Multistep Forecast Averaging with Stochastic and Deterministic Trends

Mohitosh Kejriwal* **Linh Nguyen†** **Xuewen Yu‡**

Purdue University *Purdue University* *Purdue University*

April 21, 2022

Abstract

This paper presents a new approach to constructing multistep combination forecasts in a nonstationary framework with stochastic and deterministic trends. Existing forecast combination approaches in the stationary setup typically target the in-sample asymptotic mean squared error (AMSE) relying on its approximate equivalence with the asymptotic forecast risk (AFR). Such equivalence, however, breaks down in a nonstationary setup. This paper develops combination forecasts based on minimizing an Accumulated Prediction Errors (APE) criterion that directly targets the AFR and remains valid whether the time series is stationary or not. We show that the performance of APE-weighted forecasts is close to that of the optimal, infeasible combination forecasts. Monte Carlo experiments are used to demonstrate the finite sample efficacy of the proposed procedure relative to Mallows/Cross-Validation weighting that target the AMSE. An application to forecasting US macroeconomic time series demonstrates the relevance of the proposed method in practice.

Keywords: averaging, combination, cross-validation, Mallows, unit root, accumulated prediction errors.

JEL Classification: C22, C53

*Krannert School of Management, Purdue University, 403 West State Street, West Lafayette IN 47907 (mkejriwa@purdue.edu).

†Krannert School of Management, Purdue University, 403 West State Street, West Lafayette IN 47907 (nguye535@purdue.edu).

‡Krannert School of Management, Purdue University, 403 West State Street, West Lafayette IN 47907 (yu656@purdue.edu).

1 Introduction

The pioneering work of Granger (1966) demonstrated that a large number of macroeconomic time series have a typical spectral shape dominated by a peak at low frequencies. This finding suggests the presence of relatively long run information in the current level of the variables which should be taken into account when modeling their time series evolution and can potentially be exploited to yield improved forecasts. One way to incorporate this long-run information in econometric modeling is through stochastic trends (unit roots) and/or deterministic trends. However, given that trends are slowly evolving, there is only limited information in any data set about how best to specify the trend or distinguish between alternative models of the trend. For instance, unit root tests often fail to reject a unit root despite the fact that theory does not postulate the presence of a unit root for many macroeconomic variables (see Elliott (2006b), for further discussion of this issue). Clements and Hendry (2001) documented, both analytically and numerically, the detrimental consequences of trend misspecification on the resulting forecasts in the presence of parameter estimation uncertainty. Specifically, they find that when the sample size increases at a faster rate than the forecast horizon, misspecifying a difference stationary process as trend stationary or vice-versa yields forecast error variances of a higher order of magnitude relative to the correctly specified model.

Notwithstanding the importance of the low frequency components and the uncertainty surrounding their precise nature, a common practice in the economic forecasting literature is to first apply a stationarity-inducing transformation (e.g., differencing or detrending) to the time series of interest and then attempt to forecast the transformed series. Consequently, most of the forecasting procedures in current use have been developed under the assumption of data stationarity. The traditional approach of Box and Jenkins (1970) transforms the data through differencing which amounts to modeling the low frequency peak in the spectrum as a zero frequency phenomenon and proceeds to forecast the transformed series using standard stationary autoregressive moving average (ARMA) models. More recently, Stock and Watson (2005, 2006) constructed a extensive database of 132 monthly macroeconomic time series over the period 1959-2003 and applied a variety of transformations to render them stationary before using a handful of common factors extracted from the data set using principal components as predictors (the so-called diffusion-index methodology). Similarly, McCracken and Ng (2016) assembled a publicly available database of 134 monthly time series referred to as FRED-MD and updated on a timely basis by the Federal Reserve Bank of St Louis. They also suggest a set of data transformations which is used to construct factor-based diffusion indexes for forecasting as well as analyze business cycle turning points.

This paper proposes a new forecast combination approach designed for forecasting a highly persistent time series that simultaneously addresses uncertainty about the presence of a stochastic trend as well as uncertainty about the nature of short-run dynamics within a unified autoregressive modeling framework. Given that uncertainty about the nature of the trend is likely to be

particularly important for longer horizons, we focus on constructing multistep forecasts instead of only one-step forecasts.[1] Existing forecast combination approaches employed in the stationary setup such as Mallows model averaging (MMA) and cross-validation (CV) weighting typically target the in-sample asymptotic mean squared error (AMSE) relying on its approximate equivalence with the asymptotic forecast risk (AFR) [e.g., Hansen, 2008; Hansen, 2010b; Liao and Tsay, 2020]. Such equivalence, however, breaks down in a nonstationary setup. Hansen (2010a) shows, within a local-to-unity framework, that the AMSE of unrestricted as well as restricted (imposing a unit root) one-step ahead forecasts are different from the corresponding expressions for their AFR in autoregressive models with a general lag order and a deterministically trending component (see section 3 for further discussion on the issue of equivalence or lack thereof).

To address the lack of equivalence between AMSE and AFR, we develop combination forecasts based on minimizing the so-called Accumulated Prediction Errors (APE) criterion that directly targets the AFR instead of the AMSE. Previous work in the context of model selection has shown the APE criterion to remain valid whether the process is stationary or has a unit root. Specifically, Ing (2004) shows that a normalized version of the APE converges almost surely to the AFR in the stationary case while a similar result is obtained by Ing et al. (2009) in the unit root case. Focusing on the first order autogressive case and one-step ahead forecasts, Yu et al. (2012) extend the validity of the APE to a unit root model with a deterministic time trend. Our analysis generalizes existing results by establishing the asymptotic validity of the APE for multistep forecasts in the unit root and (fixed) stationary cases, both for models with and without deterministic trends. We further show that, regardless of the presence of a unit root, the performance of APE-weighted forecasts remains close to that of the infeasible combination forecasts which assume that the optimal (i.e., AFR minimizing) weights are known. Monte Carlo experiments are used to (i) demonstrate the finite sample efficacy of the proposed procedure relative to Mallows/Cross-Validation weighting that target the AMSE; (ii) underscore the importance of accounting for uncertainty about the stochastic trend and/or the lag order. In a pseudo out-of-sample forecasting exercise applied to US monthly macroeconomic time series, we evaluate the performance of a variety of selection/combination-based approaches at horizons of one, three, six and twelve months. Consistent with the simulation results, the empirical analysis provides strong evidence in favor of a version of the advocated approach that simultaneously addresses stochastic trend and lag order uncertainty regardless of the forecast horizon considered.

Our paper is closely related to the existing literature on methods for forecasting nonstationary time series. Diebold and Kilian (2000) show that a unit root pretesting strategy can improve forecast accuracy relative to restricted or unrestricted estimation. Ng and Vogelsang (2002) found that the use of feasible generalized least squares (FGLS) estimates of the trend component can yield superior

[1] Analytically, the importance of the trend component over long horizons can be seen by noting that the trend/drift coefficient is multiplied by the forecast horizon when constructing forecasts so that any specification/estimation error is magnified linearly as the forecast horizon increases (Sampson, 1991).

forecasts relative to their ordinary least squares (OLS) counterparts. Turner (2004) recommended the use of forecasting thresholds whereby the restricted (unit root) forecast is preferred on one side of these thresholds while the unrestricted (OLS) forecast is preferred on the other. His proposal is based on median unbiased estimation of the local-to-unity parameter to determine the thresholds and is shown to dominate a unit root pretesting strategy. Ing et al. (2009) derive the AFR of plug-in and direct multistep forecasts in unit root autoregressions with a possibly unknown (finite) lag order but without a deterministic component and provide asymptotic justification for the APE criterion for selecting the best combination of model order and prediction method. Ing et al. (2012) study the impact of nonstationarity, model complexity and model misspecification on the AFR in infinite order autoregressions.

Hansen (2010a) adopts a local-to-unity framework to develop a one-step ahead combination forecast that combines forecasts from the restricted and unrestricted models with the weights obtained by minimizing a one-step Mallows criterion, designed to provide an approximately unbiased estimator of the in-sample asymptotic mean squared error. His analysis shows that the unit root pretesting strategy can be subject to high forecast risk for a range of persistence levels while his combination forecast performs favorably compared to a number of methods popular in applied work and dominates the unrestricted forecast uniformly in terms of finite sample forecast risk. Kejriwal and Yu (2021) develop improved combination forecasts that employ FGLS estimates of the trend parameters in conjunction with Mallows model averaging. Tu and Yi (2017) analyze one-step forecasting based on the Mallows averaging estimator in a cointegrated vector autoregressive model and finds that it dominates the commonly used approach that entails pretesting for cointegration (see Elliott (2006a) and Elliott and Timmermann (2016 for further discussion and references).

The present paper can be viewed as extending Hansen's (2010a) approach in two practically relevant directions. First, in addition to one-step ahead forecasts, we also analyze the statistical properties of multistep forecasts focusing on their dependence on the forecast horizon and the uncertainty pertaining to the presence of a stochastic trend in the time series. Second, in addition to Mallows weighting, we also evaluate the performance of combination forecasts based on APE/CV weights, both empirically and via simulations. Such a comparison serves to clarify the importance of directly targeting the AFR when estimating the combination weights in a nonstationary framework.

The rest of the paper is organized as follows. Section 2 presents the model and the related estimators that form the basis for the proposed combination forecasts. Section 3 analyzes the AMSE and AFR as alternative measures of forecast accuracy. Section 4 discusses the choice of combination weights based on the APE criterion. Section 5 extends the analysis to allow for lag order uncertainty in the construction of the forecasts. Monte Carlo evidence including comparisons with various existing methods are provided in Section 6. Section 7 details an empirical application to forecasting US macroeconomic time series. Section 8 offers concluding remarks and some directions for future research. Supplementary Appendices A-C (not for publication) contain the proofs, details

of forecasting methods considered, and additional simulation results.

2 Model and Estimation

We consider a univariate time series y_t generated as follows:

$$\begin{aligned} y_t &= m_t + u_t \\ m_t &= \beta_0 + \beta_1 t + ... + \beta_p t^p \\ u_t &= \alpha u_{t-1} + \alpha_1 \Delta u_{t-1} + \cdots + \alpha_k \Delta u_{t-k} + e_t \\ \alpha &= 1 + \frac{ac}{T}, \quad a = 1 - \alpha_1 - \cdots - \alpha_k, \quad c \leq 0 \end{aligned} \tag{1}$$

where $p \in \{0, 1\}$ is the order of the trend component and the stochastic component u_t follows a finite order autoregressive process of order $(k+1)$ process driven by the innovations e_t. The uncertainty about the stochastic trend is captured by the persistence parameter α that is modeled as local-to-unity with $c = 0$ corresponding to the unit root case and $c < 0$ to the stationary case. The initial observations are set at $u_0, u_{-1}, \cdots, u_{-k} = O_p(1)$.[2] This section treats the true lag order k as known. Lag order uncertainty is addressed in section 4. Our analysis is based on the following assumptions:

Assumption 1 *The sequence $\{e_t\}$ is a martingale difference sequence with $E(e_t|\mathcal{F}_{t-1}) = 0$ and $E(e_t^2|\mathcal{F}_{t-1}) = \sigma^2$, where $0 < \sigma^2 < \infty$, and $\mathcal{F}_t$ is the σ-field generated by $\{e_s;\ s \leq t\}$. Moreover, there exists small positive numbers ϕ_1 and ϕ_2 and a large positive number M_1 such that for $0 \leq s - s' \leq \phi_2$,*

$$\sup_{1 \leq m \leq t < \infty,\ \|\mathbf{v}_m\| = 1} \left| F_{t,m,\mathbf{v}_m}(s) - F_{t,m,\mathbf{v}_m}(s') \right| \leq M_1 (s - s')^{\phi_1},$$

where $\mathbf{v}_m = (v_1, ..., v_m)' \in \mathbb{R}^m$, $\|\mathbf{v}_m\| = \sum_{j=1}^m v_j^2$ and $F_{t,m,\mathbf{v}_m}(.)$ denotes the distribution of $\sum_{l=1}^m v_l e_{t+1-l}$.

Assumption 2 *All roots of $A(L) = 1 - \sum_{i=1}^k \alpha_i L^i$ lie outside the unit circle.*

The data generating process in (1) and Assumptions 1-2 are adopted from Hansen (2010a) with an additional restriction on the distribution of$\{e_t\}$ which ensures that the sample second moments of the regressors are bounded in expectation (see Ing et al., 2009). For $h \geq 1$, let the optimal (infeasible) mean squared error minimizing h-step ahead forecast of y_t be denoted μ_{t+h}. It is the conditional mean of y_{t+h} given $\mathcal{F}_t$, which is obtained from the following recursion (Hamilton, 1994,

[2] The conclusion for the subsequent analysis will not be affected as long as the initial observations are $o_p(T^{1/2})$.

p. 80-82):

$$\mu_{t+h} = z'_{t+h}\beta + \alpha(\mu_{t+h-1} - z'_{t+h-1}\beta) + \alpha_1(\Delta\mu_{t+h-1} - \Delta z'_{t+h-1}\beta) + \cdots + \alpha_k(\Delta\mu_{t+h-k} - \Delta z'_{t+h-k}\beta) \quad (2)$$

with $\mu_{t+j} = y_{t+j}$ if $j \leq 0$; $\beta = \beta_0$ and $z_t = 1$ if $p = 0$; $\beta = (\beta_0, \beta_1)'$ and $z_t = (1, t)'$ if $p = 1$. We can further rewrite (2) as

$$\mu_{t+h} = z'_{t+h}\beta^* + \alpha\mu_{t+h-1} + \sum_{j=1}^{k} \alpha_j \Delta\mu_{t+h-j} \quad (3)$$

where $\beta^* = (1-\alpha)\beta_0$ if $z_t = 1$ and $\beta^* = (\beta_0^*, \beta_1^*)'$ with $\beta_0^* = (1-\alpha)\beta_0 + (\alpha - \sum_{j=1}^{k}\alpha_j)\beta_1$, $\beta_1^* = (1-\alpha)\beta_1$ if $z_t = (1, t)$.

We consider three alternative estimators of μ_{t+h}. The first is the unrestricted estimator $\hat{\mu}_{t+h}$ obtained as

$$\hat{\mu}_{t+h} = z'_{t+h}\hat{\beta}^* + \hat{\alpha}\hat{\mu}_{t+h-1} + \sum_{j=1}^{k} \hat{\alpha}_j \Delta\hat{\mu}_{t+h-j} \quad (4)$$

with $\hat{\mu}_{t+j} = y_{t+j}$ if $j \leq 0$ where $(\hat{\beta}^*, \hat{\alpha}, \hat{\alpha}_j)$ are the OLS estimates from the regression

$$y_s = z'_s\beta^* + \alpha y_{s-1} + \sum_{j=1}^{k} \alpha_j \Delta y_{s-j} + e_s, \quad s = k+2, ..., T$$

Instead of using (4), one may consider a two-step strategy for estimating μ_{t+h} that entails regressing y_t on z_t and obtaining the estimate $\hat{\beta}$ of β and the residuals $\hat{u}_t = y_t - z'_t\hat{\beta}$ in a first step and then estimating an autoregression of order $k+1$ in $\hat{u}_t$ to obtain the estimates of $(\alpha, \alpha_1, ..., \alpha_k)$. The forecasts are obtained from (4). However, as shown in Ng and Vogelsang (2002), the one-step estimate $\hat{\mu}_{t+h}$ is preferable to the two-step estimate with persistent data.

The second estimator is the restricted estimator $\widetilde{\mu}_{t+h}$ that imposes the unit root restriction $\alpha = 1$ and is obtained as

$$\widetilde{\mu}_{t+h} = \Delta z'_{t+h}\widetilde{\beta}^* + \widetilde{\mu}_{t+h-1} + \sum_{j=1}^{k} \widetilde{\alpha}_j \Delta\widetilde{\mu}_{t+h-j}$$

with $\widetilde{\mu}_{t+j} = y_{t+j}$ if $j \leq 0$ where $(\widetilde{\beta}^*, \widetilde{\alpha}, \widetilde{\alpha}_j)$ are the OLS estimates from the regression

$$\Delta y_s = \Delta z'_s\beta^* + \sum_{j=1}^{k} \alpha_j \Delta y_{s-j} + e_s, \quad s = k+2, ..., T$$

Finally, the third estimator is based on taking a weighted average of the unrestricted and restricted forecasts. Letting $w \in [0,1]$ be the weight assigned to the unrestricted estimator, the averaging estimator is given by

$$\hat{\mu}_{t+h}(w) = w\hat{\mu}_{t+h} + (1-w)\widetilde{\mu}_{t+h}$$

The relative accuracy of the three foregoing estimators can be evaluated using the asymptotic forecast risk (AFR) which is the limit of the h-step ahead expected squared forecast error:

$$\begin{aligned}
f_0(c,p,k,h) &= \lim_{T\to\infty} \frac{T}{\sigma^2} E(\widetilde{\mu}_{T+h} - \mu_{T+h})^2 \\
f_1(c,p,k,h) &= \lim_{T\to\infty} \frac{T}{\sigma^2} E(\hat{\mu}_{T+h} - \mu_{T+h})^2 \\
f_w(c,p,k,h) &= \lim_{T\to\infty} \frac{T}{\sigma^2} E(\hat{\mu}_{T+h}(w) - \mu_{T+h})^2
\end{aligned}$$

In order to derive analytical expressions for the AFR, we introduce the following notation. Let $W(.)$ denote a standard Brownian motion on $[0,1]$ and define the diffusion process

$$dW_c(r) = cW_c(r) + dW(r)$$

For $p \in \{0,1\}$, let $X_c(r) = (r^p, W_c(r))'$ and define the stochastic processes

$$W_c^*(r,p) = \begin{cases} W_c(r) & \text{if} \quad p=0 \\ W_c(r) - \int_0^1 W_c(s)ds & \text{if} \quad p=1 \end{cases}$$

$$X_c^*(r,p) = \begin{cases} X_c(r) & \text{if} \quad p=0 \\ X_c(r) - \int_0^1 X_c(s)ds & \text{if} \quad p=1 \end{cases}$$

and the functionals

$$\begin{aligned}
T_{0c} &= -cW_c^*(1,p) + I(p=1)W(1) \\
T_{1c} &= X_c^*(1,p)' \left(\int_0^1 X_c^*(r,p) X_c^*(r,p)' \right)^{-1} \int_0^1 X_c^*(r,p) dW(r) + I(p=1)W(1)
\end{aligned}$$

Next, note that from (1), we can write

$$y_{t+h} = E_t(y_{t+h}) + \sum_{j=0}^{h-1} b_j e_{t+h-j}$$

where $E_t(.)$ denotes conditional expectation with respect to information at time t and the coefficients b_j $(j = 0, ..., h-1)$ are obtained by equating coefficients of L^j on both side of the equation

$$b(L)d(L) = 1$$

where $b(L) = \sum_{j=0}^{h-1} b_j L^j$ and $d(L) = 1 - \alpha L - (1-L)\sum_{j=1}^{k} \alpha_j L^j$. When $\alpha = 1$, $b_j = \sum_{i=0}^{j} \nu_j$, $\nu_0 = 1$ and ν_j; $j \geq 1$, satisfies $1 + \sum_{j=1}^{\infty} \nu_j L^j = 1/A(L)$ [see Ing et al., 2009].

