Supplemental Materials for Taking Variance Seriously: Visualizing the Statistical and Substantive Significance of ARCH-GARCH Models

Allyson L. Benton
Soren Jordan
Andrew Q. Philips

Contents

1 The Delta Method: An Alternative to Bootstrapping

1.1 Applying the Method

One clear alternative method to our bootstrapping techniques discussed in the main manuscript is to create a prediction and confidence intervals (or standard errors) using the delta method approximation, which can be used for a nonlinear combination of regression estimates. In other words, it will provide both an expectation (the expected value of $\hat{\sigma}_t^2$, conditional on the estimates from the regression model, as well as whatever we set the value of our regressors to in the volatility equation) as well as an uncertainty measure around this.

Starting with our vector of estimated parameters from our GARCH model, $\hat{\boldsymbol{\gamma}}$, our goal is transformation $g(\hat{\boldsymbol{\gamma}})$—the expected conditional error variance using our simulation approach—which has a corresponding estimated variance covariance matrix $\widehat{\text{Var}}[g(\hat{\boldsymbol{\gamma}})]$. The latter is a function of both the estimated variance covariance matrix of $\hat{\boldsymbol{\gamma}}$ as well as the estimated partial derivatives with respect to the parameters in $g(\hat{\boldsymbol{\gamma}})$ (Greene 2018). From this, upper and lower confidence intervals can be created using $\sqrt{\widehat{\text{Var}}[g(\hat{\boldsymbol{\gamma}})]}$ along with the appropriate quantile of the t-distribution.

In terms of actually using the delta method step-by-step, the process is quite similar to the bootstrap techniques outlined in the main paper, although a single calculation is being performed at each simulation time rather than a set of calculations across all bootstraps that are then averaged over. The process consists of the following steps:

1. *Estimate the ARCH-GARCH model*.

2. *Generate the the expected conditional error variance*. Since ARCH-GARCH models often include the lagged variance, we need a stable estimate of that variance before the counterfactual shock as well as the changes to that variance after the shock.

    - *Set the covariates to the values of interest*.

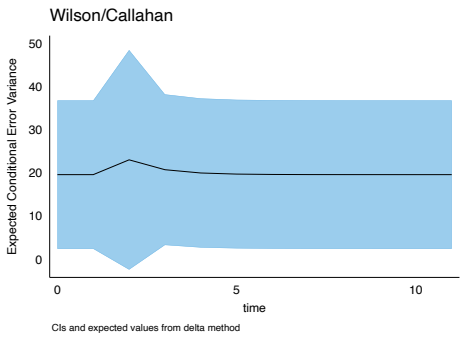
- *Conduct a "burn in" process to estimate the conditional error variance in the first period* $m = 1$. Initialize a draw of $\mathrm{E}[\tilde{\sigma}^2]$, then use this value to update the prediction at the next time point, and so on, across a sufficient number (our experience was 30) or so burn in periods to get a stable prediction that does not change from period to period. This final stable prediction becomes our first scenario prediction, $\mathrm{E}[\tilde{\sigma}_1^2]$, for time $m = 1$.
- *Simulate the conditional error variance in the remainder of the pre-shock period*. Use the delta method to create expected values (and record the confidence intervals) up until the period just before a counterfactual shock, $m = 1, 2, \cdots, m = s - 1$.
- *Simulate the impact of the covariate shock on the conditional error variance*. At $m = s$, implement a counterfactual change in one of the covariates and calculate $\mathrm{E}[\tilde{\sigma}_{m=s}^2]$ and its confidence intervals.
- *Simulate the future evolution on the conditional error variance*. Continue simulating the future evolution until the final simulation time M.

3. *Graph the predictions*.

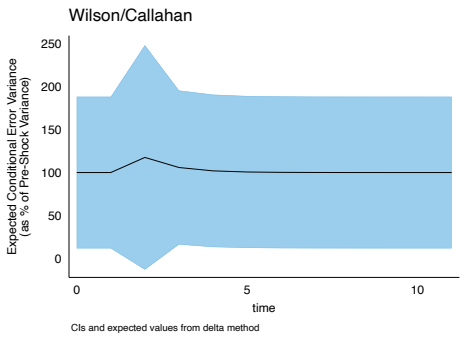
One issue with using the delta method is that while the expected value of the heteroskedastic conditional variance will be strictly positive, the associated confidence intervals might not be, which of course is nonsensical since error variances must be positive. A similar issue arose for the bootstrapping techniques as well when using the standard error approach. On the other hand, the delta method is far faster than the bootstrapping techniques that require estimation of B or $B + 1$ (maximum entropy and residual, respectively) ARCH-GARCH models; in our experience it takes only slightly longer than the parametric bootstrap, as the delta method must be used recursively (e.g., after getting a burn-in estimate of $\mathrm{E}[\tilde{\sigma}_1^2]$, this is needed because it enters into the equation as the GARCH term to estimate $\mathrm{E}[\tilde{\sigma}_2^2]$, and so on).

1.2 Results Using the Delta Method

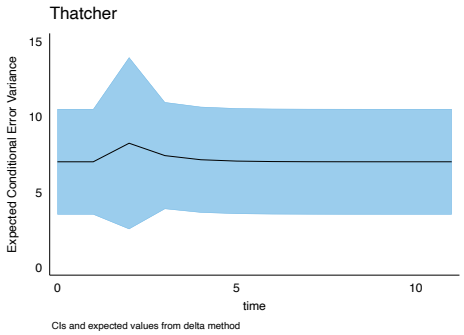
Below we replicate our results from the main manuscript, now using the delta method instead of bootstrapping. Figure SM.1 replicates Figure 2 in the main manuscript, and shows a positive +4 increase in trade for a single period, under the different UK prime ministers. The left column shows the expected conditional error variance in its original metric, while the right column rescales the expected conditional error variance as a percentage of the pre-shock variance. Figure SM.2 is similar, but shows a negative decline in trade of the same magnitude. The substantive conclusions using the delta method are exactly the same as the bootstrapping techniques discussed in the main manuscript; while volatility increases (decreases) in response to a positive (negative) trade shock, these changes are not statistically significant and persist only temporarily.



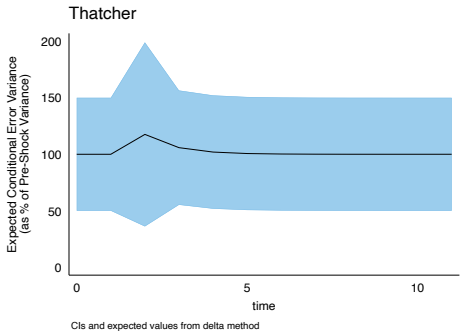
(a) Wilson/Callahan Unscaled



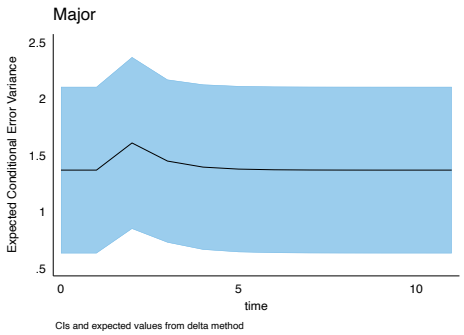
(b) Wilson/Callahan Rescaled



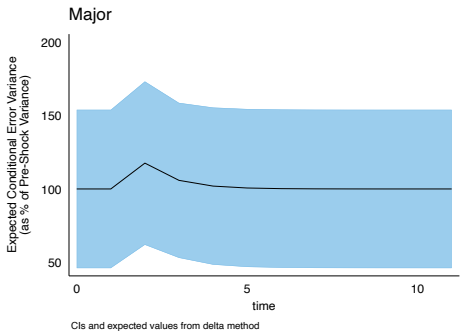
(c) Thatcher Unscaled



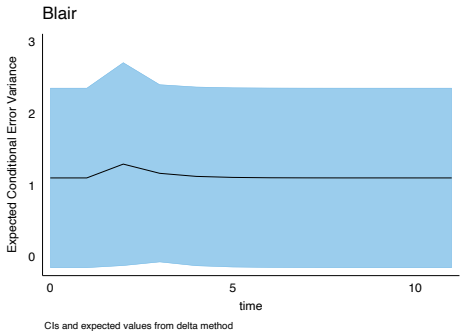
(d) Thatcher Rescaled



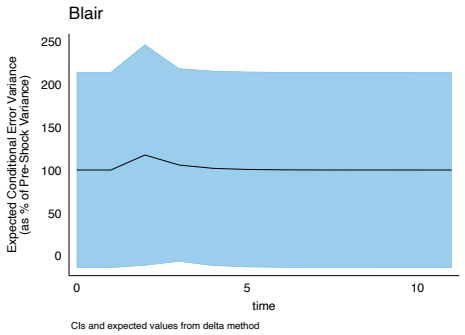
(e) Major Unscaled



(f) Major Rescaled



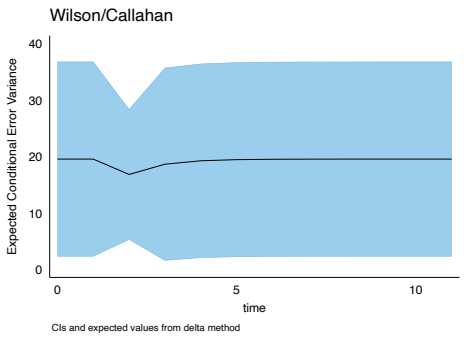
(g) Blair Unscaled



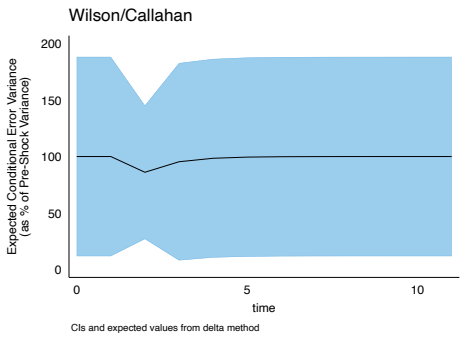
(h) Blair Rescaled

Figure SM. 1: Replication of +4 trade shock using delta method (from Hellwig example, Figure 2 in main manuscript)

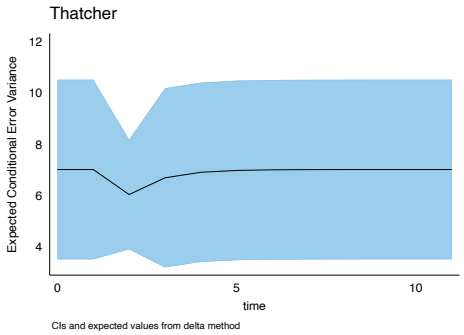
Note: For each prime minister, all other prime minister dummy variables set to 0. From left to right: delta method (unscaled), delta method (rescaled). Black line shows expected conditional error variance, with 95% confidence intervals shown.



