RESEARCH ARTICLE

Differences From Differencing: Should Local Projections With Observed Shocks Be Estimated in Levels or Differences?

Jeremy Piger[1] | Thomas Stockwell[2]

[1]Department of Economics, University of Oregon, Eugene, Oregon, USA | [2]Department of Economics, Sykes College of Business, University of Tampa, Tampa, Florida, USA

ABSTRACT

We show there are large finite-sample estimation improvements from estimating local projections (LP) in a cumulated differences (long-differenced) specification vs. a specification in levels when the impulse response of interest is to an externally identified ("observed") shock. The long-differenced specification substantially reduces, and in many cases eliminates, estimation bias, as well as significantly improves confidence interval coverage. These improvements increase for more persistent processes, at longer horizons, and for smaller sample sizes and extend to the instrumental variable LP (LP-IV) setting where the observed shock is an instrument for an endogenous variable of interest. We demonstrate these results via simulations as well as illustrative analytic results.

JEL Classification: C01, C14, C22, C26, C32, C54, E52

1 | Introduction

Following Jordá (2005), local projections (LP) have become a popular approach to estimate impulse response functions (IRFs). In the empirical macroeconomics literature specifically, LP are now commonly used as an alternative to the usual IRFs estimated via vector autoregressive (VAR) models. The popularity of LP arises in large part to their ease of use. LP are simple to estimate and draw inference on, requiring only the use of single equation linear regressions. The structure of LP also makes it straightforward to accommodate state-dependent and nonlinear specifications, though recent results in Gonçalves et al. (2024) call into question the consistency and interpretation of state-dependent IRFs estimated by LP.[1] Beyond their ease of implementation, LP place few restrictions on the shape of the IRF, contributing to confidence intervals that are substantially more robust to misspecification than VAR models (Montiel Olea et al. 2024).[2] As LP increases in

popularity, there is a growing theoretical literature studying the asymptotic properties of LP and their relation to VAR models.[3]

The popularity of LP has also benefited from a corresponding growth in macroeconomic studies that estimate the dynamic response of variables to externally identified, "observed", shocks (Ramey 2016; Stock and Watson 2018). In these studies, a researcher obtains an exogenous shock of interest through some procedure external to the estimation of the impulse response. Once this exogenous shock is in hand, a simple and natural approach to estimate the h horizon impulse response is to regress a response variable of interest at horizon h on the observed shock at time t. This type of estimation is exactly what the LP provides and as studies with observed shocks have proliferated, so has the use of LP. In the remainder of this paper, we will refer to these observed shock LP regressions as simply LP, while acknowledging that "LP" is a term used to also represent settings

where shocks are identified internally to the LP regression. An increasingly common extension of LP is to consider the observed shocks not as the ultimate shock of interest but instead as an instrument for an endogenous treatment (Jordá et al. 2015). In the following, we refer to these instrumental variable LP as LP-IV.

In this paper, we will be interested in a necessary specification choice made when estimating LP and LP-IV and how this choice impacts the performance of the associated impulse response estimators. Specifically, in the literature employing LP (LP-IV), there are differences in the way the response variable is specified, as well as how lagged response variables enter as controls. Consider a response variable y_t and suppose we are interested in the horizon h impulse response estimate, where throughout the paper we assume that h is a finite, nonnegative integer. Most studies specify the LP (LP-IV) regression in levels, where y_{t+h} is the estimand and values of $y_{t-1}, y_{t-2}, \dots, y_{t-p_L}$ are possibly included as controls. A smaller number of studies use a cumulated differences, or *long-differenced*, specification, in which the estimand is $(y_{t+h} - y_{t-1})$ and values of $\Delta y_{t-1}, \Delta y_{t-2}, \dots, \Delta y_{t-p_D}$ are possibly included as controls. In some references, both the levels and long-differenced specification are presented (Stock and Watson 2018; Jordá and Taylor 2025). However, most recent applications have focused on the levels specification, and Li et al. (2024) note that the use of data in levels represents applied practice.

The levels vs. long-differenced specification estimate the same impulse response and should provide comparable estimates in large samples. At the same time, there is now ample evidence that standard OLS estimates of IRFs via LP specified in levels are biased and produce incorrect confidence intervals in finite samples, particularly in the relatively small sample sizes used in the empirical macroeconomics literature. Using simulations, Kilian and Kim (2011) find that asymptotic confidence intervals from LP are less accurate than bias-adjusted VAR bootstrap confidence intervals, though this analysis relates to LP regressions for which shocks are identified internally to the estimation, which is not the focus of our paper. Herbst and Johannsen (2024) document that LP estimated with observed shocks are in practice often used with small samples in the time dimension and that point estimates of IRFs from LPs can be severely biased on these sample sizes. This is especially true when the process under consideration is persistent, as is the case with most macroeconomic series of interest. Building on these results, a small number of papers have presented attempts to reduce finite-sample bias and improve the accuracy of confidence intervals in LP regressions. Herbst and Johannsen (2024) use an approximate bias function to characterize and partially account for the bias in the LP regression, while Montiel Olea and Plagborg-Møller (2021) find that bootstrapped LP generates improved confidence interval accuracy in finite samples.

These studies finding finite-sample bias in LP regressions have focused on LP specified in levels and have not considered the performance of LP specified in long differences. Despite their large sample equivalence, the demonstrated poor performance of the levels LP specification in empirically relevant sample sizes leaves open the possibility that long-differenced specifications may provide improvements. In this paper, we fill this gap by conducting a simulation study to evaluate the finite-sample performance of LP and LP-IV specified in levels vs. long differences.

We begin with the example of an AR(1) with i.i.d. disturbances and demonstrate analytically that long-differencing has the promise of substantially reducing a particular source of small sample bias that exists in levels LP when the true data generating process (DGP) is stationary but persistent. In the unit root case, this particular advantage disappears, but a new one emerges, which is that the long-differenced LP imposes a correct parameter restriction on the levels specification. Then, using a wide variety of DGPs for empirically relevant sample sizes, we show using simulations that the long-differenced specification substantially reduces bias and improves confidence interval accuracy over LP regressions specified in levels. These improvements are larger as the persistence of the impulse response increases, as the sample size shrinks, and as the horizon of the response increases. Even for data that are less persistent, the long-differenced specification does not demonstrate any apparent disadvantages over the levels regression in terms of estimation bias or confidence interval coverage. Some of the DGPs considered that the long-differenced estimator displays higher simple variance than the levels estimator, though the size or existence of this difference is not uniform across DGPs and parameter calibrations. Overall, the long-differenced specification appears to be an effective approach to reduce bias and improve the accuracy of confidence intervals in LP and LP-IV estimation of IRFs.

As an application, we revisit the effects of Jarociński and Karadi (2020) monetary policy shocks and Federal Reserve information shocks on US output and prices. We find notable differences in the estimated effects of these shocks from the levels vs. long-differenced LP specifications. Further, the long-differenced specification consistently produces estimated effects that are larger in absolute value.

The rest of this paper proceeds as follows: Section 2 reviews the LP approach to estimate IRFs with externally identified, observed, shocks and discusses standard inference techniques used in the literature. Section 3 uses the stylized example of an AR(1) to demonstrate the intuition for the improvements in estimation bias that come from the long-differenced specification. Extensions of these results to a VAR(p) are presented in the Appendix B. Section 4 presents our simulation study that considers estimation bias, confidence interval accuracy, and estimation variance for a variety of DGPs and practical estimation considerations. Section 5 discusses the application to estimation of the output effects of monetary policy shocks. Section 6 concludes the study.

2 | Local Projections

Suppose one observes an exogenous shock of interest, labeled s_t, and a response variable of interest, labeled y_t. We wish to measure the impulse response at horizon h, up to a finite maximum horizon H as follows:

$$\beta_h = \frac{\partial y_{t+h}}{\partial s_t}, h = 0, 1, \dots, H.$$

The levels specification of the LP to estimate β_h is simply the direct multistep ahead prediction as follows:

$$y_{t+h} = \beta_h s_t + \rho_{1,h} y_{t-1} + \rho_{2,h} y_{t-2} + \cdots + \rho_{p_L,h} y_{t-p_L} + \gamma'_h X_t + \upsilon_{t+h} \tag{1}$$

where the exogeneity of s_t implies that $E(s_t \upsilon_{t+h}) = 0$. In most applications of LP, lagged values of the response variable appear as controls, and we have explicitly allowed for p_L lags of the response variable in Equation (1). Additional controls can appear in the vector X_t and usually include deterministic terms, such as a constant or deterministic time trends. In some applications, lags of variables other than the response variable are also included.

We can alternatively estimate β_h using a cumulated differences specification. To begin, consider an LP where the response variable is the first difference of y_{t+h} as follows:

$$\Delta y_{t+h} = \tilde{\beta}_h s_t + \tilde{\rho}_{1,h} \Delta y_{t-1} + \tilde{\rho}_{2,h} \Delta y_{t-2} \\ + \cdots + \tilde{\rho}_{p_D,h} \Delta y_{t-p_D} + \tilde{\gamma}'_h \tilde{X}_t + \tilde{\upsilon}_{t+h}, \tag{2}$$

where $\Delta y_{t+h} = y_{t+h} - y_{t+h-1}$ and $\tilde{\beta}_h$ is the impulse response of Δy_{t+h} to the shock s_t. We can then recover β_h as follows:

$$\beta_h = \sum_{i=0}^{h} \tilde{\beta}_i \tag{3}$$

One could estimate β_h by first estimating Equation (2) and then forming the sum in Equation (3). However, as pointed out by Stock and Watson (2018), we can instead first sum Equation (2), providing the following equation to estimate β_h directly:

$$\Delta_h y_{t+h} = \beta_h s_t + \theta_{1,h} \Delta y_{t-1} + \theta_{2,h} \Delta y_{t-2} \\ + \cdots + \theta_{p_D,h} \Delta y_{t-p_D} + \alpha'_h X^D_t + u_{t+h}, \tag{4}$$

where $\Delta_h y_{t+h} = y_{t+h} - y_{t-1}$. We refer to Equation (4) as the "long-differenced" specification.[4]

While the impulse responses at alternative horizons could be estimated by treating the H equations as a seemingly unrelated regression that is estimated jointly, it is common in the applied LP literature to estimate via equation by equation OLS. Also, as discussed in Jordá (2005), the disturbance terms in Equations (1) and (2) are serially correlated and follow a moving average (MA) process. Because of this, much of the literature makes use of robust standard errors to compute confidence intervals for β_h, with the Newey–West methodology being a popular choice. The disturbance term in Equation (4) is further complicated by the summation of errors from Equation (2). In the remainder of this paper, we will evaluate the performance of equation by equation OLS estimation of the LP in both the levels and long-differenced specification, as well as the performance of the Newey–West methodology for computing standard errors. Recent studies by Montiel Olea and Plagborg-Møller (2021) and Herbst and Johannsen (2024) have argued for the use of heteroskedasticity robust standard errors, rather than HAC standard errors, when conducting inference with LP. Thus, in Section 4.4, we consider the robustness of our results to the use of the Eicker–Huber–White (EHW) standard errors.

We will also be interested in the common case where s_t is endogenous, but we have an available instrument, labeled ε_t (Stock and Watson 2018). In this case, we can implement LP using instrumental variable methods, commonly known as LP-IV (Jordá et al. 2015). For example, taking a two-stage least squares approach, $\hat{s}_t$ will replace s_t in the levels and long-differenced specifications as follows:

$$y_{t+h} = \beta_h \hat{s}_t + \rho_{1,h} y_{t-1} + \rho_{2,h} y_{t-2} + \cdots + \rho_{p,h} y_{t-p} + \gamma'_h X_t + \upsilon^{IV}_{t+h} \tag{5}$$

$$\Delta_h y_{t+h} = \beta_h \hat{s}_t + \theta_{1,h} \Delta y_{t-1} + \theta_{2,h} \Delta y_{t-2} \\ + \cdots + \theta_{p,h} \Delta y_{t-p} + \alpha'_h X^D_t + u^{IV}_{t+h}, \tag{6}$$

where $\hat{s}_t$ is the fitted value from following first stage regression:

$$s_t = \gamma \varepsilon_t + \delta' X^{IV}_t + \eta_t \tag{7}$$

and X^{IV}_t contains controls, including deterministic terms, for the first stage regression. Conditions for instrument validity are outlined in Stock and Watson (2018) and include a standard relevance condition, as well as a lead-lag exogeneity condition that requires ε_t to be uncorrelated with $\upsilon_{t+h}, \forall h$.

3 | An Illustrative Example Based on an AR(1)

In this section, we consider a specific DGP, an autoregression of order 1 (AR(1)). Herbst and Johannsen (2024) derive the approximate finite-sample bias for the levels LP in the AR(1) case. Here, we present analytic expressions that will aid our intuition regarding the relative effectiveness of estimating LP via the levels vs. long-differenced specification. Appendix A presents additional details behind these expressions, while Appendix B extends the analysis to a VAR(p).

The AR(1) DGP is as follows:

$$y_t = \alpha + \beta_0 s_t + \phi y_{t-1} + \omega_t \tag{8}$$

The observed shock of interest is s_t and at this point is assumed to be strictly exogenous, so that $E(s_t \omega_{t+j}) = 0, \forall j$. By virtue of s_t being a "shock" in the traditional sense, we assume that $E(s_t s_{t+j}) = 0, \forall j \neq 0$. For simplicity, we assume that $s_t \sim$ i.i.d. (μ_s, σ^2_s) and $\omega_t \sim$ i.i.d. $(0, \sigma^2_\omega)$, where extension to the heteroskedastic case would be straightforward. We set $\mu_s = 0$ without loss of generality. We begin by assuming $|\phi| < 1$, thereby focusing attention on the stationary case where the benefits of long-differencing might be considered dubious a priori. Later in this section, we will consider the unit root case where $\phi = 1$.

The correct levels specification for the LP is as follows:

$$y_{t+h} = c^L_h + \beta_h s_t + \rho_{1,h} y_{t-1} + \upsilon_{t+h} \tag{9}$$

where $\beta_h = \beta_0 \phi^h$, $\rho_{1,h} = \phi^{h+1}$, and $\upsilon_{t+h} = \sum_{i=0}^{h-1} \beta_i s_{t+h-i} + \sum_{i=0}^{h} \phi^i \omega_{t+h-i}$. Despite the fact that s_t is uncorrelated with each of the values of $s_{t+j}, j > 0$ and $\omega_{t+j}, j \geq 0$ that contribute to the regression disturbance υ_{t+h}, the OLS estimate of β_h will be biased in finite samples in part because of an expected nonzero sample covariance between s_t and υ_{t+h}. This nonzero expected sample

covariance arises because of the interaction of the sample means of s_t and v_{t+h} in the sample covariance formula. Specifically, Appendix A shows that the expected sample covariance, $cov_{s_t,v_{t+h}}$, is as follows:

$$
\begin{aligned}
E\left(cov_{s_t,v_{t+h}}\right) &= -\sum_{i=0}^{h-1} \beta_i E\left(\overline{s}_{[0]}\overline{s}_{[h-i]}\right) \\
&= \frac{\sigma_s^2}{T^2}\left[-\sum_{i=0}^{h-1} \beta_i (T - h + i)\right],
\end{aligned}
\tag{10}
$$

where $\overline{s}_0$ is the sample mean of s_t calculated over the period $t = 1 \rightarrow T$ and $\overline{s}_{[h-i]}$ is the sample mean of s_t calculated over the period $t = (1 + h - i) \rightarrow (T + h - i)$.

Equation (10) provides several elements of intuition regarding the expected bias in the OLS estimate of β_h. First, the size of the expected covariance between s_t and v_{t+h} depends on the value of $\beta_i = \beta_0\phi^i$ for $i = 0, \dots h - 1$. In other words, the expected covariance depends on the value of the true IRF at all horizons up to horizon $h - 1$. The more persistent the IRF, the larger will be these terms in absolute value, which increases the covariance in absolute value. Second, the expected covariance will grow in absolute value with the horizon h. Third, the sample size influences the size of the expected covariance. As T grows, the denominator grows with respect to the numerator and shrinks the size of the covariance.

Herbst and Johannsen (2024) derive an analytic approximation to the finite-sample bias in the levels LP estimator of β_h for the AR(1) case, and it is worth connecting our results to their approximation. Equation (6) in Herbst and Johannsen (2024) decomposes the approximate finite-sample bias into two components. The first arises because of the need to estimate the intercept in Equation (9). This source of bias corresponds to the nonzero expected covariance in Equation (10), which would disappear if there was no need for an intercept in the levels specification of the LP. Specifically, if the true DGP did not include an intercept and the intercept was omitted from Equation (9), then the term $E\left(\overline{s}_{[0]}\overline{s}_{[h-i]}\right)$ would cease to be relevant for bias as sample means would no longer appear in the OLS estimator. The second source of bias documented by Herbst and Johannsen (2024) is related to the need to estimate $\rho_{1,h}$ in Equation (9), where this bias grows with the persistence of y_{t-1}[5].

Figure 1 shows results of an initial simulation experiment based on the AR(1) DGP in (8), where we consider two sample sizes $T = \{100; 200\}$ and three persistence levels $\phi = \{0.70, 0.90, 0.95\}$. We set $\beta_0 = 1$ so that $\beta_h = \phi^h$, $\alpha = 0$, and $\sigma_s^2/\sigma_\omega^2 = 1$.[6] Both disturbances, s_t and ω_t, are generated from normal distributions. Each panel of the figure shows the true IRF (solid line), as well as the average value of $\hat{\beta}_h$ (dashed line) across 5000 simulations, where $\hat{\beta}_h$ is the OLS estimate from the levels LP specification in (9). The results of the simulation confirm the intuition provided above. Specifically, there is finite-sample bias in $\hat{\beta}_h$, and this bias increases in magnitude as the persistence of the true IRF rises and as the sample size decreases. To assess the source of the bias, we repeat these simulation where the intercept in Equation (9) is set equal to its true value of zero. The results for this case are presented in Figure 2 and show that the bias in $\hat{\beta}_h$ is eliminated when $\phi = \{0.7, 0.9\}$. When $\phi = 0.95$, a very substantial portion of the bias is eliminated, but we begin to see bias remaining due to the high persistence of y_{t-1}, which was the second source of bias identified by Herbst and Johannsen (2024).