Denoting $\alpha(k) = (\alpha_1, ..., \alpha_k)'$, we define the following quantities:

$$S_M(k) = \begin{pmatrix} \alpha(k-1) & I_{k-1} \\ \alpha_k & \mathbf{0}'_{k-1} \end{pmatrix}, \quad S_M^0(k) = I_k$$

$$M_h(k) = \sum_{j=0}^{h-1} b_j S_M^{h-1-j}(k), \quad \Gamma(k) = \lim_{j\to\infty} E(\mathbf{s}_j(k)\mathbf{s}_j(k)'), \quad \mathbf{s}_j(k) = (\Delta y_j, ..., \Delta y_{j-k+1})'$$

$$g_h(k) = \begin{cases} 0 & \text{if } k = 0 \\ tr\left(\Gamma(k)M_h(k)\Gamma^{-1}(k)M_h'(k)\right) & \text{if } k \geq 1 \end{cases}$$

With the above notation in place, we have the following result which provides an analytical representation for the AFR of the unrestricted and restricted forecasts:

Theorem 1 *Under Assumptions 1-2 and* $\sup_t E(|e_t|^{\theta_h}) < \infty$, *where* $\theta_h = \max\{8,\ 2(h+2)\}$ *for some* $\delta > 0$,

(a) $f_1(c, p, k, h) = f_1(c, p, h) + g_h(k)$, $f_1(c, p, h) = \left(\sum_{j=0}^{h-1} b_j\right)^2 E(T_{1c}^2)$.

(b) $f_0(c, p, k, h) = f_0(c, p, h) + g_h(k)$, $f_0(c, p, h) = \left(\sum_{j=0}^{h-1} b_j\right)^2 E(T_{0c}^2)$.

Theorem 1 shows that the AFR of both the restricted and unrestricted forecasts can be decomposed into two components: the first component $f_j(c, p, h)$, $j = 0, 1$, depends on both the underlying stochastic/deterministic trends as well as the short-run dynamics through the coefficients $\{b_j\}$; the second component $g_h(k)$ is common to the restricted and unrestricted estimators and depends on the parameters governing the short-run dynamics of the time series. The result generalizes Theorem 2 of Hansen (2010a) for one-step forecasts to multistep forecasts. Interestingly, when $h = 1$, the AFR can be expressed as the sum of a purely nonstationary component representing the stochastic/deterministic trends (since $b_0 = 1$) and a stationary short-run component which is simply the number of first-differenced lags, i.e., $g_1(k) = k$. However, as Theorem 1 shows, when $h > 1$, such a stationary-nonstationary decomposition no longer holds since both components now depend on the short-run coefficients $\{\alpha_j\}$. Theorem 1 also generalizes Theorem 2.2 of Ing et al. (2009) which derives an expression for AFR assuming an exact unit root ($c = 0$) and no

deterministic component.

The next result, which follows as a direct consequence of Theorem 1, shows that the optimal combination weight is independent of the forecast horizon and the moving average coefficients $\{b_j\}$ but depends on the nuisance parameter c:

Corollary 1 *The AFR of the combination forecast is given by*

$$f_w(c,p,k,h) = \left(\sum_{j=0}^{h-1} b_j\right)^2 \left\{w^2 E(T_{1c}^2) + (1-w)^2 E(T_{0c}^2) + 2w(1-w)E(T_{1c}T_{0c})\right\} + g_h(k)$$

with optimal (i.e., AFR minimizing) weight

$$w^* = \frac{E(T_{0c}^2) - E(T_{0c}T_{1c})}{E(T_{0c}^2) + E(T_{1c}^2) - 2E(T_{0c}T_{1c})}$$

3 Asymptotic Mean Squared Error and Asymptotic Forecast Risk

An alternative measure of forecast accuracy is the in-sample asymptotic mean squared error (AMSE) defined as

$$m_u(c,p,k,h) = \lim_{T\to\infty} \frac{1}{\sigma^2} \sum_{t=1}^{T-h} E(\hat{\mu}_{t+h} - \mu_{t+h})^2$$

for the unrestricted estimator with similar expressions in place for the restricted and averaging estimators. Hansen (2008) establishes the approximate equivalence between this measure and the AFR under the assumption of strict stationarity. Accordingly, existing forecast combination approaches developed in the stationary framework are based on targeting the AMSE by appealing to its equivalence with the AFR. Hansen (2008) proposes estimating the weights by minimizing a Mallows (2000) criterion which yields an asymptotically unbiased estimate of the AMSE. Similarly, Hansen (2010b) demonstrates that a leave-h-out cross validation criterion delivers an asymptotically unbiased estimate of the AMSE.

This equivalence result, however, breaks down in a nonstationary setup. For instance, when the process has a unit root with no drift and the regression does not include a deterministic component, it follows from the results in Hansen (2010a) that the AMSE of the one-step ahead forecast coincides with the expected value of the squared limiting Dickey-Fuller t-statistic. This expectation has been shown to be about 1.141 by Gonzalo and Pitarakis (1998) and Meng (2005) using analytical and numerical integration techniques, respectively. In contrast, Ing (2001) theoretically establishes that the AFR of the one-step ahead forecast for the same data generating process and regression is 2. More recently, Hansen (2010a) demonstrates the lack of equivalence within a local-to-unity framework showing that the AMSE of unrestricted as well as restricted (imposing a unit root) one-step ahead forecasts are different from the corresponding expressions for their AFR in autoregressive

models with a general lag order and a deterministically trending component. Notwithstanding this result, he suggests using a Mallows criterion to estimate the combination weights and evaluates the adequacy of the resulting combination forecast in finite samples via simulations. A similar approach is taken by Kejriwal and Yu (2021) who also employ Mallows weighting but estimate the deterministic component by FGLS in order to improve upon the accuracy of OLS-based forecasts.

To illustrate the failure of equivalence, Figure 1 plots the AMSE and the AFR of the unrestricted estimator for the case $p = 0$ and $k = 0$.[3] The figure clearly illustrates that while the two measures of forecast accuracy follow a similar path for c sufficiently far from zero, they tend to diverge as the process becomes more persistent. This pattern remains robust across different forecast horizons and suggests that a forecast combination approach that directly targets AFR instead of AMSE can potentially generate more accurate forecasts of highly persistent time series when forecast risk is used as a metric for forecast evaluation.

4 Choice of Combination Weights

The optimal combination forecast $\hat{\mu}_{t+h}(w^*)$ is infeasible in practice since the weight w^* depends on the unknown local-to-unity parameter c that is not consistently estimable. Given the lack of equivalence between AMSE and AFR for nonstationary time series as discussed in the previous section, we pursue an alternative approach to estimating the combination weights that directly targets the AFR, which is a more direct and practical measure of forecast accuracy than AMSE. In particular, the estimated weight $\hat{w}$ is obtained by minimizing the so-called Accumulated Prediction Errors (APE) criterion defined as

$$APE(w) = \sum_{i=m_h}^{T-h} \{y_{i+h} - \hat{\mu}_{i+h}(w)\}^2 = \sum_{i=m_h}^{T-h} \{w(y_{i+h} - \hat{\mu}_{i+h}) + (1-w)(y_{i+h} - \widetilde{\mu}_{i+h})\}^2 \qquad (5)$$

with respect to w, where $w \in [0,1]$, $\hat{\mu}_{i+h}(w)$ is the h-step ahead combination forecast based only on data up to period i, and m_h denotes the smallest positive number such that the forecasts $\hat{\mu}_{i+h}$ and $\widetilde{\mu}_{i+h}$ are well-defined for all $i \geq m_h$. The solution is given by

$$\hat{w} = \frac{\sum_{i=m_h}^{T-h}(y_{i+h} - \widetilde{\mu}_{i+h})^2 - \sum_{i=m_h}^{T-h}(y_{i+h} - \hat{\mu}_{i+h})(y_{i+h} - \widetilde{\mu}_{i+h})}{\sum_{i=m_h}^{T-h}(y_{i+h} - \widetilde{\mu}_{i+h})^2 + \sum_{i=m_h}^{T-h}(y_{i+h} - \hat{\mu}_{i+h})^2 - 2\sum_{i=m_h}^{T-h}(y_{i+h} - \hat{\mu}_{i+h})(y_{i+h} - \widetilde{\mu}_{i+h})}$$

The APE criterion with $h = 1$ was first introduced by Rissanen (1986) in the context of model selection. Wei (1987) derives the asymptotic properties of APE in general regression models and specializes his results to stationary and nonstationary autoregressive processes with $h = 1$. Ing (2004) demonstrates the strong consistency of the APE-based lag order estimator in stationary

[3] The figure was obtained by simulating the AMSE and AFR assuming i.i.d. normal errors with $T = 1000$. 5000 replications were used.

autoregressive models for $h \geq 1$. In particular, he shows that a normalized version of the APE converges almost surely to the AFR in large samples. Ing et al. (2009) extends the analysis to autoregressive processes with a unit root. The results in Wei (1987), Ing (2004) and Ing et al. (2009) all rely on the law of iterated logarithm which ensure that, in large samples, APE is equivalent to $\log T$ times the AFR, almost surely. It is, however, important to note that while this convergence result holds pointwise for $|\alpha| \leq 1$, they do not hold uniformly over α. In particular, it does not hold in the local-to-unity setup considered in this paper for $c < 0$.[4] Nevertheless, the following result shows that the APE criterion remains asymptotically valid in the current framework at the two limits of c which represent the unit root and fixed stationary cases:

Theorem 2 *For a given k, let $APE_0 = \sum_{i=m_h}^{T-h} \{y_{i+h} - \widetilde{\mu}_{i+h}\}^2$, $APE_1 = \sum_{i=m_h}^{T-h} \{y_{i+h} - \hat{\mu}_{i+h}\}^2$. Under Assumptions 1-2 and $sup_t\ E(|e_t|^r) < \infty$, for some $r > 2$,*
(a) For $c = O(T)$, $lim_{T\to\infty}\ (\sigma^2 \log T)^{-1} \left(APE_1 - \sum_{i=m_h}^{T-h} \eta_{i,h}^2\right) = \lim_{c\to-\infty} f_1(c, p, k, h)$.
(b) $lim_{c\to 0}\ \lim_{T\to\infty}\ (\sigma^2 \log T)^{-1} \left(APE_0 - \sum_{i=m_h}^{T-h} \eta_{i,h}^2\right) = \lim_{c\to 0} f_0(c, p, k, h)$.

Remark 1 *In a similar vein, Hansen (2010a) develops feasible combination weights by evaluating the Mallows criterion at the two limits of c, given that the criterion depends on c and is therefore infeasible in general. Thus, while his analysis demonstrates that the infeasible Mallows criterion is an asymptotically unbiased estimate of the AMSE for any c, the feasible version of the criterion remains valid only in the two limit cases. When estimation is performed using FGLS instead of OLS, Kejriwal and Yu (2021) show that the infeasible Mallows criterion also depends on the parameter a in (1) which governs the short-run dynamics. Evaluating the criterion at the two limits, however, eliminates the dependence on both nuisance parameters.*

Figure 2 plots the AFR of the optimal (infeasible) and APE-based combination forecasts for $p = 1$ and $k = 0$.[5] For comparison, the unrestricted and restricted forecasts are also presented. As expected, the forecast risk of the restricted estimator increases with $|c|$ while the risk function of the unrestricted estimator is relatively flat as a function of c. Regardless of the forecast horizon, the feasible combination forecast maintains a risk profile close to that of the optimal forecast. In particular, the risk of the APE-weighted forecast is uniformly lower than that of the unrestricted estimator across values of c as well as lower than that of the restricted estimator unless c is very close to zero. These results suggest that the loss in forecast accuracy due to the unknown degree of persistence is relatively small when constructing the combination weights based on the APE criterion. In sections 5 and 6, we will conduct an extensive comparison of the APE-based combination forecasts with both the Mallows and cross-validation based combination forecasts.

[4] To illustrate the lack of uniformity, consider the case $p = 1$ with $k = 0$. Using the same arguments as in the proof of Theorem 2 of Yu et al. (2012), it follows that, for any finite $c \leq 0$, $\sum_{i=m_h}^{T-h} \{y_{i+h} - \hat{\mu}_{i+h}\}^2 = E(T_{10}^2) \log T + o_p(\log T)$, where $E(T_{10}^2) = 6$. The lack of uniformity follows since $E(T_{1c}^2) \neq E(T_{10}^2)$ for any $c < 0$.

[5] This figure was obtained using the same method as Figure 1.

5 Lag Order Uncertainty

This section extends the preceding analysis to the case where the lag order k is unknown. In order to accommodate lag order uncertainty, the set of models on which the combination forecast is based needs to be expanded to include models with different lag orders. Such a forecast can potentially trade off the misspecification bias inherent from the omission of relevant lags against the problem of overfitting induced by the inclusion of unnecessary lags. We include sub-models with $l \in \{0, 1, ..., K\}$, $K \geq k$, with the corresponding restricted and unrestricted forecasts given by $\widetilde{\mu}_t(l)$ and $\hat{\mu}_t(l)$, respectively. We consider two types of combination forecasts. The first is a "partial averaging" forecast that only addresses lag order uncertainty by averaging over the $K+1$ unrestricted forecasts:

$$\hat{\mu}_{t+h}(\hat{W}) = \sum_{l=0}^{K} \hat{w}_l \hat{\mu}_t(l) \tag{6}$$

The weights $\hat{W} = (\hat{w}_0, \hat{w}_1, ..., \hat{w}_K)'$ are obtained by minimizing the APE criterion

$$APE_P(W) = \sum_{i=m_h}^{T-h} \left\{ \sum_{l=0}^{K} [w_l(y_{i+h} - \hat{\mu}_{i+h}(l))] \right\}^2 \tag{7}$$

where $w_l \geq 0$ $(l = 0, ..., K)$, $\sum_{l=0}^{K} w_l = 1$. We refer to (6) as the APE-based Partial Averaging (APA) forecast.

The second forecast is a "general averaging" forecast that accounts for both persistence and lag order uncertainty and thus combines the forecasts from all $2(K+1)$ sub-models:

$$\breve{\mu}_{t+h}(\breve{W}) = \sum_{l=0}^{K} (\breve{w}_{1l} \hat{\mu}_t(l) + \breve{w}_{0l} \widetilde{\mu}_t(l)) \tag{8}$$

The weights $\breve{W} = (\breve{w}_{01}, \breve{w}_{02}, ..., \breve{w}_{0K}, \breve{w}_{11}, \breve{w}_{12}, ..., \breve{w}_{1K})'$ are obtained by minimizing a generalized APE criterion of the form

$$APE_G(W) = \sum_{i=m_h}^{T-h} \left\{ \sum_{l=0}^{K} [w_{1l}(y_{i+h} - \hat{\mu}_{i+h}(l)) + w_{0l}(y_{i+h} - \widetilde{\mu}_{i+h}(l))] \right\}^2 \tag{9}$$

where $w_{1l} \geq 0, w_{0l} \geq 0$ $(l = 0, ..., K)$, $\sum_{l=0}^{K}(w_{0l} + w_{1l}) = 1$. We refer to (8) as the APE-based General Averaging (AGA) forecast. Comparing the APA and AGA forecasts will serve to isolate the effects of the two sources of uncertainty on forecast accuracy.

6 Monte Carlo Simulations

This section reports the results of a set of Monte Carlo experiments designed to (1) evaluate the finite sample performance of the proposed approach relative to extant approaches; (2) quantify the importance of accounting for each source of uncertainty in terms of its effect on finite sample forecast risk. Section 6.1 lays out the experimental design. Section 6.2 details the different forecasting procedures included in the analysis. Sections 6.3 and 6.4 present the results. Results are obtained for $p \in \{0, 1\}$. For brevity, we report the results only for $p = 1$. The results for $p = 0$ are qualitatively similar, although the improvements offered by the proposed approach are more pronounced for $p = 1$ than $p = 0$. The full set of results is available upon request.

6.1 Experimental Design

We adopt a design similar to that in Hansen (2010a) and Kejriwal and Yu (2021) to facilitate direct comparisons. The data generating process (DGP) is based on (1) and specified as follows: (a) the innovations $e_t \overset{i.i.d.}{\sim} N(0, 1)$; (b) the trend parameters are set at $\beta_0 = \beta_1 = 0$; (c) the true lag order $k \in \{0, 6, 12\}$ with $\alpha_j = -(-\theta)^j$ for $j = 1, ..., k$ and $\theta = 0.6$. The maximum number of first-differenced lags included is set at $K = 12$. The sample size is set at $T \in \{100, 200\}$. The local-to-unity parameter c varies from -20 to 0, implying α ranging from 0.8 to 1 for $T = 100$ and α ranging from 0.9 to 1 for $T = 200$. At each c value, the finite-sample forecast risk $TE\left[(\hat{\mu}_{T+h} - \mu_{T+h})^2\right]$ is computed for all estimators considered, where $h \in \{1, 3, 6, 12\}$. All experiments are based on 10,000 Monte Carlo replications.

We report two sets of results. The first assumes k is known thereby allowing us to demonstrate the effect of persistence uncertainty on forecast accuracy while abstracting from lag order uncertainty. The second allows k to be unknown and facilitates the comparison between forecasts that address both forms of uncertainty with those that only account for lag order uncertainty.

6.2 Forecasting Methods

The benchmark forecast in both the known and unknown lag cases is calculated from a standard autoregressive model of order $K + 1$ estimated by OLS:

$$y_t = \beta_0^* + \beta_1^* t + \alpha y_{t-1} + \sum_{j=1}^{K} \alpha_j \Delta y_{t-j} + \epsilon_t \tag{10}$$

When the number of lags is assumed known (section 6.3) we compare a set of six forecasting methods: (1) Mallows Selection (Mal-Sel); (2) Cross-Validation Selection (CVh-Sel); (3) APE Selection (APE-Sel); (4) Mallows Averaging (Mal-Ave); (5) Cross-Validation Averaging (CVh-Ave); (6) APE Averaging (APE-Ave). With an unknown number of lags, the following six methods are

compared[6]: (1) Mallows Partial Averaging (MGA); (2) Cross-Validation Partial Averaging (CPA); (3) APE Partial Averaging (APA); (4) Mallows General Averaging (MGA); (5) Cross-Validation General Averaging (CGA); (6) APE General Averaging (AGA). For brevity, a detailed description of these methods is not presented here but included in Supplementary Appendix B.

Both the APE selection and combination forecasts require a choice of m_h. To our knowledge, no data-dependent methods for choosing m_h are available in the existing literature. We therefore examined the viability of alternative choices via simulations. Specifically, for each persistence level (value of c), we computed the minimum forecast risk over all values of m_h in the range $[15, 70]$ with a step-size of 5 (assuming a known number of lags k). While no single value was found to uniformly dominant across persistence levels/horizons, $m_h = 20$ turned out to be a reasonable choice overall.[7] To justify this choice, Figure C.1 in Supplementary Appendix C plots the difference between the optimal forecast risk and the risk of the APE selection forecasts for $m_h = 20$ expressed as a percentage of the forecast risk for $m_h = 20$. The corresponding results for the APE combination forecasts are presented in Figure C.2. It is evident that using $m_h = 20$ entails only a marginal increase in forecast risk (at most 5%) for the combination forecasts over the optimal forecast risk across different persistence levels and horizons. In contrast, the optimal choice of m_h for the selection forecasts is somewhat more unstable and appears to depend more heavily on the forecast horizon and the level of persistence. This robustness in behavior provides additional motivation for employing a combination approach to forecasting in practice.

