Online Supplement for ‘Prediction Intervals for Economic Fixed-Event Forecasts’

Fabian Krüger (KIT) Hendrik Plett (ETH Zürich)

April 9, 2024

Section A of the present supplement provides details on the model from Section 3 of the main paper, whereas Section B provides additional empirical analysis and results.

A Details on the autoregressive model

A.1 Comparison to Patton and Timmermann (2011)

Our model in Section 3 of the main paper differs from the one by Patton and Timmermann in three respects. First, we use weekly (rather than monthly) high-frequency observations. Second, while Patton and Timmermann consider both independent and serially correlated measurement error, we focus on the independent case which seems sufficient to capture the stylized facts of interest. Finally, in our setup the coefficients γ_j (defined below Equation 3) are either strictly increasing or strictly decreasing in j, whereas Patton and Timmermann's weight function for GDP features flat segments (see their Figure B.1). This difference arises because Patton and Timmermann assume that the quarterly GDP index is given by the index of the quarter's last month (e.g., March for the year's first quarter). By contrast, we represent a quarter's index by the average of all weeks within the quarter (e.g., weeks 1-13 for the first quarter). We prefer our functional form since it is simpler while, in our view, being at least as plausible. In particular, averaging the weekly levels within a quarter is consistent with averaging the quarterly levels within a year (which is done by both Patton and Timmermann's and our approach).

A.2 Derivation of aggregation weights

Here we derive the triangular weighting function that we use to approximate the annual-average-over-annual-average growth rate based on (hypothetical) weekly data. While this weighting function and its excellent approximation quality are well known (see e.g. Patton and Timmermann, 2011; Hepenstrick and Blunier, 2022, and the references therein), our derivation is simpler than the ones we are aware of, and we thus include it for ease of reference.

Let GDP^*_w denote the hypothetical level of GDP in week w. The predictand of interest is the growth rate in the annual average of ¸GDP. We next show that this growth rate is well approximated by a weighted sum of weekly logarithmic growth rates. For ease of presentation, we consider the percent growth rate from year $t = 1$ (weeks $w = 1, 2, \dots, 52$) to year $t = 2$ (weeks $w = 53, 54, \dots, 104$), which we denote by g_2. We obtain

$$
\begin{aligned}
g_2/100 &= \frac{\frac{1}{52}\sum_{w=53}^{104}\mathrm{GDP}_w^*}{\frac{1}{52}\sum_{w=1}^{52}\mathrm{GDP}_w^*} - 1 && (1)\\
&\approx \log\left(\frac{1}{52}\sum_{w=53}^{104}\mathrm{GDP}_w^*\right) - \log\left(\frac{1}{52}\sum_{w=1}^{52}\mathrm{GDP}_w^*\right) && (2)\\
&\approx \log\left(\prod_{w=53}^{104}\mathrm{GDP}_w^*\right)^{1/52} - \log\left(\prod_{w=1}^{52}\mathrm{GDP}_w^*\right)^{1/52} && (3)\\
&= \frac{1}{52}\sum_{w=53}^{104}\left(\log\mathrm{GDP}_w^* - \log\mathrm{GDP}_{w-52}^*\right) && (4)\\
&= \sum_{w=2}^{104}\frac{52-|53-w|}{52}\underbrace{\left(\log\mathrm{GDP}_w^* - \log\mathrm{GDP}_{w-1}^*\right)}_{\equiv Y_w^*/100}, && (5)\\
&= \sum_{j=1}^{103}\underbrace{\left(1-\frac{|52-j|}{52}\right)}_{\equiv\gamma_j} Y_{105-j}^*/100, && (6)
\end{aligned}
$$

where Y_w^* is the logarithmic GDP growth rate from week $w-1$ to w (in percent), and $\gamma_j \in [1/52, 1]$ denotes a triangle-type weighting function as described in the main text. Note that Equation (6) corresponds to Equation (3) from the main paper in the case $t = 2$.

The first approximation (Equation 2) replaces exact growth rates by logarithmic growth rates, based on the common first-order Taylor expansion $\log(1+z) \approx z$ for $z \approx 0$. The second approximation (Equation 3) replaces the arithmetic mean of the annual GDP level by the geometric mean. Note that both approximations are exactly correct if and only if GDP is constant over time, i.e. if $\mathrm{GDP}_1^* = \mathrm{GDP}_2^* = \ldots = \mathrm{GDP}_{104}^*$. In this case, the annual growth rate g_2 is exactly zero, and the arithmetic and geometric mean coincide. In this sense, approximating the arithmetic mean by the geometric mean is not an additional assumption, but represents another use of the same approximation (namely, constant GDP levels and hence zero growth rates).

Equation (4) is the unweighted arithmetic mean of 52 year-over-year log growth rates. Equation (6) is the weighted arithmetic mean of 103 month-over-month log growth rates. Finally, note that the geometric mean is weakly smaller than the arithmetic mean. The two approximations in Equation (3) hence feature errors of opposite signs: We're underestimating average GDP in year 2, but also subtract an underestimated version of GDP in year 1. This setup is beneficial in that it allows the approximation errors to partly offset each other.

