GDP FORECASTING BY CZECH INSTITUTIONS: AN EMPIRICAL EVALUATION

Jiří Šindelář*

Abstract
This paper evaluates the accuracy of real GDP growth forecasts published in the period 1995–2013 by two Czech institutions: the Ministry of Finance (MF) and the Czech National Bank (CNB). A two-stepped approach is adopted: first a battery of forecasting errors (MAE, RMSE, MASE) is calculated, complementary to evaluation papers already available. Then statistical analysis is carried out by comparing both MF and CNB forecasts with OECD, European Commission and consensus benchmarks (Kruskal-Wallis test), assessing the presence of systemic bias (Wilcoxon test) and determining their incremental improvement (Page trend test). The results show that although some error patterns might suggest performance deficiencies (*i.e.* during the recession periods), the accuracy of forecasts prepared by both the MF and CNB does not differ significantly from the benchmark forecasts; MF and CNB predictions do not contain systemic bias and their accuracy improves as the horizon shortens. The paper also highlights several methodological shortcomings in the internal evaluations conducted by both institutions, indicating a potential for further improvement.

Keywords: GDP forecasting, accuracy measures, Ministry of Finance, Czech National Bank
JEL Classification: E37, E66, H68, O47

1. Introduction

Macroeconomic forecasts are an essential input for many decision making procedures in both the public and corporate spheres. This is particularly true regarding the forecasts of main macroeconomic aggregates such as GDP or inflation produced by public or government institutions. GDP growth forecasting in particular occupies a special position. Not only does it have a vital signal function for commercial subjects, it is a crucial input for the decision making and planning of governmental organisations in areas such as tax revenues and state budget composition. Because of this, forecasting performance of public institutions is a frequent point of discussion in the forecasting community, as exhibited by *e.g.* Öller and Barot (2000), Allan (2013) and Keereman (1999).

In the Czech context, GDP forecasts prepared by the most prominent institutions, *i.e.* the Ministry of Finance (MF) and the Czech National Bank (ČNB), have faced increased public scrutiny over recent years. Initially the MF was criticised by the massmedia for allegedly compromising its predictions with an upward bias and failing to foresee the cool down period of 2009; a conjecture also indicated by some of the few empirical works available (Soukup, 2012; Boček, 2012; Polák, 2011). Lately, the CNB´s forecasting ability and past performance has also been questioned, as part of a wider discussion related to its 2013 monetary intervention. Although more of a technical matter, the overall public interest

* Jiří Šindelář, University of Finance and Administration, Prague, Czech Republic (jiri.sindelar@vsfs.cz).

in the accuracy of central growth forecasts seems to have increased substantially compared to the pre-2009 period.

In addition to the predominantly media discourse, both the MF and CNB have conducted their own (MF 2013) or related (Arnoštová *et al*., 2009; Antal *et al*., 2008; Antoničová *et al*., 2009; Novotný and Raková, 2011) evaluations over time, with substantially different, mostly positive outcomes. The already inconsistent views on the MF's and CNB´s GDP forecasting are further complicated by the fact that most contributions deal with methodological deficiencies, such as ambiguity of empirical background (Antoničová *et al*., 2009), no rigorous tests of significance (*e.g.* Soukup, 2012), or inappropriate measurement methods (*e.g.* Novotný and Raková, 2011).Both the factual inconsistencies and methodological problems of current evaluations represent a vital opportunity for a comprehensive and thorough study. The necessity of the topic is further amplified by an overriding lack of Czech context in relevant foreign literature, as demonstrated by the already mentioned Öller and Barot (2000) or Keereman (1999) papers.

With respect to the above, this paper seeks to dispel the uncertainty concerning the GDP forecasting of two major Czech institutions (MF, CNB) by empirically evaluating their forecasting performance in the years 1995–2013. In order to fulfil this goal, the paper is divided into three sections. In the first part, available evaluations of both MF and CNB forecasting performance are discussed, both from the factual and from the methodological perspective. In the second part, a battery of three forecasting errors (MAE, RMSE, MASE) is calculated, complementing measures used in the existing evaluations. Finally, in the last section, MF and CNB forecasts were tested for the presence of systemic bias (Wilcoxon test), significant difference *versus* benchmark (Kruskal-Wallis test *versus* OECD, EC and consensus forecast) and incremental improvement during revisions (Page trend test), covering forecasting traits identified as vital in the literature overview.

2. Literature Overview

Following the recent media controversy regarding their forecasting performance, both the Ministry of Finance and the Czech National Bank conducted their own evaluation of their respective forecasts. The ministry based their analysis (MF, 2013) on three methods: Average forecasting error (AE), Mean average error (MAE) and Theil´s Inequality Coefficient (TIC), benchmarked on the naïve forecast. With these tools, the MF identified strong over-forecasting in its GDP forecasts (1-year horizon) during the recession years of 2009 and 2012, and mild under-forecasting in the stable years of 2001–2006. Theil´s indicator became lower than one (~ 0.4) for 1-year GDP forecasts in the period 2007–2012, indicating better forecasting performance than the naïve benchmark. On the contrary, the results for the 1995–2000 period were significantly worse (~ 0.8) suggesting that there has been significant improvement in the forecasting accuracy.

