ARTICLE IN PRESS
International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Conditionally optimal weights and forward-looking approaches to combining forecasts[☆]

Christopher G. Gibbs [*], Andrey L. Vasnev

The University of Sydney, New South Wales, Australia

ARTICLE INFO

Keywords:
Forecast combination
Conditionally optimal weights
Forecast combination puzzle
Inflation
Phillips curve

ABSTRACT

In forecasting, there is a tradeoff between in-sample fit and out-of-sample forecast accuracy. Parsimonious model specifications typically outperform richer model specifications. Consequently, information is often withheld from a forecast to prevent over-fitting the data. We show that one way to exploit this information is through forecast combination. Optimal combination weights in this environment minimize the conditional mean squared error that balances the conditional bias and the conditional variance of the combination. The bias-adjusted conditionally optimal forecast weights are time varying and forward looking. Real-time tests of conditionally optimal combinations of model-based forecasts and surveys of professional forecasters show significant gains in forecast accuracy relative to standard benchmarks for inflation and other macroeconomic variables.

1. Introduction

The standard approach to forecast combination is to construct combination weights based on the past performance of the individual forecasts. This backward-looking approach is sensible, but it has also generated a puzzle. Empirically, these strategies are not particularly effective. Instead, it is often found that simple forecast combination strategies, such as equal weights (averaging), produce the most reliably accurate forecasts.[1] We show that a key contributor to the poor performance of optimal forecast combination strategies is that if the underlying forecasts are biased in some way, then optimal weights are misspecified. We prove that when biases exist, the optimal combination weights should minimize the mean squared error conditional on any information that may predict these biases.

To be concrete, assume that we wish to forecast y_{T+1} and have two individual forecasts available, $f_{1,T+1}$ and $f_{2,T+1}$, which can be combined linearly: $f_{c,T+1} = wf_{1,T+1} + (1 - w)f_{2,T+1}$. In the classical framework of Bates and Granger (1969), if the individual forecasts are unbiased,

☆ We would like to thank Graham Elliott, George Evans, Jeremy Piger, Ken Wallis, James Morley, Benjamin Wong, Yinchu Zhu; conference participants of the 32nd Annual International Symposium of Forecasters, 2013 Canadian Economic Associations Meetings, ISF2017 in Cairns, ESEM2017 in Lisbon, the Continuing Education in Macroeconomics Workshop hosted by the University of Tasmania; seminar participants at Massey University, Auckland University of Technology, University of Sydney, Colgate University, QUT, and at the 2016 SERG workshop; and the research staff at Reserve Bank of New Zealand for useful comments on earlier drafts of this paper. The authors are responsible for any remaining errors.

* Correspondence to: School of Economics, The University of Sydney, Australia.

E-mail address: christopher.gibbs@sydney.edu.au (C.G. Gibbs).

1 A prominent empirical example of the forecast combination puzzle for inflation is presented in Stock and Watson (2003). To the best of our knowledge, the first formal reference to the forecast combination puzzle in the literature is Stock and Watson (2004); however, the results have certainly been acknowledged in the literature, at least dating back to Bates and Granger (1969). Surveys and comments on this finding are found in Clemen (1989), Diebold and Lopez (1996), Elliott and Timmermann (2016), Granger (1989), Timmermann (2006), and Wallis (2011).

https://doi.org/10.1016/j.ijforecast.2024.03.002

i.e., the forecast errors $e_{1,T+1} = y_{T+1} - f_{1,T+1}$ and $e_{2,T+1} = y_{T+1} - f_{2,T+1}$ have zero expectation, then the error of the combined forecast $e_{c,T+1} = y_{T+1} - f_{c,T+1} = we_{1,T+1} + (1-w)e_{2,T+1}$ will have zero expectation, and its variance is

$$\mathrm{var}(e_{c,T+1}) = w^2\sigma^2_{e_1} + (1-w)^2\sigma^2_{e_2} + 2w(1-w)\rho_{e_1,e_2}\sigma_{e_1}\sigma_{e_2},$$

where $\sigma^2_{e_1} = \mathrm{var}(e_{1,T+1})$, $\sigma^2_{e_2} = \mathrm{var}(e_{2,T+1})$ and $\rho_{e_1,e_2} = \mathrm{corr}(e_{1,T+1}, e_{2,T+1})$. The variance of the combined forecast is minimized when

$$w^* = \frac{\sigma^2_{e_2} - \rho_{e_1,e_2}\sigma_{e_1}\sigma_{e_2}}{\sigma^2_{e_1} + \sigma^2_{e_2} - 2\rho_{e_1,e_2}\sigma_{e_1}\sigma_{e_2}}. \tag{1}$$

We call such a strategy backward looking because the weights are based on the in-sample moments of the forecast errors.

Now, consider the case in which a forecaster has additional information, I_T, that may be useful for predicting y_{T+1}. However, assume that the forecaster excludes this information from the individual forecasts for the sake of parsimony or because statistical tests have low power to assess its relevance for forecasting given available data. If that information is relevant, then the forecasts' errors can be decomposed as $e_{1,T+1} = b_{1,T} + \xi_{1,T+1}$ and $e_{2,T+1} = b_{2,T} + \xi_{2,T+1}$, where $b_{1,T} = \mathrm{E}(e_{1,T+1}|I_T)$, $b_{2,T} = \mathrm{E}(e_{2,T+1}|I_T)$ and $\mathrm{E}(\xi_{1,T+1}|I_T) = \mathrm{E}(\xi_{2,T+1}|I_T) = 0$. There is no contradiction here with the unbiasedness of the original forecasts because unconditionally $\mathrm{E}(b_{1,T}) = \mathrm{E}(b_{2,T}) = 0$. We stress that the unbiasedness assumption is not critical, and it is relaxed later in the paper. The mean squared error (MSE) conditional on I_t is

$$\mathrm{MSE}(w) = (wb_{1,T} + (1-w)b_{2,T})^2 + w^2\sigma^2_{\xi_1} + (1-w)^2\sigma^2_{\xi_2} + 2w(1-w)\rho_{\xi_1,\xi_2}\sigma_{\xi_1}\sigma_{\xi_2}$$

where $\sigma^2_{\xi_1} = \mathrm{var}(\xi_{1,T+1}|I_T)$, $\sigma^2_{\xi_2} = \mathrm{var}(\xi_{2,T+1}|I_T)$ and $\rho_{\xi_1,\xi_2} = \mathrm{corr}(\xi_{1,T+1}, \xi_{2,T+1}|I_T)$. The conditionally optimal weights that minimize the conditional MSE to simultaneously balance the bias and variance components are

$$w^*(I_T) = \frac{\sigma^2_{\xi_2} + b^2_{2,T} - \rho_{\xi_1,\xi_2}\sigma_{\xi_1}\sigma_{\xi_2} - b_{1,T}b_{2,T}}{\sigma^2_{\xi_1} + b^2_{1,T} + \sigma^2_{\xi_2} + b^2_{2,T} - 2\rho_{\xi_1,\xi_2}\sigma_{\xi_1}\sigma_{\xi_2} - 2b_{1,T}b_{2,T}}. \tag{2}$$

We call this strategy forward looking because it not only relies on the in-sample moments of the conditional forecast errors but also may rely on a conditional out-of-sample forecast of the bias that varies over time.

A special case illustrates the advantages of the conditional approach. If $\rho_{\xi_1,\xi_2} = 0$ and $b_{2,T} = 0$, then the conditionally optimal solution is

$$\frac{\sigma^2_{\xi_2}}{\sigma^2_{\xi_1} + b^2_{1,T} + \sigma^2_{\xi_2}}.$$

In the unconditional solution, the ratio between the variances determines the weight if the correlation is zero (i.e., $w^* = \frac{\sigma^2_{e_2}}{\sigma^2_{e_1}+\sigma^2_{e_2}}$). The conditional bias now plays a similar role. The larger $b^2_{1,T}$ is, the smaller the weight $w^*(I_T)$. If $\sigma_{\xi_1} = \sigma_{\xi_2}$, i.e., if the models showed similar performance in the past, then the weights are not equal as in (1) when $\rho_{e_1,e_2} = 0$ and $\sigma_{e_1} = \sigma_{e_2}$. Instead, less weight is placed on the model that is *expected* to be biased in the forecasting period.[2]

In practice, of course, forecasters do not consider obviously biased forecasts. Forecasts are often selected specifically because they are unbiased. Any residual bias in a forecast is usually difficult either to detect or to remove. Moreover, adding additional predictors to a forecast model to eliminate residual bias tends to increase in-sample fit at the expense of out-of-sample forecast accuracy. We prove that when the forecast bias is difficult to correct, correcting the combination weights instead, while leaving the underlying forecasts uncorrected, achieves the greatest forecast accuracy.

Estimating time-varying optimal weights is also difficult in practice. For example, Claeskens et al. (2016) and Smith and Wallis (2009) show that the estimation noise of the error variance–covariances needed to construct optimal weights is one explanation for the forecast combination puzzle. Conditionally optimal weights offer an advantage in this case. We find that estimates of the conditional bias are much more stable than the estimates of the unconditional or conditional variances of the forecast errors. We show that ignoring the conditional variances and relying solely on the predicted bias to construct weights produces more accurate combined forecasts than using equal weights. We suggest and test a number of strategies based on this idea.

There are several related papers that support our forward-looking approach. Giacomini and Rossi (2009) show that detecting changes in relative forecast accuracy is possible in real time. Timmermann and Zhu (2016) show that a forward-looking approach to model selection is useful in situations with weak predictors due to estimation error. We, however, use the predictions to construct combined forecasts rather than for model selection.[3] Giacomini and White (2006) propose a test of *conditional* predictive ability that Granziera and Sekhposyan (2019) use to establish whether relative forecasting performance is predictable and to construct *heuristic* model averaging strategies that are effective in forecasting. Our paper complements (Giacomini & White, 2006) with the *conditional* combination framework and provides theoretical foundations for the strategies used in Granziera and Sekhposyan (2019). Finally, Montero-Manso et al. (2020)

[2] We see from Eq. (2) that if $b_{1,T} = b_{2,T}$ then the biases cancel each other out, i.e., $w^*(I_T) = w^*$. If biases are strongly positively related, i.e., $b_{1,T} \approx b_{2,T}$, then $w^*(I_T) \approx w^*$. However, if the biases are very different from each other, then the conditionally optimal weights $w^*(I_T)$ are very different from the unconditional optimal weights w^*. In the special case where $\sigma^2_{\xi_1} \approx \sigma^2_{\xi_2}$, similar biases do not help much compared with the equal weight, but if the biases are substantially different, the conditional optimal weight $w^*(I_T)$ is different from equal weight.

[3] Clements and Hendry (1996) and Wallis and Whitley (1991) explore another related strategy known as intercept correction. Intercept correction uses the most recently observed forecast errors to correct the bias of a point forecast by adding the errors to the next forecast to correct the model. Wallis and Whitley (1991) find that intercept correction produces modest improvements over an uncorrected model for forecasts of UK inflation and other macroeconomic variables.

uses forward-looking (or future-based) strategies, which achieved second place in the M4 competition of Makridakis et al. (2018), and Zhang and Zhang (2023) propose a forward-validation criterion for model averaging and establish its asymptotic optimality as well as its superior forecasting performance of the annual excess equity premium over the S&P 500 index.

The basic concept of conditionally optimal combinations is approached by Aiolfi and Timmermann (2006). They note that there is persistence in relative forecast performance among linear and nonlinear time-series models used to forecast a wide range of macroeconomic variables, and that this persistence can be exploited to select forecasts and construct weights. However, their proposed weights use a restricted information set that is backward looking and based on recent historical forecasting performance. We provide a general framework that allows conditioning on any available information, which encompasses Aiolfi and Timmermann's weights and is explicitly forward looking.

We demonstrate the efficacy of conditionally optimal and forward-looking strategies in three real-time out-of-sample forecasting exercises. First, we forecast a measure of United States (US) inflation using combinations of ARMA, VAR, and direct forecasts. We choose US inflation as our base exercise because inflation forecasting is *hard* (Stock & Watson, 2007) and because inflation forecasts that rely on changes in real activity are found to have time-varying forecast accuracy (Stock & Watson, 2009), with inflation varying more with real activity during economic downturns. The time variation in the forecast accuracy of different models of inflation is an example of the type of time-varying bias that should be exploitable by conditionally optimal weights. We confirm this hypothesis and find that conditionally optimal-weights forecasts consistently outperform equal-weights forecasts in real-time forecasting comparisons.

We next test conditionally optimal weights for combining survey responses of expert forecasters from the European Central Bank's (ECB) Survey of Professional Forecasters (SPF). We combine the individual forecasts to predict the inflation rate implied by the Harmonized Index of Consumer Prices. We compare our forward-looking approach to the backward-looking bias-correction approach proposed by Issler and Lima (2009) for optimally combining panels of biased forecasts. They propose a nonparametric strategy to estimate fixed forecast biases to optimally bias-correct combined forecasts. We find that this approach uniformly produces less accurate forecasts than our forward-looking approach. We obtain this result because the biases we observe are time varying and difficult to estimate in real time, which, as our theory predicts, is an environment that is better suited to conditionally optimal weights.

To assess the robustness of our findings, we compare conditionally optimal strategies to unconditionally optimal and bias-correction strategies in hundreds of different real-time out-of-sample exercises, which we create by varying the individual forecasts included in the combined forecasts. By varying the forecasts that we combine, we can generate a distribution of out-of-sample forecasting results that act as a reality check in the spirit of White (2000). We find that on average, forward-looking strategies are superior to all considered challengers.

For the final forecasting exercise, we demonstrate the tradeoff between bias correction and forecast accuracy that we show exists in theory. We prove in Section 2 that whether forecasts should be bias corrected or conditionally combined depends on the quality of the information one has to predict the bias. When the available signal in the information set for the bias is small relative to its noise, constructing conditionally optimal weights is a superior forecasting strategy, and vice versa. We demonstrate this relationship empirically by real-time forecasting 25 macroeconomic time series, including real GDP growth, the GDP deflator measure of inflation, interest rates, consumption growth, and investment growth, for the US, Canada, the United Kingdom, Australia, and New Zealand. The exercise reveals a clear positive relationship between underlying forecast bias and the accuracy of a bias-corrected combined forecast. When the bias is small, bias correction is not an effective strategy. When the bias is large, it is effective. In contrast, the accuracy of a conditionally optimal combined forecast is unrelated to the size of the biases of the forecasts. It provides consistent improvements in accuracy relative to equal-weights forecasts for all data types and countries considered. Therefore, we find that conditionally optimal weights and forward-looking approaches work both in theory and in practice.