We now consider how bias may be mitigated by estimating the long-differenced LP. The correctly specified long-differenced LP
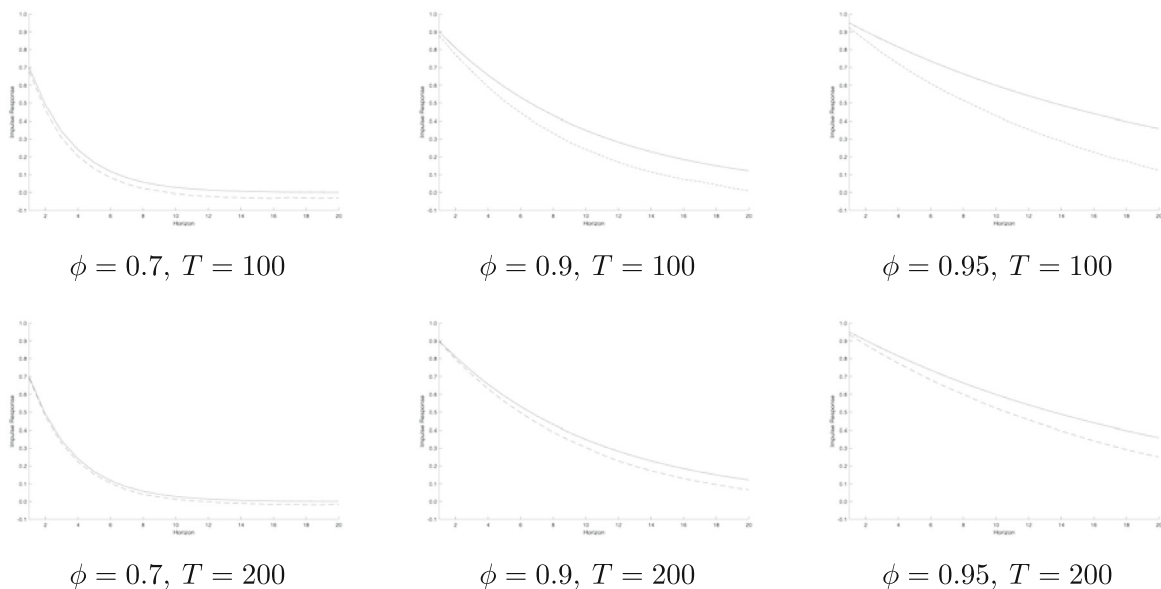

$\phi = 0.7, \; T = 100$	$\phi = 0.9, \; T = 100$	$\phi = 0.95, \; T = 100$
$\phi = 0.7, \; T = 200$	$\phi = 0.9, \; T = 200$	$\phi = 0.95, \; T = 200$

FIGURE 1 | Estimated IRF from levels LP and AR(1) DGP. *Note:* This figure displays simulation results from estimation of the levels specification of the LP when the true DGP is an AR(1) model. Results for three alternative values of the autoregressive parameter ($\phi = \{0.7, 0.9, 0.95\}$), as well as two alternative sample sizes $T = \{100, 200\}$, are displayed. Each figure shows the average impulse response function estimate for the levels specification (dashed line) and the true impulse response function (solid line).
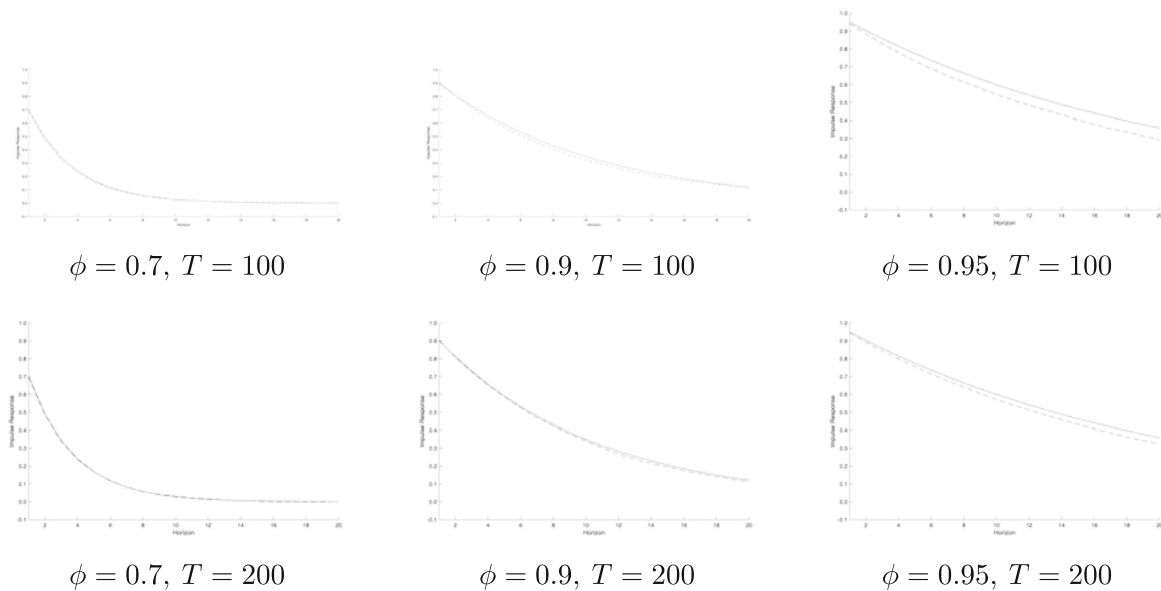
$\phi = 0.7,\ T = 100$ $\phi = 0.9,\ T = 100$ $\phi = 0.95,\ T = 100$

$\phi = 0.7,\ T = 200$ $\phi = 0.9,\ T = 200$ $\phi = 0.95,\ T = 200$

FIGURE 2 | Estimated IRF from levels LP and AR(1) DGP. *Note:* This figure displays simulation results from estimation of the levels specification of the LP when the true DGP is an AR(1) model with no intercept, and no intercept is included in the LP regression. Results for three alternative values of the autoregressive parameter ($\phi = \{0.7, 0.9, 0.95\}$), as well as two alternative sample sizes $T = \{100, 200\}$, are displayed. Each figure shows the average impulse response function estimate for the levels specification (dashed line) and the true impulse response function (solid line).

for the AR(1) DGP is as follows:

$$\Delta_h y_{t+h} = c_h^D + \beta_h s_t + \theta_{1,h}\Delta y_{t-1} + \cdots + \theta_{h+1,h}\Delta y_{t-h-1} + u_{t+h} \tag{11}$$

where $\beta_h = \beta_0 \phi^h$, $\theta_{i,h} = \phi^{h+1}$ and

$$u_{t+h} = \sum_{i=0}^{h-1} \beta_i (s_{t+h-i} - s_{t-1-i})$$
$$- \beta_h s_{t-h-1} + \sum_{i=0}^{h} \phi^i (\omega_{t+h-i} - \omega_{t-1-i}).$$

As shown in the Appendix A, the expected sample covariance between s_t and u_{t+h} is as follows:

$$E(cov_{s_t, u_{t+h}}) = \beta_h E\left(\overline{s}_{[0]}\overline{s}_{[-(h+1)]}\right)$$
$$- \sum_{i=0}^{h-1} \beta_i E\left(\overline{s}_{[0]}\left(\overline{s}_{[h-i]} - \overline{s}_{[-(i+1)]}\right)\right) \tag{12}$$
$$= \frac{\sigma_s^2}{T^2}\left[\beta_h(T-h-1) - \sum_{i=0}^{h-1}\beta_i(1-h+2i)\right].$$

The expected sample covariance from the long-differenced specification in (12) will in general be much smaller than that from the levels specification in (10). In other words, the observed shock, s_t, will display less expected correlation with the regression disturbance in the long-differenced LP than the levels LP. Figure 3 displays the expected sample covariance from (10) and (12) for the case where $T = \{100, 200\}$, $\sigma_s^2 = 1$, $\beta_0 = 1$, and for three values of persistence, $\phi = \{0.7, 0.9, 0.95\}$. The figure shows that the expected sample covariance between s_t and the levels LP regression disturbance is increasing in absolute value in both horizon and persistence, whereas this is not the case for the

long-differenced LP regression. Also, the expected sample covariance is larger in absolute value for the levels regression for all horizons beyond $h = 1$.

The source of the reduction in the expected sample covariance term can be seen through comparison of Equations (10) and (12). In Equation (10), each of the expectations $E(\overline{s}_{[0]}\overline{s}_{[h-i]})$, $i = 0, 1, \ldots, h-1$ creates $(T-h+i)$ nonzero terms due to overlap between the samples used to calculate $\overline{s}_{[0]}$ and $\overline{s}_{[h-i]}$. By contrast, in (12), each of the expectations $E\left(\overline{s}_{[0]}\left(\overline{s}_{[h-i]} - \overline{s}_{[-(i-1)]}\right)\right)$, $i = 0, 1, \ldots, h-1$ creates only $(1-h+2i) << (T-h+i)$ nonzero terms, with this reduction due to cancelation of terms caused by the differencing in the expectation. In the end, regardless of the value of h, Equation (12) includes only a single expectation that does not include such a difference, that being $\phi^h E\left(\overline{s}_{[0]}\overline{s}_{[-(h+1)]}\right)$. By contrast, Equation (10) has h such terms. As such, the reduction in the expected sample covariance will be larger for larger h. Also, since these terms in Equation (10) are scaled by $\beta_h = \beta_0\phi^i$, $i = 0, 1, \ldots, h-1$, the reduction in the expected sample covariance seen in Equation (12) will be larger for higher values of ϕ.

To see the benefits of long-differencing for reducing bias, Figure 4 repeats the simulation experiment of Figure 1 but where $\hat{\beta}_h$ is the OLS estimate from the long-differenced LP specification in (11). The average value of $\hat{\beta}_h$ across 5000 simulations is displayed with a dash-circle line, while the true value of the IRF is again a solid line. The results of the simulation are striking. In all cases considered, long-differencing essentially erases any bias present for the levels LP specification. This initial simulation evidence suggests that long-differencing may be a powerful bias reduction strategy in LPs.

The benefits of long-differencing extend to the LP-IV setting where s_t is endogenous, but we have an available instrument,
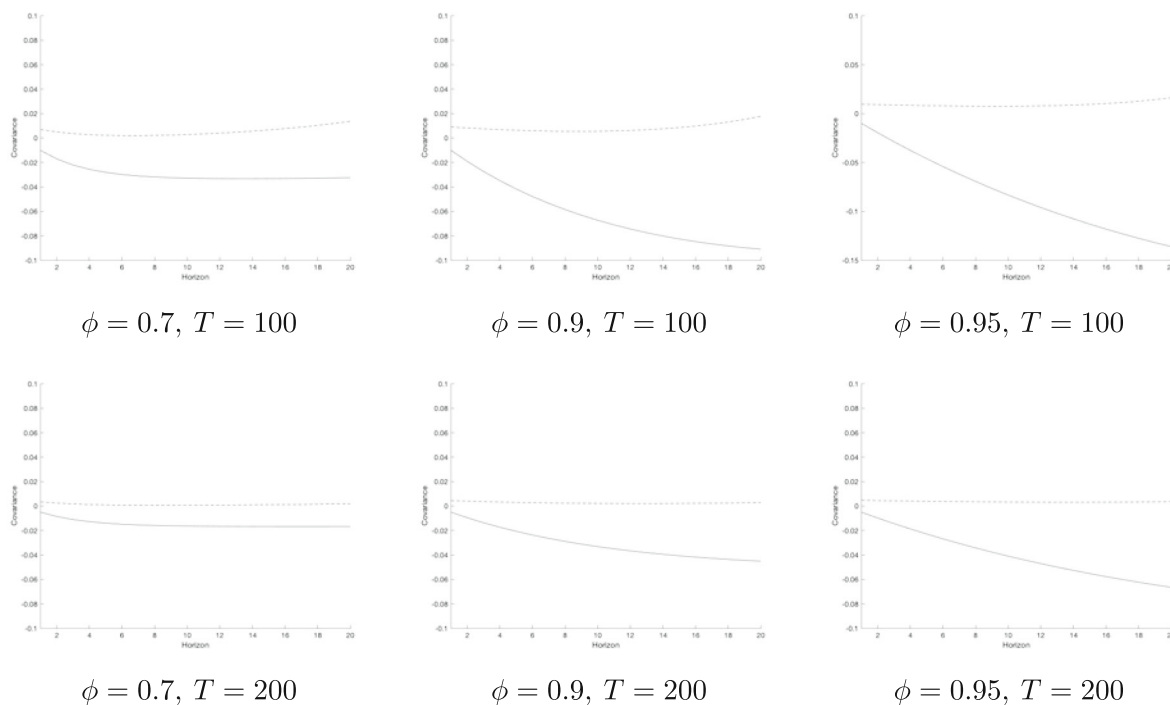
FIGURE 3 | Expected Sample Covariance between Observed Shock and LP Regression Disturbance. *Note:* This figure displays the expected sample covariance from Equations (10) and (12) when $T = \{100, 200\}$, and $\phi = \{0.7, 0.9, 0.95\}$. In each subfigure, the black solid line is the expected sample covariance from the levels specification of the LP, while the black dashed line is the expected sample covariance from the long-differenced specification of the LP.
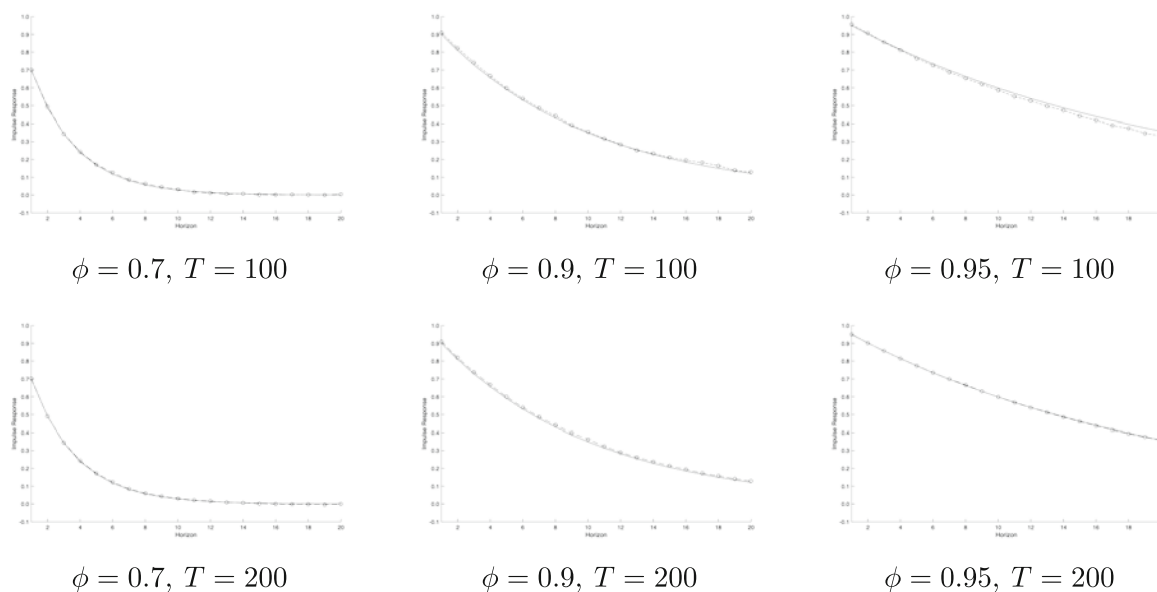

FIGURE 4 | Estimated IRF from long-differenced LP and AR(1) DGP. *Note:* This figure displays simulation results from estimation of the levels specification of the LP when the true DGP is an AR(1) model. Results for three alternative values of the autoregressive parameter ($\phi = \{0.7, 0.9, 0.95\}$), as well as two alternative sample sizes $T = \{100, 200\}$, are displayed. Each figure shows the average impulse response function estimate for the long-differenced specification (dash-circle line) and the true impulse response function (solid line).

labeled ε_t. Suppose the DGP again has AR(1) structure is as follows:

$$y_t = \alpha + \beta_0 s_t + \phi y_{t-1} + \omega_t,$$

where s_t is i.i.d. (μ_s, σ_s^2), ω_t is i.i.d. $(0, \sigma_\omega^2)$, and s_t is now endogenous such that $E(s_t \omega_t) \neq 0$. Without loss of generality, we set

$\mu_s = 0$. The instrument, ε_t, is assumed i.i.d. $(0, \sigma_\varepsilon^2)$ and satisfies a lead-lag exogeneity condition, $E(\varepsilon_t \omega_{t+j}) = 0, \forall j$. Assume that the relationship between ε_t and s_t is described by the first stage regression as follows:

$$s_t = \lambda + \gamma \varepsilon_t + \eta_t,$$

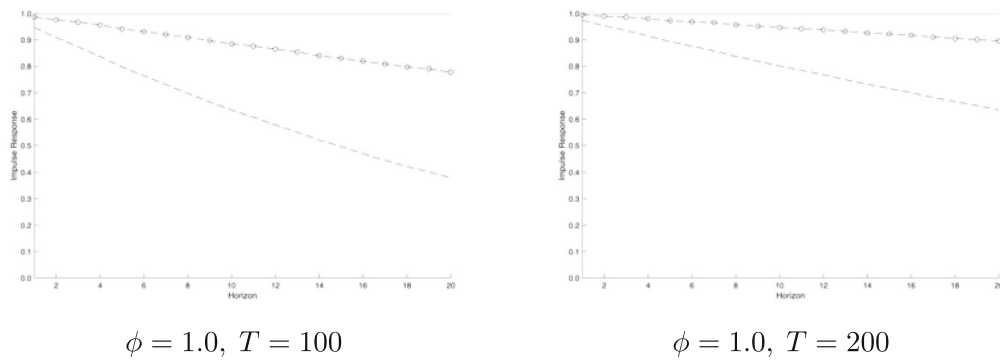$$\phi = 1.0, \ T = 100 \qquad\qquad \phi = 1.0, \ T = 200$$

FIGURE 5 | Estimated IRF for Unit Root AR(1) DGP. *Note:* This figure displays simulation results from estimation of the levels and long-differenced specification of the LP when the true DGP is an AR(1) model with $\phi = 1$. Results for two alternative sample sizes $T = \{100, 200\}$ are displayed. Each figure shows the average impulse response function estimate for the levels specification (dashed line), long-differenced specification (dash-circle line) and the true impulse response function (solid line).

where $E(\varepsilon_t \eta_t) = 0$, and without loss of generality, we set $\lambda = 0$. In order to provide comparable expectations, as well as to focus on the effects of long-differencing, we assume that γ is known. From the first-stage, we then obtain $\hat{s}_t = \gamma \varepsilon_t$, and the correctly specified LP-IV in levels is as follows:

$$y_{t+h} = c_h^L + \beta_h \hat{s}_t + \rho_{1,h} y_{t-1} + v_{t+h}^{IV}$$
$$v_{t+h}^{IV} = v_{t+h} + \beta_h \eta_t,$$

while the correctly specified long-differenced LP-IV is as follows:

$$\Delta_h y_{t+h} = c_h^D + \beta_h \hat{s}_t + \theta_{1,h}\Delta y_{t-1} + \cdots + \theta_{h+1,h}\Delta y_{t-h-1} + u_{t+h}^{IV},$$
$$u_{t+h}^{IV} = u_{t+h} + \beta_h \eta_t.$$

As shown in the Appendix A,

$$E\left(cov_{\hat{s}_t, v_{t+h}^{IV}}\right) = \kappa E\left(cov_{s_t, v_{t+h}}\right)$$
$$E\left(cov_{\hat{s}_t, u_{t+h}^{IV}}\right) = \kappa E\left(cov_{s_t, u_{t+h}}\right). \tag{13}$$

In other words, the expected covariance between the regressor of interest and the regression disturbance is proportional in the LP and LP-IV models, where this is true for both the levels and long-differenced specification with the same constant of proportionality. This implies that the relative benefits from long-differencing will be similar in the LP-IV and LP models.

The discussion above has focused on the case of the stationary AR(1) model. In the unit root case, we would not expect to see a mitigation in the bias associated with estimation of the intercept from use of the long-differenced specification. Returning to the LP case for simplicity, note that when $\phi = 1$, the correct levels specification is as follows:

$$y_{t+h} = c_h^L + \beta_h s_t + \rho_{1,h} y_{t-1} + v_{t+h},$$

where $\beta_h = \beta_0$, $\rho_{1,h} = 1$, and $v_{t+h} = \sum_{i=0}^{h-1}\beta_0 s_{t+h-i} + \sum_{i=0}^{h}\omega_{t+h-i}$. The correct long-differenced specification is as follows:

$$\Delta_h y_{t+h} = c_h^D + \beta_h s_t + u_{t+h}$$

where $\beta_h = \beta_0$, $c_h^D = c_h^L$, and $u_{t+h} = v_{t+h} = \sum_{i=0}^{h-1}\beta_0 s_{t+h-i} + \sum_{i=0}^{h}\omega_{t+h-i}$. Thus, in the unit root case, the regression disturbance is the same for the levels vs. long-differenced specification, and there is thus no difference in the finite sample expected covariance between s_t and the regression disturbance term from using one specification vs. the other. With that being said, we would still expect improved finite-sample performance from the long-differenced specification. As Herbst and Johannsen (2024) show, a component of the bias in the levels specification of the LP for the AR(1) case is due to the need to estimate $\rho_{1,h}$ in highly persistent processes. The long-differenced specification eliminates this requirement by correctly imposing the restriction $\rho_{1,h} = 1$.

As an initial investigation into the expected gains from imposing this restriction, Figure 5 extends the simulation experiments of Figures 1–4 to the case where $\phi = 1$. Here we see significant bias in both the levels and long-differenced LP. However, the long-differenced specification displays much less bias than the levels LP and indeed shows some of the largest bias *improvements* for any of the persistence levels considered. Thus, the bias reduction from long-differencing in the unit root case might be especially large.

The results in this section suggest that when the true DGP is a stationary AR(1), the long-differenced LP (LP-IV) will yield an impulse response estimate with less finite-sample bias than that produced by the levels LP (LP-IV). This reduction in bias comes through a mitigation of the correlation between the observed shock of interest and the LP regression disturbance in the long-differenced vs. the levels specification. In the unit root case, this mitigation disappears. However, we still see improved bias performance from the long-differenced specification in this case, as it correctly enforces restrictions imposed by the integration properties of the DGP.

The analytical and simulation results in this section were specific to the AR(1) process, and the simulations further assumed knowledge of the correct LP regression, including controls. With this illustrative example as motivation, in the next section, we conduct a broader range of simulation experiments to investigate the relative performance of the long-differenced LP (LP-IV) vs.

the levels LP (LP-IV) in empirically relevant settings, including a range of DGPs unknown to the econometrician, varying persistence levels, and varying sample sizes. Estimator performance will be assessed not only in terms of bias but also confidence interval coverage and estimator variance.

4 | Simulation Evidence

In this section, we present results of a simulation study using a variety of different DGPs to evaluate the performance of the levels (Equation (1)) and long-differenced (Equation (4)) LP specifications. The response variable is labeled y_{t+h} in all cases. For each of the DGPs for y considered, we assume that the true DGP is unknown to the econometrician. We begin by considering cases where the shock of interest, labeled s_t in all cases, is externally identified and available and turn in Section 4.5 to cases where s_t represents an endogenous treatment, and we observe an instrument ε_t. We consider both univariate and multivariate DGPs.

We set the control variables in Equations (1) and (4) as follows: We include p_L lags of the level of y_t in the levels specification and p_D lags of the first difference of y_t in the differences specification. For the levels specification, X_t includes a constant for univariate DGPs and additionally contains p_L lags of the level of additional endogenous variables beyond y_t for multivariate DGPs. For DGPs that produce data with trending behavior, X_t also includes a linear time trend.[7] For the differences specification, X_t^D contains a constant for univariate DGPs and additionally contains p_D lags of the first difference of additional endogenous variables beyond y_t for multivariate DGPs. When estimating each version of the LP models on the simulated data, we conduct data-based lag selection to select p_L and p_D via a test-down procedure, where the procedure begins with $p_{max} = 12$. The sensitivity of our results to lag selection is discussed in Section 4.4.1.