In the case of the Czech National Bank, it is difficult to find a similar comprehensive evaluation. The CNB obviously evaluates its forecasting performance, but its attention, at least in terms of public analyses, is focussed mainly on inflation forecasting. Principal evaluations available linked to the CNB were conducted by its analysts in external journals. Arnoštová *et al.* (2011) evaluated the performance of six different models in short horizon GDP forecasting in the period of 2001–2009. Using quarterly RMSE comparison, they found the CNB model (historical near-term forecasts) to be the second most accurate.

Antal *et al.* (2008) conducted a comprehensive evaluation of different CNB forecasts in 1998–2007, including GDP forecasts. With Mean Error (ME), MAE and t-statistics he concluded that CNB growth forecasts were most of the time undershooting the actual value ("sandbagging") and were unbiased for a 1Q (3 month) interval, yet biased for a 4Q (12 month) horizon. Antoničová *et al*. (2009) later confirmed this result for quarterly periods during 2004–2006. Using Mean Absolute Deviation (MAD), she also discovered that the forecasting error of CNB forecasts increased with lengthening horizon and gradually decreased over time – being largest for the forecasts produced in 2004 and, by contrast, lower in the case of the latest reviewed forecasts dating from 2006. Finally, Novotný and Raková (2011) compared a consensus forecast, utilized by the CNB, with those of three international bodies (European Commission, IMF, OECD). Employing a combination of five measures (ME, MAE, MAPE, MSE and RMSE), error regression and the Diebold-Mariano test, they determined that the consensus forecast beats the alternatives by a difference which is typically statistically significant and confirmed a relatively low accuracy of next-year GDP growth forecasts. Overall, the CNB analysts seem to prefer different accuracy measures than the MF ones, with the most common metrics being ME, MAE, RMSE, MSE and even pure MAPE, whose direct application to GDP forecasting remains questionable. In two of the four surveyed papers, statistical tests were also employed (incl. prolific Diebold-Mariano test).

From the research perspective, despite the praise both institutions display in relation to their own forecasting performance, serious questions remain. The MF exhibited a rather conservative and cautious approach in its internal evaluation, with a limited set of absolute errors and high reliance on TIC. Such combination does not sufficiently cover all of the accuracy aspects. According to evidence provided by Makridakis and Hibon (1995) and Hyndman and Koehler (2006), it does not capture variance in forecast error (AE, MAE) and it is sensitive to outliers (TIC), particularly because of using RMSE as its relative measure[1]. Using RMSE as a part of the TIC coefficient is a suitable strategy, because it covers errors in changes well. It should, however, be accompanied by a more sophisticated apparatus, or computed also with different error measures, to prevent the outlier deformation. This problem is closely linked to the evaluation period, which fundamentally affects the outcomes of all indicators. In the MF internal evaluation, this period was set to a default six year period in an effort to provide robust, independent intervals. That is not necessarily a bad thing, but it represents only a very basic picture of the time frame. Further disclosure or provision of different time structures (*i.e.* tied to economic, political cycles) would be helpful – as exhibited, for example, by Daníelsson (2008). Even within this valuation framework and with some time intervals missing (*e.g.* 2001–2006 TIC), the review still indicates rather poor performance in comparison with either OECD or consensus forecast, as well as ample over-forecasting in the recession years of 2007–2012. No statistical

1 This is clearly visible when comparing MAE and TIC for 1995–2000 and 2007–2012 intervals, on the 6–12-month horizon. While the MAE for both intervals remains approximately the same, the TIC reported substantial improvement. But was the MF forecast really improving? Analysis of underlying data shows that the TIC denominator (naïve RMSE of $GDP_{t-1} - GDP_t$) was in the first interval significantly lower than in the later one (outliers), leading to potentially lower proportions. The fact that the error magnitude (MAE) remained roughly the same in both periods suggests that it might have been just the shape of the GDP growth, rather than real MF forecast improvement, which "lowered" the TIC indicator.

testing was conducted, preventing the analysis of random *versus* systemic phenomena. All in all, although the MF attempted to evaluate its forecasting performance in front of public scrutiny and it used an ambitious methodological framework, its review still exhibits some deficiencies and arguably fails to dispel much of the past criticism.

In the CNB case, a lack of any comprehensive (public) evaluation of its own GDP-forecasting performance, despite heated media debate, feels inappropriate. Partial studies conducted by CNB analysts reveal more frequent combinations of error measures (ME, MAE, RMSE, MSE), but, on the other hand, no application of relative errors and even obscurities such as MAPE, whose application in the GDP environment (close to zero values) is consistently criticized (Armstrong and Collopy, 1992; Hyndman and Koehler, 2006). Although some of the papers indicate that the CNB forecasted more conservatively than the MF (Antoničová *et al*., 2009; Antal *et al*., 2008), a thorough empirical review or comparison of forecasting performance is missing. Utilizing Armstrong´s (2001) checklist on CNB-linked studies, their evaluation strategy can be characterized by either poor reliability and outlier protection (Arnoštová *et al*., 2010) or inadequate incorporation of variation aspects (Antal *et al*., 2008).

Overall, consideration of internal forecasting evaluations of both institutions raised serious questions about their reliability and validity. Combined with the lack of independent empirical literature on the topic[2] and the practical omission of Czech GDP forecasting in respected international papers (such as Öller and Barot, 2000), these deficiencies delimit the purpose and focus-points of this study. The hypotheses and methods use are detailed below.