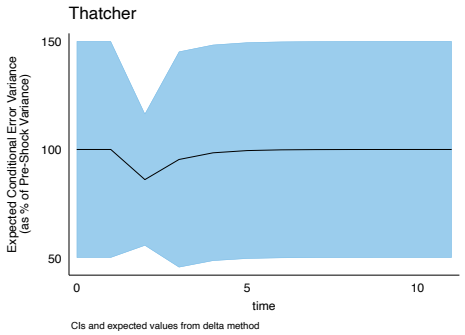
(a) Wilson/Callahan Unscaled



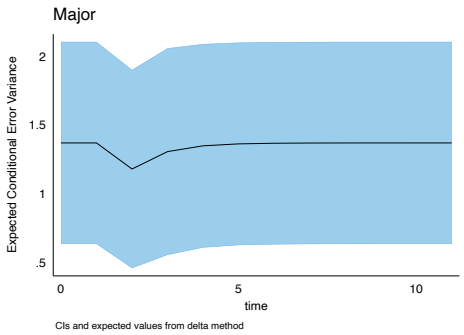
(b) Wilson/Callahan Rescaled



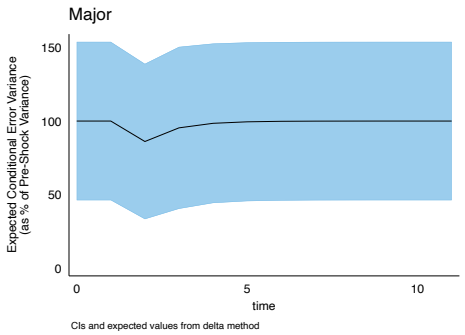
(c) Thatcher Unscaled



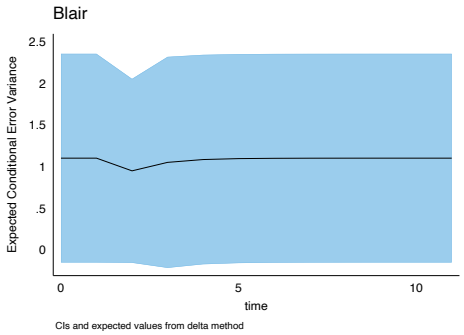
(d) Thatcher Rescaled



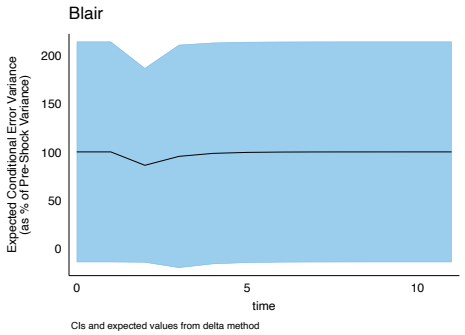
(e) Major Unscaled



(f) Major Rescaled



(g) Blair Unscaled



(h) Blair Rescaled

Figure SM. 2: Replication of -4 trade shock using delta method (from Hellwig example, Figure 2 in main manuscript)

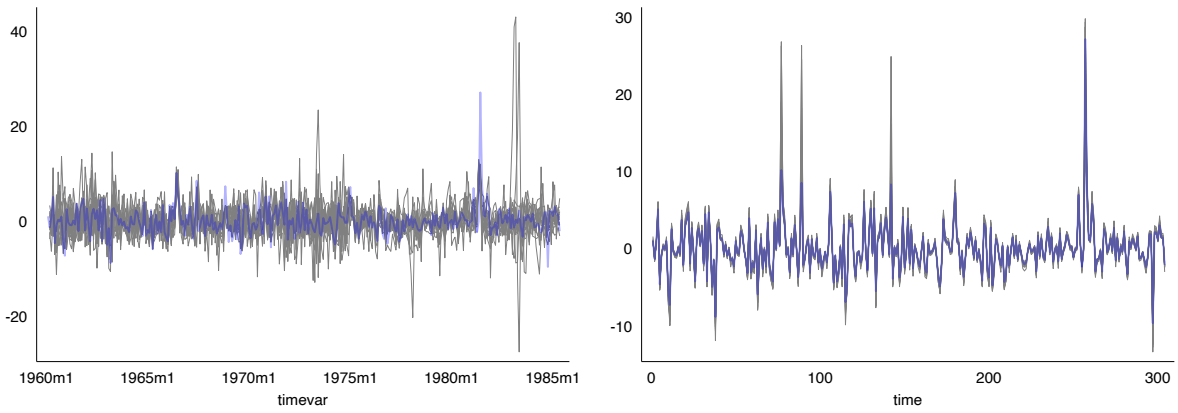
Note: For each prime minister, all other prime minister dummy variables set to 0. From left to right: delta method (unscaled), delta method (rescaled). Black line shows expected conditional error variance, with 95% confidence intervals shown.

2 Comparison of Bootstrapping Techniques

In this section we further compare some of the features of the different bootstrapping techniques described in the main paper.

2.1 General Considerations

- The parametric bootstrap takes far less time to create and simulate predictions than the other two techniques since the latter is creating B synthetic replicates of the dependent variable, while the former is creating posterior draws of parameters, given the (single) GARCH model estimated. An example of the creation of several synthetic series for the dependent variable in Hellwig's example—UK government support—are shown in Figure SM.3, for the residual bootstrap (Figure SM.3a) and maximum entropy bootstrap (Figure SM.3b). There are of course no synthetic series created with the parametric bootstrap. The original series is shown in blue, while synthetic replicates are in light gray.



(a) Residual bootstrap

(b) Maximum entropy bootstrap

Figure SM. 3: Synthetic UK government support series created through the bootstrapping techniques

Note: Original series shown in blue, replicates in gray.

- GARCH models are notorious for not converging. Therefore, while we only needed to estimate a single model for the parametric bootstrap draws, both the

residual and maximum entropy bootstrapping techniques involve estimating B individual GARCH models to obtain B parameter estimates. Not only is such a process time-intensive, but estimates do not always converge. Our solution was to simply create enough bootstrapped series until we obtained the desired number of B estimates (we set $B = 1000$).

- While the simulation figures did not differ drastically across techniques, we also examined the bootstrapped parameter estimates from our substantive examples to see whether they differed across our bootstrapping methods. Of course, this is only an applied example; below we investigate differences in performances more systematically.

  - In Figure SM.4 we show the distribution of bootstrapped estimates of the GARCH parameter from the Schneider and Troeger example. On this figure we overlay the original GARCH estimate (solid black line) and 95% confidence intervals (dashed lines). While both the parametric and maximum entropy bootstrap distributions are centered on the original estimate of around 0.7, the residual based technique appears to be consistently underestimating the magnitude of the GARCH parameter. Moreover, the distributions are much more widely dispersed than the other two techniques. If anything, the parametric and maximum entropy techniques appear to have a very small—perhaps too small—distribution around the original estimate.
  - In Figure SM.5 we present the distributions for the FTSE parameter, and in Figure SM.6 we show the distributions for the Palestinian-Israeli conflict severity coefficient. Once again the residual bootstrap appears to have attenuated parameters, on average, although the overall variance of the coefficients is similar to that of the parametric bootstrap. Here it is also clear that the maximum entropy technique is probably far too confident; its parameter distributions in both figures are extremely concentrated, which would likely lead to confidence intervals that are too small.

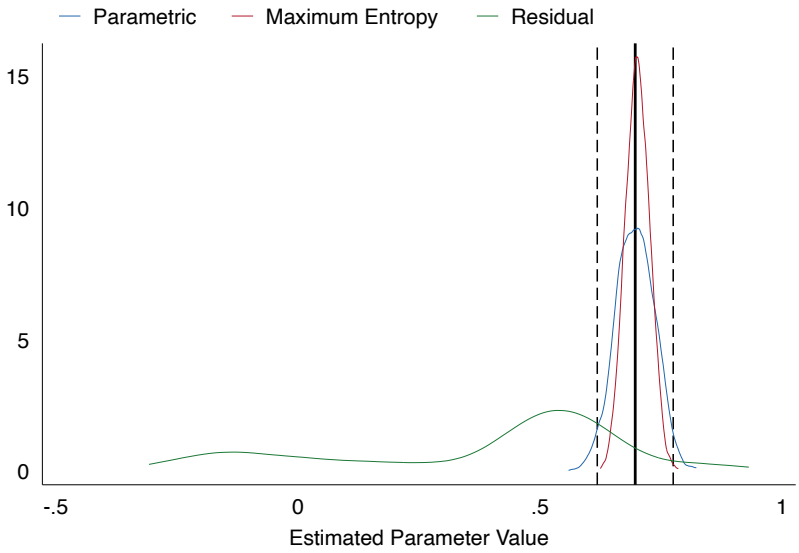


Figure SM. 4: Comparison of bootstrapping techniques, GARCH parameter from Schneider and Troeger's example

Note: Original estimated parameter shown (black solid vertical line) along with associated 95% confidence intervals (black dashed vertical lines).

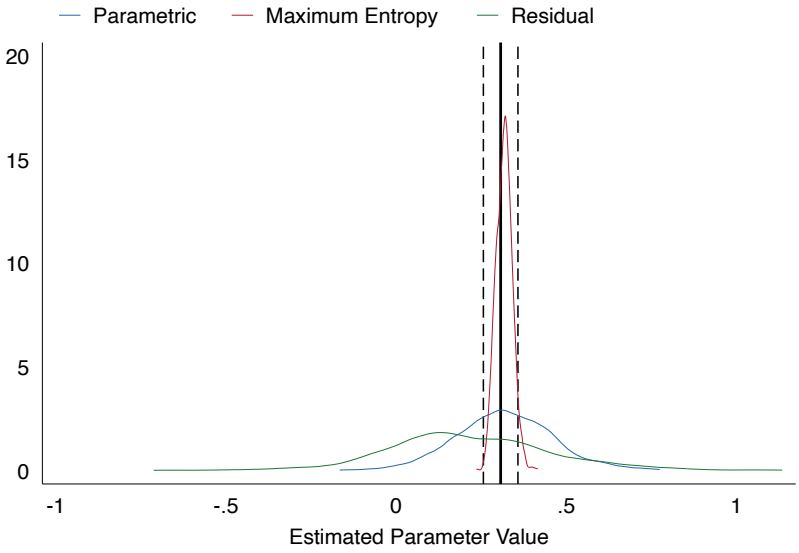


Figure SM. 5: Comparison of bootstrapping techniques, FTSE parameter from Schneider and Troeger's example

Note: Original estimated parameter shown (black solid vertical line) along with associated 95% confidence intervals (black dashed vertical lines).

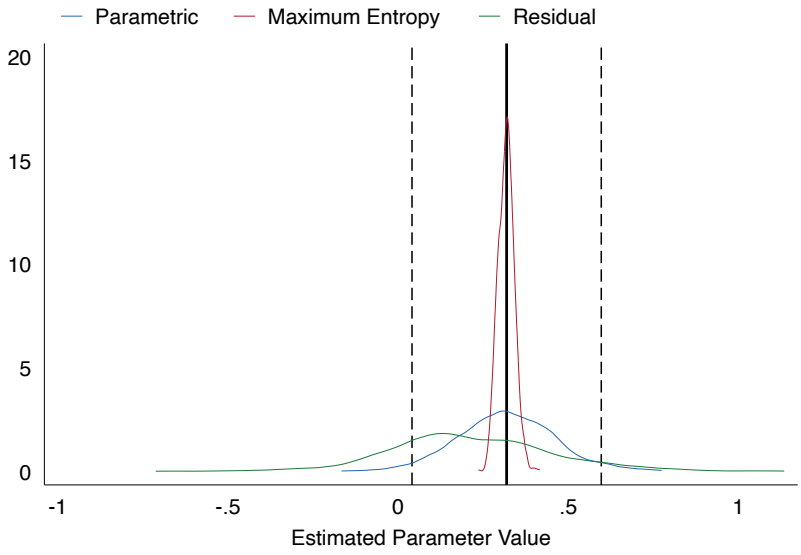


Figure SM. 6: Comparison of bootstrapping techniques, Palestinian-Israeli conflict severity parameter from Schneider and Troeger's example

Note: Original estimated parameter shown (black solid vertical line) along with associated 95% confidence intervals (black dashed vertical lines).

2.2 Monte Carlo Evidence

In order to create a more general set of suggestions on which bootstrapping method might be preferred for most applications, we also created a series of Monte Carlo experiments, using the following data-generating process (DGP):

$$y_t = \beta_0 + \phi y_{t-1} + \beta_x x_t + \beta_z z_t + \sqrt{\sigma^2}\nu_t \tag{1}$$

i.e., an autoregressive process with two regressors and, since $\sqrt{\sigma^2}\nu_t = \varepsilon_t$, a GARCH(1,1) process given as:

$$\sigma_t^2 = \omega_1(\nu_{t-1}^2\sigma_{t-1}^2) + \alpha\sigma_{t-1}^2 + \exp(\beta_0 + \beta_x x_t + \beta_z z_t) \tag{2}$$

where the GARCH(1,1) process itself is also a function of the same regressors in the mean equation. Across each simulation we consider the following:

- $\omega_1 = 0.1$ (the ARCH(1) term)

- $\alpha = 0.5$ (the GARCH(1) term)
- $x_t \sim N(0,5)$, where $\beta_x = 0.5$
- $z_t \sim B(1,0.5)$, where $\beta_z = -1$
- $\beta_0 = 0.5$, the constant appearing in both the conditional mean and conditional variance equation
- $\phi = 0.25$, the autoregressive term in the equation for y_t

We then consider the following four scenarios, which vary both the number of time points as well as the type of error process under consideration:

1. $T = 250$, and $\nu_t \sim N(0,1)$ (Experiment I)
2. $T = 1000$, and $\nu_t \sim N(0,1)$ (Experiment II)
3. $T = 250$, and ν_t are random draws from a Student's t-distribution with two degrees of freedom (Experiment III)
4. $T = 1000$, and ν_t are random draws from a Student's t-distribution with two degrees of freedom (Experiment IV)

The latter two scenarios are designed to evaluate the performance of the parametric bootstrap exactly when we fail to meet the assumption that residuals are normally distributed. It is not a priori clear how this will affect the other bootstrapping techniques. Moreover, by construction all scenarios have time-varying heteroskedasticity in the residuals since $\varepsilon_t = \sqrt{\sigma^2}\nu_t$.

