Averaging Impulse Responses Using Prediction Pools*

Paul Ho
Federal Reserve Bank of Richmond†

Thomas A. Lubik
Federal Reserve Bank of Richmond‡

Christian Matthes
Indiana University§

February 14, 2024

Abstract

Macroeconomists construct impulse responses using many competing time series models and different statistical paradigms (Bayesian or frequentist). We adapt optimal linear prediction pools to efficiently combine impulse response estimators for the effects of the same economic shock from this vast class of possible models. We thus alleviate the need to choose one specific model, obtaining weights that are typically positive for more than one model. Our Monte Carlo simulations and empirical applications illustrate how the weights leverage the strengths of each model by (i) trading off properties of each model depending on variable, horizon, and application and (ii) accounting for the full predictive distribution rather than being restricted to specific moments.

JEL Classification: C32, C52

Key Words: Prediction Pools, Model Averaging, Impulse Responses, Misspecification

*We are grateful to our editor Borağan Aruoba, the associate editor Christiane Baumeister, and an anonymous referee for comments that helped shape our paper. Mikkel Plagborg-Møller, Mark Watson, our discussants Gianni Amisano and Ed Herbst, and numerous workshop and seminar participants provided useful comments. Aubrey George, Colton Lapp, and Brennan Merone provided excellent research assistance. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

†Research Department, P.O. Box 27622, Richmond, VA 23261. Email: paul.ho@rich.frb.org.

‡Research Department, P.O. Box 27622, Richmond, VA 23261. Email: thomas.lubik@rich.frb.org.

§Wylie Hall, 100 South Woodlawn Avenue, Bloomington, IN 47405. Email: matthesc@iu.edu.

1 Introduction

Impulse responses are a key tool in macroeconomists' arsenal to trace out the effects of structural shocks on aggregate quantities and prices. When estimating these impulse responses, economists have a wide range of options. For example, a researcher can choose between local projections (LPs) and vector autoregressions (VARs), Bayesian and frequentist methods, and different specifications. Each choice has its own drawbacks and benefits, and these choices can generate significantly different results (see Ramey (2016) for several leading examples). While there is a growing literature discussing conditions under which one approach might be preferred over another (Stock and Watson, 2018; Herbst and Johannsen, 2020; Plagborg-Møller and Wolf, 2021), many of these conditions are difficult to verify in practical applications.

In this paper, we introduce a method to average impulse responses from different estimators by extending the optimal prediction pools studied by Geweke and Amisano (2011) and Hall and Mitchell (2007).[1] In particular, we compute the optimal weights that maximize the weighted average log score function for forecasts conditional on the structural shock of interest. This conditionality separates our approach from the literature. If a specific structural shock is important for forecasting variables of interest, our weights differ substantially from weights computed using traditional approaches that do not account for the shock. The only input required for our method is a set of forecast densities that trace out the model-specific effects of the shock of interest. Individual impulse responses can be based on any method that delivers such a conditional forecast density for a given variable at a given horizon.

Our approach is designed to appeal to empirical macroeconomists, who may find it difficult to choose between different methods for estimating impulse responses and who want to take into account uncertainty across methods. LPs have become popular because they allow for the introduction of extraneous variables in a straightforward manner. At the same

[1] Opinion pools, i.e., a forecast density formed by averaging over model-specific forecast densities, were first introduced by Stone (1961).

time, the confidence intervals of the LP-based responses tend to be wide. In contrast, VAR-based impulse response estimates suffer from a well-known bias when used to approximate moving average processes with many lags. Our proposed solution to these issues in practical applications is to take all of these concerns at face value and compute combined responses that take these trade-offs into account. Our use of prediction pools provides a systematic and computationally tractable method to account for these issues in a wide range of applications, where the focus is on identifying plausible and robust dynamic behavior over time, irrespective of the underlying models.

A key strength of our approach is its flexibility. In particular, it removes the necessity to choose one model or even one statistical paradigm. Moreover, the methodology is applicable to a wide range of models. Typical methods such as Bayesian model averaging are unavailable when one of the estimators considered is based on LPs as LPs are not 'generative models', that is, a set of LPs for different horizons do not form a consistent data-generating process. In addition to the aforementioned LPs and VARs, dynamic equilibrium models (Smets and Wouters, 2007), dynamic factor models (Stock and Watson, 2016), or single equation methods (Baek and Lee, 2022) can be used in our framework. Our method provides horizon- and variable-specific averages, thus exploiting each method's strength as much as possible.

As highlighted by Geweke and Amisano (2011), prediction pools have properties that make them well-suited to average over models or estimators when it is clear that all included models are misspecified. In contrast to Bayesian model averaging or related frequentist methods, more than one model will generally receive a positive weight. This helps prediction pools to outperform other model selection or model averaging approaches using various measures of forecast accuracy. Our extension inherits these properties, as we demonstrate with Monte Carlo exercises and with two empirical applications.

Prediction pools are also computationally straightforward to implement relative to alternative methods of averaging across models. Since each model-specific forecasting density can be obtained separately, the most-time consuming part of forming prediction pools can

be parallelized. The second step is then a relatively simple numerical maximization problem with a concave objective function and convex constraints. Other methods that also combine information from various models such as mixture models or composite-likelihood estimators (Qu, 2018; Canova and Matthes, 2021) do not share this modularity and thus have substantially higher computational complexity.

Overall, our paper highlights several broad messages for estimating impulse responses. The theoretical properties of individual models are not sufficient criteria for the choice of optimal weights in the prediction pool. Misspecified models can dominate correctly specified (or more flexible) models in finite samples. On the other hand, models that produce tighter estimates need not receive greater weight. The choice of models and their weights depend on the entire predictive distribution and not only the point estimates. While our examples focus on the mean and variance, higher-order moments or any other properties of the predictive distribution can be important more generally. Finally, we also find that the optimal weights on models depend on horizon, variable, and application, making it difficult to derive general guidelines or rules-of-thumb.

We illustrate our methodology using Monte Carlo experiments. The first is a stylized univariate example motivated by Herbst and Johannsen (2020), which serves as a proof of concept and implies reasonable optimal weights. They result in an average impulse response with a bias similar to one chosen by minimizing a squared error criterion. The second Monte Carlo compares a VAR and an LP in a setting where the VAR is misspecified, but the LP produces noisier estimates and has finite sample bias that is of the opposite sign from the VAR. While most of the weight is placed on the VAR, substantial weight is also placed on the LP, reducing the bias of the averaged impulse response relative to the VAR on its own, with the biases of the two models offsetting each other. In the appendix, we also consider a Monte Carlo exercise that simulates data from the DSGE model from Smets and Wouters (2007), illustrating the weighting scheme's ability to trade off bias and variance for a relatively realistic data-generating process. These exercises highlight how our approach

often gives positive weight to all competing models, but is also consistent with previous theoretical results found in Herbst and Johannsen (2020) and Li et al. (2022).

We then consider an empirical application that follows Ramey (2016), where we average across three models that use the same Romer and Romer (2004) narrative instrument for monetary shocks. We find a range of results depending on horizon and variable, emphasizing the flexibility of our methodology and the importance of considering the full predictive distribution rather than individual statistics. Notably, the averaged response rules out implausible features of the impulse response estimates from individual models.

Finally, we study the the response of unemployment to a total factor productivity (TFP) shock measured using the series from Fernald (2014). We consider three VARs, three LPs, a factor-augmented VAR (FAVAR), and a single equation regression model (as in Romer and Romer (2004)). While the estimated responses differ substantially across models, we find an averaged response that is consistent with Basu et al. (2006)—unemployment rises initially in response to the TFP shock, but rapidly returns to its long-run average.

Related Literature. Our approach is motivated by the vast array of choices for computing impulse responses available to practitioners. It allows researchers to average optimally across multiple approaches rather than choosing just one. The two main statistical models are VARs (Sims, 1980) and LPs (Jordà, 2005), which we focus on for most of the paper. Within these two classes of models there are numerous variations. For example, in VARs inference can be conducted using Bayesian or frequentist methods (Sims and Zha, 1999). The Bayesian approach requires the choice of priors (Doan et al., 1984; Del Negro and Schorfheide, 2004; Giannone et al., 2015) while the frequentist approach requires choices about bias correction and the construction of confidence intervals (Kilian, 1998; Pesavento and Rossi, 2006). With LPs, there is a growing literature providing choices on the approach to inference (Herbst and Johannsen, 2020; Montiel Olea and Plagborg-Møller, 2021; Lusompa, 2021; Bruns and Lütkepohl, 2022) and smoothing the impulse responses (Barnichon and Brownlees, 2019;

Ferreira et al., 2023).

Having a method that is flexible enough to cater to different models, variables, and horizons is particularly useful given the range of conclusions in the literature about the relative strengths of the different methods. Asymptotic results on the relative performance of VARs and LPs (Stock and Watson, 2018; Plagborg-Møller and Wolf, 2021) rely on conditions that may not be easily verifiable in practice. In finite sample settings, the literature has also compared the performance of VARs and LPs (Kilian, 1998; Marcellino et al., 2006; Li et al., 2022). However, it is difficult to draw general conclusions, especially in empirical applications when the true model is not known. Our Monte Carlos and empirical applications show that the relative weights on different models can vary substantially not only with the data but also by variable and horizon, consistent with results in the literature (Marcellino et al., 2006; Gürkaynak et al., 2013). We consider it therefore important to rely on a general method that is able to assign weights variable by variable and horizon by horizon.

Prediction pools have been used to average models since their introduction by Geweke and Amisano (2011) and subsequent follow-up work in Geweke and Amisano (2012) and Amisano and Geweke (2017). Our key innovation is that prediction pools can be used to average impulse responses by treating the impulse responses as conditional forecasts. This allows for a flexible method that inherits desirable properties of the original prediction pools.

Model averaging has a long tradition in economics, partially motivated by the observation that averages of forecasts across multiple models tend to outperform forecasts based on an individual model (Bates and Granger, 1969). In the Bayesian setting, model averaging is just an application of Bayes' theorem (for an application to VARs, see, for example, Strachan and van Dijk (2007)). As mentioned, Bayesian model averaging generally requires the use of generative models and, as such, rules out LPs. Frequentist versions of model or forecast averaging such as Hansen (2007) also focus on specific classes of models (averages of least squares estimators in that case). Hansen (2016) studies model combination of various restricted VARs estimated via least squares. He proposes to find optimal model weights to

minimize the mean squared error of a function of the VAR parameters (which can be an impulse response at a specific horizon). Dinh et al. (2023) propose averaging across different LP specifications, especially in the presence of a large number of possible controls. Unlike us, they do not solve for an optimal weight and do not consider models beyond LPs.

Outline. The rest of the paper is structured as follows. Section 2 introduces our methodology. Section 3 describes our Monte Carlo exercises. In Section 4, we apply our method to study the impulse responses to monetary and TFP shocks. Section 5 concludes.

2 Prediction Pools

We use prediction pools to average impulse responses across different models, based on Geweke and Amisano (2011). In their framework, predictive densities $p\left(z_{t+h}|X_m^t;\mathcal{M}_m\right)$ for each model $\mathcal{M}_m$ are combined to create a predictive density for an observable z_t conditional on model-specific predictive variables X_m^t, from which objects of interest, such as forecasts, can be computed.[2] The individual predictive densities are taken as given; that is, in contrast with other approaches to model averaging such as the estimation of mixture models, the parameters of the specific models and the model weights are not estimated jointly.

Formally, for any given horizon h, the goal is to maximize the log predictive score function:

$$\max_{\sum_{m=1}^{M} w_{m,h}=1, w_{m,h}\geq 0} \sum_{t=1}^{T} \log\left[\sum_{m=1}^{M} w_{m,h} p\left(z_{t+h}|X_m^t;\mathcal{M}_m\right)\right], \tag{1}$$

where z_{t+h} denotes the variable of interest, X_m^t denotes the history of variables that z_{t+h} depends on in model $\mathcal{M}_m$, and $m = 1, ..., M$ indexes different models. The framework can be extended to the multivariate case, where z_{t+h} is a vector of observables, but for ease of exposition and in our empirics later we find it useful to focus on one variable at a time.[3]

[2] Subscripts generally denote period-specific outcomes (except for the subscript m, which denotes the model at hand), whereas superscripts denote histories up to and including the period specified in the superscript.

[3] We treat the computation of the weights at different horizons as distinct problems. One could, alterna-

A key feature of prediction pools is that they generally improve forecasting ability relative to individual models as judged by the log predictive score (Geweke and Amisano, 2011, 2012). They do so by usually giving more than one model a positive weight, in contrast with posterior model probabilities in a Bayesian setting.

2.1 Adapting Prediction Pools to Impulse Response Averaging

We leverage the useful properties of prediction pools for the problem of impulse response estimation using the insight that impulse responses are nothing but conditional forecasts. The impulse responses in our framework are averages of model-specific impulse response estimators.[4] Our approach thus rewards models that forecast well.

The main feature distinguishing our approach from Geweke and Amisano (2011) is that we use a measure of the shock of interest as a conditioning argument in our predictive densities. In general, we form forecast densities that depend on observables up to time $t-1$ and a measure of the structural shock at time t. These measures of shocks can depend on time t data and model parameters. They can also incorporate identification restrictions.

This conditioning scheme makes our approach relevant for the empirical practice and choices that researchers are facing. An alternative approach would be to use the Geweke and Amisano (2011) approach for different reduced-form models directly and then impose identifying restrictions ex-post after finding the optimal weights. However, many applications of LPs directly use information on structural shocks (or instruments thereof) in the estimation, making this alternative less appealing when at least one of the models in our

tively, compute weights jointly and impose a penalty that forces the changes in the weights across horizons to be smoother than in our benchmark. For example, one could estimate the sets of weights $\{\{w_{m,h}\}_{m=1}^{M}\}_{h=1}^{H}$ by maximizing the following objective function:

$$\max_{\sum_{m=1}^{M} w_{m,h}=1, w_{m,h}\geq 0} \sum_{h=1}^{H}\sum_{t=1}^{T} \log\left[\sum_{m=1}^{M} w_{m,h} p\left(z_{t+h}|X_m^t;\mathcal{M}_m\right)\right] + \lambda \sum_{h=2}^{H}\sum_{m=1}^{M}(w_{m,h}-w_{m,h-1})^2,$$

where λ controls how much smoother the weights will be relative to our benchmark. With $\lambda = 0$, we replicate our benchmark since then each horizon's weights can be solved for independently of all other horizons.

[4] Our focus on this paper is on linear models, but our approach could also be used in nonlinear settings.

pool is based on LPs. Conditioning on the shock explicitly rewards models where the shock helps to forecast the variable of interest.

Another distinctive component of our approach is how we implement the distribution over each model's parameters, i.e., how our forecasting densities incorporate parameter uncertainty within a model. Geweke and Amisano (2011) use two approaches: The posterior distribution of parameters from a Bayesian estimation or fixed parameter values from some point estimate. We use a more general framework where the parameters of model $\mathcal{M}_m$ are collected in a vector Ω_m. We generate draws from a distribution $g_m(\Omega_m)$ that captures the parameter uncertainty we want to consider. This could be a posterior distribution, a point mass, a prior distribution, or a distribution derived using frequentist principles, say, by appealing to standard asymptotic arguments or numerical approaches such as the bootstrap.

We generally study a vector y_t of macroeconomic variables and denote the jth variable of that vector by $y_{t,j}$. Since LPs are usually estimated for one specific variable and horizon at a time, we carry out our analysis variable by variable and horizon by horizon as well. This also gives us additional flexibility as different models might have better forecasting ability for different variables or horizons.

With these definitions in hand, we define our optimization problem:

$$\max_{\sum_{m=1}^{M} w_{m,h}=1, w_{m,h}\geq 0} \sum_{t=1}^{T} \log \left[\sum_{m=1}^{M} w_{m,h} p_m^* \left(y_{t+h,j} \right) \right] \tag{2}$$

where, relative to Geweke and Amisano (2011), we simply replace the unconditional predictive density $p\left(z_{t+h}|X_m^t; \mathcal{M}_m\right)$ in (1) with the conditional predictive density p_m^* defined:

$$p_m^* \left(y_{t+h,j} \right) = \int p\left(y_{t+h,j} | y^{t-1}, \varepsilon_t(\Omega_m, y^t), \Omega_m, \mathcal{M}_m \right) g_m(\Omega_m) d\Omega_m. \tag{3}$$

As noted above, the key differences between p_m^* and $p\left(z_{t+h}|X_m^t; \mathcal{M}_m\right)$ are the dependence of p_m^* on the shock ε_t and the integration over the distribution g_m of the parameters Ω_m.

