Köhler, Tim; Döpke, Jörg

Will the last be the first? Ranking German macroeconomic forecasters based on different criteria

Empirical Economics

Provided in Cooperation with:
Springer Nature

Suggested Citation: Köhler, Tim; Döpke, Jörg (2022) : Will the last be the first? Ranking German macroeconomic forecasters based on different criteria, Empirical Economics, ISSN 1435-8921, Springer, Berlin, Heidelberg, Vol. 64, Iss. 2, pp. 797-832, https://doi.org/10.1007/s00181-022-02267-9

Empirical Economics (2023) 64:797–832
https://doi.org/10.1007/s00181-022-02267-9

Will the last be the first? Ranking German macroeconomic forecasters based on different criteria

Tim Köhler[1] · **Jörg Döpke**[2]

Received: 12 October 2021 / Accepted: 25 May 2022 / Published online: 29 June 2022

Abstract
We rank the quality of German macroeconomic forecasts using various methods for 17 regular annual German economic forecasts from 14 different institutions for the period from 1993 to 2019. Using data for just one year, rankings based on different methods correlate only weakly with each other. Correlations of rankings calculated for two consecutive years and a given method are often relatively low and statistically insignificant. For the total sample, rank correlations between institutions are generally relatively high among different criteria. We report substantial long-run differences in forecasting quality, which are mostly due to distinct average forecast horizons. In the long-run, choosing the criterion to rank the forecasters is of minor importance. Rankings based on recession years and normal periods are similar. The same does hold for rankings based on real-time vs revised data.

Keywords Macroeconomic forecasts · Ranking · Germany

1 Introduction

Ranking forecasters has a long tradition in economics and finance (see, e.g., Cowles 1933). Forecast competitions are (perhaps even increasingly) popular in the media: Several newspapers and database providers[1] evaluate forecasters more or less regularly

[1] As regards Germany, for example, Fricke (2018) for various newspapers or Handelsblatt (2014). Consensus Forecast (ed) (2020) ranks forecasters in several countries and regions, including Germany.

✉ Tim Köhler
tim.koehler@hs-merseburg.de

Jörg Döpke
joerg.doepke@hs-merseburg.de

[1] University of Applied Sciences Merseburg, Eberhard-Leibnitz-Straße 2, 06217 Merseburg, Germany

[2] University of Applied Sciences, Merseburg, Germany

(see Silvia and Iqbal (2012), and Döhrn (2015) for an overview of recent German rankings).

Behind these efforts, we can assume a broad range of possible motivations: First, of course, there is the interest of the audience and, thus, an aspect of entertainment. Second, some authors are interested in comparing forecasts stemming from a particular institution with those of others as a benchmark for their quality (see, among others, Fritsche and Tarassow (2017) for the German IMK Institute or Pagan (2003) for the Bank of England). Third, it might be relevant how policy-makers rank within the forecasting industry: Lehmann and Wollmershäuser (2021), for example, consider the possibility that governments' projections might have different properties as compared to forecasts of private institutions. In a similar vein, Gamber et al. (2014) analyze the quality of central bank forecasts. Additionally, it might be of interest from a monetary policy perspective, which institution ranks high in a list of forecasters since both FED and ECB conduct a survey of professional forecasters (see, for example, Meyler 2020; Rich and Tracy 2021, for the ECB). Finally, comparing forecast accuracy across countries (Heilemann and Müller 2018; Heilemann and Stekler 2013) might also give valuable insights, for example, in analyzing possible lower bounds of accuracy.

From a methodological perspective, several approaches have been proposed to rank forecasters (see, the literature cited in, Sinclair et al. 2012, 2016). To assess forecast quality, it is possible to use absolute or relative (that employ another prediction as a benchmark) accuracy measures. The criteria may rely on linear or quadratic loss functions. The evaluation may use numerical forecast errors or an analysis of directional change. One-dimensional rankings that refer to only one variable may lead to other results than multi-dimensional ones based on a vector of variables (Sinclair et al. 2015; Fortin et al. 2020). Our paper refers to numerical forecasts only, but it is noteworthy that Rybinski (2021) has recently suggested a ranking that relies on sentiment indices obtained from the texts of the forecast reports.

Other factors can influence the rankings as well: Do the forecasters foresee all variables equally well? Alternatively, emerge different rankings for growth, inflation, unemployment, or other variables? In other words: are there specialists among the institutions that are particularly good at predicting a specific variable, as it is considered by Timmermann and Zhu (2019)? Related, given findings according to which forecast errors strongly depend on business cycle phases (Dovern and Jannsen 2017), one might also ask whether some forecasters are specialists for specific periods, say, recessions.

From the perspective of economic policy, we also take into account some other aspects: First, not all relevant forecasters make their forecasts at the same time. Hence, the question arises, what is the impact of a longer forecast horizon in case of a fixed-event-forecast?(Knüppel and Vladu 2016) Is the horizon more important for accuracy as the institute, as found by, for example, Döhrn and Schmidt (2011)? Second, is it also crucial whether we use the most recently available data or real-time data? Döhrn (2019) argues using German data that the magnitude of forecast errors depends, to a substantial amount, on data revisions after the forecast has been made and evaluated. Does the same hold for rankings? Finally, ranking the institutes may rely not on one year only but on a more extended period. Evaluating forecasting institutions for a longer time ensures that a superior performance in a particular year is not the result of

sheer luck. Consequently, several authors ask whether any forecaster is consistently better than all others (see, among others, Stekler 1987; Batchelor 1990; Qu et al. 2019).

All in all, these different aspects raise the question, whether criteria proposed by media, practitioners, and academic literature lead to similar results. Depend conclusions such as "all forecasters are equal"(Batchelor 1990) crucially on the criterion used to rank them? Are forecasting competitions meaningful beyond the aspect of pure entertainment? Or, as Döhrn (2015) puts it, lead such efforts just to "random results"?

The primary purpose of this paper is to discuss these questions empirically and compare the forecast quality of institutions that predict the macroeconomic development in Germany. To this end, we use annual data for 17 growth, inflation, unemployment rate, ex- and import changes forecasts from 1993 to 2019, which come from 14 institutions. The choice of these institutes and organizations is motivated by their importance for economic policy and the attention they receive from the media. The length of our sample includes all years for which the institutions made forecasts explicitly for unified Germany and not for West Germany only or separately for both parts.[2] Furthermore, we tried to include forecasts that comprise at least predictions for growth, inflation, unemployment, exports, and imports. The choice of these variables is also motivated by their relevance for economic policy since they roughly relate to the so-called magic square of German economic policy.[3]

To these data, we employ various methods to rank the quality of the forecasters, which differ along the lines mentioned above. If one looks at only one particular year, the rankings vary widely: an institution that makes, say, good growth forecasts is not necessarily equally good at predicting inflation. Moreover, rankings of forecasters for a given variable vary considerably between two consecutive years: the institute that has the best growth forecast in one year might easily end up at the bottom of the ranking in the following year.

From a longer-term perspective, the rankings show a more stable picture. Using the total sample, the correlations between the forecast performances according to various criteria across institutions are pretty high. Also, some institutes are superior or inferior to others in this longer perspective. This over-or under-performance, though, is almost exclusively due to a shorter or longer average forecasting horizon.

We organize the paper as follows: Sect. 2 introduces several criteria to rank forecasters. Section 3 describes the dataset. Section 4 presents rankings of the forecasting performance of the respective institutions and compares the rankings based on different criteria. Furthermore, we discuss whether these rankings are stable over time and whether there is an institution that outperforms the other in the long-run. Section 5 concludes.

[2] Regarding some exceptions for the year 1993, see the data appendix.

[3] Some—particularly older—forecasts just referred to growth and/or inflation, limiting our sample in the time dimension.

2 Criteria to rank forecasters

As already mentioned above, it is possible to evaluate forecasters in various dimensions. More specifically, (at least) the following dimensions should be considered: (i) The number of forecasters to be ranked, (ii) the number of predictions for each institution, (iii) the number of variables considered, (iv) the length of the forecast horizon, and, last but not least, the statistical method measuring forecast quality.

By combining these dimensions, one can create numerous rankings. For example, a ranking may include three forecasters who each forecast growth and inflation in October 2017 for 2018. Another ranking may rely just on growth forecasts, but for 20 years and include 20 institutions. These examples show that there is a huge number of possibilities to create a leaderboard for forecasters. This number increases even further if one takes into account the degrees of freedom within a specific statistical evaluation method. For example, in applying multi-dimensional methods, the individual variables can be weighted differently.

Among this large number of possible rankings, we opt, first, for fixed-event forecasts with a forecast horizon of at least 8 to a maximum of 16 months[4], which allows including all arguably policy-relevant institutions, which have regularly forecasted for many years, which yields 17 different predictions.[5]

As mentioned above, we analyze these data in a first step, for one year only, because this is the state-of-the-art in most forecasting competitions. Then, we turn to the stability of yearly rankings before we consider the total sample of data. As regards the target variables, we chose five (real GDP growth, inflation, unemployment, real export and import growth), again motivated by relevance to economic policy.

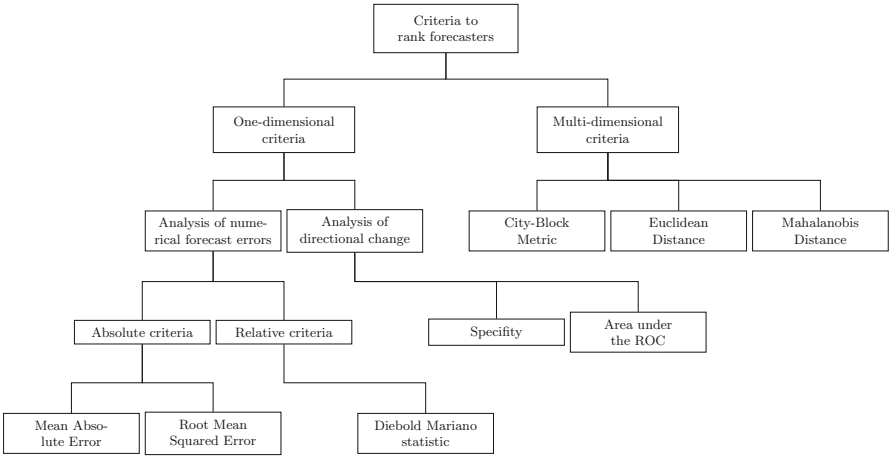
Figure 1 gives a bird's-eye view of the statistical methods to rank forecasters considered in this paper. On the one hand, we consider measures that are one-dimensional in the sense that they evaluate forecasts only for one target. On the other hand, we examine multi-dimensional measures, which assess several forecasts. Within the group of one-dimensional measures, it is possible to distinguish evaluations based on numerical forecast errors and figures relying on the analysis of directional change. Some readers might miss relative measures like Theil's Inequality Coefficient or the Scaled Mean Error (Hyndman and Koehler 2006). Note, however, that both figures divide a series of absolute errors by the same denominator. Therefore, a ranking based on Theil's inequality coefficient would be identical to a ranking based on the Root Mean Squared Error. In a similar vein, applying the Mean Absolute Errors and the Scaled Mean Error would result in the same hierarchy of forecasters.