For each DGP, the results are based on 5000 simulations. We consider three sample sizes, $T \in \{100, 200, 300\}$. Herbst and Johannsen (2024) survey 71 recent empirical papers utilizing LPs and find that the median value of T across these studies is 95, while 20% have sample sizes as large as 200 and 6% have sample sizes as large as 300.[8] As increases in the sample size tend to have fairly monotonic effects on our simulation results, we present results for $T \in \{100, 200\}$ in the main text and record the results for $T = 300$ in the Supporting Information Appendix. We assess the bias of the OLS point estimates of β_h and accuracy of the Newey–West coverage intervals for impulse responses at horizons up to and including a maximum horizon of $H = 20$. In constructing the Newey–West standard errors, the maximum autocorrelation lag is set to $H + 1$ following Jordá (2005). In order to assess potential bias variance trade-offs in the levels vs. long-differenced LP estimators, we also report the simple standard deviation of the estimates of β_h.

4.1 | Higher Order Autoregressive Models

We begin by considering AR(p) models. As Section 3 already presented simulation results for an AR(1) model, we focus here on a higher order case, namely, an AR(8) as follows:

$$y_t = \alpha + \beta_0 s_t + \sum_{i=1}^{8} \phi_i y_{t-i} + \omega_t,$$

where $\omega_t \sim \text{i.i.d.} N(0, \sigma_\omega^2)$, $s_t \sim \text{i.i.d.} N(0, \sigma_s^2)$, and $E(s_t \omega_{t+j}) = 0, \forall j$. We set $\sigma_\omega^2 = \sigma_s^2 = 1$ and discuss sensitivity to this parameterization in Section 4.4.
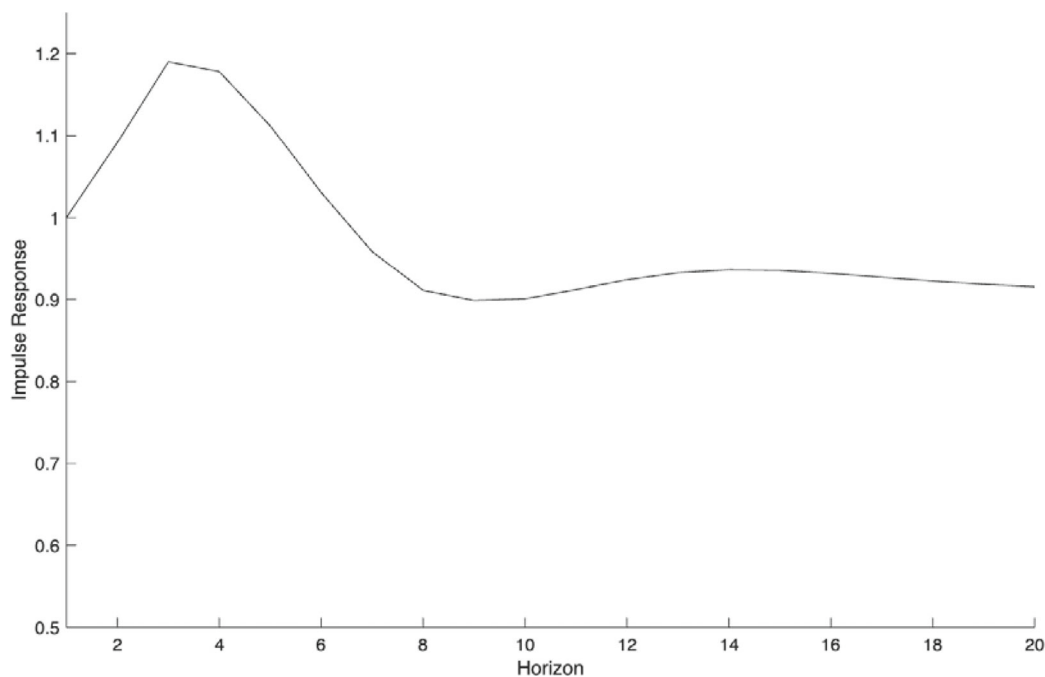

FIGURE 6 | Implied IRF from AR(8) Model Estimated on U.S. Real GDP. *Note:* This figure shows the impulse response function implied by an AR(8) model fit to U.S. log real GDP over the sample period 1947:Q1 - 2024:Q3.

We explore several different calibrations for this model, which differ in their level of persistence. To vary the level of persistence, we consider three alternative values for the sum of the AR coefficients, $\rho = \sum_{i=1}^{8} \phi_i$. The first specification features a process that is persistent, but clearly stationary in that unit root, tests will have very high power to detect the null of stationarity ($\rho = 0.70$); the second is a very persistent, though still stationary process ($\rho = 0.95$), while the third is a unit root process ($\rho = 1$). In all cases, we set the intercept $\alpha = 0$. To set the specific AR parameters, which will determine the shape of the IRF, we first fit an AR(8) model to quarterly log real US GDP from 1947:Q1 to 2024:Q3, the largest dataset available at the time of this writing. Figure 6 shows the IRF implied by this estimation, which follows the typical "hump shaped" pattern often seen in the US macroeconomic data. In our simulations, we then scale the autoregressive parameter estimates from the estimated AR(8) by a constant to achieve the desired value of ρ. This will create data with IRFs of similar shape to that implied by the data but with varying levels of persistence.

Figure 7 shows results where the sample size is $T = 100$. For each value of ρ considered, the figure contains three sets of results. The left panel shows the bias in the average estimated IRF across simulations for both the levels (dashed line) and long-differenced (dash-circle line) specification. The middle panel shows the proportion of simulations where the true value of β^h is contained inside of a 90% confidence interval constructed via the levels specification (dashed line) and long-differenced specification (dash-circle line). Finally, the right panel shows the ratio of the standard deviation of the impulse response estimate for the long-differenced specification to that for the levels specification.

Figure 7 provides a clear conclusion: For all three persistence levels for the AR(8) model, the long-differenced specification produces estimates with less bias and confidence intervals with more accurate coverage than the levels specification. As both the persistence of the process and the horizon of the IRF increase, the better the performance of the long-differenced specification *relative* to the levels specification. It is worth emphasizing that bias
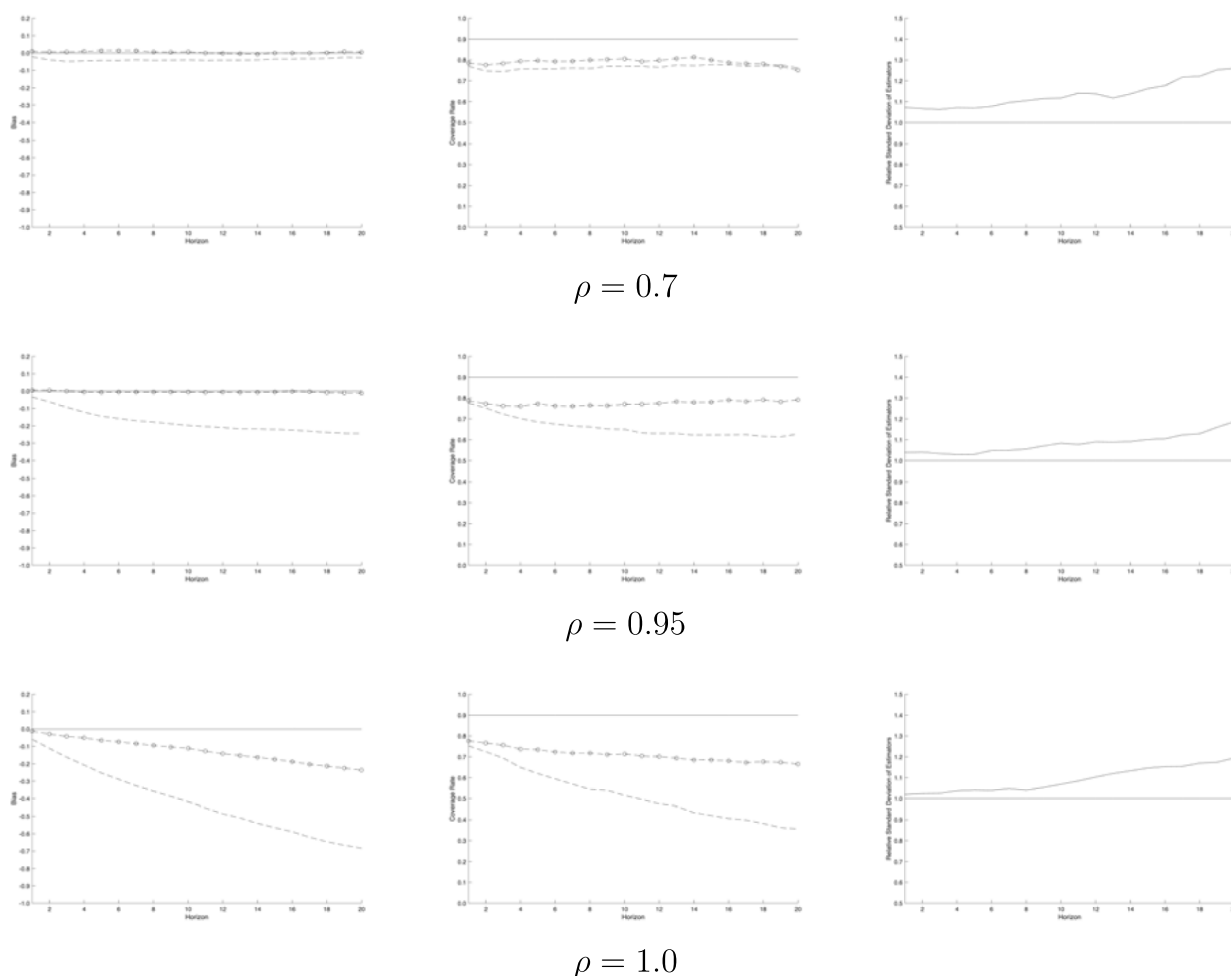

$$\rho = 0.7$$

$$\rho = 0.95$$

$$\rho = 1.0$$

FIGURE 7 | Simulation Results from levels and long-differenced LP and AR(8) DGP ($T = 100$). *Note:* This figure displays simulation results from estimation of the levels and long-differenced specification of the LP when the true DGP is an AR(8) model and $T = 100$. Results for three alternative values of the sum of the autoregressive parameters ($\rho = \{0.7, 0.95, 1.0\}$) are displayed. The left column shows the bias across simulations for the levels specification (dashed line) and long-differenced specification (dash-circle line). The middle column shows the 90% confidence interval coverage of the true impulse response function for the levels specification (dashed line) and long-differenced specification (dash-circle line). The right column shows the ratio of the standard deviation of the long-differenced estimator to the levels estimator.
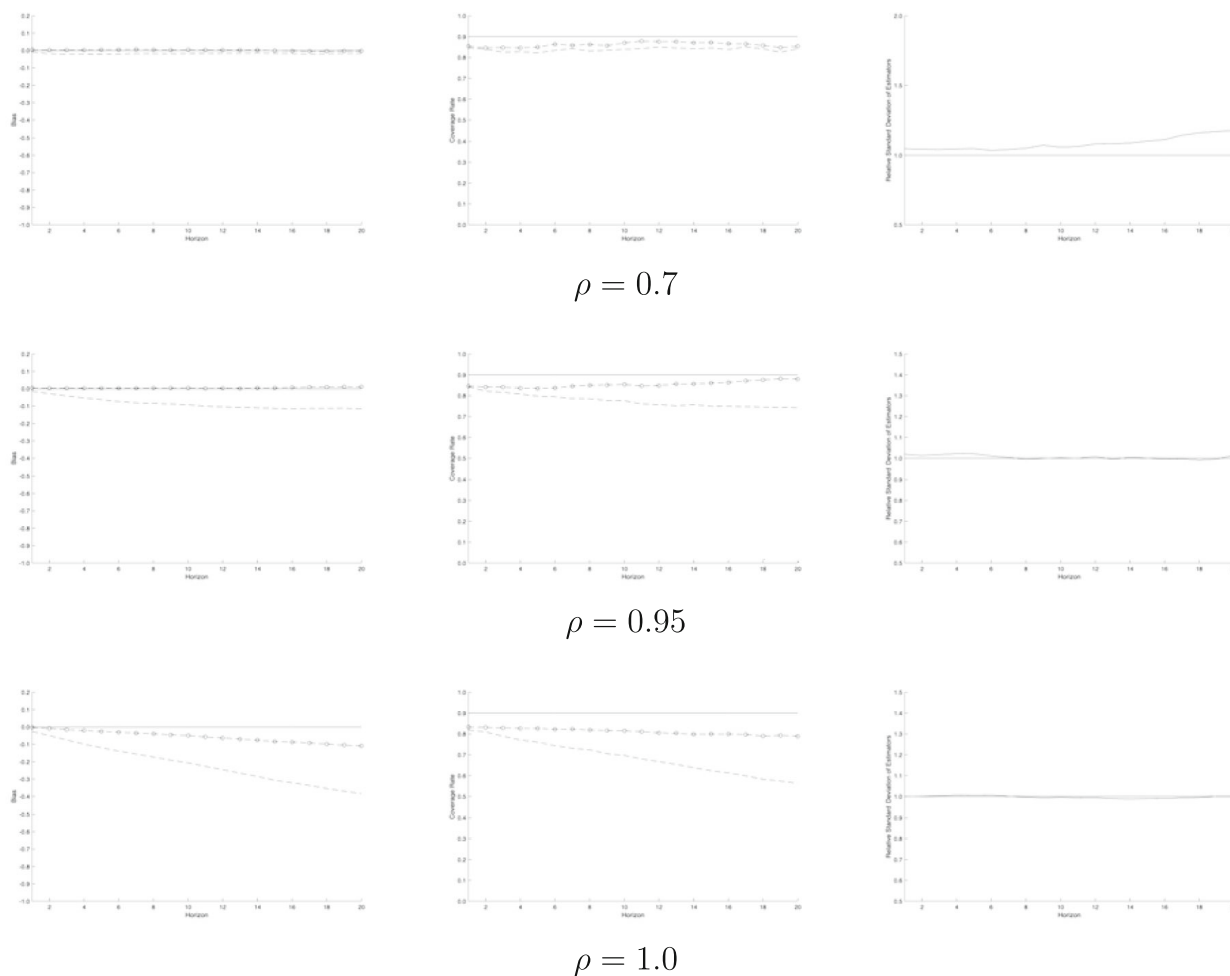
$$\rho = 0.7$$

$$\rho = 0.95$$

$$\rho = 1.0$$

FIGURE 8 | Simulation Results from levels and long-differenced LP and AR(8) DGP ($T = 200$). *Note:* This figure displays simulation results from estimation of the levels and long-differenced specification of the LP when the true DGP is an AR(8) model and $T = 200$. Results for three alternative values of the sum of the autoregressive parameters ($\rho = \{0.7, 0.95, 1.0\}$) are displayed. The left column shows the bias across simulations for the levels specification (dashed line) and long-differenced specification (dash-circle line). The middle column shows the 90% confidence interval coverage of the true impulse response function for the levels specification (dashed line) and long-differenced specification (dash-circle line). The right column shows the ratio of the standard deviation of the long-differenced estimator to the levels estimator.

reduction from using the long-differenced specification is still visible even with a process that is clearly stationary.

With this general conclusion in place, we turn to the results in more detail. For the two stationary cases, the long-differenced specification produces approximately unbiased estimates and 90% confidence intervals that are undersized (between 75% and 90% for all horizons). The levels specification performs reasonably well in the $\phi = 0.7$ case, though it still displays noticeable downward bias and less accurate coverage than the long-differenced specification. When $\phi = 0.95$, the performance of the levels specification deteriorates significantly, with estimates displaying very high levels of bias and confidence intervals with coverage far below their stated levels. These inaccuracies become larger as the horizon of the impulse response increases. Finally, in the unit root case, there is some bias introduced in the long-differenced specification, and coverage intervals fall to between 65% and 80%. However, the long-differenced specification vastly outperforms the levels specification in this case. Indeed, the level specification in the unit root case has abysmal

performance, with bias around 70% of the true value at the longest horizons and 90% confidence intervals with coverage around 35%.

Of course, an estimator with better bias properties may come at the expense of increased estimator variance (Li et al. 2024). To investigate this possibility, the third column of Figure 7 displays the ratio of the simple standard deviation of the estimate of β_h from the long-differenced specification to that from the levels specification. Here we see that the long-differenced estimator does have higher variance, with the standard deviation for this estimator 5%–20% higher than for the levels estimator at most horizons.

Figure 8 shows the results when the sample size is increased to $T = 200$. These results are qualitatively similar to the $T = 100$ case. As would be expected, the performance of both the levels and long-differenced specification improves, in terms of both bias and coverage. However, the long-differenced specification maintains a distinct performance advantage. Also, for $\rho = 0.95$ and $\rho = 1.0$, the levels and long-differenced estimators now have

very close to the same estimator variance. Figure C1 in the Supporting Information Appendix shows the results when $T = 300$ and demonstrates a continued convergence of the performance of the two estimators but with still clear bias reduction produced by the long-differenced LP, particularly at higher persistence levels. Finally, Figure C2 in the Supporting Information Appendix contains results when $\rho = 0.5$. At this lower persistence level, the levels and long-differenced estimators have similar performance.

4.2 | Unobserved Components Models

We next consider several different so-called unobserved components (UC) DGPs, which are popular descriptive models of macroeconomic aggregates.[9] We will consider both univariate and multivariate UC models. In addition to their empirical relevance, these models can provide interesting features not present in the univariate AR models studied to this point, including MA dynamics and cointegration.

Each of the UC models we consider will include a "cyclical" or "transitory" component, C_t, which will follow a covariance-stationary AR(2) process as follows:

$$C_t = \phi_1 C_{t-1} + \phi_2 C_{t-2} + s_t$$
$$s_t \sim \text{i.i.d.} N(0, \sigma_s^2),$$

where the roots of the lag polynomial $(1 - \phi_1 L - \phi_2 L^2)$ have modulus greater than one. In the simulations, s_t will correspond to the observed, exogenous, shock of interest. The alternative UC models will differ in their treatment of trends, the presence of multivariate information, and the presence of additional transitory dynamics. We will consider the following three alternative models:

Trend-Stationary UC Model

$$y_t = T_t + C_t,$$
$$T_t = \mu + T_{t-1}.$$

This UC model is equivalent to a trend-stationary AR model and as such is a natural extension of the covariance stationary AR models studied in the previous section. Specifically, this will allow us to investigate the impact of time trends on the performance of the levels and long-differenced estimators.

Stochastic Trend UC Model

$$y_t = T_t + C_t,$$
$$T_t = \mu + T_{t-1} + v_t,$$
$$v_t \sim \text{i.i.d.} N(0, \sigma_v^2).$$

By matching moments, one can see that this model is equivalent to a restricted ARIMA(2,1,2) for y_t (Harvey 1985). Thus, this model provides us with an extension in the form of MA dynamics over the AR models with unit roots (ARIMA(p,1,0)) considered earlier.

Common Trends and Common Cycles UC Model

$$y_{i,t} = a_i + T_t + b_i C_t + e_{i,t}, i = 1, 2, \ldots, 3,$$
$$T_t = \mu + T_{t-1} + v_t,$$
$$v_t \sim \text{i.i.d.} N(0, \sigma_v^2),$$
$$e_{i,t} \sim \text{i.i.d.} N(0, \sigma_i^2).$$

In this model, each series, indexed by i, shares a common stochastic trend (T_t) and a common cyclical component (C_t), and additionally contains an idiosyncratic transitory component. The model implies that the vector $Y_t = (y_{1,t}, y_{2,t}, y_{3,t})'$ is cointegrated with cointegrating vectors $(1, -1, 0)'$ and $(1, 0, -1)'$. This DGP will then provide us with some insight into how levels vs. long-differenced LPs behave when applied to data generated from a cointegrated system.

To calibrate the UC DGPs, we estimate each UC model via maximum likelihood on a sample of the US macroeconomic aggregates from 1947:Q1 to 2024:Q3, which was the longest sample available at the time of writing. The trend stationary and stochastic trend UC models are popular for decomposing real GDP into permanent and transitory components, and so we estimate these model using log real GDP for y_t. The common trends and common cycles UC model is often proposed as a model for a multivariate system of NIPA aggregates, such as real GDP, consumption and investment (Kim and Piger 2002). To estimate this model, we define y_{1t} as log real GDP, y_{2t} as log real personal consumption expenditures, and y_{3t} as log real gross private domestic investment. For the purpose of calculating IRFs, we set $y_t = y_{1t}$, so our simulation reflects the response of log real GDP to s_t.

Figures 9 and 10 contain the results of the simulations based on the calibrated UC DGPs when $T = 100$ and $T = 200$, respectively, while Figure C3 in the Supporting Information Appendix contains the $T = 300$ results. As each of these DGPs produces data for y_t that display clear trending behavior, we include a time trend in the conditioning set of the levels LP. For these various UC GDPs, there is again a clear conclusion: The impulse response estimates from the long-differenced specification exhibit essentially no bias over the entire horizon, while the estimates from the levels specification have significant bias. Also, the long-differenced specification produces confidence intervals with true coverage much closer to the nominal coverage. Thus, in this expanded set of DGPs, long-differencing continues to demonstrate impressive bias reduction properties.

