ahead, finding that realized GDP almost always falls within 90 percent confidence intervals. We, too, construct some of our forecasts with very simple parametric models.[4] 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 percent forecast intervals, with more accurate coverage at shorter rather than longer horizons. Karmakar, Chudý, and Wu (2022) evaluate a long-horizon forecasting model with many predictors. They evaluate their model with high-frequency electricity prices and find that the model we call MWd has good coverage up to 17 weeks ahead. Pesaran, Pettenuzzo, and Timmermann (2006) use a univariate model with regime changes to forecast interest rates 60 months ahead, finding that Bayesian procedures work well. Christensen, Gillingham, and Nordhaus (2018) use Müller and Watson's (2016) models and a survey of experts to estimate uncertainty around long-run GDP growth forecasts but do not evaluate these measures of uncertainty. Giannone, Lenza, and Primiceri (2019) use vector autoregressions to

[4] 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.

predict macroeconomic variables out to 10 years but only evaluate point forecasts. Müller, Stock, and Watson (2022) construct and evaluate a low-frequency model of cross-country GDP growth but do not study the 9 other variables in our data set.

Several of the papers cited above use multivariate models and models of structural change. We leave for future research evaluation of more complex forecasting models. By showing that univariate and constant parameter models can be reasonably well calibrated across many variables out to 25 years, we document readily available models for practitioners and also highlight models that can be used as comparative benchmarks in future research.

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

## 2 Forecasting Models

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

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

We use each model to produce a forecast distribution. We first discuss the iid, AR(1), and random walk models and then discuss the MW0, MW1, and MWd models. For the iid, AR(1), and random walk models, we treat the future average realization, $\bar{x}_{\tau,h}$, as normally distributed with a conditional mean and variance $f_{\tau,h}$ and $V_{\tau,h}$. That is, we use $\bar{x}_{\tau,h} \sim N(f_{\tau,h}, V_{\tau,h})$. The adjective "conditional" means dependent on data on $x_t$ realized up to time $\tau$. The three different models have different forms for $f_{\tau,h}$ and $V_{\tau,h}$, which we provide in the online appendix. We make five comments here. First, we use estimates of $f_{\tau,h}$ as our point forecasts.[5] Second, we compute $V_{\tau,h}$ parametrically assuming that the data are iid (iid model), or that the AR(1) innovation is iid (AR(1) model), or that $\Delta x_t$ is iid (random walk model). Third, while the AR(1) mean takes a familiar form, it relies on a constant term and slope coefficient that we bias adjust following Yamamoto and Kunitomo (1984). Fourth, for the AR(1) model, we construct $f_{\tau,h}$ and $V_{\tau,h}$ in such a way that if the bias-adjusted slope coefficient is exactly zero, then $V_{\tau,h}$ is exactly as in the iid model apart from a degrees of freedom adjustment. Fifth, if the estimated AR(1) slope coefficient is greater than or equal to 1, we produce forecast distributions with the random walk model instead of the AR(1) model.

The Müller and Watson (2016) forecasting models are based on extracting low-frequency patterns from the sample $\{x_1, \ldots, x_\tau\}$ by using a small number, $q$, of slowly cycling cosine waves. We discuss the choice of $q$ in Section 4 below. For each of MW0, MW1, and MWd, we present formulas and discussion in the online appendix. We make three comments here. First, the MW0 and MW1 forecast intervals have generalized Student-$t$ distributions with $q$ degrees of freedom. Second, the point forecasts of the MW0 and iid models are the same, namely, the in-sample average; forecast intervals are constructed differently and potentially (and in practice) are quite different. Third, for MWd, we follow Müller and Watson (2016) and treat $d$ as unknown and use a Bayesian approach to construct the forecast density. We set a grid of potential values of $d$, $\{-0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8, 1\}$, and use a uniform prior with equal mass on each grid point. The resulting Bayes predictive density is a weighted average of generalized Student-$t$ densities with $q$ degrees of freedom.

[5] The iid, AR(1), and random walk models as well as the MW0 and MW1 models have standard, symmetric forecast distributions with a mean equal to the median. The MWd model will have a potentially asymmetric forecast distribution with a mean and median that will not generally be equal.