A.3 State space representation

Here we represent the model from Section 3 of the main paper as a linear state space model. Our notation loosely follows Hamilton (1994). The state equation is given by

$$
\begin{aligned}
\xi_w^* &= F\xi_{w-1}^* + v_w^*,\\
v_w^* &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}_{[103,1]}, Q),
\end{aligned}
$$

where $\xi_w^* = \begin{pmatrix} Y_w^*, & Y_{w-1}^*, & \ldots, & Y_{w-102}^* \end{pmatrix}'$ is a 103×1 vector of weekly log growth rates up until week w. $F = \begin{pmatrix} F_1 \\ F_2 \end{pmatrix}$ is a 103×103 matrix, where $F_1 = \begin{pmatrix}\phi, & \mathbf{0}_{[1,102]}\end{pmatrix}$ and $F_2 = \begin{pmatrix} I_{102}, & \mathbf{0}_{[102,1]}\end{pmatrix}$, with I_{102} denoting the 102-dimensional identity matrix, and $\mathbf{0}_{[r,c]}$ denoting an $r \times c$ matrix of zeros. v_w^* is an 103×1 vector of error terms with mean zero and covariance matrix Q, where Q is an 103×103 matrix with its $[1,1]$ element equal to σ_ε^2, and all other elements equal to zero. The observation equation is given by

$$
\tilde{Y}_w^* = H'\xi_w^* + \eta_w^*,
$$

where

$$
H = \begin{pmatrix} 1, & 0 & \ldots, & 0 \end{pmatrix}
$$

is a 103×1 vector, and $\eta_w^* \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\eta^2)$ is a scalar random variable. Conceptually, $\tilde{Y}_w^*$ is a noisy version of the weekly growth rate Y_w^*. We can further write $Y_t = \Gamma' \xi^*_{w_{\text{last}}(t)}$, where $w_{\text{last}}(t)$ species the index for the last week of year t, and

$$\Gamma = \begin{pmatrix} \gamma_1, & \gamma_2, & \dots, & \gamma_{103} \end{pmatrix}'$$

is a 103×1 vector collecting the triangle-type aggregation weights defined below Equation (3) in the main paper. With the above setup, standard Kalman filter formulas can be used to compute the optimal h-step ahead forecasts of Y_t, based on the relevant information set $(\tilde{Y}_w^*)_{w=1}^{w_{\text{last}}(t)-h}$ consisting of noisy versions of weekly growth rates. In a nutshell, the Kalman filter yields an optimal ('smoothed') estimate of the sequence $(\xi_w^*)_{w=1}^{w_{\text{last}}(t)-h}$ of state vectors. This estimate is then used to predict the future state vector $\xi_{w_{\text{last}}(t)}$ and, finally, Y_t. See e.g. Hamilton (1994) for an exposition of the relevant computations.

A.4 Other parameter choices

Figure S1 explores the model's structural parameters, by changing one parameter at a time. Compared to the default case (solid curve), doubling either ρ (autoregressive persistence) or σ_ε^2 (variance of white noise error term in autoregression) increases the RMSFE at all horizons, and tends to yield a steeper increase of the RMSFE across horizons. Doubline σ_η^2 (variance of measurement error) increases the RMSFE at short horizons but has no visible effect at horizons exceeding one year.