The processes below are identical for each of the scenarios and consists of first doing the following:

1. Generate $M = 500$ series y_t and σ_t^2 using the DGP above.
2. Estimate a GARCH(1,1) model using the same specification as shown in Equations 1 and 2 on each series m. Store these estimated parameters.

3. Across each m, create an expected $\hat{\sigma}_t^2$ value using the estimated parameters from Step 2, setting $z_t = 1, x_t = 1$, assuming no ARCH effects, and using a 100-period burn-in to get a stable expected value. Store this value.

4. Calculate and save the following values from the M $\hat{\sigma}_t^2$ values created in Step 3: median, standard deviation, and upper and lower 95 percentiles. We refer to these values as *truth*, since they represent summary measures of the 500 draws from the underlying DGP. In other words, we will compare all of the bootstrapping predictions to these values.[1]

5. Perform the three bootstrapping procedures. For each procedure, do the following for each of the original m series created back in Step 1:

   (a) Create $B = 500$ bootstrap replications/parameters. Thus, each m will go on to generate 500 bootstrapped values.

   (b) Create expected $\hat{\sigma}_t^2$ for b bootstraps, using the same values from Step 3, and store the following values: median, standard deviation, and upper and lower 95 percentiles.

To summarize, our approach compares the expected conditional variance from 500 draws from the underlying DGP (our "truth" case) to each of our three bootstrapping techniques. Each of the bootstrapping techniques consists of 500 bootstraps for each of the (original 500) series.[2]

We also assess the performance using the delta method to calculate both the estimate value of $\hat{\sigma}_t^2$ as well as the corresponding standard error and 95 percent confidence intervals. The procedure for this is largely the same as above, although there are no bootstrap replicates needed, just a delta method approximation after estimating the GARCH model.

[1] Unlike standard Monte Carlo analyses where we might know the parameter value to assess (e.g., compare estimators A and B to true fixed parameter value β), here we have a conditional expected value, which depends on the GARCH component. Thus, by averaging over $M = 500$ simulations we see both the "on average" value of conditional variance as well as—perhaps even more importantly—the spread of uncertainty around it. This uncertainty is what we want to avoid over- or under-predicting using our bootstrapping techniques.

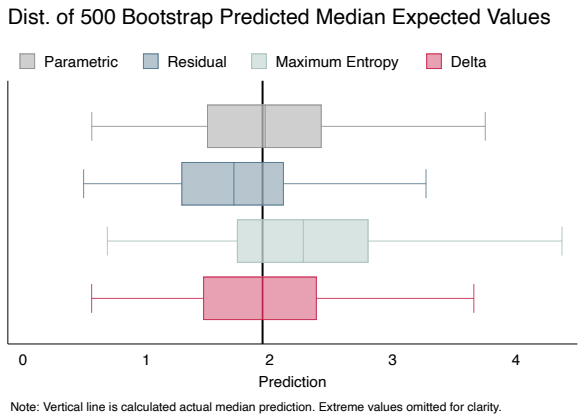
[2] The 500 bootstrapped GARCH models did not always converge, especially in Scenarios 3 and 4.

As a preliminary examination, we show box plots of each technique from Experiment I relative to the actual prediction in Figure SM. 7 for each quantity of interest (the median expected value, standard deviation, upper and lower 95% confidence intervals).[3] For instance the vertical line in Figure SM. 7a shows that the median expected value of $\hat{\sigma}_t^2$ across the 500 Monte Carlo simulations was a little less than two. Each of the box plots shows the distribution of median expected values across the 500 Monte Carlo simulations, where (for the bootstrapping techniques) 500 bootstraps were used to construct the median expected value (the delta method just provided the expected value for each of the 500 Monte Carlo simulations. The delta method and parametric bootstrap also have median predictions centered around two, in contrast to the residual bootstrap (which appears to underpredict) and maximum entropy (overprediction). However, the residual bootstrap appears to have smaller variation across the 500 simulations, which is also preferable.

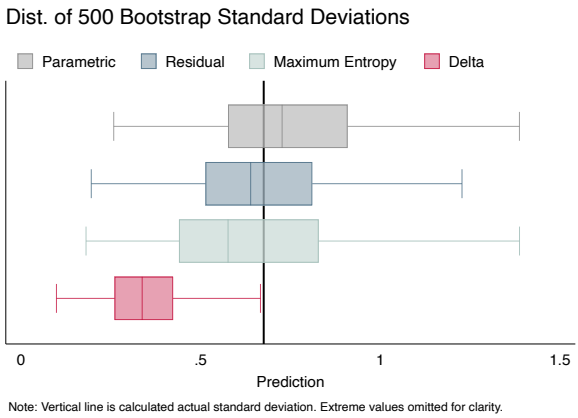
Figure SM.7b depicts the standard deviation around the original 500 GARCH median expected values from the data generating process. Thus, each of our bootstrapping techniques should—ideally—have a similar standard deviation each time we run a GARCH model and take 500 bootstraps.[4] Across the four techniques, the residual bootstrap appears to have the standard deviation that is closest to the correct value, although the other two bootstrapping techniques are close (though have larger variability). In contrast, the delta method appears to systematically underestimate the true standard deviation. Relatedly, in Figures SM.7c and SM.7d we show the upper and lower 95 percent confidence intervals. The residual bootstrap, followed by the parametric, appear to be the best performers, again having confidence intervals that are close the the true 95% confidence intervals from the data generating process. This means that these in applied examples these techniques are likely to better reflect the correct amount of uncertainty when making predictions.

[3]Our simulations sometimes resulted in extremely large values; we omit these for clarity from the box plots although they are included in the RMSE and R(Med)SE calculations in Table SM. 1 below.

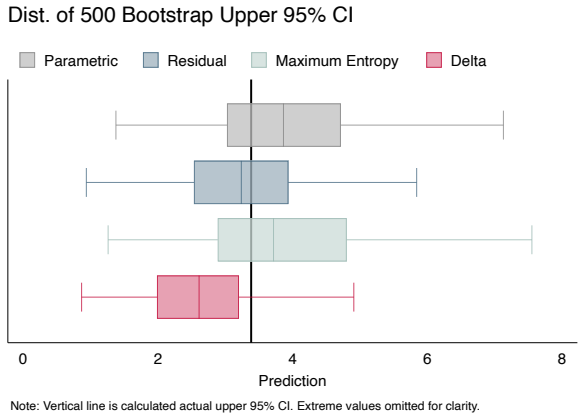
[4]Of course, the delta method is not bootstrapping but instead approximating the standard error, which is shown here.



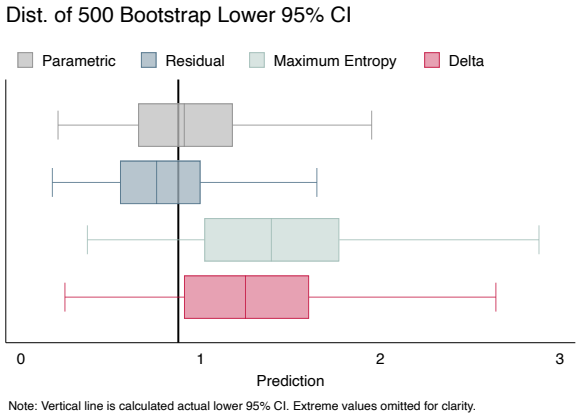
(a) Predicted Values



(b) Standard Deviation



(c) Upper 95% Confidence Interval

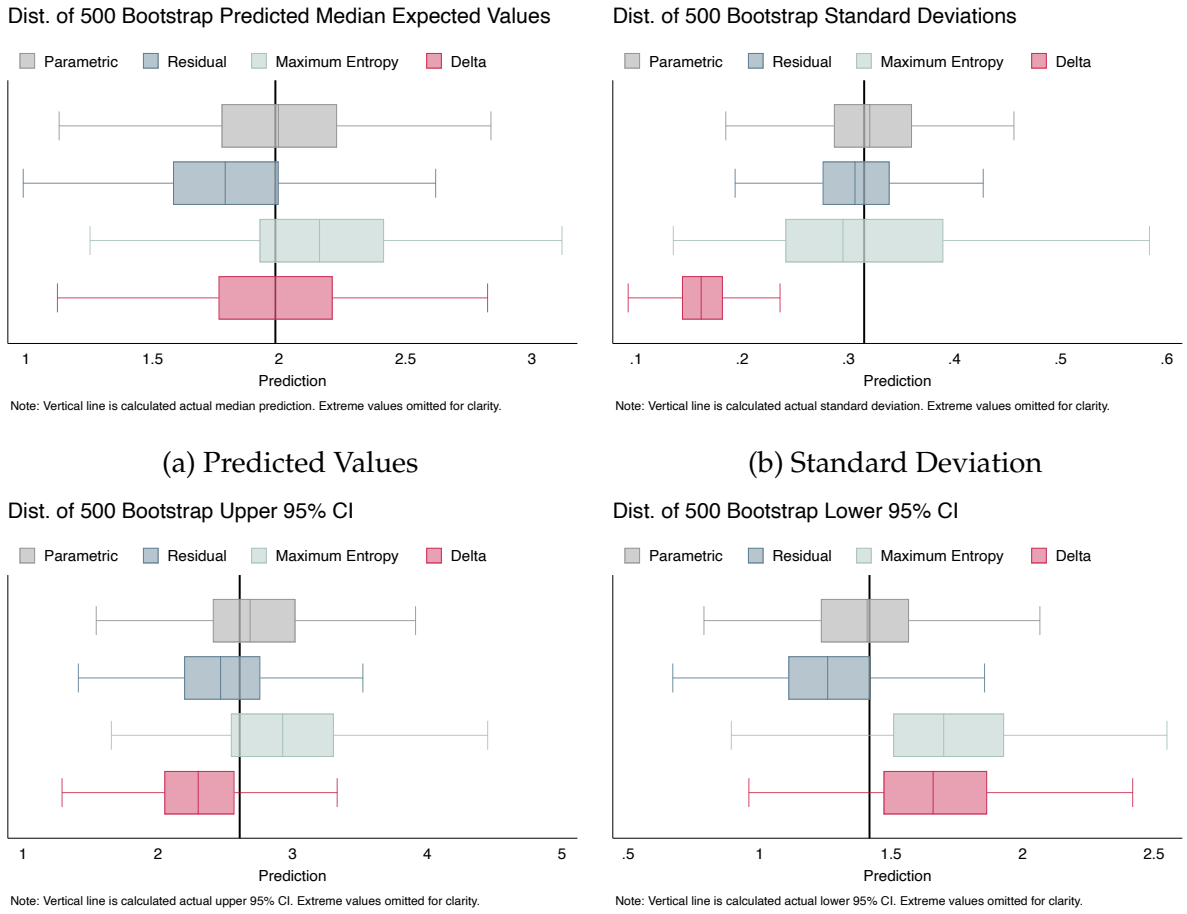


(d) Lower 95% Confidence Interval

Figure SM. 7: Experiment I: $T = 250$, and $\mathrm{v}_t \sim N(0,1)$

Note: Vertical lines show the actual 500 calculated median expected values, standard deviations, upper and lower 95% confidence intervals, from the underlying data-generating process. Each boxplot shows the distribution of median expected values, standard deviations, upper and lower 95% confidence intervals, where each observation was from constructed one draw (and then bootstrapping/delta method) of the underlying data-generating process. Extreme values omitted for clarity.

In Figure SM.8 we show the same box plots, but for the second experiment, where we increase T to 1000. In this scenario, the best performer in terms of median expected values appears to be the parametric bootstrap and delta method, although the residual bootstrap has smaller variance in standard errors. Upper and lower confidence intervals are perhaps most correctly estimated by the parametric bootstrap, although the residual bootstrap looks like the next best performer.