6.3 Forecast Risk with Known Lag Order

Figures 3-5 plot the risk of the six methods relative to the benchmark. Consider first the case $k = 0$. Several features of the results are noteworthy. First, the selection forecasts typically exhibit higher risk than the corresponding combination forecasts across sample sizes and horizons. Second, when $T = 100$, the APE combination forecast is clearly the dominant method, performing discernibly better than forecasts based on either of the two competing weighting schemes. When $T = 200$, its dominance continues except when $|c|$ is sufficiently large (the exact magnitude being horizon-dependent) in which case the benchmark delivers the most accurate forecasts and averaging over the restricted model becomes less attractive. Third, the relative performance of the Mallows and cross-validation weighting schemes depends on the horizon: at $h = 1$, the two schemes yield virtually indistinguishable forecasts; when $h \in \{3, 6\}$, Mallows weighting yields uniformly lower risk over the parameter space; at $h = 12$, Mallows weighting is preferred when persistence is high (c close to zero) while cross-validation weighting dominates for lower levels of persistence.

In the presence of higher order serial correlation ($k > 0$), the superior performance of the

[6] We do not report the results for the selection forecasts since their performance relative to the combination forecasts is qualitatively similar to the known lag order case. The results are nevertheless available upon request.

[7] This choice was also adopted by Ing and Yang (2014) in their Monte Carlo analysis of forecasting using autoregressive models with positive-valued errors.

APE combination forecast becomes even more evident: it now dominates all competing forecasts regardless of horizon and sample size. In particular, APE weighting outperforms the benchmark at all persistence levels even at $T = 200$, unlike the $k = 0$ case. The intuition for this difference in relative performance between the cases with and without higher order serial correlation is that in the former case, averaging is comparatively more beneficial since imposing the unit root restriction can potentially reduce the estimation uncertainty associated with the coefficients of the lagged differences. This reduction in sampling uncertainty in turn engenders a reduction in the overall risk of the combination forecast relative to the unrestricted benchmark forecast. Another notable difference from the $k = 0$ case is that while Mallows and cross-validation weighting are comparable for $h \in \{1, 3\}$, the former now dominates for $h \in \{6, 12\}$ uniformly over the parameter space.

6.4 Forecast Risk with Unknown Lag Order

Figures 6-8 plot the relative risk of the six combination forecasts which comprise the three partial forecasts that only account for lag order uncertainty and the three general forecasts that account for both lag order and stochastic trend uncertainty. A clear implication of these results is that general averaging methods typically exhibit considerably lower forecast risk than partial averaging methods unless the process has relatively low persistence in which case averaging over the unit root model increases the forecast risk incurred by the general averaging methods. The improvements offered by general averaging hold both across horizons and the number of lags (k) in the true DGP and become more prominent as the sample size increases.

Among the three weighting schemes, APE-based weights are the preferred choice except when $h \in \{6, 12\}$ and $T = 100$ where Mallows weighting turns out to be the dominant approach if persistence is relatively low. A potential explanation for this result is that with long horizons and a small sample size, the APE criterion is based on a relatively smaller number of prediction errors which increases the sampling variability associated with the resulting weights thereby increasing the risk of the combination forecast. As in the known lag order case, the choice between Mallows and cross-validation weighting is horizon-dependent: when $h = 1$, cross-validation weighting is preferred while when $h > 1$, Mallows weighting is preferred with the magnitude of reduction in forecast risk increasing as h increases.

In summary, the results from the simulation experiments make a strong case for employing APE weights when constructing the combination forecasts and clearly highlight the benefits of targeting forecast risk rather than in-sample mean squared error. The comparison of general and partial combination forecasts also underscore the importance of concomitantly controlling for both stochastic trend uncertainty and lag order uncertainty in generating accurate forecasts.

7 Empirical Application

This section conducts a pseudo out-of-sample forecast comparison of the different multistep forecast combination methods using a set of US macroeconomic time series. Our objectives are to empirically assess (1) the efficacy of different averaging/selection methods relative to a standard autoregressive benchmark; (2) the importance of averaging over both the persistence level and the lag order; and (3) the relative performance of alternative weight choices for constructing the combination forecasts.

Our analysis employs the FRED-MD data set compiled by McCracken and Ng (2016) containing 123 monthly macroeconomic variables over the period January 1960 - December 2018.[8] McCracken and Ng (2016) suggest a set of seven transformation codes designed to render each series stationary: (1) no transformation; (2) Δy_t; (3) $\Delta^2 y_t$; (4) $\log(y_t)$; (5) $\Delta\log(y_t)$; (6) $\Delta^2 \log(y_t)$; (7) $\Delta(y_t/y_{t-1}-1)$. To ensure that the series fit our framework that allows for highly persistent time series with/without deterministic trends, we adopt the following transformation codes as modified by Kejriwal and Yu (2021): (1') no transformation; (2') y_t; (3') Δy_t, (4') $\log(y_t)$; (5') $\log(y_t)$; (6') $\Delta \log(y_t)$; (7') $y_t/y_{t-1}-1$. For series that correspond to codes (1') and (4'), we construct the forecasts from a model with no deterministic trend ($p = 0$), while for the remaining codes, we use forecasts from a model that include a linear deterministic trend ($p = 1$). We also report results for eight core series as in Stock & Watson (2002b), comprising four real and four nominal variables.

As in the simulation experiments, four alternative forecast horizons are considered: $h \in \{1, 3, 6, 12\}$. We use a rolling window scheme with an initial estimation period between 1960:01-1969:12 so that the forecast evaluation period is 1970:01-2018:12 (588 observations). The size of the estimation window changes depending on the forecast horizon h. For example, when $h = 1$, the initial training sample contains 120 observations from 1960:01-1969:12 while for $h = 3$, it contains only 118 observations from 1960:01-1969:10. This ensures that the forecast origin is 1970:01 for all forecast horizons considered. We compare ten different methods in terms of the mean squared forecast error (MSFE) computed as the average of the squared forecast errors: (1) MPA: Partial Mallows averaging over the number of lags only in the unrestricted model; (2) MGA: General Mallows averaging over both the unit root restriction and the number of lags; (3) CPA: Leave-h-out cross-validation (CV-h) averaging over the number of lags only in the unrestricted model; (4) CGA: Leave-h-out cross-validation averaging over both the unit root restriction and the number of lags; (5) APA: Accumulated Prediction Error averaging over the number of lags only in the unrestricted model; (6) AGA: Accumulated Prediction Error averaging over both the unit root restriction and the number of lags; (7) MS: Mallows selection from all models (unrestricted and restricted) that vary with the number of lags; (8) CVhS: Leave-h-out cross-validation selection from all models (unrestricted and restricted) that vary with the number of lags; (9) APES: Accumulated Prediction Error Selection from all models (unrestricted and restricted) that vary with the number of lags; (10) AR: Unre-

[8] The data set is publicly available for download at https://research.stlouisfed.org/econ/mccracken/fred-databases/

stricted autoregressive model (benchmark). The maximum number of allowable first differenced lags in each method is set at $K = 12$. The benchmark forecast is computed from unrestricted OLS estimation of an autoregressive model of the form (10) that uses 12 first-differenced lags of the dependent variable and includes/excludes a deterministic trend depending on the transformation code the series corresponds to as discussed above.

Table 1a ($h = 1, 3$) and Table 1b ($h = 6, 12$) report the percentage wins and losses based on the MSFE for the 123 series. Specifically, it shows the percentage of 123 series for which a method listed in a row outperforms a method listed in a column, and all other methods (last column). A summary of the results in Tables 1a and 1b is given below:

1. The averaging methods uniformly dominate their selection counterparts at all forecast horizons. For instance, Mallows/cross-validation averaging outperform the corresponding selection procedures in more than 90% of the series at each horizon. The performance of AGA relative to APES is relatively more dependent on the horizon, with improvements observed in 77% (65%) of the series for $h = 1$ ($h = 12$), respectively.

2. Given a particular weighting scheme, averaging over both the unit root restriction and number of lags (general averaging) outperforms averaging over only the number of lags (partial averaging) at all horizons. For instance, when $h = 1$, MGA (CGA, AGA) dominate MPA (CPA, APA) in 95% (81%, 79%) of the series, respectively, based on pairwise comparisons. A similar pattern is observed for multi-step forecasts.

3. Across all horizons, AGA emerges as the leading procedure due to its ability to deliver forecasts with the lowest MSFE among all methods for the maximum number of series (last column of Tables 1a and 1b). This approach also dominates each of the competing approaches in terms of pairwise comparisons. The APES approach ranks second among all methods so that forecasting based on the accumulated prediction errors criterion (either AGA or APES) outperforms the other approaches for more than 50% of the series over each horizon (the specific percentages are 68.3% for $h = 1, 3$; 57.7% for $h = 6$; 55.3% for $h = 12$).

Next, we examine the performance of the forecasting methods for different types of series based on their groupwise classification by McCracken and Ng (2016) in an attempt to uncover the extent to which the best methods vary by the type of series analyzed. In particular, McCracken and Ng (2016) classify the series into eight distinct groups: (1) output and income; (2) labour market; (3) housing; (4) consumption, orders and inventories; (5) money and credits; (6) interest and exchange rates; (7) prices; (8) stock market. For each of these groups, Table 2 reports the method(s) with the lowest MSFE for the most number of series compared to all other competing methods. We also report the number of horizons in which (a) averaging outperforms selection and vice-versa; (b) averaging over both the unit root restriction and number of lags (general averaging - GA) methods

is superior to averaging over only the number of lags (partial averaging - PA) and vice-versa; (c) each of the three weighting schemes dominates the other two. The results are consistent with those in Tables 1a-b and clearly demonstrate (1) the dominance of averaging over selection (with the exception of Group 3) ; (2) the benefits of accounting for both stochastic trend uncertainty and lag order uncertainty (GA) relative to only the latter (PA) for five out of the eight groups; (3) the superiority of APE weighting over the two competing weighting schemes (the exception is Group 5 where cross-validation weighting is the dominant approach).

Finally, we present a comparison of the different methods with respect to their ability in forecasting the eight core series analyzed in Stock and Watson (2002b). Table 3 reports the MSFE of the eight methods relative to the benchmark model (10) for four real variables (industrial production, real personal income less transfers, real manufacturing and trade sales, number of employees on nonagricultural payrolls) while Table 4 reports the corresponding results for four nominal variables (the consumer price index, the personal consumption expenditure implicit price deflator, the consumer price index less food and energy, the producer price index for finished goods). To assess whether the difference between the proposed methods and the benchmark model is statistically significant, we use a two-tailed Diebold-Mariano test statistic (Diebold and Mariano, 2002). A number less than one indicates better forecast performance than the benchmark and vice versa. The method with smallest relative MSFE for a given series is highlighted in bold.

Consider first the results for real variables (Table 3). The performance of the best method is statistically significant (at the 10% level) relative to the benchmark in twelve out of the sixteen cases. Consistent with the results in Tables 1a-b and 2, general averaging typically dominates partial averaging, the exceptions being nonagricultural employment for $h \leq 6$, industrial production at $h = 12$, and real manufacturing and trade sales for $h = 6, 12$, where APES is the dominant procedure. The AGA approach turns out to have the highest relative forecast accuracy in 50% of all cases with the improvements offered over rival approaches particularly notable at $h = 12$. While cross-validation weighting does not yield the best forecasting procedure in any of the cases, Mallows weighting is the preferred approach in only two cases although the improvements are statistically insignificant. Turning to the results for nominal variables (Table 4), the best method significantly outperforms the benchmark in ten cases. Again, general averaging is usually preferred to partial averaging, the exception being the case $h = 12$ where APA outperforms all other methods for three of the four variables. As with the real variables, the AGA forecast is the most accurate in 50% of all cases though the improvements are now comparable across horizons. Finally, cross-validation weighting redeems itself to some extent by providing the best forecast in four cases while Mallows weighting is the preferred method is only one case.

In summary, the empirical results are consistent with the simulation results in finding that (1) addressing both persistence uncertainty and lag-order uncertainty are crucial for generating accurate forecasts; (2) a weighting scheme that directly targets forecast risk instead of in-sample

mean squared error yields an efficacious forecast combination approach at all horizons.

8 Conclusion

This paper develops new multistep forecast combination methods for a time series driven by stochastic and/or deterministic trends. In contrast to existing methods based on Mallows/cross-validation weighting, our proposed combination forecasts are based on constructing weights obtained from an accumulated prediction errors criterion that directly targets the asymptotic forecast risk instead of the in-sample AMSE. Our analysis finds strong evidence in favor of a version of the proposed approach that simultaneously addresses stochastic trend and lag order uncertainty. Our preferred approach can potentially serve as a useful univariate benchmark when evaluating the effectiveness of methods designed to exploit information in large data sets (e.g., Stock and Watson, 2002a).

We conclude with a discussion of three possible directions for future research. First, the APE-based combination forecasts can potentially be used in conjunction with FGLS estimation of the deterministic component, given that the latter has been shown to yield improved forecasts over OLS estimation (Kejriwal and Yu, 2021). Second, it may be useful to explore the possibility of allowing for a nonlinear deterministic component through, say, the inclusion of polynomial trends or a few low frequency trigonometric components (Gallant, 1981). To the extent that the specific nonlinear modeling structure captures the observed nonlinearities, such an approach may contribute to a further improvement in forecasting performance. Third, and perhaps most challenging, while our numerical and empirical analyses clearly document the desirability of the proposed approach based on APE weighting relative to Mallows/CV weighting, an analytical comparison may shed further light on the relative merits of the different methods. To our knowledge, such results are primarily available in the context of the standard stationary framework with Mallows/cross-validation weighting (e.g., Hansen, 2007; Zhang et al., 2013; Liao and Tsay, 2020). Extending these results to the present nonstationary framework would be a potentially fruitful endeavor.

References

Box, G. E. and Jenkins, G. M. (1970). Time series analysis: Forecasting and control. *San Francisco, Holdan-Day*.

Clements, M. P. and Hendry, D. F. (2001). Forecasting with difference-stationary and trend-stationary models. *The Econometrics Journal*, 4(1):1–19.

Diebold, F. X. and Kilian, L. (2000). Unit-root tests are useful for selecting forecasting models. *Journal of Business & Economic Statistics*, 18(3):265–273.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144.

Elliott, G. (2006a). Forecasting with trending data. *Handbook of economic forecasting*, 1:555–604.

Elliott, G. (2006b). Unit root pre-testing and forecasting. Technical report, Working Paper, UCSD.

Elliott, G. and Timmermann, A. (2016). Forecasting in economics and finance. *Annual Review of Economics*, 8:81–110.

Gallant, A. R. (1981). On the bias in flexible functional forms and an essentially unbiased form: the fourier flexible form. *Journal of Econometrics*, 15(2):211–245.

Gonzalo, J. and Pitarakis, J.-Y. (1998). On the exact moments of asymptotic distributions in an unstable ar (1) with dependent errors. *International Economic Review*, pages 71–88.

Granger, C. W. (1966). The typical spectral shape of an economic variable. *Econometrica: Journal of the Econometric Society*, pages 150–161.

Hamilton, J. D. (1994). *Time series analysis*. Princeton university press.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, 75(4):1175–1189.

Hansen, B. E. (2008). Least-squares forecast averaging. *Journal of Econometrics*, 146(2):342–350.

Hansen, B. E. (2010a). Averaging estimators for autoregressions with a near unit root. *Journal of Econometrics*, 158(1):142–155.

Hansen, B. E. (2010b). Multi-step forecast model selection. In *20th Annual Meetings of the Midwest Econometrics Group*.

Ing, C. (2001). A note on mean-squared prediction errors of the least squares predictors in random walk models. *Journal of Time Series Analysis*, 22(6):711–724.

Ing, C.-K. (2004). Selecting optimal multistep predictors for autoregressive processes of unknown order. *The Annals of Statistics*, 32(2):693–722.

Ing, C.-K., Lin, J.-L., and Yu, S.-H. (2009). Toward optimal multistep forecasts in non-stationary autoregressions. *Bernoulli*, 15(2):402–437.

Ing, C.-K., Sin, C.-y., and Yu, S.-H. (2012). Model selection for integrated autoregressive processes of infinite order. *Journal of Multivariate Analysis*, 106:57–71.

Ing, C.-K. and Yang, C.-Y. (2014). Predictor selection for positive autoregressive processes. *Journal of the American Statistical Association*, 109(505):243–253.

Kejriwal, M. and Yu, X. (2021). Generalized forecast averaging in autoregressions with a near unit root. *The Econometrics Journal*, 24(1):83–102.

Liao, J.-C. and Tsay, W.-J. (2020). Optimal multistep var forecast averaging. *Econometric Theory*, 36(6):1099–1126.

Mallows, C. L. (2000). Some comments on cp. *Technometrics*, 42(1):87–94.

McCracken, M. W. and Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.

Meng, X.-L. (2005). From unit root to stein's estimator to fisher's k statistics: If you have a moment, i can tell you more. *Statistical Science*, 20(2):141–162.

Ng, S. and Vogelsang, T. (2002). Forecasting autoregressive time series in the presence of deterministic components. *The Econometrics Journal*, 5(1):196–224.

Rissanen, J. (1986). Order estimation by accumulated prediction errors. *Journal of Applied Probability*, 23(A):55–61.

Sampson, M. (1991). The effect of parameter uncertainty on forecast variances and confidence intervals for unit root and trend stationary time-series models. *Journal of Applied Econometrics*, 6(1):67–76.

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460):1167–1179.

Stock, J. H. and Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

Stock, J. H. and Watson, M. W. (2005). An empirical comparison of methods for forecasting using many predictors. *Manuscript, Princeton University*, 46.

Stock, J. H. and Watson, M. W. (2006). Forecasting with many predictors. *Handbook of economic forecasting*, 1:515–554.

Tu, Y. and Yi, Y. (2017). Forecasting cointegrated nonstationary time series with time-varying variance. *Journal of Econometrics*, 196(1):83–98.

Turner, J. L. (2004). Local to unity, long-horizon forecasting thresholds for model selection in the ar (1). *Journal of Forecasting*, 23(7):513–539.

Wei, C. (1987). Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. *The Annals of Statistics*, pages 1667–1682.

Yu, S.-H., Lin, C.-C., and Cheng, H.-W. (2012). A note on mean squared prediction error under the unit root model with deterministic trend. *Journal of Time Series Analysis*, 33(2):276–286.

Zhang, X., Wan, A. T., and Zou, G. (2013). Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics*, 174(2):82–94.