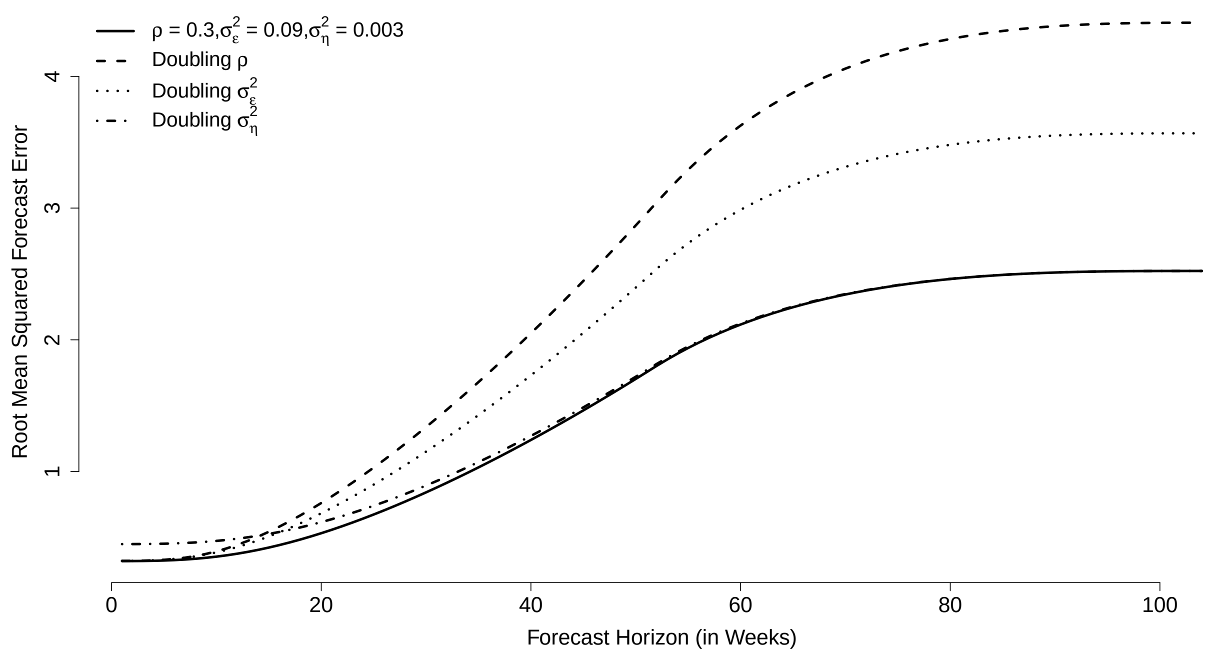

Figure S1: The setup corresponds to Figure 1 from the main paper, and the solid curve is the same as in that figure. The other curves represent different parameter values: In each case, one parameter is doubled (compared to the solid curve), whereas the others are held constant.

A.5 Correlation structure of forecast errors

Consider two forecast errors e_{t_1,h_1} and e_{t_2,h_2}, where $e_{t,h}$ is a forecast for year t, made h weeks before the end of year t. By convention, we let $t_2 \geq t_1$. Furthermore, we assume that $h_1, h_2 \leq 104$. Below we provide matrix representations of the two forecast errors. We first introduce the auxiliary matrices L and $A(h)$. The former, which is of dimension $[2, h_1 + h_2]$, is given by

$$L = \begin{pmatrix} \mathbf{g}_{[1,h_2]} & \mathbf{0}_{[1,h_1]} \\ \mathbf{0}_{[1,h_2]} & \mathbf{g}_{[1,h_1]} \end{pmatrix},$$

where $\mathbf{g} = \begin{pmatrix} \gamma_1, & \gamma_2, & \ldots, & \gamma_{103}, & 0 \end{pmatrix} = \begin{pmatrix} 1, & 2, & \ldots, & 51, & 52, & 51, & \ldots, & 2, & 1, & 0 \end{pmatrix}/52$ is of dimension 1×104, and the notation $\mathbf{g}_{[1,r]}$ denotes the first r elements of $\mathbf{g}$. $A(h)$ is an upper triangular $h \times h$ matrix with ones on the main diagonal, and $[i, j]$ element equal to ρ^{j-i} for $j > i$.

We next introduce a matrix expression for the 2×1 vector $\begin{pmatrix} e_{t_2,h_2}, & e_{t_1,h_1} \end{pmatrix}'$ of forecast errors. This expression allows to assess the errors' correlation structure. In the following expressions, we assume that the matrix $\mathbf{0}_{[r,c]}$ of zeros vanishes whenever $r \leq 0$ or $c \leq 0$.

Case 1: Both errors refer to the same target year ($t_1 = t_2$)

Here the number of relevant weekly shocks (that enter either e_{t_2,h_2} or e_{t_1,h_1} or both) is given by $T^* = \max(h_1, h_2)$.

$$\begin{pmatrix} e_{t_2,h_2} \\ e_{t_1,h_1} \end{pmatrix} = L \times \begin{pmatrix} A(h_2) & \mathbf{0}_{[h_2,T^*-h_2]} \\ A(h_1) & \mathbf{0}_{[h_1,T^*-h_1]} \end{pmatrix} \times \begin{pmatrix} \varepsilon^*_{w_{\text{last}}(t_2)} \\ \varepsilon^*_{w_{\text{last}}(t_2)-1} \\ \vdots \\ \varepsilon^*_{w_{\text{last}}(t_2)-T^*+1} \end{pmatrix}$$

Case 2: Target years differ by one ($t_2 = t_1 + 1$)

Here T^* (defined as in Case 1) is given by $T^* = \max(h_1 + 52, h_2)$.

$$\begin{pmatrix} e_{t_2,h_2} \\ e_{t_1,h_1} \end{pmatrix} = L \times \begin{pmatrix} A(h_2) & \mathbf{0}_{[h_2,T^*-h_2]} & \\ \mathbf{0}_{[h_1,52]} & A(h_1) & \mathbf{0}_{[h_1,T^*-52-h_1]} \end{pmatrix} \times \begin{pmatrix} \varepsilon^*_{w_{\text{last}}(t_2)} \\ \varepsilon^*_{w_{\text{last}}(t_2)-1} \\ \vdots \\ \varepsilon^*_{w_{\text{last}}(t_2)-T^*+1} \end{pmatrix}$$