We can approximate the integral on the right-hand side of (3) by Monte Carlo methods,

as is often necessary in practice. We can extend our definition of p^* by allowing different models to depend on different right-hand side variables. While we allow the shock measure $\varepsilon_t(\Omega_m, y^t)$ to be model-specific, in our applications we use the same shock (or instrument of a shock as a conditioning argument) in all models and assume that the observed shock is one element of the vector y_t.

Geweke and Amisano (2011) use true out-of-sample forecast densities, i.e., their densities $p\left(z_{t+h}|X_m^t;\mathcal{M}_m\right)$ are generally re-estimated every period. While this is possible in our framework as well, we use an alternative approach inspired by cross validation (Hastie et al., 2009). In particular, we split the sample in half and estimate the models for each subsample separately. We then use these estimates to obtain implied out-of-sample forecast densities for the parts of the sample that were not used for estimation. Combining the two subsamples gives us out-of-sample predictive densities for the entire sample, which we then use to obtain a single set of weights.

More specifically, we first estimate each model using the first half of the sample, and then use those parameter estimates to forecast the second half. In the next step, we estimate based on the second subsample, fix parameter estimates and forecast the first subsample. This produces the necessary out-of-sample forecast densities for each subsample and, therefore, each period, without having to re-estimate every period. We view this approach as trading off computing time and overfitting concerns, which would play a role if we did not split the sample at all. While the sample splitting can increase estimation time, it does not change the time taken to compute the weights since we solve for a single set of weights using the full sample of predictive densities.

More generally, one could split the sample into $S > 2$ subsamples and compute the predictive densities for each subsample using the remaining $S-1$ subsamples to estimate the parameters. The case of $S = T$ is analogous to leave-one-out cross validation, where an observation is predicted using all other data points. The Geweke and Amisano (2011) approach takes $S = T$ but only uses data from subsamples (or, equivalently, observations)

$1, .., s-1$ to obtain parameter estimates and predictive densities for observation t.

2.2 Properties of Prediction Pools

With forecast densities (3) in hand, the theorems stated in Geweke and Amisano (2011) all apply. In particular, as long as the expected average forecast densities do not take on the same value for different models, the true model asymptotically receives a weight of 1 if it is contained in the set of models we consider. In contrast to Bayesian model averaging, more than one model will receive positive weight even asymptotically if the true model is not contained in the set of weights (Geweke and Amisano, 2012). The individual model with the highest log predictive score might not even receive a positive weight in the optimal pool if more than two models are being considered. The pooling weights thus do not necessarily represent a ranking or evaluation of the models. Rather, the weights are chosen to optimize the performance of the *averaged* model in terms of the log-score objective function. Furthermore, the weights satisfy a number of consistency requirements that make their use appealing. We state these consistency requirements as derived by Geweke and Amisano (2011) in Appendix A.

The prediction pool framework is particularly well suited for applications in empirical macroeconomics. First, by studying each horizon separately, we overcome the issue that LPs are not generative models. In particular, there is no unique way to simulate a sample of arbitrary length from LPs estimated using different horizons. The simulation from one horizon is in general inconsistent with simulations from LPs for a different horizon. As a consequence, Bayesian model averaging is not possible. Second, prediction pools allow us to compare Bayesian and frequentist approaches. In particular, the probability distribution $g_m(\Omega_m)$ can be either Bayesian (i.e., a posterior distribution) or frequentist (i.e., an asymptotic distribution).[5] Finally, the optimization problem (2) is computationally straightforward.

[5]By allowing for both Bayesian and frequentist models to enter our model pool, we implicitly equate the interpretations of uncertainty in Bayesian and frequentist frameworks. This is in the spirit of much applied work, which compares error bands across Bayesian and frequentist approaches, disregarding philosophical

Despite the flexibility of our framework, there are some notable limitations, all of which apply more generally to model averaging techniques. First, in averaging across impulse responses, a researcher has to entertain the assumption that each response corresponds to the same shock. This may be a controversial assumption when comparing responses identified by different instruments if one views each instrument as identifying a different shock.[6] Second, the optimal weights are not informative about whether exogeneity assumptions are satisfied. Instead, arguments about such exogeneity require economic theory that is not reflected in the predictive densities. Next, when comparing Bayesian and frequentist estimates, one should be mindful of the differing interpretations of the respective credible or confidence intervals. While these are important caveats, we consider our approach as reflective of actual practice in much applied macroeconomic research.

In addition, while the predictive density p_m^* conditions on the shock of interest, it can also depend on other features of the model. By conditioning on the shock, our approach already improves on typical model averaging or comparison methods (e.g., Bayesian model averaging, Akaike information criterion, or the original Geweke and Amisano (2011) approach), which focus on the overall performance of the model without explicitly account for the shock of interest. Nevertheless, Section 2.5 shows how to modify the objective function to reward models in which the structural shock *improves* forecasting ability, hence putting more emphasis specifically on the shock.

Finally, a key question is how many models should be included in the prediction pool and along which dimensions they should differ. Heuristically, and practically, applied researchers would likely consider one LP and one VAR specification. That being said, Geweke and Amisano (2011) show that in population, models in the prediction pool that are not informative (i.e., are dominated by the other models), receive zero weight. Consequently, one decision rule would be to stop adding to the pool when the additional model's estimated

differences between the two approaches. In our applications below, we do not compare across paradigms.

[6]For instance, McKay and Wolf (2023) argue that the instruments in Romer and Romer (2004) and Gertler and Karadi (2015) identify different types of monetary shocks.

weight is below some threshold close to zero. Alternatively, a good pool of models is one where the forecast density is properly calibrated. Amisano and Geweke (2017) report on a battery of tests to check whether a pool is large enough along this dimension, while Hansen et al. (2011) use a sequential testing approach to compute the model confidence set.

2.3 Implementation

We now present a step-by-step guide that summarizes our approach.

1. **Partition:** Split the estimation sample in half, so that each subsample has $T/2$ observations (we assume for simplicity that T is even). We denote the subsample by $s = 1, 2$, where $s = 1$ means that periods dated $t = 1, ...T/2$ are used in the estimation, whereas $s = 2$ means that periods $t = T/2 + 1, ..., T$ are used. In a slight abuse of notation, we define a function $s(t)$ that is equal to 1 if $t \leq T/2$ and equal to 2 if $t > T/2$. We now give densities additional superscripts that denote the estimation sample.

2. **Estimation:** Estimate (or calibrate) each model $m = 1, ..., M$ for each subsample s. This means that for each model we get a distribution $p_m^s(\Omega_m)$ for each subsample.

3. **Predictive densities:** For each model and subsample construct $p_m^{*,s}\left(y_{t+h,j}\right)$ by first constructing the forecast density conditional on parameters and a given shock (see Section 2.4 for an example on how to do this in VAR models). Then average over draws from the relevant $p_m^s(\Omega_m)$ density. Doing this for each subsample gives us a full set of predictive densities for $t = 1$ to T.

4. **Prediction pool weights:** Compute model weights by solving the following maximization problem for each horizon h and each variable j separately:

$$\max_{\sum_{m=1}^{M} w_{m,h}^j = 1, w_{m,h}^j \geq 0} \sum_{t=1}^{T} \log \left[\sum_{m=1}^{M} w_{m,h}^j p_m^{*,3-s(t)}\left(y_{t+h,j}\right) \right] \tag{4}$$

The superscript of the density $p_m^{*,3-s(t)}$ clarifies that we use out-of-sample forecasts to construct the objective function, but the first summation goes from $t = 1$ to T, showing that we use the full sample of predictive densities to compute a single set of weights.[7]

5. **Average impulse responses:** With model weights in hand, we can construct weighted averages of impulse responses and other statistics of interest from each model. First, re-estimate each model using the entire sample to obtain a final estimate of $g_m(\Omega_m)$ and use that distribution to construct our statistics of interest. Next, take draws $k = 1, ..., K$ from the set of models $m^k \in \{1, ..., M\}$ with probabilities given by the weights. Finally, for each draw, k, take a draw of the impulse response from model m^k. This gives us K draws from the averaged impulse response, which we can then use to compute moments or quantiles.

Steps 2 and 3 are the most time-consuming steps of the algorithm, but they can both be parallelized across models.

2.4 Illustrative example: Constructing p^* for a VAR(1)

For concreteness, we now illustrate how to construct the forecasting density $p_m^*(y_{t+h,j})$ in the context of a linear Gaussian VAR(1):

$$y_t = By_{t-1} + u_t \tag{5}$$

$$u_t = C\varepsilon_t, \tag{6}$$

where $\varepsilon_t \sim \mathcal{N}(0, I)$ is a vector of structural shocks and $V[u_t] = CC'$. In terms of the notation from the previous section and assuming this VAR is model 1, we have $\Omega_1 = [vec(B)'\ vec(C)']'$, where vec denotes columnwise vectorization of a matrix. The impulse response of y_t to shock j at horizon h is then $B^h C_{\bullet,j}$, where $C_{\bullet,j}$ is the jth column of the matrix C.

[7] Geweke and Amisano (2011) provide conditions for the concavity of the objective function.

Given B and C, we can compute the on-impact conditional distributions:

$$E\left[y_t \mid y_{t-1}, \varepsilon_{t,j}\right] = By_{t-1} + C_{\bullet,j}\varepsilon_{t,j} \tag{7}$$

$$V\left[y_t \mid y_{t-1}, \varepsilon_{t,j}\right] = CC' - C_{\bullet,j}C'_{\bullet,j} \tag{8}$$

and iterate forward:

$$E\left[y_{t+h} \mid y_{t-1}, \varepsilon_{t,j}\right] = BE\left[y_{t+h-1} \mid y_{t-1}, \varepsilon_{t,j}\right] \tag{9}$$

$$V\left[y_{t+h} \mid y_{t-1}, \varepsilon_{t,j}\right] = BV\left[y_{t+h-1} \mid y_{t-1}, \varepsilon_{t,j}\right]B' + CC' \tag{10}$$

The predictive density of the vector y_t *conditional on parameters* h periods ahead is then Gaussian with conditional means and variances defined above. Furthermore, the forecast distribution of a specific variable $y_{t,j}$ conditional on parameters and the shock is given by a normal distribution where the mean and variance are the relevant elements of $E\left[y_{t+h} \mid y_{t-1}, \varepsilon_{t,j}\right]$ and $V\left[y_{t+h} \mid y_{t-1}, \varepsilon_{t,j}\right]$.

With the internal instrument VAR (Noh, 2018; Plagborg-Møller and Wolf, 2021), which we use in our Monte Carlos and empirical applications in Sections 3 and 4, an econometrician observes the shock if she knows the parameters. More generally, we replace $\varepsilon_{t,j}$ with $\hat{\varepsilon}_{t,j}$, the jth element of $\hat{\varepsilon}_t \equiv C^{-1}(y_t - By_{t-1})$, the fitted value of ε_t.

If B and C are estimated, we can account for parameter uncertainty by integrating over their posterior or asymptotic distribution. We implement this by averaging the predictive density across draws in a Bayesian framework, for example.

It is instructive to compare the forecast errors for the case where we condition on $\varepsilon_{t,j}$ with the case where we do not. In the latter, the forecast error is given by $y_t - By_{t-1}$. In the former, there is an additional adjustment due to knowledge of a structural shock, which leads to a forecast error of $y_t - By_{t-1} - C_{\bullet,j}\varepsilon_{t,j}$. Since the impulse response depends not only on the autocorrelation structure summarized by the By_{t-1} term, but also the effect of the shock on impact captured by the additional $C_{\bullet,j}\varepsilon_{t,j}$ term, it is important to condition on the

shock when computing predictive densities as input into the optimal weights.

The special case of an internal instrument VAR with two variables, i.e., an instrument ordered first followed by one endogenous variable, provides further intuition for how conditioning on the shock might matter. With B and CC' known, and the internal instrument assumption that C is lower triangular, we can show:

$$C_{\bullet,1}\varepsilon_{t,1} = \begin{bmatrix} 1 \\ \beta \end{bmatrix} u_{t,1} \text{ with } \beta \equiv \frac{\sigma_{21}}{\sigma_1^2}, \tag{11}$$

where σ_{21} is the covariance of the reduced form errors and σ_1^2 is the variance of the first reduced form error. The forecast error for the endogenous variable is thus $\begin{bmatrix} -\beta & 1 \end{bmatrix} u_t$, when conditioning on the shock, and $u_{t,2}$, when we do not condition on the shock. Once we condition on the shock, the forecast error for the endogenous variable is thus altered by $-\beta u_{t,1}$, where the parameter β captures how well the forecast error of the instrument can predict that of the endogenous variable.

How this difference propagates to longer horizons depends on the autoregressive coefficient B, as reflected by equation (9). Denote the (i,j) element of B^h by $b_{ij}^{(h)}$. The difference between the h period ahead conditional expectation for the endogenous variable with and without conditioning on the shock is $(b_{21}^{(h)} + b_{22}^{(h)}\beta)u_{t,1}$. This expression reveals two channels, through which conditioning on the shock may be important. First, the $b_{21}^{(h)}$ term captures how the persistent response of the instrument to the initial shock spills over to the endogenous variable. Second, the $b_{22}^{(h)}\beta$ term captures how the shock continues to have an effect at horizon h through the persistence of the endogenous variable.

2.5 Extensions

It is straightforward to extend our methodology to several common settings.

In many scenarios, macroeconomists use identification schemes that do not point identify the structural shock of interest (such as in the case of sign restrictions in VARs). To accom-

modate such cases, we can enlarge the parameter vector Ω_m for any model m, where the structural shock is not point-identified, to include a parameter that selects one possible value of the structural shock consistent with the other parameters of the model. In a VAR, this would be a rotation matrix that maps the covariance matrix of the one-step ahead forecast error into the matrix of impact impulse responses. While this parameter is by definition not point identified, it does not conflict with our approach.

Similarly, it is numerically straightforward to accommodate models with nonlinearities where the models are conditionally linear and Gaussian. Key examples are VAR models with parameters that follow discrete (Sims and Zha, 2006) or continuous (Cogley and Sargent, 2005; Primiceri, 2005) Markov processes, where the respective innovations are independent of other innovations in the model. In these cases, we need to enlarge the parameter vector Ω_m to include estimates of the time t state of the Markov process.[8]

Our approach typically gives larger weights to models that are better at forecasting the series of interest at a given horizon. This forecasting ability can be due to the inclusion of the structural shock or due to other features of each model. If a researcher wants to reward models with a larger weight when the inclusion of the structural shock *improves* forecast ability, the following alternative objective function could be used:

$$\underbrace{\sum_{t=1}^{T} \log \left[\sum_{m=1}^{M} w_m p_m^* \left(y_{t+h,j} \right) \right]}_{\text{standard objective}} + \underbrace{\phi \left[\sum_{t=1}^{T} \log \left[\sum_{m=1}^{M} w_m \left(p_m^* \left(y_{t+h,j} \right) / \overline{p_m} \left(y_{t+h,j} \right) \right) \right] \right]}_{\text{reward for forecast improvement due to structural shock}}, \tag{12}$$

where we define

$$\overline{p_m} \left(y_{t+h,j} \right) = \int p \left(y_{t+h,j} | y^{t-1}, \Omega_m, \mathcal{M}_m \right) g_m(\Omega_m) d\Omega_m \tag{13}$$

as the forecast density based on model $\mathcal{M}_m$ when the structural shock is not used as a predictive variable. The parameter ϕ governs how much the researcher rewards forecast

[8] We can also extend our approach to allow for time-varying weights along the lines of Waggoner and Zha (2012) or Del Negro et al. (2016).

improvement due to the inclusion of a structural shock. For simplicity, we set $\phi = 0$ and ignore the forecast improvement from the structural shock, as is typical in most model averaging in the literature. To assess the plausibility of this modification, one can also study how the weights change when we replace p_m^* with $\overline{p_m}$ in our baseline problem (2), as we do below. How much these weights differ is application-specific and depends on how important the shock of interest is for the evolution of the variables a researcher studies.