# 3 Data

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

The source for all variables except labor productivity growth is the Macrohistory Database (Jordà, Schularick, and Taylor, 2017; Jordà et al., 2019), which has data beginning as early as 1870 and running through 2020. Productivity data, running from 1871 to 2022, are from the Long-Term Productivity Database (Bergeaud, Cette, and Lecat, 2016).[6] We convert data with obvious upward trends, such as per capita real GDP, to percentage growth rates as 100 × log differences. We construct the real exchange rate from the Macrohistory Database's bilateral nominal US dollar exchange rates using its CPI data for the US and the relevant other country.

Details on sample periods and country coverage are in the online appendix. Here is a summary. Most series run 1870s-2020. The number of countries for a given series runs from a low of 7 (investment to GDP) to a high of 18 (labor productivity growth). Table 4.1 gives the exact number for each variable. To give an idea of country coverage, here are the 17 countries for per capita real GDP growth, CPI inflation, and population growth: Australia, Belgium, Canada, Switzerland, Germany, Denmark, Spain, Finland, France, Great Britain, Italy, Japan, Netherlands, Norway, Portugal, Sweden, and the US. With the exception of labor productivity growth, country coverage for all other series is a subset of these 17. The labor productivity data, which cover 18 countries, omits Australia and Portugal but adds Chile, Greece, and Ireland.

# 4 Pseudo Out-of-Sample Analysis

We evaluate the forecasting models with the following pseudo out-of-sample analysis. $R = 48$ is our initial sample size. For example, for all countries for per capita GDP growth, the initial rolling and recursive sample is 1871-1918. We use this sample to estimate model parameters and compute

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

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

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

For each estimation sample in our pseudo out-of-sample analysis, we estimate the AR(1) model's bias-adjusted slope coefficient and the maximum likelihood value of $d$ from the MWd model.[7] Hence, we have a large number of estimated persistence parameters for each variable: $P$ forecasts times 2 sample schemes (rolling and recursive) for each country. Table 4.1 summarizes these persistence estimates by showing the median estimated value and the interquartile range (IQR) of estimated values for each of our 10 variables.

Table 4.1 displays the variables in plausibly stationary (lines (2) to (7)) or plausibly non-stationary (lines (9) to (12)) groups. Plausibly non-stationary variables are those whose estimates of the fractional integration parameter $d$ tend to be above 0.5, plus the real exchange rate, whose median $d$ was 0.4 but which was very highly serially correlated.[8] We divide the variables into plausibly stationary and plausibly non-stationary because our discussion below finds this useful for presentational purposes: reducing our hundreds of forecasts to these two groups allows us to highlight central patterns.

[7] We give estimation details in the online appendix.

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

Table 4.1: Persistence statistics estimated in the $h = 10$ pseudo out-of-sample analysis

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

Notes:

1. The "no. of estimates" column gives the total number of estimates for the $h = 10$ horizon. For example, for $h = 10$, GDP growth has 17 countries with 2 sample schemes per country and $P = 93$ forecasts per sample scheme. This gives $17 \times 2 \times 93 = 3162$ total forecasts and estimates.

2. The "$\tilde{\rho}_1$" column shows the median and IQR (in parentheses) of the estimates of the bias-adjusted AR(1) slope coefficients.

3. The "$\hat{d}$" column shows the median and IQR (in parentheses) of the maximum likelihood estimates of $d$. The maximum likelihood value is chosen from the grid $\{-0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8, 1\}$.

Table 4.1 shows that our variables have a wide range of persistence estimates, ranging from near zero serial correlation in productivity growth to apparent unit roots in long-term nominal interest rates. Hence, our pseudo out-of-sample analysis should be informative about how the forecasting models perform over the range persistence values encountered in macroeconomic variables.

The main focus of our pseudo out-of-sample analysis will be to evaluate the calibration of each model's forecast distributions. To do this, we construct 68 percent equal-tailed forecast intervals from the 16th and 84th percentiles of the forecast distributions.[9] For a given variable, country, and sampling scheme, let $FI^{68}_{\tau,h}$ denote the 68 percent forecast interval made with sample $\{x_t\}_{t=1}^{\tau}$ for horizon $h$. We then compute the coverage rate, $P^{-1}\sum_{\tau=R}^{T-h}\mathbf{1}(\bar{x}_{\tau,h} \in FI^{68}_{\tau,h})$ with $\mathbf{1}(\cdot)$ being the indicator function, for each variable, country, and sampling scheme. A well-calibrated forecasting model will have a coverage rate of 0.68.