Case 3: Target years differ by two or more ($t_2 \geq t_1 + 2$)

Recalling that $h_1, h_2 \leq 104$, we have $T^* = (t_2 - t_1) \times 52 + h_1$, and

$$\begin{pmatrix} e_{t_2,h_2} \\ e_{t_1,h_1} \end{pmatrix} = L \times \underbrace{\begin{pmatrix} A(h_2) & \mathbf{0}_{[h_2,T^*-h_2]} \\ \mathbf{0}_{[h_1,T^*-h_1]} & A(h_1) \end{pmatrix}}_{\underline{A}} \times \begin{pmatrix} \varepsilon^*_{w_{\text{last}}(t_2)} \\ \varepsilon^*_{w_{\text{last}}(t_2)-1} \\ \vdots \\ \varepsilon^*_{w_{\text{last}}(t_2)-T^*+1} \end{pmatrix}$$

We next illustrate the computation of the errors' covariance matrix for Case 3. Recalling the assumption of Gaussian IID errors ε^*_m, with $\text{Var}(\varepsilon^*_m) = \sigma^*_\varepsilon$, we get

$$\text{Var} \begin{pmatrix} e_{t_2,h_2} \\ e_{t_1,h_1} \end{pmatrix} = \sigma^2_\varepsilon \ L \ \underline{A} \ \underline{A}' L'.$$

By examining the definitions of L and $\underline{A}$, it becomes clear that $\text{Var} \begin{pmatrix} e_{t_2,h_2} \\ e_{t_1,h_1} \end{pmatrix}$ is a diagonal matrix in Case 3. If $h_2 \leq 52$, the same finding obtains in Case 2. Otherwise (in Case 2 with $h_2 \geq 53$, or in Case 1), we generally obtain a non-diagonal covariance matrix for $\begin{pmatrix} e_{t_2,h_2} \\ e_{t_1,h_1} \end{pmatrix}$.

B Additional empirical results

B.1 Forecast calibration

Figure S2 presents results on the coverage of the prediction intervals, evaluated separately across forecast horizons.

B.2 Diebold-Mariano type testing

In Section B.2, we provide additional discussion and results on Diebold-Mariano type testing for equal predictive ability of postprocessing methods versus the SPF 'histograms'.

B.2.1 Implementation variants

We first repeat the Diebold-Mariano (DM) tests reported in Figure 3 of the main paper, using various test implementations. The DM test refers to the null hypothesis that $\mathbb{E}(d_{t,h}) = 0$, i.e., that the expected difference between the interval scores of two methods is zero.[1] Here t denotes the target year, and h denotes the forecast horizon (measured in weeks). The DM tests are conducted separately across horizons $h \in \{6.5, 19.5, \ldots, 97.5\}$, where the choice of horizons conforms to the timing of the SPF. We thus seek to test the null hypothesis based on a sample $(d_{t,h})_{t=1}^{n_h}$, using a t-statistic of the form

$$\frac{\bar{d}_h}{\widehat{V}(\bar{d}_h)},$$

where $\bar{d}_h = n_h^{-1} \sum_{t=1}^{n_h} d_{t,h}$ is the sample average of the loss differences, and $\hat{V}$ denotes an estimate of the variance of the sample average. Due to data availability, the size n_h of the evaluation sample ranges between 36 and 42, depending on the horizon h. Common choices of $\hat{V}$ depend on estimates or assumptions relating to the time series dependence (specifically, autocorrelation) in $(d_{t,h})_{t=1}^{n_h}$. To discuss this aspect, note that the loss differences $d_{t,h}$ are observed annually, i.e., the target years for the observations $d_{t,h}$ and $d_{t+1,h}$ are one year apart. Since four of the eight forecast horizons h we consider are less than one year, the informal benchmark setup suggested by Diebold and Mariano (1995, Section 1.1) corresponds to a truncation lag of zero (i.e., no consideration of autocorrelation) for these horizons, and a truncation lag of one (i.e., consideration of autocorrelation up to lag one) for the remaining four horizons.[2] Alternatively, the truncation lag can be determined according to data-based procedures that are popular in the literature on covariance matrix estimation in the presence of time dependence.

We consider the following DM test implementations:

1. **sandwich**: The estimator implemented in the function NeweyWest of the R package sandwich (Zeileis, 2004; Zeileis et al., 2020), implementing the data-based choice of truncation lag described in Newey and West (1994). We compare the resulting t-statistic to standard normal critical values. This choice is the same as in Figure 3 of the main paper.

2. **CM13**: An implementation variant that performs well in Monte Carlo simulations by Clark and McCracken (2013) on DM testing under squared error loss. The test uses standard normal critical values. Variance estimation is based on a rectangular kernel, with the truncation lag chosen according to the Diebold and Mariano benchmark setup mentioned above. Furthermore, the finite sample adjustment proposed by Harvey et al. (1997) is applied to the test statistic. (Referring to the authors of the latter study, Clark and McCracken denote this implementation variant by 'HLN'.)