The remainder of this paper is organized as follows. In Section 2, we provide the theory behind conditionally optimal weights. Section 3 reports the results of a Monte Carlo experiment. Section 4 provides some practical refinements for conditionally optimal weights based on the properties observed in the Monte Carlo experiment. Section 5 reports the results of a real-time out-of-sample forecasting exercise for US inflation. Section 6 reports the results for real-time forecasting of EU harmonized inflation using the ECB Survey of Professional Forecasters. Section 7 reports the results of real-time forecasting using international macroeconomic data. Section 8 concludes the paper.

2. Conditionally optimal weights

Assume that we wish to forecast $y_{T+h} \in \mathbb{R}$ and consider the vector of h-step-ahead forecasts $\boldsymbol{f}_{T+h} = (f_{1,T+h}, f_{2,T+h}, \dots, f_{n,T+h})' \in \mathbb{R}^k$, where I_T is the information set available at time T. Following Giacomini and White (2006), we assume that y_t may follow a complex process marked by measurement issues, structural changes, and nonstationarity induced by distribution changes. Following Aiolfi and Timmermann (2006), we map all forecasts to the real number line and limit our analysis to linear combinations with weights $\boldsymbol{w} = (w_1, w_2, \dots, w_n)'$ to produce the combined forecast $f_{c,T+h} = \boldsymbol{w}'\boldsymbol{f}_{T+h}$. We denote the vector of forecasting errors as $\boldsymbol{e}_{T+h} = y_{T+h}\boldsymbol{\iota} - \boldsymbol{f}_{T+h}$, where $\boldsymbol{\iota}$ is a vector of ones, and the error of the combined forecast is

$$e_{c,T+h} = y_{T+h} - f_{c,T+h} = \boldsymbol{w}'\boldsymbol{e}_{T+h}.$$

Given I_T and assuming that the loss function $L(\cdot)$ depends only on $e_{c,T+h}$, the conditionally optimal combination weights, $\boldsymbol{w}^*(I_T)$, solve the problem[4]

$$\boldsymbol{w}^*(I_T) = \arg\min_{\boldsymbol{w}} \mathrm{E}[L(e_{c,T+h})|I_T]. \quad (3)$$

Under mean squared error (MSE) loss, $L(e) = e^2$, only the first two conditional moments influence the optimal weights, and the optimization problem can be solved explicitly. We also assume that the forecasts are unbiased, $\mathrm{E}(\boldsymbol{e}_{T+h}) = \mathbf{0}$; thus we solve the optimization problem (3) subject to the restriction that the weights sum up to one, $\boldsymbol{w}'\boldsymbol{\iota} = 1$. The existence of the first two moments of the errors, $\boldsymbol{e}_{T+h}$, is sufficient for the results derived in this section to be valid under our general data assumptions.

We assume that it is possible to decompose the original forecast errors, $\boldsymbol{e}_{T+h}$, as a sum of two parts such that $\boldsymbol{e}_{T+h} = \boldsymbol{b}_T + \boldsymbol{\xi}_{T+h}$, $\boldsymbol{b}_T = \mathrm{E}(\boldsymbol{e}_{T+h}|I_T)$, and $\mathrm{E}(\boldsymbol{\xi}_{T+h}|I_T) = 0$. One simple example where this decomposition arises naturally is when an autocorrelation is present in the forecasting errors, as in the empirical example in Section 6.[5] In this case, the information set I_T consists of previous forecasting errors, and the conditional bias follows a first-order autocorrelation model with $e_{i,T+1} = \phi_i e_{i,T} + \xi_{i,T+1}$ (for $h = 1$) and $b_{i,T} = \phi_i e_{i,T}$. The remainder is $\xi_{i,T+1} = e_{i,T+1} - \phi_i e_{i,T}$.

Given the decomposition $\boldsymbol{e}_{T+h} = \boldsymbol{b}_T + \boldsymbol{\xi}_{T+h}$, we can then derive the MSE

$$\mathrm{MSE}(\boldsymbol{w}) = (\boldsymbol{w}'\boldsymbol{b}_T)^2 + \boldsymbol{w}'\Sigma_\xi \boldsymbol{w} = \boldsymbol{w}'(\Sigma_\xi + \boldsymbol{b}_T\boldsymbol{b}_T')\boldsymbol{w},$$

with $\Sigma_\xi = \mathrm{var}(\boldsymbol{\xi}_{T+h}|I_T) = \mathrm{E}(\boldsymbol{\xi}_{T+h}\boldsymbol{\xi}_{T+h}'|I_T)$. The MSE is minimized by *the conditionally optimal weights*

$$\boldsymbol{w}^*(I_T) = \frac{[\Sigma_\xi + \boldsymbol{b}_T\boldsymbol{b}_T']^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'[\Sigma_\xi + \boldsymbol{b}_T\boldsymbol{b}_T']^{-1}\boldsymbol{\iota}}, \quad (4)$$

where $\boldsymbol{\iota}$ is a vector of ones. The minimum MSE that is achieved by the conditionally optimal weights (4) is

$$\mathrm{MSE}(\boldsymbol{w}^*(I_T)) = \frac{1}{\boldsymbol{\iota}'[\Sigma_\xi + \boldsymbol{b}_T\boldsymbol{b}_T']^{-1}\boldsymbol{\iota}}. \quad (5)$$

Naturally, $\boldsymbol{b}_T$, Σ_ξ, $\mathrm{MSE}(\boldsymbol{w})$, and the optimal solution depend on I_T. Without loss of generality, we assume Σ_ξ is constant for the remainder of this section, and to keep the notation simple we note dependency on I_T explicitly only for the optimal solution $\boldsymbol{w}^*(I_T)$. This highlights the fact that the conditionally optimal weights $\boldsymbol{w}^*(I_T)$ are time varying because they depend on the information available at time T.

For comparison, the well-established classical results (see Elliott, 2011) for the *unconditionally optimal weights*

$$\boldsymbol{w}^* = \frac{\Sigma_e^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\Sigma_e^{-1}\boldsymbol{\iota}} \quad (6)$$

are based on the unconditional variance of the errors $\Sigma_e = \mathrm{var}(\boldsymbol{e}_{T+1}) = \Sigma_\xi + \mathrm{E}(\boldsymbol{b}_T\boldsymbol{b}_T')$, and the minimum of $\mathrm{var}(e_{c,T+h}) = w'\Sigma_e w$ at $\boldsymbol{w}^*$ is

$$\frac{1}{\boldsymbol{\iota}'\Sigma_e^{-1}\boldsymbol{\iota}}. \quad (7)$$

The following central result formalizes the intuitive idea that using more information allows us to construct a better combined forecast.

Theorem 1. *Given that the first and the second conditional and unconditional moments exist, the following inequalities hold:*

(a) for the conditional and unconditional MSE, $\mathrm{E}(\mathrm{MSE}(\boldsymbol{w}^*(I_T))) \leq \mathrm{MSE}(w^*)$, *or equivalently* $\mathrm{E}\big[\min_{\boldsymbol{w}} \boldsymbol{w}'\,\mathrm{E}(\boldsymbol{e}_{T+h}\boldsymbol{e}_{T+h}'|I_T)\boldsymbol{w}\big] \leq \min_{\boldsymbol{w}} \boldsymbol{w}'\,\mathrm{E}(\boldsymbol{e}_{T+h}\boldsymbol{e}_{T+h}')\boldsymbol{w}$, *or equivalently*

$$\mathrm{E}\left(\frac{1}{\boldsymbol{\iota}'[\Sigma_\xi + \boldsymbol{b}_T\boldsymbol{b}_T']^{-1}\boldsymbol{\iota}}\right) \leq \frac{1}{\boldsymbol{\iota}'\Sigma_e^{-1}\boldsymbol{\iota}};$$

(b) for the conditional MSE when two information sets $J_T \subset I_T$ *are available,* $\mathrm{E}(\mathrm{MSE}(\boldsymbol{w}^*(I_T))|J_T) \leq \mathrm{MSE}(\boldsymbol{w}^*(J_T))$, *or equivalently* $\mathrm{E}\big[\min_{\boldsymbol{w}} \boldsymbol{w}'\,\mathrm{E}(\boldsymbol{e}_{T+h}\boldsymbol{e}_{T+h}'|I_T)\boldsymbol{w}\big|J_T\big] \leq \min_{\boldsymbol{w}} \boldsymbol{w}'\,\mathrm{E}(\boldsymbol{e}_{T+h}\boldsymbol{e}_{T+h}'|J_T)\boldsymbol{w}$, *or equivalently*

$$\mathrm{E}\left[\frac{1}{\boldsymbol{\iota}'\left[\mathrm{E}(\boldsymbol{e}_{T+h}\boldsymbol{e}_{T+h}'|I_T)\right]^{-1}\boldsymbol{\iota}}\middle|J_T\right] \leq \frac{1}{\boldsymbol{\iota}'\left[\mathrm{E}(\boldsymbol{e}_{T+h}\boldsymbol{e}_{T+h}'|J_T)\right]^{-1}\boldsymbol{\iota}};$$

(c) for a convex loss function $L(\cdot)$ *and* $J_T \subset I_T$, *if the conditional expectations and the solutions of the minimization problems exist, then* $\mathrm{E}\big[\min_{\boldsymbol{w}} \mathrm{E}[L(e_{c,T+h})|I_T]\big|J_T\big] \leq \min_{\boldsymbol{w}} \mathrm{E}[L(e_{c,T+h})|J_T]$.

Proof. See online Appendix A1.1

There are several observations that help us understand the new conditionally optimal weights. First, using unconditional weights $\boldsymbol{w}^*$ is equivalent to using *no information* to predict the errors, i.e., $\boldsymbol{w}^* = \boldsymbol{w}^*(\emptyset)$. Second, if $\boldsymbol{b}_T$ is proportional to $\boldsymbol{\iota}$, i.e., the predictable parts are the same for all forecasts, then by applying the Sherman–Morrison formula it is possible to show that $\boldsymbol{b}_T$ does not play a role in the conditionally optimal weight, i.e., $\boldsymbol{w}^*(I_T) = \boldsymbol{w}^\dagger = \frac{\Sigma_\xi^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\Sigma_\xi^{-1}\boldsymbol{\iota}}$. Third, without predictability, i.e., if $\boldsymbol{b}_T = \mathbf{0}$, we have $\boldsymbol{w}^*(I_T) = \boldsymbol{w}^*$ because $\Sigma_\xi = \Sigma_e$ in this case. The presence of bias increases the MSE, i.e., $\frac{1}{\boldsymbol{\iota}'\Sigma_\xi^{-1}\boldsymbol{\iota}} < \frac{1}{\boldsymbol{\iota}'[\Sigma_\xi + \boldsymbol{b}_T\boldsymbol{b}_T']^{-1}\boldsymbol{\iota}}$. Next, if $\boldsymbol{w}^*$ is used rather than $\boldsymbol{w}^*(I_T)$ when $\boldsymbol{b}_T \neq \mathbf{0}$, then the minimum of the MSE given by (5) is not achieved. That is,

$$\mathrm{MSE}(\boldsymbol{w}^*(I_T)) = \frac{1}{\boldsymbol{\iota}'[\Sigma_\xi + \boldsymbol{b}_T\boldsymbol{b}_T']^{-1}\boldsymbol{\iota}} < \boldsymbol{w}^{*'}[\Sigma_\xi + \boldsymbol{b}_T\boldsymbol{b}_T']\boldsymbol{w}^*$$
$$= \frac{\boldsymbol{\iota}'\Sigma_e^{-1}[\Sigma_\xi + \boldsymbol{b}_T\boldsymbol{b}_T']\Sigma_e^{-1}\boldsymbol{\iota}}{[\boldsymbol{\iota}'\Sigma_e^{-1}\boldsymbol{\iota}]^2}.$$

[4] For the optimization problem to be well defined, we need the existence of the conditional moment $\mathrm{E}[L(e_{c,T+h})|I_T]$ and well-behaved $L(\cdot)$ for the minimum to exist; see Elliott and Timmermann (2004) for general loss functions and forecast error distributions. Using the loss function of Lima and Meng (2017), it is possible to extend our approach to combinations of quantile forecasts.

[5] See also Figure A9 in the online appendix. Another example is non-nested models, as in Timmermann and Zhu (2016). Assume the data generating process $y_{t+1} = (\alpha_1, \alpha_2)'X_t + (X_t'\gamma)\eta_{t+1}$, where $X_t = (x_{1,t}, x_{2,t})'$, and two forecasts $f_{1,t+1} = \alpha_1 x_{1,t}$ and $f_{2,t+1} = \alpha_2 x_{2,t}$ are available. We have the following decomposition of the forecasting errors: $e_{1,t+1} = \alpha_2 x_{2,t} + (X_t'\gamma)\eta_{t+1} = b_{1,t} + \xi_{1,t+1}$ and $e_{2,t+1} = \alpha_1 x_{1,t} + (X_t'\gamma)\eta_{t+1} = b_{2,t} + \xi_{2,t+1}$ with $b_{1,t} = \alpha_2 x_{2,t}$, $\xi_{1,t+1} = (X_t'\gamma)\eta_{t+1}$, $b_{2,t} = \alpha_1 x_{1,t}$, $\xi_{2,t+1} = (X_t'\gamma)\eta_{t+1}$.