In our pseudo out-of-sample analysis, we also compute average forecast interval lengths and average Winkler (1972) interval scores. The Winkler score is a proper and consistent scoring (or loss) function for equal-tailed forecast intervals (Gneiting and Raftery, 2007; Brehmer and Gneiting, 2021), and we use it to assess the trade-off between the calibration and the interval length of a forecasting model.[10]

We make three final comments before presenting our empirical results. First, we compute coverage rates and the other statistics for 136 series (10 variables with 7 to 18 countries per variable) times 2 sampling schemes, giving 272 values of each statistic. Doubling the number of series in lines (8) and (13) in Table 4.1 to reflect the 2 sampling schemes, we will have 184 values of each statistic for stationary series and 88 for non-stationary series. To summarize this large number of statistics, we report medians and IQRs over the 184 and 88 observations in each of these two groups.

Second, our pseudo out-of-sample forecasts have a heavily overlapping nature. Using per capita GDP growth as an example, the number of forecasts that we have for each country and sampling scheme is $P = 93$ for $h = 10$, $P = 78$ for $h = 25$, and $P = 53$ for $h = 50$ as noted above. Thus, we only have at most 10 non-overlapping forecasts for $h = 10$, 4 non-overlapping forecasts for $h = 25$,

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

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

and 2 non-overlapping forecasts for $h = 50$. These small samples lead us to follow Müller and Watson (2018) and our own work (Lunsford and West, 2019) in using 68 percent forecast intervals. Because realizations in the tails of a distribution are infrequent, observing behavior in the tails, as is required for evaluation of 90 or 95 percent intervals, would likely require longer samples than we have available.

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

# 5 Pseudo Out-of-Sample Results

## 5.1 Coverage Rates of 68 Percent Forecast Intervals

For each forecasting model and forecast horizon, we compute 184 coverage rates for the plausibly stationary variables and 88 coverage rates for the plausibly non-stationary variables. Table 5.1 summarizes these coverage rates by reporting the median coverage rate and IQR of the coverage rates for each forecasting model and forecast horizon across the two data categories.

For arguably stationary variables, we see in columns (3a) and (3b) of Table 5.1 that the MW0, AR(1), and MWd models are reasonably well calibrated for $h = 10$. We show in the online appendix that the iid model performs comparably to these three models for the subset of arguably stationary variables that have very little serial correlation (i.e., for GDP growth, productivity growth, and equity returns). We say "reasonably well" calibrated in light of the huge literature documenting the difficulties of inference when overlapping data are used (see, e.g., Lazarus et al. (2018)). In the context of this literature, it is a good outcome for a nominal 68 percent forecast interval to yield actual coverage rates of 0.64 to 0.73 (the range of median values for the MW0, AR(1), and MWd models for $h = 10$ in column (3a) of Table 5.1). The IQRs in column (3b) indicate that half or more of the samples yielded coverage rates tolerably close to 0.68. By contrast, the random walk and MW1 models perform poorly.

We see in columns (3a) and (3b) that as $h$ increases from 10 to 25 to 50, performance degrades for all models. For the models that performed well for $h = 10$, actual size falls as the horizon

Table 5.1: Coverage rates of nominal 68 percent forecast intervals: medians and IQRs

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

Notes:

1. See Table 4.1 for categorization of variables as plausibly stationary or plausibly non-stationary.

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

3. The number of samples for each group in row (2) is the sum of the number of variables in that group times the two sampling schemes (rolling and recursive).

4. "Median coverage" refers to the median across the number of samples in row (2). Consider the figure of 0.50 in row (4), column (3a). Per columns (1) and (2), the iid model is used to produce 184 sets (92 series × 2 sampling schemes) of pseudo out-of-sample forecasts for $h = 10$. Actual coverage rates of forecast intervals with nominal 68 percent coverage are computed for each set of pseudo out-of-sample forecasts. 0.50 is the median value across these 184 coverage rates.

5. IQR is interquartile range. It shows the 25th and 75th percentiles of the coverage rates across the number of samples in row (2). These percentiles are computed analogously to the median in the previous note.

increases. The median coverage rates in column (3a) for the AR(1) and MWd models stay quite close to 0.68 for $h = 25$; however, the lengths of their IQRs in column (3b) have gone up, indicating wider dispersion away from 0.68. For $h = 50$, it is hard to describe any model as well calibrated. For MWd, probably the best performing model for $h = 50$, we see from the median and upper end of the IQR that only 25 percent of the samples had actual coverage between 0.58 and 0.81.