2.1 One-dimensional evaluation of business cycle forecasts

The forecast error is defined as $e_t = A_t - F_t$, where A_t is the actual value in period t minus the forecast F_t made in period $t-1$ for period t. Hence, a negative forecast error

[4] See Appendix Table 16 for details.

[5] This choice is also motivated by data availability: We could not consider more forecasters and observation years due to missing forecasts.



Source: Own compilation.

Fig. 1 Methods to rank forecasters—overview

corresponds to overestimating the variable at hand, whereas a positive value represents underestimating. We consider the following statistics:

Simple accuracy measures

– The Mean Absolute Error indicates the average absolute distance between the forecast value and the one that actually occurred. A smaller value indicates a better forecast, and the assumed underlying loss function is a linear one.

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|e_t| \tag{1}$$

– The Root Mean Squared Error is calculated from the square root of the average squared forecast error. By squaring them, large forecast errors are weighted more heavily, referring to a quadratic loss function. Again, a smaller value corresponds to a better forecast.

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}e_t^2} \tag{2}$$

Measures of relative accuracy

One weakness of the rankings based on simple accuracy measures is that these numbers do not have a natural scaling. Therefore, it is necessary to compare it with a competing prediction. For forecast competitions, this shortcoming is of limited relevance, since, in most cases, the benchmark is identical for all institutions.

– We follow Timmermann and Zhu (2019), who use the value of the Diebold and Mariano (1995) test (as compared to an AR(1) in their case or naive forecast, in

Table 1 Notation for direction-of-change forecast evaluation

	Actual i	Actual j	Marginal
Observed cell counts			
Forecast i	O_{ii}	O_{ij}	$O_{i.}$
Forecast j	O_{ji}	O_{jj}	$O_{j.}$
Marginal	$O_{.i}$	$O_{.j}$	Total: O

Source: (in part taken from Diebold and Lopez (1996), Fig. 1)

our case) to classify the quality of a prediction. Hence, we calculate, based on the forecast error of a naive (no change) forecast (e_{Naive}) and of the respective institution ($e_{\text{Institution}}$), the Squared Errors (SE_{Naive} and $\text{SE}_{\text{Institutions}}$, respectively). The difference between these time series—the loss differential ($d_t = \text{SE}_{\text{Institution}} - \text{SE}_{\text{Naive}}$)—is used to obtain the test statistic:

$$\frac{\frac{1}{T}\sum_1^t d_t}{\hat{\sigma_d}} \rightarrow N(0,1) \tag{3}$$

with $\hat{\sigma_d}$ representing a consistent estimate of the standard deviation of d. It should be estimated robustly since loss differentials are likely to be serially correlated(Diebold 2015). Equation 3 allows testing the hypothesis of equal accuracy of both predictions. Hence, the lower the implied test-statistic for this test is, the better is the forecast at hand compared to the naive benchmark.

Measures of directional change

The underlying notation for measures of the accuracy of directional change is taken from Diebold and Lopez (1996) and depicted in Table 1:

A forecast and an actual output can assume the state 'i' (for acceleration) and 'j' (for deceleration). An accurate acceleration forecast, for example, falls into cell O_{ii}. If acceleration is predicted, but a deceleration occurs, then this case falls into cell O_{ij}, and so forth. In particular, we calculate:

- The Area under the Receiver Operating Characteristic Curve (AUROC), which plots the true positive rate (TPR) vs the false positive rate (FPR) at different classification thresholds on the Receiver Operating Characteristic Curve.

$$\text{TPR} = \frac{O_{ii}}{O_{ii} + O_{ij}} \tag{4}$$

$$\text{FPR} = \frac{O_{ji}}{O_{ji} + O_{jj}} \tag{5}$$

The value range of AUROC is 0 to 1, where 1 means that all forecasts are correct and 0 means that all forecasts are wrong.

- Some authors suggest using the specificity of the forecast(Bailey et al. 2018) to compare forecasts, i.e., the share of correct forecast in relation to all forecasts:

$$\text{SPE} = \frac{O_{jj}}{O_{ij} + O_{jj}} \tag{6}$$

2.2 Multi-dimensional evaluation of business cycle forecasts

The measures sketched above refer to one dimension only. As Sinclair et al. (2016) and Döhrn (2015) argue, an evaluation can rest on more than one variable. Hence, it is necessary to refer to a vector of predicted variables. Döhrn (2015) discusses three possible criteria to judge a vector of forecasts:

– The City-Block Metric:

$$D_{\text{CB}} = \sum_{m=1}^{M} |A_m - F_m| \tag{7}$$

where M is the number of predicted series included, for example, M equals five, if we include the five variables mentioned above. The City Block Metric sums the absolute forecast errors across all variables. A lower value of D_{CB} indicates better forecast. The measure rests on the assumption of an underlying linear loss function.

– The Euclidean Distance:

$$D_{\text{EU}} = \sqrt{\sum_{m=1}^{M} (A_m - F_m)^2} \tag{8}$$

The Euclidean Distance assumes an underlying quadratic loss function. Similar to the Root Mean Squared Error, the method weights larger forecast errors more heavily. Again, a smaller value signals a better forecast.

– The Mahalanobis (1936) Distance:

$$D_{\text{MA}} = (F - A)' W (F - A) \tag{9}$$

with F and A as vectors of predictions and actuals, respectively, and W as the inverse variance–covariance matrix, which must be estimated from historical data of the forecasted time series. The primary motivation for this modification is that the forecast errors might be correlated with each other. Consequently, if this is not the case, the covariance matrix is the unit matrix and the Mahalanobis (1936) Distance equals the Euclidean Distance. The estimation of the covariance matrix is based on the last 10 years of actual outcomes (see Sinclair et al. (2016), who use 20 years instead).

2.3 Testing for long-run superiority

To check for a long-run superiority of an institution's forecasting efficiency, we refer to a simple measure of long-run relative performance suggested by Stekler (1987) (see also D'Agostino et al. (2012); Meyler (2020); Rich and Tracy (2021)). In the first step, a score (R_{it}) is assigned to every forecast, which takes the value of the rank according to the respective criterion, for example, the Absolute Forecast Error. In the second step, the cumulated rank-sum of these scores is calculated:

$$S_i = \sum_{t=1}^{T} R_{it} \tag{10}$$

Under the null hypothesis that each institution has the same predictive ability, each institution should have an expected cumulative sum of scores of:

$$S_i^e = \frac{T(N+1)}{2} \tag{11}$$

In our case, we have 27 years and thus an expected value of $\frac{27(17+1)}{2} = 243$. To calculate the test statistic, we use the corrected standard deviation proposed by Batchelor (1990):

$$\sigma = \sqrt{\frac{TN(N+1)}{12}} \tag{12}$$

Additionally, we follow the approach suggested by Bürgi and Sinclair (2017). They calculate for each institution a dummy variable "that takes value 1 in a given period if that forecaster has a lower squared error in that period than the simple average and 0 otherwise." (Bürgi and Sinclair 2017,][p. 106). Over time, the average of the dummy equals the percentage share of periods each institution has beaten the simple average in the past. Equipped with this number, it is possible to check, whether a specific forecaster has been better than the average over a particular period, say, five years. Only the forecasters that meet this criterion will be taken into account in the next period.

3 Data

We use growth, inflation, unemployment, ex-, and import changes forecasts of 14 different forecasting institutions that have dominated German macroeconomic forecasting for a long time. Some institutions regularly provide more than one major report for a given year (usually "spring" and "autumn"). In these cases, we take both forecasts into account. All in all, we use 17 forecasting reports. The sample runs from 1993 to 2019. For more details on the dataset, compare the Data appendix.

In this paper, we refer to the "forecasting season" (see Fig. 2), which is usual in Germany. Therefore, we attribute some forecasts published in the recent year as

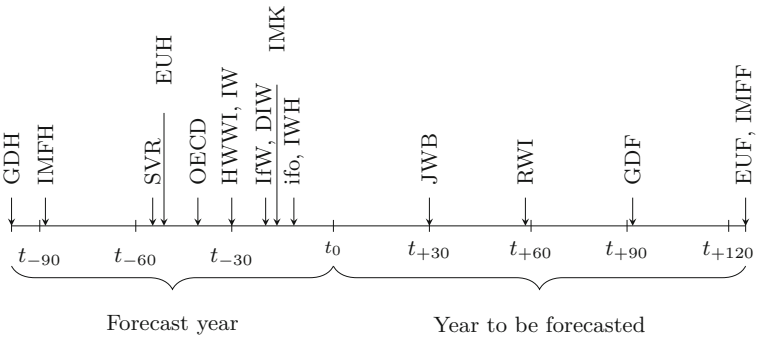


Fig. 2 The "Forecasting season"for 2019 in Germany. *Source:* Own compilation from the forecasting reports. Legend: GDH: Joint diagnosis, autumn; GDF: Joint diagnosis, spring; SVR: Council of Economic Experts; IfW: Kiel Institute; DIW: Berlin Institute, HWWI: Hamburg Institute; ifo: Munich Institute; RWI: Essen Institute; IW: Cologne Institute; IMK: Düsseldorf Institute, EUH: European Commission, autumn, EUF: European Commission, spring, JWB: Governments' Economic Report, IMFF: International Monetary Fund, spring, IMFH: International Monetary Fund, autumn

"predictions." Consequently, some forecasts made in period t are labelled as made in $t-1$ in the following because they have been made between January and April.