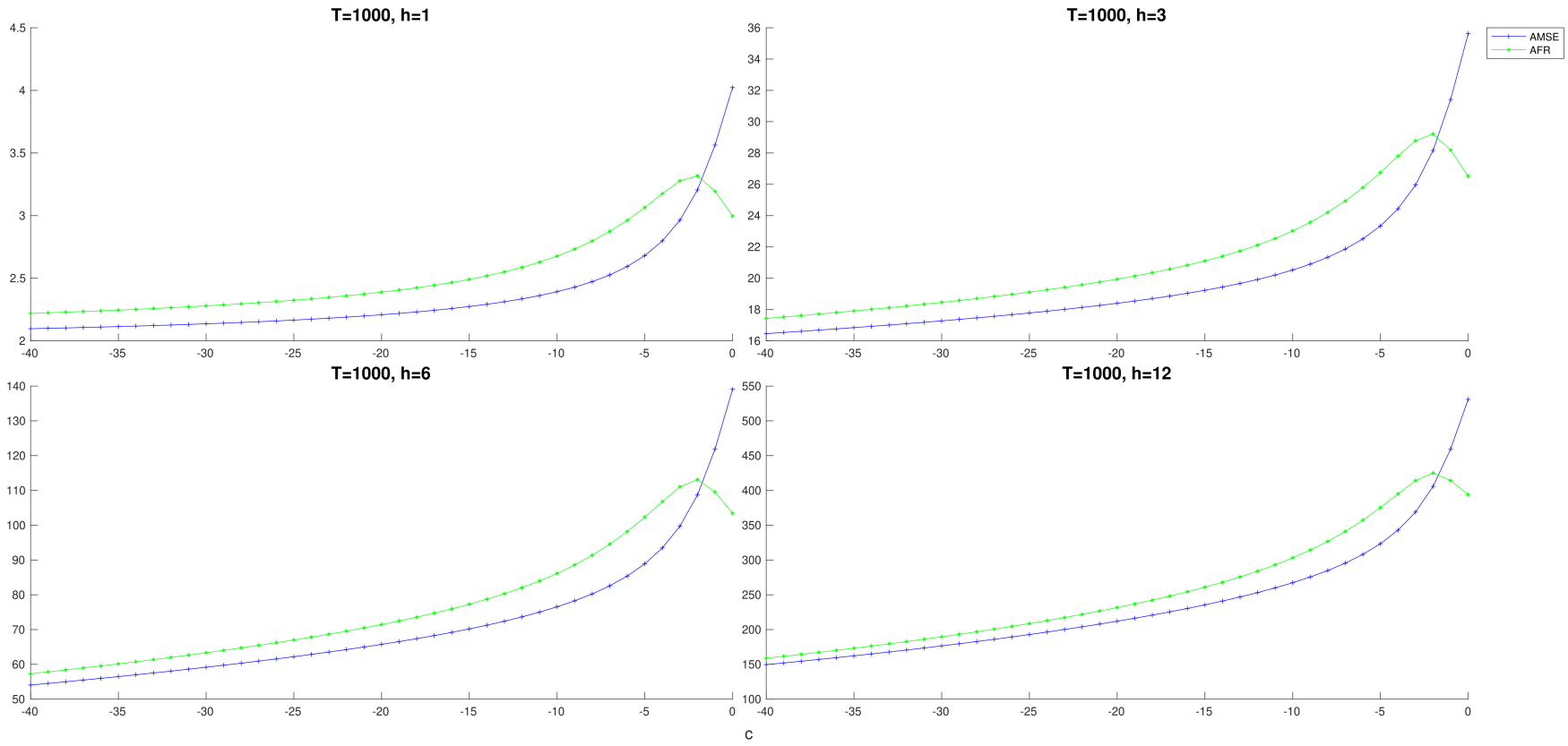

Figure 1: In-sample AMSE versus asymptotic forecast risk ($p = 0$, $k = 0$)

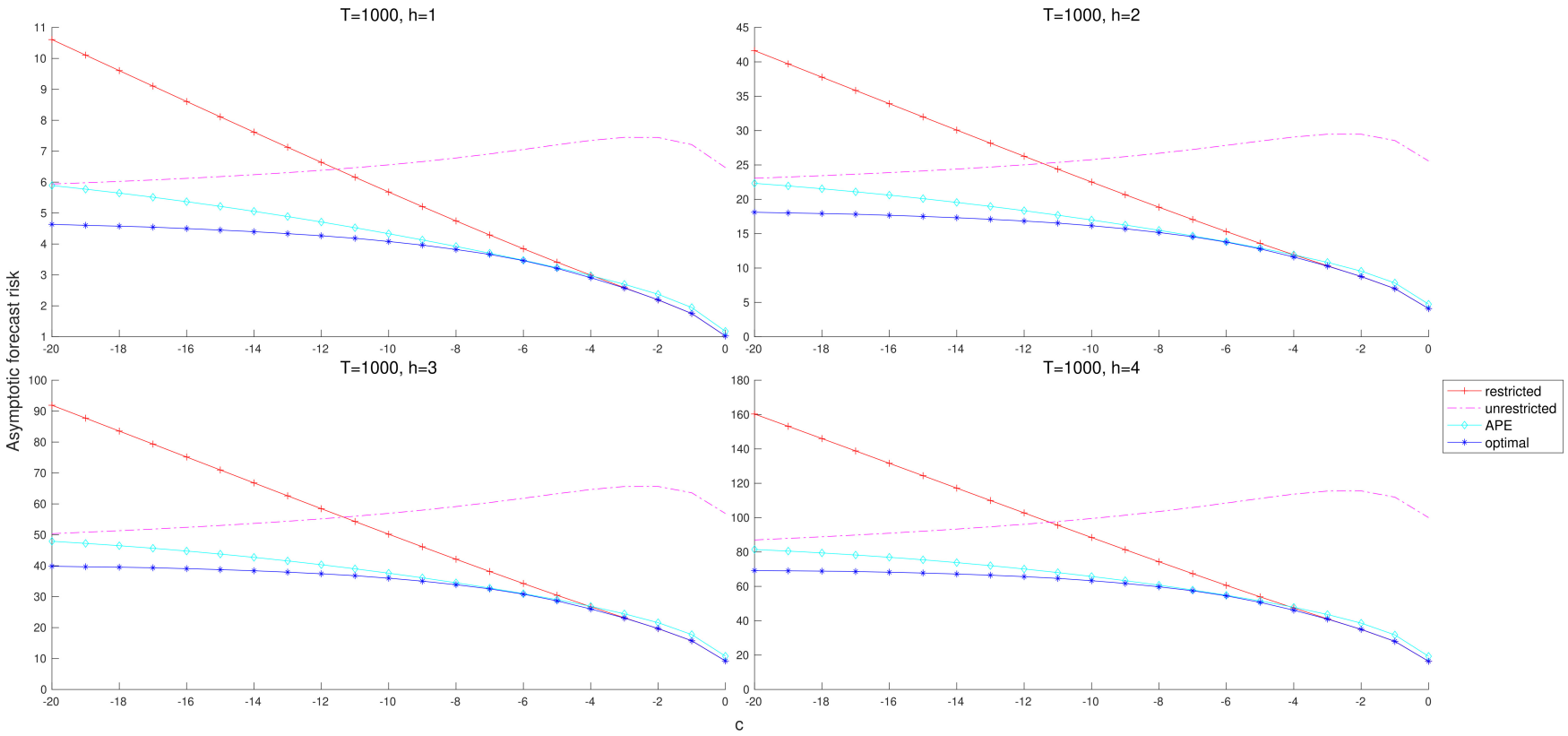

Figure 2: Asymptotic forecast risk of infeasible (optimal) and feasible (APE-based) combination forecasts ($p = 1, k = 0$)

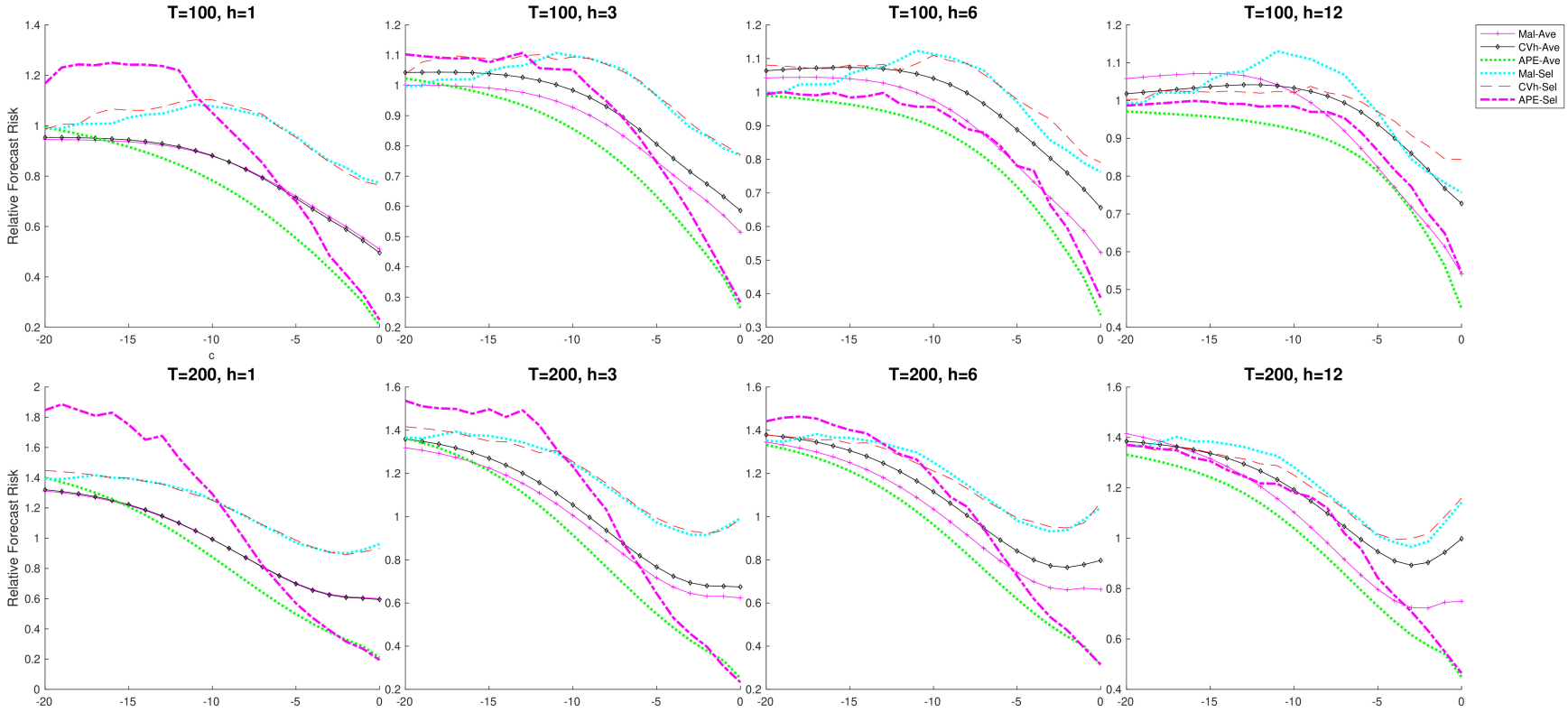

Figure 3: Forecast risk with known lag order ($k = 0$)

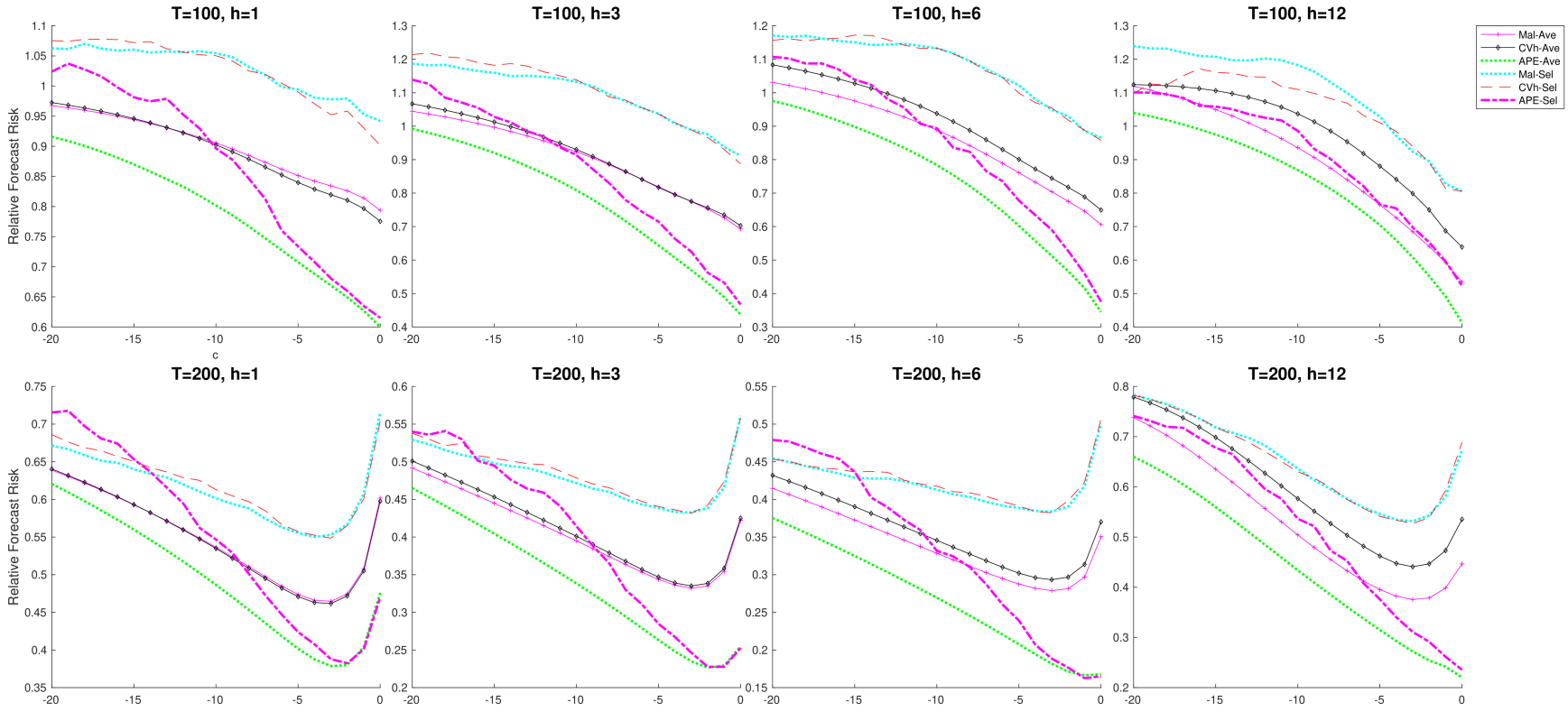

Figure 4: Forecast risk with known lag order ($k = 6$)

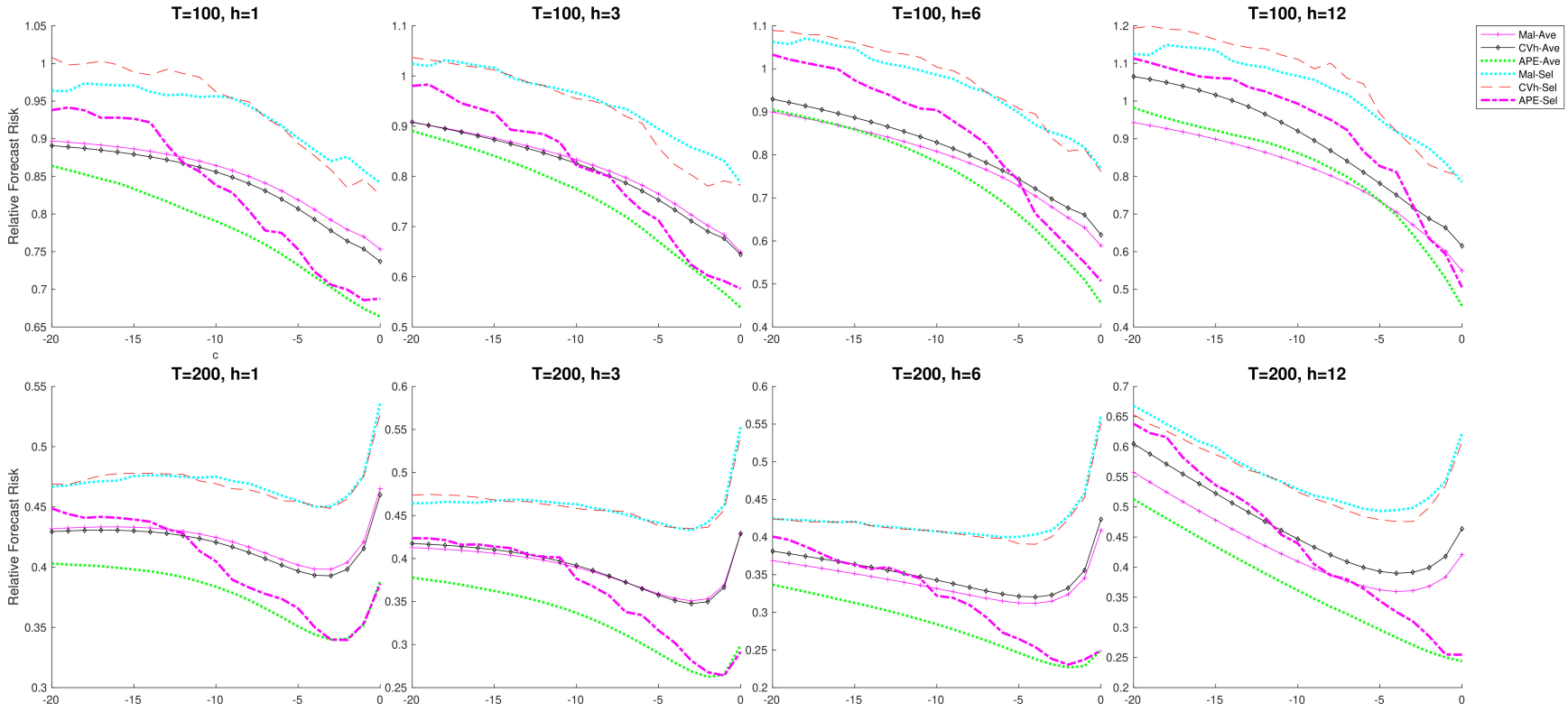

Figure 5: Forecast risk with known lag order ($k = 12$)

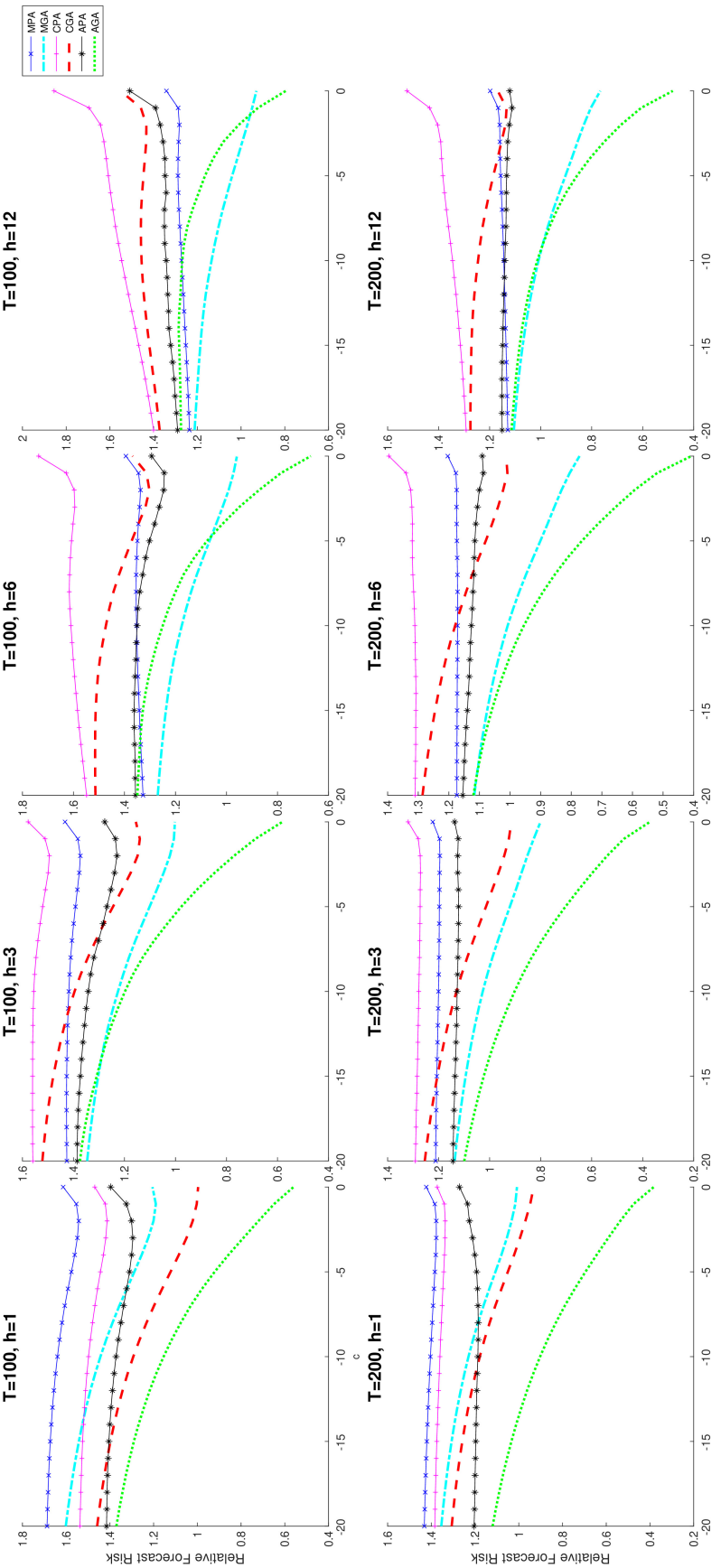

Figure 6: Forecast risk with unknown lag order ($k = 0$)