(a) Predicted Values

(b) Standard Deviation

(c) Upper 95% Confidence Interval

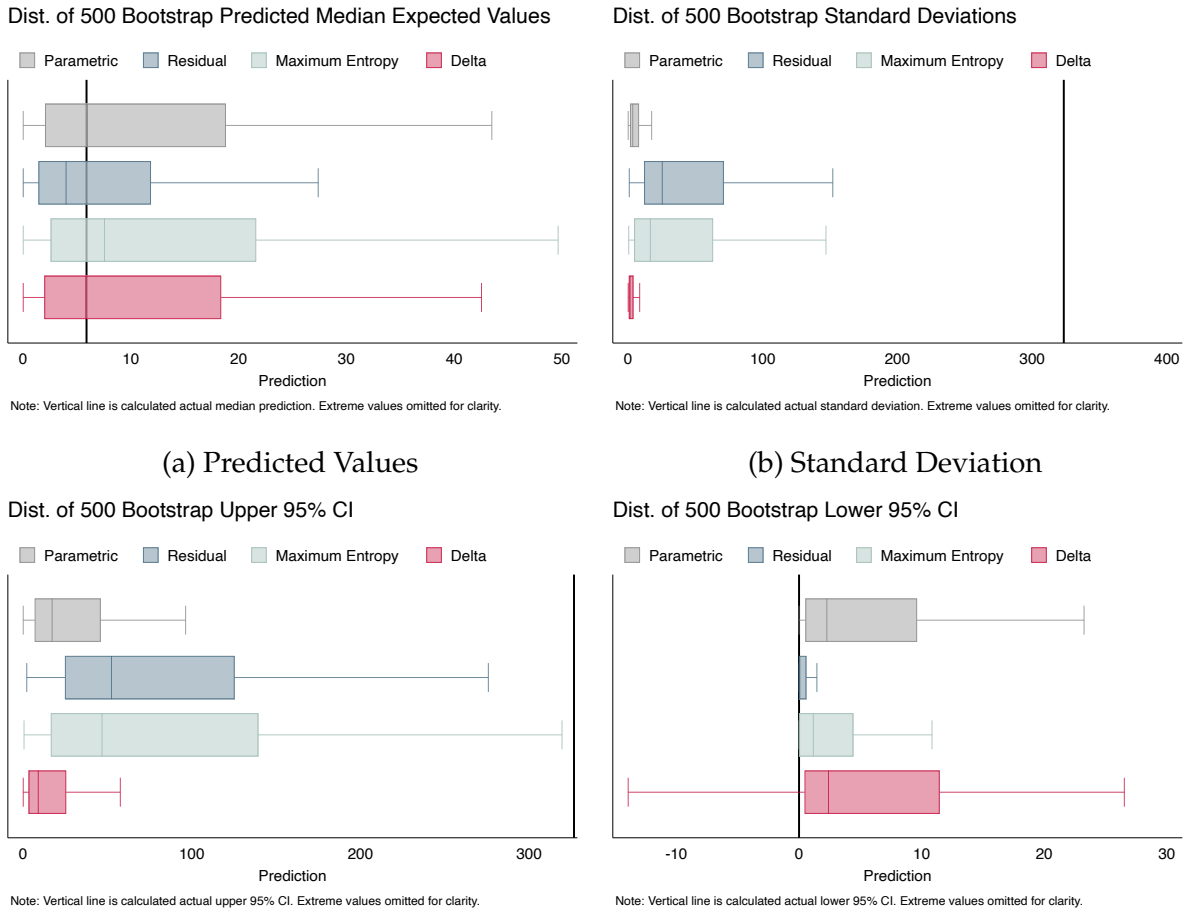
(d) Lower 95% Confidence Interval

Figure SM. 8: Experiment II: $T = 1000$, and $\nu_t \sim N(0, 1)$

Note: Vertical lines show the actual 500 calculated median expected values, standard deviations, upper and lower 95% confidence intervals, from the underlying data-generating process. Each boxplot shows the distribution of median expected values, standard deviations, upper and lower 95% confidence intervals, where each observation was from constructed one draw (and then bootstrapping/delta method) of the underlying data-generating process. Extreme values omitted for clarity.

In Figures SM.9 ($T = 250$) and SM.10 ($T = 1000$) we change the distribution of ν_t—part of the error component—from a standard normal to a Student-t distribution with two degrees of freedom, which should produce much heavier tails than in the standard

normal. Here it is clear our techniques struggle much more, especially with constructing measures of uncertainty; estimated standard deviations tend to be much smaller than those in the underlying data generating process. Across both experiments, the upper confidence level tends to be underestimated (much too small) while the lower confidence interval is slightly overestimated (although this looks least bad for the residual bootstrap). In terms of expected values, the residual bootstrap has the smallest variation although median predictions of both the parametric and delta method are correctly centered on the underlying true median.



(a) Predicted Values

(b) Standard Deviation

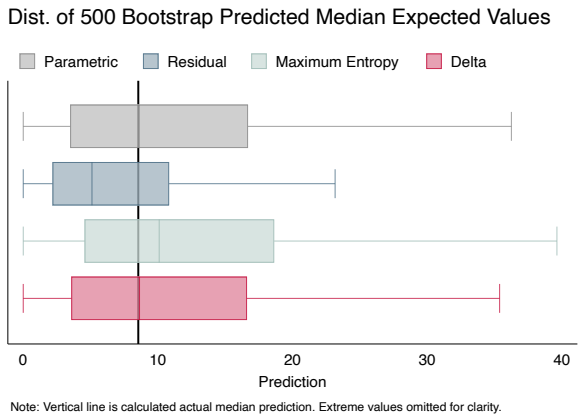
(c) Upper 95% Confidence Interval

(d) Lower 95% Confidence Interval

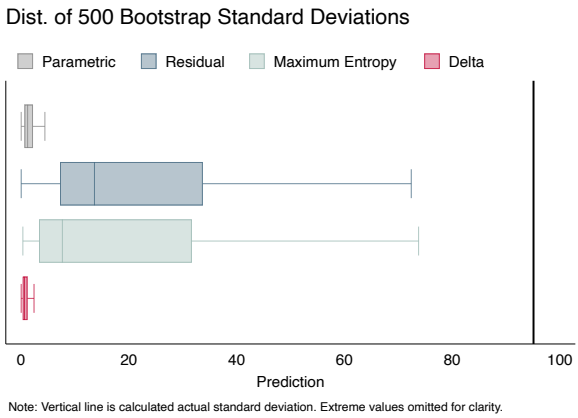
Figure SM. 9: Experiment III: $T = 250$, and $\nu_t \sim t(2)$

Note: Vertical lines show the actual 500 calculated median expected values, standard deviations, upper and lower 95% confidence intervals, from the underlying data-generating process. Each boxplot shows the distribution of median expected values, standard deviations, upper and lower 95% confidence intervals, where each observation was from constructed one draw (and then bootstrapping/delta method) of the underlying data-generating process. Extreme values omitted for clarity.

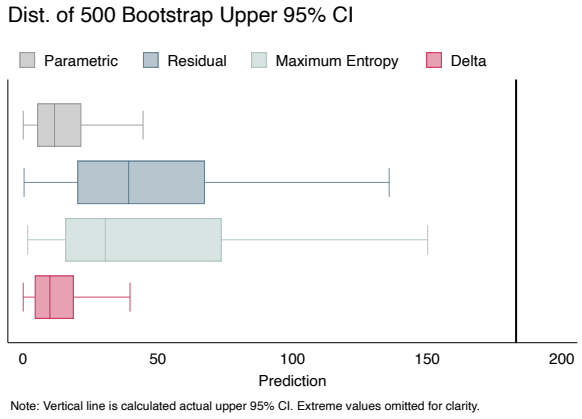
In Table SM. 1, we summarize these same findings for each bootstrapping tech-



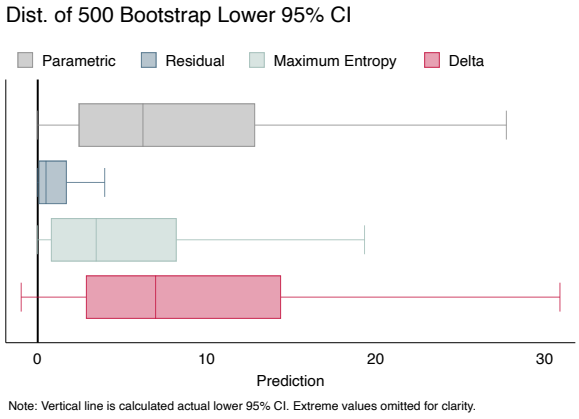
(a) Predicted Values



(b) Standard Deviation



(c) Upper 95% Confidence Interval



(d) Lower 95% Confidence Interval

Figure SM. 10: Experiment IV: $T = 1000$, and $\nu_t \sim t(2)$

Note: Vertical lines show the actual 500 calculated median expected values, standard deviations, upper and lower 95% confidence intervals, from the underlying data-generating process. Each boxplot shows the distribution of median expected values, standard deviations, upper and lower 95% confidence intervals, where each observation was from constructed one draw (and then bootstrapping/delta method) of the underlying data-generating process. Extreme values omitted for clarity.

nique using two quantities of interest in order to provide more clear suggestions about the best performing technique across quantities of interest as well as the different experiments. Root mean squared error (RMSE) measures both bias and efficiency of a quantity of interest, and is calculated as (Hopkins et al. 2024):

$$\text{RMSE}[\hat{\theta}] = \sqrt{\frac{1}{M}\sum_{m=1}^{M}[(\hat{\theta}_m - \theta)^2]} \tag{3}$$

Where θ is the quantity originally obtained in Step 4 (for example, the standard error or upper 95% confidence interval calculated from the expected values simulated from the original 500 GARCH models), and $\hat{\theta}_m$ is the quantity of interest obtained from B bootstraps for a single one of the m series. Lower values are more preferred; for instance, if looking at the RMSE of the upper 95% confidence interval, an RMSE of zero for a bootstrapping technique would suggest that the bootstrapping technique exactly recovers the correct upper 95% confidence interval (i.e., not too small or too large).

In addition to RMSE, we also calculate root *median* square error, or R(Med)SE, which is useful when we suspect the quantity of interest is not normally distributed (Hopkins et al. 2024); this is likely the case for our experiments, since quantities like standard deviation or the GARCH expected values tend to be right skewed since they cannot be zero. R(Med)SE should be less sensitive to outliers as well (again, these are likely to be very large positive values).

The columns of Table SM. 1 display RMSE and R(Med)SE for each bootstrapping technique, while the rows are the different quantities of interest from each bootstrapping procedure (median prediction, standard deviation, upper and lower 95% confidence intervals) for each experiment. We highlight the "best performer" bootstrapping technique for each quantity in each experiment, either by **bolding** (for RMSE) or *italicizing* (for R(Med)SE) the smallest value. Table SM. 1 has several important findings worth noting:

- In terms of RMSE, the residual bootstrap is often the best performer, and, if using

Table SM. 1: Monte Carlo Results

	Parametric		Delta		Residual		Maximum Entropy	
	RMSE	R(Med)SE	RMSE	R(Med)SE	RMSE	R(Med)SE	RMSE	R(Med)SE
Experiment I: $T = 250$, and $\nu_t \sim N(0,1)$								
Prediction	0.683	*0.460*	0.671	0.466	**0.639**	0.475	0.860	0.533
Std. Dev.	**0.260**	0.159	0.347	0.339	0.322	*0.150*	5.518	0.221
Upper 95% CI	1.342	0.829	1.125	0.861	**1.04**	*0.735*	21.861	0.946
Lower 95% CI	0.401	0.272	0.664	0.401	**0.345**	*0.256*	0.799	0.517
Experiment II: $T = 1000$, and $\nu_t \sim N(0,1)$								
Prediction	0.316	0.226	**0.314**	*0.225*	0.341	0.246	0.395	0.261
Std. Dev.	0.050	0.035	0.154	0.154	**0.045**	*0.033*	0.313	0.076
Upper 95% CI	0.416	0.285	0.460	0.337	**0.394**	*0.276*	0.921	0.407
Lower 95% CI	**0.245**	*0.169*	0.378	0.256	0.266	0.191	0.421	0.289
Experiment III: $T = 250$, and $\nu_t \sim t(2)$								
Prediction	341.373	4.786	340.848	4.863	**115.735**	*4.546*	515.377	5.179
Std. Dev.	7.498e16	320.226	**318.883**	321.985	1.291e7	*304.082*	691.493	309.729
Upper 95% CI	6.809e9	311.272	**463.886**	317.919	3.865e5	*281.079*	2087.655	295.022
Lower 95% CI	291.61	2.402	299.063	2.846	**1.748**	*0.060*	34.651	1.214
Experiment IV: $T = 1000$, and $\nu_t \sim t(2)$								
Prediction	92.086	5.763	90.618	*5.610*	55.918	5.625	**55.568**	5.792
Std. Dev.	**93.48**	93.914	93.938	94.504	8.711e8	*83.764*	144.317	88.642
Upper 95% CI	**180.418**	172.148	181.261	173.852	304.507	*149.189*	319.292	156.083
Lower 95% CI	87.994	6.146	87.946	6.935	**5.421**	*0.559*	16.374	3.589

Note: Root mean squared error (RMSE) and root median square error (R(Med)SE) shown for each bootstrapping procedure across each quantity of interest and each experiment. Best performer for each experiment highlighted in **bold** (for RMSE) and *italics* (for R(Med)SE.