The growth forecast is the predicted rate of change of real GDP. Regarding inflation, we use the predicted change of the consumer price index or—if it is not available—the deflator of private consumption. Other forecast values that are being investigated are the unemployed rate and the rate of change of real exports and imports. In the case of interval forecasts, we use the simple average of the upper and lower bound of the interval.

4 Empirical results

We organize our results[6] by the length of the analyzed period, starting with one year as an example, turning to the correlation of rankings for two subsequent years, and discussing a possible long-run superiority for an institution in the full sample.

4.1 Rankings for just one year: 2019 as an example

We start with one year—2019—as an example. This procedure relates to most of the forecasting competitions in the media that have been analyzed, for example, by Döhrn (2015). Table 2 starts with the arguably most popular measure—the Mean Absolute Error. For each of the five variables under investigation, we report a ranking. These rankings differ substantially across the variables: The institution with, say, the best growth forecast is not necessarily the one with the best inflation or unemployment prediction. This result comes as a slight surprise since several papers (see, for example, Casey 2020) suggest that forecasters rely on prominent relationships between

[6] We use R Core Team (2021) for the calculations in this paper. Among others, we use the packages of Robin et al. (2011), Dowle and Srinivasan (2021), Conigrave (2020), and Hyndman and Khandakar (2008).

Table 2 Ranking of forecast quality based on the mean absolute errors and multi-dimensional criteria for 2019

	Mean Absolute Error					Multidimensional criteria		
	Growth	Inflation	Unemployment	Export	Import	City_Block	Euclidean	Mahalanobis
Joint diag., autumn	16.5	11.5	13	17	16	17	15	10
Joint diag., spring	2.5	3.5	13	6.5	15	5	5	4
Council of Econ. Exp.	9	14	13	10.5	10.5	14	13	12
Kiel Institute	14.5	14	13	14	14	16	16	13
Berlin Institute	11	11.5	13	6.5	5	13	11	9
Hamburg Institute	11	14	2	2	4	7	12	16
Munich Institute	5	16.5	5	10.5	8	12	14	14
Essen Institute	7.5	1	13	15.5	17	6	7	8
Halle Institute	7.5	5.5	17	13	10.5	9	6	5
OECD	11	16.5	13	8.5	12	15	17	15
Cologne Institute	6	8	2	3	1.5	3	4	7
Düsseldorf Institute	13	8	5	15.5	13	11	9	6
Governments Ec. Rep.	4	5.5	5	4.5	6.5	4	3	2
Europ. Com., autumn	14.5	10	2	8.5	9	8	10	11
Europ. Com., spring	1	3.5	7	1	1.5	1	1	1
IMF, autumn	16.5	8	8.5	12	6.5	10	8	17
IMF, spring	2.5	2	8.5	4.5	3	2	2	3

Notes: Own calculations

variables, like Okun's law or the Phillips curve, in making their forecasts. Furthermore, this finding calls for a ranking that considers all variables to figure out the best forecaster.

Therefore, the table also comprises the multi-dimensional methods—namely the City Block Metric, the Euclidean Distance, and the Mahalanobis Distance. While rankings based on the City Block Metric and the Euclidean Distance are pretty similar, there are some differences to the ranking based on the Mahalanobis Distance. Nevertheless, as regards the winner for 2019 the multi-dimensional methods point to the same institution: In all cases, the spring forecast of the European Commission ranks on top.

Table 3 further underlines the difficulties in creating a unique leaderboard for a given year. The exhibit compares the rankings based on the Spearman Rank Correlation Coefficient and Kendall's Tau. The coefficients for the variables differ, in some cases, considerably. The correlation between two consecutive rankings is often low and not significantly different from zero. In one case (remarkably, unemployment and inflation), it is even slightly negative.

4.2 Are the rankings stable over time?

Based on a similar exercise for one year only, Döhrn (2015) considered the possibility that forecasting competition results are "purely random." Therefore, we will look at yearly rankings and how they change from one year to the other. For this purpose, we have calculated rankings as in the previous section for each year from 1993 to 2019 and the Spearman Rank correlation of the rankings of two consecutive years. The result of this task is in Fig. 3.

If a particular group of forecasters would be regularly better than their competitors, the rank correlation coefficient should be (significantly) positive for most years. For any of the rankings relying on just one prediction, this does not appear to be the case. The Spearman Rank Correlation Coefficient is significantly positive for the multi-dimensional methods, only for a few years.

Instead, the correlations are about as often positive as negative, and the confidence bands regularly contain a zero correlation. This suggests that it is not possible using the ranking for a specific year to guess the forecasters you should listen to in the following year.

4.3 Has any institution superior forecasting skills in the long-run?

4.3.1 Rankings based on the total sample

For the following rankings, we consider the entire sample: 27 observations covering the period from 1993 to 2019 and calculate the measures already used in the previous sections. Additionally, we include now the Root Mean Squared Error since, in the case of more than one observation, a different ranking than according to the Mean Absolute Error may occur. Also, it is now possible to include direction-of-change measures. We use the specificity and the Area under the Receiver Operating Curve

Table 3 Spearman (1906) (lower triangular) and Kendall (1938) (upper triangular) rank correlation coefficients across different ranking criteria for 2019

	Mean absolute error for...					City Block Metric	Euclidean distance
	Growth	Inflation	Unemployment	Exports	Imports		
Mean absol. error for...							
Growth	(-)	0.337	0.017	0.431*	0.305	0.581**	0.506**
Inflation	0.504**	(-)	−0.069	0.086	0.039	0.550**	0.749***
Unemployment	0.072	−0.062	(-)	0.387	0.409	0.330	0.149
Export	0.597**	0.133	0.490**	(-)	0.654***	0.502**	0.352
Import	0.370	0.064	0.570**	0.813***	(-)	0.454*	0.320
City Block Metric	0.728***	0.756***	0.445*	0.671***	0.551**	(-)	0.779***
Euclidean Dist.	0.679***	0.882***	0.232	0.493**	0.470*	0.919***	(-)
Mahalanobis Dist.	0.686***	0.775***	−0.036	0.285	0.152	0.647***	0.779***

Source: own calculation. *(**,***) denotes rejection of the null hypothesis of a zero rank correlation coefficient at the 10 (5, 1) % level)

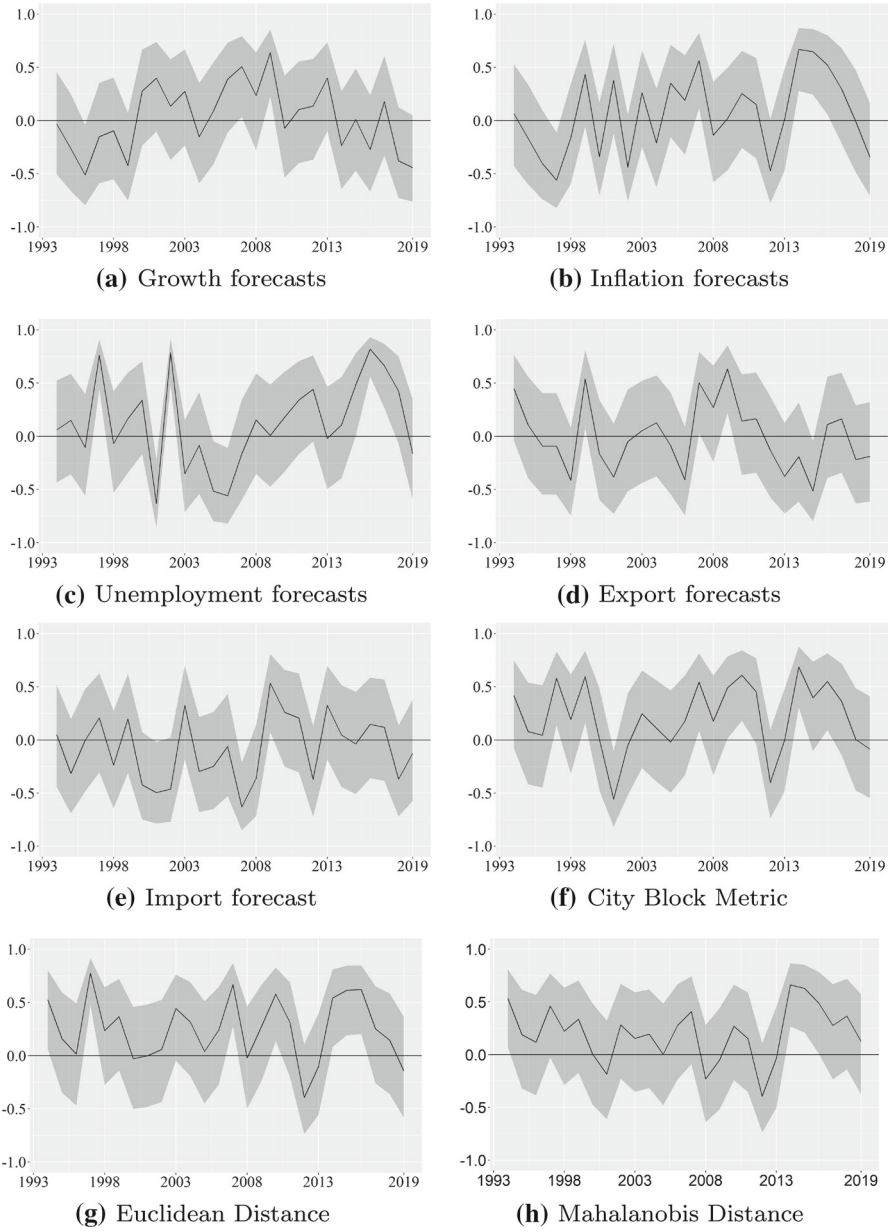


Fig. 3 Rank correlation coefficients for two subsequent years by ranking method, 1993 to 2019 . *Notes:* Source: Our own calculation. Shaded areas represent a 95% confidence interval calculated by $\tanh(\arctan r \pm 1.96/\sqrt{n-3})$ with r as the empirical Spearman rank correlation coefficient and n as number of observations

(AUROC). Furthermore, following Timmermann and Zhu (2017), we add the Diebold and Mariano (1995) statistic to rank the forecast.