[1] In more detail, let $d_{t,h}$ denote the difference in interval scores between the two methods of interest: Combination of postprocessing methods, henceforth abbreviated by 'comb.', and the average of the histogram type forecasts from the Survey of Professional Forecasters, henceforth 'SPF'. For a given target year t and forecast horizon h, this difference is given by

$$d_{t,h} = \mathrm{IS}(l_{t,h}^{\text{comb.}}, u_{t,h}^{\text{comb.}}, y_t) - \mathrm{IS}(l_{t,h}^{\text{SPF}}, u_{t,h}^{\text{SPF}}, y_t),$$

where $l_{t,h}^{j}$ denotes the lower end of the prediction interval produced by method $j \in \{\text{comb.}, \text{SPF}\}$, for target year t and horizon h. The definition of the upper end $u_{t,h}^{j}$ of the interval is analogous, and y_t denotes the realizing observation. The interval score is defined in Equation (6) of the main paper.

[2] Diebold and Mariano (1995) denote the difference in losses by d_t, and the forecast horizon by k. Unlike in our notation, t refers to an arbitrary time scale (e.g., monthly, quarterly, or annually), and the forecast horizon k is measured on the same scale as t. Based on properties of optimal k-step ahead forecast errors, Diebold and Mariano propose using $(k - 1)$ as a benchmark choice of truncation lag.