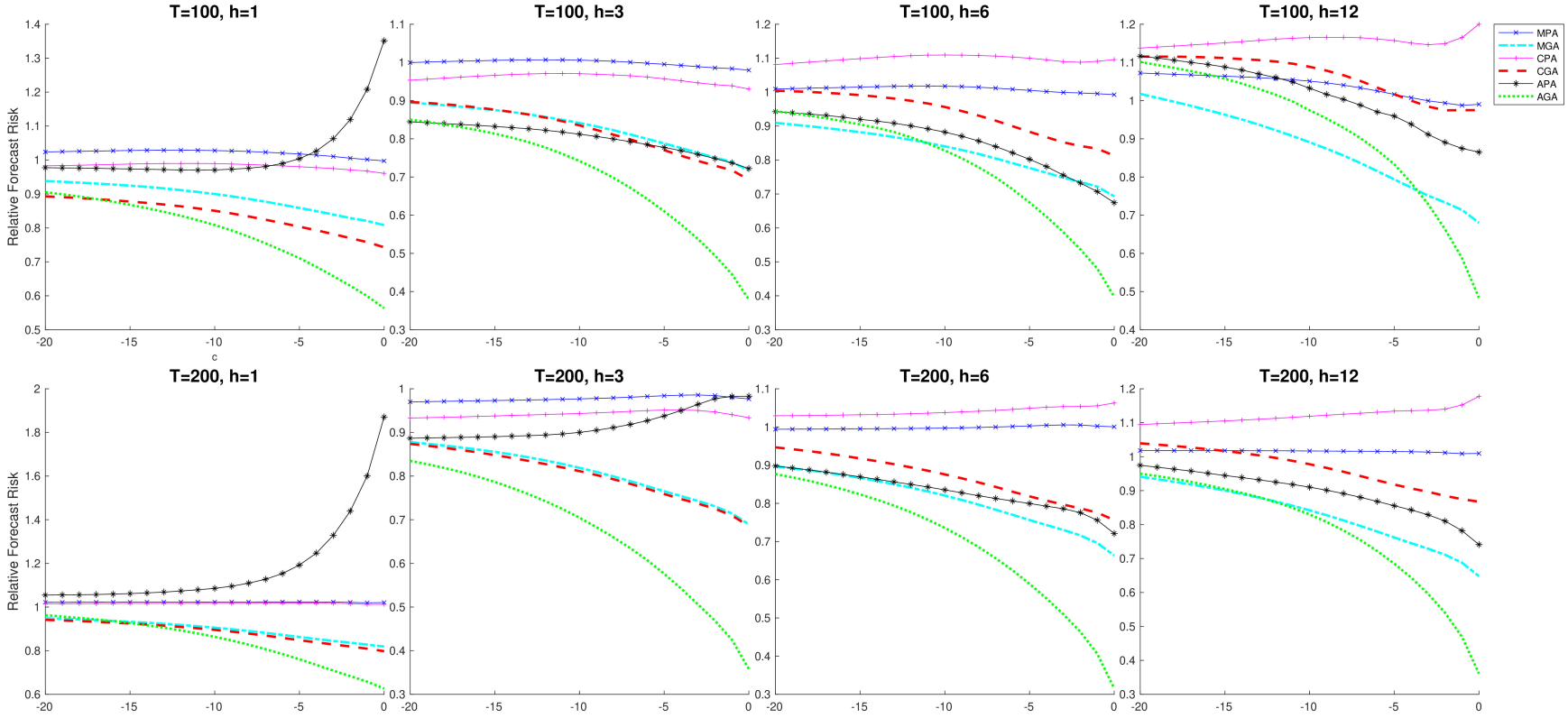

Figure 7: Forecast risk with unknown lag order ($k = 6$)

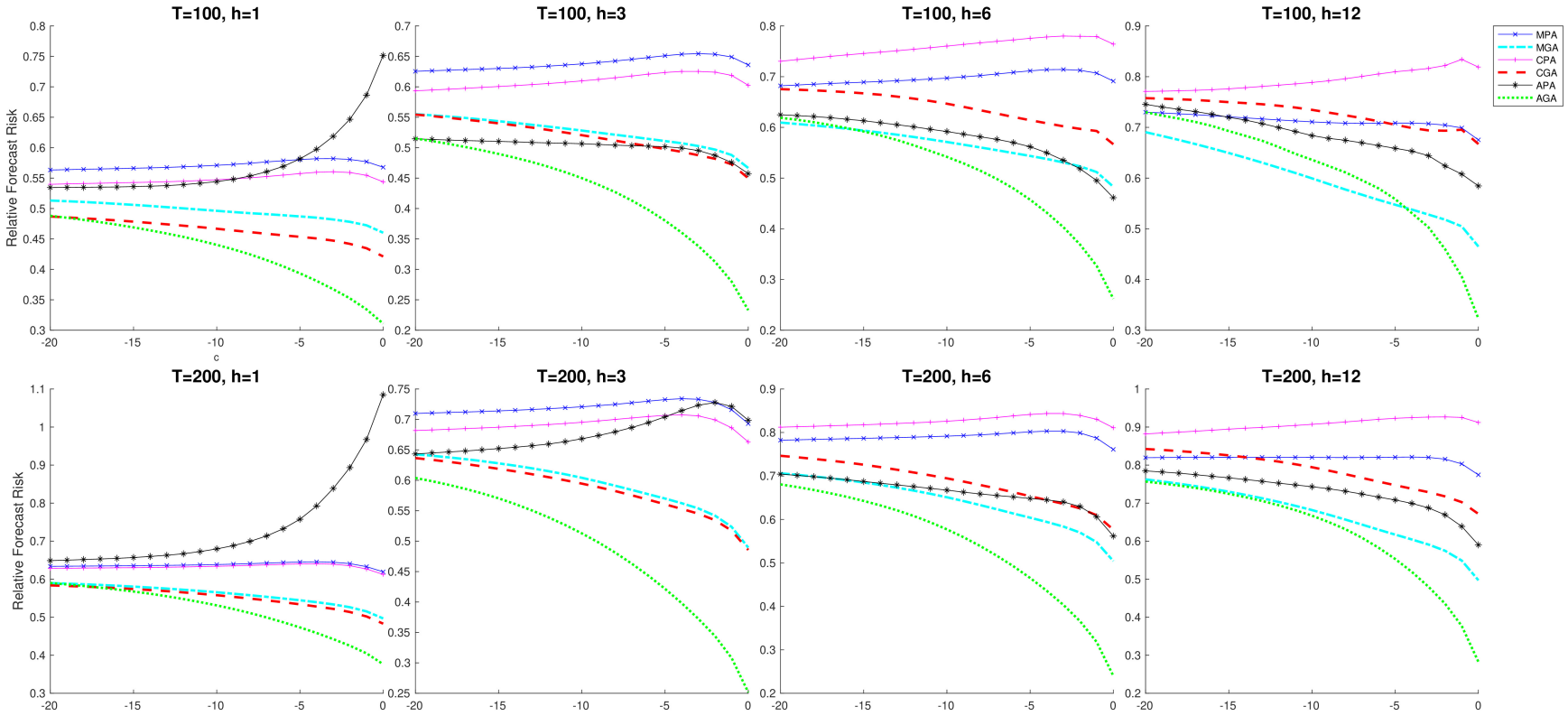

Figure 8: Forecast risk with unknown lag order ($k = 12$)

Table 1a: Percentage wins/losses of different forecasting methods for $h = 1$ & $h = 3$

Methods	MPA	MGA	CPA	CGA	APA	AGA	MS	CVhS	APES	AR	All
MPA	0.0	4.9	22.0	3.3	43.9	7.3	96.7	58.5	25.2	100.0	0.0
MGA	95.1	0.0	61.8	13.8	61.8	13.0	99.2	71.5	40.7	100.0	4.9
CPA	78.0	38.2	0.0	18.7	66.7	16.3	95.1	81.3	36.6	100.0	5.7
CGA	96.7	86.2	81.3	0.0	78.0	23.6	98.4	95.9	53.7	100.0	8.9
APA	56.1	38.2	33.3	22.0	0.0	21.1	71.5	54.5	35.0	78.9	11.4
AGA	92.7	87.0	83.7	76.4	78.9	0.0	97.6	93.5	77.2	100.0	**54.5**
MS	3.3	0.8	4.9	1.6	28.5	2.4	0.0	18.7	9.8	90.2	0.0
CVhS	41.5	28.5	18.7	4.1	45.5	6.5	81.3	0.0	22.0	95.1	0.8
APES	74.8	59.3	63.4	46.3	65.0	22.8	90.2	78.0	0.0	95.9	13.8
AR	0.0	0.0	0.0	0.0	21.1	0.0	9.8	4.9	4.1	0.0	0.0
MPA	0.0	6.5	43.1	8.1	50.4	10.6	87.0	58.5	23.6	97.6	0.8
MGA	93.5	0.0	65.9	43.1	67.5	17.1	97.6	73.2	36.6	99.2	8.9
CPA	56.9	34.1	0.0	15.4	59.3	16.3	81.3	68.3	30.1	96.7	4.1
CGA	91.9	56.9	84.6	0.0	84.6	23.6	95.1	93.5	43.1	99.2	10.6
APA	49.6	32.5	40.7	15.4	0.0	14.6	62.6	48.0	29.3	71.5	6.5
AGA	89.4	82.9	83.7	76.4	85.4	0.0	95.9	93.5	69.9	100.0	**41.5**
MS	13.0	2.4	18.7	4.9	37.4	4.1	0.0	34.1	10.6	80.5	0.0
CVhS	41.5	26.8	31.7	6.5	52.0	6.5	65.9	0.0	14.6	89.4	0.8
APES	76.4	63.4	69.9	56.9	70.7	30.1	89.4	85.4	0.0	95.1	26.8
AR	2.4	0.8	3.3	0.8	28.5	0.0	19.5	10.6	4.9	0.0	0.0

Note: Percentage of the 123 series for which a method listed in a row outperforms a method in a column, and all other methods (last column). AR refers to the benchmark autoregressive model that uses 12 lags of the first differences (see section 7 of the main text for details).

Table 1b: Percentage wins/losses of different forecasting methods for $h = 6$ & $h = 12$

Methods	MPA	MGA	CPA	CGA	APA	AGA	MS	CVhS	APES	AR	All
MPA	0.0	7.3	61.8	12.2	57.7	21.1	82.9	57.7	32.5	95.9	3.3
MGA	92.7	0.0	83.7	52.0	83.7	34.1	98.4	83.7	46.3	98.4	13.0
CPA	38.2	16.3	0.0	13.8	60.2	21.1	65.0	60.2	30.1	87.0	8.1
CGA	87.8	48.0	86.2	0.0	86.2	35.8	93.5	96.7	44.7	98.4	13.8
APA	42.3	16.3	39.8	13.8	0.0	21.1	56.1	47.2	30.9	65.9	3.3
AGA	78.9	65.9	78.9	64.2	78.9	0.0	92.7	82.9	69.9	99.2	**34.1**
MS	17.1	1.6	35.0	6.5	43.9	7.3	0.0	36.6	17.1	78.0	0.0
CVhS	42.3	16.3	39.8	3.3	52.8	17.1	63.4	0.0	19.5	78.0	0.8
APES	67.5	53.7	69.9	55.3	69.1	30.1	82.9	80.5	0.0	96.7	23.6
AR	4.1	1.6	13.0	1.6	34.1	0.8	22.0	22.0	3.3	0.0	0.0
MPA	0.0	10.6	60.2	16.3	74.8	21.1	85.4	46.3	23.6	91.9	1.6
MGA	89.4	0.0	75.6	48.0	84.6	27.6	97.6	70.7	30.9	97.6	10.6
CPA	39.8	24.4	0.0	13.8	66.7	25.2	58.5	47.2	26.8	81.3	8.9
CGA	83.7	52.0	86.2	0.0	81.3	36.6	91.9	91.9	40.7	98.4	10.6
APA	25.2	15.4	33.3	18.7	0.0	22.8	47.2	35.8	28.5	65.9	8.1
AGA	78.9	72.4	74.8	63.4	77.2	0.0	87.8	77.2	65.0	95.9	**31.7**
MS	14.6	2.4	41.5	8.1	52.8	12.2	0.0	33.3	15.4	78.9	0.8
CVhS	53.7	29.3	52.8	8.1	64.2	22.8	66.7	0.0	26.8	82.9	4.1
APES	76.4	69.1	73.2	59.3	71.5	35.0	84.6	73.2	0.0	91.9	23.6
AR	8.1	2.4	18.7	1.6	34.1	4.1	21.1	17.1	8.1	0.0	0.0

Note: Percentage of the 123 series for which a method listed in a row outperforms a method in a column, and all other methods (last column). AR refers to the benchmark autoregressive model that uses 12 lags of the first differences (see section 7 of the main text for details).

Table 2: Best forecasting methods by group

		Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
h	1	AGA	AGA	AGA	AGA	APA	AGA	AGA	AGA
	3	AGA	AGA	APES	AGA	CGA	APES	APA AGA	AGA
	6	AGA	APES	APES	AGA	CPA CGA	AGA	MGA AGA	AGA
	12	APES	AGA	AGA APES	AGA	CGA	AGA	APA	AGA
GA $\succ$ PA		3	3	1	4	1	3	2	4
PA $\succ$ GA		0	0	0	0	2	0	1	0
AVE $\succ$ SEL		3	3	1	4	4	3	4	4
SEL $\succ$ AVE		1	1	2	0	0	1	0	0
M $\succ$ (CVh,APE)		0	0	0	0	0	0	0	0
CVh $\succ$ (APE,M)		0	0	0	0	3	0	0	0
APE $\succ$ (M,CVh)		4	4	4	4	1	4	3	4

Note: *The groups are defined as in McCracken and Ng (2016): (1) Output and income; (2) labour market; (3) housing; (4) consumption, orders, and inventories; (5) money and credits; (6) interest and exchange rates; (7) prices; (8) stock market.
**The last 7 row counts excluded pairwise ties.

Table 3: Relative MSFE of core real macroeconomic time series

Method	Industrial Production	Personal Income	Mfg & trade sales	Nonag. employment	Industrial Production	Personal Income	Mfg & trade sales	Nonag. employment
h			1				6	
MPA	.971**	.964	.983	.956***	.978	.991	.985	.952*
MGA	.965***	.959*	.970**	.948***	**.945**	.949	.934	.918***
CPA	.967**	.919*	.983	.961***	.993	.999	1.009	.961*
CGA	.957***	.908*	.968**	.949***	.952	.968	.953	.936**
APA	1.040	.930	.995	1.262***	1.122	1.033	1.043	1.360***
AGA	**.947****	**.889****	**.956****	.923***	.974	**.891***	.916*	.925*
MS	.984	.992	.992	.946***	1.009	1.021	.983	.939**
CVhS	.985	.946	.986	.945***	1.007	1.003	1.037	.953**
APES	.972	.905*	.973	**.922*****	.981	.918	**.915**	**.911***
h			3				12	
MPA	.976	.979	.986	.955**	.970	.972	.995	.968
MGA	**.962**	.961	.957	.937***	.906***	.905**	.886**	.895***
CPA	.987	.978	.997	.963**	1.027	1.007	1.023	1.005
CGA	.972	.957*	.965	.950***	.852*	.957	.848	.912*
APA	1.121*	.997	1.031	1.419***	1.173	1.010	1.023	1.305***
AGA	.970	**.922****	**.955**	.921**	.837**	**.775****	.751**	**.841****
MS	.990	1.006	.983	.942***	.988	1.001	.985	.962
CVhS	.999	1.004	1.017	.957***	.914	1.020	.926	.966
APES	.974	.941	.969	**.902*****	**.814****	.777**	**.737***	.851**

Note: Here, *denotes 10%, **denotes 5%, and ***denotes 1% significance level for a twsided Diebold and Mariano (1995) test. The benchmark is an unrestricted OLS estimation method with 12 lags (see section 7 of the main text for details).

Table 4: Relative MSFE of core nominal macroeconomic time series

Method	CPI	Consumption deflator	CPI exc. food	PPI	CPI	Consumption deflator	CPI exc. food	PPI
h			1				6	
MPA	.966*	.963**	.965*	.952**	.985	.975	.970	.957
MGA	.957**	.957**	.961**	.944**	.964	.954	.964	.950*
CPA	.962	.955**	.957*	.938**	.973	.984	.961	.953
CGA	.955*	.952**	.957*	.937**	.952	.952	**.943***	**.942***
APA	.958	.953**	.960	**.930****	.971	.966	.959	.944*
AGA	**.948***	**.949****	**.949***	.941**	**.949**	**.947**	.945	.955
MS	1.001	.994	.999	.999	1.038	1.009	1.011	.983
CVhS	.980	.979	.967	.946*	.977	.992	.984	.969
APES	.989	.982	.968	.939*	.967	.985	.964	.968
h			3				12	
MPA	.969	.966	.963	.946**	.972	.984	.955*	.954*
MGA	.957	.956	.959	.938**	.948*	.961	.945	**.949***
CPA	.957	.959	.952	.933*	.955	.967	.939**	.954
CGA	.942	.946	**.945**	**.926****	.947*	.949	.928**	.952*
APA	.956	.951	.962	.927**	**.937****	**.935**	**.920****	.958
AGA	**.938**	**.936**	**.945**	.936**	.944	.939	.938	.972
MS	1.029	1.017	.989	.991	1.006	1.035	.998	.992
CVhS	.983	.987	.981	.938*	.960	.977	.946*	.956
APES	.970	.957	.974	.942*	.958	.957	.968	.972

Note: Here, *denotes 10%, **denotes 5%, and ***denotes 1% significance level for a twsided Diebold and Mariano (1995) test. The benchmark is an unrestricted OLS estimation method with 12 lags (see section 7 of the main text for details).