Our results for the analysis of numerical forecast errors are in Table 4. For the sake of brevity, we only report the findings for growth and inflation and leave the respective statistics for other variables in the appendix (Tables 9 to 12 report the rankings, while Tables 13 and 14 show the numerical results for the respective criterion).

As becomes apparent, in contrast to the results based on the somewhat limited information of just one year, the rankings seem to be closer to each other. Still, there are some significant deviations from this rule. For example, the Essen Institute ranks relatively good regarding the predictions for growth, inflation, and unemployment, while more at the bottom for exports and imports. Rankings resulting from the three multi-dimensional criteria are also very similar.

Table 4 also reveals that the rankings based on specificity and AUROC differ in some cases. However, as is shown in Appendix Table 13, the measures are quite close to each other. The rankings for the Mean Absolute Error, the Root Mean Square Error, and the Diebold and Lopez (1996)-statistic are very alike. However, when comparing the rankings based on numerical forecast errors with the ones based on directional change analyses, a few differences in the rankings become apparent, because the direction-of-change measures show a smaller variance across institutions, i.e., several institutions show the same rank number.

Table 5 contains the Spearman Correlation Coefficients and Kendalls's Tau for the rankings methods based on the total sample.[7] We find that the correlations among numerical criteria are generally stronger than the correlations within direction-of-change measures. However, the correlations are always positive, generally much higher than in the case for just one year, and in almost every case, significantly different from zero at standard significance levels.

4.3.2 Measures of long-run superiority

After demonstrating in Sect. 4.2 that rankings are not stable over two consecutive years, it is still possible that certain institutions are significantly more often better than the average in the long-run. To analyze this problem, we calculate the statistics proposed by Stekler (1987) and Sinclair et al. (2016) for the full sample.

Figure 4 shows the development of the total cumulative rank-sum based on the Absolute Forecast Errors for each variable and the multi-dimensional methods. The last value of each time series represents the test statistics proposed by Stekler (Equation 10). The black diagonal line shows the rank-sum for each year expected for a forecaster that always makes the median rank. The vertical black line represents two standard deviations. In the exhibit, we can identify some institutions that perform significantly better or worse than the average, which is visible in the case of the multi-dimensional criteria. In the case of the one-dimensional criteria, in particular for ex- and imports, one has the impression of a relatively similar long-run performance. Appendix Table 15 also confirms this finding since it shows the rank correlation coef-

[7] For the sake of brevity, we left out some variables we have considered for the ranking based on just one year. These numbers are available upon request from authors.

Table 4 Rankings based on the full sample, 1993 to 2019

	Numerical forecast errors						Directional change				Multivariate criteria		
	MAE, growth	MAE, Inflation	RMSE, growth	RMSE, inflation	D/M Stat., growth	D/M Stat., inflation	AUROC, growth	AUROC, inflation	Specificity, growth	Specificity, inflation	City-Block metric	Euclidean Distance	Mahabilonis Distance
Joint diagnosis, autumn	15	17	15	17	15	16	11.5	12.5	15	15.5	16	15	15
Joint diagnosis, spring	2	1.5	1	3	2	1	6.5	4	4.5	5	1	1	1
Council of Economic Experts	13	14	13	14	14	14	13.5	12.5	10	15.5	13	14	14
Kiel Institute	9	7	7	9	6	11	3.5	9.5	12.5	11	7	7	7
Berlin Institute	8	12.5	11	10	11	12	10	15	7.5	9	11	11	12
Hamburg Institute	7	11	10	11	8	10	13.5	9.5	10	11	10	10	11
Munich Institute	5	8	5	8	5	9	1	5	2	7	5	5	6
Essen Institute	4	3	4	5	4	4	6.5	2.5	4.5	2.5	6	6	3
Halle Institute	10	12.5	8	13	9	13	2	12.5	1	15.5	8	8	10
OECD	12	9	12	7	12	8	9	8	10	8	14	13	9
Cologne Institute	16	15	16	15	17	15	16	17	15	11	15	16	17
Düsseldorf Institute	11	6	9	6	10	6	3.5	6.5	12.5	5	9	9	8
Governments Economic Report	6	4	6	4	7	3	15	6.5	7.5	5	4	4	4
European Com., autumn	14	10	14	12	13	7	11.5	12.5	15	15.5	12	12	13
European Com., spring	1	5	2	1	1	5	6.5	2.5	4.5	2.5	2	2	5
IMF, autumn	17	16	17	16	16	17	17	16	17	13	17	17	16
IMF, Spring	3	1.5	3	2	3	2	6.5	1	4.5	1	3	3	2

Notes: Own calculations

Table 5 Spearman (1906) (lower triangular) and Kendall (1938) (upper triangular) rank correlation coefficients across different ranking criteria based on the full sample (1993 to 2019)

	MAE, growth	MAE, inflation	RMSE, growth	RMSE, inflation	D/M Stat, growth	D/M Stat., inflation	AUROC, growth	AUROC, inflation	Specificity, growth	Specificity, inflation	City Block Metric	Euclidean Distance
MAE, growth	(-)	0.667***	0.897***	0.706***	0.897***	0.632***	0.373	0.708***	0.668***	0.611***	0.794***	0.824***
MAE, Inflation	0.833***	(-)	0.726***	0.830***	0.696***	0.859***	0.353	0.759***	0.430*	0.709***	0.711***	0.711***
RMSE, growth	0.961***	0.890***	(-)	0.721***	0.912***	0.647***	0.464*	0.738***	0.637***	0.580**	0.897***	0.926***
RMSE, Inflation	0.870***	0.950***	0.882***	(-)	0.721***	0.809***	0.342	0.753***	0.481*	0.781***	0.735***	0.735***
D/M Stat., growth	0.968***	0.872***	0.983***	0.870***	(-)	0.618***	0.449*	0.753***	0.590**	0.564**	0.838***	0.868***
D/M Stat., Inflation	0.787***	0.960***	0.811***	0.926***	0.792***	(-)	0.266	0.708***	0.404	0.642***	0.691***	0.691***
AUROC, growth	0.518**	0.492**	0.647***	0.462*	0.625***	0.371	(-)	0.378	0.514**	0.256	0.418*	0.449***
AUROC, Inflation	0.850***	0.905***	0.887***	0.899***	0.886***	0.866***	0.553**	(-)	0.506**	0.712***	0.647***	0.677***
Specificity, growth	0.811***	0.570**	0.813***	0.610***	0.767***	0.526**	0.633***	0.635***	(-)	0.433*	0.590**	0.590**
Specificity, Inflation	0.782***	0.851***	0.769***	0.913***	0.751***	0.805***	0.342	0.836***	0.510**	(-)	0.549**	0.549**
City Block	0.939***	0.885***	0.978***	0.868***	0.958***	0.824***	0.582**	0.834***	0.773***	0.730***	(-)	0.971***
Euclidean Dis.	0.944***	0.892***	0.983***	0.880***	0.968***	0.836***	0.604**	0.856***	0.773***	0.738***	0.995***	(-)
Mahalanobis Dis.	0.900***	0.961***	0.956***	0.924***	0.934***	0.897***	0.587***	0.927***	0.717***	0.821***	0.926***	0.944***

Sources: Own calculations. *,(**,***) denotes rejection of the null hypothesis of a zero correlation coefficient at the 10 (5,1) % level

ficients for the methods to be positive, relatively high and statistically different from zero.

Table 6 contains the Stekler statistic, i.e., the final (2019) cumulative rank-sums for each institution and method. Compared to the results in Table 2, which lists the rankings for 2019, the results suggest more essential differences in the forecasting quality since many institutions show values outside the two standard error bands. Hence, in the long-run, some institutions seem to be superior to the average forecaster.

As mentioned above, however, the forecasters under investigation provide their predictions typically on diverging dates. Appendix Table 16 provides an overview of all forecasters and shows the average forecast horizon in days and its standard deviation. Since the ranking deals with fixed-event rather than with fixed-horizon forecasts (see Knüppel and Vladu 2016, for these concepts), the differing forecast horizons will influence accuracy. Figure 5 shows the relation between the average forecast horizon and the rank-sums based on the four selected measures. The ranks decline, signalling a better long-run relative performance with a shorter forecast horizon. The correlation is pretty strongly negative, and only a few deviations from the estimated trend are visible.

4.3.3 Selecting successful forecasters

Finally, we adopt the Bürgi and Sinclair (2017) approach outlined above to select successful forecasters for a given year. To this end, we refer to the last five years and demand that an institution should have been better than the average at least half of the time, i.e., the percentage threshold in our case is 50 %. Because—for example, in case of a pronounced downturn—the advantage of forecasting lately is significant, we restricted the sample to institutions with an average forecast horizon of more than 300 days. Figure 6 shows the results. The group of successful forecasters, according to the described rule, changes over time. In some years, only five of the 14 possible forecasts are selected by the procedure. Remarkably the group of forecasters selected by the approach changes quite often, and there is no institution that is in it for the total sample.

4.4 Forecast horizons and real-time data

The most recently available revised data can be compared with the first published data ("real-time data") to determine the result. It is even possible that forecasters aim at different targets, i.e., one institution tries to predict the last available data vintage, while another institution seeks to anticipate real-time data (Clements 2019). To assess whether selecting one of these databases affects the ranking, we compared rankings based on the Mean Absolute Error and either database in Table 7. The results hardly show any variations in the rank order, suggesting that the data vintage does not matter very much in ranking German forecasters.