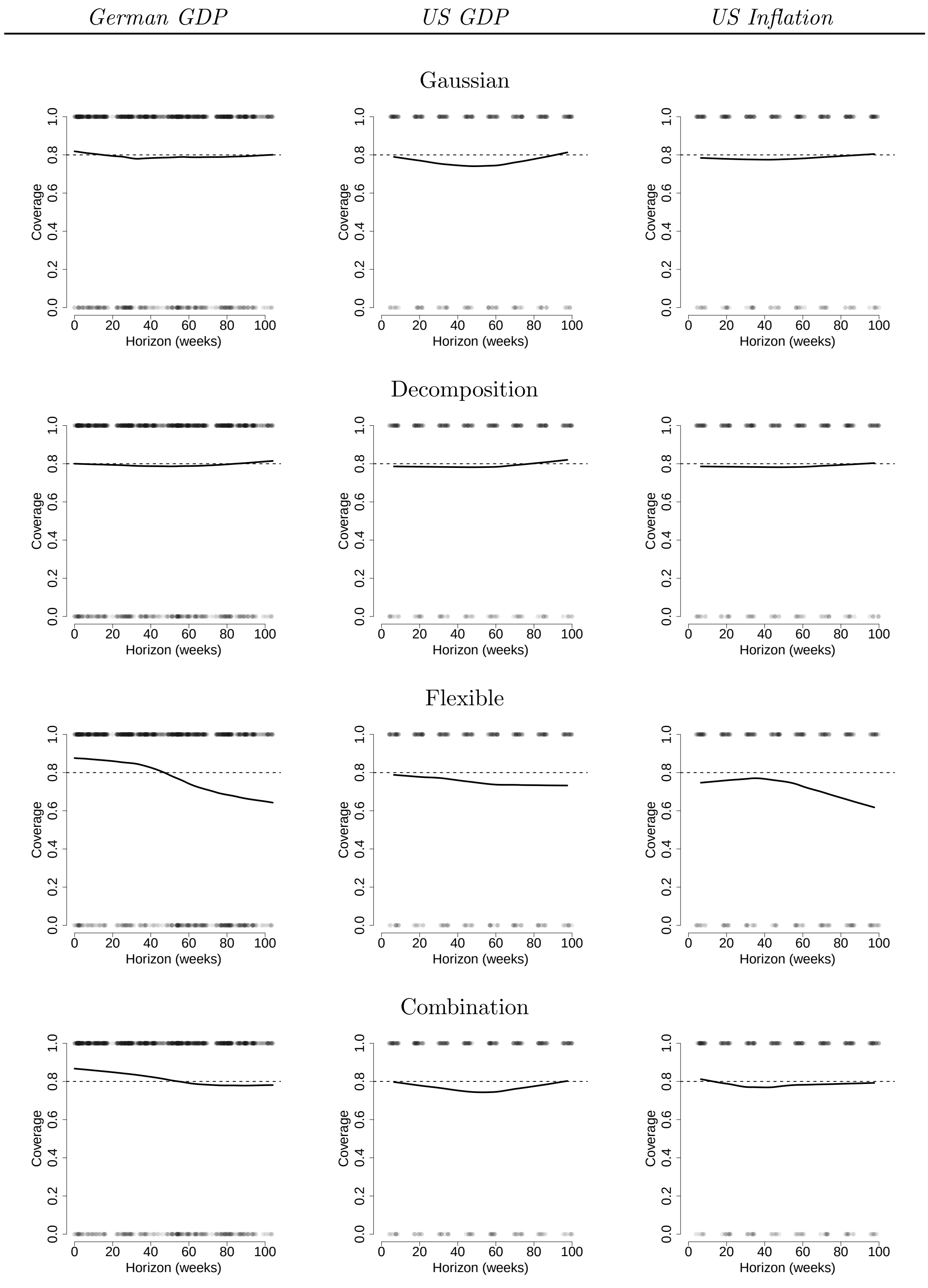

Figure S2: Coverage of the prediction interval (vertical axis; 0 = no, 1 = yes) at various horizons (horizontal axis; jittering used for better display). Dots represent individual forecast/outcome pairs. Solid line shows nonparametric estimate of coverage probability as a function of the horizon. Dashed horizontal line indicates the target coverage probability of 80%.

3. **EWC**: A test statistic using the equally weighted cosine (EWC) variance estimator proposed by Lazarus et al. (2018). We compare the test statistic against critical values from a t distribution with ν degrees of freedom, where ν is an implementation parameter that is set according to the rule in Lazarus et al. (2018, Equation 4 and Footnote 3). For our choices of sample size n_h, the rule prescribes to set $\nu = 4$.

4. **IID**: A standard t-test that does not account for autocorrelation. We compare the test statistic against critical values from a t distribution with $n_h - 1$ degrees of freedom. This test serves as a simple baseline, and illustrates the impact of ignoring autocorrelation.

While this list of DM implementation variants is quite diverse, it is far from exhaustive. See Coroneo and Iacone (2020) and Coroneo et al. (2024) for further options, including the use of bootstrap techniques and different asymptotic frameworks.

The top row of Figure S3 shows the different p-values for GDP (left panel) and inflation (right panel). The four test implementations generally show a high level of agreement. The main differences occur for GDP at horizons $h \in \{19.5, 32.5, 45.5\}$. Even for these settings, however, all tests reject the null at the ten percent level except in one case (EWC, $h = 45.5$). Generally, there is a tendency for EWC to be more conservative (i.e., produce larger p-values) than the other tests.

As noted, the four test implementations differ in their handling of potential autocorrelation in the loss differences. The bottom row of Figure S3 presents empirical evidence on autocorrelation, by showing the empirical first-order autocorrelation coefficients of the series $(d_{t,h})_{t=1}^{n_h}$, for different horizons h. For both variables, these coefficients attain small positive or even negative values at the four shortest horizons. They are somewhat larger at the four longer horizons, but remain moderate (below 0.36) even there. The figures also show that there is substantial sampling uncertainty, with few of the autocorrelation coefficients being significantly different from zero at the five percent level. As a consequence, it is not obvious which DM test variant is most appropriate in terms of size and power in the current situation.

B.2.2 Performance comparisons on average across horizons

Instead of conducting DM tests separately for each horizon h, it is also possible to evaluate forecasting performance jointly across all horizons. Similar to Quaedvlieg (2021), one strategy for doing so is to consider average forecast performance across horizons. To this end, let $\bar{d}_t = \frac{1}{8} \sum_{h \in \mathcal{H}} d_{t,h}$, where the set $\mathcal{H} = \{6.5, 19.5, \ldots, 97.5\}$ comprises the eight forecast horizons of interest. Figure S4 shows the time series of d_t for GDP (left panel) and inflation (right panel). The loss differences are defined such that positive values are in favor of the SPF, whereas negative values are in favor of the combination.

We can now test the null hypothesis that $\mathbb{E}(\bar{d}_t) = 0$ using a DM test like the ones considered above. (Alternatively, Quaedvlieg (2021) considers bootstrap techniques for testing this hypothesis.) In order to implement the test, we require complete data on loss differences at all eight horizons h, for any given year t. This requires us to remove years for which one or more horizons are missing, so that our final data set includes 33 observations (i.e., years).

Table S1 presents the results of this analysis. For GDP, the EWC test variant yields a p-value of 19.2%, whereas the other variants attain p-values around five percent. For inflation, all four test variants yield very small p-values below 0.5%. In all cases, rejections are due to negative t-statistics (i.e., in favor of the combination method).

B.3 Comparisons to additional forecasting methods

Table S2 presents additional empirical comparisons to a Gaussian model under the assumption of a zero mean ($\mu = 0$), and to quantile-based combinations of survey forecasts ('quantile mean' and 'quantile median'). See the last two paragraphs of Section 7.2 in the main paper for details.