If the bias is ignored, i.e., $\boldsymbol{w}^{\dagger}=\frac{\Sigma_{\xi}^{-1}\iota}{\iota'\Sigma_{\xi}^{-1}\iota}$ is used rather than $\boldsymbol{w}^{*}(I_T)$ when $\boldsymbol{b}_T\neq\boldsymbol{0}$, then the minimum of the MSE given by (5) is not achieved. That is,

$$\mathrm{MSE}(\boldsymbol{w}^{*}(I_T))=\frac{1}{\iota'[\Sigma_{\xi}+\boldsymbol{b}_T\boldsymbol{b}_T']^{-1}\iota}<\boldsymbol{w}^{\dagger'}[\Sigma_{\xi}+\boldsymbol{b}_T\boldsymbol{b}_T']\boldsymbol{w}^{\dagger}=\frac{1}{\iota'\Sigma_{\xi}^{-1}\iota}+\left[\frac{\iota'\Sigma_{\xi}^{-1}\boldsymbol{b}_T}{\iota'\Sigma_{\xi}^{-1}\iota}\right]^2.$$

Or using the Cauchy–Schwarz inequality, $(\iota'\Sigma_{\xi}^{-1}\boldsymbol{b}_T)^2\leq(\iota'\Sigma_{\xi}^{-1}\iota)(\boldsymbol{b}_T'\Sigma_{\xi}^{-1}\boldsymbol{b}_T)$, we have another upper bound MSE $(\boldsymbol{w}^{*}(I_T))<\frac{1}{\iota'\Sigma_{\xi}^{-1}\iota}\left[1+\boldsymbol{b}_T'\Sigma_{\xi}^{-1}\boldsymbol{b}_T\right]$. Finally, if h increases, then it is natural to expect that the predictable part, $\boldsymbol{b}_T=\mathrm{E}(\boldsymbol{e}_{T+h}|I_T)$, decreases (see Breitung & Knüppel, 2021). In this situation, the conditionally optimal weights converge to the unconditionally optimal weights as $h\to\infty$.

When the conditional forecast bias is persistent, i.e., $I_T=\{\boldsymbol{e}_T\}$, $e_{i,T+1}=\phi e_{i,T}+\xi_{i,T+1}$ (for $h=1$), and the forecasts are equally precise after the conditional bias is taken into account, $\Sigma_{\xi}=I$, then by applying the Sherman–Morrison formula to the conditionally optimal weights (4), we have

$$\boldsymbol{w}^{*}_{\mathrm{AR1}}(I_T)=\frac{\iota-\phi^2/\delta\times(\iota'\boldsymbol{e}_T)\times\boldsymbol{e}_T}{n-\phi^2/\delta\times(\iota'\boldsymbol{e}_T)^2}$$

with $\delta=1+\phi^2(\boldsymbol{e}_T'\boldsymbol{e}_T)$. If the average forecast was unbiased in the previous period, $(\iota'\boldsymbol{e}_T)=0$, then there is no correction necessary in the next period, and the optimal weights are just equal weights, i.e., $\boldsymbol{w}^{*}_{\mathrm{AR1}}(I_T)=\iota/n$. If the average forecast was positively biased in the previous period, $(\iota'\boldsymbol{e}_T)>0$, but a particular forecast i_0 was unbiased, $e_{i_0,T}=0$, then this forecast does not need correction except for normalization, i.e., $w^{*}_{\mathrm{AR1},i_0}(I_T)=1/(n-\phi^2/\delta\times(\iota'\boldsymbol{e}_T)^2)$. However, the forecasts with positive biases in the previous period, $e_{j,T}>0$, are downgraded with smaller weights $w^{*}_{\mathrm{AR1},j}(I_T)<w^{*}_{\mathrm{AR1},i_0}(I_T)$, and the biggest adjustment is applied to the forecast with the largest error. At the same time, the forecasts with the negative biases in the previous period, i.e., $e_{k,T}<0$, receive larger weights $w^{*}_{\mathrm{AR1},k}(I_T)>w^{*}_{\mathrm{AR1},i_0}(I_T)$. Symmetrical adjustments will be observed for the case when $(\iota'\boldsymbol{e}_T)<0$.

In addition to the simplifications above, the conditionally optimal framework allows many extensions and generalizations. Appendix A1.3 gives examples where an ARMA model is used to capture autocorrelation in forecasting errors, or if the forecasts have an unconditional bias in addition to the conditional bias, or if other loss functions and forecast error distributions considered in Elliott and Timmermann (2004) are used instead of the MSE.

2.1. Bias correction approach

We now investigate the bias correction approach. In theory, this is a superior forecasting strategy if the bias can be precisely estimated.[6] To see this, note that perfectly bias-corrected forecasts $\widetilde{\boldsymbol{f}}_{T+h}=\boldsymbol{f}_{T+h}+\boldsymbol{b}_T$ have forecast errors $\widetilde{\boldsymbol{e}}_{T+h}=y_{T+h}\iota-\widetilde{\boldsymbol{f}}_{T+h}=\boldsymbol{e}_{T+h}-\boldsymbol{b}_T=\boldsymbol{\xi}_{T+h}$, which are unbiased. Therefore, $\boldsymbol{w}^{\dagger}=\frac{\Sigma_{\xi}^{-1}\iota}{\iota'\Sigma_{\xi}^{-1}\iota}$ is optimal, and it achieves the MSE $=\frac{1}{\iota'\Sigma_{\xi}^{-1}\iota}$, which is smaller than the MSE achieved by the conditionally optimal solution given by (5). However, this conclusion follows only when $\boldsymbol{b}_T$ is available. If only an estimate of $\widehat{\boldsymbol{b}}_T$ is available, then the relationship becomes more complicated.

Assume that the estimated bias is given by $\widehat{\boldsymbol{b}}_T=\boldsymbol{b}_T+\boldsymbol{\eta}_T$, where $\boldsymbol{\eta}_T$ is the estimation or measurement error. For example, a relevant variable for forecasting many macroeconomic quantities is the output gap. However, as shown by Orphanides and van Norden (2005), the measurement error of the output gap is so large in real time that it often provides little value to a forecaster. The adjusted forecasts $\breve{\boldsymbol{f}}_{T+h}=\boldsymbol{f}_{T+h}+\widehat{\boldsymbol{b}}_T$ have forecasting errors $\breve{\boldsymbol{e}}_{T+h}=y_{T+h}\iota-\breve{\boldsymbol{f}}_{T+h}=\boldsymbol{e}_{T+h}-\widehat{\boldsymbol{b}}_T=-\boldsymbol{\eta}_T+\boldsymbol{\xi}_{T+h}$. Since $\boldsymbol{\eta}_T$ and $\widehat{\boldsymbol{b}}_T$ are related, $\breve{\boldsymbol{e}}_{T+h}$ will still have some predictable part conditional on $\widehat{\boldsymbol{b}}_T$. Ignoring the predictable information and using the weights $\boldsymbol{w}^{\dagger\dagger}=\frac{[\Sigma_{\xi}+\Sigma_{\eta}]^{-1}\iota}{\iota'[\Sigma_{\xi}+\Sigma_{\eta}]^{-1}\iota}$ is suboptimal and produces

$$\mathrm{MSE}(\boldsymbol{w}^{\dagger\dagger})=\frac{\iota'[\Sigma_{\xi}+\Sigma_{\eta}]^{-1}\Omega_1[\Sigma_{\xi}+\Sigma_{\eta}]^{-1}\iota}{(\iota'[\Sigma_{\xi}+\Sigma_{\eta}]^{-1}\iota)^2}$$

(where $\Omega_1=\Sigma_{\xi}+\mathrm{var}(\boldsymbol{\eta}_T|\widehat{\boldsymbol{b}}_T)+\mathrm{E}(\boldsymbol{\eta}_T|\widehat{\boldsymbol{b}}_T)\,\mathrm{E}(\boldsymbol{\eta}_T'|\widehat{\boldsymbol{b}}_T)$), while the conditional solution $\boldsymbol{w}(\widehat{\boldsymbol{b}}_T)=\frac{[\Sigma_{\xi}+\Sigma_{\eta}+\widehat{\boldsymbol{b}}_T\widehat{\boldsymbol{b}}_T']^{-1}\iota}{\iota'[\Sigma_{\xi}+\Sigma_{\eta}+\widehat{\boldsymbol{b}}_T\widehat{\boldsymbol{b}}_T']^{-1}\iota}$ is applied to the original forecasts and achieves $\mathrm{MSE}(\boldsymbol{w}(\widehat{\boldsymbol{b}}_T))=$

$$\frac{\iota'[\Sigma_{\xi}+\Sigma_{\eta}+\widehat{\boldsymbol{b}}_T\widehat{\boldsymbol{b}}_T']^{-1}\Omega_2[\Sigma_{\xi}+\Sigma_{\eta}+\widehat{\boldsymbol{b}}_T\widehat{\boldsymbol{b}}_T']^{-1}\iota}{(\iota'[\Sigma_{\xi}+\Sigma_{\eta}+\widehat{\boldsymbol{b}}_T\widehat{\boldsymbol{b}}_T']^{-1}\iota)^2}$$

(where $\Omega_2=\Sigma_{\xi}+\mathrm{var}(\boldsymbol{b}_T|\widehat{\boldsymbol{b}}_T)+\mathrm{E}(\boldsymbol{b}_T|\widehat{\boldsymbol{b}}_T)\,\mathrm{E}(\boldsymbol{b}_T'|\widehat{\boldsymbol{b}}_T)$). If we vary $\boldsymbol{\eta}_T$, the $\mathrm{MSE}(\boldsymbol{w}^{\dagger\dagger})$ is unbounded while the $\mathrm{MSE}(\boldsymbol{w}(\widehat{\boldsymbol{b}}_T))$ is bounded.

Theorem 2. *If $\boldsymbol{b}_T$ and $\boldsymbol{\eta}_T$ are elliptically distributed (e.g., have a normal or t-distribution), $\boldsymbol{\eta}_T=\gamma\boldsymbol{\eta}_0$, and $\gamma\to\infty$, then $\mathrm{MSE}(\boldsymbol{w}^{\dagger\dagger})\xrightarrow{p}\infty$ and*

$$\mathrm{MSE}(\boldsymbol{w}(\widehat{\boldsymbol{b}}_T))\xrightarrow{p}\frac{\iota'[\Sigma_{\eta_0}+\boldsymbol{\eta}_0\boldsymbol{\eta}_0']^{-1}\left[\Sigma_{\xi}+\Sigma_{b}\right][\Sigma_{\eta_0}+\boldsymbol{\eta}_0\boldsymbol{\eta}_0']^{-1}\iota}{(\iota'[\Sigma_{\eta_0}+\boldsymbol{\eta}_0\boldsymbol{\eta}_0']^{-1}\iota)^2}.$$

Proof. See online Appendix A1.1.

In practice, forecasts with large estimation or measurement issues would be discarded, and forecasts with perfectly correctable biases would be corrected. Theorem 2, though, shows that when a forecaster is uncertain

[6] The panel-data approach, in particular, delivered several important findings in this area. Issler and Lima (2009) investigate bias-corrected average forecasts when the elements of $\boldsymbol{b}_T$ are independent of time and drawn from identical (but perhaps dependent) distributions with the mean B. In this case, the collective bias B can be estimated consistently when $n\to\infty$. The bias-corrected average forecast is asymptotically optimal in this situation. Davies and Lahiri (1995) use the generalized method of moments (GMM) to efficiently estimate individual biases and standard errors. Davies (2006) expands the framework by including the forecast horizon as an additional panel dimension and conducts GMM tests for forecaster biases in the expanded framework. Gaglianone and Issler (2015) derive sequential asymptotic results for 'big data' applications.

about these quantities, conditionally optimal weights are a better choice. Noisy correction may significantly affect the forecast performance of the individual forecasts. Combining the corrected forecasts can somewhat offset the effect of noise, but ultimately, the forecast accuracy of the combined forecast is driven by the underlying forecasts that are combined. Since conditionally optimal weights leave the underlying forecasts unchanged, the combined forecast is less affected by this noise.

There are two special cases that illustrate this point. First, if $\boldsymbol{b}_T = 0$, so that only $\widehat{\boldsymbol{b}}_T = \boldsymbol{\eta}_T$ is observed, then

$$\text{MSE}(\boldsymbol{w}^{\dagger\dagger}) = \frac{\boldsymbol{\iota}'[\Sigma_\xi + \Sigma_\eta]^{-1}\left[\Sigma_\xi + \boldsymbol{\eta}_T\boldsymbol{\eta}_T'\right][\Sigma_\xi + \Sigma_\eta]^{-1}\boldsymbol{\iota}}{(\boldsymbol{\iota}'[\Sigma_\xi + \Sigma_\eta]^{-1}\boldsymbol{\iota})^2},$$

which can become arbitrarily large when $\boldsymbol{\eta}_T$ is inflated. In contrast, the conditionally optimal solution,

$$\text{MSE}(\boldsymbol{w}(\widehat{\boldsymbol{b}}_T)) = \frac{\boldsymbol{\iota}'[\Sigma_\xi + \Sigma_\eta + \boldsymbol{\eta}_T\boldsymbol{\eta}_T']^{-1}\ \Sigma_\xi\ [\Sigma_\xi + \Sigma_\eta + \boldsymbol{\eta}_T\boldsymbol{\eta}_T']^{-1}\boldsymbol{\iota}}{(\boldsymbol{\iota}'[\Sigma_\xi + \Sigma_\eta + \boldsymbol{\eta}_T\boldsymbol{\eta}_T']^{-1}\boldsymbol{\iota})^2},$$

cannot become arbitrarily large because $\boldsymbol{\eta}_T$ balances the numerator and the denominator.

Second, the unconditional MSE of the combined forecast is $\boldsymbol{w}'\,\text{var}(\boldsymbol{e}_{T+h})\boldsymbol{w}$. Assuming that there is no unconditional bias, i.e., $\text{E}(\boldsymbol{e}_{T+h}) = 0$, and that there is no correlation between $\boldsymbol{b}_T$ and $\boldsymbol{\xi}_{T+h}$, the optimal weight is $\boldsymbol{w}^* = \frac{[\Sigma_\xi + \text{E}(\boldsymbol{b}_T\boldsymbol{b}_T')]^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'[\Sigma_\xi + \text{E}(\boldsymbol{b}_T\boldsymbol{b}_T')]^{-1}\boldsymbol{\iota}}$ and the corresponding MSE is $\text{MSE}(\boldsymbol{w}^*) = \frac{1}{\boldsymbol{\iota}'[\Sigma_\xi + \text{E}(\boldsymbol{b}_T\boldsymbol{b}_T')]^{-1}\boldsymbol{\iota}}$.

If $\boldsymbol{b}_T$ is available, then the bias correction approach will give us $\widetilde{\boldsymbol{e}}_{T+h} = \boldsymbol{\xi}_{T+h}$, the optimal weight $\boldsymbol{w}^\dagger = \frac{\Sigma_\xi^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\Sigma_\xi^{-1}\boldsymbol{\iota}}$, and the corresponding $\text{MSE}(\boldsymbol{w}^\dagger) = \frac{1}{\boldsymbol{\iota}'\Sigma_\xi^{-1}\boldsymbol{\iota}}$. Clearly, the bias correction is preferable in this case, as $\text{MSE}(\boldsymbol{w}^\dagger) < \text{MSE}(\boldsymbol{w}^*)$.