The final columns of Figures 9 and 10 display the relative standard deviation of the long-differenced LP estimator to the levels LP estimator. Here, there is mixed evidence. Despite the significant bias reductions, for the trend-stationary UC model, there does not appear to be any bias/variance tradeoff in this case, as the long-differenced estimator has lower variance than the levels estimator at most horizons. In contrast, for the stochastic trend UC DGP, the long-differenced estimator has higher variance. Finally, in the UC DGP with cointegration, the direction of the relative variance depends on the sample size and horizon. Thus, whether the long-differenced or levels LP estimator has higher variance appears to be DGP, horizon, and sample size dependent.
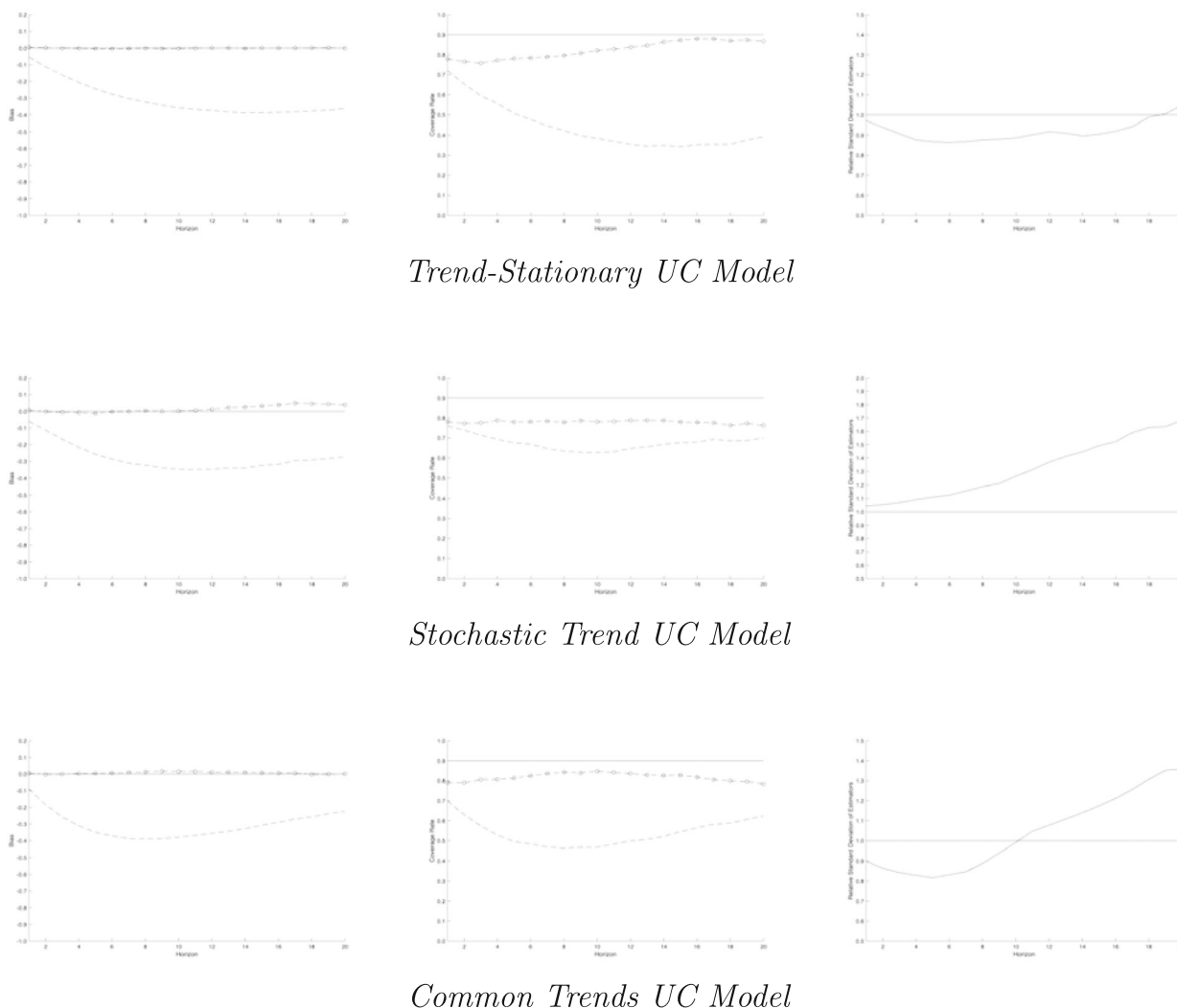
Trend-Stationary UC Model

Stochastic Trend UC Model

Common Trends UC Model

FIGURE 9 | Simulation Results from Levels and Long-Differenced LP and UC-Model DGP (T=100). *Note:* This figure displays simulation results from estimation of the levels and long-differenced specification of the LP when the true DGP is an Unobserved Components model and the sample size $T = 100$. The left column shows the bias across simulations for the levels specification (dashed line) and long-differenced specification (dash-circle line). The middle column shows the 90% confidence interval coverage of the true impulse response function for the levels specification (dashed line) and long-differenced specification (dash-circle line). The right column shows the ratio of the standard deviation of the long-differenced estimator to the levels estimator.

4.3 | VAR Models

In this section, we consider VAR DGPs. We begin with a simple bivariate VAR(1) used by Kilian and Kim (2011) in their simulations evaluating inference from the levels LP. This VAR will allow us to study the performance of the levels and long-differenced LP estimators in a stylized setting where we can clearly control certain model features. We then move to a larger VAR DGP calibrated to the US macroeconomic data.

4.3.1 | Kilian and Kim (2011) Bivariate VAR(1)

The Kilian and Kim (2011) bivariate VAR(1) is specified as follows. Consider a vector of variables defined as $Y_t = (x_t, y_t)'$ that follows a VAR(1) process:

$$Y_t = \Phi_0 + \Phi_1 Y_{t-1} + W_t$$
$$W_t \sim \text{i.i.d.} N(0, \Sigma),$$

where

$$\Phi_0 = \begin{bmatrix} \phi_1^0 \\ \phi_2^0 \end{bmatrix}, \Phi_1 = \begin{bmatrix} \phi_{11}^1 & 0 \\ \phi_{12}^1 & \phi_{22}^1 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

The structural shocks are equal to $U_t = Q * W_t$, where Q is the inverse of the Cholesky factorization of Σ. Define the first component of U_t as s_t, which will serve as our observed, exogenous, shock of interest. Our interest is then on the response of y_{t+h} to s_t. For all calibrations, we set $\phi_1^0 = 0$, $\phi_2^0 = 0$, $\phi_{12}^1 = 0.5$, $\phi_{22}^1 = 0.5$, $\sigma_1^2 = 1$, $\sigma_{12} = 0.3$, and $\sigma_2^2 = 1$. We present results for three alternative values for $\phi_{11}^1 = \{0.7, 0.95, 1.0\}$, which will serve to vary the persistence of the effect of s_t on y_{t+h}.
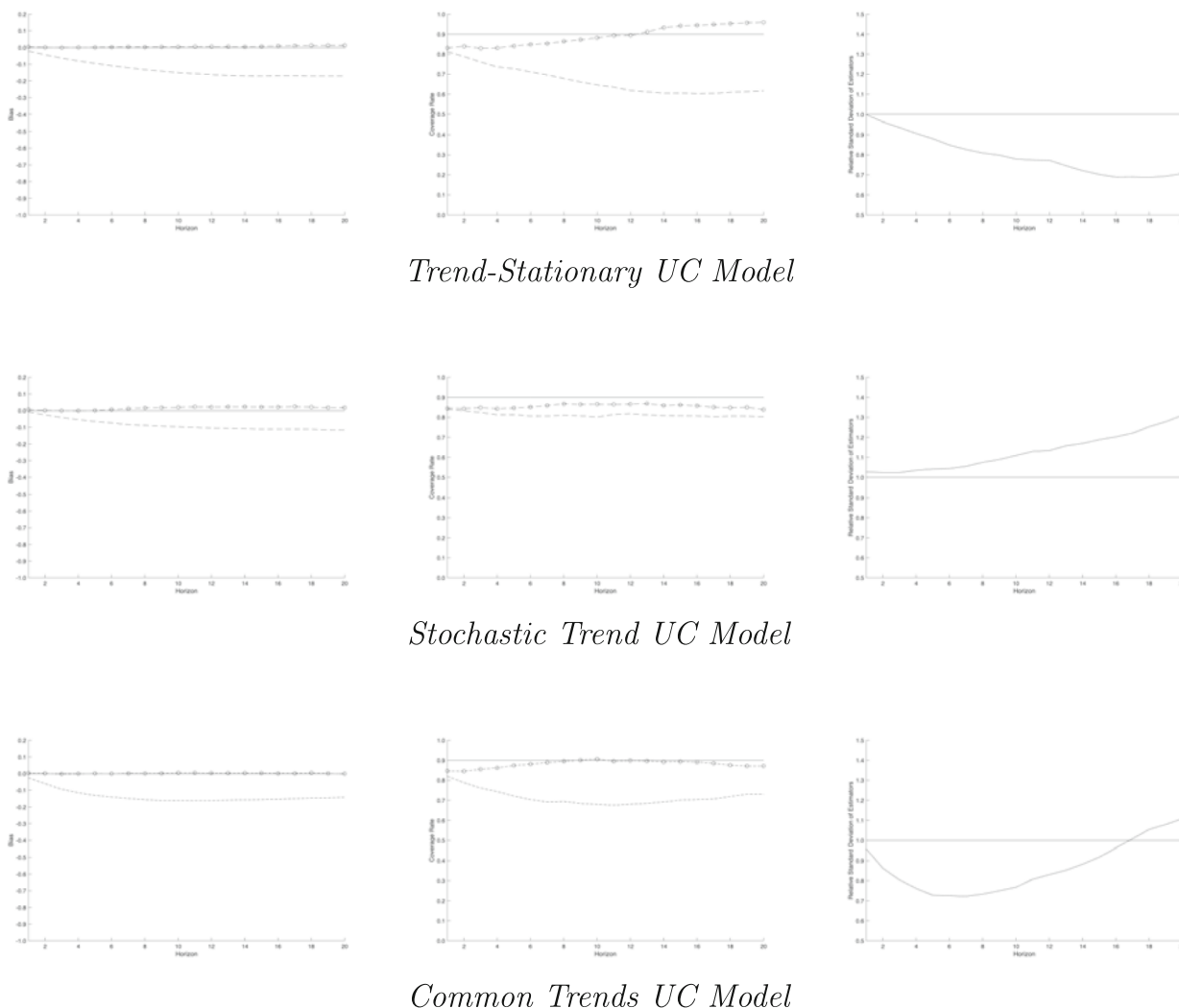
Trend-Stationary UC Model

Stochastic Trend UC Model

Common Trends UC Model

FIGURE 10 | Simulation Results from Levels and Long-Differenced LP and UC-Model DGP (T=200). *Note:* This figure displays simulation results from estimation of the levels and long-differenced specification of the LP when the true DGP is an Unobserved Components model and the sample size $T = 200$. The left column shows the bias across simulations for the levels specification (dashed line) and long-differenced specification (dash-circle line). The middle column shows the 90% confidence interval coverage of the true impulse response function for the levels specification (dashed line) and long-differenced specification (dash-circle line). The right column shows the ratio of the standard deviation of the long-differenced estimator to the levels estimator.

Figures 11 and 12 show the results of simulations for this VAR model when $T = 100$ and $T = 200$ respectively. The results for the VAR DGP are very similar to the other DGPs we have seen thus far. The levels specification has a small downward bias at the lowest calibration of ϕ_{11}^1, with the bias increasing as ϕ_{11}^1 increases and as the horizon increases. The long-differenced specification has much less bias than the levels IRF for all three levels of persistence and is approximately unbiased for the stationary versions of the DGP. The long-differenced specification produces confidence intervals that are generally undersized for all values of ϕ_{11}^1. However, the coverage of these intervals is much closer to the stated 90% size than those produced from the levels specification, which are extremely undersized. Finally, the long-differenced specification produces estimates with generally similar variance to, and in some cases less than, the estimator from the levels specification. Results for a larger sample size, $T = 300$, as well as a lower value of ϕ_{11}^1, are contained in Figures C4 and C5 in the Supporting

Information Appendix. These results show progressions similar to that seen for the univariate AR models.

4.3.2 | Christiano et al. (2005) 9-Variable VAR(4)

We now turn to the performance of the levels and long-differenced LP estimators when the DGP is a medium scale VAR estimated on U.S. macroeconomic data. In their study of bias correction for the levels LP, Herbst and Johannsen (2024) simulate data from the Christiano et al. (2005) 9-variable VAR(4), where the model parameters are estimated on a sample spanning from 1965:Q3-1995:Q3, the same dates as in Christiano et al. (2005).[10] Here we estimate the same VAR over the same sample period. Following Herbst and Johannsen (2024), our focus is on the effects of a monetary policy shock on log real GDP, the log real GDP Deflator, and the Federal Funds Rate. The monetary policy shock is identified recursively. As
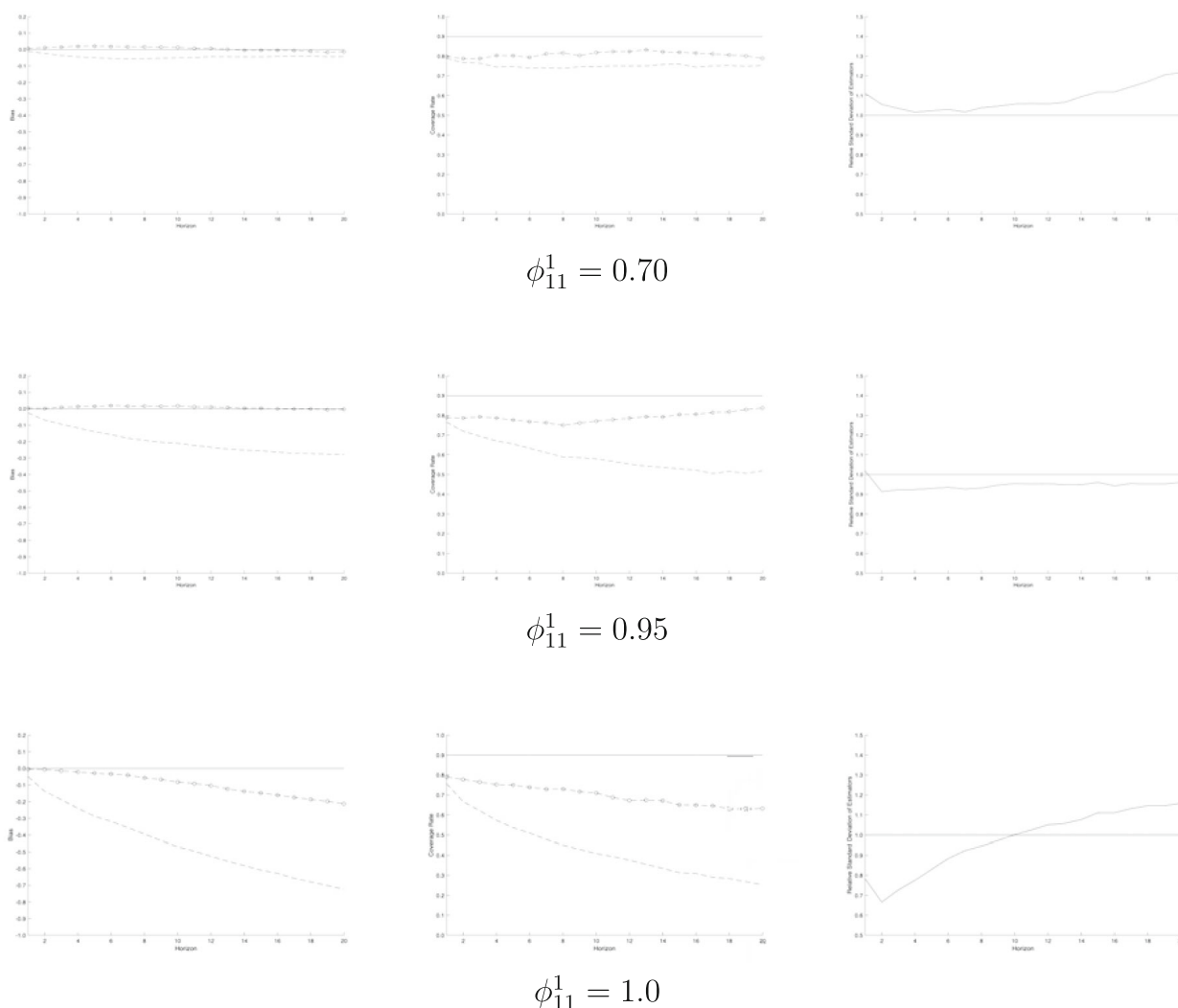
$$\phi_{11}^1 = 0.70$$

$$\phi_{11}^1 = 0.95$$

$$\phi_{11}^1 = 1.0$$

FIGURE 11 | Simulation results from levels and long-differenced LP and Killian and Kim (2011) VAR(1) DGP ($T = 100$). *Note:* This figure displays simulation results from estimation of the levels and differences specification of the LP when the true DGP is the Kilian and Kim (2011) bivariate VAR(1) model and $T = 100$. Results are shown for three alternative values of $\phi_{11}^1 = \{0.7, 0.95, 1.0\}$. Other parameters are set as described in Section 4.3.1. The left column shows the bias across simulations for the levels specification (dashed line) and long-differenced specification (dash-circle line). The middle column shows the 90% confidence interval coverage of the true impulse response function for the levels specification (dashed line) and long-differenced specification (dash-circle line). The right column shows the ratio of the standard deviation of the long-differenced estimator to the levels estimator.

in Herbst and Johannsen (2024), the monetary policy shock, which serves as s_t in our notation, is assumed observed in the generated data. To maintain an accurate comparison to Herbst and Johannsen (2024), we also follow their assumption that the econometrician uses four lags of all variables in the levels LP regressions. We further set the lag order for the long-differenced LP also to four. Thus, for this DGP, we do not conduct lag selection.

Figures 13 and 14 display the results for sample sizes $T = 100$ and $T = 200$ respectively, while Figure C6 in the Supporting Information Appendix shows results for the $T = 300$ case. In the figures, "Real GDP Response" indicates the response of the first variable in the VAR to the monetary policy shock, which was the position of log Real GDP in the estimated VAR. The other panels are similarly defined. Beginning with the levels LP, there is noticeable bias in the IRF estimator for the levels LP for all three

responses considered and for both sample sizes. The amount of bias matches closely to that reported in Figure 7 of Herbst and Johannsen (2024). In contrast, the long-differenced estimator produces estimates with very little bias in the $T = 100$ case and essentially no bias in the $T = 200$ case.