R(Med)SE, is nearly always the best (in terms of minimizing RMSE/R(Med)SE). Some important exceptions to this are the uncertainty calculations (SD and confidence intervals) when ε_t is not normally distributed, although this appears to improve as T increases.

- The maximum entropy technique is always outperformed by some other bootstrapping technique for every scenario across all quantities of interest. Given the length of time it takes to compute the maximum entropy bootstrap GARCH models, it therefore seems wiser to either use parametric or delta—and thus save substantially on computing time—or wait slightly longer for the residual bootstrap results which tend to perform best.

- All bootstrapping techniques perform well when ε_t is normally distributed, and improve as T grows longer. If computing time is not an issue, using the residual based bootstrap seems advisable, although the quick methods (parametric, delta) can be used without major drops in performance. However, the poor performance for virtually all bootstrapping procedures when calculating uncertainty measures when ε_t is t-distributed indicates that users should proceed with caution if this part of the error is not normal. However, it is easy to check for normality using common tests such as Shapiro-Wilk or Shapiro-Francia, or examining quantile plots.[5]

- The delta method and parametric bootstrap are nearly identical in terms of performance. Still, one would likely prefer the latter, since, if using the percentile method to construct confidence intervals, by construction it cannot lead to negative (and thus, nonsensical) values, in contrast with the former, which will never result in a negative prediction, but could easily construct a negative lower 95% confidence interval. For instance, see the negative lower 95% confidence intervals that sometimes occurred across our simulations with the delta method in Figure SM.10d.

[5]Recall that the total residual component is $\sqrt{\sigma^2}\nu_t$, so users wishing to test for normality could also divide the residual by σ_t to isolate ν and proceed to test the latter.

- Outliers appear to be a clear issue in Experiments III and IV, as evidenced by the extremely high RMSE (and slightly lower R(Med)SE) values. While this largely does not affect the prediction, it does affect the estimated measures of uncertainty. This might be mitigated somewhat in real-world applications by increasing the number of bootstrap replications (the Monte Carlos here only used 500), placing better defined starting/priming values when estimating the GARCH models (we simply used Stata's default in our experiments), or perhaps throwing out bootstraps that create non-sensical results (e.g., a GARCH model that technically converges and provides estimates but does not show standard errors because they were estimated to be infinite could be dropped). Still, as discussed above we encourage careful model specification in order to ensure predicted residual components are approximately normal.

3 Additional Results

In this section we present several additional results.

3.1 Hellwig

- In Figure SM.11 we replicate the +4 point trade shock and its effect on the conditional error variance of governing party support in the UK (the Hellwig example shown in Figure 2 in the main manuscript), but now do not perform the rescaling procedure as done in the main manuscript. In other words, the vertical axis shows the actual value of the conditional error variance.

- In Figure SM.12 we show the same trade shock as in Figure SM.11 but now use our standard deviation approach to calculating confidence intervals.

- In Figure SM.13 we present an alternative plotting strategy where we show both the "short-run" change between the expected conditional error variance in the period in which the counterfactual shock occurs, and the expected conditional error variance in the period just before the shock. We also include the "long-run" change between the expected conditional error variance in the final simulation period and the expected conditional error variance in the period just before the shock (i.e., after 9 periods). Such an approach has been used before in political science in order to show "cumulative" or "total" effects after a certain amount of time has elapsed (Breunig and Busemeyer 2012; Adolph, Breunig and Koski 2020).[6] This alternative strategy has the advantage of being able to easily compare both short- and long-run changes in volatility as a consequence of a shock; the former describes immediate movements in conditional error variance in response to a change in a covariate, while the latter shows the total, permanent shift that has occurred as a result of the shock after a longer period of time. Moreover, it can be performed for both the bootstrapping and delta method techniques, and

[6] Neither of these articles examine contemporaneous effects in addition to a longer-period change, as we do.

more easily allows us to discern whether effects are statistically significantly different from the pre-shock expected conditional error variance.

- In Figure SM.14 we replicate the -4 point trade shock and its effect on the conditional error variance of governing party support in the UK (Hellwig example from Figure 3 in the main manuscript), but do not rescale the conditional error variance.
- In Figure SM.15 we show the negative trade shock, but now use our standard deviation approach to calculating confidence intervals.
- In Figure SM.16 we replicate our short- and long-run plotting strategy, but now for the -4 point trade shock.

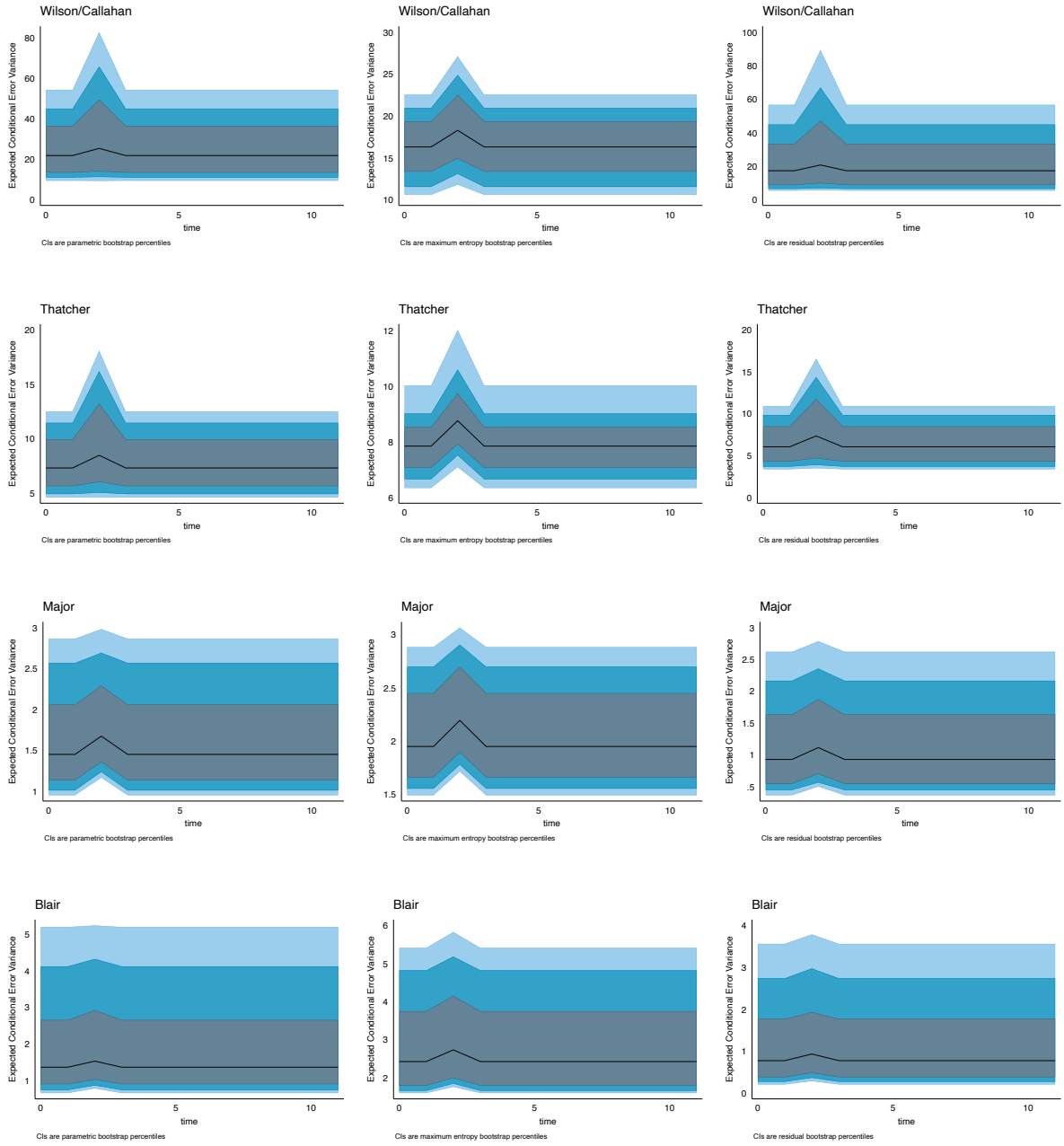


Figure SM. 11: Replication of Figure 2 (in main manuscript), no rescaling

Note: For each prime minister, all other prime minister dummy variables set to 0. From left to right: parametric bootstrap, maximum entropy bootstrap, residual bootstrap. Black line shows median expected conditional error variance. Grey: 75% confidence interval, medium blue: 90% confidence interval, light blue: 95% confidence interval.

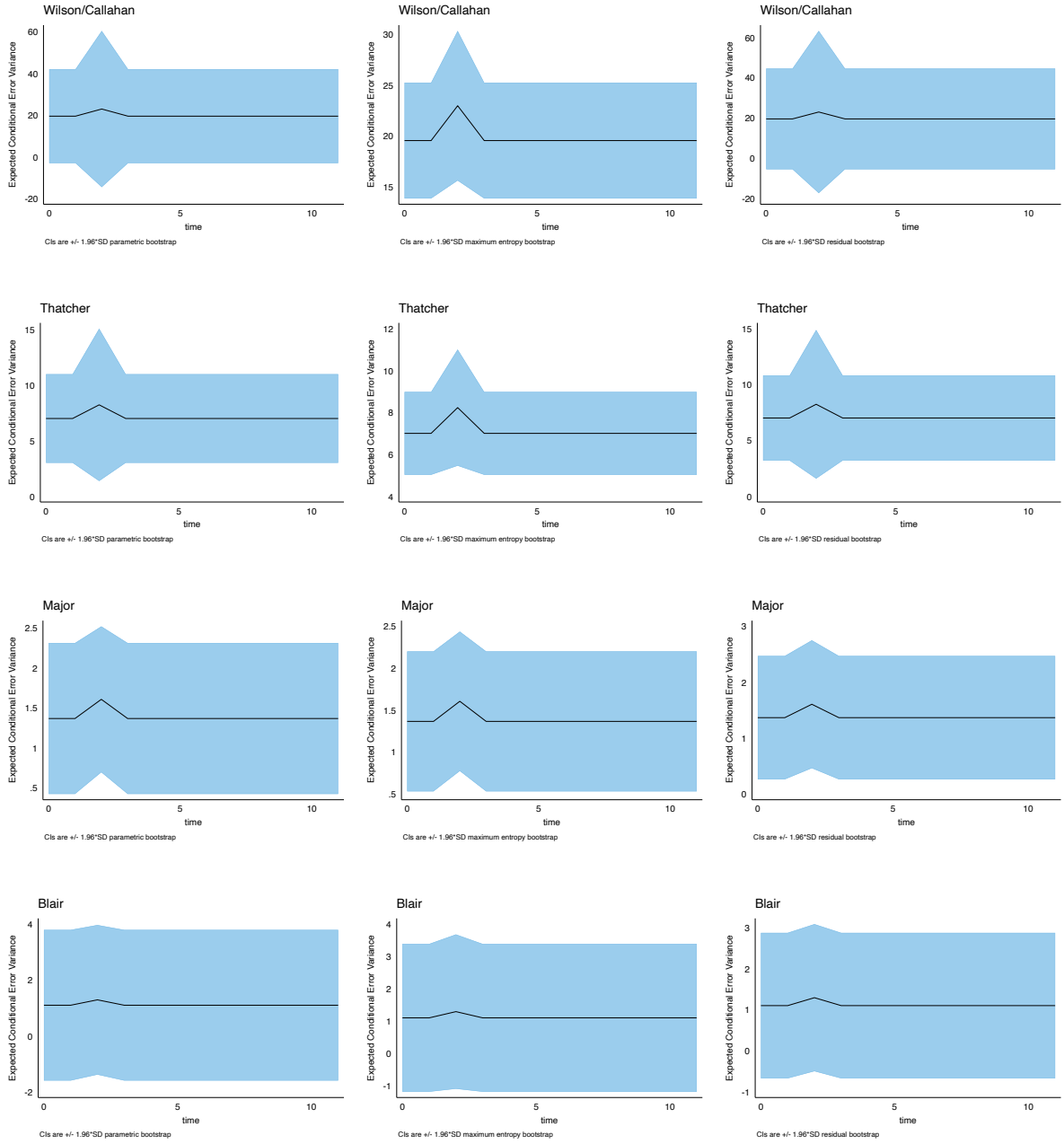


Figure SM. 12: Replication of Figure 2 (in main manuscript), using the standard deviation approach to confidence intervals

Note: For each prime minister, all other prime minister dummy variables set to 0. From left to right: parametric bootstrap, maximum entropy bootstrap, residual bootstrap. Black line shows expected conditional error variance, with 95% confidence intervals calculated using the standard error approach shown.