However, if only $\widehat{\boldsymbol{b}}_T$ is available, then ignoring it and using $\boldsymbol{e}_{T+h}$ will still produce the same $\boldsymbol{w}^*$ and the corresponding $\text{MSE}(\boldsymbol{w}^*)$, but the bias correction will move us to $\breve{\boldsymbol{e}}_{T+h} = -\boldsymbol{\eta}_T + \boldsymbol{\xi}_{T+h}$, the optimal weight $\boldsymbol{w}^{\dagger\dagger} = \frac{[\Sigma_\xi + \Sigma_\eta]^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'[\Sigma_\xi + \Sigma_\eta]^{-1}\boldsymbol{\iota}}$, and $\text{MSE}(\boldsymbol{w}^{\dagger\dagger}) = \frac{1}{\boldsymbol{\iota}'[\Sigma_\xi + \Sigma_\eta]^{-1}\boldsymbol{\iota}}$.

Theorem 3. *In the unconditional framework,* $\text{MSE}(\boldsymbol{w}^{\dagger\dagger}) < \text{MSE}(\boldsymbol{w}^*)$ *if and only if* $\Sigma_\eta < \text{E}(\boldsymbol{b}_T\boldsymbol{b}_T')$.

Proof. See online Appendix A1.1.

In other words, the bias correction is a preferable option if the estimation noise $\boldsymbol{\eta}_T$ has a relatively small variance. However, if the noise dominates the signal, then using the original forecast will produce a better outcome.

Our empirical application shows that the bias correction strategy is not superior to the conditionally optimal weights in most cases. This is because bias correction inflates the forecast error variances to $\Sigma_\xi + \Sigma_\eta$, while the conditionally optimal weights are used with the original forecasts that are not contaminated by noise. Clements and Hendry (1996) show that a similar tradeoff exists when intercept correction strategies are pursued, which use past forecast errors to correct for bias. In online Appendix A1.4, we conduct two Monte Carlo experiments to illustrate the tradeoff numerically in an i.i.d. environment.

3. Monte Carlo exercise

Theory shows that conditionally optimal weights should perform well in situations where bias is present but small. When forecast bias is large, correcting individual forecasts first and then combining should be the best strategy. We investigate this result using a Monte Carlo exercise following Lima and Meng (2017). We consider the following location-scale model:

$$y_{t+1} = \beta_0 + \sum_i \beta_i x_{i,t} + \left(\gamma_0 + \sum_i \gamma_i x_{i,t}\right)\eta_{t+1} \qquad (8)$$
$$i = 1, 2, 3, \ldots, 6;\ \ t = 1, 2, \ldots, 1000,$$

where $\beta_0 = 1$, $\eta_{t+1} \sim N(0, \sigma_\eta^2)$, and $\sigma_\eta = 0.75$. The sample size is set to 1000. For pseudo-forecasting purposes, the same is partitioned into three subsets: 1) $t < 500$, 2) $500 \le t < 901$, and 3) $t > 900$. The first subset is used to provide initial estimates for our individual forecast models, which are described below. The second subset is used to generate a sample of forecast errors for each individual forecast model to construct the initial conditional bias estimates required for conditionally optimal weights and bias-correction forecasts. The final subset is the evaluation period, which we recursively forecast with the individual models and combined forecasts.

The number of potential predictors, $x_{i,t}$, is fixed at six. Predictors are drawn from a uniform distribution over (0, 1), where, as in Lima and Meng (2017), we consider the Spearman correlation among the predictors of $\rho \in (0, 0.1, 0.25, 0.5, 0.95)$. Following Elliott et al. (2013), we consider all distinct subsets of the six predictors of size 1, 2, and 3 as forecasting models. The forecast models take the form of

$$y_{t+1} = b_0 + b_i x_{i,t} + b_j x_{j,t} + b_k x_{k,t} + e_t,$$

where for subsets of size one, $b_j = b_k = 0$ and $i = 1, \ldots, 6$; for subsets of size two, $b_k = 0$ and there are 15 different combinations of $x_{i,t}$ and $x_{j,t}$; and for subsets of size 3, we have 20 different combinations of $x_{i,t}$, $x_{j,t}$, and $x_{k,t}$.

We construct combined forecasts of the six individual models ($k = 1$ in the notation of Complete Subset Regression (CSR) of Elliott et al., 2013), 21 models comprising all subsets of $n \le 2$ ($k = 2$), and 41 models comprising all subsets of $n \le 3$ ($k = 3$) using the following:

1. conditionally optimal weights (COW),
2. bias-corrected optimal weights (BC-OW),
3. bias-corrected equal weights (BC-EW), and
4. equal weights (CSR).

Bias-corrected optimal and equal weights apply conditional bias correction to the individual forecasts first, before the forecasts are pooled using the designating strategy. Conditional bias is modeled as an AR(1) with a constant

$$\hat{y}_{i,t+1} - y_{t+1} = e_{i,t+1} = c_i + \phi e_{i,t} + \zeta_{t+1}.$$

Table 1
Monte Carlo results for scale-location DGP – Weak and partially weak predictors.

Model	$\rho = 0$		$\rho = 0.1$		$\rho = 0.25$		$\rho = 0.5$		$\rho = 0.95$	
	Rel. MSFE	CW-stat	Rel. MSFE	CW-stat	Rel. MSFE	CW-stat	Rel. MSFE	CW-stat	Rel. MSFE	CW-stat
x_1	0.9637	1.92	0.9730	1.65	0.9833	1.31	0.9951	0.74	1.0015	−0.27
x_2	0.9641	1.92	0.9733	1.66	0.9815	1.37	0.9923	0.84	1.0016	−0.25
x_3	1.0012	−0.12	1.0012	−0.14	1.0006	−0.22	1.0009	−0.15	1.0015	−0.23
x_4	1.0006	−0.21	1.0009	−0.23	1.0011	−0.29	1.0011	−0.22	1.0014	−0.25
x_5	1.0015	−0.28	1.0015	−0.31	1.0012	−0.18	1.0018	−0.22	1.0016	−0.24
x_6	1.0011	−0.24	1.0011	−0.31	1.0011	−0.15	1.0016	−0.21	1.0015	−0.27
EW - 6/ CSR ($k = 1$)	0.9757	2.73	0.9816	1.64	0.9881	1.38	0.9957	0.98	1.0014	−0.26
CSR ($k = 2$)	0.9615	2.73	0.9699	2.46	0.9786	2.09	0.9883	1.43	1.0008	0.03
CSR ($k = 3$)	0.9511	2.73	0.9610	2.47	0.9712	2.13	**0.9830**	1.56	**1.0008**	0.09
COW - 6	**0.9366**	2.65	**0.9491**	2.40	0.9641	2.04	0.9851	1.39	1.0138	−0.12
COW - 21	0.9694	2.27	0.9868	2.03	1.0062	1.57	1.0260	1.14	1.0599	0.12
COW - 41	1.0024	2.05	1.0290	1.71	1.0526	1.32	1.0662	0.88	1.1153	0.00
BC-OW - 6	0.9393	2.62	0.9608	2.18	**0.9623**	2.02	0.9875	1.40	1.0195	−0.04
BC-OW - 21	0.9706	2.28	1.0048	1.74	1.0101	1.59	1.0281	1.13	1.0720	0.08
BC-OW - 41	1.0166	1.95	1.0596	1.45	1.0645	1.26	1.0891	0.83	1.1308	0.04
BC-EW - 6	0.9797	1.81	0.9885	1.33	0.9917	1.05	0.9986	0.45	1.0056	−0.09
BC-EW - 21	0.9651	2.28	0.9769	1.79	0.9817	1.57	0.9907	1.02	1.0043	0.05
BC-EW - 41	0.9544	2.47	0.9681	2.01	0.9738	1.82	0.9851	1.29	1.0036	0.14
# simulations:	250		250		250		250		250	

Notes: Monte Carlo simulations for combined forecasts of a location-scale model following Lima and Meng (2017) given by Eq. (8). Each simulation recursively forecasts 100 periods, and the MSFE is recorded. The (Clark & West, 2007) test statistic (CW-stat) for equal forecast accuracy of the combined forecast relative to a forecast from a simple recursive average of past observations (y_t) is calculated for the 100 forecasts. The table shows the mean relative MSFE and mean CW-stat for 250 simulations.

The univariate benchmark is a simple recursive average of y_t.

We assume weak and partially weak predictors with $\beta_i = \gamma_i = 0$ for $i = 3, 4, \ldots, 6$ for all t, $\beta_1 = -1.5$ and $\gamma = 5$ if $\eta_{t+1} \leq \Phi^{-1}(0.5)$, and $\beta_2 = 1.5$ and $\gamma_2 = 5$ if $\eta_{t+1} > \Phi^{-1}(0.5)$, where $\Phi^{-1}(x)$ refers to the $x \times 100$ percentile of the distribution of η_t.

Table 1 shows the mean Monte Carlo results from 250 experiments that each include 100 out-of-sample forecasts. All results are reported relative to the univariate benchmark (a recursive average of y_t) with the average (Clark & West, 2007) test statistic from the 250 experiments shown on the right. Note that by construction only $x_{1,t}$ and $x_{2,t}$ have any ability to forecast y_{t+1}. The remaining predictors are noise and provide no forecasting power above the benchmark.

The combined forecasts all show some ability to forecast y_{t+1} relative to the benchmark forecast. Looking across the different forecast combination strategies, there is a clear relationship between combined forecast accuracy and whether the combination strategy nests optimal weights or equal weights. The former relies on a variance–covariance estimate of the forecast errors, while the latter specifications do not. Combinations based on optimal methods perform best when the number of models combined (n) is small, and the correlation among the predictors is low. For example, COW and BC-OW combinations of the six single $x_{i,t}$ prediction models provide the lowest MSFE among the strategies tested when $\rho \leq 0.25$. In contrast, when $\rho \geq 0.5$, the best strategies rely on equal weights and combine all 41 models.

The explanation for this pattern is straightforward. The majority of the forecasts that are combined in this exercise are noise. Only two of the six predictors contain any information about the DGP by construction. The rest are pure noise. Averaging over that noise with CSR, or BC-EW, zeroes it out and leads to better forecast accuracy. With COW or BC-OW, however, the variance–covariance of the errors from the forecasts is exploited. The addition of many nearly identical noise forecasts leads to highly correlated forecast errors. The highly correlated forecast errors lead to near multicollinearity in the forecast error variance–covariance estimate, and to imprecision in the estimated weights. In addition, the optimal weights can be negative, which generates a range of issues that, when unchecked, affect the combined forecast performance; see Radchenko et al. (2023).

Lastly, we find that combinations that bias-correct the forecasts before combining them systematically yield higher MSFEs than the combinations that use the predicted bias to construct weights for combining the uncorrected forecasts. This illustrates the insights discussed in Section 2.1. Bias-correcting the individual forecasts can introduce noise that the combined forecast cannot remove. However, that same information can be used to construct combination weights that improve upon other combination methods, such as CSR.

4. Practical considerations and operational strategies

The Monte Carlo exercise reveals that conditionally optimal weights are sensitive to the inclusion of many irrelevant forecasts, especially if the forecasts are correlated. Optimal strategies perform well when there are fewer forecasts and when the forecast errors are less correlated. The equal-weights strategy performs better when there are many forecasts and the forecast errors are correlated. In practice, therefore, conditionally optimal weights may benefit from small modifications that provide more robust estimates of Σ_ξ, shrinking the weights towards equal weights, or ignoring Σ_ξ and using only the estimates of the bias.

We consider two different shrinkage strategies to explore in our empirical application. The first relies on constructing a linear combination of the estimates with a fixed stabilization matrix Σ_0 such that $\alpha\Sigma_0+(1-\alpha)[\widetilde{\Sigma}_\xi+\widehat{\boldsymbol{b}}_T\widehat{\boldsymbol{b}}_T^{\prime}]$, $0<\alpha<1$, and

$$\widehat{\boldsymbol{w}}^*_{\mathrm{COWS}}(I_T)=\frac{(\alpha\Sigma_0+(1-\alpha)[\widetilde{\Sigma}_\xi+\widehat{\boldsymbol{b}}_T\widehat{\boldsymbol{b}}_T^{\prime}])^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}^{\prime}(\alpha\Sigma_0+(1-\alpha)[\widetilde{\Sigma}_\xi+\widehat{\boldsymbol{b}}_T\widehat{\boldsymbol{b}}_T^{\prime}])^{-1}\boldsymbol{\iota}}. \tag{9}$$

We refer to $\widehat{\boldsymbol{w}}^*_{\mathrm{COWS}}(I_T)$ as *conditionally optimal weights with shrinkage* (COWS). The second relies on ignoring the $\widehat{\Sigma}_\xi$ estimate entirely and using only the conditional bias estimates. The weights in this case are

$$\widehat{\boldsymbol{w}}^*_b(I_T)=\frac{[\widehat{\boldsymbol{b}}_T\widehat{\boldsymbol{b}}_T^{\prime}]^{-}\boldsymbol{\iota}}{\boldsymbol{\iota}^{\prime}[\widehat{\boldsymbol{b}}_T\widehat{\boldsymbol{b}}_T^{\prime}]^{-1}\boldsymbol{\iota}},$$

where the generalized inverse $[\widehat{\boldsymbol{b}}_T\widehat{\boldsymbol{b}}_T^{\prime}]^{-1}$ can be employed in the case of singularity. This strategy is explicitly forward looking because its efficacy depends only on the predictability of the bias.

Two natural refinements to make when using the predicted bias are

$$\widehat{\boldsymbol{w}}^*_{\mathrm{PB}}(I_T)=\frac{1}{\sum_{l=1}^{n}\widehat{b}_{l,T}^{-2}}\left(\widehat{b}_{1,T}^{-2},\ldots,\widehat{b}_{n,T}^{-2}\right)^{\prime}, \tag{10}$$

which intuitively places larger weights on the forecasts with smaller biases, and

$$\widehat{\boldsymbol{w}}^*_{\mathrm{PE}}(I_T)=\frac{1}{\sum_{l=1}^{n}\exp(-\gamma\widehat{b}_{l,T}^{2})}\left(\exp(-\gamma\widehat{b}_{1,T}^{2}),\ldots,\exp(-\gamma\widehat{b}_{n,T}^{2})\right)^{\prime}, \tag{11}$$

which uses the same information but decreases the weights faster as the bias increases.[7] The parameter γ acts as a tuning parameter, governing the speed at which the weights change, with $\gamma=0$ leading to equal weights and $\gamma\to\infty$ placing all weight on a single forecast. The exponential function is also bounded when the bias is near zero, which stabilizes the weights further relative to the previous strategies. In addition, in the absence of a predictable bias in any of the forecasts ($\widehat{\boldsymbol{b}}_T=\boldsymbol{0}$), the weights collapse to equal weights. We refer to $\widehat{\boldsymbol{w}}^*_{\mathrm{PB}}(I_T)$ as *predicted bias weights* (PBW) and $\widehat{\boldsymbol{w}}^*_{\mathrm{PE}}(I_T)$ as *predicted exponential weights* (PEW).