Supplementary Appendix A: Proofs

Let $W(.)$ denote a standard Brownian motion on $[0,1]$ and define the diffusion process: $dW_c(r) = cW_c(r) + dW(r)$. For $p \in \{0,1\}$, let $X_c(r) = (r^p, W_c(r))'$ and define the detrended processes

$$W_c^*(r,p) = \begin{cases} W_c(r) & \text{if} \quad p=0 \\ W_c(r) - \int_0^1 W_c(s)ds & \text{if} \quad p=1 \end{cases}$$

$$X_c^*(r,p) = \begin{cases} X_c(r) & \text{if} \quad p=0 \\ X_c(r) - \int_0^1 X_c(s)ds & \text{if} \quad p=1 \end{cases}$$

and the functionals

$$T_{0c} = -cW_c^*(1,p) + I(p=1)W(1)$$
$$T_{1c} = X_c^*(1,p)' \left(\int_0^1 X_c^*(r,p) X_c^*(r,p)' \right)^{-1} \int_0^1 X_c^*(r,p) dW(r) + I(p=1)W(1).$$

Let $\beta = (\beta_0, \beta_1)'$, $z_t = (1,t)'$. Without loss of generality, we assume that $\beta_0 = \beta_1 = 0$ in the true data generating process. For a matrix A, $\|A\|^2 = \sup_{\|v\|=1} v'A'Av$ with $\|v\|$ denoting the Euclidean norm for vector v. Unless otherwise defined, for any variable x, we use x^* to denote its demeaned version. For a random quantity δ, we write $\delta = \delta_0 + o_p(\delta_0)$ as $\delta = \delta_0 + s.o.$, where $s.o.$ represents a term of smaller order in probability. For brevity, all proofs are provided only for the case $p=1$. The proofs for $p=0$ are simpler and follow analogous arguments.

We start by noting that if u_t is generated by (1), it has the $AR(k+1)$ representation $u_t = \sum_{i=1}^{k+1} a_i u_{t-i} + e_t$, where $a_1 = \alpha + \alpha_1$, $a_i = \alpha_i - \alpha_{i-1}$ $(i = 2, ..., k)$, $a_{k+1} = -\alpha_k$. The companion VAR(1) form of the model is expressed as

$$Y_t = B'Y_{t-1} + \nu_t$$

where

$$\underset{(k+3)\times 1}{Y_t} = (1, t+1, y_t, ..., y_{t-k})', \quad \underset{(k+3)\times 1}{\nu_t} = (0, 0, e_t, 0, ..., 0)'$$

$$\underset{(k+3)\times(k+3)}{B(k+1)} = \begin{pmatrix} B_1 & B_2 \\ \mathbf{0}_{(k+1)\times 2} & F \end{pmatrix}, \quad \underset{(2\times 2)}{B_1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \underset{2\times(k+1)}{B_2} = \begin{pmatrix} 0 & 0 & ... & 0 \\ 0 & 0 & ... & 0 \end{pmatrix}$$

$$\underset{(k+1)\times(k+1)}{F} = \left(\begin{array}{c|c} a(k) & I_k \\ \hline & \mathbf{0}_k' \end{array} \right), \quad F^0 = I_{k+1}, \quad a(k) = (a_1, ..., a_{k+1})'$$

With "hat" and "tilde" denoting the unrestricted and restricted OLS estimates, respectively, the unrestricted and restricted forecasts can then be expressed as (see, e.g., Ing, 2003):

$$\hat{\mu}_{T+h} = \boldsymbol{y}_T(k+1)'\hat{B}^{h-1}\hat{\gamma} \tag{A.1}$$

$$\widetilde{\mu}_{T+h} = \boldsymbol{y}_T(k+1)'\widetilde{B}^{h-1}\widetilde{\gamma} \tag{A.2}$$

where $\boldsymbol{y}_T(k+1) = (1,\ T+1,\ y_T, ..., y_{T-k})'$, $\hat{\gamma} = (\hat{\beta}_0^*, \hat{\beta}_1^*, \hat{a}_1, ..., \hat{a}_{k+1})'$ and

$$\underset{(k+3)\times(k+3)}{\hat{B}(k+1)} = \begin{pmatrix} B_1 & \hat{B}_2 \\ \mathbf{0}_{(k+1)\times 2} & \hat{F} \end{pmatrix},$$

$$\underset{(2\times 2)}{B_1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \underset{2\times(k+1)}{\hat{B}_2} = \begin{pmatrix} \hat{\beta}_0^* & 0 & ... & 0 \\ \hat{\beta}_1^* & 0 & ... & 0 \end{pmatrix}$$

$$\underset{(k+1)\times(k+1)}{\hat{F}} = \left(\begin{array}{c|c} \hat{a}(k) & I_k \\ \hline & \mathbf{0}_k' \end{array}\right), \ \hat{F}^0 = I_{k+1}, \quad \hat{a}(k) = (\hat{a}_1, ..., \hat{a}_{k+1})'$$

$$\underset{(k+3)\times(k+3)}{\widetilde{B}(k+1)} = \begin{pmatrix} B_1 & \widetilde{B}_2 \\ \mathbf{0}_{(k+1)\times 2} & \widetilde{F} \end{pmatrix}, \quad \underset{2\times(k+1)}{\widetilde{B}_2} = \begin{pmatrix} \widetilde{\beta}_0^* & 0 & ... & 0 \\ 0 & 0 & ... & 0 \end{pmatrix} \tag{A.3}$$

The matrix $\widetilde{F}$ is constructed in the same way as $\hat{F}$ with $\hat{a}(k)$ replaced by $\widetilde{a}(k)$, where $\widetilde{a}(k) = (\widetilde{a}_1, ..., \widetilde{a}_{k+1})' = (1+\widetilde{\alpha}_1, \widetilde{\alpha}_2 - \widetilde{\alpha}_1, ..., \widetilde{\alpha}_k - \widetilde{\alpha}_{k-1}, -\widetilde{\alpha}_k)'$ with $\widetilde{\gamma} = (\widetilde{\beta}_0^*, 0, \widetilde{a}_1, ..., \widetilde{a}_{k+1})'$.

Next, we state a set of lemmas that will be useful in developing the proofs of the main results. Lemmas A.1-A.4, A.7-A.9 below parallel Lemmas A.1-A.4, B.1-B.3 in Ing et al. (2009) who assume an exact unit root ($c = 0$). Since the sample moments have the same order whether $c = 0$ or $c < 0$, the proofs of the following lemmas also follow directly those in Ing et al. (2009) and are hence omitted.

Lemma A.1 *Suppose $\{y_t\}$ satisfies (1) and Assumptions (1)-(2). Then for any $q > 0$,*

$$E||\hat{R}_T^{-1}||^q = O(1)$$

where

$$\hat{R}_T = T^{-1}D_T \sum_{j=k+1}^{T-1} \boldsymbol{y}_j(k+1)\boldsymbol{y}_j(k+1)'D_T{}'$$

with

$$\underset{(k+3)\times(k+3)}{D_T} = diag(1, T^{-1}, \bar{D}_T),$$

$$\underset{(k+1)\times(k+1)}{\bar{D}_T} = \begin{pmatrix} \frac{1}{\sqrt{T}} & \frac{-\alpha_1}{\sqrt{T}} & \cdots & \cdots & \frac{-\alpha_k}{\sqrt{T}} \\ 1 & -1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 & -1 \end{pmatrix}$$

Lemma A.2 *Suppose* $\{y_t\}$ *satisfies (1) and Assumptions (1)-(2) and for some* $q_1 \geq 2$, $\sup_{-\infty \leq t \leq \infty} E|e_t|^{2q_1} < \infty$. *Then for any* $0 < q < q_1$,

$$E||\hat{R}_T^{-1} - \hat{R}_T^{*-1}||^q = O(T^{-q/2})$$

where

$$\underset{(k+3)\times(k+3)}{\hat{R}_T^*} = diag(\underset{3\times 3}{\hat{R}_c^*}, \underset{k\times k}{\hat{\Gamma}_T(k)})$$

$$\hat{R}_c^* = \begin{pmatrix} T^{-1}(T-1-k) & T^{-1}\sum_{j=k+1}^{T-1} X_t' \\ T^{-1}\sum_{j=k+1}^{T-1} X_t & T^{-2}\sum_{j=k+1}^{T-1} X_t X_t' \end{pmatrix}$$

$$X_t = [T^{-1}(t+1),\ T^{-1/2}N_t]',\ N_j = A(L)y_j$$

$$\hat{\Gamma}_T(k) = T^{-1}\sum_{j=k+1}^{T-1} \boldsymbol{s}_j(k)\boldsymbol{s}_j(k)'$$

Lemma A.3 *Suppose* $\{y_t\}$ *satisfies (1) and Assumptions (1)-(2) with* $\sup_{-\infty \leq t \leq \infty} E|e_t|^q < \infty$ *for some* $q \geq 2$. *Then,*

$$E||T^{-1/2}D_T \sum_{j=k+1}^{T-1} \boldsymbol{y}_j(k+1)e_{j+1}||^q = O(1)$$

Lemma A.4 *Suppose* $\{y_t\}$ *satisfies (1) and Assumptions (1)-(2) with* $\sup_{-\infty \leq t \leq \infty} E|e_t|^r < \infty$ *for some* $r > 4$.. *Then,*

$$\lim_{T\to\infty} E(F_{T,k}) = 0$$

where

$$F_{T,k} = \boldsymbol{s}_T(k)' M_h(k) \hat{\Gamma}_T^{-1}(k) \Big\{ \sum_{j=k+1}^{T-1} \boldsymbol{s}_j(k) e_{j+1} \Big\} X_T' \big(\sum_{j=k+1}^{T-1} X_j X_j{}' \big)^{-1} \left\{ \sum_{j=k+1}^{T-1} X_j e_{j+1} \right\}$$

Lemma A.5 *Let* $\underset{T\times(p+1)}{X} = [\underset{T\times 1}{X_1}, \underset{T\times p}{X_2}]$, $X_1 = (1, \cdots, 1)'$, *and assume* $X'X$ *is invertible. Define* $M_1 = \underset{T\times T}{I} - X_1(X_1'X_1)^{-1}X_1'$, $X_2^* = M_1 X_2$. *For any* $T \times 1$ *vector* e *and any* $p \times 1$ *vector* x_2, *we have* $x'(X'X)^{-1}X'e = x_1(X_1'X_1)^{-1}X_1'e + x_2^{*\prime}(X_2^{*\prime}X_2^*)^{-1}X_2^{*\prime}e$, *where* $x = (x_1, x_2')'$, $x_1 = 1$, $x_2^* = x_2 - (X_1'X_1)^{-1}X_2'X_1$.

Lemma A.6 *Under Assumptions (1)-(2),* $\frac{\sqrt{T}\widetilde{\beta}_0^*}{\sigma} \xrightarrow{d} W_c(1)$.

Lemma A.7 *Under Assumptions (1)-(2) and* $\sup_{-\infty\le t\le\infty} E|e_t|^q < \infty$ *for some* $q > 2$,
(i) For some $\kappa_1 > 0$, $||\hat{\Gamma}(k) - \Gamma(k)|| = o(T^{-\kappa_1})$ *a.s.;*
(ii) For some $\kappa_2 > 0$, $||\hat{R}_T - \hat{R}_T^*|| = o(T^{-\kappa_2})$ *a.s.;*
(iii) $||\hat{R}_T^{-1}|| = O(\log\log T)$ *a.s..*

Lemma A.8 *Under Assumptions (1)-(2) and* $\sup_{-\infty\le t\le\infty} E|e_t|^q < \infty$ *for some* $q > 2$, $\sum_{i=m_h}^{T-h} F_{i,k} = o(T)$ *a.s., where*

$$F_{i,k} = \boldsymbol{s}_i(k)' M_h(k) \hat{\Gamma}_i^{-1}(k) \Big\{ \sum_{j=k+1}^{i-1} \boldsymbol{s}_j(k) e_{j+1} \Big\} X_i' \big(\sum_{j=k+1}^{i-1} X_j X_j{}' \big)^{-1} \left\{ \sum_{j=k+1}^{i-1} X_j e_{j+1} \right\}$$

Lemma A.9 *Let* $\{x_T\}$ *be a sequence of real numbers.*
(i) If $x_T \ge 0, T^{-1}\sum_{j=1}^T x_j = O(1)$, *and for some* $\xi > 1, \liminf_{T\to\infty} \nu_T/T^\xi > 0$, *then,* $\sum_{j=1}^T x_j/\nu_j = O(1)$*;*
(ii) If $T^{-1}\sum_{j=1}^T x_j = o(1)$, *then,* $\sum_{j=1}^T x_j/j = o(\log T)$.

Proof of Lemma A.5. Note, by block matrix inversion,

$$(X'X)^{-1} = \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} (X_1'X_1)^{-1} + (X_1'X_1)^{-1}X_1'X_2(X_2'M_1X_2)^{-1}X_2'X_1(X_1'X_1)^{-1} & -(X_1'X_1)^{-1}X_1'X_2(X_2'M_1X_2)^{-1} \\ -(X_2'M_1X_2)^{-1}X_2'X_1(X_1'X_1)^{-1} & (X_2'M_1X_2)^{-1} \end{pmatrix}$$

then

$$(X'X)^{-1}X'e = \begin{pmatrix} (X_1'X_1)^{-1}X_1'[I - X_2(X_2'M_1X_2)^{-1}X_2'M_1]e \\ (X_2'M_1X_2)^{-1}X_2'M_1e \end{pmatrix}$$

Recall $x = (x_1, x_2')' = (x_1, x_2^{*\prime})' + (0, X_1'X_2(X_1'X_1)^{-1})'$, we have,

$$\begin{aligned}
x'(X'X)^{-1}X'e =& \underbrace{[(x_1, x_2^{*\prime})](X'X)^{-1}X'e}_{\text{Term 1}} + \underbrace{[(0, X_1'X_2(X_1'X_1)^{-1})](X'X)^{-1}X'e}_{\text{Term 2}} \\
=& \underbrace{x_1(X_1'X_1)^{-1}X_1'[I - X_2(X_2'M_1X_2)^{-1}X_2'M_1]e + x_2^{*\prime}(X_2'M_1X_2)^{-1}X_2'M_1e}_{\text{Term 1}} \\
&+ \underbrace{X_1'X_2(X_1'X_1)^{-1}(X_2'M_1X_2)^{-1}X_2'M_1e}_{\text{Term 2}} \\
=& x_1(X_1'X_1)^{-1}X_1'e + x_2^{*\prime}(X_2'M_1X_2)^{-1}X_2'M_1e \\
&\underbrace{-x_1(X_1'X_1)^{-1}X_1'X_2(X_2'M_1X_2)^{-1}X_2'M_1e + X_1'X_2(X_1'X_1)^{-1}(X_2'M_1X_2)^{-1}X_2'M_1e}_{=0,\text{ since } x_1=1,\ (X_1'X_1)^{-1}=1/T,\text{ which is a constant}} \\
=& x_1(X_1'X_1)^{-1}X_1'e + x_2^{*\prime}(X_2^{*\prime}X_2^*)^{-1}X_2^{*\prime}e
\end{aligned}$$

■

Proof of Lemma A.6. The true DGP can be expressed as

$$\Delta y_t = \beta_0^* + \sum_{j=1}^{k} \alpha_j \Delta y_{t-j} + e_t^*$$

where $\beta_0^* = 0$ and $e_t^* = \frac{ac}{T}u_{t-1} + e_t$. Let $\dot{Z}_t = (\dot{Z}_1, \dot{Z}_{2,t}')$, $\dot{Z}_1 = 1$, $\dot{Z}_{2,t} = (\Delta y_{t-1}, \cdots, \Delta y_{t-k})'$, $\boldsymbol{\iota}_1 = (1, 0, \cdots, 0)'$, $\boldsymbol{\iota}_{[2:k+1]} = (0, 1, \cdots, 1)'$. Now

$$\begin{aligned}
\frac{\sqrt{T}\widetilde{\beta}_0^*}{\sigma} =& \frac{\sqrt{T}}{\sigma}\boldsymbol{\iota}_1'(\sum_{t=k+1}^{T}\dot{Z}_t\dot{Z}_t')^{-1}\sum_{t=k+1}^{T}\dot{Z}_t'(\frac{ac}{T}u_{t-1} + e_t) \qquad \text{(A.4)} \\
=& \frac{\sqrt{T}}{\sigma}(\sum_{t=k+1}^{T}\dot{Z}_1^2)^{-1}\sum_{t=k+1}^{T}\dot{Z}_1(\frac{ac}{T}u_{t-1} + e_t) + o_p(1) \\
=& \frac{ca}{\sigma\sqrt{T}}\sum_{t=k+1}^{T}u_{t-1} + \frac{1}{\sigma\sqrt{T}}\sum_{t=k+1}^{T}e_t + o_p(1) \xrightarrow{d} c\int_0^1 W_c + W(1) = W_c(1)
\end{aligned}$$

■

Proof of Theorem 1. (a) Defining $\gamma = (\beta_0^*, \beta_1^*, a_1, ..., a_{k+1})'$, $\hat{L}_h = \sum_{j=0}^{h-1} b_j \hat{B}^{h-1-j}$, and

$L_h = \sum_{j=0}^{h-1} b_j B^{h-1-j}$, we can write

$$\begin{aligned}
\frac{T}{\sigma^2}E(\hat{\mu}_{T+h} - \mu_{T+h})^2 =& \frac{T}{\sigma^2}E\big[\boldsymbol{y}_T(k+1)'\hat{L}_h(\hat{\gamma}-\gamma)\big]^2 \\
=& \frac{T}{\sigma^2}\Bigg[E\big[\boldsymbol{y}_T(k+1)'L_h(\hat{\gamma}-\gamma)\big]^2 \\
& + E\big[\boldsymbol{y}_T(k+1)'\big\{\hat{L}_h - L_h\big\}(\hat{\gamma}-\gamma)\big]^2 + o(1)\Bigg] \\
=& \frac{1}{\sigma^2}E\Big[\boldsymbol{y}_T(k+1)'L_h D_T'(\hat{R}_T^{*-1})\frac{D_T}{\sqrt{T}}\sum_{j=k+1}^{T-1}\boldsymbol{y}_j(k+1)e_{j+1}\Big]^2 \\
& + \frac{1}{\sigma^2}E\Big[\boldsymbol{y}_T(k+1)'L_h D_T'(\hat{R}_T^{-1} - \hat{R}_T^{*-1})\frac{D_T}{\sqrt{T}}\sum_{j=k+1}^{T-1}\boldsymbol{y}_j(k+1)e_{j+1}\Big]^2 \\
& + \frac{T}{\sigma^2}E\Big[\boldsymbol{y}_T(k+1)'\big\{\hat{L}_h - L_h\big\}(\hat{\gamma}-\gamma)\Big]^2 + o(1) \\
=& (I) + (II) + (III)
\end{aligned} \tag{A.5}$$

The (II) and (III) terms in (A.5) are each $o(1)$ by Lemmas A.1-A.3 and Holder's inequality [see, e.g. the proof of Theorem 2.2 in Ing et al., 2009].