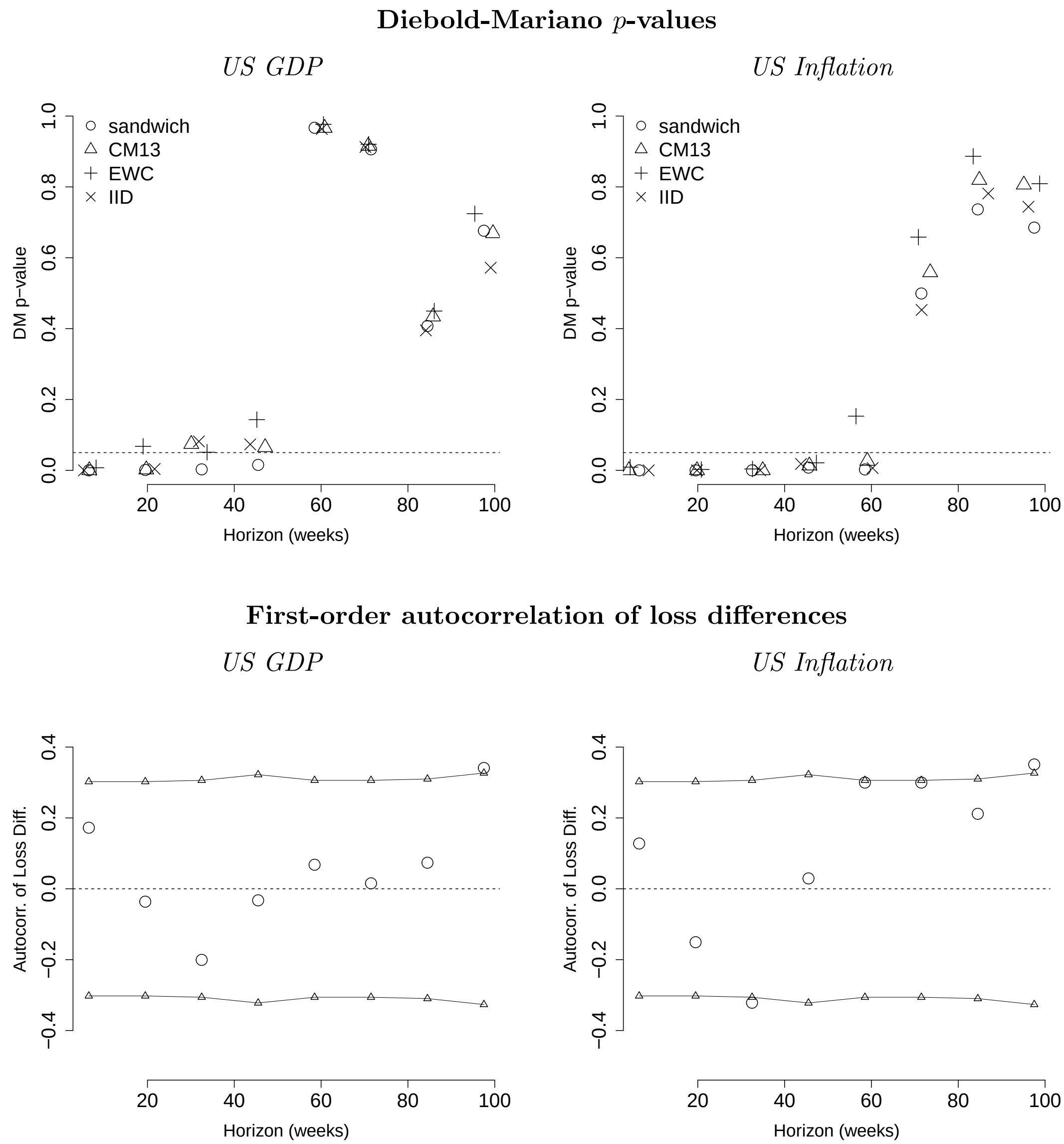

Figure S3: Top row: Diebold-Mariano p-values obtained from different test implementations (two-sided tests). Horizontal location of points is jittered for better display. Horizontal line marks a p-value of five percent. Bottom row: Circles mark first-order autocorrelation of loss differences $d_{t,h}$. Triangles (connected by lines) mark asymptotic critical values for the null hypothesis of zero autocorrelation (see Lütkepohl, 2005, Section 4.4.1) at the 5% level.

	US GDP	*US Inflation*
sandwich	0.050	0.000
CM13	0.047	0.000
EWC	0.192	0.000
IID	0.055	0.001

Table S1: p-values for the null hypothesis that $\mathbb{E}(\bar{d}_t) = 0$. Two-sided tests. Rows represent four DM test implementations.