[7] The predicted bias weights are similar to inverse MSE weights when the mean is calculated over a short sample. In Appendix A2.2.2 and A2.3, we compare this type of strategy and other related inverse MSE strategies to the conditionally optimal weights. We find that the conditionally optimal weights offer forecasting advantages over these alternatives.

5. Real-time forecasting: US inflation

We go into some detail in this section on how we conduct and evaluate real-time forecasts. We use the same methodology for the subsequent two exercises.

5.1. Data, models, and inference

We use data from the Philadelphia Federal Reserve's Real-Time Macroeconomic data set for this exercise, which allows us to restrict the information in each period to that which would have been available to a forecaster at the time a forecast is made.[8] The measure of inflation that we consider is constructed from the Price Index for Personal Consumption Expenditure (PCE). Quarterly inflation is defined as $\pi_t=\ln(p_t/p_{t-1})\times400$, where p_t is the price index.

We use five different variants of real GDP and unemployment to predict inflation and forecast errors. The real GDP measure is used to create three predictors: GDP growth, constructed as log-differenced GDP; the output gap, constructed using the Hodrick–Prescott (HP) filter; and a GDP growth gap measure, which is constructed as the difference between the current GDP growth rate and the maximum GDP growth rate observed over the previous 12 quarters. The unemployment rate is used in levels and as a one-sided unemployment gap measure, which is constructed using the same method as the GDP growth gap. The unemployment gap and growth gap measures follow Stock and Watson (2010) and provide one-sided measures of the business cycle to capture possible nonlinearities in the Phillips curve. We construct each of these measures anew every period using the appropriate vintage of data to ensure that information not available to a forecaster does not contaminate our predictions.

The forecast of interest is the four-quarter-ahead forecast of quarterly inflation, which is expressed as an annual rate.[9] In the real-time data set, quarterly PCE inflation often appears with a lag even though some additional monthly data are often available. We take the data in the Philadelphia Fed's data set as the sole source of information available at any given point in time. The target measure of inflation used to evaluate forecasts is a composite series constructed from the second-release quarterly observations of inflation as they appear in the real-time data set. Comparisons of the composite series to the most recent vintage of inflation are available in online Appendix A2.

5.1.1. Models

We select a collection of parsimonious ARMA, VAR, and direct forecast (DF) models to forecast inflation. The ARMA models we consider are an AR(1), AR(2), AR(4), ARMA(1,1), ARMA(4,4), and the naïve random walk model (AO) proposed by Atkeson and Ohanian (2001), which

[8] A detailed description of the data set and an explanation of its usefulness for evaluating forecasting strategies are provided by Croushore and Stark (2001).

[9] We explored horizons of one to eight quarters and cumulative definitions of inflation, and the results were qualitatively the same.

is the average of the previous four quarters of inflation. These models reflect commonly used benchmark forecasts employed in the literature. For example, the AO forecast is found to forecast inflation well since the mid-1980s. The AO forecast compares well to the inflation gap forecasting strategy proposed by Stock and Watson (2007), as shown by Faust and Wright (2013). The ARMA(1,1) is the benchmark forecast employed by Ang et al. (2007), who compared dozens of different forecast specifications covering surveys, ARMA models, regressions using real-activity measures, and term structure models. The AR(4) is the benchmark in Stock and Watson (2010).

We consider six VAR specifications that we refer to as Phillips curve (PC) models because they use lagged inflation and real activity measures to predict inflation. Five of the six VAR specifications are bi-variate VARs with two lags of inflation and two lags of a real activity measure. The final VAR specification (VAR All) includes four variables: inflation, output gap, GDP growth, and the unemployment gap. The VAR All specification is a robustness check. The model includes all the information that we find forecasts the biases well. If this information is more useful to forecast inflation directly rather than to construct weights, then this specification should reflect that reality.

The conditional bias in the forecast errors is modeled as

$$e_{i,t+4} = c_i + \beta_i x_t + \xi_{i,t+4}, \tag{12}$$

where $e_{i,t+4} = \pi_{t+4} - E_{i,t}\pi_{t+4}$ is the four-quarter-ahead forecast error, and x_t is a real activity measure. This parsimonious specification is chosen to keep our forecasting exercise as transparent as possible. We use the same real activity measures here as in the individual forecast models. This way we are not intentionally biasing our individual forecasts by leaving out relevant predictors. Instead, we are simply using the same information in a different way.

To further verify that we are not intentionally biasing our forecasts, we also include five direct forecast specifications in our combinations. The direct forecasts are OLS regressions of a given real activity measure on four-quarter-ahead inflation, which mimics the exact way we model conditional bias, but used to forecast inflation directly. The DF forecasts along with the ARMA and PC forecasts brings the total number of forecasts we wish to combine to 17.

5.1.2. Out-of-sample forecasts: Evaluation and inference

We evaluate forecasts based on bias, measured by the mean forecast error (MFE), and accuracy, measured by the root mean squared forecast error (RMSFE). Inferences on the observed differences in RMSFE are obtained using the Diebold and Mariano (1995) (DM) test for equal within-sample forecast accuracy, with the Harvey et al. (1997) small sample size and long horizon correction. There is not much guidance in the literature on the correct test statistic for evaluating combined forecasts. Most test statistics assume asymptotic convergence to stationary distributions for the estimated regression coefficients of the models considered and of the forecast errors (see West, 2006). Neither holds in our case. The DM test statistic follows the recommendations provided by Clark and McCracken (2013) for evaluating forecasts on real-time data. Inference on the bias results is obtained using a t-test with Newey and West (1987) standard errors.

The forecast error series is constructed using real-time errors obtained from comparing past real-time forecasts to the composite series of second-release information, which is supplemented with the first-release information for the most recent observations at each point when a forecast is made. This is consistent with a forecaster who is keeping track of the revisions as new data become available. Importantly, no data that would not have been available to the forecaster at the time the forecast is made are used. This convention makes it so that the individual forecast model's bias, which informs the combination weights, will closely match the real-time bias that we report for each individual model. We also explored using each new vintage of data at each point in time as the benchmark for creating a new history of forecast errors. Under this approach, the real-time estimates of a forecast's bias that we report, and the bias used to inform the combination weights, may differ over time. However, we found that the results changed little under this alternative scenario.

The out-of-sample forecasting exercise requires the data to be separated into three periods. The required divisions are (1) a training period to estimate the initial parameters of the forecast models, (2) an in-sample forecasting period to recursively construct an initial series of forecast errors to estimate Eq. (12), and (3) an out-of-sample period to conduct out-of-sample forecasts. For the full sample experiment, the three periods are 1947Q2–1965Q4, 1966Q1–1969Q4, and 1970Q1–2018Q1, respectively, where the last forecast is made in 2017Q1 and compared with the realization in 2018Q1. In addition, we consider two subsamples with periods 1947Q2–1965Q4, 1966Q1–1982Q4, and 1983Q1–2007Q3 and periods 1947Q2–1965Q4, 1966Q1–2007Q3, and 2007Q4–2018Q1.

5.2. Results

Table 2 reports the real-time forecasts of inflation and their bias-corrected counterparts. The first column shows the MFE of each model for the full sample. All of the forecasts are found to be unbiased by this measure. However, the opposite result is obtained for the two subsample periods, where both show a positive bias. The two results are reconciled by the fact that inflation was generally rising during the first part of the sample and falling or steady during the two subsamples. The relatively stable inflation observed in the 1960s did not prove useful to forecast the dramatic rise in inflation seen in the 1970s, leading to negative biases. Likewise, the low and stable inflation of the latter part of the sample was a surprise to forecasts based on the experience of the 1970s and 1980s.

The bottom panel of Table 2 shows the results for bias-corrected forecasts, where we use Eq. (12) and the

Table 2
Individual uncorrected and bias-corrected real-time forecasts of US inflation.

Uncorrected	Horizon: Four quarters								
	Full sample: 1970Q1–2017Q1			Subsample: 1983Q1–2007Q3			Subsample: 2007Q4–2017Q1		
	MFE	RMSFE	Rel. RMSFE	MFE	RMSFE	Rel. RMSFE	MFE	RMSFE	Rel. RMSFE
AO	**0.06**	2.29	1.00	**0.02**	1.48	1.00	**0.06**	1.97	1.00
AR(1)	**−0.11**	2.56	1.12	0.84	1.58	1.07	1.11	2.21	1.12
AR(2)	**−0.07**	2.35	1.03	0.64	1.51	1.02	0.89	2.16	1.10
AR(4)	**−0.21**	2.46	1.08	0.63	1.51	1.02	0.85	2.15	1.09
ARMA(1,1)	**−0.08**	2.30	1.01	0.57	1.47	0.99	0.79	2.12	1.07
ARMA(4,4)	**−0.16**	2.48	1.08	0.70	1.66	1.12	0.64	2.05	1.04
DF Output Gap	**−0.04**	3.17	1.39	1.30	2.17	1.46	1.70	2.51	1.27
DF Unemployment Gap	**0.07**	2.92	1.28	1.22	1.85	1.25	1.84	2.51	1.27
DF GDP Growth	**−0.11**	2.93	1.28	1.10	1.76	1.19	1.61	2.34	1.19
DF Growth Gap	**0.07**	2.91	1.27	1.27	1.87	1.26	1.74	2.43	1.23
DF Unemployment Rate	**0.37**	2.98	1.30	1.42	2.12	1.43	2.13	2.82	1.43
VAR ALL	**−0.21**	2.49	1.09	0.71	1.61	1.09	0.81	2.20	1.12
PC Output Gap	**0.02**	2.35	1.03	0.62	1.57	1.06	0.97	2.28	1.15
PC Unemployment Gap	**−0.28**	2.39	1.04	0.54	1.46	0.98	0.72	2.13	1.08
PC GDP Growth	**−0.04**	2.30	1.01	0.53	1.51	1.02	0.76	2.15	1.09
PC Growth Gap	**0.21**	2.45	1.07	0.81	1.73	1.17	1.23	2.40	1.22
PC Unemployment Rate	**−0.25**	2.45	1.07	0.54	1.78	1.20	**0.48**	2.13	1.08
Bias-corrected									
AO	−2.23	5.44	2.38	−0.37	2.70	1.82	**−0.14**	3.56	1.81
AR(1)	−2.46	5.29	2.32	**−0.90**	2.84	1.91	**−0.48**	3.63	1.84
AR(2)	−2.25	5.23	2.29	−0.62	2.74	1.85	**−0.33**	3.59	1.82
AR(4)	−2.51	5.25	2.30	−1.05	2.87	1.93	**−0.60**	3.64	1.85
ARMA(1,1)	−2.26	5.13	2.24	−0.72	2.84	1.92	**−0.59**	3.61	1.83
ARMA(4,4)	−3.31	5.44	2.38	−2.40	3.63	2.45	−1.78	4.07	2.06
DF Output Gap	−1.69	5.71	2.50	**−0.64**	2.97	2.00	**0.72**	3.55	1.80
DF Unemployment Gap	−1.75	5.52	2.41	**0.34**	2.73	1.84	**0.58**	3.46	1.75
DF GDP Growth	−1.78	5.57	2.43	**0.38**	2.70	1.82	**0.57**	3.45	1.75
DF Growth Gap	−1.70	5.46	2.39	**0.39**	2.72	1.84	**0.50**	3.44	1.74
DF Unemployment Rate	−1.71	5.40	2.36	**0.08**	2.92	1.97	**0.59**	3.53	1.79
VAR ALL	−2.22	5.29	2.32	**0.47**	2.74	1.84	**−0.19**	3.62	1.83
PC Output Gap	−2.39	5.19	2.27	−1.05	2.91	1.96	**−0.42**	3.63	1.84
PC Unemployment Gap	−2.32	5.26	2.30	−0.69	2.73	1.84	**−0.34**	3.62	1.83
PC GDP Growth	−2.41	5.11	2.24	−1.15	2.93	1.98	**−0.58**	3.65	1.85
PC Growth Gap	−2.27	5.10	2.23	−1.04	2.96	2.00	**−0.32**	3.65	1.85
PC Unemployment Rate	−2.38	5.25	2.30	−0.78	3.06	2.07	**−0.58**	3.67	1.86

Notes: The table reports the mean forecasting error (MFE), root mean squared forecasting error (RMSFE), and relative RMSFE compared to the AO forecast for recursive real-time out-of-sample forecasts for US PCE inflation. The in-sample data start in 1947Q2 in all reported results. Significance for the relative results ($^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$) is only indicated for improvements over the benchmark. Bolded values of MFEs indicate a failure to reject the null hypothesis of unbiasedness at the 10% significance level ($p > 0.1$).

output gap to predict the bias and correct each of the point forecasts such that[10]

$$E_t\pi^{BC}_{i,t+4} = E_t\pi_{i,t+4} + E_t e_{i,t+4}. \quad (13)$$

Bias correction fails to maintain unbiasedness on the full sample but does unbias many of the forecasts for the subsample periods. The reasoning mirrors the original bias results. The rise and fall of inflation in the early part of the sample was difficult to predict. The persistent positive bias in the latter sample is correctable. However, the correction comes at the cost of a near universal doubling of RMSFE. Therefore, although the later bias-corrected forecasts are indeed unbiased, they are inaccurate, which illustrates the issue shown in Section 2.1 with bias correction in practice.

5.2.1. Combined forecast results

We turn now to combinations of the 17 forecasts. We compare conditionally optimal weights (COW), conditionally optimal weights with shrinkage (COWS), predicted bias weights (PBW), and predicted exponential weights (PEW) to combine the individual forecasts. We also test the outcome when we select the forecast with the lowest expected bias in each period, which is equivalent to a PEW forecast with $\gamma \rightarrow \infty$. This forecast removes the hedging benefit of combining forecasts and tests whether our bias predictions truly provide useful forecasts period-by-period. We apply shrinkage to the estimated weights by taking a linear combination of the estimated variance–covariance matrices and an identity matrix with a weight equal to one-half, which shrinks towards equal weights.[11]

10 Similar results are obtained for each of the real activity measures we consider, which can be seen in Table 3 in the next section by looking at the bias-corrected optimal-weights forecast results or bias-corrected equal-weights forecast results.