3 Monte Carlo Simulations

We now present Monte Carlo exercises to illustrate our methodology. First, we consider a univariate example with two alternative models that produce consistent estimates but differ in finite sample. Second, we consider a model in which the VAR is misspecified but the LP produces consistent estimates. In Appendix C, we also consider data simulated from a DSGE model, such that both the VAR and LP are misspecified. We report biases and standard deviations of our approach vis-a-vis individual models, following the common focus of the literature on first and second moments. However, our approach targets the entire forecast distribution and is not restricted to these moments.

3.1 AR(1)

As an initial proof of concept, we first consider the AR(1) Monte Carlo exercise from Herbst and Johannsen (2020). We show that in this setting, our model averaging approach performs close to optimally on a number of dimensions.

Data-Generating Process. We generate data from the univariate model:

$$y_t = \rho y_{t-1} + v_{1,t} + v_{2,t}, \tag{14}$$

where $(v_{1,t}, v_{2,t})' \overset{iid}{\sim} \mathcal{N}(0, I)$. We take $\rho = 0.95$ and use $T = 150$ observations. We seek the impulse response of y_t to a shock $v_{1,t}$.

Models. We estimate models of the form:

$$y_{t+h} = \beta_m^{(h)\prime} x_{m,t} + \varepsilon_{m,t+h}^{(h)}, \tag{15}$$

and compare the two specifications for $x_{m,t}$ considered by Herbst and Johannsen (2020):

- **With Controls:** $x_{m,t} = (v_{1,t}, y_{t-1})'$.
- **Without Controls:** $x_{m,t} = v_{1,t}$.

Both specifications produce consistent estimates $\beta_{m,1}^{(h)}$ of the impulse response at horizon h. However, the second specification does not control for lagged y_t, resulting in differing finite sample performances between models. Herbst and Johannsen (2020) show that the two specifications produce different finite sample biases. The variances of the estimated impulse responses also differ. Appendix B.1 shows that adding a VAR(1) into the pool of models does not materially change the relative weights on each of these models.

Results. Figure 1 shows the results averaged across 5×10^4 simulations. The weights produced are intuitive and perform well on a number of dimensions. The top left panel shows that optimal weights tend to favor the model with a smaller bias. The weights are closer to 0.5 when the biases of the two models are closer.

The remaining panels show that the resulting mixture model performs well. First, the bias of the mixture model is close to the optimum that one could get with each individual model horizon by horizon. Second, the standard deviation of the mean estimate from the mixture model is also close to the lower envelope of the two individual models. Finally, the mixture model generally performs as well as the better model in terms of coverage as well.

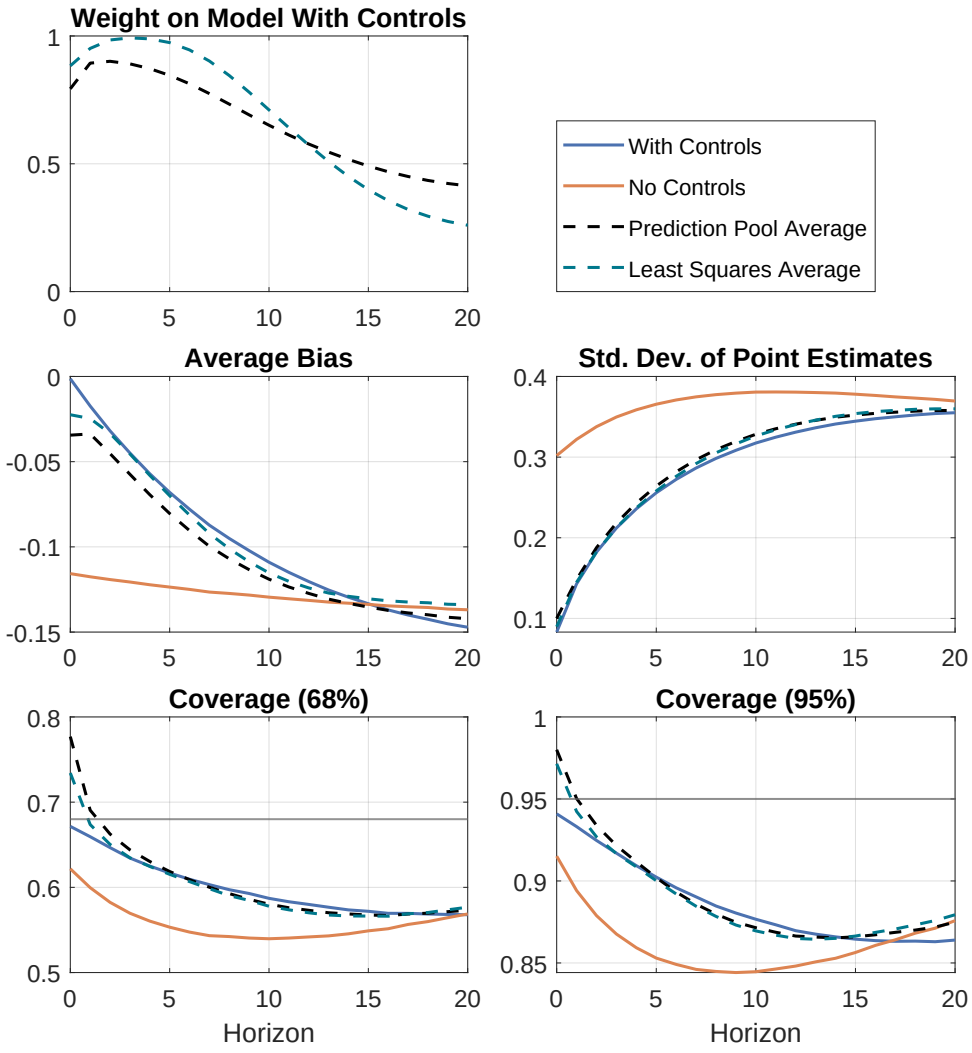

Figure 1: Prediction pool weights, biases, standard deviations, and coverage from Monte Carlo with AR(1) model from Herbst and Johannsen (2020). **Top left:** Prediction pool and least-squares weights on model with controls; **Middle left:** Bias of impulse responses under each specification and averaged model; **Middle right:** Standard deviation of point estimate of impulse response estimates; **Bottom:** Coverage of equal-tailed 68% (left) and 95% (right) error bands. Dashed lines correspond to individual models, solid lines correspond to averaged model. All plots show averages across all Monte Carlo repetitions.

In Appendix B.2, we specifically show that the mixture model improves more substantially on the coverage of the individual models once we introduce omitted variable bias.

We also compare the results to optimal weights computed using a least-squares objective function, which replaces the optimization problem in equation (2) with:

$$\min_{\sum_{m=1}^{M} w_{m,h}=1, w_m, h \geq 0} \sum_{t=1}^{T-h} \left(y_{t+h} - \sum_{m=1}^{M} w_m \hat{y}_{m,t+h}^{(h)} \right)^2, \tag{16}$$

where $\hat{y}_{m,t+h}^{(h)} = \beta_m^{(h)\prime} X_{m,t}$ is the fitted value of y_{t+h} in model m. We use the same sample-splitting scheme as with the prediction pool weights.[9]

Even though the least-squares objective function directly targets the bias and the standard deviation of the averaged point estimates, the prediction pool performs similarly on both these measures. The prediction pool is thus able to obtain close to optimal point estimates according to this criterion while taking into account the entire probability distribution for the estimated impulse response in each simulation. In situations where the forecasting density is more complicated, this is not necessarily guaranteed to be true and weights based on such a least squares objective could miss important features of the data.

3.2 Misspecified Shock

We now present an example in which the VAR is misspecified but the LP produces consistent estimates. The weights trade off the flexibility of the LP and the structure and relatively tighter estimates of the VAR. In addition, the two models produce impulse responses with biases of opposite signs in finite sample, offsetting each other once we average them.

[9] Even in the case of Gaussian predictive densities, a comparison of this least squares objective function with our benchmark is not straightforward because our objective then computes the natural logarithm of a sum of Gaussian densities, which cannot be written as a weighted sum of the logarithms of Gaussian densities. Furthermore, in practice, we use a bootstrap procedure to construct the predictive densities in this section. We describe this procedure in Appendix E.

Data-Generating Process. We consider data generated from a model similar to (14), but where the shock is AR(1) instead of iid:

$$y_t = \rho y_{t-1} + v_{1,t} + v_{2,t} \tag{17}$$

$$v_{2,t} = \gamma v_{2,t-1} + e_{2,t}, \tag{18}$$

where

$$\begin{bmatrix} v_{1,t} \\ e_{2,t} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1-\gamma^2 \end{bmatrix} \right).$$

Our parameterization ensures that the long-run variance of both $v_{1,t}$ $v_{2,t}$ are 1. We take $(\rho, \gamma) = (0.97, 0.75)$ and simulate the model over 250 periods. We seek the impulse response of y_t to a shock $v_{1,t}$, which is ρ^h at horizon h under the data-generating process.

Models. The first model we estimate is an internal instrument VAR:

$$\begin{bmatrix} z_t \\ y_t \end{bmatrix} = B \begin{bmatrix} z_{t-1} \\ y_{t-1} \end{bmatrix} + u_t, \tag{19}$$

where z_t is the shock of interest and u_t is assumed to be independent over time. The impulse response at horizon h is $B^h C_{\bullet,1}$, where C is the lower triangular matrix satisfying $CC' = V[u_t]$, obtained using a Cholesky decomposition.[10] As before, $C_{\bullet,1}$ is the first column of C. We assume for simplicity that the shock is perfectly observed, i.e., $z_t = v_{1,t}$. We estimate the model equation by equation using least squares, with standard errors computed using the "wild" bootstrap (Gonçalves and Kilian, 2004).

The second model we consider is an LP:

$$y_{t+h} = \beta^{(h)} v_{1,t} + \gamma_v^{(h)} v_{1,t-1} + \gamma_y^{(h)} y_{1,t-1} + \varepsilon_{t+h}^{(h)}. \tag{20}$$

[10] This is closely related to the VARX model from Bagliano and Favero (1999) and Paul (2020), where the instrument is included as an exogenous variable in a VAR.

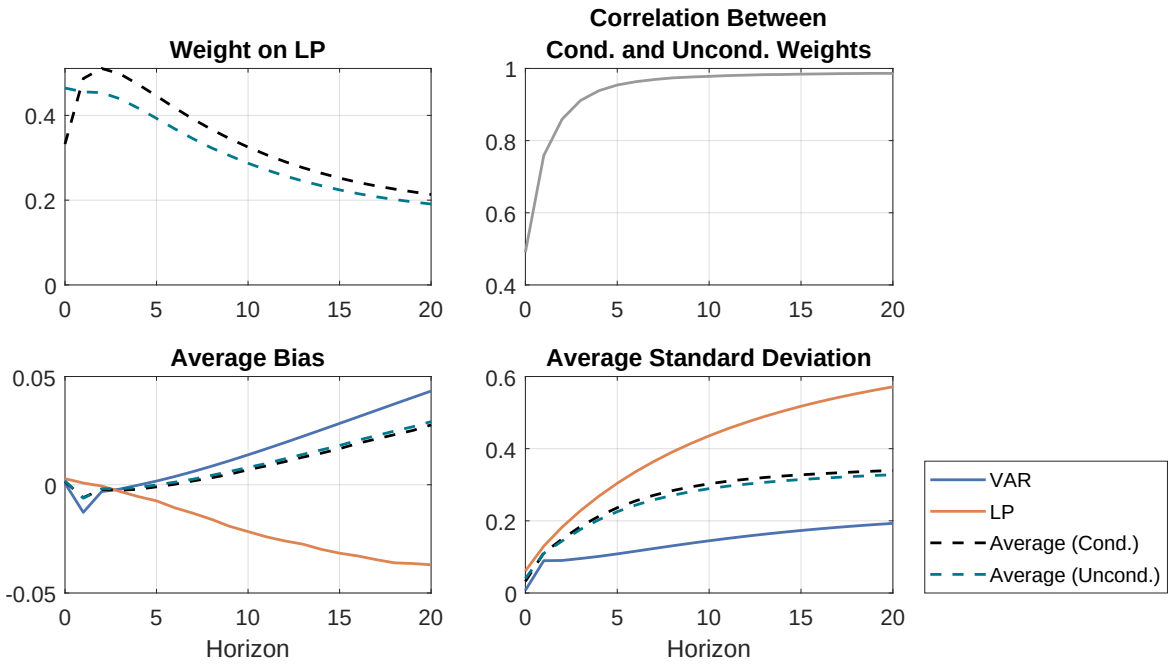

Figure 2: Prediction pool weights, biases, and asymptotic standard deviations from Monte Carlo with persistent shocks. **Top left:** Optimal weights on LP; **Top right:** Correlation between conditional and unconditional weights across Monte Carlo repetitions; **Bottom left:** Bias of impulse responses; **Bottom right:** Standard deviation of impulse responses. Weights, bias, and standard deviation averaged across all Monte Carlo repetitions.

The estimated impulse response at horizon h is $\beta^{(h)}$. The model is estimated using least squares, with White standard errors (Montiel Olea and Plagborg-Møller, 2021).

The two models face a bias-variance trade-off highlighted by Li et al. (2022). The VAR (19) is misspecified because the autocorrelation of the shock u_t is assumed to be zero. This induces bias even asymptotically. The LP produces consistent estimates, with a finite sample bias that vanishes as the sample size goes to infinity. However, the structure of the VAR induces a smaller variance than the LP. Our averaging approach balances both considerations while also taking into account the finite sample performance of each method.

Results. The results, averaged across 2.5×10^4 simulations, are summarized in Figure 2. While the majority of the weight is placed on the VAR, there is substantial weight of up to 0.51 placed on the LP. The weight on the LP peaks around $h = 3$, but remains above 0.2 for all horizons after impact. By averaging the two models, we obtain an impulse response that

has a lower standard deviation and only slightly larger bias than the LP. Since the VAR and LP have biases of opposite signs, averaging them tends to offset each other. In this case, the difference in standard deviations leads to a larger weight on the VAR.[11]

More generally, a correctly-specified or more flexible model need not dominate a misspecified model in finite sample. The finite sample performance of each model may not correspond to their asymptotic behavior. Furthermore, these properties may differ across impulse response horizon or the variable of interest. Our impulse response averaging approach flexibly accounts for these by constructing an optimal composite impulse response variable by variable and horizon by horizon. Figure 2 also highlights the trade-off between bias and standard deviation that is present in practically any model-averaging exercise (unless one model dominates in terms of both bias and standard deviation). Our approach reduces the bias relative to the VAR, but does so by increasing the standard deviation.

To highlight the role of conditioning on the shock, we also compare the weights from our methodology to those implied by problem (1), as initially proposed by Geweke and Amisano (2011). On impact, the weights (averaged across Monte Carlo repetitions) are noticeably different—the unconditional weight on the LP is 0.46 while the one that conditions on the shock is 0.33. When we look across simulations, the correlation between the conditional and unconditional weights is 0.49, suggesting that the two tend to differ in a given sample. For longer horizons, the weights become more similar and the correlation across simulations increases to close to 1. Intuitively, at short horizons, the shock plays a larger role in the conditional forecast. Since the shock is transitory, the forecast at longer horizons depends more on the autocorrelation structure of y_t, which is already captured by the unconditional forecasts. This is not at odds with our goal—the shape of the impulse response itself also depends more on these autocorrelations at longer horizons.

As further evidence of the behavior of the optimal weights in empirically relevant settings,

[11] We compute the standard deviation of the impulse responses for each Monte Carlo sample and then average across all samples. Both the sample specific bias and standard deviation depend on the estimated weights for that sample. The averages we report in our figure for the Monte Carlo experiments thus take into account sample variation in the estimated weights.

Appendix C reports a Monte Carlo exercise data simulated from the New Keynesian model of Smets and Wouters (2007). The results there highlight features of prediction pool that may underlie the empirical results here. First, the flexibility to trade off bias and variance concerns variable by variable and horizon by horizon allows one to make full use of the relative strengths of each model. Second, even when models have similar asymptotic properties, there can be substantial gains from averaging over them in finite sample. In particular, the bias of the average impulse response can be lower than that of any individual model.