We also compute the probability integral transforms of the forecast distributions for each model, forecast horizon, and data series. Let $F_{\tau,h}(\cdot)$ be the forecast cumulative distribution function made with sample $\{x_t\}_{t=1}^{\tau}$ for horizon $h$. Then, $F_{\tau,h}(\bar{x}_{\tau,h})$ is the probability integral transform. Well-calibrated models have probability integral transforms that are uniform on $[0, 1]$ (Dawid, 1984; Diebold, Gunther, and Tay, 1998). We find that the probability integral transforms are consistent with our coverage rate results and so report the probability integral transforms in the online appendix. For the plausibly stationary variables, the probability integral transforms are reasonably close to uniform for the MW0, AR(1), and MWd models for $h = 10$ and are tolerably close to uniform for the AR(1) and MWd models for $h = 25$. No model appears to have probability integral transforms that are even roughly uniform for $h = 50$.

We turn now to plausibly non-stationary variables. For $h = 10$, we see in columns (4a) and (4b) that the random walk and maybe the MW1 model perform tolerably, with median coverage rates that are 0.70 and 0.63. Unsurprisingly, the iid and MW0 models perform quite poorly for this set of highly persistent variables. The AR(1) model (perhaps surprisingly) performs comparably to the MWd model. Indeed, the AR(1) and MWd models often perform similarly across data types, forecast horizons, and evaluation statistics, and we typically discuss these models together. For longer horizons, performance degrades for all models. For $h = 25$ and 50, the median coverage rates are tolerable for the random walk and MW1 models (e.g., 0.72 for random walk for $h = 25$ and 0.64 for MW1 for $h = 50$), but the IQRs are quite long.

The probability integral transforms (presented in the online appendix) reinforce our coverage rate results for $h = 10$. The uniformity of the probability integral transforms for the random walk and MW1 models for the plausibly non-stationary variables appears similar to the uniformity of the probability integral transforms for the AR(1) model for the plausibly stationary variables. For $h = 25$, the probability integral transforms for the random walk model for the plausibly non-stationary variables appear reasonably uniform and are similar to the uniformity of the probability integral transforms for the AR(1) and MWd models for the plausibly stationary variables. No model appears to have uniform probability integral transforms for $h = 50$ for the plausibly non-stationary variables.

We conclude that for $h = 10$ or $h = 25$, the AR1 and MWd models are tolerably well calibrated for arguably stationary variables. For arguably non-stationary variables, the random walk and MW1 models are tolerably well calibrated for $h = 10$, and the probability integral transforms indicate that the random walk model is tolerably well calibrated for $h = 25$. For $h = 50$, we do not find a well-calibrated model for either type of variable.

### 5.2 Interval Lengths

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

Table 5.2 summarizes our results. For both the plausibly stationary and plausibly non-stationary variables, the models can generally be ordered from smallest to largest relative length (with a few exceptions) as follows: iid, MW0, AR(1), MWd, MW1, and random walk. The iid model almost always has the smallest average interval length.

To relate the results in Table 5.2 to those in Table 5.1, consider first the iid and MW0 models. These two models have identical point forecasts by construction. In Table 5.1, judging by either median or IQR, MW0 is better calibrated. Thus, this is a case where, unambiguously, better calibration comes from generally larger interval lengths. Next consider the MW0, AR(1), and MWd models. For $h = 10$, the three models are similarly well calibrated for the stationary variables while having similar interval lengths. For $h = 25$, the AR(1) and MWd models remain well calibrated,

Table 5.2: Average lengths: medians and IQRs of relative values and fraction with minimum value

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

Notes:

1. See Table 4.1 for categorization of variables as plausibly stationary or plausibly non-stationary.

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

3. The number of samples for each group in row (2) is the sum of the number of variables in that group times the two sampling schemes (rolling and recursive).

4. For each model in each sample, average interval length is expressed relative to average interval length for the iid model in that sample. Medians and IQRs (interquartile ranges) of the resulting relative average interval lengths were constructed as described in the notes to Table 5.1.