The term (I) can be written as:

$$\begin{aligned}
& \frac{1}{\sigma^2}E\big[\boldsymbol{y}_T(k+1)'L_h D_T'\hat{R}_T^{*-1}\frac{D_T}{\sqrt{T}}\sum_{j=k}^{T-1}\boldsymbol{y}_j(k+1)e_{j+1}\big]^2 \\
&= \frac{1}{\sigma^2}E\big[\boldsymbol{y}_T(k+1)'D_T'\bar{L}_h\hat{R}_T^{*-1}\frac{D_T}{\sqrt{T}}\sum_{j=k}^{T-1}\boldsymbol{y}_j(k+1)e_{j+1}\big]^2
\end{aligned} \tag{A.6}$$

where $\bar{L}_h = \sum_{j=0}^{h-1} b_j diag(G_T^{h-1-j}, \bar{F}^{h-1-j})$ with $G_T = \begin{pmatrix} 1 & T^{-1} \\ 0 & 1 \end{pmatrix}$, $\bar{F} = diag(1, S_M(k))$ and $S_M(k) = \begin{pmatrix} \alpha(k-1) & I_{k-1} \\ \alpha_k & \mathbf{0}_{k-1}' \end{pmatrix}$, $S_M^0(k) = I_k$.

Note that $\boldsymbol{y}_T(k+1)'D_T' = \big(1, T^{-1}(T+1), T^{-1/2}N_T, \boldsymbol{s}_T(k)\big)$. Further, since G_T is upper trian-

gular, (A.6) converges to

$$\frac{1}{\sigma^2}\Big(\sum_{j=0}^{h-1} b_j\Big)^2 \lim_{T\to\infty} E\Big\{T^{-1/2}\sum_{j=k+1}^{T-1} e_{j+1} + X_T^{*\prime}\Big(\sum_{j=k+1}^{T-1} X_j^* X_j^{*\prime}\Big)^{-1}\sum_{j=k+1}^{T-1} X_j^* e_{j+1}\Big\}^2$$
$$+ \lim_{T\to\infty}\frac{1}{\sigma^2}E\Big\{\boldsymbol{s}_T'(k)M_h(k)\hat{\Gamma}_T^{-1}(k)T^{-1/2}\sum_{j=k+1}^{T-1}\boldsymbol{s}_j(k)e_{j+1}\Big\}^2 + \frac{2}{\sigma^2}\Big(\sum_{j=0}^{h-1} b_j\Big)\lim_{T\to\infty} E(F_{T,k})$$
$$= B.1 + B.2 + B.3 \tag{A.7}$$

where $B.1$ utilizes Lemma A.5. Since $B.2 = g_h(k)$ by Theorem 1 of Ing (2003) and $B.3 = 0$ by Lemma A.4, (A.7) simplifies to:

$$B.1 + B.2 = \Big(\sum_{j=0}^{h-1} b_j\Big)^2 \lim_{T\to\infty} E\Big\{W(1) + X_c^*(1)'\Big(\int_0^1 X_c^* X_c^{*\prime}\Big)^{-1}\int_0^1 X_c^* dW\Big\}^2 + g_h(k)$$
$$= \Big(\sum_{j=0}^{h-1} b_j\Big)^2 E\big[T_{1c}^2\big] + g_h(k) \tag{A.8}$$

The required result then follows from (A.5), (A.7) and (A.8).

(b) Defining $\widetilde{L}_h = \sum_{j=0}^{h-1} b_j \widetilde{B}^{h-1-j}$, with similar arguments as in (a), we can write:

$$\frac{T}{\sigma^2}E(\widetilde{\mu}_{T+h} - \mu_{T+h})^2 = \frac{T}{\sigma^2}E\Big[\boldsymbol{y}_T(k+1)'\widetilde{L}_h(\widetilde{\gamma} - \gamma)\Big]^2$$
$$= \frac{T}{\sigma^2}E\Big[\boldsymbol{y}_T(k+1)'L_h(\widetilde{\gamma} - \gamma)\Big]^2 + o(1) \tag{A.9}$$

Note that

$$L_h = \sum_{j=0}^{h-1} b_j B^{h-1-j} = \sum_{j=0}^{h-1} b_j \begin{pmatrix} B_1 & 0 \\ 0 & F \end{pmatrix}^{h-1-j} = \sum_{j=0}^{h-1} b_j \begin{pmatrix} B_1^{h-1-j} & 0 \\ 0 & F^{h-1-j} \end{pmatrix}$$

Since B_1 is upper triangular with $B_1(1,1) = 1$,

$$(A.9) = \frac{T}{\sigma^2}E\Big[\boldsymbol{y}_T(k+1)'\begin{pmatrix} \sum_{j=0}^{h-1} b_j \begin{bmatrix} \widetilde{\beta}_0^* \\ 0 \end{bmatrix} \\ \sum_{j=0}^{h-1} b_j F^{h-1-j}[\widetilde{a}(k) - a(k)] \end{pmatrix}\Big]^2 + o(1)$$
$$= \frac{T}{\sigma^2}E\Big[\Big(\sum_{j=0}^{h-1} b_j\Big)\widetilde{\beta}_0^* + (y_T, ..., y_{T-k})\sum_{j=0}^{h-1} b_j F^{h-1-j}[\widetilde{a}(k) - a(k)]\Big]^2 + o(1) \tag{A.10}$$

Now, consider the term

$$\begin{aligned}
&\frac{\sqrt{T}}{\sigma}(y_T, ..., y_{T-k})\sum_{j=0}^{h-1} b_j F^{h-1-j}[\widetilde{a}(k) - a(k)] \\
=& \frac{\sqrt{T}}{\sigma}(y_T, ..., y_{T-k}) L_h^{(F)}[\widetilde{a}(k) - a(k)] \\
=& \frac{\sqrt{T}}{\sigma}(y_T, ..., y_{T-k}) L_h^{(F)}\Big[\hat{a}(k) - a(k) + HD_T'\hat{R}_T^{-1}D_T R'(RD_T'\hat{R}_T^{-1}D_T R')^{-1}(r - R\hat{\gamma})\Big] \qquad \text{(A.11)}
\end{aligned}$$

where

$$L_h^{(F)} = \sum_{j=0}^{h-1} b_j F^{h-1-j}, \underset{(k+1)\times(k+3)}{H} = \begin{bmatrix} \mathbf{0}_{(k+1)\times 2} & I_{(k+1)} \end{bmatrix}, \underset{2\times(k+3)}{R} = \begin{bmatrix} 0 & 1 & 0 & ... & 0 \\ 0 & 0 & 1 & ... & 1 \end{bmatrix}, \underset{2\times 1}{r} = (0,1)'$$

Next, defining $\bar{L}_h^{(F)} = diag(\sum_{j=0}^{h-1} b_j, M_h(k))$, $\underset{(k+1)\times 1}{\hat{\theta}} = (T(1-\hat{\alpha})/a, 0, \cdots, 0)'$, we have

$$\begin{aligned}
(A.11) =& \frac{1}{\sigma}(y_T, ..., y_{T-k})\Big[\sqrt{T}L_h^{(F)}\{\hat{a}(k) - a(k)\} + L_h^{(F)}\bar{D}_T'\hat{\theta}\Big] \\
=& \frac{1}{\sigma}(y_T, ..., y_{T-k})\Big[\sqrt{T}L_h^{(F)}\{\hat{a}(k) - a(k)\} + \bar{D}_T'\bar{L}_h^{(F)}\hat{\theta}\Big] \\
=& \frac{1}{\sigma}(y_T, ..., y_{T-k})\sqrt{T}L_h^{(F)}(\hat{a}(k) - a(k)) + \frac{1}{\sigma}(N_T/\sqrt{T}, \boldsymbol{s}_T(k))diag(\sum_{j=0}^{h-1} b_j, M_h(k))\hat{\theta} \\
=& \frac{1}{\sigma}(y_T, ..., y_{T-k})\sqrt{T}L_h^{(F)}(\hat{a}(k) - a(k)) + \frac{N_T}{\sigma\sqrt{T}}\sum_{j=0}^{h-1} b_j\big[-c - T(\hat{\alpha} - \alpha)/a\big] \\
=& \frac{N_T}{\sigma\sqrt{T}}\sum_{j=0}^{h-1} b_j\big\{T(\hat{\alpha} - \alpha)/a\big\} + \frac{1}{\sigma}\boldsymbol{s}_T(k)'M_h(k)\hat{\Gamma}_T^{-1}(k)T^{-1/2}\sum_{j=k}^{T-1}\boldsymbol{s}_j(k)e_{j+1} \\
&+ \frac{N_T}{\sigma\sqrt{T}}\sum_{j=0}^{h-1} b_j\big[-c - T(\hat{\alpha} - \alpha)/a\big] \\
=& -c\frac{N_T}{\sigma\sqrt{T}}\sum_{j=0}^{h-1} b_j + \frac{1}{\sigma}\boldsymbol{s}_T(k)'M_h(k)\hat{\Gamma}_T^{-1}(k)T^{-1/2}\sum_{j=k}^{T-1}\boldsymbol{s}_j(k)e_{j+1} \qquad \text{(A.12)}
\end{aligned}$$

Then, combining (A.10) with (A.12) and using Lemma A.6, we finally get

$$\begin{aligned}\lim_{T\longrightarrow\infty}\frac{T}{\sigma^2}E(\widetilde{\mu}_{T+h}-\mu_{T+h})^2 &= E\big[\sum_{j=0}^{h-1}b_j(W_c(1)-cW_c(1))\big]^2 \\ &+\frac{1}{\sigma^2}\lim_{T\longrightarrow\infty}E\Big\{\boldsymbol{s}_T'^{*}(k)M_h(k)\hat{\Gamma}_T^{-1}(k)T^{-1/2}\sum_{j=k}^{T-1}\boldsymbol{s}_j(k)e_{j+1}\Big\}^2 \\ &=\big(\sum_{j=0}^{h-1}b_j\big)^2E\big[T_{0c}^2\big]+g_h(k)\end{aligned}$$

which uses the fact that $W_c(1)-cW_c(1)=W(1)-cW_c^*(1)$, thereby proving the result. ■

Proof of Theorem 2. Henceforth, estimated parameters and quantities with subscript i denotes the estimates using observations from 1 to i. We prove (a) first. It follows from Chow (1965) and Ing (2004) that

$$APE_1-\sum_{i=m_h}^{T-h}\eta_{i,h}^2=\sum_{i=m_h}^{T-h}\Big[\boldsymbol{y}_i'(k+1)\hat{L}_{i,h}(\hat{\gamma}_i-\gamma)\Big]^2(1+o(1))+O(1)$$

Using similar algebra as in Theorem 1, we have:

$$\begin{aligned}\sum_{i=m_h}^{T-h}\Big[\boldsymbol{y}_i'(k+1)\hat{L}_{i,h}(\hat{\gamma}_i-\gamma)\Big]^2 &= \sum_{i=m_h}^{T-h}\Bigg[\big[\boldsymbol{y}_i(k+1)'L_h(\hat{\gamma}_i-\gamma)\big]^2 \\ &\quad+\big[\boldsymbol{y}_i(k+1)'\big\{\hat{L}_{i,h}-L_h\big\}(\hat{\gamma}_i-\gamma)\big]^2\Bigg]+s.o. \\ &=\sum_{i=m_h}^{T-h}\frac{1}{i}\Big[\boldsymbol{y}_i(k+1)'L_hD_i'(\hat{R}_i^{*-1})\frac{D_i}{\sqrt{i}}\sum_{j=k+1}^{i-1}\boldsymbol{y}_j(k+1)e_{j+1}\Big]^2 \\ &\quad+\sum_{i=m_h}^{T-h}\frac{1}{i}\Big[\boldsymbol{y}_i(k+1)'L_hD_i'(\hat{R}_i^{-1}-\hat{R}_i^{*-1})\frac{D_i}{\sqrt{i}}\sum_{j=k+1}^{i-1}\boldsymbol{y}_j(k+1)e_{j+1}\Big]^2 \\ &\quad+\sum_{i=m_h}^{T-h}\Big[\boldsymbol{y}_i(k+1)'\big\{\hat{L}_{i,h}-L_h\big\}(\hat{\gamma}_i-\gamma)\Big]^2+s.o. \\ &=(IV)+(V)+(VI)\end{aligned} \tag{A.13}$$

The (V) and (VI) terms in (A.13) are each $O(1)$ following similar arguments in Ing et al. (2009) which build on Lemmas A.7-A.9.

Analogous to (A.6) and (A.7) in the proof of Theorem 1, (IV) can be rewritten as:

$$\begin{aligned}(IV) =&\Big(\sum_{j=0}^{h-1} b_j\Big)^2 \sum_{i=m_h}^{T-h} \Big\{ Z_i'\Big(\sum_{j=k+1}^{i-1} Z_j Z_j'\Big)^{-1} \sum_{j=k+1}^{i-1} Z_j e_{j+1}\Big\}^2 \\ &+ \sum_{i=m_h}^{T-h} \Big\{ \boldsymbol{s}_i'(k) M_h(k) \hat{\Gamma}_i^{-1}(k) \frac{1}{i} \sum_{j=k+1}^{i-1} \boldsymbol{s}_j(k) e_{j+1}\Big\}^2 + 2\Big(\sum_{j=0}^{h-1} b_j\Big) \sum_{i=m_h}^{T-h} \frac{1}{i} F_{i,k} \\ &= C.1 + C.2 + C.3\end{aligned}$$

where $Z_j = (1, t+1, N_j)'$. In analogy with Theorem 3.1 of Ing (2004),

$$C.2 = g_h(k)\sigma^2 \log T + o_p(\log T) \tag{A.14}$$

By Lemmas A.8 and A.9, $C.3 = o_p(\log T)$. Now we focus on $C.1$. By Theorem 4 of Wei (1987), we have

$$C.1 = \Big(\sum_{j=0}^{h-1} b_j\Big)^2 \sigma^2 \log\det\Big(\sum_{j=k+1}^{T-1} Z_j Z_j'\Big) + o_p(\log T)$$

Defining the 3×3 matrix $\Upsilon_T = diag(T, T^3, T^2/|c|)$ and using Lemma A of Phillips (2014) in conjunction with the fact that $|c|\, T^{-2} = O(T^{-1})$, we can calculate

$$\begin{aligned}\log\det\Big(\sum_{j=k+1}^{T-1} Z_j Z_j'\Big) &= \log\det\Big(\Upsilon_T^{1/2}\Upsilon_T^{-1/2} \sum_{j=k+1}^{T-1} Z_j Z_j' \Upsilon_T^{-1/2}\Upsilon_T^{1/2}\Big) \\ &= \log\det(\Upsilon_T) + O_p(1) = \log(T^5) + O_p(1) \\ &= 5\log(T) + O_p(1)\end{aligned} \tag{A.15}$$

which leads to $C.1 = 5\sigma^2\big(\sum_{j=0}^{h-1} b_j\big)^2 \log(T) + o_p(\log T)$. Thus,

$$\lim_{T\to\infty} \frac{1}{\sigma^2 \log T}\Big(APE_1 - \sum_{i=m_h}^{T-h} \eta_{i,h}^2\Big) = 5\Big(\sum_{j=0}^{h-1} b_j\Big)^2 + g_h(k) \tag{A.16}$$

where the right hand side of (A.16) is the limit of $f_1(c,p,k,h) = f_1(c,p,h) + g_h(k)$ as $c \to$-∞.

We next prove (b). Following similar steps as in the proof of (a) and the proof of Theorem 1 for the restricted case, we can derive

$$\begin{aligned}APE_0 - \sum_{i=m_h}^{T-h} \eta_{i,h}^2 =&\Big(\sum_{j=0}^{h-1} b_j\Big)^2 \sum_{i=m_h}^{T-h} (\widetilde{\beta}_{0,i}^* - c\frac{N_i}{i})^2 \\ &+ \sum_{i=m_h}^{T-h} \Big\{ \boldsymbol{s}_i'(k) M_h(k) \hat{\Gamma}_i^{-1}(k) \frac{1}{i} \sum_{j=k+1}^{i-1} \boldsymbol{s}_j(k) e_{j+1}\Big\}^2 + o_p(\log T) = D.1 + D.2\end{aligned}$$

In view of (A.4), taking the limit $c \to 0$, we have

$$\begin{aligned}\sum_{i=m_h}^{T-h}(\widetilde{\beta}^*_{0,i} - c\frac{N_i}{i})^2 &= \sum_{i=m_h}^{T-h}[\boldsymbol{\iota}_1'(\sum_{t=k+1}^{i}\dot{Z}_t\dot{Z}_t')^{-1}\sum_{t=k+1}^{i}\dot{Z}_t'e_t]^2 \\ &= \sum_{i=m_h}^{T-h}[(\sum_{t=k+1}^{i}\dot{Z}_1^2)^{-1}\sum_{t=k+1}^{i}\dot{Z}_1e_t]^2 + s.o. \\ &= \log\det(\sum_{j=k+1}^{T-1}\dot{Z}_1^2) + o_p(\log T) = \sigma^2\log T + o_p(\log T)\end{aligned}$$

Further, using the same argument as in (A.14), we have $D.2 = g_h(k)\sigma^2\log T + o_p(\log T)$. Thus,

$$\lim_{c\to 0}\lim_{T\to\infty}\frac{1}{\sigma^2\log T}(APE_0 - \sum_{i=m_h}^{T-h}\eta_{i,h}^2) = \big(\sum_{j=0}^{h-1}b_j\big)^2 + g_h(k) \tag{A.17}$$

where the right hand side of (A.17) is the limit of $f_0(c,p,k,h) = f_0(c,p,h) + g_h(k)$ as $c \to 0$ since $\lim_{c\to 0}E(T_{0c}^2) = E[W(1)^2] = 1$. ■

References

Chow, Y. S. (1965). Local convergence of martingales and the law of large numbers. *The Annals of Mathematical Statistics*, 36(2):552–558.

Ing, C.-K. (2003). Multistep prediction in autoregressive processes. *Econometric Theory*, 19(2):254–279.

Ing, C.-K. (2004). Selecting optimal multistep predictors for autoregressive processes of unknown order. *The Annals of Statistics*, 32(2):693–722.

Ing, C.-K., Lin, J.-L., and Yu, S.-H. (2009). Toward optimal multistep forecasts in non-stationary autoregressions. *Bernoulli*, 15(2):402–437.

Phillips, P. C. (2014). On confidence intervals for autoregressive roots and predictive regression. *Econometrica*, 82(3):1177–1195.

Wei, C. (1987). Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. *The Annals of Statistics*, pages 1667–1682.

Supplementary Appendix B: Description of Methods

This Appendix provides a detailed description of the forecasting methods compared in the Monte Carlo analysis presented in Section 6 and the empirical analysis presented in Section 7.