4 Empirical Applications

We now apply our methodology using actual data to estimate impulse responses to monetary and TFP shocks. Overall, our applications indicate that prediction pools offer a plausible assessment of the dynamic effects of the shocks as they optimally resolve the bias-variance trade-off, especially when, as is likely, the underlying models are misspecified.

4.1 Monetary Shock

Our first empirical application follows the study of the Romer and Romer (2004) shocks in Ramey (2016). We use monthly data on the log of industrial production (IP), the unemployment rate, the log of the consumer price index (CPI), the log of a commodity price index, the federal funds rate, and the Romer and Romer (2004) instrument for March 1969 through December 1996 as endogenous variables.

We consider three models, each estimated using frequentist methods:

1. **Internal Instrument VAR:** We estimate a VAR with the Romer and Romer (2004) instrument as the first variable as in equation (19), followed by the endogenous variables. The monetary shock is assumed to be the first shock from a Cholesky decomposition.

2. **LP With Contemporaneous Controls:** We follow Ramey (2016) and estimate regressions of the form

$$z_{t+h} = \alpha_h + \theta_h \cdot \text{shock}_t + \text{control variables} + \varepsilon_{t+h}, \tag{21}$$

where z_{t+h} is the variable of interest and shock_t is the Romer and Romer (2004) instrument. The control variables include lags of the Romer and Romer (2004) shock and the endogenous variables, as well as contemporaneous values of all endogenous variables except the federal funds rate.

3. **LP With Lagged Controls Only:** This is identical to the previous LP specification, except that we do not control for contemporaneous variables. We rely on the assumption that the Greenbook forecasts used by Romer and Romer (2004) already include all information used by the Fed for setting interest rates.[12]

Following Ramey (2016), the VAR uses twelve lags and both LPs use two lags. All three models aim to estimate the same impulse response using the same instrument. However, the models make different assumptions, include different controls, and have different numbers of lags. Importantly, this means that the LPs do not nest the VARs.

Results. The results are summarized in Figure 3. The averaged impulse responses are reported in black with gray error bands - they are the same in each of the three plots that show the individual model impulse responses for a given variable. Overall, average impulse responses show a contraction, peaking around the one-year horizon, followed by a return to trend. Inflation displays a small initial increase followed by an insignificant response.[13]

[12] Such an assumption is not necessarily valid, which would result in a misspecified model. For instance, Aruoba and Drechsel (2023) find evidence that Greenbook forecast errors are correlated with economic information available to the Fed. Nevertheless, as pointed out in the Monte Carlo exercise of Section 3.2, our approach appropriately downweights misspecified models to the extent that their conditional forecasts do not perform as well as the other models in the pool.

[13] Aruoba and Drechsel (2023) find similar results in a VAR with additional controls. They use an alternative text-based instrument constructed with natural language processing and machine learning techniques.

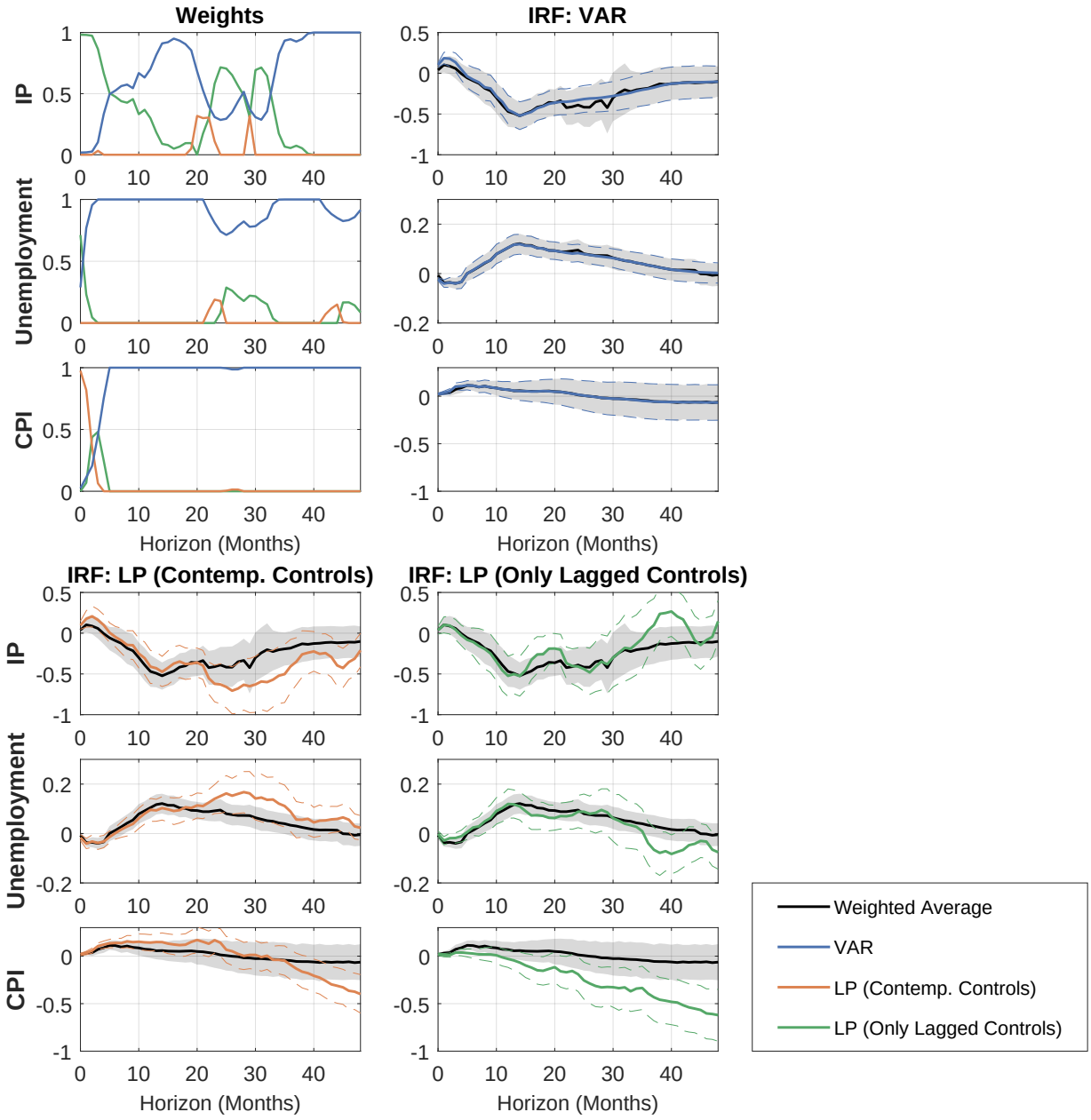

Figure 3: Prediction pool weights and impulse responses from monetary shock empirical application. **Top left:** Optimal weights on each model; **Top right, bottom:** Impulse responses and 68% error bands.

While the three models have fairly similar responses at shorter horizons, the impact of the averaging becomes more notable as the responses begin to diverge at longer horizons. In general, a majority of weight is placed on the LPs on impact, while the VAR gets assigned successively greater weight in the following periods. Appendix C shows a similar result in a Monte Carlo with data simulated from a DSGE model. Nevertheless, there are substantial differences in the weights across variables, with mostly extreme weights for CPI and more even weights for IP.

We also observe that minimal weight is put on impulse responses that seem a priori less plausible. For example, the LPs estimate that a decline in CPI continues to deepen even after four years, a feature that seems ex ante unlikely based on theory. Correspondingly, almost all the weight is placed on the VAR after the initial months.

One notable instance where an LP gets higher weight at a longer horizon is for IP at the two- to three-year horizon. In particular, even though the LP with contemporaneous controls has higher variance and a similar point estimate to the VAR, more weight is put on the LP. The weights are therefore not solely tied to the precision of the estimate and we find an average response that has wider error bands than the VAR.

In Appendix D, we show several additional results. First, we consider different approaches to splitting the sample. While the weights and average responses change substantially when the predictive densities are computed in-sample (i.e., no sample splitting), the difference is much smaller when we consider more subsamples. Second, we add an additional model into the pool and show that the weights are consistent with the theoretical results we state in Appendix A. Finally, we replace the narrative instrument with high frequency instruments from Gertler and Karadi (2015), Miranda-Agrippino and Ricco (2021), and Jarociński and Karadi (2020). While the exact average differs across instruments, we continue to find a large weight on the VAR.

4.2 Total Factor Productivity Shock

We now consider the effect of a TFP shock on unemployment as a second example. We revisit the results of Gali (1999) and Basu et al. (2006) that suggest that positive productivity shocks lead to a decline in labor inputs.

We take quarterly data from 1968Q1 through 2019Q4. As a measure of TFP, we use the updated version of the TFP series from Fernald (2014).[14] The remaining data come from the updated FRED-QD database (McCracken and Ng, 2020).[15]

We consider eight models:

- **VARs:** Recursive identification with TFP ordered last.
  - **All Controls:** Control for the first three principal components, GDP growth, and utilization.
  - **Factors:** Control for the first three principal components.
  - **No Controls:** Only include unemployment and TFP.
- **FAVAR:** Following Bernanke et al. (2005), we estimate:

$$Y_t = \Lambda F_t + u_t \tag{22}$$

$$F_t = \sum_{\ell=1}^{L} B_\ell F_{t-\ell} + C\varepsilon_t \tag{23}$$

  We take the first three principal components and TFP as observed factors, F_t, and identify TFP shock recursively with TFP ordered last.
- **LPs:** Use TFP as an instrument and include the same sets of controls as the VARs.
- **Single Equation:** Following Romer and Romer (2004) and Baek and Lee (2022), we

[14] Available on John Fernald's website.
[15] Available on Michael McCracken's website.

estimate:

$$\Delta U_t = \gamma + \sum_{\ell=0}^{A} \alpha_\ell \Delta TFP_{t-\ell} + \sum_{\ell=1}^{B} \beta_\ell \Delta U_{t-\ell} + v_t, \tag{24}$$

with $A = 20$ and $B = 4$, from which we can recursively compute the impulse responses.

All models include 4 lags of the endogenous variables.

The wide variety of models emphasizes the flexibility of the methodology. The range of controls reflects the question of whether and how one should control for capacity utilization when estimating the effects of productivity shocks, a point emphasized by Basu et al. (2006). The inclusion of this many models does not substantially change the computation time, as the individual model estimation and computations can be parallelized and the calculation of the weights is relatively efficient.

Results. We plot the individual model and averaged impulse responses panel by panel in Figure 4. The models find a wide range of possible impulse responses, with no consistent pattern, especially at 1-2 year horizon. For example, the VARs have zero response on impact by assumption, the FAVAR estimates a negative response of unemployment on impact, while the LPs and single equation model find mixed results. At the five-year horizon, there are again substantial differences in point estimates and error bands across models.

Despite the differences across models, the averaged response delivers a clear message: unemployment rises over the first year, but returns to trend, with a tightly estimated zero response at the five-year horizon. Notably, the uncertainty region in the middle of the horizon is moderately wide, reflecting the different responses across models. Yet, the average remains resolutely positive. This is consistent with results in Basu et al. (2006). Importantly, no single model replicates all the features of the averaged response.

Figure 5 shows the weights underlying these average responses. While most of the weight is placed on the LPs at the one- to two-year horizon, the VARs and FAVAR are more heavily weighted subsequently. The difference across horizons is again consistent with the monetary shock application and DSGE Monte Carlo in Appendix C. As expected, the additional

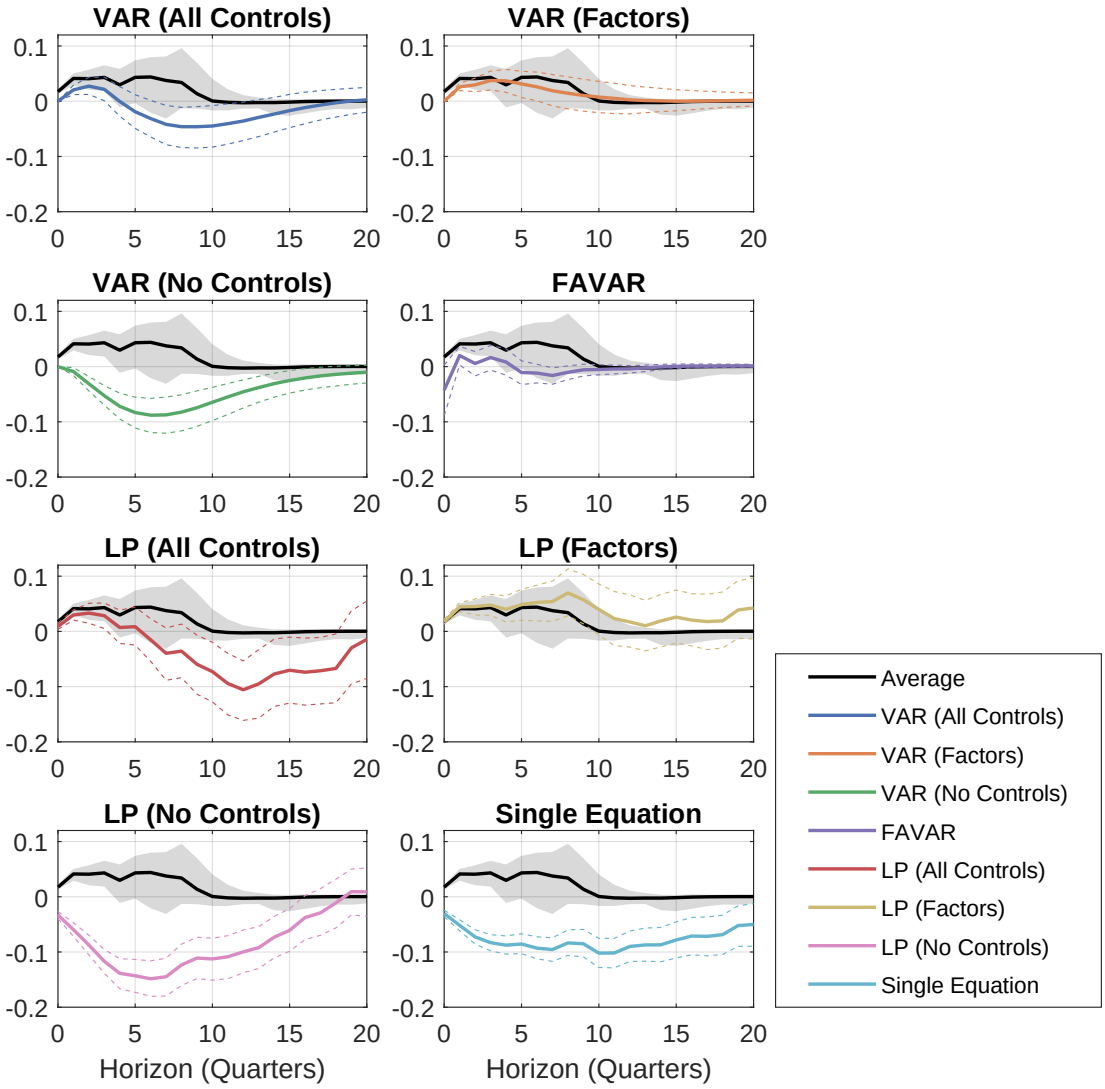

Figure 4: Impulse responses for each model, mean and 68% error bands. Colored lines correspond to individual models, black line and gray shade regions correspond to averaged model.

structure of multiple equation models such as the VARs and FAVAR tighten estimates of the longer run response to the shock, resulting in predictive densities that are supported by the data. We also find that models that use the principal components as controls are favored.