3. Research Hypotheses

Based on the above, the paper seeks to verify a total set of three research hypotheses:

> **H_1:** MF/CNB forecasting accuracy is not significantly different from the accuracy of other inquired institutions.

The first hypothesis represents the public discourse related to MF/CNB forecasting in general. Based on popular opinion, the MF is expected to achieve significantly different (lower) forecasting accuracy than other inquired institutions.

> **H_2:** MF/CNB forecasts do not contain any systematic bias.

The second hypothesis verifies the existence of systemic bias, *i.e.* observed error above or below zero not due to random fluctuations, particularly in relation to over- and under-forecasting.

> **H_3:** MF/CNB forecasting accuracy does not increase with the shortening of horizon.

The last hypothesis is related to the general rule of accuracy increase with the shortening forecasting horizon. This presumption, in terms of an alternative hypothesis, represents a criterion validity indicator of some sort as well.

2 Polák (2011) and Soukup´s (2012) analyses represent a rare exception, with their outcomes indicating worse forecasting performance of the MF compared to international organizations (OECD, IMF).

4. Method

The evaluation strategy of this paper reflects Hibon´s *et al.* (2012) two-stepped approach: (i) first, calculation of selected institutions´ forecasting error was conducted and (ii) secondly, statistical analysis (testing) was undertaken, in order to validate the set of hypotheses.

In the first step, the accuracy of selected institutions´ forecasts is evaluated with the combination of three error measures[3]:

- **Mean Absolute Error (MAE)**

$$MAE = mean\left(\left|E_t\right|\right) \tag{1}$$

- **Root Mean Squared Error (RMSE)**

$$RMSE = \sqrt{mean\left(E_t^{\,2}\right)} \tag{2}$$

- **Mean Average Scaled Error (MASE)**

$$MASE = mean\frac{E_t}{\frac{1}{n-1}\sum_{i=2}^{n}\left|Y_i - Y_{i-1}\right|} \tag{3}$$

This combination reflects empirical findings of most influential forecast-evaluation papers to date and mentioned earlier (*i.e.* prominently Hyndman and Koehler (2006) and Armstrong and Collopy (1992)). It enables coverage of all the crucial aspects of forecasting performance, such as magnitude of forecasting error, systemic bias and performance in changes. As documented by the GDP forecasting literature overview, two of the selected measures (MAE, RMSE) represent the standard in the surveyed field, while the third (MASE) represents recent development in forecasting theory. The measures used also represent a complementary addition to the measures already used by the institutions (MF) in their internal evaluations, enabling the paper to undertake a different-point-of-view evaluation.

The second step of the paper method deals with the analysis of attributes and differences between the forecasting accuracy (error) of surveyed institutions. In relation to the set of hypotheses, a battery of three tests is applied:

(i) In order to compare performance (accuracy) between the institutions, the standard **Kruskal-Wallis test** is utilized. The choice of method here is limited, as it needs to deal with two specific traits of underlying GDP data: finite time-series and also – as with most macro indicators – serial persistence. As documented by Christensen *et al.* (2007), most of the traditional methods, *i.e.* Diebold-Mariano test, exhibit substantial problems in dealing with such setup (rejecting null too often – oversized type I error), making them unsuitable. For this reason we decided to use the less restrictive non-parametric K-W test, preceded by assessment of its two main assumptions: independence and symmetry. We tested the first one using the Box-Pierce independence test and the second with the Miao, Gel and Gastwirth symmetry test (Miao *et al*., 2006), with mutually positive outcomes: the independence (insignificance of serial correlation) and symmetry was not rejected for

3 Let us denote the real value at any given time t as Y_t, while F_t stands for the forecast value and E_t the subsequent forecasting error $(Y_t - F_t)$.

every included time series on every horizon (*i.e.* H_0 was not rejected on common $p = 0.05$; full disclosure of the p-values can be found in Appendix).

(ii) For the systemic bias analysis, the paired **Wilcoxon signed rank test** is used (Wilcoxon, 1945), examining, whether an observed error above or below zero indicated bias could be due to random fluctuations. Application of the Wilcoxon test for forecasting bias has a well-established lineage, as documented in papers by Mühleisen *et al.* (2005), Campbell and Ghyssels (1995) and Daníelsson (2008). Compared to mostly parametric alternatives (paired t-test), Wilcoxon has the obvious advantage of not requiring the errors to be normally or t-distributed. As with the preceding method, it demands that the errors are independently drawn from a continuous and symmetric population – both assumptions were not rejected in the previous paragraph. The control method here consists of paired t-test accompanied by the necessary test of normality (Shapiro-Wilk test) as well as its variant developed specifically for weakly dependent processes (Bai and Ng, 2005).

(iii) In order to detect whether the average error size reduces as the horizon advances, the **Page trend test** is applied, in a classical (Page, 1963) variant. As with the previous, the Page trend test also belongs to the nonparametric test family, resulting in limited data prerequisites. By employing the Page test, statistical significance of the incremental improvement would be verified.

5. Data

As stated in the introduction, this study provides an analysis of forecasting performance of two Czech institutions (MF, CNB) in comparison with three other international or benchmark bodies (OECD, EC and consensus forecast compiled by the MF).We examine annual (real) GDP growth forecasts produced quarterly from the second quarter 1995 to the fourth quarter 2013. In addition, two different sub-periods were also defined:

(i) Six-year period, analogous with the MF internal evaluation (2.1995–4.2001, 2.2002–4.2007, 2.2008–4.2013). This period would enable us to confront results achieved by the MF in its own assessment[4].