11 The inclusion of shrinkage is important for the full-sample results, as the volatility of forecast errors in the 1970s and early 1980s introduces significant instability into the variance–covariance estimates, leading to very poor forecasts. For the post-1983 sample, when inflation is more stable, small or no shrinkage can produce superior forecasts to the reported results in many instances. For

For the PEW specification, we set $\gamma = 5$, which was determined by searching over whole number values of γ to minimize the MSFE in a pre-sample (1967Q1–1969Q4) period.

We compare the conditionally combined forecasts to five different combined benchmark forecasts. The first is the equal-weights forecast. The second is the classic unconditional optimal-weights forecast with shrinkage (OWS) and without shrinkage (OW). The third is the bias-corrected equal-weights forecast (BC-EW). The fourth is the bias-corrected optimal-weights forecast with shrinkage (BC-OWS). The latter two strategies use equal weights and classical optimal weights, respectively, to combine the individual bias-corrected forecasts. The same bias predictions that we use to construct conditionally optimal weights are used to bias-correct the individual forecasts. We also apply the same shrinkage method to the optimal weights as we used for the conditional weights to create fair comparisons across all forecasts.

Table 3 shows the combined forecast results for four-quarter-ahead real-time forecasts of US PCE inflation. We report results grouped by the predictor used in Eq. (12) to forecast the bias. When comparing the bias-corrected combinations to the conditionally optimal and forward-looking combinations, distinct patterns emerge with respect to bias (MFE) and forecast accuracy (RMSFE). The conditionally optimal strategy produces unbiased forecasts over the full sample, while bias-corrected combinations remain biased. In the subsamples, the conditionally optimal strategies are more likely to be unbiased than their bias-corrected counterparts. When both are biased, conditionally optimal strategies produce smaller biases on average. The COW and COWS strategies, in particular, stand out in this regard by being unbiased in nearly all cases considered. Similarly, the conditionally optimal strategies produce more accurate forecasts than bias-corrected optimally combined forecasts. Not a single bias-corrected combined forecast significantly outperforms an equal-weights forecast in any of the samples tested, and in only two of the 15 considered cases does it show a qualitative improvement, i.e., a lower but statistically insignificant result.

The disparity in the effectiveness between the COW(S) forecasts and bias-corrected forecasts is explained by the different way the bias predictions are used. The COW(S) strategies primarily exploit the relative ranking of the forecasts implied by the predicted biases, whereas bias correction relies on the accuracy of the actual point forecast of the bias. This insight underlies the intuition of Theorem 2 in Section 2.1. Bias correction can introduce more noise into the forecasts than the actual bias generates, which results in significantly worse forecasts. Leaving the underlying forecasts unchanged and using the predicted bias to correct weights, however, provides a robust way to use this information when forecasting.

The PEW forecast in particular illustrates the role that bias prediction plays in aiding combination weights. The weights, in this case, do not shrink toward equal weights but rather increase toward the forecast with the lowest expected squared error. Therefore, in each period, we rely on the individual bias predictions to identify the best forecasts. The forecast performance of this strategy is more varied than the performance of strategies that rely more on averaging, but also yields some of the best forecasts.

The PEW forecast with $\gamma \to \infty$ pushes bias prediction to its limit. This approach removes the hedging advantage of a combined forecast and demonstrates how forward-looking bias prediction actually drives improvements in forecast accuracy.[12] We find that choosing the expected best model in each period leads to reductions in the relative RMSFE compared to equal weights in a majority of the out-of-sample forecast experiments. It also leads to a reduction in bias compared to the combined forecast benchmarks in most cases.

Although the overall gains in forecast accuracy of conditionally optimal weights and the combined strategies are small, the timing of the gains is economically relevant. In online Appendix A2.2, we show that the largest improvements in forecast accuracy occur around US recessions, when it is arguably most important to make accurate forecasts.

5.3. Robustness

Comparisons of combined forecast strategies face an external validity problem because the results are sensitive to the set of forecast models considered for a combination. This concern is particularly acute when comparing an optimal combination strategy to an equal-weights forecast. An equal-weights forecast has no mechanisms to filter out obviously poor forecasts, which makes it easy to construct a straw man equal-weights forecast by including poorly performing forecasts that a more sophisticated model combination strategy can easily disregard and which an actual forecaster would not consider.

To address this issue, we study a type of value-at-risk calculation for combined forecasts by varying the underlying forecasts that we combine. We select the 12 best individual models found in Table 2. We then choose n models at a time to form every unique combination of the models for $n = 2, 3, \ldots, 12$, which provides us with 4083 unique sets of forecast models to combine.[13] We conduct a real-time out-of-sample forecast exercise using the full sample (1970–2018) comparing the RMSFE obtained from all 4083 sets combined using the OWS, COWS, and PEW specifications to an equal-weights

example, if the out-of-sample period from 1990–2018 is analyzed, then the relative MSFE of COW using the output gap to forecast the bias is 0.94 compared to the AO forecast of 0.96.

12 This case is of particular interest because Timmermann (2006) notes that choosing a single model in every period typically results in poor out-of-sample forecast accuracy. In online Appendix A2.2.2, we offer an in-depth exploration of backward- versus forward-looking forecasts, showing why active prediction trumps using a static estimate of the bias.

13 The number of distinct combinations of the 12 models for each n is as follows: $n = 2 \to 66$ sets, $n = 3 \to 220$ sets, $n = 4 \to 495$ sets, $n = 5 \to 792$ sets, $n = 6 \to 924$ sets, $n = 7 \to 792$ sets, $n = 8 \to 495$ sets, $n = 9 \to 220$ sets, $n = 10 \to 66$ sets, $n = 11 \to 12$ sets, and $n = 12 \to 1$ set.

Table 3
Combined and benchmark real-time forecasts of US inflation.

Benchmarks	Horizon: Four quarters								
	Full Sample: 1970Q1–2018Q1			Subsample: 1983Q1–2007Q3			Subsample: 2007Q4–2018Q1		
	MFE	RMSFE	Rel. to EW	MFE	RMSFE	Rel. to EW	MFE	RMSFE	Rel. to EW
AO	**0.06**	2.29	0.94	**0.02**	1.48	0.95	**0.06**	1.97	0.92*
Equal Weights (EW)	−**0.04**	2.42	1.00	0.80	1.57	1.00	1.08	2.14	1.00
Optimal Weights (OW)	−1.35	6.07	2.50	−0.87	2.46	1.57	−0.93	2.49	1.16
Optimal Weights with Shrinkage (OWS)	0.50	2.39	0.99	**0.19**	1.49	0.95	**0.07**	2.09	0.97
Output Gap Bias Prediction									
Bias-Corrected EW	1.30	2.84	1.17	2.27	3.01	1.92	1.42	2.48	1.16
Bias-Corrected OWS	**0.55**	2.60	1.08	1.12	2.28	1.45	**0.43**	2.25	1.05
Conditionally Optimal Weights (COW)	**0.29**	3.31	1.37	**0.12**	2.01	1.28	**0.12**	1.96	0.92*
Conditionally Optimal Weights with Shrinkage (COWS)	−**0.23**	2.22	0.92*	**0.05**	1.58	1.01	**0.01**	2.10	0.98
Predicted Bias Weights (PBW)	−**0.06**	2.15	0.98*	0.72	1.67	0.95***	0.72	2.05	0.96
Predicted Exponential Weights (PEW)	−**0.02**	2.36	0.89**	0.29	1.42	0.91**	0.74	2.01	0.94***
Best Predicted Model	−**0.07**	2.19	0.90***	**0.27**	1.54	0.98	**0.22**	1.99	0.93
Unemployment Gap Bias Prediction									
Bias-Corrected EW	1.29	2.79	1.15	2.22	2.66	1.70	1.42	2.48	1.16
Bias-Corrected OWS	0.66	2.88	1.19	1.51	2.28	1.45	**0.59**	2.19	1.02
Conditionally Optimal Weights	−0.64	4.22	1.74	−**0.36**	2.20	1.40	−**0.20**	2.33	1.09
Conditionally Optimal Weights with Shrinkage	−**0.40**	2.36	0.97	**0.21**	1.59	1.01	**0.34**	2.05	0.96
Predicted Bias Weights	−**0.18**	2.38	0.98	0.99	1.79	1.02	1.32	2.21	1.03
Predicted Exponential Weights	**0.35**	2.35	0.97	0.53	1.59	1.01	1.08	2.12	0.99
Best Predicted Model	0.50	2.43	1.00	0.51	1.67	1.07	1.65	2.45	1.14
GDP Growth Bias Prediction									
Bias-Corrected EW	1.20	2.58	1.07	2.07	2.54	1.62	1.37	2.33	1.09
Bias-Corrected OWS	0.52	2.37	0.98	1.06	1.93	1.23	**0.42**	2.10	0.98
Conditionally Optimal Weights	−**0.14**	3.64	1.50	−**0.06**	1.98	1.26	−**0.06**	2.02	0.94
Conditionally Optimal Weights with Shrinkage	−**0.19**	2.19	0.90**	**0.11**	1.47	0.94	**0.37**	2.02	0.94
Predicted Bias Weights	−**0.07**	2.38	0.98***	0.80	1.78	1.01	1.21	2.33	1.09
Predicted Exponential Weights	**0.24**	2.25	0.93*	0.43	1.49	0.95	1.11	2.17	1.01
Best Predicted Model	**0.21**	2.35	0.97	**0.27**	1.50	0.96	1.30	2.49	1.16
GDP Growth Gap Bias Prediction									
Bias-Corrected EW	1.29	2.68	1.10	2.10	2.55	1.63	1.39	2.35	1.10
Bias-Corrected OWS	0.70	2.56	1.06	1.30	2.04	1.30	0.63	2.14	1.00
Conditionally Optimal Weights	−0.46	3.78	1.56	−**0.33**	2.21	1.41	−**0.02**	2.09	0.98
Conditionally Optimal Weights with Shrinkage	−**0.22**	2.25	0.93*	**0.19**	1.50	0.95	**0.46**	2.02	0.94
Predicted Bias Weights	−**0.05**	2.28	0.94***	0.78	1.76	1.00	1.34	2.36	1.10
Predicted Exponential Weights	**0.31**	2.31	0.95	0.45	1.48	0.94	1.19	2.19	1.02
Best Predicted Model	**0.33**	2.38	0.98	**0.27**	1.49	0.95	1.50	2.49	1.16
Unemployment Rate Bias Prediction									
Bias-Corrected EW	1.05	2.67	1.10	2.12	2.54	1.62	1.20	2.22	1.04
Bias-Corrected OWS	0.56	2.52	1.04	1.37	2.08	1.33	**0.39**	2.17	1.01
Conditionally Optimal Weights	−**0.15**	3.57	1.47	−**0.26**	2.06	1.32	−**0.24**	2.31	1.08
Conditionally Optimal Weights with Shrinkage	−**0.15**	2.24	0.93**	**0.15**	2.24	0.93***	**0.21**	2.07	0.97
Predicted Bias Weights	−**0.19**	2.20	0.91***	0.75	1.67	0.94***	0.96	2.06	0.96
Predicted Exponential Weights	**0.38**	2.37	0.98	0.52	1.55	0.99	1.06	2.11	0.98***
Best Predicted model	**0.31**	2.38	0.98	0.48	1.63	1.04	0.81	2.12	0.99

Notes: The table reports the mean forecasting error (MFE), root mean squared forecasting error (RMSFE), and relative RMSFE compared to EW for recursive real-time out-of-sample forecasts for US PCE inflation. The in-sample data start in 1947Q2 in all reported results. Significance for the relative results ($^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$) is only indicated for improvements over the benchmark. Bolded values of MFEs indicate a failure to reject the null hypothesis of unbiasedness at the 10% significance level ($p > 0.1$).

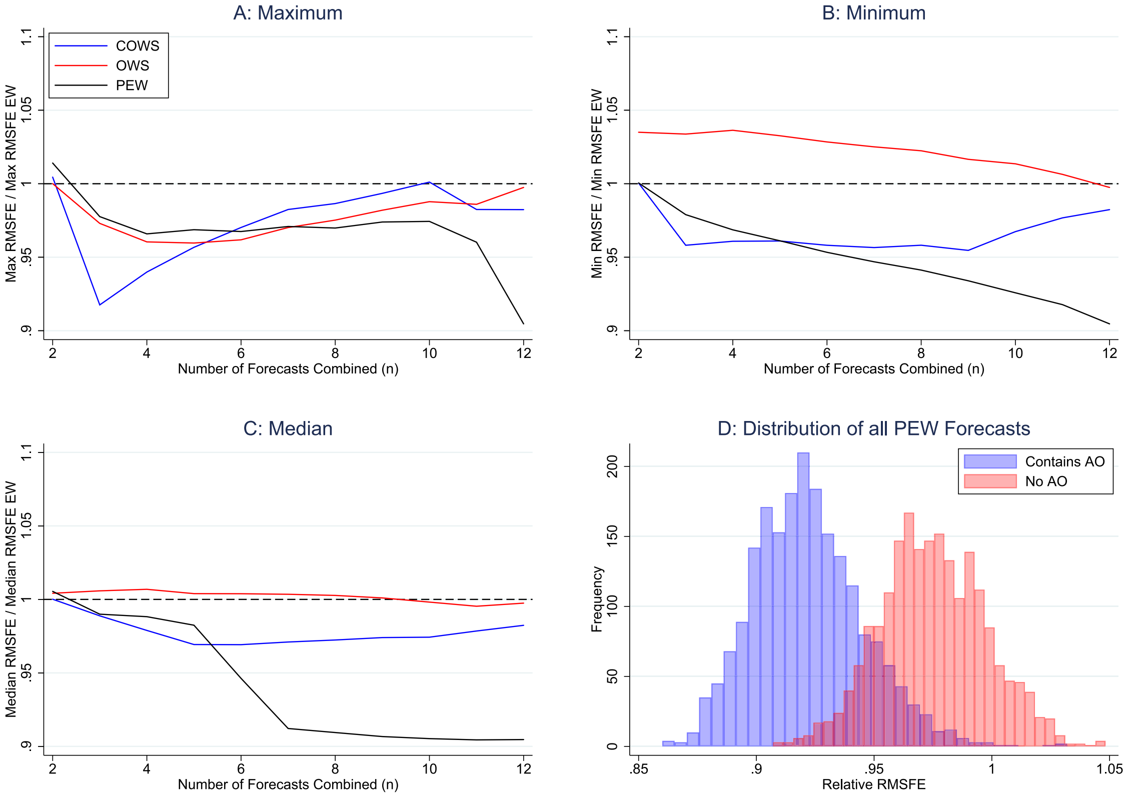

Fig. 1. Forecasting tournament results for US inflation. *Notes:* Panels A (maximum), B (minimum), and C (median) show the maximum, minimum, and median observed RMSFEs for conditional optimally weighted (COWS), unconditional optimally weighted (OWS), and predicted exponentially weighted (PEW) combinations of all unique sets of n forecasts compared to the maximum, minimum, and median observed RMSFEs of equally weighted (EW) combinations of all unique sets of n forecasts, where $n = 2, \ldots, 12$. Panel D shows a histogram of the RMSFEs of all PEW combined forecasts relative to the RMSFEs of an equally weighted forecast of the same unique set. Outcomes are separated by whether the set of forecasts contains the AO forecast.

forecast.[14] To summarize our results for each n, we compute the maximum, median, and minimum observed RMSFE for optimal combinations of all unique sets of n forecasts and compare them to the maximum, median, and minimum observed RMSFE for equally weighted forecasts of all unique sets of n forecasts. By comparing forecasts using the number of models combined, we can characterize the worst-, median, and best-case scenarios for each strategy if a researcher chooses n forecast models at random to combine.