Herbst and Johannsen (2024) provide an analytic approximate to the finite-sample bias in the levels LP estimator and use this approximation to propose a bias-corrected estimator. While their procedure produces noticeable bias improvements, it leaves behind a significant portion of the bias produced by the LP estimator. As one example, in their Figure 7, the bias-corrected estimator eliminates about one-third of the bias observed in the real GDP response for the $T = 100$ case. Thus, it is striking that long-differencing, which is trivial to implement, produces such drastic improvements, eliminating essentially all of the bias observed for the Christiano et al. (2005) VAR DGP.
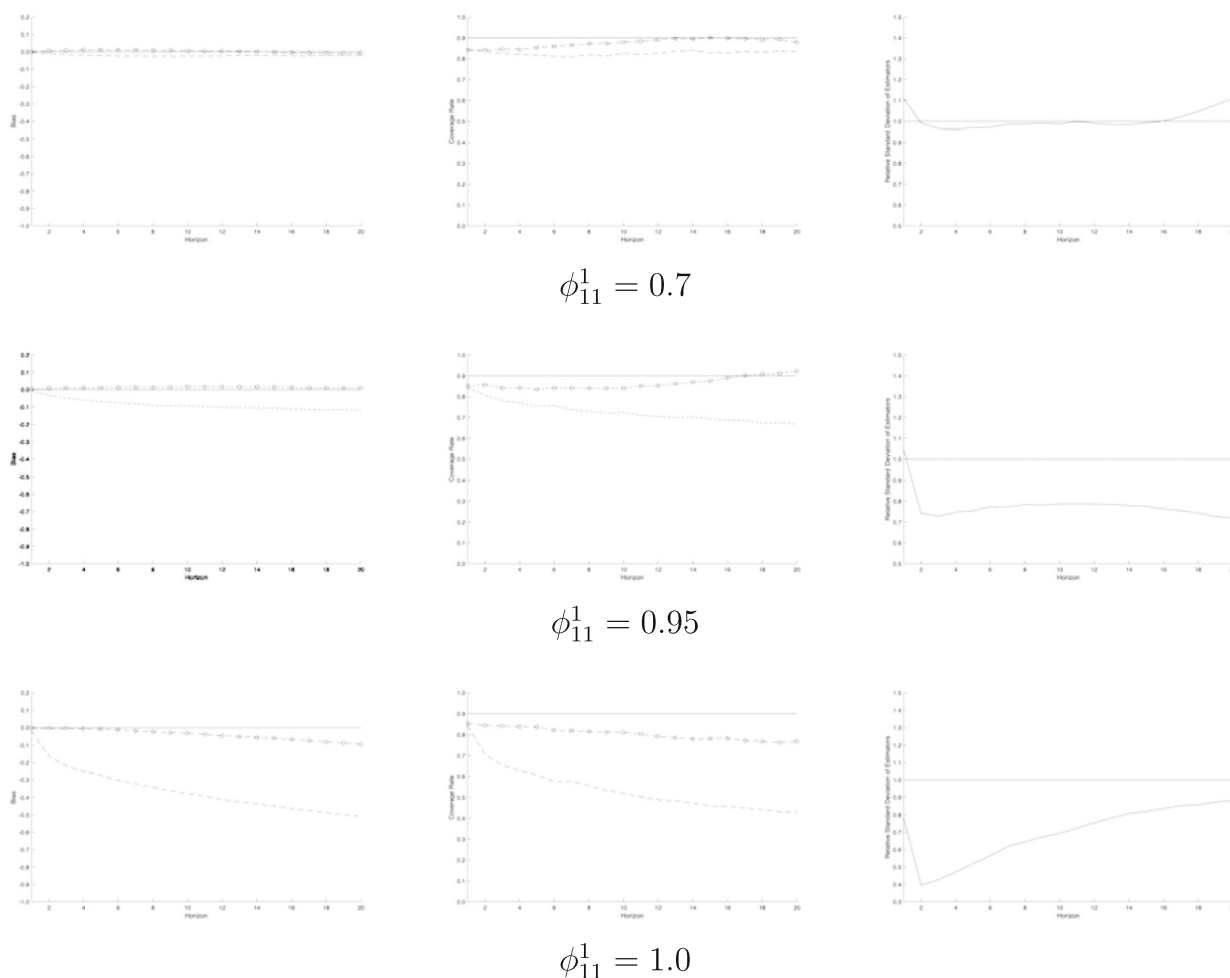
Journal of Applied Econometrics, 2025

$$\phi_{11}^1 = 0.7$$

$$\phi_{11}^1 = 0.95$$

$$\phi_{11}^1 = 1.0$$

FIGURE 12 | Simulation results from levels and long-differenced LP and Killian and Kim (2011) VAR(1) DGP ($T = 200$). *Note:* This figure displays simulation results from estimation of the levels and differences specification of the LP when the true DGP is a the Kilian and Kim (2011) bivariate VAR(1) model and $T = 200$. Results are shown for three alternative values of $\phi_{11}^1 = \{0.7, 0.95, 1.0\}$. Other parameters are set as described in Section 4.3.1. The left column shows the bias across simulations for the levels specification (dashed line) and long-differenced specification (dash-circle line). The middle column shows the 90% confidence interval coverage of the true impulse response function for the levels specification (dashed line) and long-differenced specification (dash-circle line). The right column shows the ratio of the standard deviation of the long-differenced estimator to the levels estimator.

Moving beyond bias, the second columns of Figures 13 and 14 show that confidence intervals for the long-differenced LP estimator are more accurate than those from the levels estimator in all cases, and often very substantially so. Finally, the third columns show that for this DGP, the long-differenced LP estimator does have larger variance than the levels LP estimator. The extent of this difference varies significantly over sample size, horizon, and which variable the response is being measured for.

4.4 | Additional Robustness Checks

In this section, we consider three additional robustness checks. The first is the sensitivity of the results to the choice of lag order in the levels and long-differenced LPs. The second is the effect of changing the variance of the observed shock (σ_s^2) relative to other noise in the process. Finally, we assess the effects of an alternative approach to standard error calculation on the coverage properties of confidence intervals.

4.4.1 | Choice of Lag Order

With one exception, the preceding results have been generated assuming that lag selection is performed by the econometrician when implementing the levels and long-differenced LPs. For this reason, the results represent a mixture of cases where the estimated model contains less than, equal to, and more than the true number of lags. In this section, we will investigate the importance of lag order for the performance of both the levels and long-differenced LP. We focus on the case of the AR(8) DGP, where we can both precisely control the correct lag order, as well as evaluate significant departures from the correct lag order in both directions. For all simulations, we set the value of $\rho = 0.95$ and $T = 100$. The results when $T = \{200, 300\}$ are qualitatively similar.

We conduct several experiments. In the first, we assume that the correct lag order for the long-differenced and levels LP are known. For stationary versions of the AR(8), these lag orders will be $p_L = 8$ for the levels LP, and $p_D = 8 + h$ for the
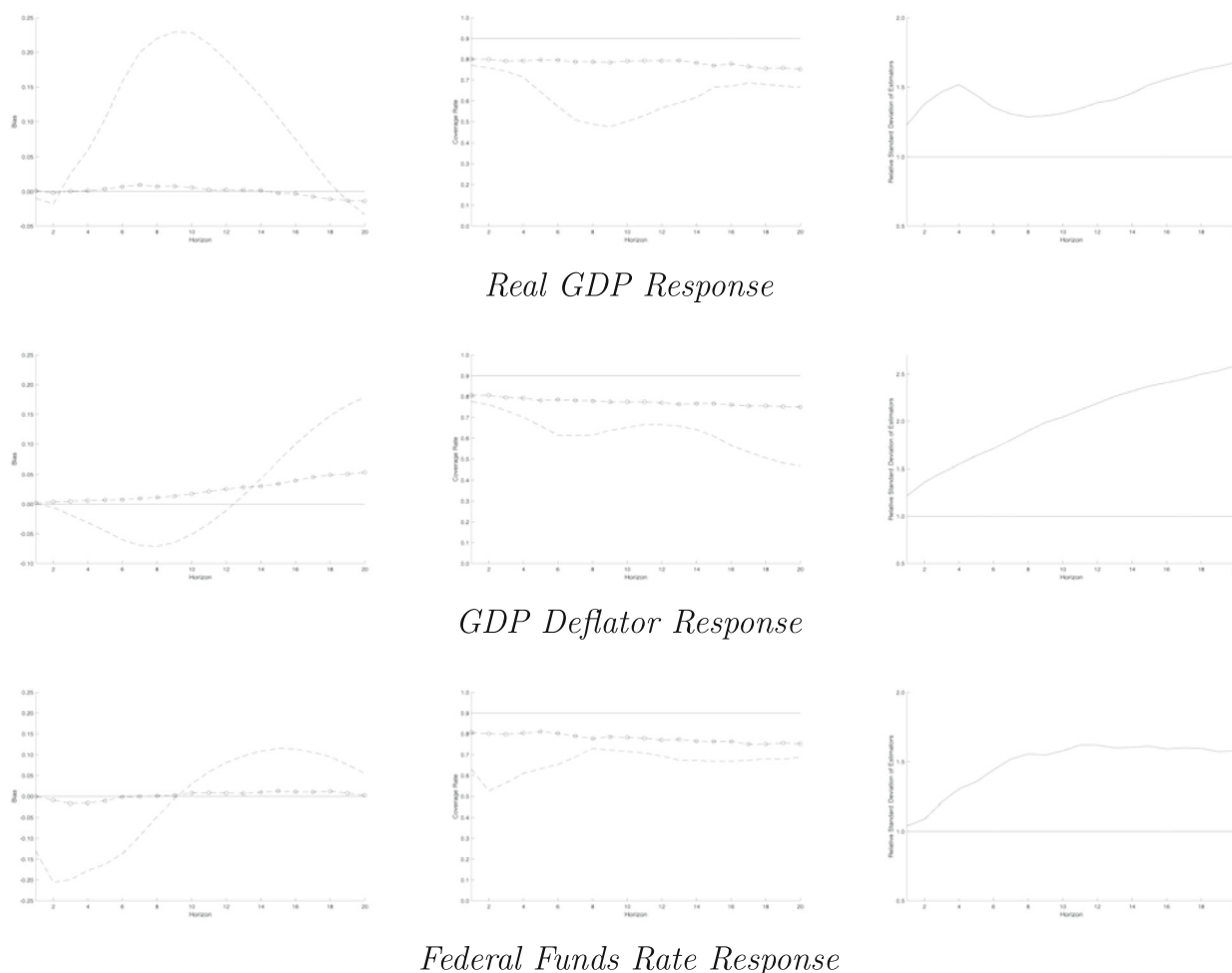
Real GDP Response

GDP Deflator Response

Federal Funds Rate Response

FIGURE 13 | Simulation results from levels and long-differenced LP and CEE VAR(4) model ($T = 100$). *Note:* This figure displays simulation results from estimation of the levels and long-differenced specification of the LP when the true DGP is the 9-variable VAR(4) from Christiano et al. (2005), estimated as described in Herbst and Johannsen (2024). The sample size is $T = 100$. The left column shows the bias across simulations for the levels specification (dashed line) and long-differenced specification (dash-circle line). The middle column shows the 90% confidence interval coverage of the true impulse response function for the levels specification (dashed line) and long-differenced specification (dash-circle line). The right column shows the ratio of the standard deviation of the long-differenced estimator to the levels estimator.

long-differenced LP. The top row of Figure 15 contains the results under this correct lag order assumption. Interestingly, the estimator bias and confidence interval coverage are similar to those shown in Figure 7 when lag selection is performed. The main difference seen is for the relative standard deviation of the estimators, which shows a significant increase in the variance of the long-differenced LP relative to the levels LP, especially at the longer horizons. This likely comes due to the large number of parameters being estimated in the correctly specified long-differenced LP at longer horizons, since the number of lags in this specification grows with the horizon. As discussed in footnote 4, there are implied parameter restrictions on the long-differenced LP that could be imposed to reduce this proliferation of parameters. However, given that the bias reduction is largely unchanged from when using lag selection, it is unclear what the value added of this approach would be.

Montiel Olea and Plagborg-Møller (2021) argue for "lag-augmented" levels LPs, which refers to the use of level LP regressions with lags beyond the correct lag order.

Lag-augmented LPs have a number of favorable properties, including an asymptotic distribution that is uniform over persistence. The middle row of Figure 15 evaluates the case where the levels LP is lag-augmented with $p_L = 9$. As we have already seen that the long-differenced LP has significantly higher estimator variance with "correct" lag orders, we here simply set $p_D = 9$ for this set of simulations. The results here are largely unchanged from those in Figure 7, suggesting that lag augmentation does not play a significant role in affecting the types of finite-sample bias we are focused on in this paper.

Our final experiment investigates the effects of simply controlling for a single lag, so that $p_L = p_D = 1$, which is not an uncommon practice in the literature.[11] The final panel of Figure 15 demonstrates that the effects of this significant under-parameterization are minimal on the bias, confidence interval coverage, or relative variance observed in the levels and long-differenced LP estimators.
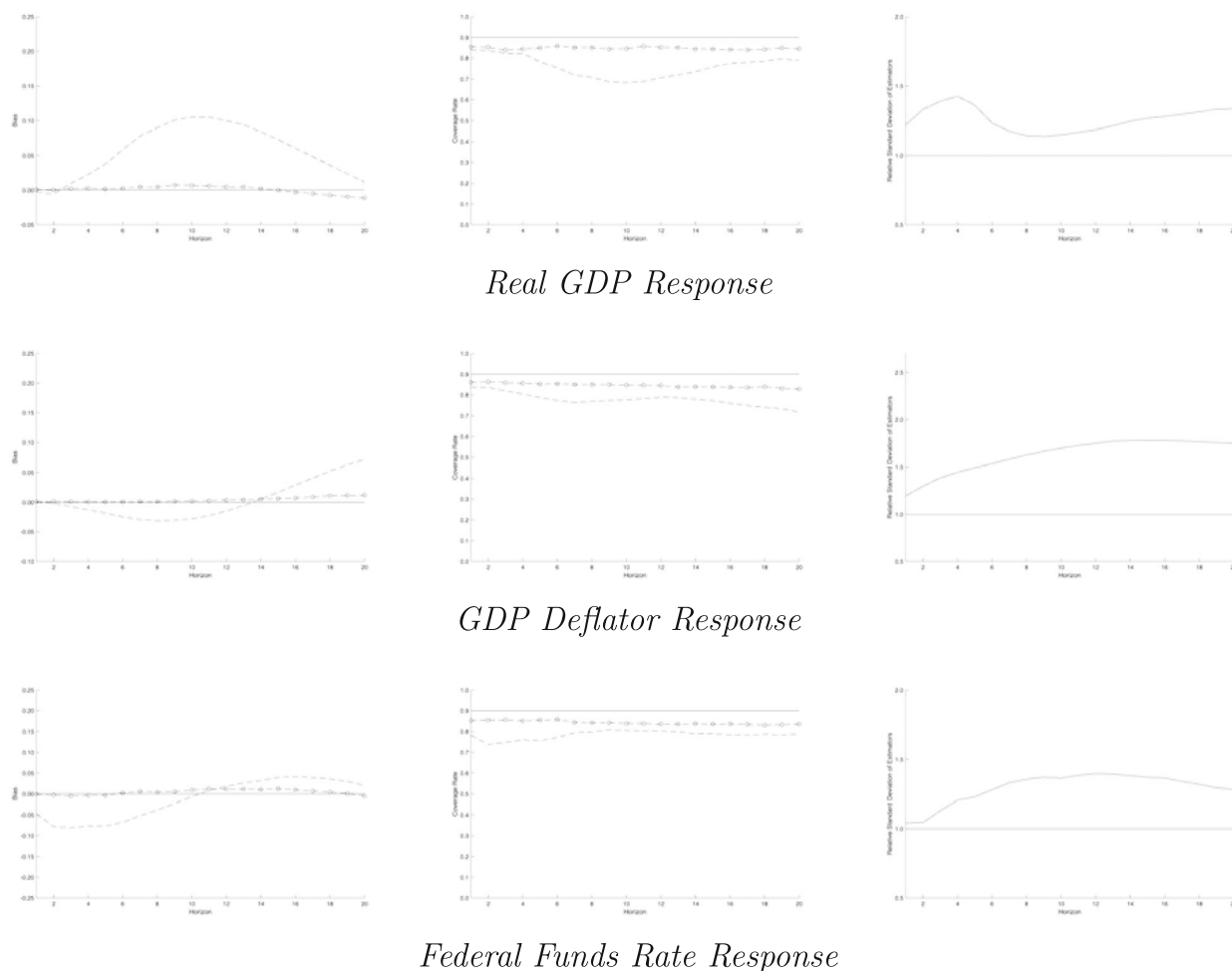
Real GDP Response

GDP Deflator Response

Federal Funds Rate Response

FIGURE 14 | Simulation results from levels and long-differenced LP and CEE VAR(4) model ($T = 200$). *Note:* This figure displays simulation results from estimation of the levels and long-differenced specification of the LP when the true DGP is the 9-variable VAR(4) from Christiano et al. (2005), estimated as described in Herbst and Johannsen (2024). The sample size is $T = 200$. The left column shows the bias across simulations for the levels specification (dashed line) and long-differenced specification (dash-circle line). The middle column shows the 90% confidence interval coverage of the true impulse response function for the levels specification (dashed line) and long-differenced specification (dash-circle line). The right column shows the ratio of the standard deviation of the long-differenced estimator to the levels estimator.

Overall, the results of this section suggest that for LPs with observed shocks, performing lag selection generates results that are no worse than, and sometimes better than, the use of the correct lag order, lag augmentation, or simply fitting a model with one lag. Most importantly for this paper, the long-differenced estimator continues to be a powerful bias reduction strategy under all of these lag order choice assumptions.

4.4.2 | Relative Variance of the Observed Shock

For each of the DGPs used above, there has been a calibration of the variance of the observed shock (σ_s^2) to the variance of the other contemporaneous stochastic elements of the DGP. Stated differently, there was a calibration of the amount of variation in y_t explained by s_t. In most cases, this was empirically estimated, but in other cases, such as the AR(8), this was set in a nondata-based manner.

Here we investigate the robustness of our results to changes in this assumption, focusing on the case of the AR(8) model. In the baseline results for the AR(8) DGP, we set $\sigma_s^2/\sigma_\omega^2 = 1$. Here we conduct two alternative experiments, corresponding to $\sigma_s^2/\sigma_\omega^2 = \{0.5, 2.0\}$. To economize on figures, we focus on the case where $T = 100$ and $\rho = 0.95$, though the conclusions drawn from alternative sample sizes and values for ρ are similar.

In their study of finite-sample bias in levels LPs, Herbst and Johannsen (2024) demonstrate that an approximation to the bias observed in the levels LP is not a function of the variance of the observed shock relative to the overall variance of y_t. The results from our simulations, presented in Figure 16, suggest that this approximation is accurate. For ease of reference, the top panel of 16 shows the baseline results when $\sigma_s^2/\sigma_\omega^2 = 1$. A comparison of this baseline case to the alternative values of $\sigma_s^2/\sigma_\omega^2$ finds little change in the results.
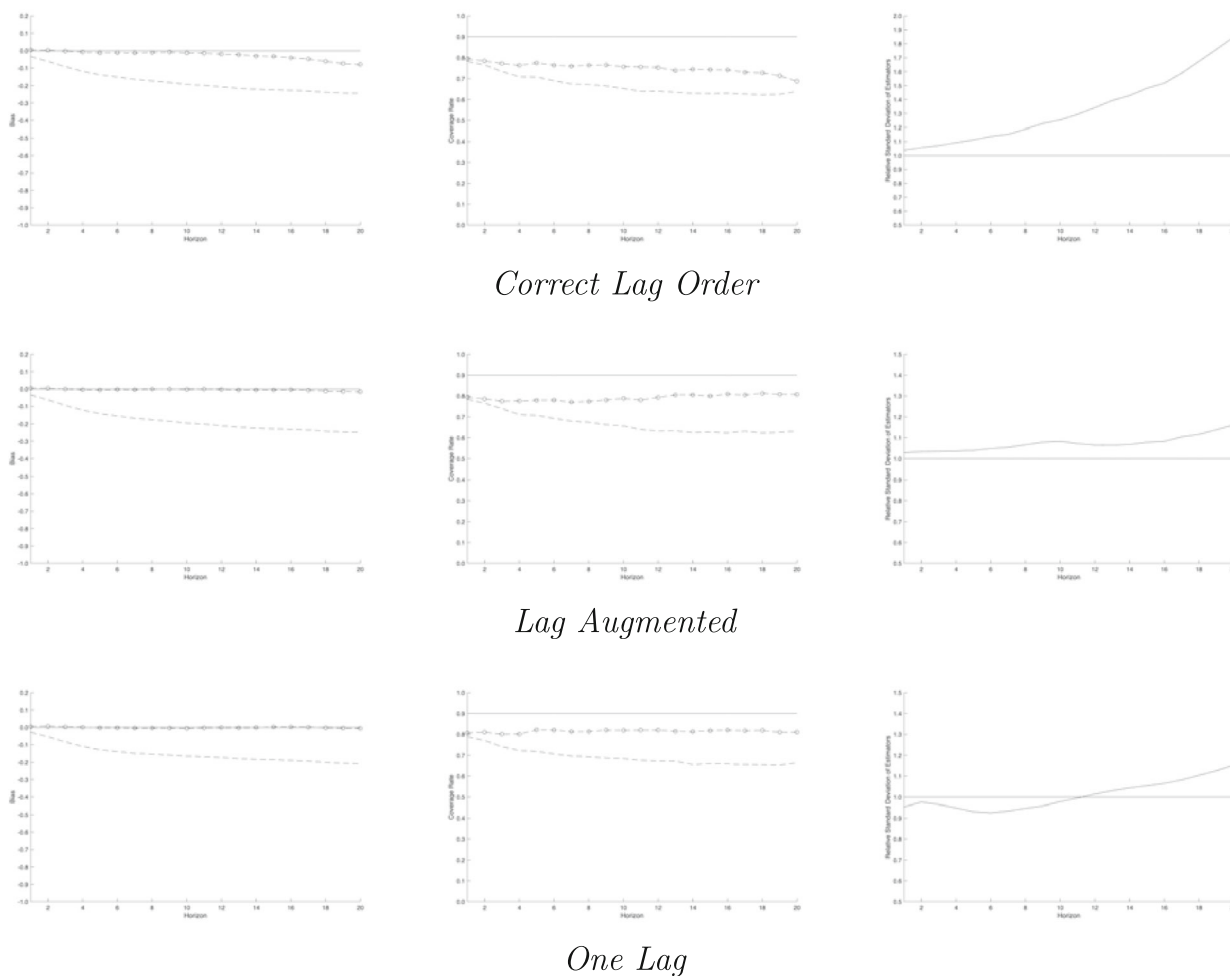
Correct Lag Order

Lag Augmented

One Lag

FIGURE 15 | Simulation Results from Levels and Long-Differenced LP and AR(8) DGP with Alternative Lag Order Selection ($T = 100$) and $\rho = 0.95$. *Note:* This figure displays simulation results from estimation of the levels and long-differenced specification of the LP when the true DGP is an AR(8) model, $T = 100$, $\rho = 0.95$, and three alternative lag orders are considered. "Correct Lag Order" indicates 8 lags in the levels specification and $8 + h$ lags in the long differenced specification. "Lag-augmented" indicates 9 lags in both the levels and long-differenced specification. "One lag" indicates 1 lag in both the levels and long-differenced specification. The left column shows the bias across simulations for the levels specification (dashed line) and long-differenced specification (dash-circle line). The middle column shows the 90% confidence interval coverage of the true impulse response function for the levels specification (dashed line) and long-differenced specification (dash-circle line). The right column shows the ratio of the standard deviation of the long-differenced estimator to the levels estimator.