(ii) Supplementary, more granular two to four-year period, compatible with cycles of the Czech Republic economy (2.1995–4.1996, 2.1997–4.1998, 2.1999–4.2002, 2.2003–4.2007, 2.2008–4.2010, 2.2011–4.2013). This period would enable us to evaluate MF/CNB performance in different stages of business cycles[5].

Taking into account the frequency with which the surveyed institutions produce their forecasts[6], the analysis is based on forecasts published in the second (spring) and fourth (autumn) quarter of the year. The most common month for the spring (Q2) forecast is April, and October for the autumn (Q4) forecast. Subsequently, four different forecast horizons were evaluated: 3 months, 9 months, 15 months and 21 months. While the short and medium

4 Because of the short-time frame available, the test results established in the sub-periods have only partial statistical conclusiveness. Still, they enable us to indicatively assess both MF and CNB performance under different (economic) conditions, which is a vital research feature, as evinced by *e.g.* the Daníelsson (2008) or Öller and Barott (2000) papers.

5 The stages of Czech Republic business cycle were defined on the basis of MF economy analyses (*e.g.* MF, 2014) and in order to reflect economy (GDP growth) turning points.

6 The surveyed institutions produce GDP forecasts in two different frequency schemes: quarterly (winter, spring, summer, autumn – MF, CNB, consensus) and biannually (spring, autumn – EC, OECD).

ones (3–9 months) are vital as an indicator of a forecast´s incremental improvement, the longer ones (15 and 21 months) represent the horizon of national budgetary policy and therefore hold a significant position in the planning apparatus (MF, 2013).

Finally, not all of the data were available for the whole surveyed period. Only the MF and OECD forecasts cover all of the 1995–2013 interval, while the rest (CNB, OECD and consensus) can be traced back only for years 1998–2000. This limitation would be addressed during the testing and interpretation of the results. As for real GDP growth (Y_t) value, the study utilizes actual (2014) data published by the Czech Statistical Office (CZSO). Utilizing the most recent data enables us to conduct the most accurate *ex-post* evaluation, similarly to Allan (2013), Daníelsson (2008) or Arnoštová *et al.* (2010) papers[7], at the expense of *ex-ante* information efficiency perspective, which remains a matter of additional research.

6. Results

6.1 Error measures

In the first part, forecasting errors for the whole period and its sub-periods were calculated, as summarized in Table 1 and Table 2.

Going through the conventional error measures (MAE, RMSE), we can derive three basic observations. Firstly, the results show increase of forecast error with lengthening horizon in narrow majority of periods (1995–1996, 2003–2013), yet this is not confirmed universally, with other years (1997–2002) exhibiting higher error with shorter-term predictions. Secondly, the results suggest that MF forecasting might have performed better in the periods of stable growth than in the recession period. From a more granular point of view, MF forecasts offered the lowest accuracy in the imminent recession (breaking) periods of 1997–1998 and 2008–2010. Interestingly enough, however, the MF also produced notably inaccurate (next year) forecasts in the preceding growth period (2003–2007) and in the subsequent third recession (2011–2013). Combined with the previous analysis of mean deviation, which revealed ample over-forecasting in recession years and sandbagging in growth periods, this fact immediately implies the possibility of systemic bias within the forecasts. Finally, from the most aggregate perspective, the results suggest that MF forecasting performance might have somehow improved over time.

Evaluation with MASE method offers a different picture. Bearing in mind that MASE is a scaled error based on the in-sample MAE from the naïve (random walk) forecast method, its results require a different treatment and interpretative approach. Essentially, a value lower than one indicates that a forecast is better than the average one step naïve forecast computed in sample, while an outcome higher than one indicates worse performance than the simple naïve forecast. Looking at the MF results, the MASE evaluation largely corresponds to the first and last observation made with the conventional measures: decrease in accuracy with the lengthening horizon and a certain level of improvement over time. Turning to the predictive ability in stable *versus* breaking periods, we get back to the lower/higher than one rule. While in the case of the first recession (1997–1998), all of the forecasts except for the shortest one exhibited lower accuracy than the in-sample naïve forecast, in the second

7 The so called "most recent outturn" approach, aiming at comparing the forecast with the most recent (accurate) GDP estimate. For the methodological rationale, see *e.g.* Mc Nees and Ries (1983).

decline (2008–2010), this indicator improved and MF forecasts were better than the naïve ones. In general, however, the added value of the MF forecasting seems debatable given that, except for the shortest 3M predictions, the predictions in any sub-period rarely improved on the in-sample forecast.