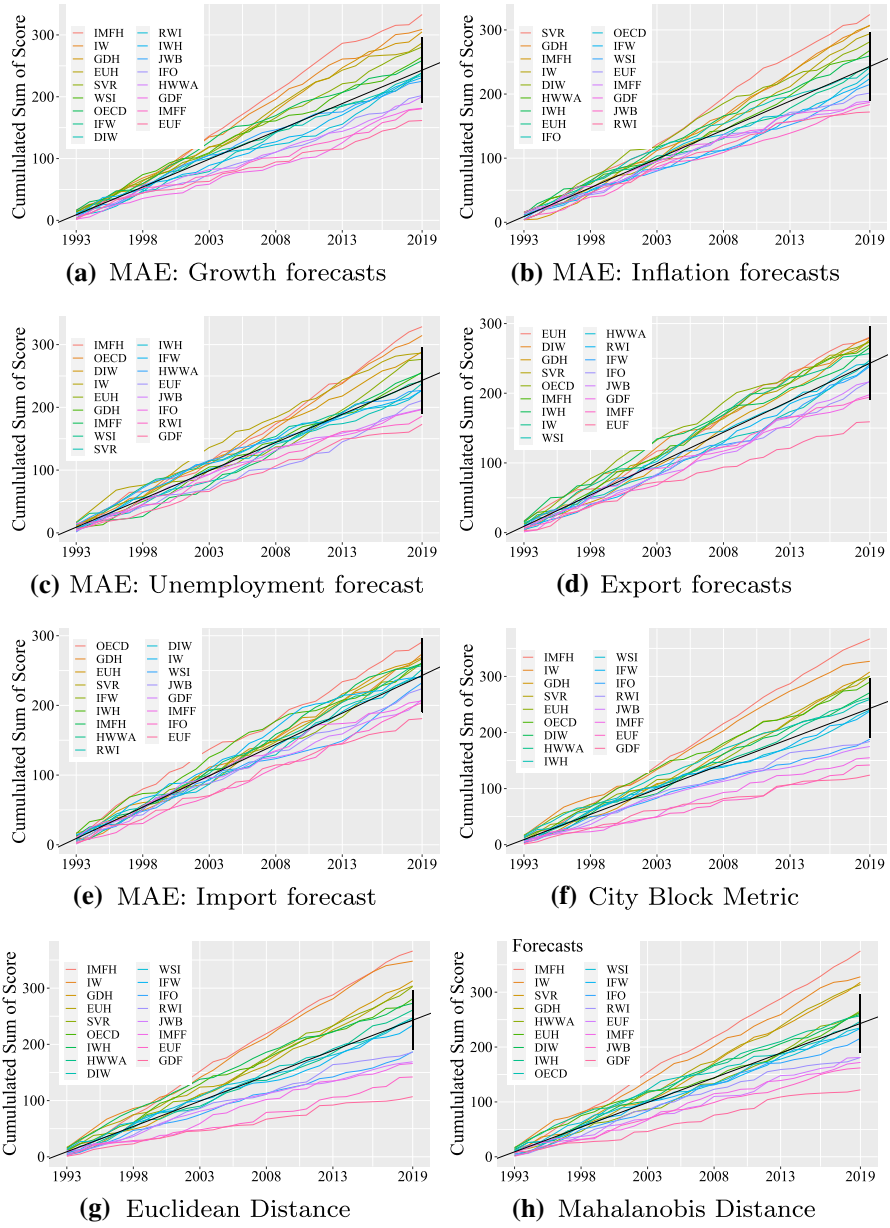


Fig. 4 The evolution of cumulated rank-sums based on the Absolute Forecast Errors and multi-dimensional criteria, 1993 to 2019 . Source: Our own calculation. The black line represents the expected development for a forecaster with a median rank in each period. The vertical black line represents two standard deviations. Legend: GDH: Joint diagnosis, autumn; GDF: Joint diagnosis, spring; SVR: Council of Economic Experts; IfW: Kiel Institute; DIW: Berlin Institute, HWWI: Hamburg Institute; ifo: Munich Institute; RWI: Essen Institute; IW: Cologne Institute; IMK: Düsseldorf Institute, EUH: European Commission, autumn, EUF: European Commission, spring, JWB: Governments' Economic Report, IMFF: International Monetary Fund, spring, IMFH: International Monetary Fund, autumn

Table 6 Stekler Statistic, 1993 to 2019

	Growth	Inflation	Unemployment	Export	Import	City_Block	Euclidean	Mahalanobis
Joint diagnosis, autumn	305*	307.5*	256.5	276.5	275	308*	314*	314*
	(15)	(17)	(12)	(14)	(16)	(15)	(15)	(15)
Joint diagnosis, spring	181.5*	189*	173.5*	197.5	209	125*	108*	124*
	(3)	(3)	(1)	(2.5)	(3)	(1)	(1)	(1)
Council of Economic Experts	280.5	301*	234	258	241.5	289	289	296*
	(13)	(15)	(9)	(10)	(7)	(12)	(13)	(14)
Kiel Institute	239	224	227.5	239	261.5	238	234	236
	(10)	(7)	(7)	(6)	(13)	(8)	(7)	(7)
Berlin Institute	238.5	283.5	289	280	245.5	269	247	263
	(9)	(13)	(15)	(17)	(9)	(11)	(9)	(11)
Hamburg Institute	200	267.5	227	246.5	258.5	264	260	265
	(4)	(12)	(6)	(8)	(11)	(10)	(10)	(13)
Munich Institute	202	242.5	196.5	214	204.5	188*	190*	218
	(5)	(9)	(3)	(4)	(2)	(6)	(6)	(6)
Essen Institute	234.5	174.5*	186.5*	244.5	253.5	186*	188*	183*
	(8)	(1)	(2)	(7)	(10)	(5)	(5)	(5)
Halle Institute	231	261	229	266.5	264.5	262	276	257
	(7)	(11)	(8)	(12)	(14)	(9)	(11)	(10)
OECD	258.5	234	315*	277	290.5	293	282	249
	(11)	(8)	(16)	(15)	(17)	(13)	(12)	(9)
Cologne Institute	309*	299	286.5	260.5	243.5	327*	350*	332*
	(16)	(14)	(14)	(11)	(8)	(16)	(16)	(16)

Table 6 continued

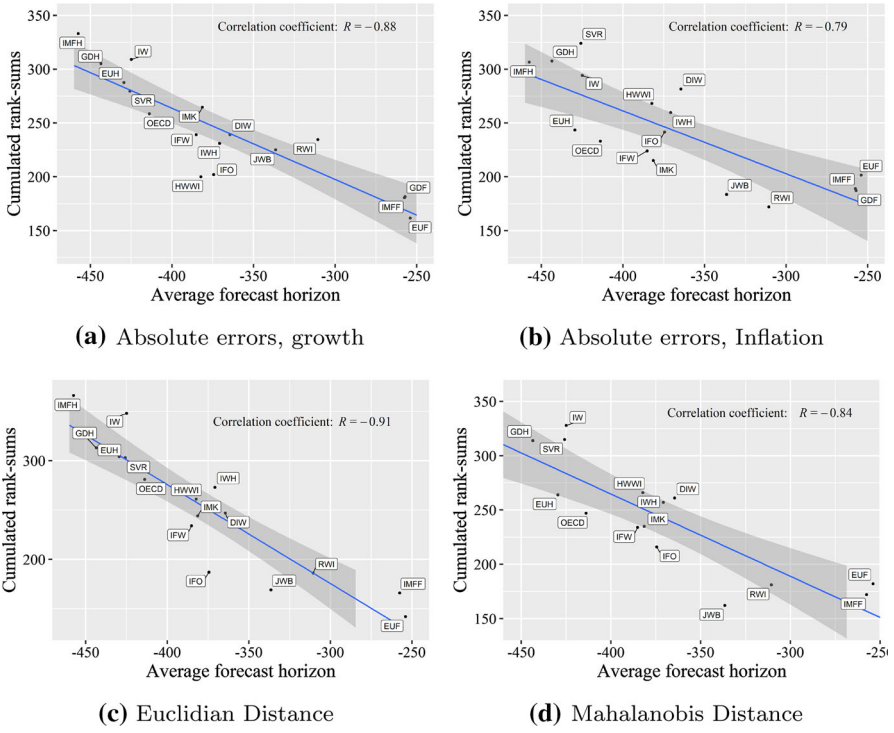
	Growth	Inflation	Unemployment	Export	Import	City_Block	Euclidean	Mahalanobis
Düsseldorf Institute	264.5	218	241.5	247.5	233.5	237	240	237
	(12)	(6)	(10)	(9)	(6)	(7)	(8)	(8)
Governments Economic Report	225	185*	197.5	216	226	180*	171*	164*
	(6)	(2)	(4)	(4)	(5)	(4)	(4)	(2)
European Comm., autumn	287	245.5	276.5	279.5	270	300*	306*	264
	(14)	(10)	(13)	(16)	(15)	(14)	(14)	(12)
European Comm., spring	161.5*	201.5	211	160*	185*	143*	141*	180*
	(1)	(5)	(5)	(1)	(1)	(2)	(2)	(4)
Int. Monetary Fund, autumn	333*	307*	328.5*	270.5	260.5	368*	369*	375*
	(17)	(16)	(17)	(13)	(12)	(17)	(17)	(17)
Int. Monetary Fund, Spring	180.5*	191.5*	255	197.5	209.5	154*	166*	174*
	(2)	(4)	(11)	(2.5)	(4)	(3)	(3)	(3)

Own calculations. The upper number shows the Stekler (1987) statistic, the expected sum under the null hypothesis 243 +- 26.2*2. The lower number in parentheses shows the rank according to this criterion. A asterisk (*) denotes that the number is outside a two standard deviation band

Table 7 Average rank based on mean absolute error and real-time versus final data, 1993 to 2019

Institution	GDP growth		Inflation	
	Based on real-time data	Based on final data	Based on real-time data	Based on final data
Joint diagnosis, autumn	11.3	11.8	11.4	10.9
Joint diagnosis, spring	6.7	6.0	7.0	6.4
Council of ec. experts	10.4	10.6	11.1	10.4
Kiel Institute	8.9	9.2	8.3	8.7
Berlin Institute	8.8	8.3	10.5	10.5
Hamburg Institute	7.4	8.1	9.9	9.1
Munich Institute	7.5	6.8	9.0	8.7
Essen Institute	8.7	8.9	6.5	7.4
Halle Institute	8.6	8.0	9.7	9.6
OECD	9.6	9.3	8.7	10.0
Cologne Institute	11.4	12.1	11.1	10.8
Düsseldorf Institute	9.8	8.3	8.1	8.8
Governments economic report	8.3	8.0	6.9	8.4
European comm., autumn	10.6	11.0	9.1	9.4
European comm., spring	6.0	6.5	7.4	7.2
IMF, autumn	12.3	12.9	11.4	10.8
IMF, spring	6.7	7.3	7.1	6.0