4.4.3 | EHW Standard Errors

Montiel Olea and Plagborg-Møller (2021) show that when LPs are lag-augmented, the EHW heteroskedasticity robust standard errors are asymptotically sufficient. Further, Herbst and Johannsen (2024) find that in finite samples, there is downward bias in the Newey–West standard errors for LPs, and this bias is alleviated through the use of EHW standard errors. These are important results, as most of the literature considers heteroskedasticity and autocorrelation consistent standard errors, typically the Newey–West standard errors that we have considered here (Jordá 2005; Ramey 2016).

Figures 17 and 18 show the 90% confidence interval coverage for each of the DGPs we have considered, where EHW standard errors are used to construct confidence intervals, and $T = 100$ and 200, respectively. Given the results of the previous section,

we do not assume that LPs are lag-augmented but instead focus on LPs with lag selection.

There are two main conclusions from the results in Figures 17 and 18. First, if one compares the coverage intervals computed using EHW to those from the corresponding figures based on the Newey–West, the EHW intervals generally have improved coverage for both the levels and long-differenced LP. For example, consider the AR(8) DGP where $\rho = 0.95$ and $T = 100$, which is displayed in the middle panel of the top row of Figure 17. In this case, the long-differenced LP produces confidence intervals with approximately correct coverage, and the levels LP produces confidence intervals with undersized coverage ranging from 0.85 to 0.7. In contrast, from the middle panel of the second row of Figure 7, the long-differenced LP produces confidence intervals that are undersized and around 0.8, while the levels LP produces confidence intervals with undersized coverage ranging from 0.7 to 0.6.
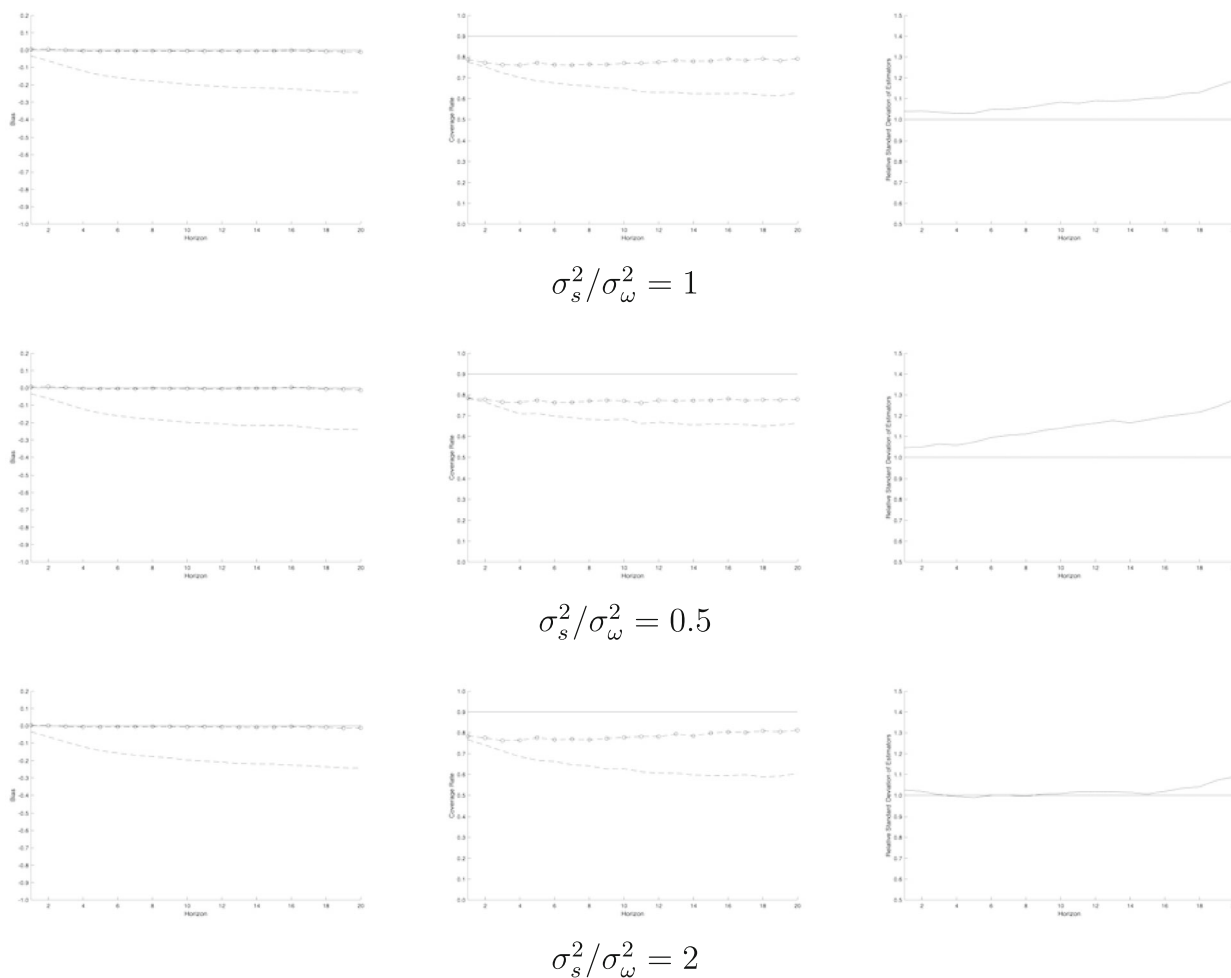
$$\sigma_s^2/\sigma_\omega^2 = 1$$

$$\sigma_s^2/\sigma_\omega^2 = 0.5$$

$$\sigma_s^2/\sigma_\omega^2 = 2$$

FIGURE 16 | Simulation results from levels and long-differenced LP and AR(8) DGP with alternative $\sigma_s^2/\sigma_\omega^2$ ($T = 100$) and $\rho = 0.95$. *Note:* This figure displays simulation results from estimation of the levels and long-differenced specification of the LP when the true DGP is an AR(8) model, $T = 100$, $\rho = 0.95$, and three alternative values for $\sigma_s^2/\sigma_\omega^2 = \{0.5, 1.0, 2.0\}$ are considered. The left column shows the bias across simulations for the levels specification (dashed line) and long-differenced specification (dash-circle line). The middle column shows the 90% confidence interval coverage of the true impulse response function for the levels specification (dashed line) and long-differenced specification (dash-circle line). The right column shows the ratio of the standard deviation of the long-differenced estimator to the levels estimator.

The second conclusion is that despite the improved coverage performance seen with EHW standard errors, the long-differenced LP continues to produce more accurate coverage than the levels-LP in nearly all cases. This improvement continues to be, in many cases, quite significant. Overall, these results are supportive of recent arguments made for the use of EHW standard errors for LPs.

4.5 | LP-IV

In the preceding simulations, we have assumed that the shock of interest, s_t, is exogenous. We now turn to simulations where s_t is endogenous, but we have an instrument available, and can thus implement LP-IV. The results of Section 3 are suggestive that the bias correction provided by long-differencing will extend to this case.

To economize on results, we restrict our attention to the AR(8) DGP as follows:

$$y_t = \alpha + \beta_0 s_t + \sum_{i=1}^{8} \phi_i y_{t-i} + \omega_t.$$

To model endogeneity, we assume that s_t is described by the following first-stage regression:

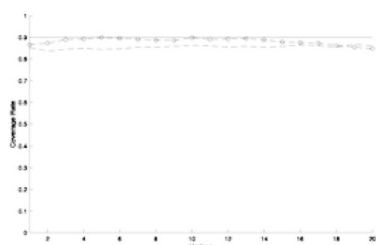$$s_t = \lambda + \gamma \varepsilon_t + \eta_t,$$

where $\varepsilon_t \sim$ i.i.d. $N\left(0, \sigma_\varepsilon^2\right)$ and $\left(\omega_t, \eta_t\right)' \sim$ i.i.d. $MVN\left(0_2, \Omega\right)$ with the following:

$$\Omega = \begin{bmatrix} \sigma_\omega^2 & \sigma_{\omega\eta} \\ \sigma_{\omega\eta} & \sigma_\eta^2 \end{bmatrix}.$$
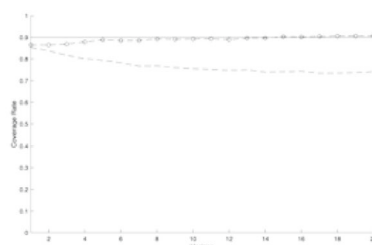
By construction, the instrument in this DGP satisfies the lead-lag exogeneity condition for instrument validity in LP (Stock and Watson 2018).

To calibrate the parameters of the AR(8) DGP, we follow the strategy outlined in Section 4.1. That is, we consider three alternative
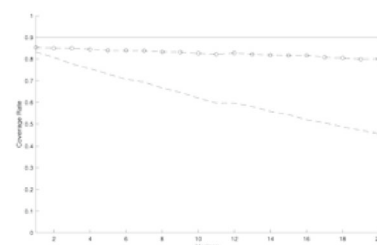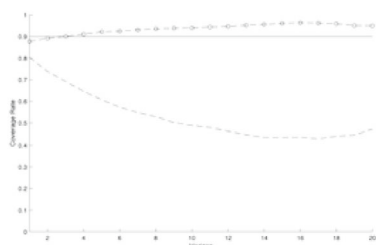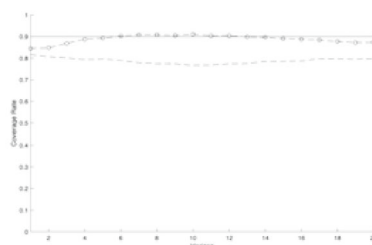
AR(8) DGP

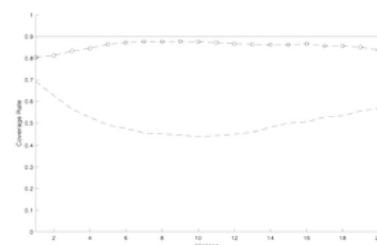| $\rho = 0.7$ | $\rho = 0.95$ | $\rho = 1.0$ |

Unobserved Components Model DGPs

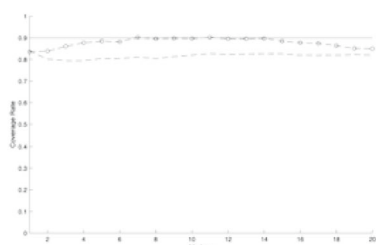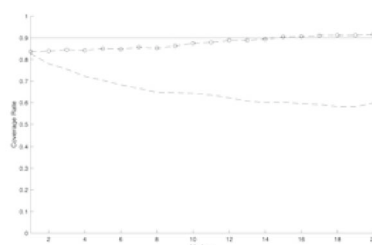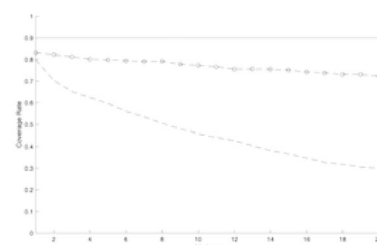**Trend Stationary UC** **Stochastic Trend UC** **Common Trends UC**
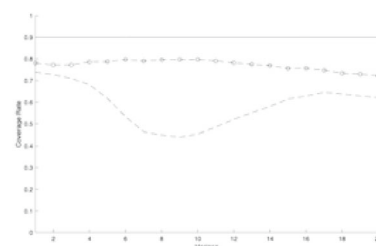
Killian and Kim (2011) VAR(1) DGP
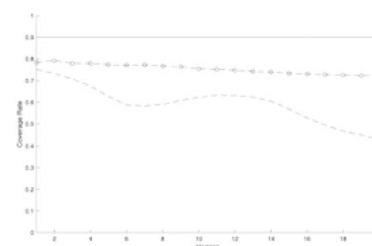

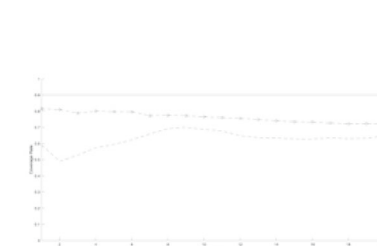$\phi_{11}^1 = 0.7$ $\phi_{11}^1 = 0.95$ $\phi_{11}^1 = 1.0$

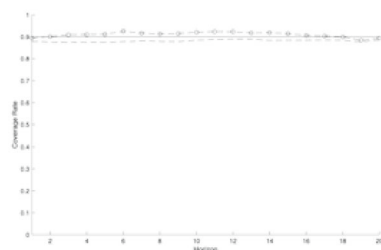## Christiano et. al (2005) VAR(4) DGP

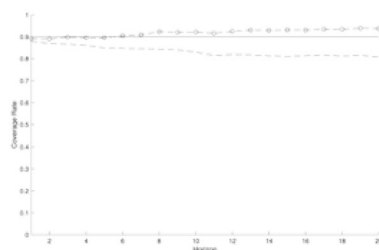
Real GDP Response **GDP Deflator Response** **Federal Funds Response**

FIGURE 17 | Coverage Rates from Levels and Long-Differenced LP Based on Eicker-Huber-White Standard Errors ($T = 100$). *Note:* This figure shows the 90% confidence interval coverage of the true impulse response function for the levels specification (dashed line) and long-differenced specification (dash-circle line) for alternative DGPs and $T = 100$.
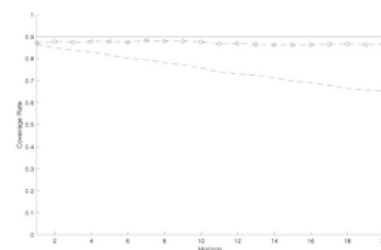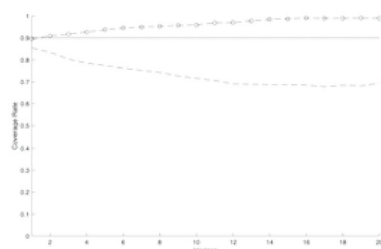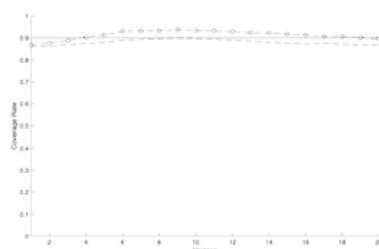
AR(8) DGP

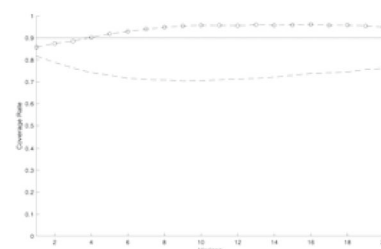$\rho = 0.7$ $\rho = 0.95$ $\rho = 1.0$

Unobserved Components Model DGPs

Trend Stationary UC   **Stochastic Trend UC**   **Common Trends UC**

Killian and Kim (2011) VAR(1) DGP
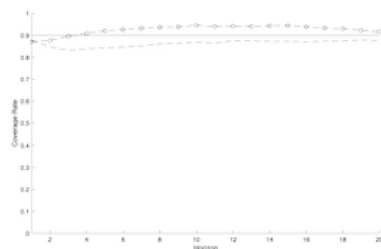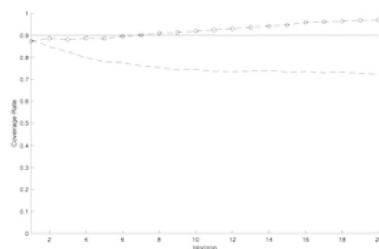
$\phi_{11}^1 = 0.7$ $\phi_{11}^1 = 0.95$ $\phi_{11}^1 = 1.0$
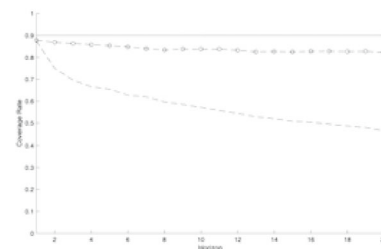
Christiano et. al (2005) VAR(4) DGP
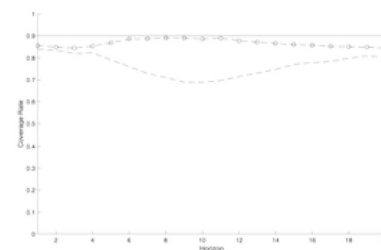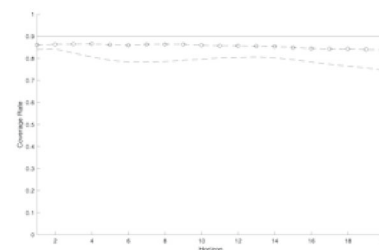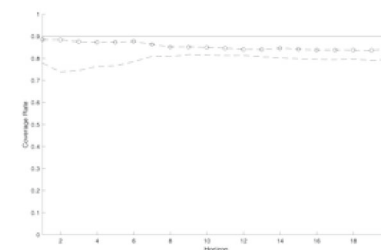
Real GDP Response   **GDP Deflator Response**   **Federal Funds Response**

FIGURE 18 | Coverage Rates from Levels and Long-Differenced LP Based on Eicker-Huber-White Standard Errors ($T = 200$). *Note:* This figure shows the 90% confidence interval coverage of the true impulse response function for the levels specification (dashed line) and long-differenced specification (dash-circle line) for alternative DGPs and $T = 200$.

values for the sum of the AR coefficients, $\rho = \sum_{i=1}^{8} \phi_i$, corresponding to $\rho = \{0.7, 0.95, 1.0\}$, set $\alpha = 0$, $\beta_0 = 1$, and set the ϕ_i based on an AR(8) model fit to quarterly log real US GDP, scaled to achieve the desired value of ρ. Finally, we again set $\sigma_\omega^2 = \sigma_s^2 = 1$. To parameterize the first stage regression, we set $\gamma = 1$ and set σ_ε^2 and σ_η^2 to simultaneously achieve $\sigma_s^2 = 1$ and a population F-statistic for the first stage regression of 10.5. This later choice is meant to maintain an empirically reasonable instrument strength while also largely avoiding simulation draws in which instruments are weak (Staiger and Stock 1997). We set the covariance parameter $\sigma_{\omega\eta} = -0.5$.

Figure 19 presents results for the bias in the LP and LP-IV estimators for various values of $\rho = \{0.7, 0.95, 1.0\}$ and $T = \{100, 200\}$. In each figure, the dashed line and dashed-circle line shows the median bias for the levels and long-differenced LP estimators respectively.[12] Looking across these graphs, we see significant bias in both estimators arising in part from the endogeneity of the LP. The dashed-x and dashed-square lines show the median bias for the levels and long-differenced LP-IV. Here we see that LP-IV eliminates significant portions of the bias observed for the LP estimators. Also, once LP-IV is used, we again see that long-differencing produces significant additional bias reduction over the levels LP-IV, with the amount of this bias reduction of a similar size as for the case where the shock was exogenous.

5 | The Effects of US Monetary Policy Shocks vs. Federal Reserve Information Shocks

In this section, we provide an application to illustrate the estimation differences one can obtain in empirical practice from the levels vs. long-differenced LP specification. In particular, we revisit the results of Jarociński and Karadi (2020) on the effects of Federal Reserve monetary policy shocks vs. information shocks. Specifically, Jarociński and Karadi (2020) separate high-frequency surprises in Federal Reserve interest rate announcements into two components, one being a traditional monetary policy shock and the other being a revelation of Federal Reserve private information regarding the future direction of the economy. Jarociński and Karadi (2020) identify this "central bank information" (CBI) shock separately from the monetary policy (MP) shock via a set of restrictions including high-frequency identification and sign restrictions. These restrictions are imposed inside of a Bayesian VAR, from which impulse response functions can be obtained. Here we revisit the effects of these shocks by directly incorporating the estimated MP and CBI shocks of Jarociński and Karadi (2020) into LPs.

To estimate the LP regressions, we focus on a monthly specification with five variables. Three of these variables are used in the baseline specification of Jarociński and Karadi (2020) and include the logarithm of the monthly average of the S&P 500 index, the Gilchrist and Zakrajsek (2012) excess bond premium, and the 1-year treasury bond yield. The other two variables measure output and the price level and consist of the logarithm of industrial production and the logarithm of the consumer price index. In their baseline specification, Jarociński and Karadi (2020) use monthly interpolations of real GDP and the GDP deflator as their measures of output and prices and use

industrial production and the consumer price index as a robustness check, obtaining similar results. Here, we use the latter as these data are more readily accessible.[13] Updated values for the Jarociński and Karadi (2020) MP and CBI shocks are obtained from Marek Jarociński's website.[14]. We produce results for the effects of both shocks on 100 times the logarithms of industrial production and the CPI, and for each response variable, we use 12 lags of the remaining four monthly variables as controls for both the levels and long-differenced LP, matching the VAR lag length used in Jarociński and Karadi (2020). Our sample begins in February 1990, which is the first date that the Jarociński and Karadi (2020) shock series is available, and ends in December 2019 to avoid contamination from the global pandemic. We use data on lagged controls prior to February 1990 as needed to avoid eliminating early observations of the shock series from the estimation.