Table 1 | Error Measures - MF

Period	3M Forecast			9M Forecast			15M Forecast			21M Forecast		
	MAE	RMSE	MASE	MAE	RMSE	MASE	MAE	RMSE	MASE	MAE	RMSE	MASE
2.1995–4.2013 (total period)	0.8	1.084	0.722	1.6	1.827	2.14	2.1	2.874	2.654	2.6	3.339	3.482
2.1995–4.2001	1.1	1.339	0.596	2.2	2.255	1.998	1.9	2.693	1.292	2.8	3.149	1.957
2.2002–4.2007	1.1	1.131	1.206	1.7	1.636	1.819	1.8	2.046	2.5	1.8	2.188	2.643
2.2008–4.2013	0.4	0.591	0.384	1.1	1.4	2.602	2.5	3.677	4.396	2.9	4.352	5.846
2.1995–4.1996 Post- transformation growth	0.5	0.595	0.682	1.0	0.679	3.689	3.1	4.142	1.043	6.6	4.633	1.215
2.1997–4.1998 First recession	1.1	1.452	0.242	3.2	3.251	2.782	1.3	1.725	2.007	2.8	2.939	3.505
2.1999–4.2002 Recovery	1.2	1.408	0.726	1.7	1.942	1.014	1.2	1.636	0.808	1.3	1.521	0.905
2.2003–4.2007 Steep growth	1.1	1.214	1.331	1.5	1.702	1.919	2.1	2.231	2.942	2.2	2.395	3.151
2.2008–4.2010 Second recession	0.7	0.818	0.199	1.7	1.737	0.371	3.5	4.902	0.556	4.0	5.64	0.799
2.2011–4.2013 Stagnation – third recession	0.1	0.173	0.569	0.8	0.95	4.833	1.7	1.733	8.236	2.3	2.465	10.894

Source: Own research.

Error measures of the second surveyed institution, the Czech National Bank (CNB), offer a broadly similar picture. The bank´s forecasting performance appeared to follow the same lines as the ministry´s, with conventional errors found to increase with the lengthening horizon in three out of four business-cycle periods and forecasts generally recording higher error values during the recession period. Particularly in the growth years of 2003–2007, the conservative approach of the CNB forecasters resulted in lower forecasting errors, with the same being true for the second (2008–2010) recession as well. On the other hand, the bank produced less accurate 15M and 21M forecasts during the third recession years (2011–2013). All in all, the bank exhibited similar accuracy patterns, with more favourable aggregate errors.

Using supplementary MASE metrics, the bank was able to achieve close-to-one values in more intervals than the MF did. In almost the whole recovery (1999–2002) and second recession (2008–2010) period, the bank fared better than the naïve forecast, with growth

period (2003–2007) 3M and 9M errors being notably lower than its MF counterpart. On the other hand, the bank exhibited inferior naïve forecast-related performance during the third recession (2011–2013) 15M and 21M forecasts, with its MASE errors exaggerating the ministry ones. Obviously, CNB forecasters expected a more optimistic scenario than a continued stagnation.

Table 2 | Error Measures - CNB

Period	3M Forecast			9M Forecast			15M Forecast			21M Forecast		
	MAE	RMSE	MASE	MAE	RMSE	MASE	MAE	RMSE	MASE	MAE	RMSE	MASE
2.1998–4.2013 (total period)	0.8	0.963	0.811	1.3	1.406	1.417	2.1	2.443	2.705	2.3	2.742	3.911
2.1998–4.2001	1.3	1.116	0.885	1.8	1.55	1.404	1.9	1.621	1.166	1.8	1.48	1.073
2.2002–4.2007	0.9	1.044	0.995	1.3	1.358	1.261	2.0	2.143	2.894	1.7	2.13	2.663
2.2008–4.2013	0.4	0.63	0.577	1.0	1.268	1.582	2.4	3.352	3.541	2.9	4.09	7.051
2.1999–4.2002 Recovery	1.3	1.458	0.735	1.5	1.847	0.779	2.0	2.217	1.259	1.4	1.77	0.874
2.2003–4.2007 Steep growth	0.9	1.111	1.066	1.2	1.474	1.418	1.9	2.167	3.223	2.2	2.332	3.178
2.2008–4.2010 Second recession	0.6	0.874	0.203	1.6	1.648	0.346	3.0	4.336	0.693	3.2	4.95	0.433
2.2011–4.2013 Stagnation – third recession	0.2	0.171	0.952	0.6	0.708	2.818	1.9	1.918	6.39	2.9	2.991	13.669

Source: Own research.

6.2 Statistical tests of research hypotheses

As pointed out in the method description, the institutions´ forecasts were further subject to three independent statistical analyses: performance comparison (with consensus forecast, OECD and EC), systemic bias test and incremental improvement analysis. Table 3 summarizes the results of the first one of them, in terms of p values of the Kruskal-Walis test carried out.

Table 3 | Differences in Accuracy

Method	3M Forecast				9M Forecast				15M Forecast				21M Forecast			
	1995–2013	1995–2001	2002–2007	2008–2013	1995–2013	1995–2001	2002–2007	2008–2013	1995–2013	1995–2001	2002–2007	2008–2013	1995–2013	1995–2001	2002–2007	2008–2013
Kruskal-Wallis test (all together)[8]	0.943	0.932	0.586	0.428	0.976	0.804	0.945	0.797	0.842	0.661	0.999	0.966	0.978	0.587	0.960	0.994

Source: Own research.

8 Consensus forecast data cover period of 2000–2013, OECD data for1995–2013 and EC data for 2000–2013.

The empirical evidence provides a clear conclusion to the previously raised questions: both CNB and MF GDP forecasts are no less or more accurate compared to their Western counterparts. In the whole sample and across all horizons, the null hypothesis (H_1) was not rejected on the most common $p = 0.05$ level, or the more benevolent $p = 0.10$ level. In the most economically exposed periods, *i.e.* 2008–2013, the K-W test was unambiguous in not finding any significant difference. The statistically most important results for the whole period of 1995–2013 offered, for all horizons and comparisons, the same indisputable result.