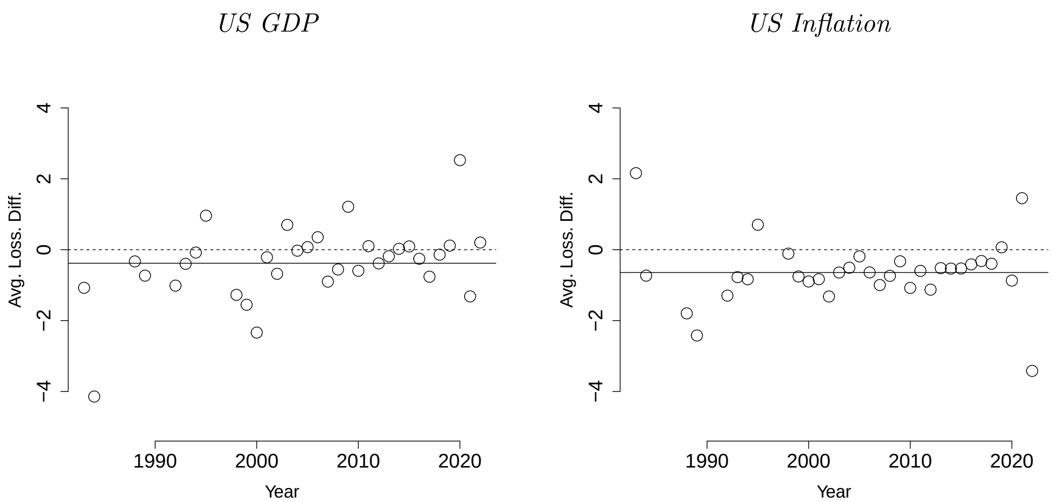

Figure S4: Average loss differences across horizons, for various target years t. Dashed horizontal line marks zero. Solid horizontal line marks the average loss difference over time.

Model	Coverage	PI length	Interval Score
German GDP (1991-2022, $n = 1307$)			
Gaussian	79.11%	2.71	5.81
Gaussian ($\mu = 0$)	78.50%	2.74	5.84
US GDP (1981-2022, $n = 320$)			
Gaussian	76.56%	2.30	4.11
Gaussian ($\mu = 0$)	78.12%	2.30	4.09
SPF Histogram	85.94%	2.97	4.48
SPF Histogram (quantile mean)	79.38%	2.35	4.58
SPF Histogram (quantile median)	77.50%	2.15	4.51
US Inflation (1981-2022, $n = 320$)			
Gaussian	78.44%	1.30	2.65
Gaussian ($\mu = 0$)	77.81%	1.37	2.67
SPF Histogram	85.94%	2.25	3.35
SPF Histogram (quantile mean)	81.88%	1.88	3.19
SPF Histogram (quantile median)	80.00%	1.68	3.28

Table S2: The setup corresponds to Table 3 in the main paper, from which the 'Gaussian' and 'SPF Histogram' results are copied for easier reference. 'Gaussian ($\mu = 0$)' denotes a restricted variant of the Gaussian method. The two additional SPF methods (US data only) represent different survey aggregation schemes as described in the text.

References

Clark, T. and M. McCracken (2013): "Chapter 20 - Advances in forecast evaluation," in *Handbook of Economic Forecasting*, ed. by G. Elliott and A. Timmermann, Elsevier, vol. 2, 1107–1201.

Coroneo, L. and F. Iacone (2020): "Comparing predictive accuracy in small samples using fixed-smoothing asymptotics," *Journal of Applied Econometrics*, 35, 391–409.

Coroneo, L., F. Iacone, and F. Profumo (2024): "Survey density forecast comparison in small samples," *International Journal of Forecasting*, forthcoming.

Diebold, F. X. and R. S. Mariano (1995): "Comparing predictive accuracy," *Journal of Business & Economic Statistics*, 13, 253–263.

Hamilton, J. D. (1994): "State-space models," in *Handbook of Econometrics*, ed. by R. F. Engle and D. L. McFadden, Elsevier, vol. 4, 3039–3080.

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

Hepenstrick, C. and J. Blunier (2022): "What were they thinking? Estimating the quarterly forecasts underlying annual growth projections," Swiss National Bank Working Paper 5/2022.

Lazarus, E., D. J. Lewis, J. H. Stock, and M. W. Watson (2018): "HAR inference: Recommendations for practice," *Journal of Business & Economic Statistics*, 36, 541–559.

Lütkepohl, H. (2005): *New Introduction to Multiple Time Series Analysis*, Springer.

Newey, W. K. and K. D. West (1994): "Automatic lag selection in covariance matrix estimation," *The Review of Economic Studies*, 61, 631–653.

Patton, A. J. and A. Timmermann (2011): "Predictability of output growth and inflation: A multi-horizon survey approach," *Journal of Business & Economic Statistics*, 29, 397–410.

——— (2012): "Forecast rationality tests based on multi-horizon bounds," *Journal of Business & Economic Statistics*, 30, 1–17.

Quaedvlieg, R. (2021): "Multi-horizon forecast comparison," *Journal of Business & Economic Statistics*, 39, 40–53.

Zeileis, A. (2004): "Econometric computing with HC and HAC covariance matrix estimators," *Journal of Statistical Software*, 11, 1–17.

Zeileis, A., S. Köll, and N. Graham (2020): "Various versatile variances: An object-oriented implementation of clustered covariances in R," *Journal of Statistical Software*, 95, 1–36.