Fig. 1 presents a summary of the full sample tournament results. The RMSFEs are shown relative to the equal-weights forecast RMSFEs for combinations of n models. The maximum relative RMSFEs observed for combinations of n models are shown in Panel A. All three strategies perform reasonably well by this metric by producing lower RMSFEs in nearly every instance compared to an equal-weights forecast. This means that the worst possible optimally combined forecasts obtained for any combination of n models result in a lower RMSFE than the worst equal-weights forecasts of n models for all three considered strategies. The PEW strategy, though, is the only strategy whose relative accuracy appears to be increasing in n.

For the minimum and median results respectively shown in Panels B and C, we find the familiar result that optimal-weights strategies are usually not optimal in practice when compared to equal weights. The minimum and median OWS forecast accuracy is worse than the minimum and median EW accuracy. This failure though does not carry over to the COW strategies. Both consistently produce more accurate combined forecasts than equal weights, with PEW increasing in accuracy as n increases.

The positive relationship between n and the forecast accuracy of the PEW forecast is significant because it is driven by its ability to exploit time variation in PC-type forecasts relative to the AO forecast. We show this explicitly in an exercise in the online Appendix in Section A2.2.2. Increases in accuracy are the result of weights shifting toward PC specifications during economic downturns. To illustrate here, Panel D in Fig. 1 presents a histogram of all 4083 subsets of PEW versus equal-weights forecasts for the same sets of models. Combinations that include the AO forecast significantly outperform those that do not. As n increases, the number of combinations that include both the AO forecast and at least one Phillips curve forecast increases, which generates the increasing

[14] For COWS and PEW, we use the output gap to predict the bias. Similar results are obtained for each of the previously studied predictors.

relative accuracy of PEW as n increases. This figure also illustrates the robustness of a simple forward-looking strategy. PEW generates a lower RMSFE in 3836 of the 4083 subsets.

6. Real-time forecasting: ECB SPF

In this section, we investigate the usefulness of conditionally optimal strategies to combine the individual EU harmonized inflation forecasts from the ECB Survey of Professional Forecasters. We pivot to Europe but continue to use inflation to show that our results are not unique to US inflation and that our approach is applicable when combining large numbers of forecasts. The survey includes 106 individual forecasters. For consistency, we again consider one-year-ahead forecasts, but similar results were obtained at other horizons. Real-time data for inflation begin in 2001Q2, and we consider forecasts through 2018Q1. We use the same real-time forecasting procedure here as in Section 5 and compare the forecasts with a target series composed of second-release vintages.

It remains true in this environment that simple aggregations of survey forecasts, such as the mean or median of forecasts, are typically difficult to beat Genre et al. (2013) despite persistent biases in the individual forecasts (Capistrán & Timmermann, 2009). The combination of survey forecasts is also complicated by the fact that these surveys usually produce an unbalanced panel data set. This means that optimal strategies that rely on consistent estimates of the variance–covariance of past forecast errors are often not feasible without excluding some of the forecasts or imputing observations. Conditionally optimal strategies are particularly advantageous when there is an unbalanced panel because, as we have seen, using forward-looking bias estimates alone and ignoring the variance–covariance of the past forecasts still offers a route to improve the combined forecast accuracy.

Since it is unknown how each participant arrives at his or her forecast, we do not attempt to add a predictor to create a forecast of the bias. Instead, we use a simple AR(1) process to capture any serial dependency in the participants' forecast errors. Bias estimates are then constructed iteratively using the AR(1) estimates of the observed error for each forecaster.[15] Due to the missing survey responses, we use the element-by-element method of Matsypura et al. (2018) to estimate the necessary variance–covariance matrices. We find the estimates to be highly erratic, so we apply a large shrinkage value of $\alpha = 0.1$. However, we leave the expected bias portion of the conditionally optimal weights unshrunk such that

$$\widehat{\boldsymbol{w}}^{*}_{\text{COWS}}(I_T) = \frac{(\alpha\Sigma_0 + (1-\alpha)[\widetilde{\Sigma}_\xi] + \widehat{\boldsymbol{b}}_T\widehat{\boldsymbol{b}}_T')^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'(\alpha\Sigma_0 + (1-\alpha)[\widetilde{\Sigma}_\xi] + \widehat{\boldsymbol{b}}_T\widehat{\boldsymbol{b}}_T')^{-1}\boldsymbol{\iota}}. \tag{14}$$

We omit forecasters for whom we do not observe sufficient past forecasts to estimate forecast bias. Because of these omissions, we also test a PEW forecast that uses all available forecasts. To accomplish this, we assume that all forecasts without sufficient data have a bias equal to the mean bias for all forecasters observed from the previous year. We assume that the shrinkage parameter is the same as in the US exercise ($\gamma = 5$). We compare the conditional forecast combinations to an equally weighted forecast, the median forecast, the bias-correction strategy of Issler and Lima (2009), and the bias-corrected equal-weights forecast, which uses the AR(1) bias predictions to correct individual forecasts before combining. Issler and Lima's approach is identical to our bias-corrected equal-weights forecast except that a nonparametric method is used to estimate a fixed bias for each forecast, which is described in online Appendix A3. Their approach is an interesting benchmark to consider because they show it provides an optimal forecast if the data are stationary and the biases are fixed.

Table 4 shows the out-of-sample results. All of the combined forecasts with the exception of BC-IL are unbiased. The COWS forecast provides a statistically significant 4% improvement over equal weights.[16] The PEW forecast also returns a qualitative improvement of approximately 1.5%. The bias-corrected combinations fail to improve on equal weights or the median forecast.

To analyze the robustness of this result, we construct subsamples of forecasts by randomly selecting 40 of the 106 surveyed forecasters with replacement and combining their forecasts using COWS, PEW, or BC-IL.[17] Fig. 2 shows the distribution of real-time relative RMSFEs for 300 draws of 40 forecasters evaluated on the full out-of-sample period. The majority of the COWS and PEW combined forecasts outperform an equal-weights forecast. The opposite is true of the BC-IL forecasts.

7. Real-time forecasting: International macroeconomic data

For the final exercise, we conduct a real-time recursive out-of-sample forecast comparison using 25 macroeconomic time series to illustrate the tradeoff between bias correction and forecast accuracy highlighted by Theorems 2 and 3 in Section 2. In particular, the accuracy of the conditional bias forecast affects whether one should bias-correct or conditionally optimally combine the forecasts. When the noise in the conditional bias outweighs the signal, conditionally optimal weights produce combined forecasts with a lower mean squared error than bias-corrected combined forecasts.

We forecast real-time data from the Organization for Economic Cooperation and Development for the US, Canada, the United Kingdom, Australia, and New Zealand comprising inflation (GDP deflator), GDP growth, consumption growth, and investment growth. In addition, we

[15] We estimated the AR(1) parameters by maximum likelihood using the Kalman filter. We used filtered estimates to fill in missing values when forecasters have missing observations. In online Appendix A3, we show there is substantial autocorrelation in the forecast errors.

[16] In the online Appendix, in Figure A9, we use the COW forecast and the equal-weights forecast against the target inflation series to illustrate how the forecasts differ.

[17] Because survey participants do not respond in every quarter and some participants join later in the sample, the number of individual forecasts combined varies quarter-by-quarter. Therefore, despite selecting 40 survey participants for each forecasting exercise, fewer than 20 forecasts are combined in each quarter in most cases.

Table 4
Combined forecasts of EU harmonized inflation.

Simple combinations	Horizon: Four quarters		
	Sample: 2001Q2–2018Q1		
	MFE	RMSFE	Rel. RMSFE
Equal weights (EW)	**−0.06**	0.84	1.00
Median	**−0.07**	0.84	1.00
Bias-corrected combinations			
BC-EW	**−0.33**	1.04	1.23
BC-IL	0.35	0.88	1.04
Conditional combinations			
COW	**−0.00**	0.81	0.96*
PEW	**−0.05**	0.83	0.99

Notes: The table reports the mean forecasting error (MFE), root mean squared forecasting error (RMSFE), and relative RMSFE compared to EW for recursive real-time out-of-sample forecasts for EU harmonized inflation constructed using the ECB's SPF. Significance for the relative results (***$p < 0.01$, **$p < 0.05$, *$p < 0.1$) is only indicated for improvements over the benchmark. Bolded values of MFEs indicate a failure to reject the null hypothesis of unbiasedness at the 10% significance level ($p > 0.1$).

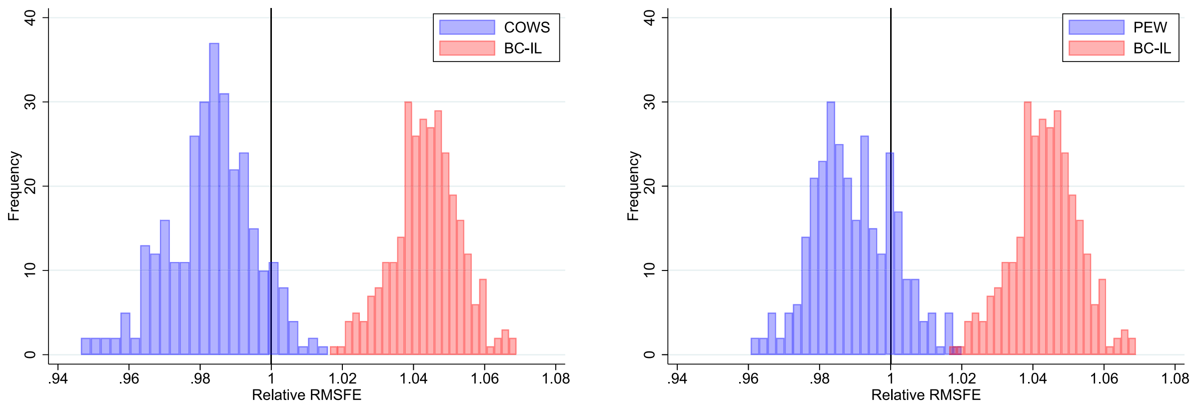

Fig. 2. Distribution of combined four-quarter-ahead forecast results for EU harmonized inflation. *Notes:* Each observation represents the full-sample real-time RMSFE of a combined forecast of 40 randomly selected participants of the ECB's Survey of Professional Forecasters relative to the equally weighted forecasts of those same 40 selected forecasters.

supplement these data with data on 10-year government bond interest rates obtained from the St. Louis Federal Reserve's Economic Database. We use comparable data for each country from 1980Q2 to 2017Q2 with real-time data starting in the first quarter of the year 2000. We follow the same procedure for real-time forecasting in this exercise as outlined in Section 5.1. We forecast each of these series at a four-quarter horizon using 12 simple forecast specifications: four univariate ARMA models, four direct forecasts, and four bivariate VARs.[18]

The simple forecasting models we consider produce varying degrees of bias as measured by MFEs across the different data series. For example, all the models assume that the time series are stationary, which is a reasonable assumption ex ante but problematic ex post, as both interest rates and inflation experienced significant declines from the 1980s to the present in all of the considered countries. This results in out-of-sample forecasts that are persistently positively biased for these variables. In contrast, real GDP growth, real consumption growth, and real investment growth are mostly stationary series, and the considered forecasts have little to no discernible bias. Figure A11 in online Appendix A4 plots the data against the forecasts to provide a visualization of different forecast biases.

To quantify the bias here, we pool the data by type, τ (e.g., inflation, interest rates), and estimate a model that nests Eq. (12) for predicting the bias:

$$e^{\tau}_{i,j,t+4} = c_i + \mu_j + \beta x_{j,t} + \Gamma_i(x_{j,t}) + \epsilon_{i,j,t+4} \quad (15)$$

where $e^{\tau}_{i,j,t+4}$ is the forecast error of model $i \in \{1, 2, \ldots, 12\}$ from country j's data, μ_j is a country fixed effect, $x_{j,t}$ is a real activity measure, and $\Gamma_i(x_{j,t})$ is a full set of interaction and dummy variables that indicate the difference between ARMA, VAR, and direct forecasts. We use real GDP growth to predict the bias for inflation, interest rates, consumption growth, and investment growth. We use consumption growth to predict real GDP growth. The results are shown in Table 5.

From these regressions, we see that there are varying amounts of predictable information in the forecast

18 The four univariate models are the AO, AR(1), ARMA(1,1), and AR(4). For each variable, we use the four other macroaggregates as the predictors in the DF and VAR specifications. We adopt the same two-lag specification as used in the previous exercises.

Table 5
Bias prediction regressions for international data.