5. “fraction with min length” reports the fraction of the samples for a given horizon in which the corresponding model has the lowest average length among the six models.

while the MW0 model becomes undersized. Table 5.2 shows that the AR(1) and MWd models' relative interval lengths increase slightly from $h = 10$ to $h = 25$, while the MW0 model's relative lengths do not. This example again shows that calibration is related to interval length.

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

### 5.3 Winkler Interval Scores

We have just seen that the random walk model has relatively long intervals and the iid model has relatively short intervals. In our discussion of calibration in Subsection 5.1, the random walk model received some support (for non-stationary variables), while the iid model received little support for either the stationary or the non-stationary series. These two models highlight that forecasters may face a trade-off between (1) models that are reasonably well calibrated but have relatively long forecast intervals, and (2) models that are poorly calibrated but have relatively short forecast intervals.[11] To quantify this trade-off, we use Winkler (1972) interval scores to assign a loss to each model (for each variable, forecast horizon, and sampling scheme). We compute Winkler scores as follows. Let $u_{\tau,h}$ and $\ell_{\tau,h}$ be the upper (84th percentile) and lower (16th percentile) bounds of the 68 percent forecast interval made using $\{x_t\}_{t=1}^{\tau}$ for horizon $h$. Then, the Winkler score is

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

in which $\mathbf{1}(\cdot)$ is the indicator function. We average these Winkler scores over $\tau = R, \ldots, T-h$ and all of our discussion is in terms of these averages. This loss penalizes both interval length and inaccurate values for the 16th and 84th percentiles. Hence, it can be used to adjudicate the calibration-length trade-off. As we noted above, this measure is both consistent and proper (Gneiting and Raftery, 2007; Brehmer and Gneiting, 2021). Lower Winkler scores are better.

As with average interval lengths, we compute relative (to the iid model) average Winkler scores for each variable, country, and sampling scheme in order to divide out units. We report medians and IQRs of these relative values.[12] We also compute the fraction of samples in which a given model has the smallest average Winkler score.

[11] This is a well-known trade-off that may also be referred to as being between calibration and "sharpness" (Gneiting, Balabdaoui, and Raftery, 2007; Gneiting and Katzfuss, 2014).

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

Table 5.3 shows these results. For the plausibly stationary variables, at $h = 10$, we see in columns (3a) and (3b) of Table 5.3 that the relative Winkler scores for the iid, MW0, AR(1), and MWd models are quite similar. We also see in column (3c) that the AR(1) and MWd models produce the lowest Winkler scores at rates similar to those of the iid model. Taken together with the coverage rate results in Table 5.1, we find that the MW0, AR(1), and MWd models are better calibrated than the iid model but without a noticeable increase in average Winkler scores. That Winkler performance for iid is comparable to MW0, AR(1) and MWd, despite the poor results for coverage, thus reflects the contribution of interval length to a Winkler score. Or, to state the flip side of the same point, the longer intervals of the MW0, AR(1), and MWd models do not outweigh their better calibration according to the metric of Winkler scores.

For the stationary variables, at $h = 25$, Table 5.3 shows that the iid, MW0, AR(1), and MWd models again have similar relative Winkler scores and similar rates of producing the lowest Winkler score. We saw above that the AR(1) and MWd models are the best calibrated at this horizon. Hence, forecasters can use the AR(1) and MWd models to achieve reasonably good calibration with no degradation in Winkler loss relative to any of our other models.

For the plausibly stationary variables, the random walk and MW1 models usually have the highest Winkler scores at all horizons. These high values result from wide forecast intervals and poor calibration. For example, for $h = 10$, column (3a) in Table 5.2 shows that the median relative interval length for the random walk is 6.64 (as opposed to, for example, 1.39 for MW0). Large interval lengths are consistent with the random walk and MW1 models generally having interval coverage well above 68 percent in Table 5.1.

For the plausibly non-stationary variables for $h = 10$, we see in columns (4a) and (4b) of Table 5.3 that the relative Winkler scores for the AR(1), MWd, random walk, and MW1 models are similar. The relative scores for the iid and MW0 models are noticeably higher. Taken together with the coverage rate results in Table 5.1, we find that the random walk and MW1 models are better calibrated than the AR(1) and MWd models but without a noticeable increase in relative Winkler scores. Column (4c) in Table 5.3 shows that the random walk model often produces the lowest average Winkler scores.