Our own calculations



(a) Absolute errors, growth **(b)** Absolute errors, Inflation

(c) Euclidian Distance **(d)** Mahalanobis Distance

Fig. 5 Forecast horizon and forecast performance . Source: Own calculation. Legend: GDH: Joint diagnosis, autumn; GDF: Joint diagnosis, spring; SVR: Council of Economic Experts; IfW: Kiel Institute; DIW: Berlin Institute, HWWI: Hamburg Institute; ifo: Munich Institute; RWI: Essen Institute; IW: Cologne Institute; IMK: Düsseldorf Institute, EUH: European Commission, autumn, EUF: European Commission, spring, JWB: Governments' Economic Report, IMFF: International Monetary Fund, spring, IMFH: International Monetary Fund, autumn

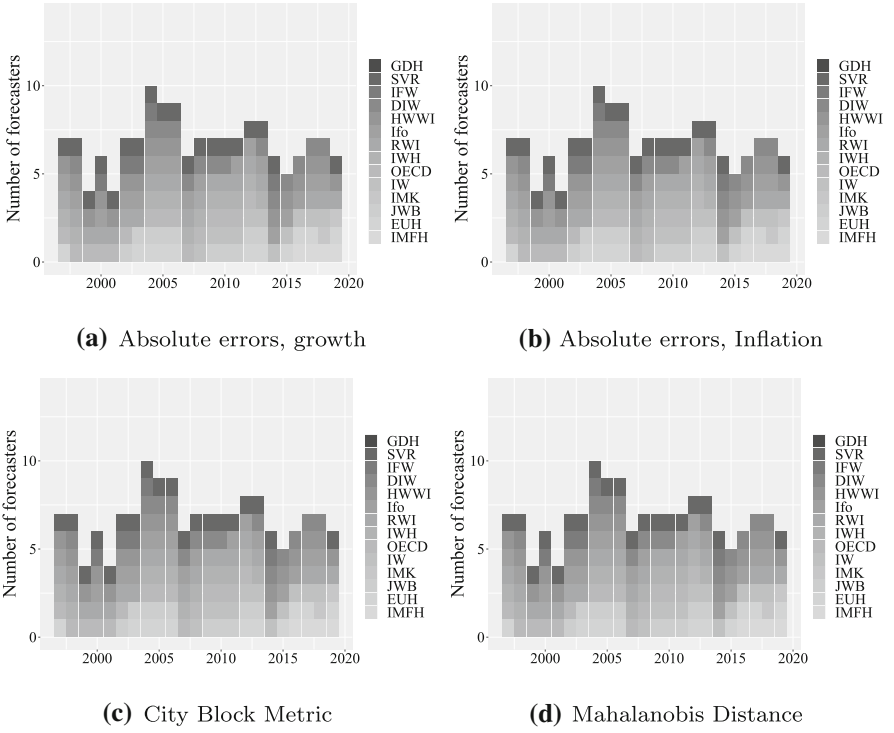
4.5 Rankings and recessions

Some papers (for example, Dovern and Jannsen 2017) have shown that the magnitude of forecast errors crucially depends on the phase of the business cycle in which the forecasts have been made. This fact raises the question of whether the relative position of the institute may also rely on that. For example, a certain institute may be a "recession specialist," while others may be better in regular times. To check this idea with our dataset, we have restricted ourselves to institutes with a forecast horizon of at least 300 days and calculated the average rank for recession and non-recession years. In this context, we define a recession year as one in which real GDP shrunk, i.e., the growth rate of real GDP was negative. Table 8 shows the results of this task. For the sake of brevity, we consider just one one-dimensional (the Mean Absolute Forecast Error) and one multi-dimensional measure (the Mahalanobis (1936) distance) of accuracy. The differences between the business cycle phases are generally rather small and seem to be not systematic. Again, the forecast horizon is the prime suspect in explaining the changing differences between the forecasters.

Table 8 Average rank in recession and non-recession years

Institution	Absolute forecast error		Mahabilonis distance	
	In recession years	In non-recession years	In recession years	In non-recession years
Joint diagnosis, autumn	9.2	6.3	8.0	6.3
Council of ec. experts	5.5	6.2	4.7	6.7
Kiel Institute	4.8	5.1	3.3	4.9
Berlin Institute	2.8	5.3	4.7	5.5
Hamburg Institute	4.0	4.4	3.3	5.8
Munich Institute	4.5	4.2	5.3	4.5
Essen Institute	2.2	5.5	2.0	3.8
Halle Institute	5.8	4.8	6.0	4.8
European comm., autumn	6.8	6.1	9.3	4.7
IMF, autumn	9.3	7.0	8.3	8.0

Our own calculations



(a) Absolute errors, growth

(b) Absolute errors, Inflation

(c) City Block Metric

(d) Mahalanobis Distance

Fig. 6 Selecting successful forecasters by different criteria . Source: Own calculation. Legend: GDH: Joint diagnosis, autumn; GDF: Joint diagnosis, spring; SVR: Council of Economic Experts; IfW: Kiel Institute; DIW: Berlin Institute, HWWI: Hamburg Institute; ifo: Munich Institute; RWI: Essen Institute; IW: Cologne Institute; IMK: Düsseldorf Institute, EUH: European Commission, autumn, EUF: European Commission, spring, JWB: Governments' Economic Report, IMFF: International Monetary Fund, spring, IMFH: International Monetary Fund, autumn

5 Conclusion

We have turned not literally all but quite a few stones to rank the quality of German macroeconomic forecasts. To this end, we refer to the arguably "leading" institutions that provide predictions for the German economy. Using annual data from 1993 to 2019 for predictions of growth, inflation, the unemployment rate, and ex- and import changes, we base our analyses on 17 different forecasts from 13 separate institutions. To these data, we apply a variety of criteria to rank these institutions according to their predictive ability. We consider different horizons for the comparisons, from just one year over several subsequent years to the full sample.

To rank the forecasts, we use, first, simple accuracy measures such as the Mean Absolute Error, the Root Mean Squared Error, and the Diebold and Mariano (1995) statistic. Second, we apply directional change statistics: the Area Under the Receiving Operating Curve and the specificity of the forecasts. Third, we consider rankings based on multi-dimensional criteria: the City-Block Metric, the Euclidean Distance, and the Mahalanobis Distance.

Our main findings suggest that—for just one year—rankings based on different methods vary widely and, in some cases, correlate only weakly with each other. In other words, rankings for growth, inflation, unemployment, and so forth do not point to the same winner. Hence, we see not much value in year-by-year forecasting competitions beyond the aspect of pure entertainment.

Furthermore, correlations of rankings calculated for two consecutive years and a given method are often relatively low and statistically insignificant. Therefore, an interested audience, in particular policymakers, cannot guess from the results from one year to whom they should listen for the next year.

Analyses based on the entire sample, however, show a much more stable pattern: First, rank correlations between institutions are generally relatively high among different criteria. Also, we find substantial long-run differences in forecasting quality as reflected, for example, by the cumulative rank-sums. However, further inspections suggest that these differences are mostly due to distinct average forecast horizons. Third, we report that the rank correlations across the several ranking methods are, in the long-run, quite high and statistically different from zero, which implies that the choice of the criterion to rank the forecasters is of minor importance. The same does hold for using real-time vs finally revised data-sets. We also find no large differences in the rankings based on recession years and normal periods.

All in all, on the substantial side, we are not able to single out an institution that is superior in predicting the German economy. On the methodological side, we find only minor differences across the ranking methods.

Further research may try to broaden the database of the investigation. In particular, privately financed institutions may show different behavior as compared to the forecasters included in this study. Also, a higher data frequency may render it possible to find differences across forecasters in adjusting to new information.

Acknowledgements The authors thank Lars Tegtmeier, Karsten Müller, Ulrich Fritsche, seminar participants at the 40th International Symposium on Forecasting, and an anonymous referee for helpful comments on a previous version of this paper. We thank the *Deutsche Forschungsgemeinschaft* (DFG) for financial support (projects FR 2677-4-1 and FR 2677-4-2 within the DFG Priority Program 1859 *Experience and Expectation: Historical Foundations of Economic Behavior*).

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

To compile our dataset, we have to make a couple of additional assumptions, we list below:

- In case of interval forecasts, we take the average between the upper and lower bound of the interval.
- The growth forecast is the predicted rate of change in real GNP (1983–1989) and in real GDP (for all other years). In doing so, we follow the headline figure of the Statistical Office for the respective year. Note that, frequently, the forecasts refer to growth rather than to either GDP or GNP. In these cases, we assume that the forecasters made no distinction between the concepts and had the same forecast for both figures.
- Also, in some cases, the forecast report included no explicit reference on whether a mentioned inflation forecast referred to the consumption deflator or the CPI. In these cases, we assume that the forecaster did not wish to make a difference between the two concepts.
- The Essen and the Düsseldorf Institute refer to West Germany in 1993.
- In one case, the trade-union financed institute, we have combined two institutes to one. Up to 2004, the macroeconomic forecasts have been provided by the WSI, until then the IMK was responsible for the prediction in behalf of the trade unions. This is motivated by the fact that both institutes are formally departments of the Hans-Böckler-Stiftung.
- In a similar vein, we have treated the (privately financed) HWWI Institute in Hamburg as the successor of the former HWWA Institute (financed by public money).
- For the inflation forecast, we use the predicted change in the deflator of private consumption when this figure was available. In some cases, however, the forecast report included no explicit reference on whether a mentioned inflation forecast refers to the consumption deflator or the CPI. In such cases, we assume that no distinction between the figures was intended by the forecaster and use the available inflation forecast.
- The subset still contained isolated missing forecast values. For these values, we have assumed the average absolute forecast error calculated from the other forecast errors of the respective report. Missing forecast values concerned the number of unemployment for the Cologne Institute for the years from 1994 to 1996 and unemployment, export, import forecasts of the Düsseldorf Institute for the year 2005.
- For the European Commission, we also observe a pause in publishing forecast in 1997. In this case, we refer to forecasts with a two-year-forecast-horizon made in 1996.
- Some forecasters predict the unemployment rate according to the ILO definition, some aim at the unemployment rate according to the national definition. Thus, for the OECD, the forecast error is calculated based on ILO data. See Tables 9, 10, 11, 12, 13, 14, 15, 16