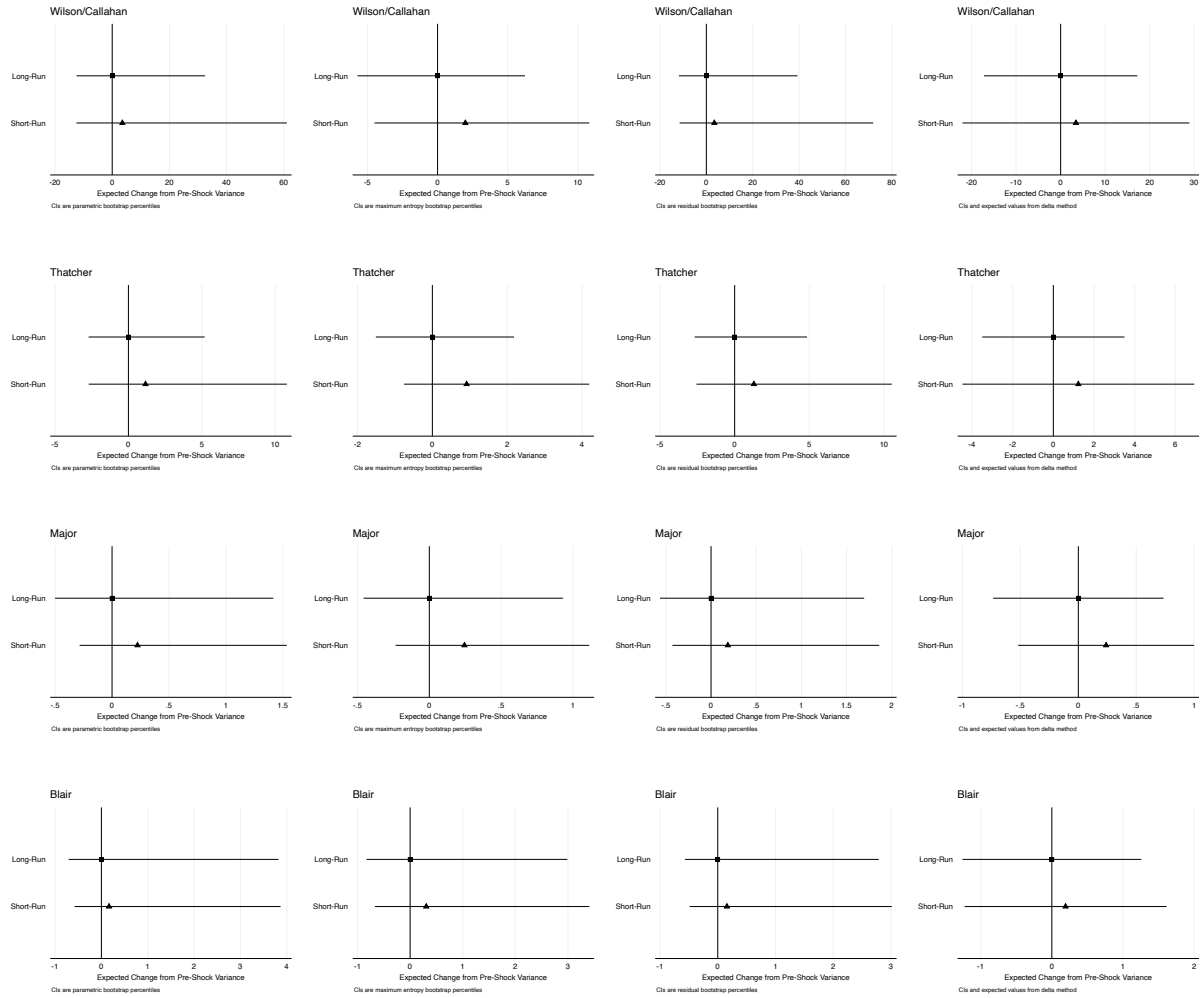


Figure SM. 13: Replication of Figure 2 (in main manuscript) but showing only the short- and long-run effect

Note: For each prime minister, all other prime minister dummy variables set to 0. From left to right: parametric bootstrap, maximum entropy bootstrap, residual bootstrap, delta method. Black line shows expected short-run (contemporaneous shock period) and long run (after 9 periods) change in conditional error variance from the expected conditional error variance in the pre-shock period, with 95% confidence intervals calculated using the percentile approach.

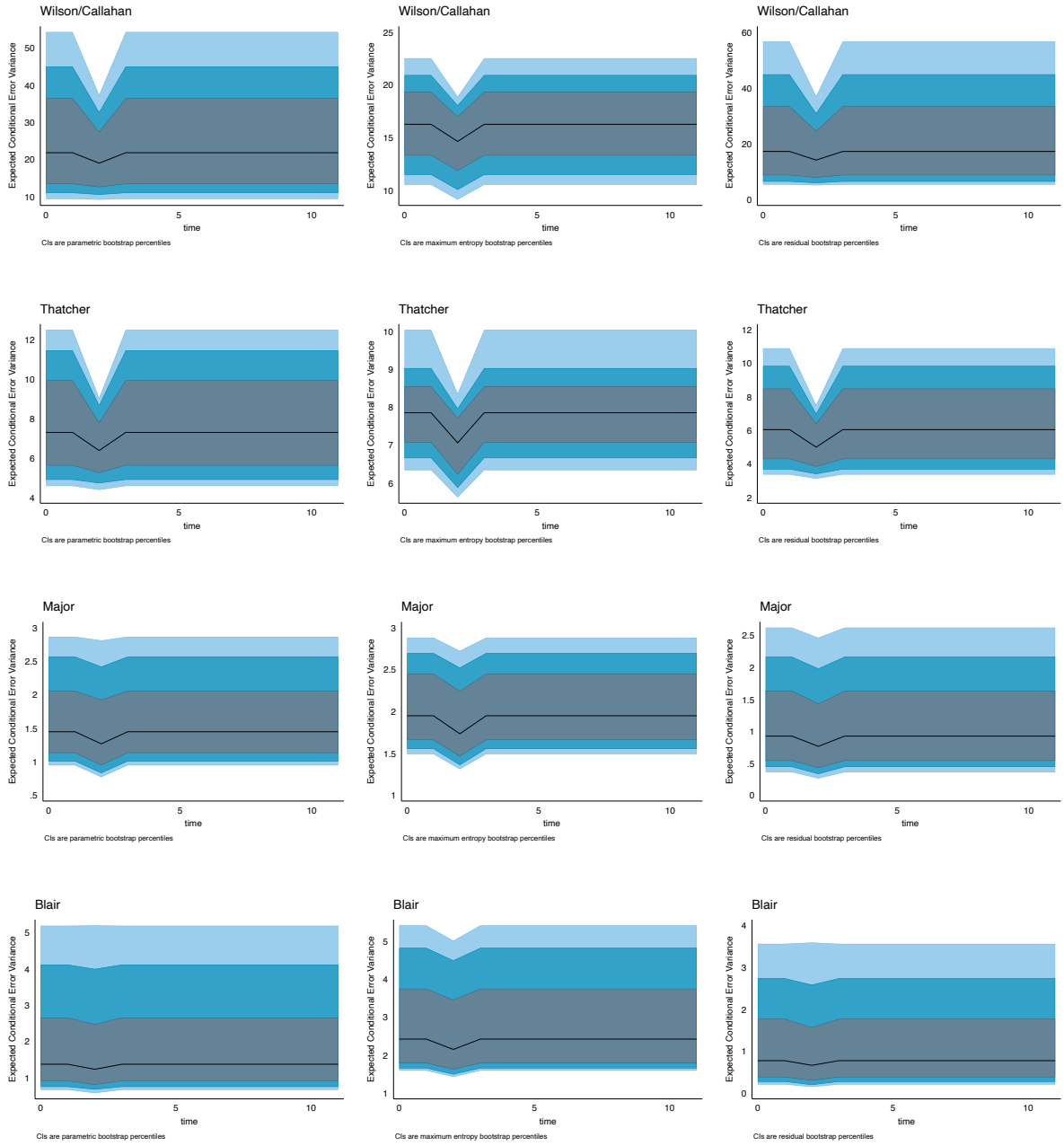


Figure SM. 14: Replication of Figure 3 (in main manuscript), no rescaling

Note: For each prime minister, all other prime minister dummy variables set to 0. From left to right: parametric bootstrap, maximum entropy bootstrap, residual bootstrap. Black line shows median expected conditional error variance. Grey: 75% confidence interval, medium blue: 90% confidence interval, light blue: 95% confidence interval.

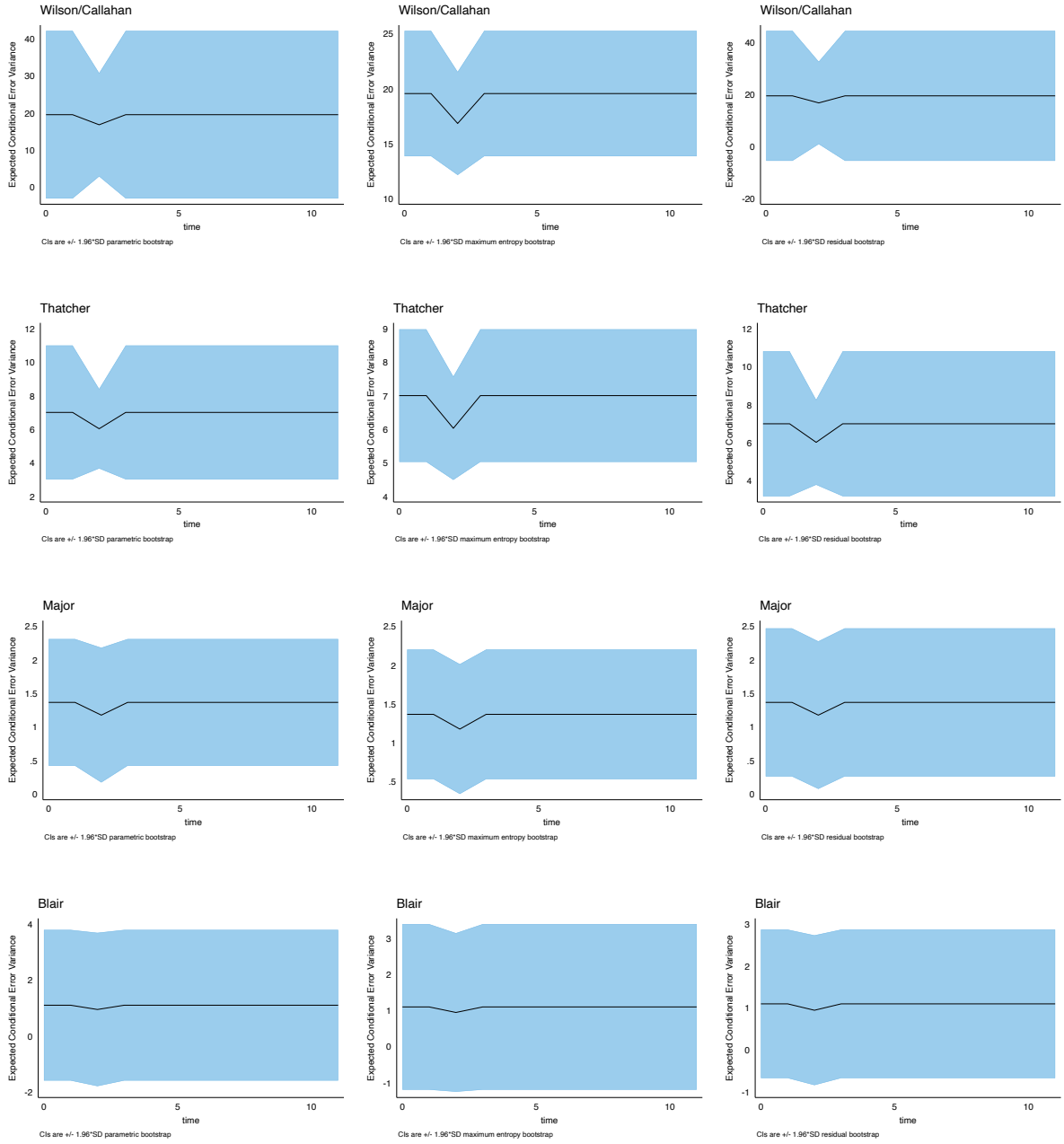


Figure SM. 15: Replication of Figure 3 (in main manuscript), using the standard deviation approach to confidence intervals

Note: For each prime minister, all other prime minister dummy variables set to 0. From left to right: parametric bootstrap, maximum entropy bootstrap, residual bootstrap. Black line shows expected conditional error variance with 95% confidence intervals shown calculated using the standard deviation approach.