Table 4 | Systemic Bias

Method	3M Forecast				9M Forecast				15M Forecast				21M Forecast			
	1995–2013	1995–2001	2002–2007	2008–2013	1995–2013	1995–2001	2002–2007	2008–2013	1995–2013	1995–2001	2002–2007	2008–2013	1995–2013	1995–2001	2002–2007	2008–2013
MF_Wilcoxon test	0.374	0.688	0.156	0.688	0.865	0.563	0.156	0.094	0.799	0.297	0.156	0.438	0.378	0.313	0.219	0.188
MF_T-test	0.279	0.665	0.108	0.565	0.823	0.693	0.123	0.116	0.421	0.333	0.172	0.308	0.244	0.239	0.186	0.207
MF_ SWNormality test	0.936	0.964	0.627	0.301	0.563	0.340	0.892	0.674	0.069	0.689	0.634	0.463	0.258	0.927	0.930	0.585
MF_BN Normality test	0.822	–			0.298	–			0.204	–			0.273	–		
CNB_Wilcoxon test	0.144	0.250	0.063	0.156	0.433	0.375	0.094	0.313	0.454	0.625	0.094	0.312	1.000	0.750	0.219	0.312
CNB_T-test	0.128	0.174	0.103	0.238	0.418	0.456	0.099	0.234	0.809	0.427	0.143	0.290	0.677	0.538	0.263	0.166
CNB_SW Normality test	0.331	0.456	0.334	0.006	0.511	0.384	0.668	0.898	0.048	0.976	0.328	0.290	0.104	0.965	0.330	0.666
CNB_BN Normality test	0.427	–			0.462	–			0.326	–			0.29	–		

Source: Own research.

Systemic bias, *i.e.* the presence of systematic over-forecasting or sandbagging, is an important quality indicator of any forecasting function. In this case, the data revealed that neither the MF nor the CNB produced systematically biased forecasts on the $p = 0.05$ level, hence the null hypothesis (H_2) was not rejected in all entries. With the more benevolent $p = 0.10$, there are several occasions where – surprisingly – the bank´s predictions can be considered systematically inaccurate: 3M, 9M and 15M forecasts in the 2002–2007 period. However, the supplementary methods (T-test, Normality test) do not confirm this outcome, as only in one horizon (9M) did at least two of the three tests agree on a significant (less than 0.1) result. As for direction of the bias, a quick analysis reveals that in the 2002–2007 period, the CNB was regularly under-forecasting (sandbagging) the real GDP growth. In the rest of the field, the results countered previous negative expectations. According to the tests employed, there is no systemic pattern in forecasting errors produced by the MF and CNB.

The final analysis deals with the incremental improvement of both MF and CNB forecasts, *i.e.* the ability to improve their accuracy in the shortening horizon.

Table 5 | Incremental Improvement

Page´s trend test	1995–2013	1995–2001	2002–2007	2008–2013
CNB	0.000	0.001	0.001	0.001
MF	0.000	0.001	0.001	0.001

Source: Own research.

Given the results in Table 5, we reject the null hypothesis (H_3) in favour of the alternate one for every surveyed horizon for both the MF and CNB (p = 0.05 and 0.10), meaning that with shortening horizon, the institutions´ forecast becomes significantly more accurate.

7. Discussion

Overall, the results shed a contrasting light on both institutions´ forecasting performance. Firstly, the triangle of errors calculated (MAE, RMSE, MASE) implied the possibility of systemic bias in the recession periods (1997–1998, 2008–2010) and inferior forecasting accuracy in relation to naïve forecast. This outcome would corroborate the previous criticism published in the massmedia and also corresponded to the critical assessment of the MF and CNB´s internal evaluations. The MF in particular, according to MAE and RMSE error measures, exhibited just the performance commented on in newspapers: ample over-optimism in the recession period (2008–2010), as well as dubious results in the consistent growth (2002–2007) sub-period.

The second part of the results, however, revealed substantially different outcomes. Examinations performed with the Kruskal-Wallis and Wilcoxon tests determined that the MF´s GDP forecasts were not significantly less or more accurate than the Western (EC, OECD) ones and contained no systemic bias, thus largely confirming findings outlined in the ministry´s own internal evaluations. As for the CNB, the study found that the visually higher accuracy of the bank´s short-forecasting concluded by Arnoštová *et al*. (2009) is not statistically significant, but, on the other hand, it upheld the bank´s performance in terms of systemic bias, not confirming the findings of Antal *et al.* (2008). As for incremental improvement, Page test results corroborate the paper of Antoničová *et al*. (2009), suggesting a significant increase inaccuracy with shortening horizon.

From an international perspective, the study largely conformed to the rich findings provided by Öller and Barott (2000). Although their paper dealt with a different time period (1975–1997), both studies agree that there is no significant difference *versus* the supranational forecasts (OECD), no systemic bias and overall better performance *versus* naïve forecast. Comparing the MF and CNB with other national evaluations at hand (Daníelsson, 2008; Milburn, 1978; Kirschgässner, 1993; Barker, 1985), the paper revealed even more positive findings as (i) systemic bias is commonly found in European institutions´ GDP forecasting and (ii) there are (in other evaluations) regular periods where short-term revisions did not yield any significantly higher accuracy.