Figure 5 also shows how the weights and average response change when we instead consider predictive densities without conditioning on the shock, as in Geweke and Amisano (2011). The weights for the first three years look substantially different. Most noticeably,

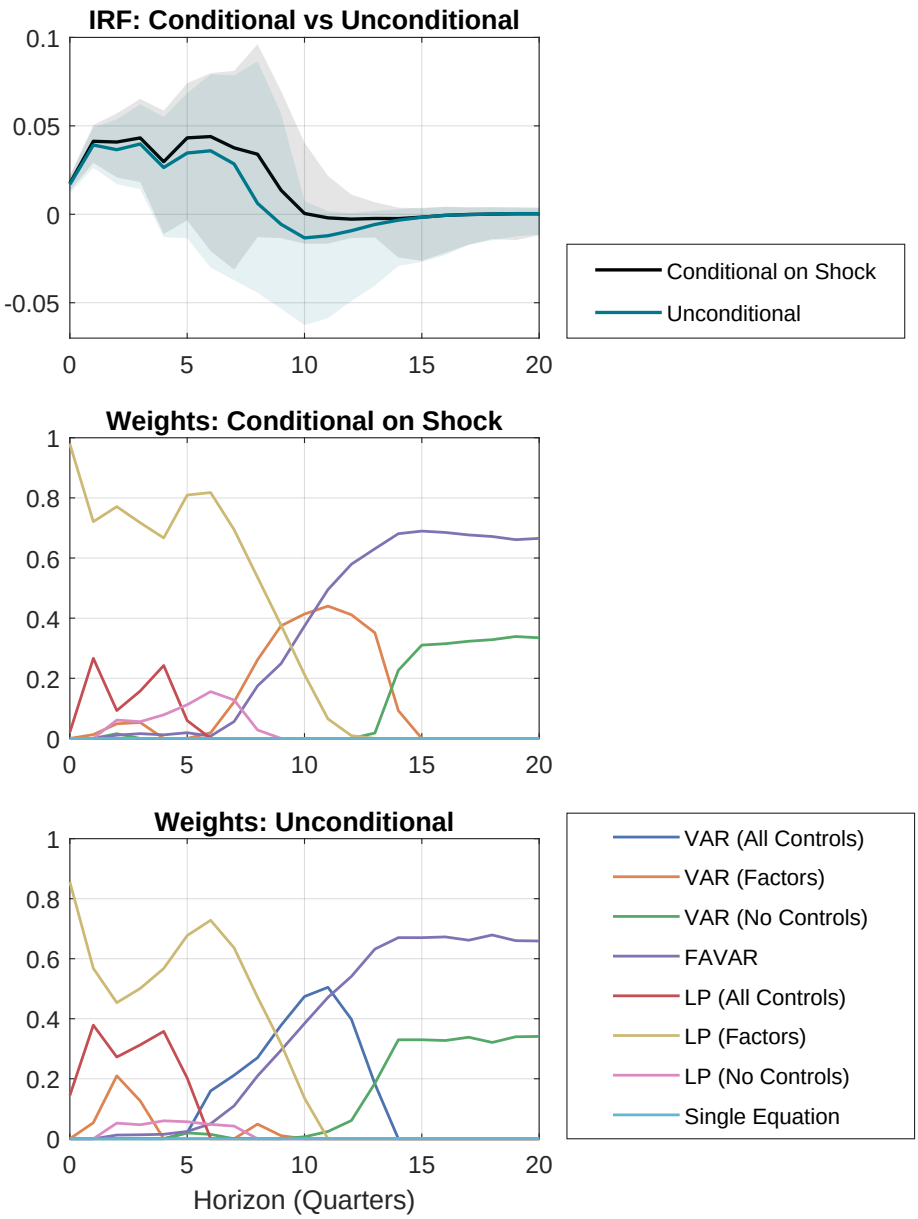

Figure 5: Averaged impulse response, optimal weights conditional on shock, and optimal unconditional weights. **Top:** Averaged impulse response, median and 68% error bands; **Middle:** Weights conditional on shock; **Bottom:** Unconditional weights.

there is less weight on the VAR and LP that only control for the principal components, with more weight placed on the models that also control for GDP growth and utilization. As a result, the averaged impulse response finds a slight dip in unemployment after ten quarters, in contrast to the flat response under the weights that condition on the shock.

On the other hand, the weights at the four- to five-year horizon do not materially depend on whether we condition on the shock. The weights that condition on the shock suggest that the effect of the TFP shock is only transitory. As a result, the most preferred models at these horizons are also the ones that forecast well unconditionally. In practice, it may not be clear ex ante how much conditioning on the shock might matter since this could depend on the data, models, and shock, and could vary across variables and horizons.

5 Conclusion

In this paper, we develop a methodology that delivers an encompassing approach to computing dynamic responses of macroeconomic variables to shocks. Building on the idea of prediction pools, as in Geweke and Amisano (2011), we average across impulse responses from multiple models by leveraging their close connection with conditional forecasts. Our framework thus presents an approach to incorporate evidence from a variety of models in a consistent and plausible manner to get closer to the actual truth in the data.

Our methodology tackles the fundamental challenge of choosing a specific impulse response estimator. Each estimator comes with its own issues such as bias or large standard errors, but general theorems about which class of models should be used are hard to come by once we make realistic assumptions. This has led to many alternative impulse response estimators coexisting in the literature. We exploit that each of these can be useful in particular situations, making empirical macroeconomics an ideal setting for flexible model-averaging schemes. Our approach makes model-averaging in these scenarios possible.

The key differences relative to existing methods are (i) much greater flexibility in the range of possible estimators; (ii) a relatively small computational burden; and (iii) ability to exploit each method's strength as much as possible by computing horizon- and variable-specific weights based on predictive densities instead of a few selected statistics.

Overall, our Monte Carlos and empirical applications highlight several broad messages:

1. The optimal weights on models depend on horizon, variable, and application.

2. The optimal weights depend on the entire predictive distribution, not only the point estimates. Our examples focus on the mean and variance, but skewness, kurtosis, or any other property of the predictive distribution could be important more generally.

3. Theoretical properties of individual models are not sufficient criteria for the choice of weights. For instance, misspecified models may dominate correctly specified (or more flexible) models in finite sample. On the other hand, models that produce tighter estimates need not receive greater weight.

Our use of prediction pools provides a systematic and computationally tractable way to account for these issues in a wide range of applications.

References

Amisano, Gianni and John Geweke (2017), "Prediction Using Several Macroeconomic Models." *Review of Economics and Statistics*, 99, 912–925.

Aruoba, S Boragan and Thomas Drechsel (2023), "Identifying monetary policy shocks: A natural language approach." Working paper.

Baek, ChaeWon and Byoungchan Lee (2022), "A Guide to Autoregressive Distributed Lag Models for Impulse Response Estimations." *Oxford Bulletin of Economics and Statistics*, 84, 1101–1122.

Bagliano, Fabio C and Carlo A Favero (1999), "Information from Financial Markets and VAR Measures of Monetary Policy." *European Economic Review*, 43, 825–837.

Barnichon, Regis and Christian Brownlees (2019), "Impulse Response Estimation by Smooth Local Projections." *Review of Economics and Statistics*, 101, 522–530.

Basu, Susanto, John G Fernald, and Miles S Kimball (2006), "Are Technology Improvements Contractionary?" *American Economic Review*, 96, 1418–1448.

Bates, J. M. and C. W. J. Granger (1969), "The Combination of Forecasts." *Journal of the Operational Research Society*, 20, 451–468.

Bernanke, Ben S, Jean Boivin, and Piotr Eliasz (2005), "Measuring the Effects of Monetary Policy: A Factor-augmented Vector Autoregressive (FAVAR) Approach." *Quarterly Journal of Economics*, 120, 387–422.

Bruns, Martin and Helmut Lütkepohl (2022), "Comparison of Local Projection Estimators for Proxy Vector Autoregressions." *Journal of Economic Dynamics and Control*, 134, 104277.

Canova, Fabio and Christian Matthes (2021), "A Composite Likelihood Approach for Dynamic Structural Models." *Economic Journal*, 131, 2447–2477.

Cogley, Timothy and Thomas J. Sargent (2005), "Drift and Volatilities: Monetary Policies and Outcomes in the Post WWII U.S." *Review of Economic Dynamics*, 8, 262–302.

Coibion, Olivier (2012), "Are the Effects of Monetary Policy Shocks Big or Small?" *American Economic Journal: Macroeconomics*, 4, 1–32.

Del Negro, Marco, Raiden B. Hasegawa, and Frank Schorfheide (2016), "Dynamic Prediction Pools: An Investigation of Financial Frictions and Forecasting Performance." *Journal of Econometrics*, 192, 391–405.

Del Negro, Marco and Frank Schorfheide (2004), "Priors from General Equilibrium Models for VARs." *International Economic Review*, 45, 643–673.

Dinh, Viet Hoang, Didier Nibbering, and Benjamin Wong (2023), "Random subspace local projections." CAMA Working Paper 34/2023.

Doan, Thomas, Robert Litterman, and Christopher Sims (1984), "Forecasting and Conditional Projection Using Realistic Prior Distributions." *Econometric Reviews*, 3, 1–100.

Fernald, John (2014), "A Quarterly, Utilization-adjusted Series on Total Factor Productivity." Federal Reserve Bank of San Francisco Working Paper Series 2012-19.

Ferreira, Leonardo N, Silvia Miranda-Agrippino, and Giovanni Ricco (2023), "Bayesian Local Projections." *The Review of Economics and Statistics*, 1–45.

Gali, Jordi (1999), "Technology, Employment, and the Business Cycle: Do Technology Shocks Explain Aggregate Fluctuations?" *American Economic Review*, 89, 249–271.

Gertler, Mark and Peter Karadi (2015), "Monetary Policy Surprises, Credit Costs, and Economic Activity." *American Economic Journal: Macroeconomics*, 7, 44–76.

Geweke, John and Gianni Amisano (2011), "Optimal Prediction Pools." *Journal of Econometrics*, 164, 130–141.

Geweke, John and Gianni Amisano (2012), "Prediction with Misspecified Models." *American Economic Review: Papers & Proceedings*, 102, 482–86.

Giannone, Domenico, Michele Lenza, and Giorgio E Primiceri (2015), "Prior Selection for Vector Autoregressions." *Review of Economics and Statistics*, 97, 436–451.

Gonçalves, Sılvia and Lutz Kilian (2004), "Bootstrapping Autoregressions with Conditional Heteroskedasticity of Unknown Form." *Journal of Econometrics*, 123, 89–120.

Gürkaynak, Refet S, Burçin Kısacıkoğlu, and Barbara Rossi (2013), "Do DSGE Models Forecast More Accurately Out-Of-Sample than VAR Models?" In *VAR Models in Macroeconomics–New Developments and Applications: Essays in Honor of Christopher A. Sims*, volume 32, 27–79, Emerald Group Publishing Limited.

Hall, Stephen G. and James Mitchell (2007), "Combining Density Forecasts." *International Journal of Forecasting*, 23, 1–13.

Hansen, Bruce E. (2007), "Least Squares Model Averaging." *Econometrica*, 75, 1175–1189.

Hansen, Bruce E (2016), "Stein Combination Shrinkage for Vector Autoregressions." Working paper, University of Wisconsin, Madison.

Hansen, Peter R., Asger Lunde, and James M. Nason (2011), "The Model Confidence Set." *Econometrica*, 79, 453–497.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd edition. Springer.

Herbst, Edward and Benjamin K. Johannsen (2020), "Bias in Local Projections." Finance and Economics Discussion Series 2020-010, Washington: Board of Governors of the Federal Reserve System.

Jarociński, Marek and Peter Karadi (2020), "Deconstructing Monetary Policy Surprises—The Role of Information Shocks." *American Economic Journal: Macroeconomics*, 12, 1–43.

Jordà, Òscar (2005), "Estimation and Inference of Impulse Responses by Local Projections." *American Economic Review*, 95, 161–182.

Kilian, Lutz (1998), "Small-Sample Confidence Intervals for Impulse Response Functions." *Review of Economics and Statistics*, 80, 218–230.

Li, Dake, Mikkel Plagborg-Møller, and Christian K. Wolf (2022), "Local Projections vs. VARs: Lessons From Thousands of DGPs." Working paper.

Lusompa, Amaze (2021), "Local Projections, Autocorrelation, and Efficiency." Federal Reserve Bank of Kansas City Working Paper 21-01.

Marcellino, Massimiliano, James H Stock, and Mark W Watson (2006), "A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series." *Journal of Econometrics*, 135, 499–526.

McCracken, Michael and Serena Ng (2020), "FRED-QD: A Quarterly Database for Macroeconomic Research." Federal Reserve Bank of St. Louis Working Paper 2020-005.

McKay, Alisdair and Christian K Wolf (2023), "What Can Time-Series Regressions Tell Us About Policy Counterfactuals?" *Econometrica*, 91, 1695–1725.

Miranda-Agrippino, Silvia and Giovanni Ricco (2021), "The Transmission of Monetary Policy Shocks." *American Economic Journal: Macroeconomics*, 13, 74–107.

Montiel Olea, José Luis and Mikkel Plagborg-Møller (2021), "Local Projection Inference is Simpler and More Robust Than You Think." *Econometrica*, 89, 1789–1823.

Noh, Eul (2018), "Impulse-Response Analysis with Proxy Variables." Working paper, University of California San Diego.

Paul, Pascal (2020), "The Time-Varying Effect of Monetary Policy on Asset Prices." *Review of Economics and Statistics*, 102, 690–704.

Pesavento, Elena and Barbara Rossi (2006), "Small-Sample Confidence Intervals for Multivariate Impulse Response Functions at Long Horizons." *Journal of Applied Econometrics*, 21, 1135–1155.

Plagborg-Møller, Mikkel and Christian K. Wolf (2021), "Local Projections and VARs Estimate the Same Impulse Responses." *Econometrica*, 89, 955–980.

Primiceri, Giorgio E. (2005), "Time Varying Structural Vector Autoregressions and Monetary Policy." *Review of Economic Studies*, 72, 821–852.

Qu, Zhongjun (2018), "A Composite Likelihood Framework for Analyzing Singular DSGE Models." *The Review of Economics and Statistics*, 100, 916–932.

Ramey, Valerie A. (2016), "Macroeconomic Shocks and Their Propagation." *Handbook of Macroeconomics*, 2, 71–162.

Romer, Christina D and David H Romer (2004), "A New Measure of Monetary Shocks: Derivation and Implications." *American Economic Review*, 94, 1055–1084.

Sims, Christopher A (1980), "Macroeconomics and Reality." *Econometrica*, 1–48.

Sims, Christopher A and Tao Zha (1999), "Error Bands for Impulse Responses." *Econometrica*, 67, 1113–1155.

Sims, Christopher A. and Tao Zha (2006), "Were There Regime Switches in U.S. Monetary Policy?" *American Economic Review*, 96, 54–81.

Smets, Frank and Rafael Wouters (2007), "Shocks and Frictions in US Business Cycles: A Bayesian DSGE approach." *American Economic Review*, 97, 586–606.

Stock, James H. and Mark W. Watson (2018), "Identification and Estimation of Dynamic Causal Effects in Macroeconomics Using External Instruments." *The Economic Journal*, 128, 917–948.

Stock, J.H. and M.W. Watson (2016), "Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics." In *Handbook of Macroeconomics* (J. B. Taylor and Harald Uhlig, eds.), volume 2, 415–525, Elsevier.

Stone, M. (1961), "The Opinion Pool." *The Annals of Mathematical Statistics*, 32, 1339 – 1342.

Strachan, R.W. and H.K. van Dijk (2007), "Bayesian Model Averaging in Vector Autoregressive Processes With an Investigation of Stability of the US Great Ratios and Risk of a Liquidity Trap in the USA, UK and Japan." Econometric Institute Research Papers EI 2007-11, Erasmus University Rotterdam, Erasmus School of Economics (ESE), Econometric Institute.

Waggoner, Daniel F. and Tao Zha (2012), "Confronting Model Misspecification in Macroeconomics." *Journal of Econometrics*, 171, 167–184.

A Theoretical Results of Geweke and Amisano (2011)

In this section we state major theoretical results from Geweke and Amisano (2011) for the sake of completeness. Let us first restate our objective function:

$$f_T(\mathbf{w}_h) = \sum_{t=1}^{T} \log \left[\sum_{m=1}^{M} w_{m,h} p_m^* \left(y_{t+h,j} \right) \right] \tag{A.1}$$

where $\mathbf{w}_h = [w_{1,h}\ w_{2,h} \cdots w_{M,h}]'$ is the vector of model weights for a given horizon h (and a given variable j, which we have not made explicit in this notation). We will generally assume that $f_T(\mathbf{w}_h)$ is concave, i.e., $\partial^2 f_T / \partial \mathbf{w}_h \partial \mathbf{w}_h'$ is negative definite. For the case of two models ($M = 2$), Geweke and Amisano (2011) show that that the objective function will be concave if the expected difference between the two predictive densities will not be zero as the same size increases.[16] We will call a subset of models *dominant* if its weights sum to 1. A subset of models is *excluded* if each of these models has a weight of 0. With the assumption of concavity, Geweke and Amisano (2011) show the following results:

1. If $\{\mathcal{M}_1, \ldots, \mathcal{M}_m\}$ dominates the pool $\{\mathcal{M}_1, \ldots, \mathcal{M}_n\}$ then $\{\mathcal{M}_1, \ldots, \mathcal{M}_m\}$ dominates $\{\mathcal{M}_1, \ldots, \mathcal{M}_m, \mathcal{M}_{j_1}, \ldots, \mathcal{M}_{j_k}\}$ for all $\{j_1, \ldots, j_k\} \subseteq \{m+1, \ldots, n\}$.