Table 9 Ranking based on numerical forecast errors, 1993 to 2019

	Mean absolute error					Root Mean squared error					Diebold Mariano statistic				
	Growth	Inflation	Unemployment	Export	Import	Growth	Inflation	Unemployment	Export	Import	Growth	Inflation	Unemployment	Export	Import
Joint diag., autumn	15	17	10	15	14	15	17	12	13	13	15	16	13	13	13
Joint diag., spring	2	1.5	1	2	4	1	3	1	2	3	2	1	1	4	5
Council Econ. Exp.	13	14	8	11	9	13	14	8	12	11	14	14	10	12	9
Kiel Institute	9	7	7	7	7	7	9	7	4	4	6	11	7	5	12
Berlin Institute	8	12.5	14	12	13	11	10	10	9	10	11	12	12	10	11
Hamburg Institute	7	11	6	10	12	10	11	6	8	7	8	10	6	8	8
Munich Institute	5	8	4	5	3	5	8	3	3	5	5	9	4	6	6
Essen Institute	4	3	3	13	15	4	5	5	16	16	4	4	2	16	15
Halle Institute	10	12.5	9	6	2	8	13	9	6	2	9	13	11	17	17
OECD	12	9	16	16	17	12	7	16	14	15	12	8	15	15	16
Cologne Institute	16	15	12	14	11	16	15	15	15	14	17	15	14	14	10
Düsseldorf Institut	11	6	13	9	8	9	6	13	10	9	10	6	9	7	7
Govern. Econ. Rep.	6	4	5	4	5	6	4	4	7	6	7	3	5	3	2
Eur. Com., autumn	14	10	15	8	10	14	12	14	11	8	13	7	17	11	14
Eur. Com., spring	1	5	2	1	1	2	1	2	1	1	1	5	3	2	3
IMF, autumn	17	16	17	17	16	17	16	17	17	17	16	17	16	9	4
IMF, spring	3	1.5	11	3	6	3	2	11	5	12	3	2	8	1	1

Source: Own calculations

Table 10 Ranking based on directional change, 1993 to 2019

	Specificity					Area under the ROC curve				
	Growth	Inflation	Unemployment	Export	Import	Growth	Inflation	Unemployment	Export	Import
Joint diag, autumn	15	15.5	11	8	9	11.5	12.5	17	7.5	14
Joint diag, spring	4.5	5	1	13	1	6.5	4	2	14	1
Council of Econ. Exp.	10	15.5	11	3.5	9	13.5	12.5	3	11.5	14
Kiel Institute	12.5	11	8.5	8	12.5	3.5	9.5	6	1	12
Berlin Institute	7.5	9	14	16.5	4	10	15	7	17	7
Hamburg Institute	10	11	3.5	8	11	13.5	9.5	11	7.5	6
Munich Institute	2	7	5.5	13	4	1	5	4.5	4.5	3.5
Essen Institute	4.5	2.5	5.5	13	4	6.5	2.5	4.5	14	3.5
Halle Institute	1	15.5	2	16.5	17	2	12.5	15	10	8
OECD	10	8	15	8	12.5	9	8	12	7.5	17
Cologne Institute	15	11	7	1	7	16	17	16	16	16
Düsseldorf Institut	12.5	5	8.5	8	9	3.5	6.5	10	7.5	14
Govern. Econ. Rep.	7.5	5	3.5	13	15	15	6.5	1	14	10
Eur. Com., autumn	15	15.5	16	13	15	11.5	12.5	13	4.5	10
Eur. Com., spring	4.5	2.5	13	3.5	4	6.5	2.5	8	2.5	3.5
IMF, autumn	17	13	17	3.5	15	17	16	14	2.5	10
IMF, spring	4.5	1	11	3.5	4	6.5	1	9	11.5	3.5

Source: Own calculations

Table 11 Numerical forecast errors, 1993 to 2019

	Mean absolute error					Root Mean squared error					Diebold Mariano statistic				
	Growth	Inflation	Unemployment	Export	Import	Growth	Inflation	Unemployment	Export	Import	Growth	Inflation	Unemployment	Export	Import
Joint diagnosis, autumn	1.06	0.54	0.49	3.63	3.41	1.48	0.70	0.67	4.96	4.64	1.53	0.77	0.13	1.81	1.89
Joint diagnosis, spring	0.56	0.28	0.25	2.80	2.86	0.73	0.38	0.30	3.55	3.74	1.89	2.33	2.91	2.01	2.05
Council of Economic Experts	0.98	0.50	0.43	3.51	3.15	1.41	0.64	0.59	4.95	4.45	1.54	0.94	0.82	1.85	2.00
Kiel Institute	0.80	0.37	0.42	3.09	3.03	1.05	0.51	0.56	3.96	3.80	1.79	1.48	1.07	1.98	1.95
Berlin Institute	0.80	0.47	0.50	3.59	3.34	1.17	0.56	0.63	4.50	4.37	1.68	1.47	0.53	1.88	1.97
Hamburg Institute	0.79	0.46	0.41	3.34	3.26	1.13	0.58	0.55	4.41	4.25	1.75	1.48	1.47	1.91	2.01
Munich Institute	0.73	0.40	0.35	3.01	2.86	0.98	0.51	0.47	3.95	3.84	1.81	1.73	2.00	1.96	2.05
Essen Institute	0.69	0.29	0.34	3.60	3.48	0.89	0.41	0.47	5.54	4.96	1.82	2.21	2.36	1.63	1.86
Halle Institute	0.82	0.47	0.43	3.02	2.83	1.08	0.60	0.63	3.99	3.68	1.74	0.98	0.57	1.58	1.56
OECD	0.95	0.40	0.65	3.71	3.64	1.32	0.49	0.84	5.19	4.92	1.67	1.75	−0.62	1.77	1.78
Cologne Institute	1.07	0.50	0.49	3.61	3.25	1.53	0.65	0.74	5.36	4.75	1.44	0.81	−0.25	1.78	1.97
Düsseldorf Institut	0.86	0.34	0.50	3.30	3.05	1.13	0.46	0.69	4.56	4.32	1.73	1.87	0.84	1.94	2.02
Governments Economic Report	0.78	0.30	0.37	3.01	2.97	1.04	0.39	0.47	4.11	4.07	1.78	2.22	1.84	2.04	2.13
European Commission, autumn	1.01	0.42	0.54	3.30	3.22	1.44	0.59	0.71	4.79	4.28	1.54	1.80	−0.75	1.87	1.89
European Commission, spring	0.54	0.31	0.33	2.34	2.53	0.78	0.36	0.43	3.33	3.66	1.89	2.13	2.06	2.08	2.11
International Monetary Fund, autumn	1.22	0.52	0.78	3.90	3.59	1.68	0.66	1.08	5.66	5.18	1.45	0.65	−0.73	1.91	2.09
International Monetary Fund, spring	0.57	0.28	0.49	2.86	2.97	0.82	0.38	0.64	3.97	4.53	1.87	2.29	0.96	2.08	2.25

Source: Own calculations

Table 12 Ranking based on multivariate forecast errors, 1993 to 2019

	City block metric	Euclidean distance	Mahalanobis distance
Joint diagnosis, autumn	16	15	15
Joint diagnosis, spring	1	1	1
Council of Economic Experts	13	14	14
Kiel Institute	7	7	7
Berlin Institute	11	11	12
Hamburg Institute	10	10	11
Munich Institute	5	5	6
Essen Institute	6	6	3
Halle Institute	8	8	10
OECD	14	13	9
Cologne Institute	15	16	17
Düsseldorf Institut	9	9	8
Governments Economic Report	4	4	4
European Commission, autumn	12	12	13
European Commission, spring	2	2	5
International Monetary Fund, autumn	17	17	16
International Monetary Fund, spring	3	3	2

Source: Own calculations

Table 13 Measures of directional change accuracy, 1993 to 2019

	Specificity					Area under the ROC curve				
	Growth	Inflation	Unemployment	Export	Import	Growth	Inflation	Unemployment	Export	Import
Joint diagnosis, autumn	0.60	0.44	0.78	0.81	0.67	0.71	0.57	0.45	0.86	0.70
Joint diagnosis, spring	0.73	0.75	0.94	0.75	0.87	0.78	0.78	0.85	0.82	0.89
Council of Economic Experts	0.67	0.44	0.78	0.88	0.67	0.70	0.57	0.83	0.84	0.70
Kiel Institute	0.67	0.56	0.83	0.81	0.60	0.79	0.63	0.79	0.91	0.71
Berlin Institute	0.73	0.62	0.67	0.69	0.73	0.73	0.56	0.77	0.74	0.78
Hamburg Institute	0.67	0.56	0.89	0.81	0.67	0.70	0.63	0.69	0.86	0.79
Munich Institute	0.80	0.69	0.89	0.75	0.73	0.81	0.74	0.82	0.88	0.82
Essen Institute	0.73	0.81	0.89	0.75	0.73	0.78	0.81	0.82	0.82	0.82
Halle Institute	0.87	0.44	0.94	0.69	0.53	0.80	0.57	0.54	0.84	0.77
OECD	0.67	0.69	0.61	0.81	0.60	0.74	0.64	0.64	0.86	0.66
Cologne Institute	0.60	0.56	0.88	0.94	0.73	0.66	0.53	0.52	0.82	0.68
Düsseldorf Institut	0.67	0.75	0.83	0.81	0.67	0.79	0.72	0.73	0.86	0.70
Governments Economic Report	0.73	0.75	0.89	0.75	0.60	0.68	0.72	0.94	0.82	0.75
European Commission, autumn	0.60	0.44	0.60	0.75	0.60	0.71	0.57	0.63	0.88	0.75
European Commission, spring	0.73	0.81	0.70	0.88	0.73	0.78	0.81	0.77	0.89	0.82
Intern. Monetary Fund, autumn	0.53	0.50	0.44	0.88	0.60	0.63	0.55	0.59	0.89	0.75
Intern. Monetary Fund, spring	0.73	0.94	0.78	0.88	0.73	0.78	0.82	0.76	0.84	0.82