8. Conclusions

The study examined GDP growth forecasting of two Czech institutions: the Ministry of Finance and the Czech National Bank. In summary, we have found that:

– The highest standard errors (MAE, RMSE) were concentrated in the recession periods (1997–1998, 2008–2010) with the long-term forecasts (15M and budgetary vital 21M). The comparatively error was universally offered by the short-term (3M) predictions in the latest surveyed sub-period of third recession (2011–2013).
– Although pooling of the bias suggested a possible systemic pattern, this was not confirmed by exact methods used in the second stage (Wilcoxon test, t-test), both in the total period and in the sub-periods.
– In comparison with naïve in-sample forecast (MASE), both institutions exhibited generally better performance with short-term (3M) forecasts. In longer horizons, the overall results indicate that the value added over the naïve forecast could be limited to some sub-periods, with most of the results suggesting poorer performance in that order. This is a serious issue worth further exploration.
– Difficult predictability of GDP indicator, resulting from the business cycle discontinuities, is underlined by the lack of serial correlation (Box-Pierce test) and anticipated persistence (*i.e.* "memory") of the time series. This may have attributed to limited performance over naïve forecast described above (MASE).
– Comparison with other international institutions (OECD, EC) and consensus forecast revealed that neither the MF nor the CNB achieved significantly different (lower or higher) accuracy (Kruskal-Wallis test) in any of the surveyed time-periods.
– Assessing the revisions from long to short-term forecast, the results showed improvement inaccuracy with shortening horizon for most of the periods. This outcome was confirmed with an exact test (Page test) meaning that there is a significant increase in forecast accuracy with shorter-time revisions.

Overall, the results create a positive picture of MF and CNB growth forecasting – contrary to popular expectations. Still, the study indicated some weak points that need to be addressed. First, both institutions exhibit notable deficiencies in their internal evaluations, in terms of methods employed and lack of significance tests. This opportunity for improvement is highlighted by this paper. Nevertheless, more methodological justification and clarification of the accuracy assessment (*e.g.* MAPE found in Novotný and Raková, 2011) are vital. Secondly, there was a visible difference detected between the outcome of Theil´s TIC in the MF's own evaluation and MASE employed in this study. Because performance *versus* the naïve benchmark is a vital indicator of forecasting value added, this discrepancy needs to be further investigated and possibly compared to a third, different indicator. Finally, the MF´s ability to forecast structural changes in the budgetary vital 21M horizon still seems limited. Although the study did not confirm any of the negative hypotheses regarding the ministry´s GDP forecasts, additional discrepancies might be discovered during the prospective (long-term) observations.

Appendix

Independence and Symmetry Tests Results

Institution/ Horizon	Box-Pierce independence test	Miao, Gel and Garswith symmetry test
MF 3M	0.239	0.243
MF 9M	0.089	0.700
MF 15M	0.438	0.371
MF 21M	0.324	0.473
CNB 3M	0.136	0.640
CNB 9M	0.382	0.961
CNB 15M	0.323	0.359
CNB 21M	0.319	0.283
OECD 3M	0.130	0.452
OECD 9M	0.055	0.727
OECD 15M	0.212	0.518
OECD 21M	0.331	0.135
EC 3M	0.436	0.152
EC 9M	0.161	0.971
EC 15M	0.510	0.359
EC 21M	0.794	0.223
Consensus 3M	0.269	0.096
Consensus 9M	0.189	0.710
Consensus 15M	0.510	0.358
Consensus 21M	0.402	0.142

Source: Own research.

References

Allan, G. (2013). *Evaluating the Usefulness of Forecasts of Relative Growth*. Strathclyde Discussion Paper in Economics No. 12–14.

Antal, J., Hlaváček, M., Horvath, R. (2008). Do Central Bank Forecast Errors Contribute to the Missing of Inflation Targets? The Case of the Czech Republic. *Czech Journal of Economics and Finance (Finance a uver)*, 58(09–10), 434–453. Available at: http://journal.fsv.cuni.cz/storage/1142_1142_3__antal-hlavacek-horvath.pdf

Antoničová, Z., Musil, K., Růžička, L., Vlček, J. (2009). Evaluation of the CNB's Forecasts. *Economic Research Bulletin*, 7(1), 8–10. Available at: http://www.ČNB.cz/en/research/research_publications/erb/download/ERB_No1_2009.pdf

Armstrong, J. S., Collopy, F. (1992). Error Measures for Generalizing about Forecasting Methods: Empirical Comparisons. *International Journal of Forecasting*, 8(1), 69–80, http://dx.doi.org/10.1016/0169-2070(92)90008-w

Armstrong, J. S. (2001). Evaluating Forecasting Methods, in Armstrong, J. S., ed., *Principles of Forecasting*. Kluwer Academic Publishers, Norwell, pp. 443–472.