2. If $\{\mathcal{M}_1, \ldots, \mathcal{M}_m\}$ dominates all pools $\{\mathcal{M}_1, \ldots, \mathcal{M}_m, \mathcal{M}_j\}$ $(j = m+1, \ldots, n)$ then $\{\mathcal{M}_1, \ldots, \mathcal{M}_m\}$ dominates the pool $\{\mathcal{M}_1, \ldots, \mathcal{M}_n\}$.

3. The set of models $\{\mathcal{M}_1, \ldots, \mathcal{M}_m\}$ is excluded in the pool $\{\mathcal{M}_1, \ldots, \mathcal{M}_n\}$ if and only if $\mathcal{M}_j$ is excluded in each of the pools $\{\mathcal{M}_j, \mathcal{M}_{m+1}, \ldots, \mathcal{M}_n\}$ $(j = 1, \ldots, m)$.

4. If the model $\mathcal{M}_1$ is excluded in all pools $(\mathcal{M}_1, \mathcal{M}_i)$ $(i = 2, \ldots, n)$ then $\mathcal{M}_1$ is excluded in the pool $(\mathcal{M}_1, \ldots, \mathcal{M}_n)$.

[16] Note that even in the case of LPs and VARs with the same right-hand side variables, it is unlikely (at least at larger horizons) that the implied predictive densities are the same even though the VAR specification is nested in LPs.

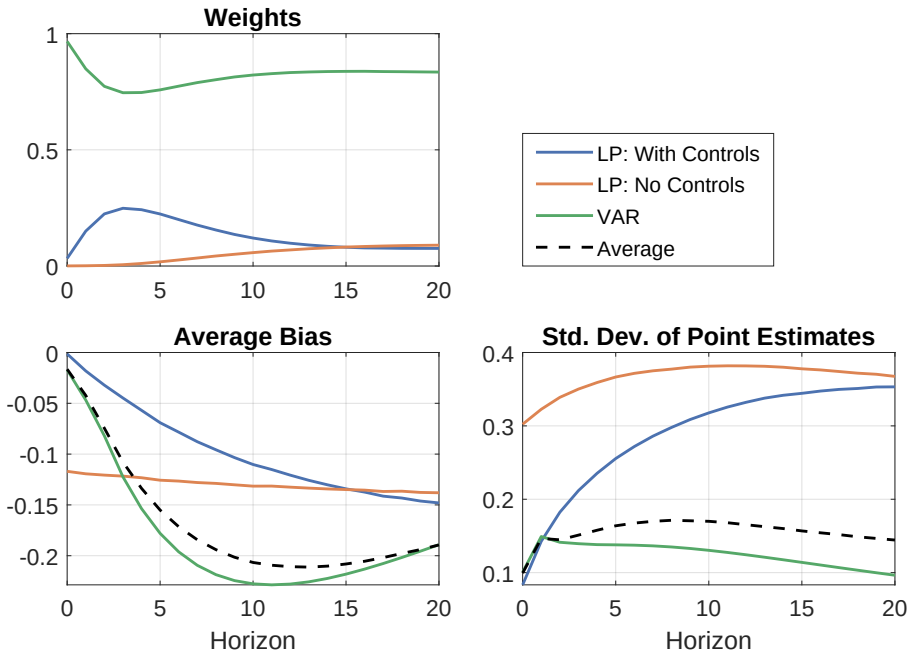

Figure A.1: Prediction pool weights, biases, and standard deviations from Monte Carlo with AR(1) model from Herbst and Johannsen (2020) in pool with internal instrument VAR. **Top left:** Prediction pool and least-squares weights on model with controls; **Bottom left:** Bias of impulse responses under each specification and averaged model; **Bottom right:** Standard deviation of point estimate of impulse response estimates. Dashed lines correspond to individual models, solid lines correspond to averaged model. All plots show averages across all Monte Carlo repetitions.

B AR(1) Monte Carlo Supplementary Results

B.1 Including a VAR(1) in the Pool of Models

We now add a VAR(1) with $v_{1,t}$ as an internal instrument (i.e., equation (19) with $z_t = v_{1,t}$) into the pool of models. On the one hand, the VAR assumes an autoregressive structure that is consistent with the data-generating process. On the other hand, including the lagged instrument $v_{1,t-1}$ introduces additional parameters to estimate. Moreover, while the innovation is directly observed in the LPs, it has to be inferred from the estimated parameters in the VAR.

The results are shown in Figure A.1. A large weight is placed on the VAR. In particular, additional structure of the VAR substantially reduces the standard deviation of the estimates,

especially at longer horizons. However, the VAR also generates finite sample bias that is larger than that of the LPs, except for short horizons. The resulting optimal weights reflect the bias-variance tradeoff, and tend to favor the VAR.

The ratio of the weights on the two LPs remains relatively unchanged after the introduction of the VAR. Appendix D shows that we find a similar result when we add a model to the monetary policy application in Section 4.1. The comparability of the relative weights across nested prediction pools is a desirable feature of the methodology, formalized theoretical results in Appendix A for the limit case where particular models receive zero weight.

B.2 Coverage with Omitted Variables

As an additional example of how our approach can improve coverage, we augment the data-generating process (14) in Section 3.1 with an omitted variable. Specifically, we generate data from the model:

$$y_t = \rho y_{t-1} + w_t + v_{1,t} + v_{2,t} \tag{A.2}$$

$$w_t = \gamma w_{t-1} + v_{w,t} \tag{A.3}$$

$$\begin{bmatrix} v_{w,t} \\ v_{1,t} \\ v_{2,t} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \eta & 0 \\ \eta & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right). \tag{A.4}$$

However, we estimate the same models as in Section 3.1, i.e., equation (15) with $x_{m,t} = (v_{1,t}, y_{t-1})'$ or $x_{m,t} = v_{1,t}$. Importantly, the exclusion of w_t in these estimated models leads to omitted variable bias when $\eta \neq 0$. The parameter η controls the bias on impact and the persistence γ of w_t controls how this bias spills over to longer horizons. We set $\gamma = 0.75$ and $\eta \in \{-0.05, 0.05\}$. As in Section 3.1, we take $\rho = 0.95$ and use $T = 150$ observations. We present results averaged over 5×10^5 simulations.

Figures A.2 and A.3 show the results with $\eta = -0.05$ and $\eta = 0.05$, respectively. Using

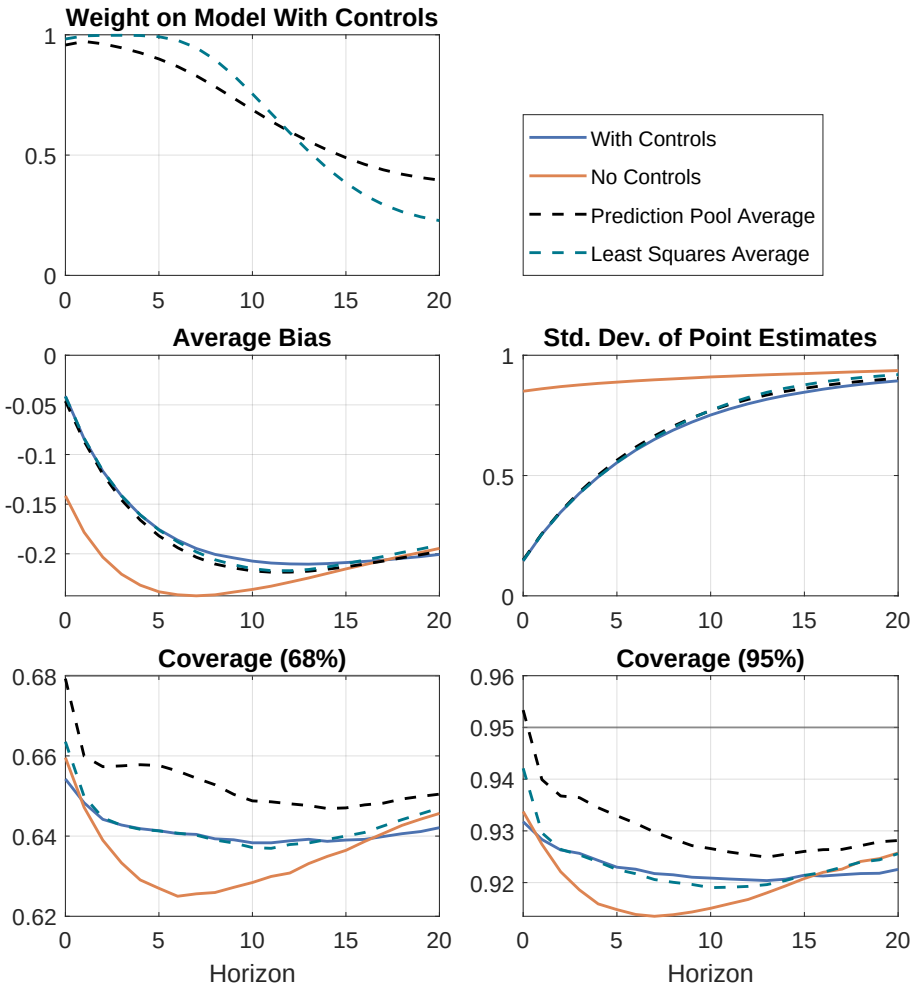

Figure A.2: Prediction pool weights, biases, standard deviations, and coverage from Monte Carlo with AR(1) model with negative omitted variable bias ($\eta = -0.05$). **Top left:** Prediction pool and least-squares weights on model with controls; **Middle left:** Bias of impulse responses under each specification and averaged model; **Middle right:** Standard deviation of point estimate of impulse response estimates; **Bottom:** Coverage of equal-tailed 68% (left) and 95% (right) error bands. Dashed lines correspond to individual models, solid lines correspond to averaged model. All plots show averages across all Monte Carlo repetitions.

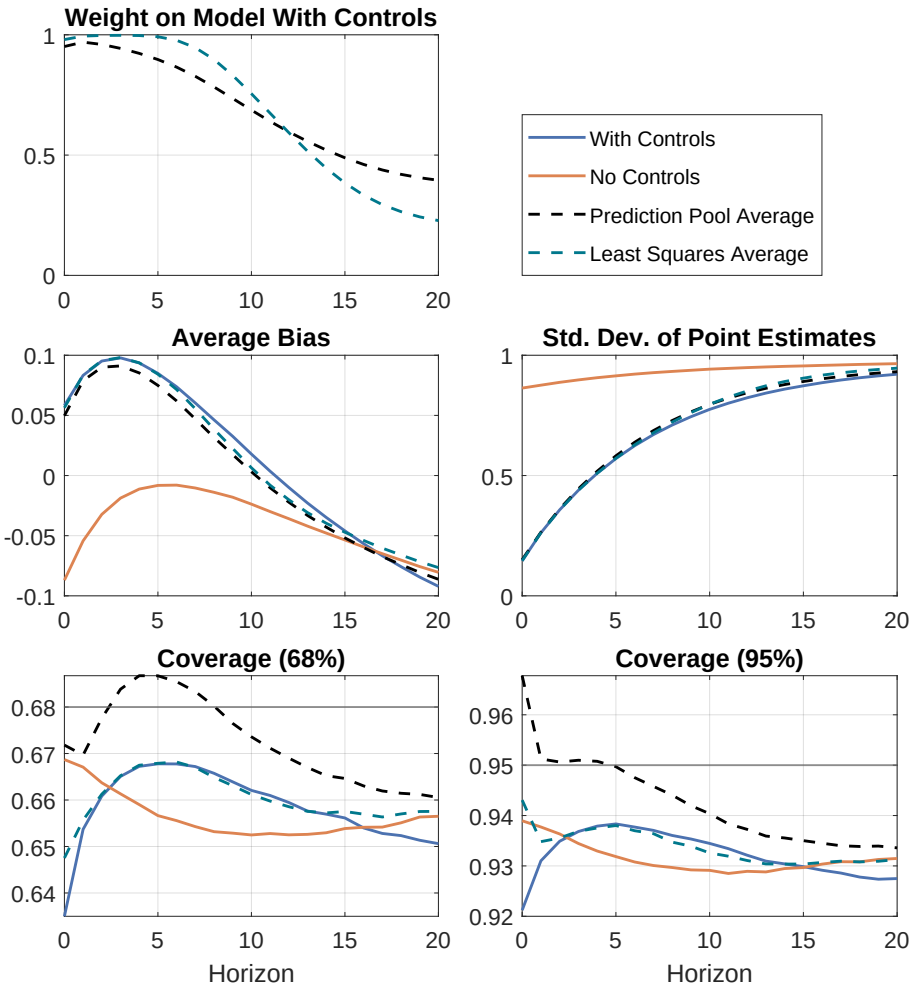

Figure A.3: Prediction pool weights, biases, standard deviations, and coverage from Monte Carlo with AR(1) model with positive omitted variable bias ($\eta = 0.05$). **Top left:** Prediction pool and least-squares weights on model with controls; **Middle left:** Bias of impulse responses under each specification and averaged model; **Middle right:** Standard deviation of point estimate of impulse response estimates; **Bottom:** Coverage of equal-tailed 68% (left) and 95% (right) error bands. Dashed lines correspond to individual models, solid lines correspond to averaged model. All plots show averages across all Monte Carlo repetitions.

the prediction pool to average across models improves coverage across almost all horizons. In contrast, the least squares average traces the upper envelope of the coverage across the two models, thus yielding poorer coverage than the prediction pool average even though their average bias and standard deviation of point estimates are similar.

Since the prediction pools objective function does not explicitly target coverage, an estimator with correct coverage but relative large bias or wide error bands could receive low weight because it yields low predictive densities. In this example, we find for the impulse response on impact that most of the weight is placed on the model with controls despite its poorer coverage. Nevertheless, this seemingly small amount of averaging still leads to a noticeable improvement in coverage compared to the model with controls on its own.

C Medium-Scale New Keynesian Model Monte Carlo

We next consider a Monte Carlo exercise with data generated from a quantitative dynamic stochastic general equilibrium (DSGE) model to connect the simple simulation examples in Section 3 more closely to actual empirical settings. We use a DSGE model as our data-generating process because it implies vector autoregressive moving average dynamics for the vector of observables, so that both models we consider, VARs and LPs, are misspecified. Despite using closely related models, we find different estimates in finite sample. The averaged impulse response balances the bias-variance trade-off, and in some cases even has a smaller bias than either individual model.

Data-Generating Process. We simulate data from the log-linearized medium-scale New Keynesian model from Smets and Wouters (2007) with parameters fixed at the posterior mode reported in the paper. We use the model to generate 150 periods of simulated data for the seven observables used by Smets and Wouters (2007) to estimate the model: GDP growth, consumption growth, investment growth, wage growth, hours, inflation, and the federal funds rate. We focus on the impulse response of each variable to a monetary shock,

which we assume to be observed by the econometrician. We obtain results across 2.5×10^4 simulations.

Models. We compare two models: an internal instrument VAR (19) estimated using Bayesian methods and the Bayesian LP from Ferreira et al. (2023). The Bayesian LP estimates:

$$\begin{bmatrix} z_{t+h} \\ y_{t+h} \end{bmatrix} = B^{(h)} \begin{bmatrix} z_t \\ y_t \end{bmatrix} + u_{t+h}^{(h)}. \tag{A.5}$$

for each horizon $h > 0$. The impulse response at horizon h is $B^{(h)}C_{\bullet,1}$, where $C_{\bullet,1}$ is obtained from (19). Ferreira et al. (2023) show how to impose a prior on the model and estimate the LP impulse response analogously to a Bayesian VAR.[17] Both models have one lag and use the same Minnesota prior. In addition, we assume that the shock z_t is perfectly observed.