For the plausibly non-stationary variables, at $h = 25$, the relative Winkler scores for the random walk model appear slightly higher than those for the AR(1) and MWd models. However, the

Table 5.3: Winkler scores: medians and IQRs of relative values and fraction with minimum value

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

Notes:

1. See Table 4.1 for categorization of variables as plausibly stationary or plausibly non-stationary.
2. In the list of models in column (2), we use the shorthand “RW” for the random walk model.
3. The number of samples for each group in row (2) is the sum of the number of variables in that group times the two sampling schemes (rolling and recursive).
4. For each model in each sample, the average Winkler score is expressed relative to the average Winkler score for the iid model in that sample. Medians and IQRs (interquartile ranges) of the resulting relative average Winkler scores were constructed as described in the notes to Table 5.1.
5. “fraction with min Winkler” reports the fraction of the samples for a given horizon in which the corresponding model has the lowest average Winkler score among the six models.

random walk model produces the lowest Winkler scores in the highest fraction of samples. Overall, the random walk model's scores do not appear materially worse than those for the AR(1) and MWd models despite the random walk model's materially larger interval lengths.

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

### 5.4 Point Forecasts

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

For the stationary variables at all horizons, the AR(1) and MWd models have similar RMSPEs and MAPEs relative to the iid and MW0 models (which have the same point forecasts). The random walk and MW1 models have noticeably higher RMSPEs and MAPEs relative to the iid and MW0 models.[13] Hence, practitioners can use the AR(1) and MWd models to achieve good calibration (as documented in Subsection 5.1) without sacrificing point forecast accuracy or Winker loss.

For the non-stationary variables for $h = 10$, the AR(1), MWd, random walk, and MW1 models have low and similar RMSPEs and MAPEs relative to the iid and MW0 models, again paralleling our Winkler score results. Further, the random walk model produces the lowest RMSPEs and MAPEs more often than the other models, again indicating that the random walk model generally performs well for non-stationary variables for $h = 10$. The point forecast results for non-stationary variables for $h = 25$ parallel the Winkler score results less closely. The AR(1) and MWd models

[13] If we measure the quality of point forecasts by bias rather than RMSPE, the random walk and MW1 models do quite well for reasons explained and illustrated in Lunsford and West (2023).

have similar RMSPEs and MAPEs relative to the iid and MW0 models, but the random walk and MW1 models have slightly higher relative values. This result indicates that while the random walk model appears reasonably well calibrated for non-stationary variables for $h = 25$, its point forecasts may be less accurate than those from several of our other models.

## 6 Conclusions

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

We use long-run annual data for 10 macroeconomic variables with up to 18 countries per variable, allowing us to construct pseudo out-of-sample forecasts for 136 different series with widely varying degrees of persistence. For plausibly stationary variables, an AR(1) model and Müller and Watson's (2016) MWd model appear reasonably well calibrated for forecast horizons of both 10 and 25 years. For plausibly non-stationary variables, a random walk model appears reasonably well calibrated for forecast horizons of both 10 and 25 years.

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

## References

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.

Brehmer, Jonas R. and Tilmann Gneiting. 2021. "Scoring Interval Forecasts: Equal-tailed, Shortest, and Modal Interval." *Bernoulli* 27 (3):1993–2010. URL https://doi.org/10.3150/20-BEJ1298.

Christensen, P., K. Gillingham, and W. Nordhaus. 2018. "Uncertainty in Forecasts of Long-Run Economic Growth." *Proceedings of the National Academy of Sciences* 115 (21):5409–5414. URL https://doi.org/10.1073/pnas.1713628115.

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.

Dawid, Alexander Philip. 1984. "Statistical Theory: The Prequential Approach." *Journal of the Royal Statistical Society: Series A (General)* 147 (2):278–292. URL https://doi.org/10.2307/2981683.

Diebold, Francis X., Todd A. Gunther, and Anthony S. Tay. 1998. "Evaluating Density Forecasts with Applications to Financial Risk Management." *International Economic Review* 39 (4):863–883. URL https://doi.org/10.2307/2527342.

Giannone, Domenico, Michele Lenza, and Giorgio E. Primiceri. 2019. "Priors for the Long Run." *Journal of the American Statistical Association* 114 (526):565–580. URL https://doi.org/10.1080/01621459.2018.1483826.