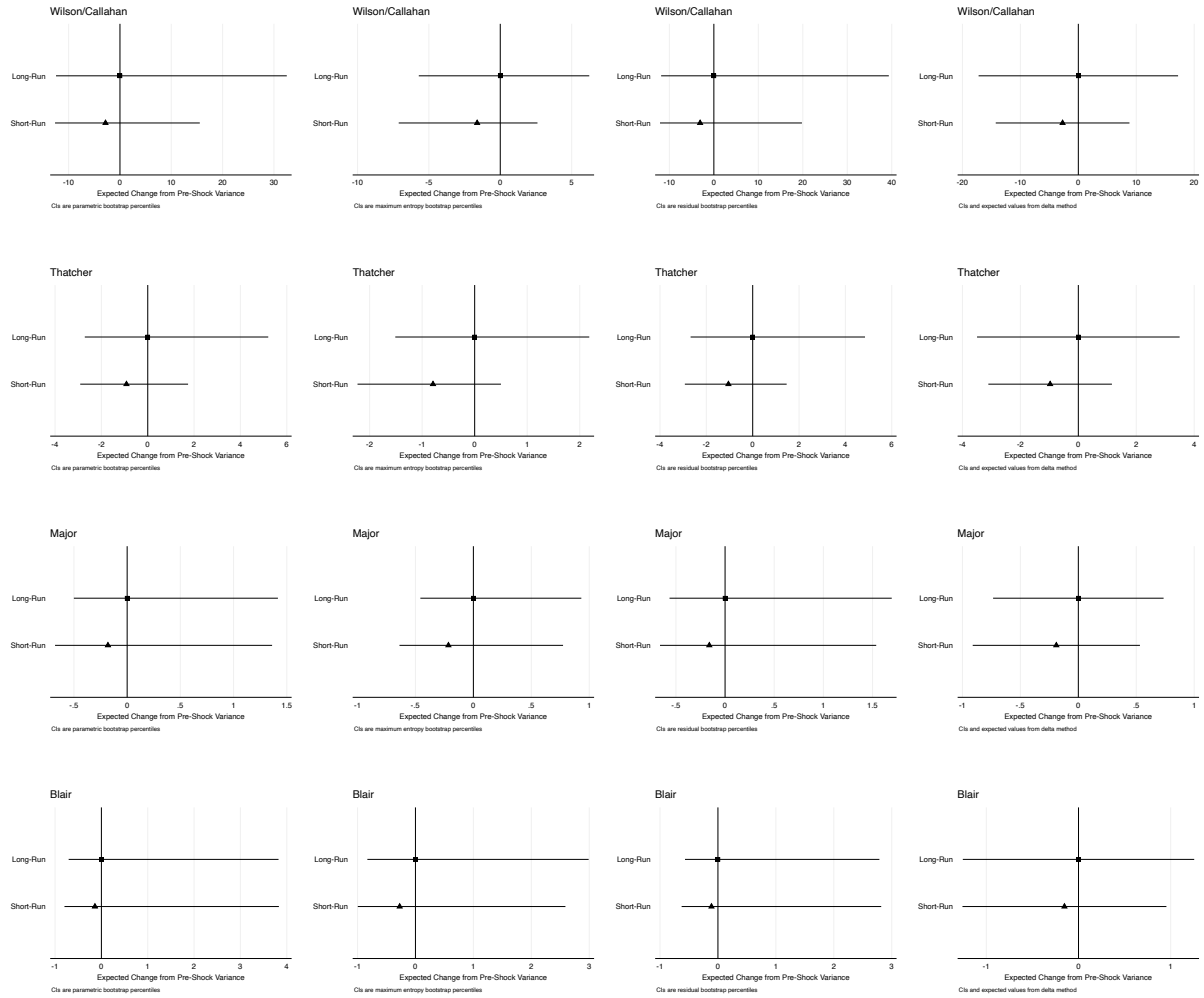
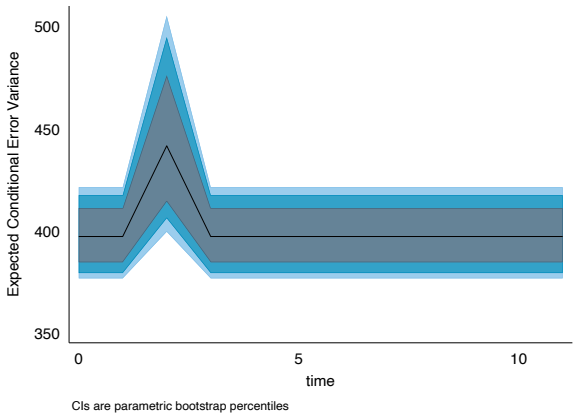


Figure SM. 16: Replication of Figure 3 (in main manuscript) but showing only the short- and long-run effect

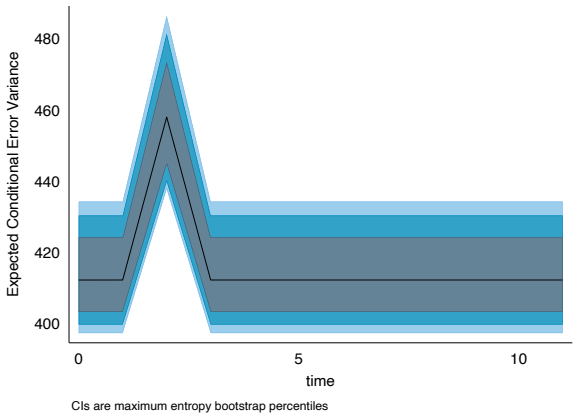
Note: For each prime minister, all other prime minister dummy variables set to 0. From left to right: parametric bootstrap, maximum entropy bootstrap, residual bootstrap, delta method. Black line shows expected short-run (contemporaneous shock period) and long run (after 9 periods) change in conditional error variance from the expected conditional error variance in the pre-shock period, with 95% confidence intervals calculated using the percentile approach.

3.2 Schneider-Troeger

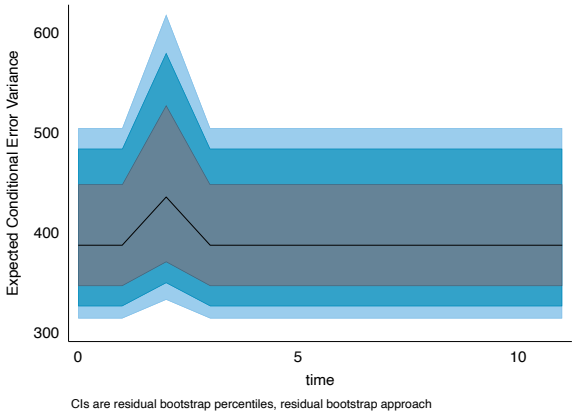
- In Figure SM.17 we recreate Figure 4 from the main manuscript, which shows the effect of one day of Palestinian/Israeli conflict severity, but do not use the rescaling method.
- In Figure SM.18 we replicated Figure 4 from the main manuscript but now use our standard deviation approach to calculating confidence intervals.
- In Figure SM.19 we show the alternative short- and long-run changes plotting approach, replicating Figure 4 from the main manuscript.
- In Figure SM.20 we recreate Figure 5 from the main manuscript, but do not perform the rescaling procedure.
- In Figure SM.21 we replicate Figure 5 but use the standard deviation approach to calculating confidence intervals.
- Last, in Figure SM.22 we replicate Figure 5, but instead depict it as short- and long-run changes.



(a) Parametric bootstrap



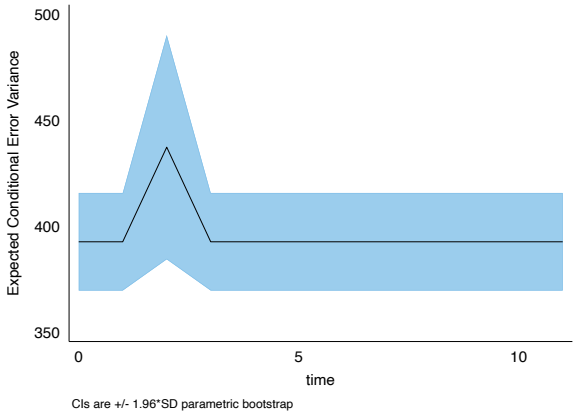
(b) Maximum entropy bootstrap



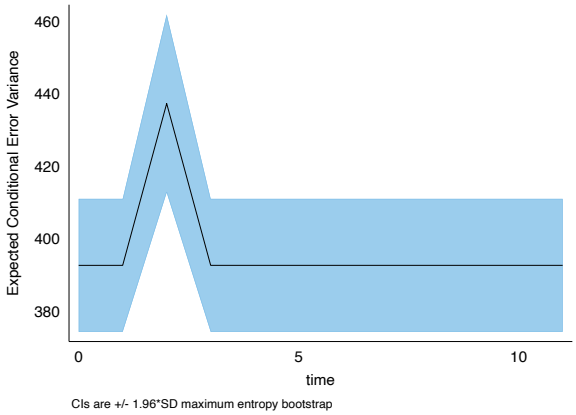
(c) Residual bootstrap

Figure SM. 17: Replication of Figure 4 (in main manuscript), without rescaling

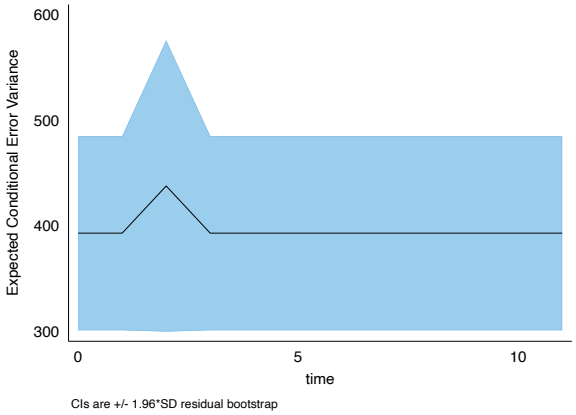
Note: Black line shows median expected conditional error variance. Grey: 75% confidence interval, medium blue: 90% confidence interval, light blue: 95% confidence interval.