The two models are closely connected. First, if $B^{(h)} = B^h$, then the VAR and LP produce identical impulse responses. In particular, given the same priors, the two models would produce identical on-impact impulse responses. Next, as pointed out by Plagborg-Møller and Wolf (2021), under the appropriate regularity conditions, the two models asymptotically produce identical impulse responses. However, as our Monte Carlo exercises show, in finite sample and under misspecification the two models can lead to substantially different estimates despite their close connections. Arguably, this requires a systematic way to average across models.[18]

Results. The weights, averaged across simulations, are summarized in the left panels of Figure A.4. We start the horizontal axis for each panel at horizon 1 since the two models

[17] Ferreira et al. (2023) do not explicitly model autocorrelation in the residual of their LP specification, but instead use a sandwich-type estimator for the posterior covariance matrix.

[18] There are two differences of note relative to Plagborg-Møller and Wolf (2021). First, because we impose a prior, the estimated impulse responses at horizon $h > 0$ differ even if the least squares estimates are equivalent. In particular, for longer horizons, the likelihood of the LP becomes more dispersed, bringing the posterior closer to the prior. Second, the estimated system (A.5) differs from the LP setup used in Plagborg-Møller and Wolf (2021).

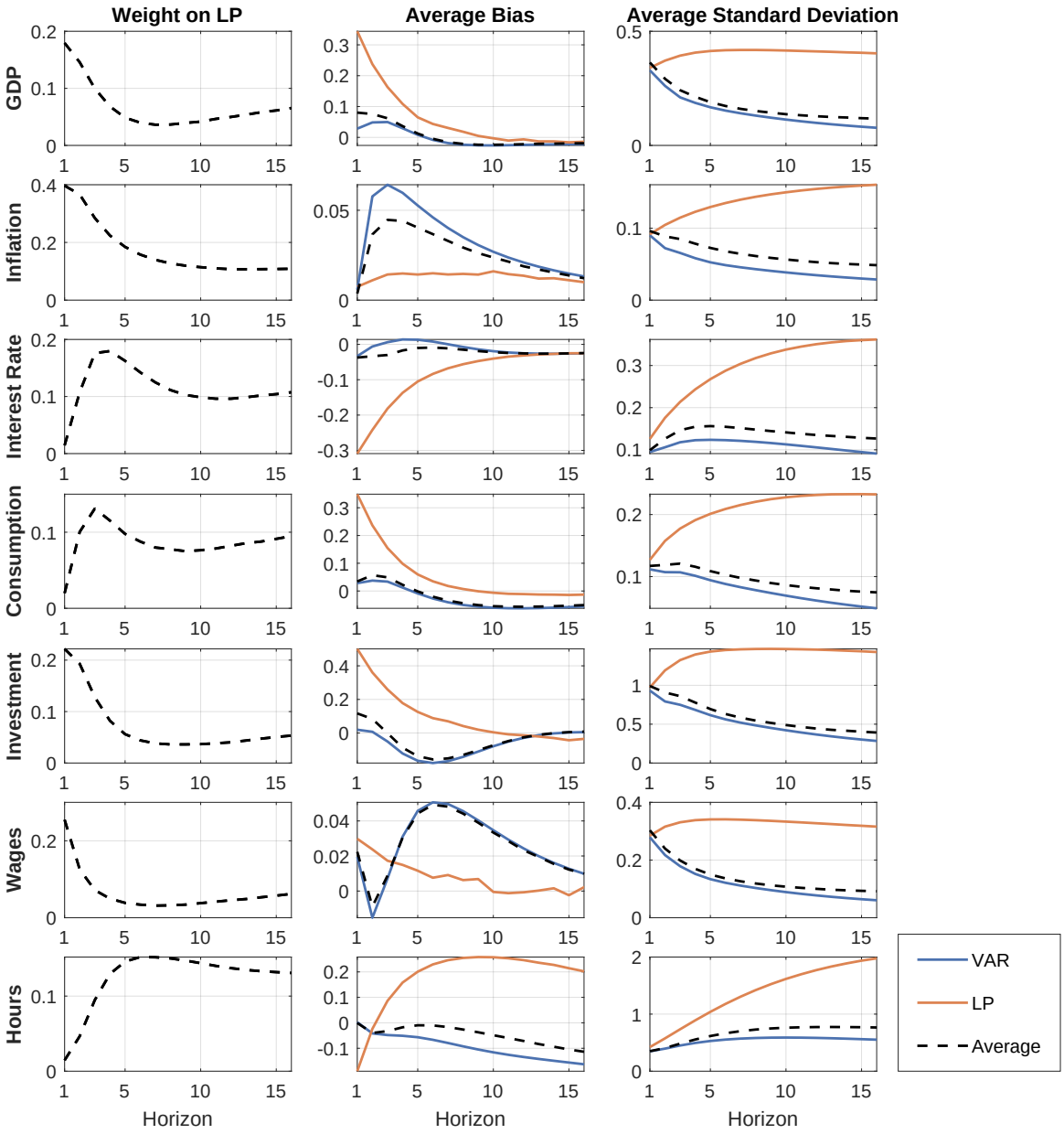

Figure A.4: Prediction pool weights, biases, and posterior standard deviations from Smets and Wouters (2007) Monte Carlo. Biases and standard deviations averaged across simulations. **Left:** Optimal weights on LP; **Middle:** Bias of impulse responses; **Right:** Posterior standard deviation of impulse responses. All plots show averages across all Monte Carlo repetitions.

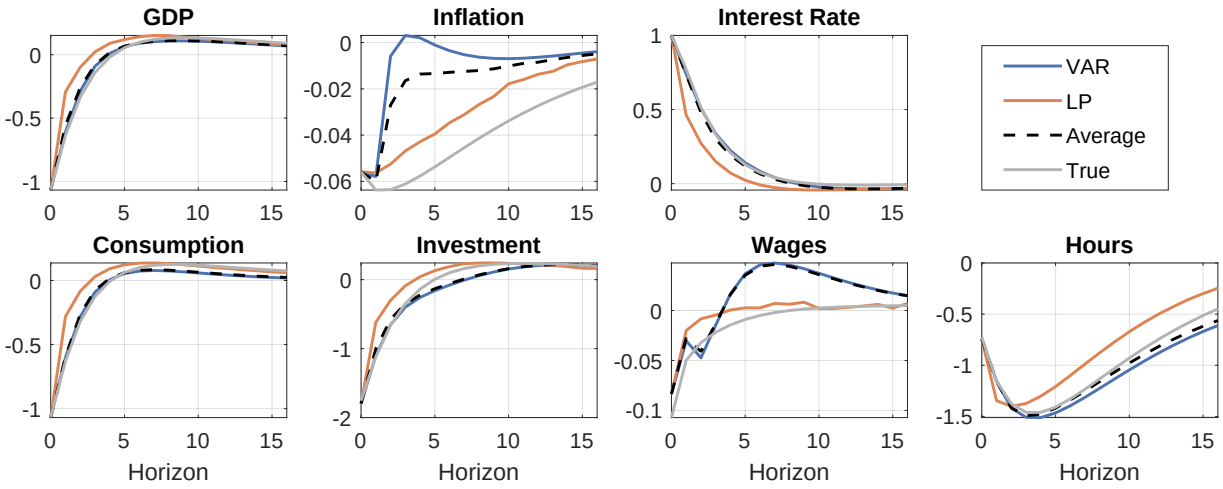

Figure A.5: Posterior mean estimates of impulse response to monetary policy shock in New Keynesian model Monte Carlo, average across simulations. GDP, consumption, investment, and wages in growth rates.

imply identical impulse responses on impact. Overall, the prediction pools place greater weight on the VAR, with the LP typically getting a weight of 0.2 or less. The weight on the LP tends to fall at longer horizons. Nevertheless, there are non-trivial weights on the LP, especially for inflation.

The middle and right panels of Figure A.4 plot the average biases and standard deviations of the impulse response functions, providing an explanation for the small weights on the LP. First, even though the LP has greater flexibility, in many cases its bias tends to be larger or is at most of similar magnitude relative to the VAR. This arises partly due to the relatively short sample of 150 periods. Second, the right panels show that the LP has substantially larger posterior standard deviations, which is consistent with evidence reported in the literature (Miranda-Agrippino and Ricco, 2021; Li et al., 2022). The difference in standard deviations is especially large at longer horizons, which accounts for the lower weights on the LP at those horizons.

To get a better sense of what is driving the weights, we focus first on the impulse response for consumption. At short horizons, the LP has a substantially larger bias than the VAR, resulting in almost all the weight being placed on the VAR. Over the initial periods, the

relative bias of the LP decreases rapidly, and the weight on the LP correspondingly increases. However, after horizon $h = 3$ the weight on the LP decreases again, as the optimal weights trade off the declining bias of the LP with its increasing standard deviation. Finally, the weight increases again at longer horizons as the VAR begins to have a larger bias than the LP. In particular, the Figure A.5 shows that the stationarity imposed by the VAR generates a VAR that converges more quickly to zero than the true response from the DSGE model.

The impulse response for hours further illustrates the behavior of the prediction pools. The weights on the LP increases over the first four periods but does not decay as quickly as for other variables. Even though the standard deviations are similar initially, the LP displays a substantially larger bias than the VAR at short horizons, reducing its optimal weight. Subsequently, the LP and VAR have biases of opposite signs that offset each other when averaged, as was the case in the Monte Carlo exercise in Section 3.2. By averaging the impulse responses, the prediction pool can produce an average impulse response that has a smaller bias than either model, with the bias almost completely eliminated at horizon $h = 15$. At longer horizons, the weights trade off two forces. First, the VAR bias begins to increase while the LP bias begins to decrease. Second, the LP posterior standard deviation increases while the VAR standard deviation remains relatively constant. In balance, the weights begin to favor the LP less at longer horizons, but with a decline that is less steep than in other variables.

Overall, the results here emphasize two key messages. First, the relative biases and variances of the models differ depending on variable and horizon. Prediction pools offer the flexibility to trade off these properties variable by variable and horizon by horizon, thus making full use of the relative strengths of each model. Second, even when models have similar asymptotic properties, there can be substantial gains from averaging over them in finite sample. In particular, the bias of the average impulse response can in some cases be lower than that of either individual model.

D Monetary Shock Supplementary Results

We present results from several supplementary exercises to the monetary policy application in Section 4.1.

D.1 Sample Splits

Throughout the main text, we computed the optimal prediction pool weights by splitting our sample into half. We estimated the models for each subsample separately and used the implied out-of-sample forecasting densities for the parts of the sample that were not used for estimation to obtain model weights. Using the monetary shock application, we now illustrate the role of the sample splitting scheme.

First, we ask how much our results change the predictive densities were taken in-sample rather than out-of-sample. To that end, instead of splitting the sample, we estimate the model on the full data sample and compute in-sample predictive densities using the full sample estimates. Next, we ask how the number of subsamples matters. In particular, we split the sample into five blocks instead of two. For each block, we compute the predictive densities using estimates from the remaining 4 blocks.

Figures A.6 and A.7 show that using in-sample predictive densities substantially change the results. In particular, with in-sample predictive densities, majority of the weight is placed on the LPs, especially at longer horizons. This reverses the result in Section 4.1 that the VAR tended to receive more weight at longer horizons. Intuitively, the additional flexibility the LPs offer allows them to fit the data more closely, whereas the VAR tightly links the forecasts of different horizons. However, a better in-sample fit does not necessarily translate to better out-of-sample forecasts, as evidenced by the contrast between in-sample and out-of-sample weights. The resulting averaged impulse responses also differ markedly, especially for CPI, where the in-sample weights produce an averaged response that continues to decline even five years after the initial shock.

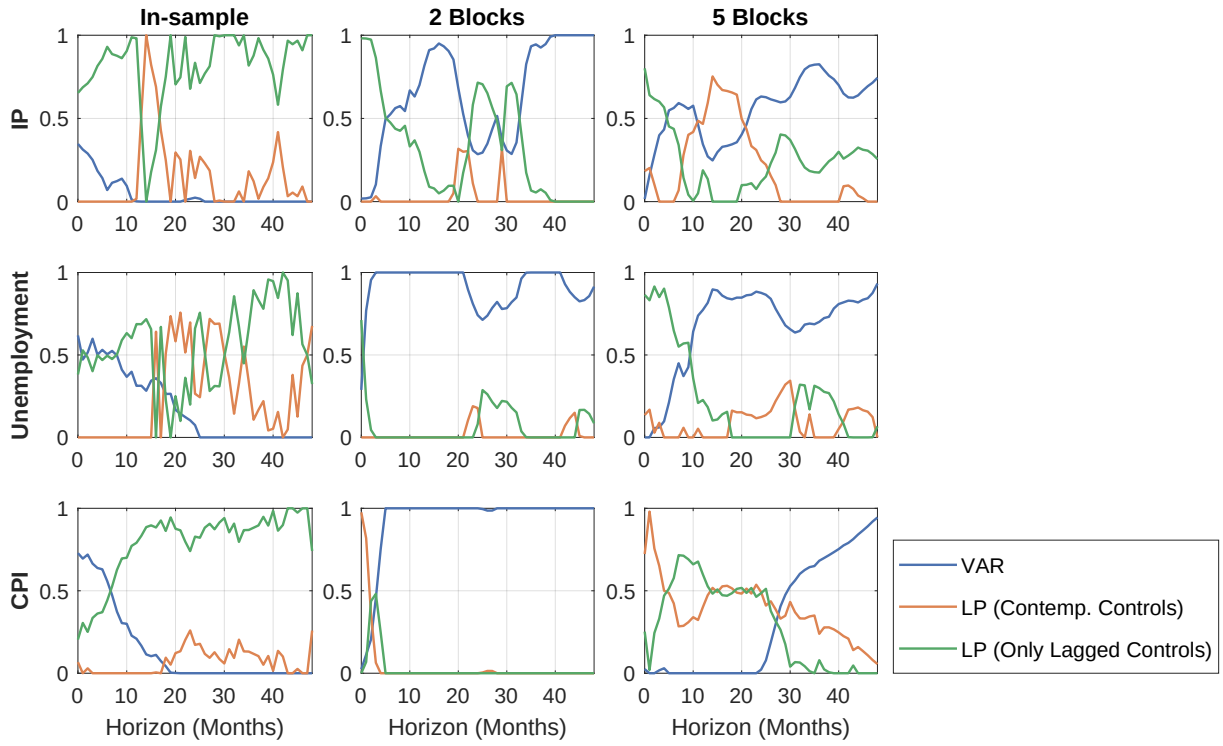

Figure A.6: Prediction pool weights from monetary shock empirical application with different sample splits. **Left:** No sample splitting; **Middle:** Two blocks; **Right:** Five blocks.

Splitting the sample into more blocks impacts the results much less. While overall there is more weight placed on the LPs, the VAR continues to receive a majority of the weight at long horizons. The corresponding averaged impulse responses look similar to the two block case, but with slightly wider error bands.