Gneiting, Tilmann, Fadoua Balabdaoui, and Adrian E. Raftery. 2007. "Probabilistic Forecasts, Calibration and Sharpness." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 69 (2):243–268. URL https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Gneiting, Tilmann and Matthias Katzfuss. 2014. "Probabilistic Forecasting." *Annual Review of Statistics and Its Application* 1:125–151. URL https://doi.org/10.1146/annurev-statistics-062713-085831.

Gneiting, Tilmann and Adrian E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102 (477):359–378. URL https://doi.org/10.1198/016214506000001437.

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.

Haubrich, Joseph, George Pennacchi, and Peter Ritchken. 2012. "Inflation Expectations, Real Rates, and Risk Premia: Evidence from Inflation Swaps." *Review of Financial Studies* 25 (5):1588–1629. URL https://doi.org/10.1093/rfs/hhs003.

Jordà, Òscar, Katharina Knoll, Dmitry Kuvshinov, Moritz Schularick, and Alan M. Taylor. 2019. "The Rate of Return on Everything, 1870-2015." *Quarterly Journal of Economics* 134 (3):1225–1298. URL https://doi.org/10.1093/qje/qjz012.

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.

Karmakar, Sayar, Marek Chudý, and Wei Biao Wu. 2022. "Long-Term Prediction Intervals with Many Covariates." *Journal of Time Series Analysis* 43 (4):587–609. URL https://doi.org/10.1111/jtsa.12629.

Kitsul, Yuriy and Jonathan H. Wright. 2013. "The Economics of Options-Implied Inflation Probability Density Functions." *Journal of Financial Economics* 110 (3):696–711. URL https://doi.org/10.1016/j.jfineco.2013.08.013.

Lazarus, Eben, Daniel J. Lewis, James H. Stock, and Mark W. Watson. 2018. "HAR Inference: Recommendations for Practice." *Journal of Business & Economic Statistics* 36 (4):541–559. URL https://doi.org/10.1080/07350015.2018.1506926.

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.

———. 2021. "An Empirical Evaluation of Some Long-Horizon Macroeconomic Forecasts." NBER Retirement and Disability Research Center Paper No. NB21-18. URL https://www.nber.org/programs-projects/projects-and-centers/retirement-and-disability-research-center/center-papers/nb21-18.

———. 2023. "Random Walk Forecasts of Stationary Processes Have Low Bias." Federal Reserve Bank of Cleveland Working Paper No. 23-18. URL https://doi.org/10.26509/frbc-wp-202318.

Makridakis, Spyros, Rob J. Hyndman, and Fotios Petropoulos. 2020. "Forecasting in Social Settings: The State of the Art." *International Journal of Forecasting* 36 (1):15–28. URL https://doi.org/10.1016/j.ijforecast.2019.05.011.

Müller, Ulrich K., James H. Stock, and Mark W. Watson. 2022. "An Econometric Model of International Growth Dynamics for Long-Horizon Forecasting." *Review of Economics and Statistics* 104 (5):857–876. URL https://doi.org/10.1162/rest_a_00997.

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.

Nordhaus, William. 2018. "Projections and Uncertainties about Climate Change in an Era of Minimal Climate Policies." *American Economic Journal: Economic Policy* 10 (3):333–360. URL https://doi.org/10.1257/pol.20170046.

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.

Pesaran, M. Hashem, Davide Pettenuzzo, and Allan Timmermann. 2006. “Forecasting Time Series Subject to Multiple Structural Breaks.” *Review of Economic Studies* 73 (4):1057–1084. URL https://doi.org/10.1111/j.1467-937X.2006.00408.x.

West, Kenneth D. 2006. “Forecast Evaluation.” In *Handbook of Economic Forecasting, Volume 1*, edited by Graham Elliott, Clive W. J. Granger, and Allan Timmermann, chap. 3. Elsevier B.V., 99–134. URL https://doi.org/10.1016/S1574-0706(05)01003-7.

Winkler, Robert L. 1972. “A Decision-Theoretic Approach to Interval Estimation.” *Journal of the American Statistical Association* 67 (337):187–191. URL https://doi.org/10.1080/01621459.1972.10481224.

Yamamoto, Taku and Naoto Kunitomo. 1984. “Asymptotic Bias of the Least Squares Estimator for Multivariate Autoregressive Models.” *Annals of the Institute of Statistical Mathematics* 36:419–430. URL https://doi.org/10.1007/BF02481980.

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