Source: Own calculations

Table 14 Average multi-dimensional measures of forecast accuracy, 1993 to 2019

	City block metric	Euclidean distance	Mahalanobis distance
Joint diagnosis, autumn	3.54	1.88	4.19
Joint diagnosis, spring	2.08	1.10	1.90
Council of Economic Experts	3.31	1.78	3.98
Kiel Institute	2.73	1.46	2.98
Berlin Institute	3.14	1.62	3.44
Hamburg Institute	2.98	1.60	3.37
Munich Institute	2.62	1.42	2.92
Essen Institute	2.65	1.44	2.28
Halle Institute	2.76	1.49	3.21
OECD	3.37	1.75	3.05
Cologne Institute	3.44	1.89	4.53
Düsseldorf Institut	2.91	1.57	3.00
Governments Economic Report	2.54	1.35	2.38
European Commission, autumn	3.23	1.70	3.77
European Commission, spring	2.09	1.13	2.42
International Monetary Fund, autumn	4.02	2.12	4.50
International Monetary Fund, spring	2.42	1.30	2.15

Source: Own calculations

Table 15 Spearman (1906) (lower triangular) and Kendall (1938) (upper triangular) rank correlation coefficients based on the Stekler Statistic (1993 to 2019)

	Growth	Inflation	Unemployment	Export	Import	City block	Euclidean distance
Growth	(-)	0.441*	0.515**	0.583**	0.441*	0.735***	0.750***
Inflation	0.657***	(-)	0.456*	0.450*	0.265	0.676***	0.632***
Unemployment	0.689***	0.640***	(-)	0.583**	0.397	0.603**	0.588***
Export	0.749***	0.671***	0.775***	(-)	0.627***	0.657***	0.642***
Import	0.603**	0.461*	0.547**	0.781***	(-)	0.529**	0.544**
City Block	0.880***	0.858***	0.797***	0.849***	0.711***	(-)	0.926***
Euclidean Dist.	0.880***	0.863***	0.752***	0.819***	0.694***	0.985***	(-)
Mahalanobis Dist.	0.794***	0.931***	0.696***	0.748***	0.583**	0.946***	0.956***

Our own calculation. *,(**, ***) denotes rejection of the hypothesis of a zero correlation coefficient

Table 16 Forecasting dates and horizons of the institutions under investigation, average 1993 to 2019

Institution	Number of observations	Forecast horizon in days[a), b)]	Standard deviation
Joint diagnosis, autumn	27	444	8
Joint diagnosis, spring	27	257	11
Council of Economic Experts	27	426	1
Kiel Institute	27	385	4
Berlin Institute	27	365	9
Hamburg Institute	27	382	31
Munich Institute	27	374	5
Essen Institute	27	311	20
Halle Institute	26	371	25
OECD	26	414	7
Cologne Institute	27	425	23
Düsseldorf Institute	27	381	11
Governments Economic Report	27	337	10
European commission, autumn	25	429	11
European commission, spring	26	254	23
International Monetary Fund, autumn	27	457	10
International Monetary Fund, spring	27	258	11

Our own compilation based on the publications of the institutions. [a)]: Refers to the date of publication of the forecast, rather than to the 'cut-off date,' up to which information had been taken into account, which was mentioned sometimes in the text. [b)]: Calculated as difference to the first of January of the year after the forecast year

References

Bailey DH, Borwein JM, Salehipour A, López de Prado M (2018) Evaluation and ranking of market forecasters. J Invest Manag 16(2):47–64

Batchelor RA (1990) All forecasters are equal. J Bus Econ Stat 8(1):143–144

Bürgi C, Sinclair TM (2017) A nonparametric approach to identifying a subset of forecasters that outperforms the simple average. Empir Econ 53(1):101–115

Casey E (2020) Do macroeconomic forecasters use macroeconomics to forecast? Int J Forecast 36(4):1439–1453

Clements MP (2019) Do forecasters target first or later releases of national accounts data? Int J Forecast 35(4):1240–1249

Conigrave J (2020) corx: create and format correlation matrices. https://CRAN.R-project.org/package=corx, r package version 1.0.6.1, last access: 2/24/2022

Consensus Forecast (ed) (2020) G7 & Western Europe 2020 Forecast Accuracy Award winners. https://www.consensuseconomics.com/cf-2020-forecast-accuracy-award-winners, last access 8/9/2021

Cowles A (1933) Can stock market forecasters forecast? Econometrica 1(3):309–324

Diebold FX (2015) Comparing predictive accuracy, twenty years later: a personal perspective on the use and abuse of Diebold-Mariano tests. J Bus Econ Stat 33(1):1–1

Diebold FX, Lopez JA (1996) Forecast evaluation and combination. In: Maddala G, Rao C (eds) Handbook of Statistics, vol 14, Elsevier, chap 8, pp 241–268

Diebold FX, Mariano RS (1995) Comparing pedictive accuracy. J Bus Econ Stat 13(13):235–265

Döhrn R (2015) Der Prognostiker des Jahres: Ein Zufallsergebnis? Möglichkeiten einer mehrdimensionalen Evaluierung von Konjunkturprognosen. Diskussionsbeitrag 208, University of Duisburg-Essen, Institute of Business and Economic Studie (IBES)
Döhrn R (2019) Revisionen der Volkswirtschaftlichen Gesamtrechnungen und ihre Auswirkungen auf Prognosen. AStA Wirtschafts-und Sozialstatistisches Archiv 13(2):99–123
Döhrn R, Schmidt CM (2011) Information or institution?: on the determinants of forecast accuracy. Jahrbücher für Nationalökonomie und Statistik 231(1):9–27
Dovern J, Jannsen N (2017) Systematic errors in growth expectations over the business cycle. Int J Forecast 33(4):760–769
Dowle M, Srinivasan A (2021) data.table: Extension of 'data.frame'. https://CRAN.R-project.org/package=data.table, r package version 1.14.2, last access: 2/24/2022
D'Agostino A, McQuinn K, Whelan K (2012) Are some forecasters really better than others? J Money Credit Bank 44(4):715–732
Fortin I, Koch SP, Weyerstrass K (2020) Evaluation of economic forecasts for Austria. Emp Econ 58(1):107–137
Fricke T (2018) Langzeitwertung der besten Prognostiker. Blog-entry "Neue Wirtschaftswunder", https://neuewirtschaftswunder.de/2018/12/20/langzeitwertung-der-besten-prognostiker/, last access: 8/9/2021
Fritsche U, Tarassow A (2017) Vergleichende Evaluation der Konjunkturprognosen des Instituts für Makroökonomie und Konjunkturforschung an der Hans-Böckler-Stiftung für den Zeitraum 2005-2014. IMK Study 54, Institut für Makroökonomie und Konjunkturforschung
Gamber EN, Smith JK, McNamara DC (2014) Where is the Fed in the distribution of forecasters? J Policy Model 36(2):296–312
Handelsblatt, NN (2014) Die besten Prognostiker im Land. Article "Handelsblatt", https://www.handelsblatt.com/infografiken/ranking-die-besten-prognostiker-im-land/9584620.html?ticket=ST-3358778-WEPGzZpcA4Hsif0udum7-ap1, last access: 8/9/2021
Heilemann U, Müller K (2018) Wenig Unterschiede-Zur Treffsicherheit Internationaler Prognosen und Prognostiker. AStA Wirtschafts-und Sozialstatistisches Archiv 12(3–4):195–233
Heilemann U, Stekler HO (2013) Has the accuracy of macroeconomic forecasts for Germany improved? German Econ Rev 14(2):235–253
Hyndman RJ, Khandakar Y (2008) Automatic time series forecasting: the forecast package for R. J Stat Softw 26(3):1–22
Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. Int J Forecast 22(4):679–688
Kendall MG (1938) A new measure of rank correlation. Biometrika 30(1/2):81–93
Knüppel M, Vladu A (2016) Approximating fixed-horizon forecasts using fixed-event forecasts. Discussion Paper 28/2016, Deutsche Bundesbank
Lehmann R, Wollmershäuser T (2021) The macroeconomic projections of the german government: a comparison to an independent forecasting institution. German Econ Rev 21(2):235–270
Mahalanobis CP (1936) On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India
Meyler A (2020) Forecast performance in the ECB SPF: ability or chance? Working Paper 2371, European Central Bank, https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp2371~4edce8ed72.en.pdf?cd8f1ebfee28d30cca20ae6d4ecc6aee, last access: 8/17/2021
Pagan A (2003) Report on modelling and forecasting at the bank of england/bank's response to the pagan report. Bank Engl Q Bull 43(1):60
Qu R, Timmermann A, Zhu Y (2019) Do any economists have superior forecasting skills? Discussion Paper 14112, Centre for Economic Policy Research (CEPR), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3496601, last access: 2/19/21
R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/
Rich RW, Tracy J (2021) All forecasters are not the same: time-varying predictive ability across forecast environments. Working Paper 21/06, Federal Reserve Bank of Cleveland, https://www.clevelandfed.org/en/newsroom-and-events/publications/working-papers/2021-working-papers/wp-2106-all-forecasters-are-not-the-same.aspx, last access: 8/3/2021
Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M (2011) proc: an open-source package for r and s+ to analyze and compare roc curves. BMC Bioinform 12:77

Rybinski K (2021) Ranking professional forecasters by the predictive power of their narratives. Int J Forecast 37(1):186–204
Silvia J, Iqbal A (2012) A new approach to rank forecasters in an unbalanced panel. Int J Econ Financ 4(9)
Sinclair T, Stekler HO, Muller-Droge HC (2016) Evaluating forecasts of a vector of variables: a German forecasting competition. J Forecast 35(6):493–503
Sinclair TM, Stekler HO, Carnow W, Hall M (2012) A new approach for evaluating economic forecasts. Econ Bull 32(3):2332–2342
Sinclair TM, Stekler HO, Carnow W (2015) Evaluating a vector of the Fed's forecasts. Int J Forecast 31(1):157–164
Spearman C (1906) Footrule for measuring correlation. Br J Psychol 2(1):89
Stekler HO (1987) Who forecasts better? J Bus Econ Stat 5(1):155–158
Timmermann A, Zhu Y (2017) Monitoring forecasting performance. Working paper, Rady School of Management, https://rady.ucsd.edu/docs/faculty/timmermann/Monitoring_performance_August_29__2017.pdf, last access: 2/19/21
Timmermann A, Zhu Y (2019) Comparing forecasting performance with panel data. Discussion Paper 13746, Center of Economic Policy Research (CEPR), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3395183, last access: 2/4/2022