Figure 20 shows the responses of industrial production and the CPI to 1 standard deviation positive MP and CBI shocks, where the dashed line indicates the levels LP estimate and the dash-circle line indicates the long-differenced LP estimate. We consider monthly horizons up to 3 years ($H = 36$). The graph also contains shading to indicate 90% confidence intervals produced using the Newey–West standard errors, where the darkest shading indicates areas of the parameter space included in the confidence intervals for both the levels and long-differenced specifications, the lightest shading indicates inclusion in only the levels specification confidence interval, and medium shading indicates inclusion in only the long-differenced specification confidence interval.

There are several items of interest from these results. Most importantly for our study, there are noticeable differences in the estimated IRF produced by the levels vs. long-differenced specification. For all cases and horizons considered, the long-differenced specification produces an estimate farther from zero, suggestive of larger effects. The magnitude of the differences in these estimates over the 3-year horizon is large. For example, the cumulated levels specification estimate of the effects of a monetary policy shock on industrial production is −1.3%, while this climbs to −8.6% using the long-differenced specification. For the CPI effects of a monetary policy shock, these numbers are −1.2% and −3.4%, respectively. Finally, the long-differenced specification confidence intervals are wider than those from the levels specification, consistent with the simulation evidence we have presented indicating that the levels confidence intervals are undersized.

Though LP- and VAR-based IRF will estimate the same IRF asymptotically, they can give very different answers in finite samples, particularly at longer horizons (Plagborg-Møller and Wolf 2021; Montiel Olea et al. 2024). Given this, it is interesting to compare the results we obtain via LP to those reported in Jarociński and Karadi (2020) via their Bayesian VAR. For the monetary policy shock, both levels and long-differenced LP produce IRF that match the shape and sign of the estimated responses reported in Appendix C of Jarociński and Karadi (2020). The LP estimated response of industrial produc-
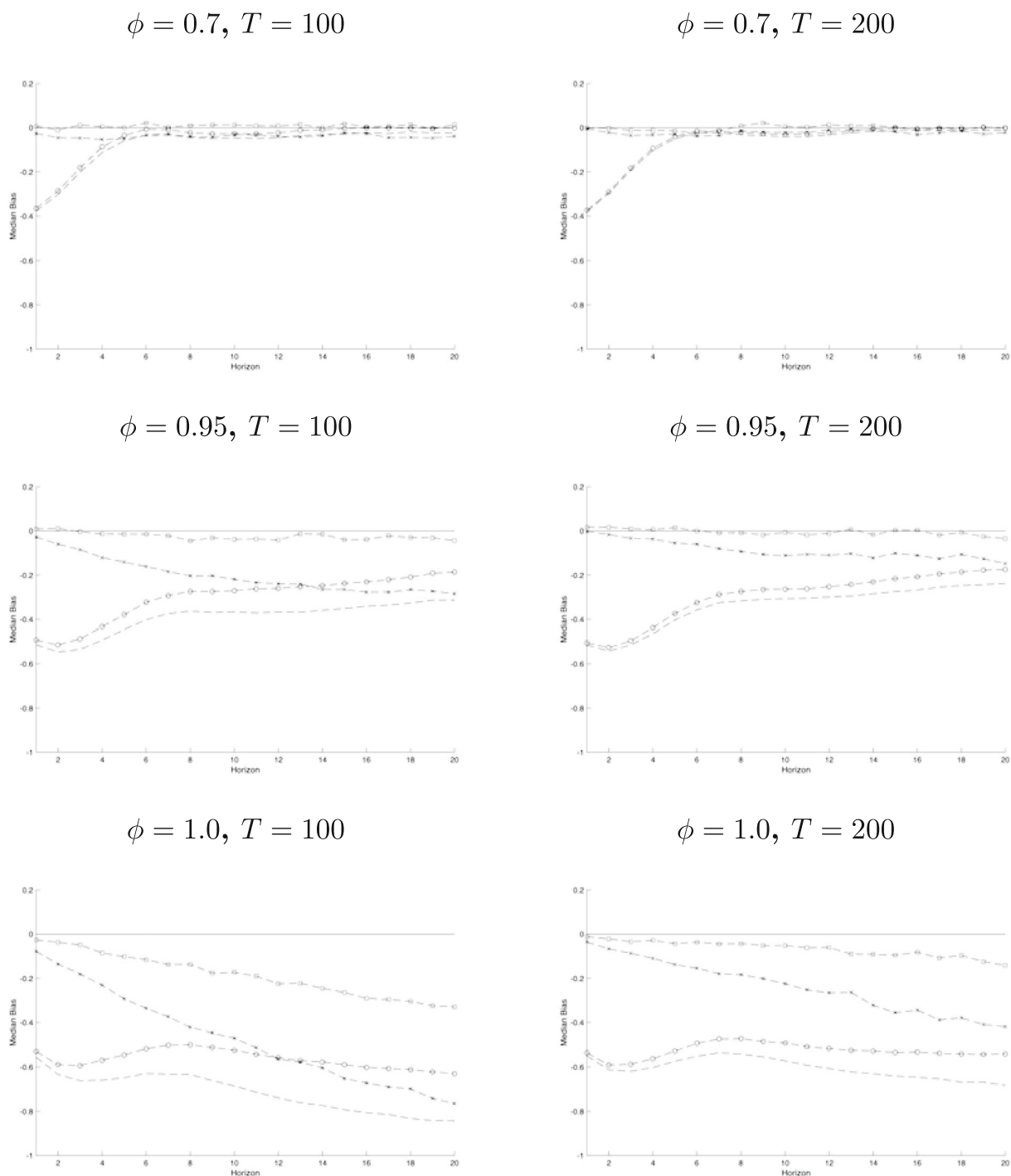
$$\phi = 0.7,\ T = 100 \qquad\qquad \phi = 0.7,\ T = 200$$

$$\phi = 0.95,\ T = 100 \qquad\qquad \phi = 0.95,\ T = 200$$

$$\phi = 1.0,\ T = 100 \qquad\qquad \phi = 1.0,\ T = 200$$

FIGURE 19 | Simulation results from levels and long-differenced LP / LP-IV and AR(8) DGP. *Note:* This figure displays simulation results from estimation of the levels and long-differenced specification of the LP / LP-IV when the true DGP is an AR(8) model and the shock of interest is endogenous. Results for three alternative values of the sum of the autoregressive parameters ($\rho = \{0.7, 0.95, 1.0\}$) and two sample sizes ($T = \{100, 200\}$) are displayed. Each figure shows the bias across simulations for the levels LP specification (dashed line), the long-differenced LP specification (dash-circle line), the levels LP-IV specification (dash-x line), and the long-differenced LP-IV specification (dash-square line).

tion to a Federal Reserve information shock is also similar to that estimated from the Bayesian VAR. The most notable difference is the CPI response to a Federal Reserve information shock. By virtue of this being an information shock, Jarociński and Karadi (2020) predict that a positive Federal Reserve information shock will be associated with higher future prices. Consistent with this, they estimate a positive response of the CPI to a positive information shock. Likewise, the response estimated from the levels LP varies between positive and negative values depending on horizon but with an overall cumulative positive response over the 3-year horizon. However, the response estimated from the long-differenced specification is negative over most horizons,
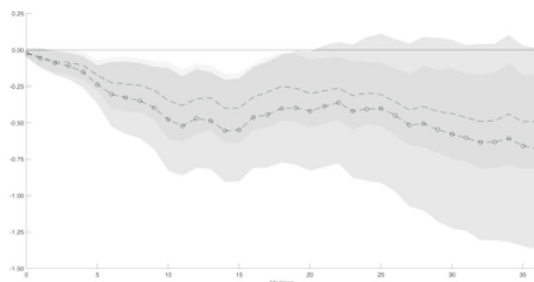
Jarociński and Karadi (2020) Monetary Policy Shock
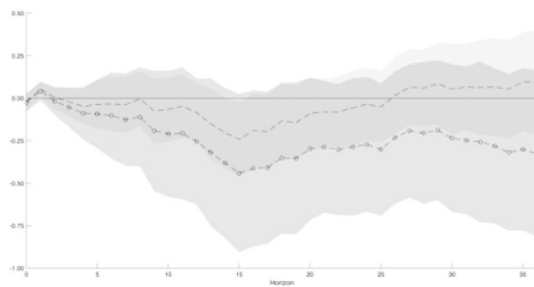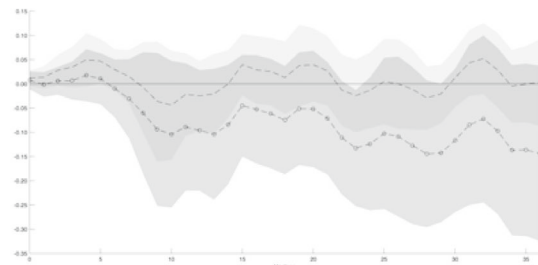

Industrial Production Response
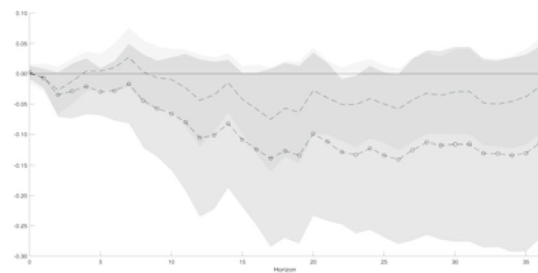
Consumer Price Index Response

Jarociński and Karadi (2020) Federal Reserve Information Shock

Industrial Production Response

Consumer Price Index Response

FIGURE 20 | Impulse response to Jarociński and Karadi (2020) shocks. *Note:* This figure shows the impulse response function to a 1 standard deviation positive Jarociński and Karadi (2020) monetary policy shock and Federal Reserve information shock. Results are shown for both 100 times the log of the US monthly industrial production and the log of the US consumer price index. The estimated response using the levels specification is the dashed line, while the estimated response using the long-differenced specification is the dash-circle line. The lightest shade of gray indicates areas of the parameter space included only in the levels specification 90% confidence interval, the medium shade indicates areas included only in the long-differenced specification 90% confidence interval, and the darkest shade indicates areas included in the confidence interval for both specifications. The sample period extends from February 1990 to December 2019.

suggesting a different directional response of prices to this information shock.

6 | Conclusion

We have investigated the finite-sample performance of LPs estimated in levels vs. long differences when there is an exogenous shock of interest available (LP) or there is an instrument for such a shock (LP-IV). We present analytic results suggestive that long-differencing should reduce the finite-sample bias visible in levels LP and LP-IV estimates. We then conduct a simulation experiment on a variety of different DGPs including AR models, unobserved components models, and VAR models.

The simulations confirm that the estimates from the levels LP and LP-IV specification are severely biased and have confidence intervals that are significantly undersized, with these deficiencies growing larger as both the persistence of the process and

the horizon of the impulse response increases. In contrast, the long-differenced LP and LP-IV specifications provide striking improvements over the levels specification in both the amount of bias and confidence interval coverage. In absolute terms, for most DGPs and impulse response horizons considered, the long-differenced specification produces close to unbiased estimates and confidence intervals with still undersized, but close to correct coverage. Importantly, the long-differenced specification provides improved inference even in cases where the underlying DGP is well inside the stationary region. Overall, our results suggest that long-differencing is a powerful tool to improve the estimation performance of LP and LP-IV regressions.

It is worth emphasizing that the results in this paper are focused specifically on the empirically popular practice of using LPs to estimate IRFs when the observed shock of interest, or instrument for such a shock, is available. In cases where this is not true and shocks are identified internally to the estimation of the LP, it is

unclear whether long-differencing would yield fruitful improvements. Indeed, in this case, the lessons from the VAR literature are likely salient, in that differencing in stationary or cointegrated systems can create noninvertibility issues for recovering structural shocks (Gospodinov et al. [2013]).

We conclude by discussing two directions for future research. First, the application of long-differenced LP and LP-IV in this paper has freely estimated the parameters on the lagged first differences that enter as controls. A variance reduction device, through shrinkage or imposition of parameter restrictions as discussed in footnote 4, may be useful for lowering estimation variance. Second, as the effect of the shock of interest became nearly permanent or exactly permanent, the long-differenced estimator retains a portion of the finite-sample bias exhibited in the levels estimator, especially at longer horizons. In these cases, it may prove fruitful to employ a local-to-unity device tailored specifically for long horizon estimation in highly persistent processes to improve the approximated small sample distribution of the long-differenced estimator (Rossi ([2005]), Rossi ([2007]), Pesavento and Rossi ([2006])).

Acknowledgments

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in the Journal Data Archive (https://doi.org/10.15456/jae.2025086.1303778166).

Endnotes

[1] The literature using state-dependent LP is crowded, with early examples provided by Auerbach and Gorodnichenko ([2013]), Tenreyro and Thwaites ([2016]), and Ramey and Zubairy ([2018]).

[2] Li et al. ([2024]) show that this robustness comes at the cost of significantly higher estimation variance as compared to VARs. The choice of LP vs. VAR then partially depends on a researcher's preferences over estimation bias and variance.

[3] Examples include Montiel Olea and Plagborg-Møller ([2021]), Plagborg-Møller and Wolf ([2021]), Montiel Olea et al. ([2024]), Gonçalves et al. ([2024]) and Xu ([2023]). Jordá and Taylor ([2025]) provide a recent survey of the LP literature.

[4] If we derive equation ([4]) directly from equation ([1]) there will be $p_D = p_L + h$ lagged first differences of the response variable on the right-hand side of the long-differenced specification. The parameters on these lagged variable will follow the mapping $\theta_{j,h} = \sum_{i=1}^{p_L} \rho_{i,h} I(i \leq j \leq h + i)$, where $I(\cdot)$ is the indicator function. These restrictions could be enforced on the long-differenced specification to ensure the same number of parameters are present in the lag structure for both the levels and long-differenced specification. This could potentially enhance efficiency for the long-differenced specification, though we do not explore this possibility here.

[5] See also Rossi ([2005]).

[6] This signal to noise ratio has very little effect on the bias observed in the levels and long-differenced LP estimators. This is discussed in more detail in Section [4.4].

[7] This will include both trend stationary DGPs and DGPs with stochastic trends that include drift.

[8] Herbst and Johannsen ([2024]) consider a sample size of $T = 50$ for simulated LP estimation based on an AR(1) DGP. All DGPs we consider in this section generate higher order dynamics than an AR(1) and would in many cases lead to heavily parameterized LP regressions relative to a sample size of $T = 50$. Thus, we focus on minimum sample sizes of $T = 100$.

[9] The literature applying UC models to macroeconomic time series is vast. Early citations include Harvey ([1985]), Watson ([1986]), Clark ([1987]), Harvey and Jaeger ([1993]) and Kuttner ([1994]).

[10] The nine variables in the Christiano et al. ([2005]) VAR are, in this order, quarterly U.S. quarterly log real GDP, log real consumption, log real investment, log GDP deflator prices, log real wages, log labor productivity, the federal funds rate, log real profits, and the growth rate of M2.

[11] See, for example, Tenreyro and Thwaites ([2016]).

[12] We focus on median, rather than mean, bias as the simulations infrequently draw a case with weak instruments, which causes an estimated parameter with outsized effects on the mean.

[13] We obtain data for industrial production, the consumer price index, and the 1-year treasury yield from the Federal Reserve Economic Database (FRED), with codes INDPRO, CPIAUCSL, and DGS1 respectively. Data on the S&P 500 index were obtained using a combination of Yahoo Finance data and FRED code SP500. Finally, the excess bond premium was obtained from the dataset provided for Bauer and Swanson ([2023]) on Michael Bauer's website (https://www.michaeldbauer.com/research/).

[14] https://marekjarocinski.github.io/jkshocks/jkshocks.html

References

Auerbach, A., and Y. Gorodnichenko. 2013. "Fiscal Multipliers in Recession and Expansion." In *Fiscal Policy After the Financial Crisis*, edited by A. Alesina and F. Giavazzi. University of Chicago Press.

Bauer, M. D., and E. T. Swanson. 2023. "A Reassessment of Monetary Policy Surprises and High-Frequency Identification." *NBER Macroeconomics Annual* 37: 87–155.

Christiano, L., M. Eichenbaum, and C. Evans. 2005. "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy." *Journal of Political Economy* 113, no. 1: 1–45.

Clark, P. K. 1987. "The Cyclical Component of U.S. Economic Activity." *Quarterly Journal of Economics* 102, no. 4: 797–814.

Gilchrist, S., and E. Zakrajsek. 2012. "Credt Spreads and Business Cycle Fluctuations." *American Economic Review* 102, no. 4: 1692–1720.

Gonçalves, S., A. M. Herrera, L. Kilian, and E. Pesavento. 2024. "State-Dependent, Local Projections." *Journal of Econometrics* 244, no. 2: 105702.

Gospodinov, N., A. M. Herrera, and E. Pesavento. 2013. "Unit Roots, Cointegration, and Pretesting in VAR Models." In *VAR Models in Macroeconomics - New Developments and Applications: Essays in Honor of Christopher A. Sims*, edited by T. B. Fomby, L. Kilian, and A. Murphy, 81–115. Emerald Group Publishing Limited.

Harvey, A. 1985. "Trends and Cycles in Macroeconomic Time Series." *Journal of Business & Economic Statistics* 3, no. 3: 216–227.

Harvey, A., and A. Jaeger. 1993. "Detrending, Stylized Facts and the Business Cycle." *Journal of Applied Econometrics* 8, no. 3: 231–247.

Herbst, E. P., and B. K. Johannsen. 2024. "Bias in Local Projections." *Journal of Econometrics* 240, no. 1: 105655.

Jarociński, M., and P. Karadi. 2020. "Deconstructing Monetary Policy Surprises - The Role of Information Shocks." *American Economic Journal: Macroeconomics* 12, no. 2: 1–43.

Jordá, O. 2005. "Estimation and Inference of Impulse Responses by Local Projections." *American Economic Review* 95, no. 1: 161–182.

Jordá, O., M. Schularick, and A. M. Taylor. 2015. "Betting the House." *Journal of International Economics* 96, no. S1: 2–18.

Jordá, O., and A. M. Taylor. 2025. "Local Projections." *Journal of Economic Literature* 63, no. 1: 59–110.

Kilian, L., and Y. J. Kim. 2011. "How Reliable are Local Projection Estimators of Impulse Responses?" *Review of Economics and Statistics* 93, no. 4: 1460–1466.

Kim, C.-J., and J. Piger. 2002. "Common Stochastic Trends, Common Cycles, and Asymmetry in Economic Fluctuations." *Journal of Monetary Economics* 49, no. 6: 1189–1211.

Kuttner, K. N. 1994. "Estimating Potential Output as a Latent Variable." *Journal of Business and Economic Statistics* 12, no. 3: 361–368.

Li, D., M. Plagborg-Møller, and C. K. Wolf. 2024. "Local Projections vs. VARs: Lessons From Thousands of DGPs." *Journal of Econometrics* 244, no. 2: 105722.

Montiel Olea, J. L., and M. Plagborg-Møller. 2021. "Local Projection Inference is Simpler and More Robust Than You Think." *Econometrica* 89, no. 4: 1789–1823.

Montiel Olea, J. L., M. Plagborg-Møller, E. Qian, and C. K. Wolf. 2024. "Double Robustness of Local Projections and Some Unpleasant Varithmetic. NBER Working Paper No. 32495".

Pesavento, E., and B. Rossi. 2006. "Small-Sample Confidence Intervals for Multivariate Impulse Response Functions at Long Horizons." *Journal of Applied Econometrics* 21, no. 8: 1135–1155.

Plagborg-Møller, M., and C. K. Wolf. 2021. "Local Projections and VARs Estimate the Same Impulse Responses." *Econometrica* 89, no. 2: 955–980.

Ramey, V., and S. Zubairy. 2018. "Government Spending Multipliers in Good Times and in Bad: Evidence From U.S. Historical Data." *Journal of Political Economy* 126: 850–901.

Ramey, V. A. 2016. "Macroeconomic Shocks and Their Propagation." In *Handbook of Macroeconomics*, vol. 2, 71–162. Elsevier.

Rossi, B. 2005. "Testing Long-Horizon Predictive Ability With High Persistence, and the Meese-Rogoff Puzzle." *International Economic Review* 46, no. 1: 61–92.

Rossi, B. 2007. "Expectations Hypotheses Tests at Long Horizons." *Econometrics Journal* 10, no. 3: 554–579.

Staiger, D., and J. Stock. 1997. "Instrumental Variables Regression With Weak Instruments." *Econometrica* 65, no. 3: 557–586.

Stock, J. H., and M. W. Watson. 2018. "Identification and Estimation of Dynamic Causal Effects in Macroeconomics Using External Instruments." *Economic Journal* 128, no. 610: 917–948.

Tenreyro, S., and G. Thwaites. 2016. "Pushing on a String: U. S. Monetary Policy Is Less Powerful in Recessions." *American Economic Journal: Macroeconomics* 8, no. 4: 43–74.