(a) Parametric bootstrap



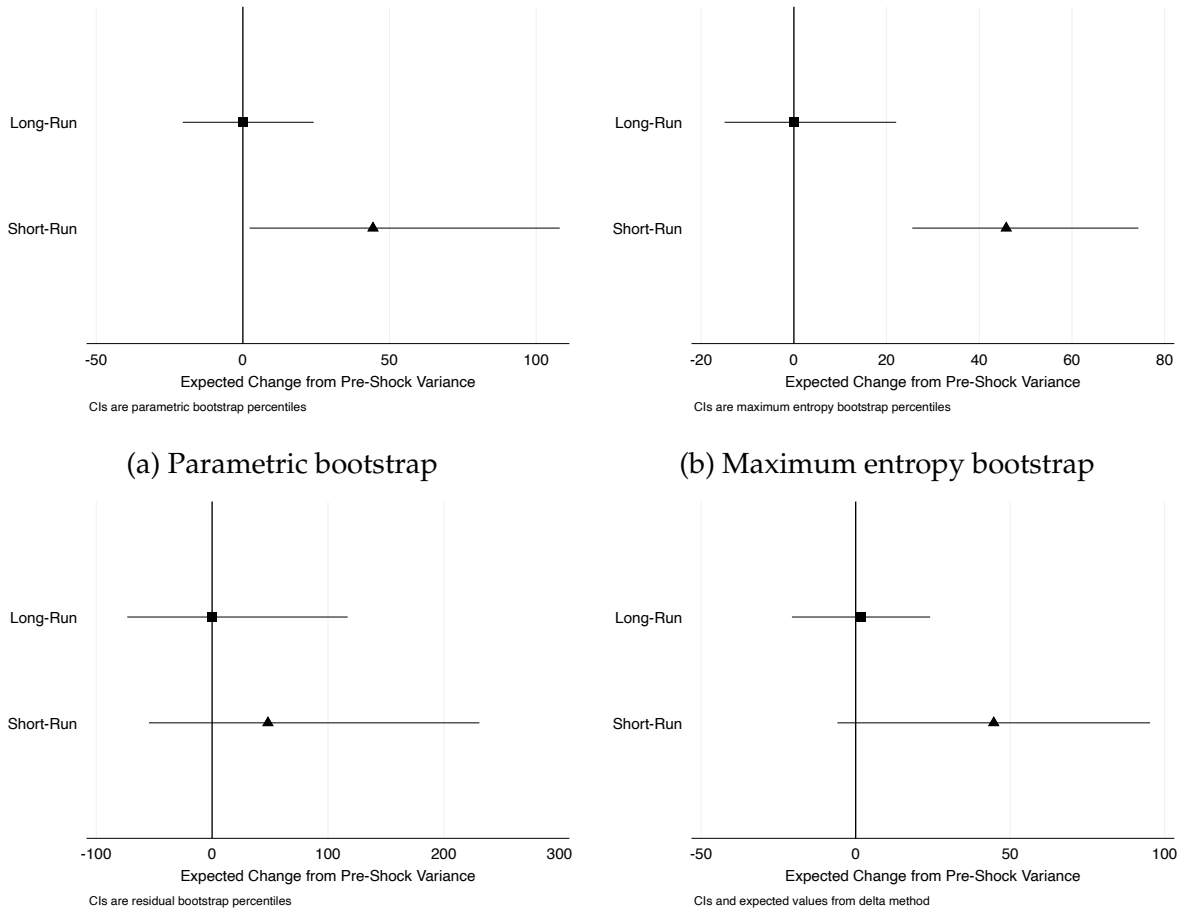
(b) Maximum entropy bootstrap



(c) Residual bootstrap

Figure SM. 18: Replication of Figure 4 (in main manuscript), using the standard deviation approach to confidence intervals

Note: Black line shows median expected conditional error variance with 95% confidence intervals calculated using the standard deviation approach.



(a) Parametric bootstrap

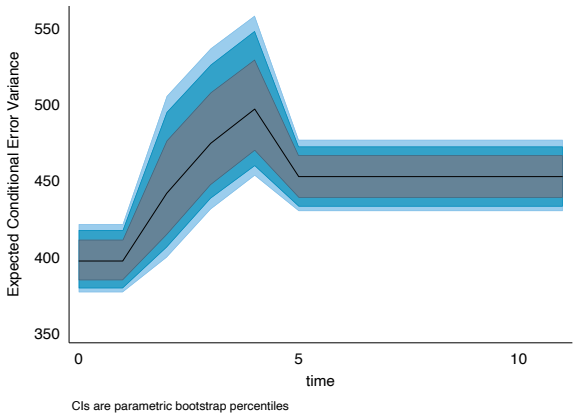
(b) Maximum entropy bootstrap

(c) Residual bootstrap

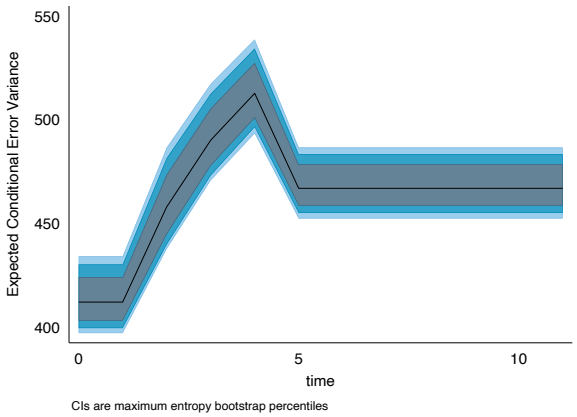
(d) Delta method

Figure SM. 19: Replication of Figure 4 (in main manuscript), but showing only the short- and long-run effect

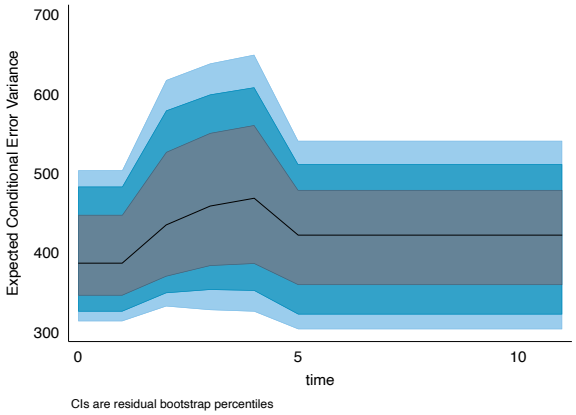
Note: Black line shows expected short-run (contemporaneous shock period) and long run (after 9 periods) change in conditional error variance from the expected conditional error variance in the pre-shock period, with 95% confidence intervals calculated using the percentile approach.



(a) Parametric bootstrap



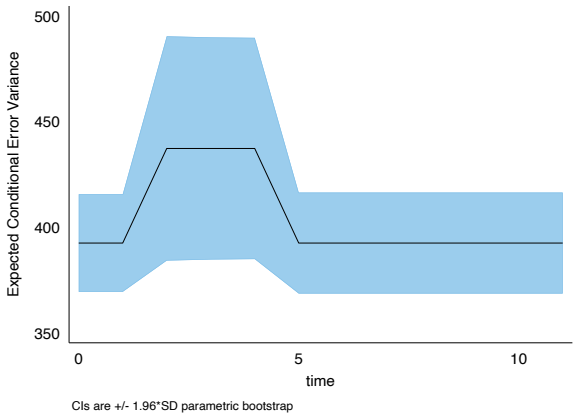
(b) Maximum entropy bootstrap



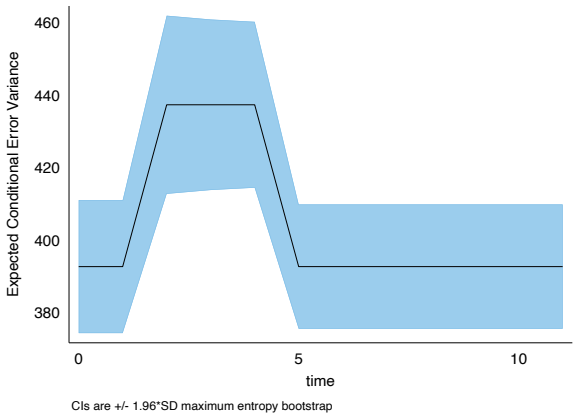
(c) Residual bootstrap

Figure SM. 20: Replication of Figure 5 (in main manuscript), without rescaling

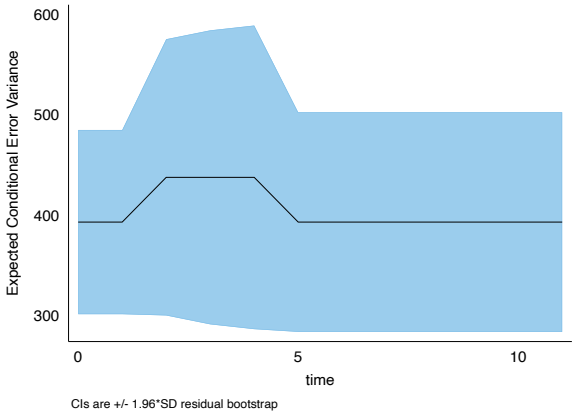
Note: Black line shows median expected conditional error variance. Grey: 75% confidence interval, medium blue: 90% confidence interval, light blue: 95% confidence interval.



(a) Parametric bootstrap



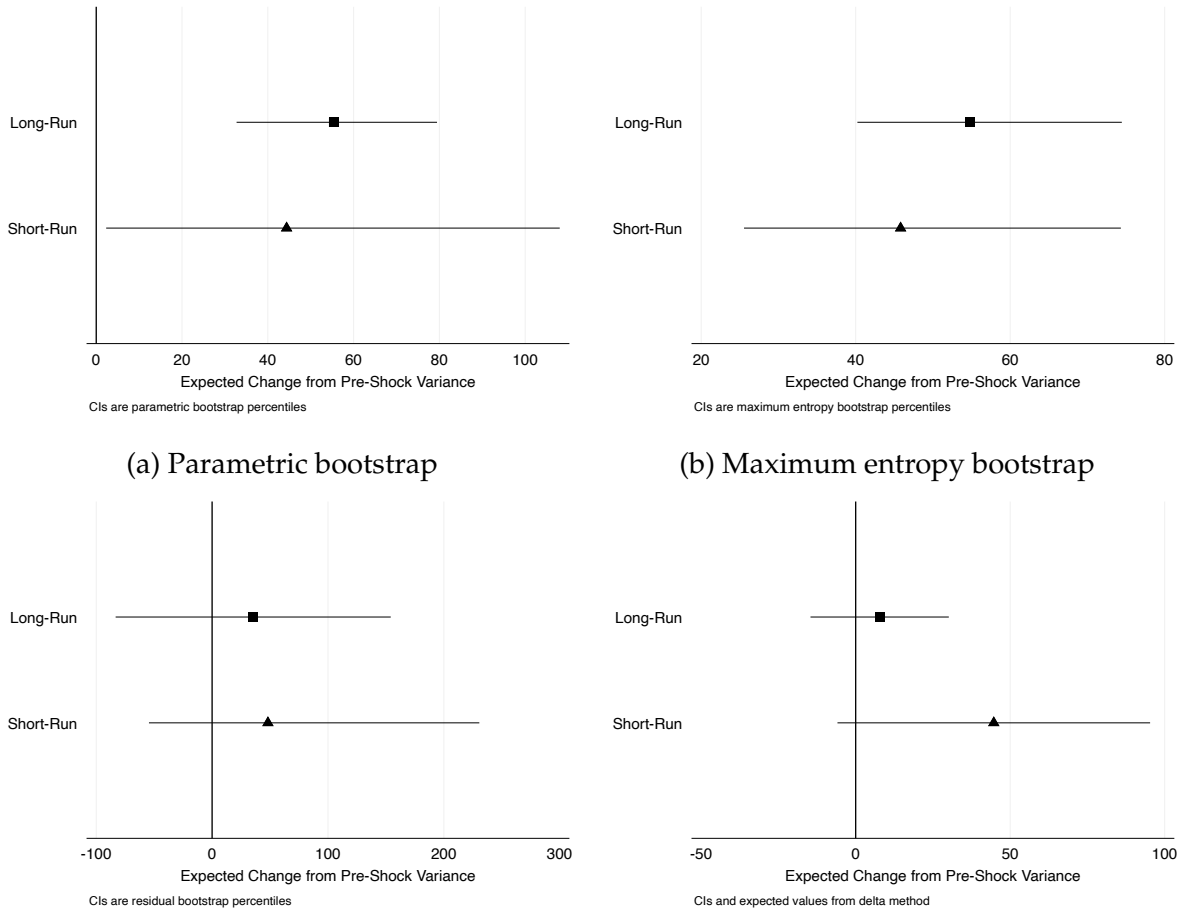
(b) Maximum entropy bootstrap



(c) Residual bootstrap

Figure SM. 21: Replication of Figure 5 (in main manuscript), without rescaling

Note: Black line shows expected conditional error variance with 95% confidence intervals calculated using the standard deviation approach.



(a) Parametric bootstrap

(b) Maximum entropy bootstrap

(c) Residual bootstrap

(d) Delta method

Figure SM. 22: Replication of Figure 5 (in main manuscript), but showing only the short- and long-run effect

Note: Black line shows expected short-run (contemporaneous shock period) and long run (after 9 periods) change in conditional error variance from the expected conditional error variance in the pre-shock period, with 95% confidence intervals calculated using the percentile approach.

References

Adolph, Christopher, Christian Breunig and Chris Koski. 2020. "The political economy of budget trade-offs." *Journal of Public Policy* 40(1):25–50.

Breunig, Christian and Marius R Busemeyer. 2012. "Fiscal austerity and the trade-off between public investment and social spending." *Journal of European Public Policy* 19(6):921–938.

Greene, William H. 2018. *Econometric analysis*. 8 ed. Pearson.

Hopkins, Vincent, Ali Kagalwala, Andrew Q Philips, Mark Pickup and Guy D Whitten. 2024. "How Do We Know What We Know? Learning from Monte Carlo Simulations." *The Journal of Politics* 86(1):36–53.