Arnoštová, K., Havrlant, D., Růžička, L., Tóth, P. (2011). Short-Term Forecasting of Czech Quarterly GDP Using Monthly Indicators. *Czech Journal of Economics and Finance (Finance a úvěr)*, 6, 566–583. Available at: https://www.ČNB.cz/cs/vyzkum/vyzkum_publikace/ČNB_wp/2010/ČNBwp_2010_12.html

Bai, B., Ng, S. (2005). Tests for Skewness, Kurtosis, and Normality for Time Series Data. *Journal of Business & Economic Statistics*, 23(1), 49–60, http://dx.doi.org/10.1198/073500104000000271

Barker, T. (1985). Forecasting the Economic Recession in the UK 1979-1982: A Comparison of Model-Based ex ante Forecasts. *Journal of Forecasting*, 4(2), 133–151, http://dx.doi.org/10.1002/for.3980040204

Boček, J. (2012). *Prognózy HDP: Kdo je nejpřesnější?* GDP Forecast: Who is the Most Accurate? Available at: http://data.blog.ihned.cz/c1-58291130-prognozy-hdp-kdo-je-nejpresnejsi

Campbell, B., Ghysels, E. (1995). Federal Budget Projections: A Nonparametric Assessment of Bias and Efficiency. *The Review of Economics and Statistics*, 77(1), 17–31, http://dx.doi.org/10.2307/2109989

Daníelsson, Á. (2008). *Accuracy in Forecasting Macroeconomic Variables in Iceland*. The Central Bank of Iceland Working Paper No. 39.

Hyndman, R. J., Koehler, A. B. (2006). Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting*, 22(4), 679–688, http://dx.doi.org/10.1016/j.ijforecast.2006.03.001

Hibon, M., Crone, S., Kourentzes, N. (2012). *Statistical Significance of Forecasting Methods*. Paper presented at the 32nd Annual International Symposium on Forecasting, Boston, MA, USA.

Christensen, J. H., Diebold, F. X., Rudebusch, G., Strasser, G. (2007). *Multivariate Comparisons of Predictive Accuracy*. University of Pennsylvania working paper. Available at: http://www.econ.uconn.edu/Seminar%20Series/strasser08.pdf

Keereman, F. (1999). *The Track Record of the Commission Forecasts (No. 137)*. Directorate General Economic and Monetary Affairs (DG ECFIN), European Commission.

Kirchgässner, G. (1993). Testing Weak Rationality of Forecasts with Different Time Horizons. *Journal of Forecasting*, 12(7), 541–558, http://dx.doi.org/10.1002/for.3980120702

Makridakis, S., Hibon, M. (1995). *Evaluating Accuracy (or Error) Measures*. INSEAD Working Paper. Available at: http://www.insead.edu/facultyresearch/research/doc.cfm?did=46875

McNees, S. K., Ries, J. (1983). The Track Record of Macroeconomic Forecasts. *New England Economic Review*, 18(5), 25–42.

Miao, W., Gel, Y. L., Gastwirth, J. L. (2006). A New Test of Symmetry about an Unknown Median, in Hsiung, A. C., Ying, Z., Zhang, C. H., eds., *Random Walk, Sequential Analysis and Related Topics – A Festschrift in Honor of Yuan-Snih Chow*. World Scientific Publisher, Singapore.

Milburn, T. W. (1978). *Successful and Unsuccessful Forecasting in International Relations. Forecasting in International Reations: Theory, Methods, Problems, Prospects*. Freeman, San Francisco, 79–91.

Ministry of Finance, Czech Republic (2013). *Makroekonomické predikce na MF ČR – pohled do zpětného zrcátka*. Macroeconomic Forecasts at MFCR – Rear – view Mirror. Prague: Ministry of Finance. Available at: http://www.mfcr.cz/assets/cs/media/Makro-ekonomicka-predikce_2013-Q3_Makroekonomicke-predikce-na-MF-CR-pohled-do-zpetneho-zrcatka-cervenec-2013.pdf

Ministry of Finance, Czech Republic (2014). *Makroekonomická predikce – leden 2014*. Macroeconomic Forecasts – January 2014. Prague: Ministry of Finance. Available at: http://www.mfcr.cz/assets/cs/media/Makro-ekonomicka-predikce_2014-Q1_Makroekonomicka-predikce-komplet-ke-stazeni.pdf

Mühleisen, M., Danninger, S., Hauner, D., Krajnyák, K., Sutton, B. (2005). *How do Canadian Budget Forecasts Compare with Those of Other Industrial Countries?* IMF Working Papers, 66(5), 1–49, http://dx.doi.org/10.5089/9781451860856.001

Novotný, F., Raková, M. (2011). Assessment of Consensus Forecasts Accuracy: The Czech National Bank Perspective. *Czech Journal of Economics and Finance (Finance a Úvěr)*, 61(4), 348–366. Available at: http://journal.fsv.cuni.cz/storage/1218_str_348_366._-_novotnypdf.pdf

Öller, L. E., Barot, B. (2000). The Accuracy of European Growth and Inflation Forecasts. *International Journal of Forecasting*, 16(3), 293–315, http://dx.doi.org/10.1016/s0169-2070(00)00044-3

Page, E. B. (1963). Ordered Hypotheses for Multiple Treatments: A Significance Test for Linear Ranks. *Journal of the American Statistical Association*, 58(301), 216–230, http://dx.doi.org/10.1080/01621459.1963.10500843

Polák, Z. (2011). *Evaluation of Macroeconomic Forecasting Accuracy.* Bachelor Thesis, Charles University in Prague. Available at: http://ies.fsv.cuni.cz/default/file/download/id/18390

Soukup, J. (2012). The Accuracy of Macroeconomic Forecasts in the years 2006–2011, in Löster, T., Pavelka, T., ed., The *6th International Days of Statistics and Economics – Conference Proceedings*. Melandrium, Prague, pp. 1043–1053.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6), 80–83, http://dx.doi.org/10.2307/3001968