Unrestricted Autoregressive Model (Benchmark). The benchmark forecast is calculated from a standard autoregressive model of order $K+1$ estimated by OLS:

$$y_t = \beta_0^* + \beta_1^* t + \alpha y_{t-1} + \sum_{j=1}^{K} \alpha_j \Delta y_{t-j} + \epsilon_t,$$

Mallows Selection. Hansen (2010a) demonstrates the validity of the Mallows criterion for selecting between the restricted and unrestricted models when $h = 1$. When the number of lags k is known, the criteria for the restricted and unrestricted models are, respectively, given by

$$M_0 = T\widetilde{\sigma}^2 + 2\hat{\sigma}^2(p+k)$$
$$M_1 = T\hat{\sigma}^2 + 2\hat{\sigma}^2(2+p+k)$$

where $\widetilde{\sigma}^2 = T^{-1}\sum_{t=1}^{T}(y_t - \widetilde{\mu}_t)^2$ and $\sigma^2 = T^{-1}\sum_{t=1}^{T}(y_t - \hat{\mu}_t)^2$. The Mallows selection estimator picks the restricted model if $M_0 \leq M_1$ and the unrestricted model otherwise. This is equivalent to picking the unrestricted model when $F_T = T(\frac{\widetilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2}) > 4$. The Mallows selection forecast can then be expressed as $\hat{\mu}_{t+h,M} = \hat{\mu}_{t+h}\mathbf{1}(F_T > 4) + \widetilde{\mu}_{t+h}\mathbf{1}(F_T \leq 4)$. When the number of lags is unknown, the relevant Mallows criteria are obtained as (see Kejriwal and Yu, 2021):

$$M_0(l) = T\widetilde{\sigma}_l^2 + 2\hat{\sigma}_K^2(p+l)$$
$$M_1(l) = T\hat{\sigma}_l^2 + 2\hat{\sigma}_K^2(2+p+l)$$

for $l = 0, 1, ..., K$, where $\hat{\sigma}_j^2 = T^{-1}\sum_{t=1}^{T}(y_t - \hat{\mu}_t(j))^2$, $j = l, K$ and $\widetilde{\sigma}_l^2 = T^{-1}\sum_{t=1}^{T}(y_t - \widetilde{\mu}_t(l))^2$. Then, defining $\widetilde{l} = \arg\min_{l\in S}\{M_0(l)\}$, $\hat{l} = \arg\min_{l\in S}\{M_1(l)\}$, where $S = \{0, 1, ..., K\}$, the Mallows selection forecast is obtained as

$$\breve{\mu}_{t+h,M} = \begin{cases} \hat{\mu}_{t+h}(\hat{l}), & if \quad \min_{l\in S}\{M_1(l)\} < \min_{l\in S}\{M_0(l)\} \\ \widetilde{\mu}_{t+h}(\widetilde{l}), & if \quad \min_{l\in S}\{M_1(l)\} \geq \min_{l\in S}\{M_0(l)\} \end{cases}$$

Mallows Averaging. As an alternative to Mallows selection, Hansen (2010a) develops the Mallows combination forecast that entails taking a weighted average of the unrestricted and restricted forecasts where the weights are chosen by minimizing a Mallows criterion. When the number of lags is known, the criterion is

$$M_w = \sum_{t=1}^{T}(y_t - \hat{\mu}_t(w))^2 + 2\hat{\sigma}^2(2w+p+k) \tag{B.1}$$

with $\hat{\mu}_t(w) = w\hat{\mu}_t + (1-w)\widetilde{\mu}_t$ and $\hat{\sigma}^2 = T^{-1}\sum_{t=1}^{T}(y_t - \hat{\mu}_t)^2$. The Mallows selected weight $\hat{w}$ is derived from minimizing (B.1) over $w \in [0,1]$. The solution is

$$\hat{w} = \begin{cases} 1 - 2/F_T & if \quad F_T > 2 \\ 0 & otherwise \end{cases}$$

The Mallows averaging estimator is then defined as

$$\hat{\mu}_{t+h,M}(\hat{w}) = \hat{w}\hat{\mu}_{t+h} + (1-\hat{w})\widetilde{\mu}_{t+h} = \begin{cases} (1 - \frac{2}{F_T})\hat{\mu}_{t+h} + \frac{2}{F_T}\widetilde{\mu}_{t+h} & if \quad F_T > 2 \\ \widetilde{\mu}_{t+h} & otherwise \end{cases} \tag{B.2}$$

When the number of lags is unknown, Hansen (2010a) considers two alternative Mallows combination forecasts. The first is the so-called partial averaging forecast that averages only over unrestricted forecasts that vary according to the number of first-differenced lags included. With a maximum of K lags, this forecast is given by

$$\hat{\mu}_{t+h,M}(\hat{W}) = \sum_{l=0}^{K} \hat{w}_l \hat{\mu}_{t+h}(l) \tag{B.3}$$

where $\hat{W} = (\hat{w}_0, \hat{w}_1 ..., \hat{w}_K)'$ minimizes the criterion (with $\hat{\mu}_t(W) = \sum_{l=0}^{K} w_l \hat{\mu}_t(l)$),

$$M_P(W) = \sum_{t=1}^{T}(y_t - \hat{\mu}_t(W))^2 + 2\hat{\sigma}_K^2 \left(\sum_{l=0}^{K} [w_l(2 + l + p)] \right)$$

subject to the restrictions $w_j \geq 0$ $(j = 0, 1, ...K)$, $\sum_{j=0}^{K} w_j = 1$. The second combination forecast is the so-called general averaging forecast that averages over the forecasts from all $2(K+1)$ models that include the $(K+1)$ restricted models. This forecast is given by

$$\breve{\mu}_{t+h,M}(\breve{W}) = \sum_{l=0}^{K} \left(\breve{w}_{1l} \hat{\mu}_{t+h}(l) + \breve{w}_{0l} \widetilde{\mu}_{t+h}(l) \right) \tag{B.4}$$

with $\breve{W} = (\breve{w}_{00}, \breve{w}_{01}, ..., \breve{w}_{0K}, \breve{w}_{10}, \breve{w}_{11}, \breve{w}_{12}, ..., \breve{w}_{1K})'$ minimizing the criterion,

$$M_G(W) = \sum_{t=1}^{T}(y_t - \breve{\mu}_t(W))^2 + 2\hat{\sigma}_K^2 \left(\sum_{l=0}^{K} [w_{0l} l + w_{1l}(2 + l)] + p \right)$$

where $\breve{\mu}_t(W) = \sum_{l=0}^{K} \left(w_{0l} \widetilde{\mu}_t(l) + w_{1l} \hat{\mu}_t(l) \right)$ and the weights satisfy: $w_{1l} \geq 0, w_{0l} \geq 0$, $\sum_{l=0}^{K}(w_{0l} + w_{1l}) = 1$. In what follows, we will refer to (B.3) and (B.4) as the MPA (Mallows Partial Averaging) and MGA (Mallows General Averaging) forecasts, respectively.

Leave-h-out Cross Validation Selection. Hansen (2010b) provides theoretical justification for constructing h-step ahead forecasts using leave-h-out cross validation under the assumption

that the data are strictly stationary. For model selection with a known number of lags, let CV_0 and CV_1 denote the cross-validation criteria for the restricted and unrestricted models, respectively. These criteria are computed as

$$CV_0 = \sum_{t=k+1}^{T-h} (y_{t+h} - \widetilde{\mu}_{t+h}^{(t)})^2 \tag{B.5}$$

$$CV_1 = \sum_{t=k+1}^{T-h} (y_{t+h} - \hat{\mu}_{t+h}^{(t)})^2 \tag{B.6}$$

where $\widetilde{\mu}_{t+h}^{(t)}$ and $\hat{\mu}_{t+h}^{(t)}$ are the restricted and unrestricted leave-h-out forecasts, respectively. Specifically, $\widetilde{\mu}_{t+h}^{(t)}$ is obtained using parameter estimates from the restricted model after leaving out the observations $\{t+1, ..., t+h\}$[1]:

$$\Delta y_j = \beta_0^* + \sum_{s=1}^{k} \alpha_k \Delta y_{j-s} + \epsilon_j, \quad j \neq t+1, ..., t+h$$

Similarly, $\hat{\mu}_{t+h}^{(t)}$ is obtained from estimating the unrestricted model after leaving out the observations $\{t+1, ..., t+h\}$:

$$y_j = \beta_0^* + \beta_1^* j + \alpha y_{j-1} + \sum_{s=1}^{k} \alpha_k \Delta y_{j-s} + \epsilon_j, \quad j \neq t+1, ..., t+h$$

Then the cross-validation based forecast is

$$\hat{\mu}_{t+h,CV} = \hat{\mu}_{t+h} \mathbf{1}(CV_0 > CV_1) + \widetilde{\mu}_{t+h} \mathbf{1}(CV_0 \leq CV_1)$$

When the number of lags is unknown, the cross-validation criterion is computed for each of the $2(K+1)$ possible models and the selected forecast is the one that corresponds to the model with the minimum value of this criterion.

Leave-h-out Cross Validation Averaging. When the number of lags is known, the cross validation weights $(\hat{w}, 1-\hat{w})$ are obtained by minimizing the criterion

$$CV_w = \sum_{t=k+1}^{T-h} \left\{ w(y_{t+h} - \hat{\mu}_{t+h}^{(t)}) + (1-w)(y_{t+h} - \widetilde{\mu}_{t+h}^{(t)}) \right\}^2$$

and the resulting forecast is $\hat{\mu}_{t+h,CV}(\hat{w}) = \hat{w}\hat{\mu}_{t+h} + (1-\hat{w})\widetilde{\mu}_{t+h}$. When the number of lags is unknown, the partial combination forecast that only combines the unrestricted forecasts with different lags is obtained as

$$\hat{\mu}_{t+h,CV}(\hat{W}) = \sum_{l=0}^{K} \hat{w}_l \hat{\mu}_{t+h}(l) \tag{B.7}$$

[1] Hansen (2010b) instead leaves out the $2h-1$ observations $\{t-h+1, ..., t, t+1, ..., t+h-1\}$. The difference emanates from the fact that he constructs direct forecasts while our forecasts are constructed iteratively which exploit the autoregressive structure and hence necessitate leaving out only the h observations $\{t+1, ..., t+h\}$.

where $\hat{W} = (\hat{w}_0, \hat{w}_1 ..., \hat{w}_K)'$ minimizes the criterion

$$CV_P(W) = \sum_{t=k+1}^{T-h} \left\{ \sum_{l=0}^{K} w_l (y_{t+h} - \hat{\mu}_{t+h}^{(t)}(l)) \right\}^2 \tag{B.8}$$

subject to the restrictions $w_j \geq 0$ $(j = 0, 1, ...K)$, $\sum_{j=0}^{K} w_j = 1$, and $\hat{\mu}_{t+h}^{(t)}(l)$ is the unrestricted leave-h-out forecast assuming l first-differenced lags. As with weight selection using the Mallows criterion, we also construct a general combination forecast that combines forecasts from the $K+1$ unrestricted models as well as the $K+1$ restricted models. This forecast is given by

$$\breve{\mu}_{t+h,CV}(\breve{W}) = \sum_{l=0}^{K} \left(\breve{w}_{1l} \hat{\mu}_{t+h}(l) + \breve{w}_{0l} \widetilde{\mu}_{t+h}(l) \right) \tag{B.9}$$

with $\breve{W} = (\breve{w}_{01}, \breve{w}_{02}, ..., \breve{w}_{0K}, \breve{w}_{11}, \breve{w}_{12}, ..., \breve{w}_{1K})'$ minimizing the criterion

$$CV_G(W) = \sum_{t=k+1}^{T-h} \left\{ \sum_{l=0}^{K} \left[w_{1l} (y_{t+h} - \hat{\mu}_{t+h}^{(t)}(l)) + w_{0l} (y_{t+h} - \widetilde{\mu}_{t+h}^{(t)}(l)) \right] \right\}^2$$

where $w_{1l} \geq 0, w_{0l} \geq 0, \sum_{l=0}^{K} (w_{0l} + w_{1l}) = 1$, $\hat{\mu}_{t+h}^{(t)}(l)$ is as defined in (B.8) and $\widetilde{\mu}_{t+h}^{(t)}(l)$ is the restricted leave-h-out forecast assuming l first-differenced lags. In what follows, we will refer to (B.7) and (B.9) as the CPA (Cross-Validation Partial Averaging) and CGA (Cross-Validation General Averaging) forecasts, respectively.

APE Selection. With a known number of lags, this forecast is computed from the model that corresponds to the lower APE between the restricted and unrestricted models:

$$\hat{\mu}_{t+h,S} = \hat{\mu}_{t+h} I(APE_0 > APE_1) + \widetilde{\mu}_{t+h} I(APE_0 \leq APE_1)$$
$$APE_0 = \sum_{i=m_h}^{T-h} \left\{ y_{i+h} - \widetilde{\mu}_{i+h} \right\}^2, \quad APE_1 = \sum_{i=m_h}^{T-h} \left\{ y_{i+h} - \hat{\mu}_{i+h} \right\}^2$$

In the unknown lags case, the forecast is computed from the model that minimizes the APE criterion among all $2(K+1)$ possible models, comprising the $K+1$ restricted and $K+1$ unrestricted models.

References

Hansen, B. E. (2010a). Averaging estimators for autoregressions with a near unit root. *Journal of Econometrics*, 158(1):142–155.

Hansen, B. E. (2010b). Multi-step forecast model selection. In *20th Annual Meetings of the Midwest Econometrics Group*.

Kejriwal, M. and Yu, X. (2021). Generalized forecast averaging in autoregressions with a near unit root. *The Econometrics Journal*, 24(1):83–102.

Supplementary Appendix C: Additional Simulation Results

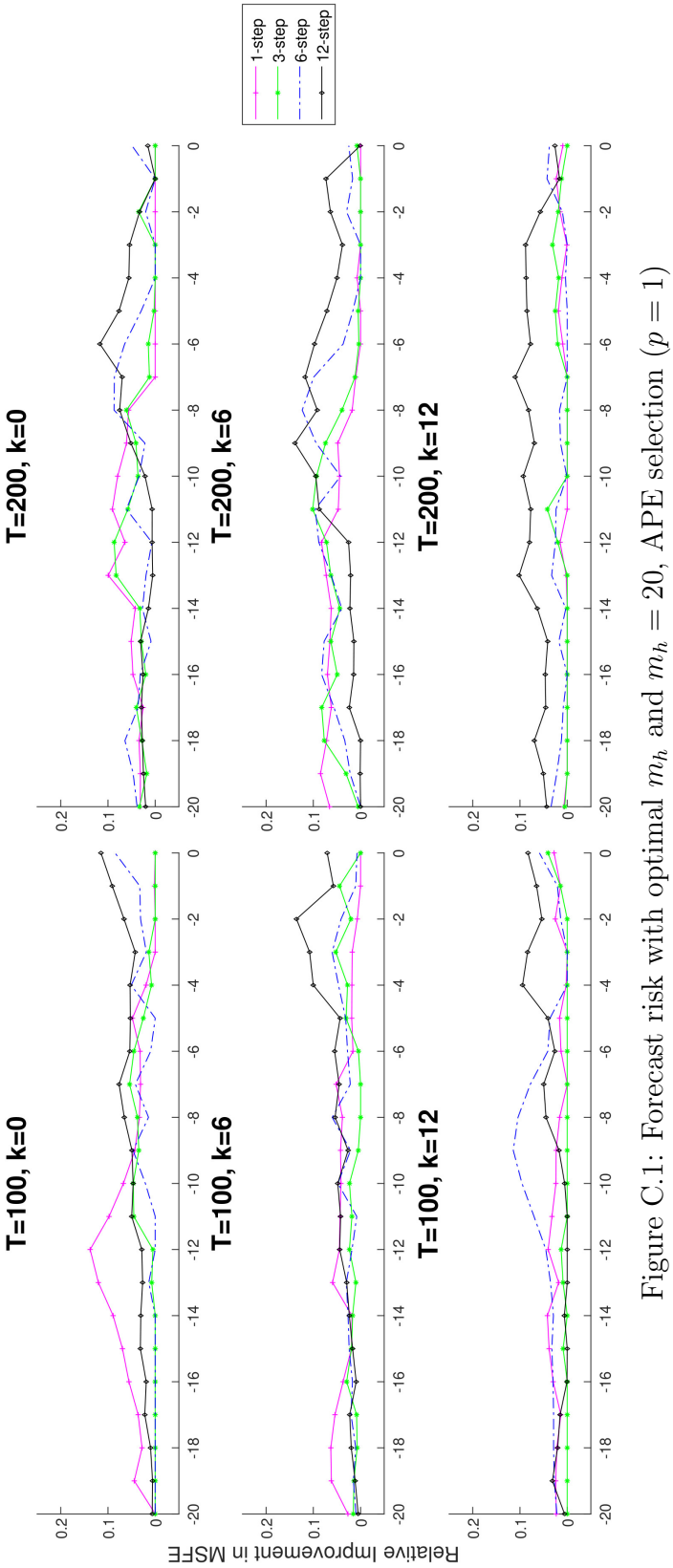

Figure C.1: Forecast risk with optimal m_h and $m_h = 20$, APE selection ($p = 1$)

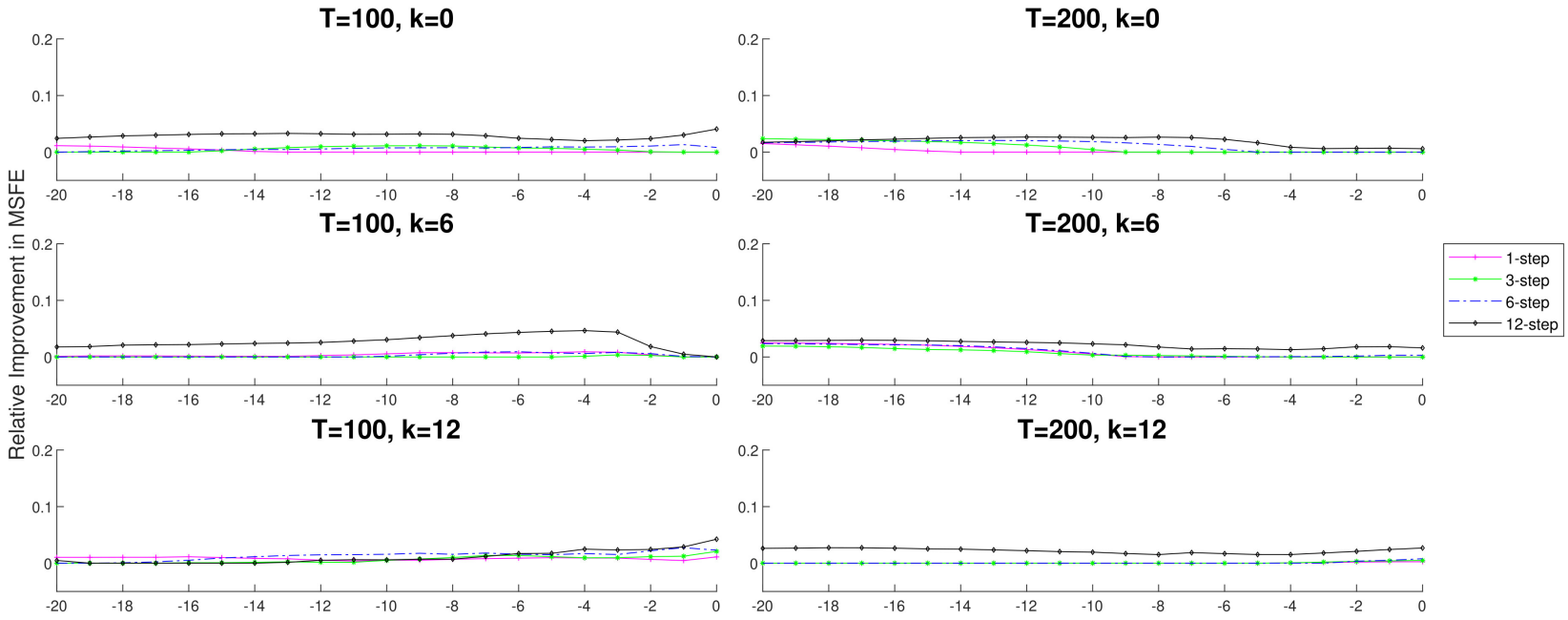

Figure C.2: Forecast risk with optimal m_h and $m_h = 20$, APE average ($p = 1$)