Variables	Forecast errors Inflation	Forecast errors Interest rates	Forecast errors RGDP Growth	Forecast errors Cons. Growth	Forecast errors Inv. Growth
Constant	−0.712***	−3.409***	−0.271**	−0.452***	−1.144**
	(0.123)	(0.042)	(0.111)	(0.095)	(0.447)
RGDP ($x_{j,t}$)	0.027	0.008	–	0.074***	0.161*
	(0.024)	(0.009)	–	(0.019)	(0.090)
Consumption ($x_{j,t}$)	–	–	−0.048**	–	–
	–	–	(0.023)	–	–
RGDP × ARMA	0.006	−0.026**	–	−0.036	−0.115
	(0.034)	(0.012)	–	(0.027)	(0.126)
Consumption × ARMA	–	–	−0.050	–	–
	–	–	(0.032)	–	–
RGDP × VAR	−0.006	−0.039***	–	−0.015	−0.015
	(0.034)	(0.012)	–	(0.027)	(0.126)
Consumption × VAR	–	–	−0.022	–	–
	–	–	(0.032)	–	–
ARMA	0.272**	3.102***	0.506***	0.364***	1.134**
	(0.133)	(0.047)	(0.127)	(0.105)	(0.489)
VAR	0.322**	3.182***	0.054	0.120	0.060
	(0.133)	(0.047)	(0.127)	(0.105)	(0.489)
Observations	5688	5052	5088	5160	5160
R-squared	0.010	0.691	0.014	0.018	0.005
Country FE	Yes	Yes	Yes	Yes	Yes

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

Notes: OLS regression estimates of pooled forecast errors from the 12 proposed models for out-of-sample forecasts of data from the US, Canada, the United Kingdom, Australia, and New Zealand. Real GDP is the predictor for the forecast errors of inflation, the interest rate, consumption growth, and investment growth. Consumption is the predictor used in the real GDP growth (RGDP) forecast error regression. 'VAR' and 'ARMA' are indicator variables that take a value of one if the forecast error is produced by that model type and take a value of zero otherwise.

errors across the five types of data. The interest rate forecast errors are large and contain the most predictable information with an R-squared of nearly 0.7. The interest rate forecasts also show a high degree of forecast disagreement among the different types of forecasts that produced the errors. The investment growth forecasts show nearly the opposite relationship. The regression estimates in this case indicate minimal differences among the three types of forecasts, and the R-squared is just 0.005. The remaining three types of data fall between these two cases.

Fig. 3 illustrates the relationship between the predictability of the bias and combined forecast accuracy for bias-corrected and COW combined forecasts. The conditional bias is calculated using Eq. (12) with the real activity measures used in Eq. (15) for these forecasts. Each point on the graph shows the relative MSFE for the full out-of-sample period of a combination of two individual forecast types plotted against a measure of the ex post bias of the individual forecasts. We quantify the bias of the underlying forecasts as

$$\left| \frac{1}{2N} \left(\sum_{t=1}^{N} \left(e^{\tau}_{i,j,t+4} + e^{\tau}_{k,j,t+4} \right) \right) \right|,$$

where $i \neq k$ and $i, k \in \{1, 2, \dots, 12\}$ for each country j. To allow data from different countries to be pooled, we also divide this measure by its standard deviation for all forecasts $i, k \in \{1, 2, \dots, 12\}$ of data of type τ for each country j. There are 66 unique combinations of two out of the 12 models for the five different countries, which yields the 330 data points for each combined forecast of each type of data shown in the figure.

Fig. 3 shows that there is a clear negative relationship between the relative MSFE and the bias of the individual forecasts for bias-corrected optimally combined forecasts with shrinkage (BC-OWS).[19] When the forecasts are unbiased, bias correction and the optimal combination of the forecasts results in a larger MSFE than combining the forecasts with equal weights. However, as the absolute size of the bias grows, so does the accuracy of the bias-correction strategy. The COW forecasts shown in blue, on the other hand, have either no relationship or a slightly negative relationship with the bias of the underlying forecasts. As predicted by Theorem 2 in Section 2, the conditionally optimal-weights forecasts are more robust than bias-corrected optimal-weights forecasts in this respect. COW forecasts do not experience the same loss in forecast accuracy relative to equal weights when the biases are small or medium sized. COW provides a robust forecast that improves upon equal-weights forecasts in the overwhelming majority of cases tested, spanning different types of data from different countries.

8. Conclusion

We showed that when there is predictable information in forecast errors, a combined forecast should be constructed to minimize a conditional expected loss function. We proved that forecast combinations constructed in this way improve upon unconditional combinations commonly used in the literature and that the improvements are greater when more information becomes

[19] The shrinkage employed here follows the same specification used in Section 5.2.1 for US inflation.

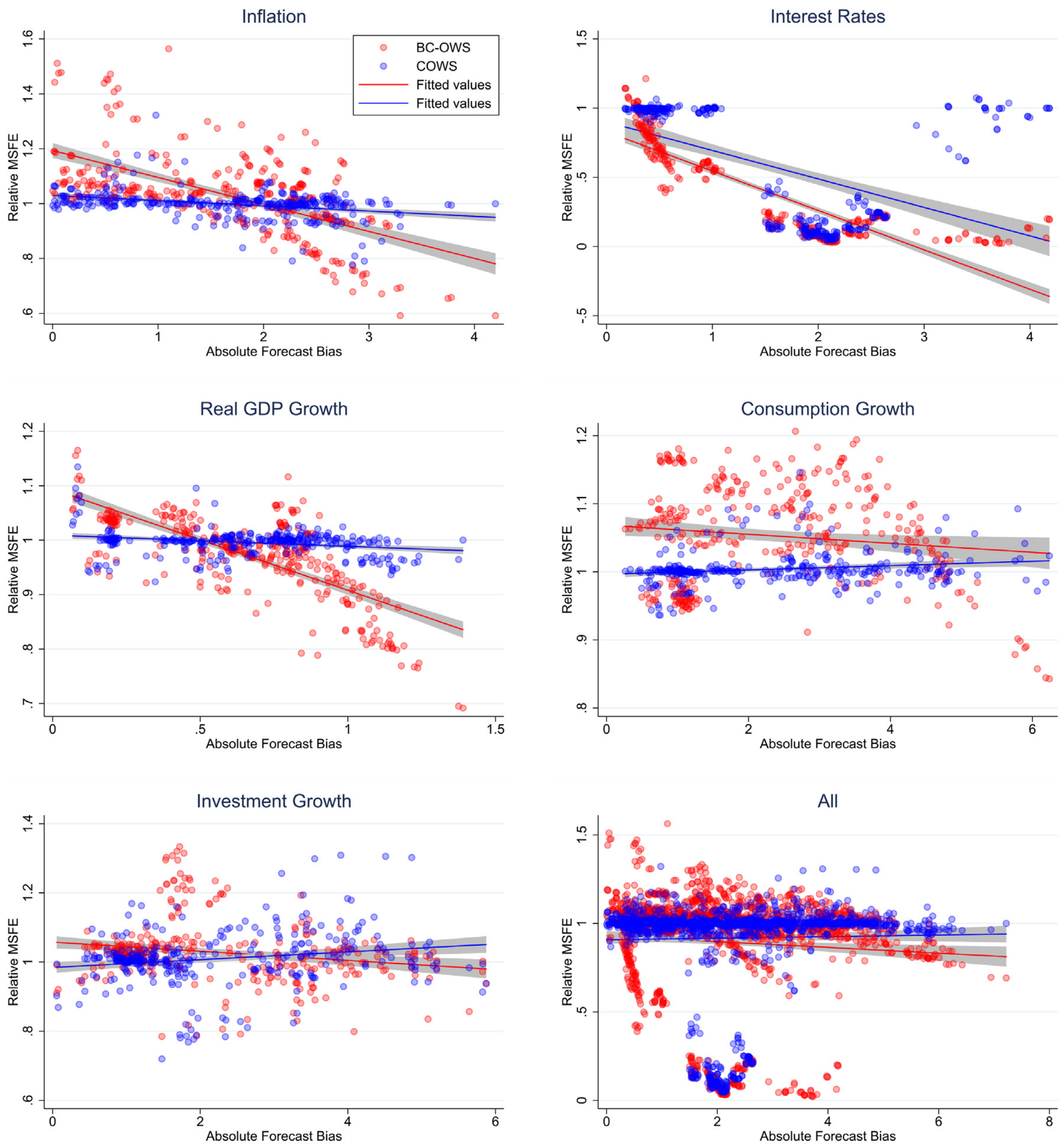

Fig. 3. Tradeoff between accuracy and bias for real-time forecasts of international data. *Notes:* Each dot represents the full sample (2000–2018) four-quarter-ahead MSFE for forecasts of data from one of the following countries: the US, Canada, the United Kingdom, Australia, or New Zealand. Bias-corrected optimally combined forecasts with shrinkage (BC-OWS) and conditionally optimal-weights forecasts with shrinkage (COWS) results are shown relative to the MSFE of an equal-weights forecast, which are plotted against a measure of the absolute forecast bias of the underlying forecasts that are combined. The absolute forecast bias measure is defined in the text. The line represents OLS regressions of relative MSFEs on the absolute forecast bias. The shaded area shows the 95% confidence intervals.

available. Our theoretical findings support forward-looking approaches to combining forecasts, where forecasts are weighted by their expected performance rather than their past performance. Our empirical results represent a proof of concept that forward-looking approaches work in practice.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Christopher Gibbs gratefully acknowledges funding support from the Australian Research Council Grant DP210101204.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2024.03.002.

References

Aiolfi, M., & Timmermann, A. (2006). Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics*, *135*(1–2), 31–53.

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

Atkeson, A., & Ohanian, L. E. (2001). Are Phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis Quarterly Review*, *25*(1), 2–11.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operations Research Quarterly*, *20*(4), 451–468.

Breitung, J., & Knüppel, M. (2021). How far can we forecast? Statistical tests of the predictive content. *Journal of Applied Econometrics*, *36*(4), 369–392.

Capistrán, C., & Timmermann, A. (2009). Disagreement and biases in inflation expectations. *Journal of Money, Credit and Banking*, *41*(2–3), 365–396.

Claeskens, G., Magnus, J. R., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, *32*(3), 754–762.

Clark, T., & McCracken, M. (2013). Advances in forecast evaluation. In G. Elliott, & A. Timmermann (Eds.), *Handbook of economic forecasting. Vol. 2B* (pp. 1107–1201). Horth-Holland: Elsevier.

Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, *138*(1), 291–311.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

Clements, M. P., & Hendry, D. F. (1996). Intercept corrections and structural change. *Journal of Applied Econometrics*, *11*(5), 475–494.

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

Davies, A. (2006). A framework for decomposing shocks and measuring volatilities derived from multi-dimensional panel data of survey forecasts. *International Journal of Forecasting*, *22*, 373–393.

Davies, A., & Lahiri, K. (1995). A new framework for analyzing survey forecasts using three-dimensional panel data. *Journal of Econometrics*, *68*, 205–227.

Diebold, F. X., & Lopez, J. A. (1996). Forecast evaluation and combination. *Handbook of Statistics*, *14*, 241–268.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 134–144.

Elliott, G. (2011). *Averaging and the optimal combination of forecasts*: *UCSD working paper*, available at https://econweb.ucsd.edu/~grelliott/AveragingOptimal.pdf.

Elliott, G., Gargano, A., & Timmermann, A. (2013). Complete subset regressions. *Journal of Econometrics*, *177*(2), 357–373.

Elliott, G., & Timmermann, A. (2004). Optimal forecast combinations under general loss functions and forecast error distributions. *Journal of Econometrics*, *122*, 47–79.

Elliott, G., & Timmermann, A. (2016). *Economic forecasting*. Princeton University Press.

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

Gaglianone, W. P., & Issler, J. V. (2015). *Microfounded forecasting*. Mimeo, available at http://hdl.handle.net/10438/13730.

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, *29*(1), 108–121.

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

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

Granger, C. W. J. (1989). Invited review combining forecasts—twenty years later. *Journal of Forecasting*, *8*(3), 167–173.

Granziera, E., & Sekhposyan, T. (2019). Predicting relative forecasting performance: An empirical investigation. *International Journal of Forecasting*, *35*, 1636–1657.

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, *13*(2), 281–291.

Issler, J. V., & Lima, L. R. (2009). A panel data approach to economic forecasting: The bias-corrected average forecast. *Journal of Econometrics*, *152*, 153–164.

Lima, L. R., & Meng, F. (2017). Out-of-sample return predictability: A quantile combination approach. *Journal of Applied Econometrics*, *32*(4), 877–895.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*(4), 802–808.

Matsypura, D., Thompson, R., & Vasnev, A. L. (2018). Optimal selection of expert forecasts with integer programming. *Omega*, *78*, 165–175.

Montero-Manso, P., Athanasopoulos, G., & Talagala, T. S. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*, 86–92.

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *545*(3), 703–708.

Orphanides, A., & van Norden, S. (2005). The reliability of inflation forecasts based on output gap estimates in real time. *Journal of Money, Credit and Banking*, *37*(3), 583–601.

Radchenko, P., Vasnev, A. L., & Wang, W. (2023). Too similar to combine? on negative weights in forecast combination. *International Journal of Forecasting*, *39*(1), 18–38.

Smith, J., & Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, *71*(3), 331–355.

Stock, J. H., & Watson, M. W. (2003). Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature*, *41*(3), 788–829.

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*(6), 405–430.

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

Stock, J. H., & Watson, M. W. (2009). Phillips curve inflation forecasts. In J. Fuhrer, Y. K. Kodrzycki, J. S. Little, & G. P. Olivei (Eds.), *Understanding inflation and the implications for Monetary policy* (pp. 101–186). Cambridge: MIT Press.

Stock, J. H., & Watson, M. W. (2010). *Modeling inflation after the crisis*: *NBER Working Paper Series, Working Paper 16488*.

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting, Volume 1. Handbook of economics. Vol. 24* (pp. 135–196). Horth-Holland: Elsevier.

Timmermann, A., & Zhu, Y. (2016). *Monitoring forecasting performance*: *UCSD working paper*.

Wallis, K. F. (2011). Combining forecasts—forty years later. *Applied Financial Economics*, *21*(1–2), 33–41.

Wallis, K. F., & Whitley, J. D. (1991). Sources of error in forecasts and expectations: UK economic models, 1984-88. *Journal of Forecasting*, *10*(3), 231–253.

West, K. D. (2006). Forecast evaluation. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting, Volume 1. Handbook of economics. Vol. 24* (pp. 99–134). Horth-Holland: Elsevier.

White, H. (2000). A reality check for data snooping. *Econometrica*, *68*(5), 1097–1126.

Zhang, X., & Zhang, X. (2023). Optimal model averaging based on forward-validation. *Journal of Econometrics*, *237*, Article 105295.