D.2 Changing the Pool of Models

To show how the addition of models affects the weights and averaged impulse response, we repeat the monetary shock exercise in Section 4.1 with an additional model:

4. **Cholesky VAR.** Following Coibion (2012), we estimate a VAR with the log of IP, the unemployment rate, the log of the CPI, and the log of the commodity price index in the first block, followed by the cumulated Romer and Romer (2004) instrument ordered last. The monetary shock is assumed to be the last shock from a Cholesky

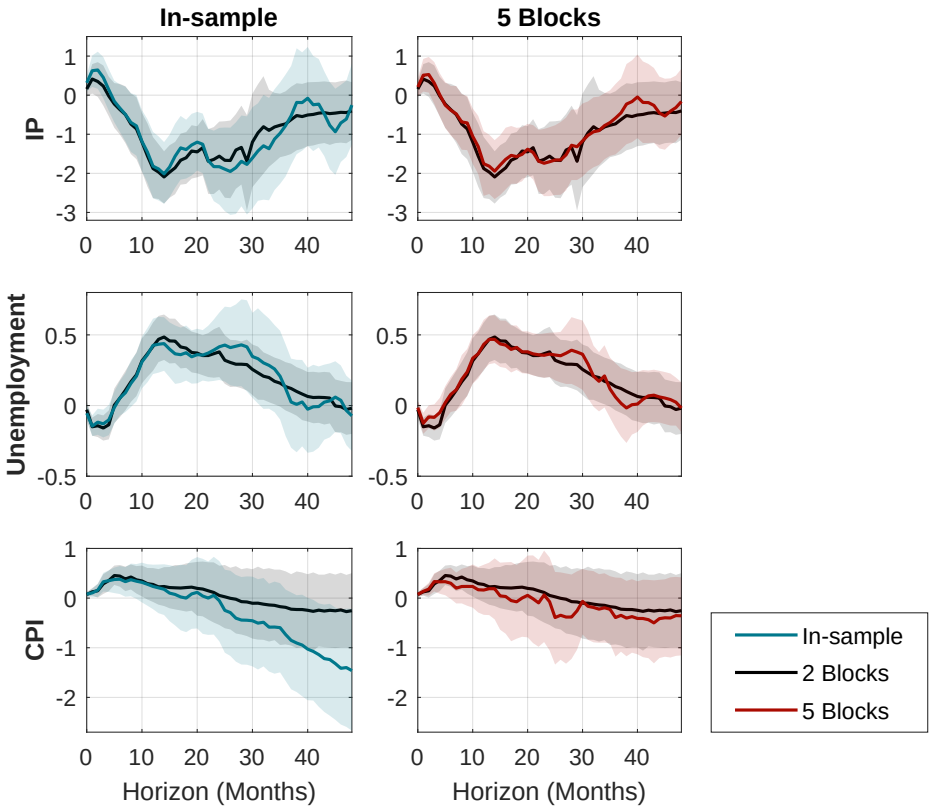

Figure A.7: Average impulse responses (mean and 68% error bands) from monetary shock empirical application with different sample splits. **Left:** No sample splitting; **Right:** Five blocks. Black lines and gray shaded regions correspond to benchmark with two blocks; colored lines and shaded regions correspond to alternative sample splits.

decomposition.

We are assuming that using the cumulated instrument ordered last in a Cholesky decomposition estimates the same shock as the internal instrument VAR and LPs from the original exercise in Secction 4.1. We view this as reasonable since the same underlying instrument is used in each model. However, the using the instrument in a Cholesky decomposition does make the assumption less straightforward, motivating our focus on the three models in the main text.

The results are shown in Figure A.8. The main difference in weights is on the internal instrument VAR, which is the closest model to the Cholesky VAR both in terms of specification and estimated impulse responses. We find that the weights on the LP remain similar

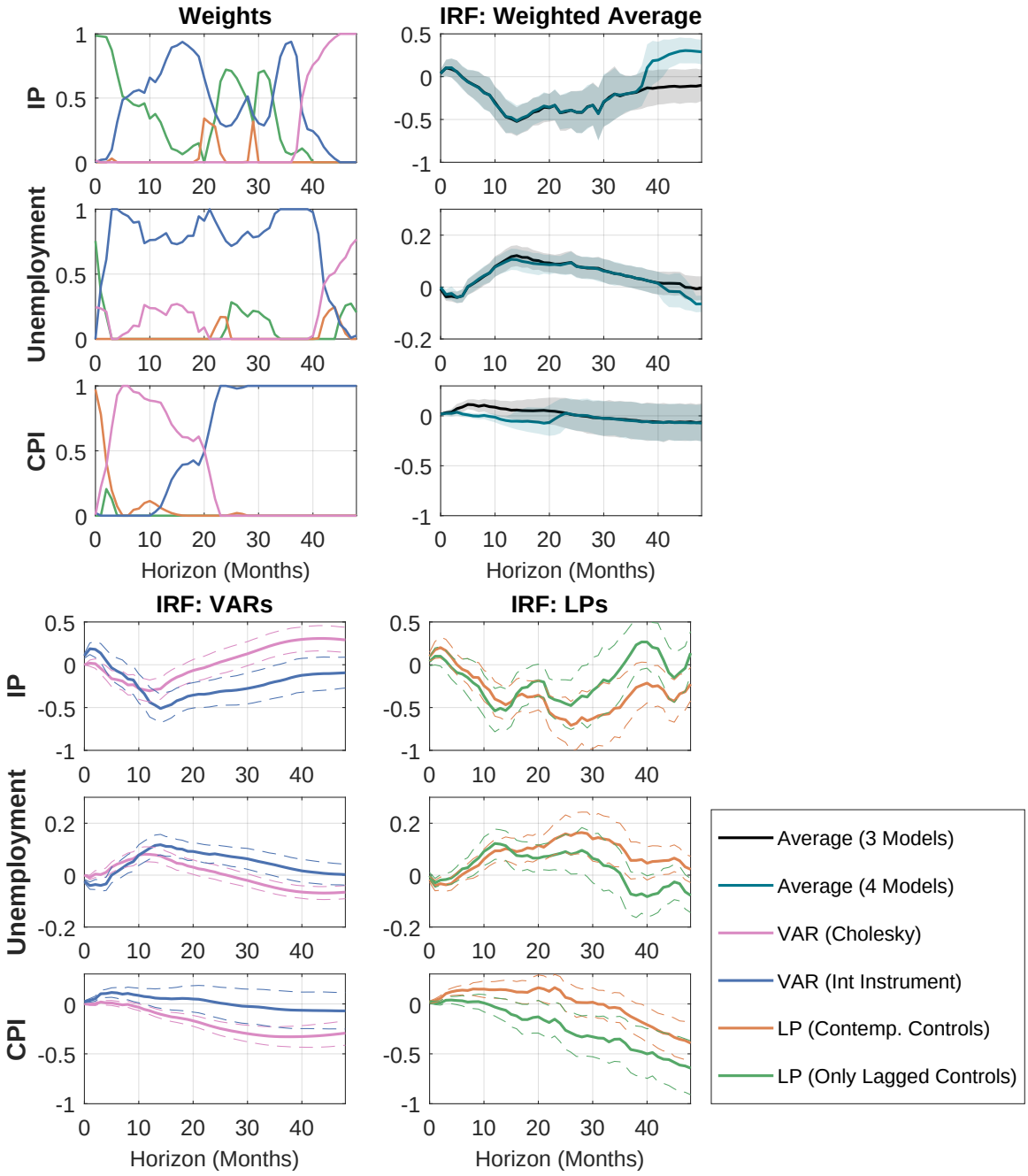

Figure A.8: Prediction pool weights and impulse responses from monetary shock empirical application with additional Cholesky VAR model. **Top Left:** Optimal weights on each model; **Top Right:** Mean and 68% error bands for averaged impulse responses; **Bottom:** Mean and 68% error bands for individual models.

to the benchmark exercise in Section 4.1 with three models. The averaged impulse responses remain relatively similar, illustrating the consistency of the weights assigned as we change the set of models (see Appendix A for details).

D.3 Alternative Instruments

Finally, we repeat the exercise for three different high-frequency identification instruments for monetary shocks instead of the narrative instruments in Section 4.1. In each case, we use the same models and variables.

The three instruments we consider are from Gertler and Karadi (2015), Miranda-Agrippino and Ricco (2021), and Jarociński and Karadi (2020). The corresponding sample periods are January 1990 through June 2012, January 1991 through December 2015, and February 1990 through June 2019, respectively, which are chosen to maximize the sample length for each instrument.

Figure A.9 shows the model weights for each of the instruments. While there are some differences across instruments, the common feature is that the VAR tends to be favored more at longer horizons. Notably, this downweights some less plausible impulse responses, such as the LPs finding a negative response of unemployment at the three- to four-year horizon with the Jarociński and Karadi (2020) instrument.

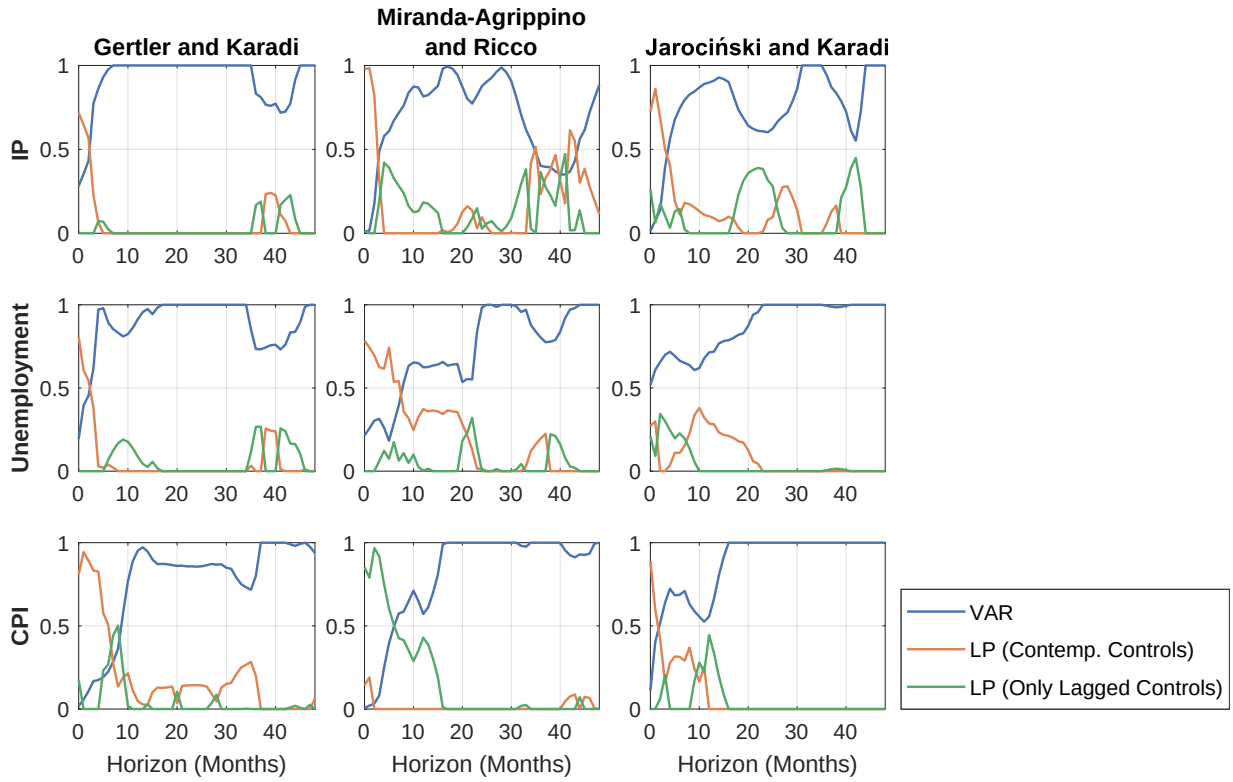

Figure A.9: Prediction pool weights from monetary shock empirical application with alternative instruments. **Left:** Gertler and Karadi (2015); **Middle:** Miranda-Agrippino and Ricco (2021); **Right:** Jarociński and Karadi (2020).

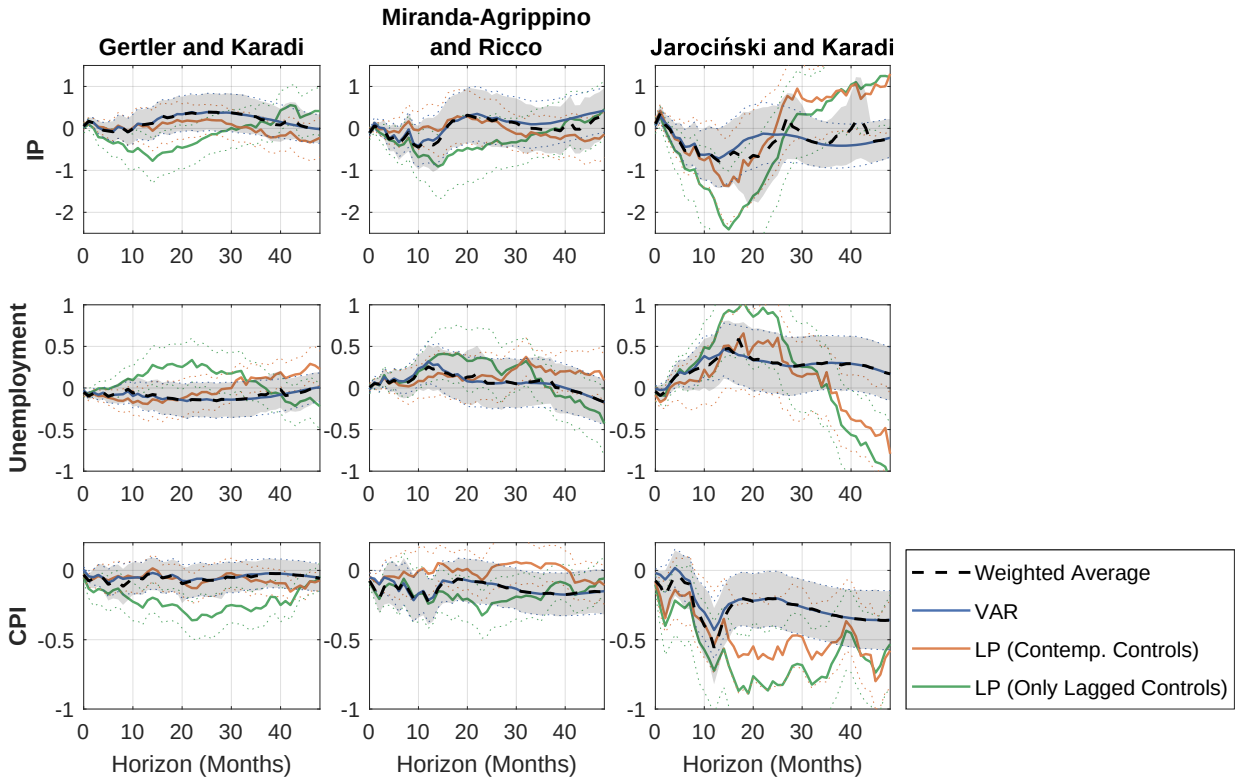

Figure A.10: Average impulse responses to 25bp monetary policy shock, identified with high frequency instruments. **Left:** Gertler and Karadi (2015); **Middle:** Miranda-Agrippino and Ricco (2021); **Right:** Jarociński and Karadi (2020).

E Bootstrap Procedure to Construct Predictive Densities for Frequentist Local Projections

In order not to have to assume Gaussianity of the shocks for the frequentist local projection estimators, we use a bootstrap procedure, which we describe below. We utilize the fact that given the distribution of β, we know the distribution of $X'\beta$. Therefore, we can use the bootstrap to obtain draws of the shocks, which in turn gives us a sample of $X'\beta$. We then integrate over this distribution of shocks.

For each horizon and each model, we run OLS and compute White standard errors for the relevant subsample. This gives us a normal distribution $N(\hat{\beta}, \hat{\Sigma})$ for the coefficients. To compute the predictive densities, we do the following:

1. Take N draws of the parameters from $\mathcal{N}(\hat{\beta}, \hat{\Sigma})$ and call these draws $\beta_1, \ldots, \beta_N$.

2. For each draw i, compute the fitted shocks. This gives us a sample of $N \times \frac{T}{2}$ fitted shocks $e_{i,t}$, where $i = 1, \ldots, N$ and $t = 1, \ldots, \frac{T}{2}$.

3. Draw from the sample of shocks with replacement to get a set of shocks $u_{1,\frac{T}{2}+1}, \ldots, u_{N,T}$, where $u_{i,t} \in \{e_{1,1}, \ldots, e_{N,\frac{T}{2}}\}$.

4. Compute the predicted value $pred_{i,t} \equiv Y_{i,t} - u_{i,t}$ for $X_t\beta$.

5. Compute the predictive density of $pred_{i,t} \sim \mathcal{N}(\hat{X_t}'\hat{\beta}, \hat{X_t}'\hat{\Sigma}\hat{X_t})$, where $\hat{X_t}$ is the predicted value of X_t.[19]

6. Average this density over all N draws.

The intuition behind the steps is as follows. Step 1 gives draws from the distribution for β. Step 2 then gives draws from the approximate marginal distribution for the shocks. Steps 3-5 then compute the density conditional on a draw of the shock. Finally, Step 6 integrates over the distribution of shocks.

[19] For simplicity, we can take $X_t = \hat{X}_t$, which is exact when X_t does not contain contemporaneous endogenous variables.