Watson, M. W. 1986. "Univariate Detrending Methods With Stochastic Trends." *Journal of Monetary Economics* 18, no. 1: 49–75.

Xu, K.-L. 2023. "Local Projection Based Inference Under General Conditions. Working Paper, Center for Applied Economics and Policy Research".

Supporting Information

Additional supporting information can be found online in the Supporting Information section. JAE_MS14771_Online_Appendix.pdf

Appendix A

This appendix provides additional detail behind the derivation of Equations (10), (12), and (13). Consider the AR(1) data generating process where s_t represents an observed, exogenous, shock of interest as follows:

$$y_t = \alpha + \beta_0 s_t + \phi y_{t-1} + \omega_t.$$

We assume that ω_t is i.i.d. $(0, \sigma_\omega^2)$, s_t is i.i.d. (μ_s, σ_s^2), and $E(s_t \omega_{t+j}) = 0, \ \forall \ j$. Without loss of generality, we set $\mu_s = 0$. The correctly specified levels LP for the AR(1) case is as follows:

$$y_{t+h} = c_h^L + \beta_h s_t + \rho_h y_{t-1} + v_{t+h},$$

where $v_{t+h} = \sum_{i=0}^{h-1} \beta_i s_{t+h-i} + \sum_{i=0}^{h} \phi^i \omega_{t+h-i}$ Consider the expected value of the sample covariance between s_t and v_{t+h} computed over the time period $\{t = 1, 2, \dots, T + h\}$ as follows:

$$E(cov_{s_t, v_{t+h}}) = E\left(\frac{1}{T} \sum_{t=1}^{T} (s_t - \overline{s}_{[0]})(v_{t+h} - \overline{v}_{[h]}) \right),$$

where for a random variable ξ, $\overline{\xi}_{[\tau]} = \frac{1}{T} \sum_{t=1+\tau}^{T+\tau} \xi_t$. In this notation, τ reflects the offset of the sample period from $1 \to T$ used to compute a sample mean. Expanding and noting that $E(s_t v_{t+h}) = 0$, we have the following:

$$E(cov_{s_t, v_{t+h}}) = -E\left(\overline{s}_{[0]} \overline{v}_{[h]} \right).$$

From the definition of v_{t+h},

$$-E\left(\overline{s}_{[0]} \overline{v}_{[h]} \right) = -\sum_{i=0}^{h-1} \beta_i E\left(\overline{s}_{[0]} \overline{s}_{[h-i]} \right) - \sum_{i=0}^{h} \phi^i E\left(\overline{s}_{[0]} \overline{\omega}_{[h-i]} \right).$$

There are $T - |h - i|$ overlapping values of s_t used in the calculation of $\overline{s}_{[0]}$ and $\overline{s}_{[h-i]}$. Given this and recognizing that $h \geq i$ in the equations above, it follows that

$$E\left(\overline{s}_{[0]} \overline{s}_{[h-i]} \right) = \frac{\sigma_s^2}{T^2}(T - h + i).$$

Also, from the exogeneity of s_t,

$$E\left(\overline{s}_{[0]} \overline{\omega}_{[h-i]} \right) = 0.$$

Combining gives us Equation (10) as follows:

$$E(cov_{s_t, v_{t+h}}) = \frac{\sigma_s^2}{T^2} \left[-\sum_{i=0}^{h-1} \beta_i (T - h + i) \right].$$

Turning to the long-differenced LP, the correctly specified LP for the AR(1) DGP is as follows:

$$\Delta_h y_{t+h} = c_h^D + \beta_h s_t + \theta_{1,h} \Delta y_{t-1} + \dots + \theta_{h+1,h} \Delta y_{t-h-1} + u_{t+h},$$

where $\beta_h = \beta_0 \phi^h$, $\theta_{i,h} = \phi^{h+1}$ and

$$u_{t+h} = \sum_{i=0}^{h-1} \beta_i (s_{t+h-i} - s_{t-1-i}) - \beta_h s_{t-h-1}$$
$$+ \sum_{i=0}^{h} \phi^i (\omega_{t+h-i} - \omega_{t-1-i}).$$

Consider the expected value of the sample covariance between s_t and u_{t+h} computed over the time period $\{-h, -h+1, \dots, 0, 1, \dots, T+h\}$:

$$E(cov_{s_t,u_{t+h}}) = E\left(\frac{1}{T}\sum_{t=1}^{T}(s_t - \bar{s}_{[0]})(u_{t+h} - \bar{u}_{[h]})\right).$$

Using similar calculations as for the levels case, we have the following:

$$E(cov_{s_t,u_{t+h}}) = -E\left(\bar{s}_{[0]}\bar{u}_{[h]}\right).$$

From the definition of u_{t+h},

$$-E\left(\bar{s}_{[0]}\bar{u}_{[h]}\right) = -\sum_{i=0}^{h-1}\beta_i E\left(\bar{s}_{[0]}\left(\bar{s}_{[h-i]} - \bar{s}_{[-(i+1)]}\right)\right)$$
$$+ \beta_h E\left(\bar{s}_{[0]}\bar{s}_{[-(h+1)]}\right)$$
$$-\sum_{i=0}^{h}\phi^i E\left(\bar{s}_{[0]}\left(\bar{\omega}_{[h-i]} - \bar{\omega}_{[-(i+1)]}\right)\right).$$

There are $T - |\tau|$ overlapping values of s_t used in the calculation of $\bar{s}_{[0]}$ and $\bar{s}_{[\tau]}$. Given this and noting that $h \geq 0$, $i \geq 0$, and $h \geq i$, we have the following:

$$E\left(\bar{s}_{[0]}\left(\bar{s}_{[h-i]} - \bar{s}_{[-(i+1)]}\right)\right) = \frac{\sigma_s^2}{T^2}(T-h+i) - \frac{\sigma_s^2}{T^2}(T-i-1)$$
$$= \frac{\sigma_s^2}{T^2}(1-h+2i)$$

and

$$E\left(\bar{s}_{[0]}\bar{s}_{[-(h+1)]}\right) = \frac{\sigma_s^2}{T^2}(T-h-1).$$

Finally, from the exogeneity of s_t,

$$E\left(\bar{s}_{[0]}\left(\bar{\omega}_{[h-i]} - \bar{\omega}_{[-(i+1)]}\right)\right) = 0.$$

Substituting and rearranging, we have Equation (12) as follows:

$$E(cov_{s_t,u_{t+h}}) = \frac{\sigma_s^2}{T^2}\left[\beta_h(T-h-1) - \sum_{i=0}^{h-1}\beta_i(1-h+2i)\right]$$

We now turn to the instrumental variables (LP-IV) case. We again assume the AR(1) data generating process but now assume that S_t is an observed endogenous variable of interest, such that $E(s_t\omega_t) \neq 0$. The instrument $\varepsilon_t \sim$ i.i.d. $(0, \sigma_\varepsilon^2)$ is such that $E(\varepsilon_t\omega_{t+j}) = 0$, $\forall j$ and has first-stage regression:
$$s_t = \lambda + \gamma\varepsilon_t + \eta_t,$$

where $E(\varepsilon_t\eta_t) = 0$ and we assume $\lambda = 0$ without loss of generality. Assuming γ is known, the correctly specified LP-IV in levels is as follows:

$$y_{t+h} = c_h^L + \beta_h\hat{s}_t + \rho_{1,h}y_{t-1} + v_{t+h}^{IV},$$

where $\hat{s}_t = \gamma\varepsilon_t$ and $v_{t+h}^{IV} = v_{t+h} + \beta_h\eta_t$. Consider the expected value of the sample covariance between $\hat{s}_t$ and v_{t+h}^{IV} computed over the time period $\{t = 1, 2, \dots, T+h\}$:

$$E(cov_{\hat{s}_t,v_{t+h}^{IV}}) = E\left(\frac{1}{T}\sum_{t=1}^{T}\left(\hat{s}_t - \bar{\hat{s}}_{[0]}\right)\left(v_{t+h}^{IV} - \bar{v}_{[h]}^{IV}\right)\right).$$

Expanding and noting that $E\left(\hat{s}_t v_{t+h}^{IV}\right) = 0$, we have the following:

$$E\left(cov_{\hat{s}_t,v_{t+h}^{IV}}\right) = -E\left(\bar{\hat{s}}_{[0]}\bar{v}_{[h]}^{IV}\right).$$

From the definition of v_{t+h}^{IV},

$$-E\left(\bar{\hat{s}}_{[0]}\bar{v}_{[h]}^{IV}\right) = -\sum_{i=0}^{h-1}\beta_i E\left(\bar{\hat{s}}_{[0]}\bar{s}_{[h-i]}\right)$$
$$-\sum_{i=0}^{h}\phi^i E\left(\bar{\hat{s}}_{[0]}\bar{\omega}_{[h-i]}\right).$$

There are $T - |h-i|$ overlapping values of $\gamma\varepsilon_t$ used in the calculation of $\bar{\hat{s}}_{[0]}$ and $\bar{s}_{[h-i]}$. Given this and recognizing that $h \geq i$ in the equations above, it follows that

$$E\left(\bar{\hat{s}}_{[0]}\bar{s}_{[h-i]}\right) = \frac{\gamma^2\sigma_\varepsilon^2}{T^2}(T-h+i).$$

Also, from the exogeneity of ε_t,

$$E\left(\bar{\hat{s}}_{[0]}\bar{\omega}_{[h-i]}\right) = 0.$$

Combining gives us the first proportionality result in Equation (13) as follows:

$$E(cov_{\hat{s}_t,v_{t+h}^{IV}}) = \frac{\gamma^2\sigma_\varepsilon^2}{T^2}\left[-\sum_{i=0}^{h-1}\beta_i(T-h+i)\right]$$
$$= \kappa E(cov_{s_t,v_{t+h}}),$$

where

$$\kappa = \frac{\gamma^2\sigma_\varepsilon^2}{\sigma_s^2}.$$

The correctly specified LP-IV in long differences is as follows:

$$\Delta_h y_{t+h} = c_h^D + \beta_h\hat{s}_t + \theta_{1,h}\Delta y_{t-1} + \cdots + \theta_{h+1,h}\Delta y_{t-h-1} + u_{t+h}^{IV},$$

where $\beta_h = \beta_0\phi^h$, $\theta_{i,h} = \phi^{h+1}$, and $u_{t+h}^{IV} = u_{t+h} + \beta_h\eta_t$. Using similar arguments as above, it can be shown that

$$E(cov_{\hat{s}_t,u_{t+h}^{IV}}) = \frac{\gamma^2\sigma_\varepsilon^2}{T^2}\left[\beta_h(T-h-1) - \sum_{i=0}^{h-1}\beta_i(1-h+2i)\right]$$
$$= \kappa E(cov_{s_t,u_{t+h}}),$$

which provides the second proportionality result in (13).

Appendix B

This appendix generalizes the analytical results in Appendix A to the case of a VAR(p). Suppose the $N \times 1$ vector of endogenous variables $Y_t = (y_{1,t}, y_{2,t}, \dots, y_{N,t})'$ follows a VAR(p) process, where $S_t = (s_{1,t}, s_{2,t}, \dots, s_{N,t})'$ represents an $N \times 1$ vector of the following exogenous shocks:

$$Y_t = C + B_0 S_t + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-1} + \cdots + \Phi_p Y_{t-p} + W_t.$$

In this notation, $W_t = (w_{1,t}, w_{2,t}, \dots, w_{N,t})'$ represents an $N \times 1$ vector of disturbances that are assumed i.i.d.$(0_N, \Sigma_W)$, where 0_N represents an $N \times 1$ vector of zeros. The vector of exogenous shocks, S_t, is i.i.d.(μ_S, Σ_S), with $E\left(S_t W_{t+j}'\right) = 0, \forall j$. Without loss of generality, we set $\mu_S = 0$. We further assume that the exogenous shocks are orthogonal, such that $E(s_{a,t}s_{b,t}) = 0$, $\forall a \neq b$. This implies that $\Sigma_S = diag\left(\sigma_{s,1}^2, \sigma_{s,2}^2, \dots, \sigma_{s,N}^2\right)$. Finally, B_0 is an $N \times N$ matrix of initial responses to the exogenous shocks, while the Φ_i are $N \times N$ matrices of lag parameters.

Cast this VAR in companion form as follows:

$$\widetilde{Y}_t = \widetilde{C} + B\widetilde{S}_t + F\widetilde{Y}_{t-1} + \widetilde{W}_t \tag{B1}$$

where $\widetilde{Y}_t = \left(Y_t', Y_{t-1}', \ldots, Y_{t-p+1}'\right)'$, $\widetilde{S}_t = \left(S_t', 0'_{(N*(p-1))}\right)'$, and $\widetilde{W}_t = \left(W_t', 0'_{(N*(p-1))}\right)'$. The matrix B is $Np \times Np$:

$$B = \begin{bmatrix} B_0 & 0_N & \ldots & 0_N \\ 0_N & 0_N & \ldots & 0_N \\ \vdots & \vdots & \ddots & \vdots \\ 0_N & 0_N & \ldots & 0_N \end{bmatrix},$$

and the matrix F is the standard $Np \times Np$ companion matrix:

$$F = \begin{bmatrix} \Phi_1 & \Phi_2 & \ldots & \Phi_p - 1 & \Phi_p \\ I_N & 0_N & \ldots & 0_N & 0_N \\ 0_N & I_N & \ddots & \vdots & 0 \\ \vdots & \ddots & \ddots & 0_N & \vdots \\ 0_N & \ldots & 0_N & I_N & 0_N \end{bmatrix}.$$

In the following discussion, $F^a_{b,c}$ indicates the (b,c) element of F^a.

Suppose that we observe the dth structural shock ($s_{d,t}$), and we are interested in estimating the horizon h response of the kth endogenous variable $y_{k,t+h}$ to s_{dt}. By iterating Equation (B1) forward and identifying the kth row, we arrive at the correct specification of the levels LP as follows:

$$y_{k,t+h} = c_{k,h}^L + \beta_{h,k,d} s_{d,t}$$
$$+ \sum_{j=1}^p \left(\sum_{n=1}^N F_{k,(n+N(j-1))}^{h+1} y_{n,t-j} \right) + v_{k,t+h},$$

where $\beta_{h,k,d}$ is the (k,d) element of $F^h B$. Also,

$$v_{k,t+h} = \sum_{i=0}^{h-1} \left(\sum_{n=1}^N \beta_{i,k,n} s_{n,t+h-i} \right)$$
$$+ \sum_{i=0}^h \left(\sum_{n=1}^N F_{k,n}^i w_{n,t+h-i} \right) + \sum_{n \neq d} \beta_{h,k,n} s_{n,t}.$$

Consider the expected value of the sample covariance between $s_{d,t}$ and $v_{k,t+h}$ computed over the time period $\{t = 1, 2, \ldots, T+h\}$:

$$E(cov_{s_{d,t}, v_{k,t+h}}) = E\left(\frac{1}{T} \sum_{t=1}^T \left(s_{d,t} - \overline{s}_{d,[0]} \right) \left(v_{k,t+h} - \overline{v}_{k,[h]} \right) \right),$$

where the notation $\overline{\xi}_{[\tau]}$ is defined in Appendix A. Given the assumptions made on the stochastic components above, it is straightforward to see that $E(s_{d,t} v_{k,t+h}) = 0$. It follows that

$$E(cov_{s_{d,t}, v_{k,t+h}}) = -E\left(\overline{s}_{d,[0]} \overline{v}_{k,[h]} \right).$$

Given the orthogonality of the exogenous shocks, it is apparent that $E\left(\overline{s}_{d,[0]} \overline{s}_{j,[0]} \right) = 0$, $\forall j \neq d$, while the exogeneity of $s_{d,t}$ implies that $E\left(\overline{s}_{d,[0]} \overline{w}_{j,[h-i]} \right) = 0$, $\forall j$. With these two results and from the definition of $v_{k,t+h}$, we have the following:

$$-E\left(\overline{s}_{d,[0]} \overline{v}_{k,[h]} \right) = -\sum_{i=0}^{h-1} \beta_{i,k,d} E\left(\overline{s}_{d,[0]} \overline{s}_{d,[h-i]} \right).$$

There are $T - |h - i|$ overlapping values of s_t used in the calculation of $\overline{s}_{d,[0]}$ and $\overline{s}_{d,[h-i]}$. Given this and recognizing that $h \geq i$ in the equations above, it follows that

$$E\left(\overline{s}_{d,[0]} \overline{s}_{d,[h-i]} \right) = \frac{\sigma_{s,d}^2}{T^2}(T - h + i).$$

Combining gives us the analogous result as for the AR(1) case in Equation (10) as follows:

$$E(cov_{s_{d,t}, v_{k,t+h}}) = \frac{\sigma_{s,d}^2}{T^2} \left[-\sum_{i=0}^{h-1} \beta_{i,k,d}(T - h + i) \right] \tag{B2}$$

Equation (B2) demonstrates very similar features to the expected covariance for the AR(1) case. Specifically, the size of the expected covariance between $s_{d,t}$ and $v_{k,t+h}$ depends on the value of $\beta_{i,k,d}$ for $i = 0, \ldots h - 1$. That is, the expected covariance depends on the value of the true IRF for the response of $y_{k,t+i}$ to $s_{d,t}$ at all horizons up to $h - 1$. Notably, the expected covariance does not depend on dynamic multipliers for other shocks. Second, for IRFs where the dynamic multipliers have the same sign, the expected covariance will grow in absolute value with the horizon h. Third, the sample size influences the size of the expected covariance. As T grows, the denominator grows with respect to the numerator and shrinks the size of the covariance.

Turning to the long-differenced LP, the correctly specified LP for the VAR(p) DGP is as follows:

$$\Delta_h y_{k,t+h} = c_{k,h}^D + \beta_{h,k,d} s_{d,t} + \sum_{j=1}^{p+h} \left(\sum_{n=1}^N \psi_{h,j,k,n} \Delta y_{n,t-j} \right) + u_{k,t+h},$$

where

$$u_{k,t+h} = \sum_{i=0}^{h-1} \left(\sum_{n=1}^N \beta_{i,k,n}(s_{n,t+h-i} - s_{n,t-1-i}) \right) - \beta_{h,k,d} s_{d,t-h-1}$$
$$+ \sum_{i=0}^h \left(\sum_{n=1}^N F_{k,n}^i \left(w_{n,t+h-i} - w_{n,t-1-i} \right) \right)$$
$$+ \sum_{n \neq d} \beta_{h,k,n}\left(s_{n,t} - s_{n,t-h-1} \right).$$

Consider the expected value of the sample covariance between $s_{d,t}$ and $u_{k,t+h}$ computed over the time period $\{-h, -h + 1, \ldots, 0, 1, \ldots, T + h\}$ as follows:

$$E(cov_{s_{d,t}, u_{k,t+h}}) = E\left(\frac{1}{T} \sum_{t=1}^T \left(s_{d,t} - \overline{s}_{d,[0]} \right) \left(u_{k,t+h} - \bar{u}_{k,[h]} \right) \right).$$

Again, given the assumptions above, it is straightforward to see that

$$E(cov_{s_{d,t}, u_{k,t+h}}) = -E\left(\overline{s}_{d,[0]} \bar{u}_{k,[h]} \right).$$

Then, using the orthogonality of the exogenous shocks and exogeneity of $s_{d,t}$, as well as the definition of $u_{k,t+h}$, we have the following:

$$E\left(\overline{s}_{d,[0]} \bar{u}_{k,[h]} \right) - = -\sum_{i=0}^{h-1} \beta_{i,k,d} E\left(\overline{s}_{d,[0]}\left(\overline{s}_{d,[h-i]} - \overline{s}_{d,[-(i+1)]} \right) \right)$$
$$+ \beta_{h,k,d} E\left(\overline{s}_{d,[0]} \overline{s}_{d,[-(h+1)]} \right).$$

There are $T - |\tau|$ overlapping values of $s_{d,t}$ used in the calculation of $\overline{s}_{d,[0]}$ and $\overline{s}_{d,[\tau]}$. Given this and noting that $h \geq 0$, $i \geq 0$, and $h \geq i$, we have the following:

$$E\left(\overline{s}_{d,[0]}\left(\overline{s}_{d,[h-i]} - \overline{s}_{d,[-(i+1)]} \right) \right)$$
$$= \frac{\sigma_{s,d}^2}{T^2}(T - h + i) - \frac{\sigma_{s,d}^2}{T^2}(T - i - 1)$$
$$= \frac{\sigma_{s,d}^2}{T^2}(1 - h + 2i)$$

and

$$E\left(\overline{s}_{[0]}\overline{s}_{[-(h+1)]}\right) = \frac{\sigma_s^2}{T^2}(T - h - 1).$$

Substituting and rearranging, we have the following:

$$E(cov_{s_{d,t},u_{k,t+h}}) = \frac{\sigma_{s,d}^2}{T^2}\left[\beta_{h,k,d}(T - h - 1) - \sum_{i=0}^{h-1}\beta_{i,k,d}(1 - h + 2i)\right] \quad (B3)$$

The expected covariance in Equation (B3) is very similar to the analogous Equation (12) for the AR(1) case. Further, using similar arguments to those in Section 3, we can see that the expected covariance in Equation (B3) will in general be much smaller than the expected covariance in Equation (B2).