LATVIJAS BANKA
EIROSISTĒMA

ISBN 978-9934-578-23-6

BORISS SILIVERSTOVS
DANIEL S. WOCHNER

WORKING PAPER
2 / 2020

RECESSIONS AS BREADWINNER FOR FORECASTERS STATE-DEPENDENT EVALUATION OF PREDICTIVE ABILITY: EVIDENCE FROM BIG MACROECONOMIC US DATA

Latvijas Banka
K. Valdemāra iela 2A, Riga, LV-1050
Tel.: +371 67022300 info@bank.lv
http://www.bank.lv https://www.macroeconomics.lv

CONTENTS

ABBREVIATIONS

AR – autoregressive
BVAR – Bayesian vector autoregressive
CADL – Combined Autoregressive Distributed Lag
CPI – Consumer Price Index
CSSFED – Cumulated Squared Sum of Forecast Error Difference
DFM – Dynamic Factor Model
DM – Diebold–Mariano
DSGE – Dynamic Stochastic General Equilibrium
FRED-QD – Federal Reserve Economic Database (Quarterly Data)
GDP – Gross Domestic Product
GLS – Generalised Least Squares
HMN – Historic Mean
MSFE – Mean Squared Forecast Error
NBER – National Bureau of Economic Research
OLS – Ordinary Least Squares
RMSFE – root mean squared forecast error
rRMSFE – relative root mean squared forecast error
rrRMSFE – ratio of relative root mean squared forecast errors
RW – random walk
SFED – Squared Forecast Error Difference
SURE – Seemingly Unrelated Regression Equations
UNI-AR – univariate autoregressive
US – United States of America
VAR – vector autoregression
vs. – versus

ABSTRACT

This paper re-examines the findings of Stock and Watson (2012b) who assessed the predictive performance of DFMs over AR benchmarks for hundreds of target variables by focusing on possible business cycle performance asymmetries in the spirit of Chauvet and Potter (2013) and Siliverstovs (2017a; 2017b; 2020). Our forecasting experiment is based on a novel big macroeconomic dataset (FRED-QD) comprising over 200 quarterly indicators for almost 60 years (1960–2018; see, e.g. McCracken and Ng (2019b)). Our results are consistent with this nascent state-dependent evaluation literature and generalize their relevance to a large number of indicators. We document systematic model performance differences across business cycles (longitudinal) as well as variable groups (cross-sectional). While the absolute size of prediction errors tend to be larger in busts than in booms for both DFMs and ARs, DFMs relative improvement over ARs is typically large and statistically significant during recessions but not during expansions (see, e.g. Chauvet and Potter (2013)). Our findings further suggest that the widespread practice of relying on *full sample* forecast evaluation metrics may not be ideal, i.e. for at least two thirds of all 216 macroeconomic indicators full sample rRMSFEs systematically over-estimate performance in expansionary subsamples and under-estimate it in recessionary subsamples (see, e.g. Siliverstovs (2017a; 2020)). These findings are robust to several alternative specifications and have high practical relevance for both consumers and producers of model-based economic forecasts.

Keywords: forecast evaluation, dynamic factor models, business cycle asymmetries, big macroeconomic datasets, US

JEL codes: C32, C45, C52, E17

The authors gratefully acknowledge helpful inputs and discussions from Jan-Egbert Sturm, Jacqueson Galimberti, Simon Hilber, Philipp Baumann and Stefan Pichler. As explained in the text, this paper builds upon previous work of one of the authors on forecasting performance asymmetries across business cycle phases, and parts of the paper therefore share similarities with these contributions (see, e.g. Siliverstovs (2017a; 2017b; 2020)). The paper has further benefitted from valuable inputs and comments from anonymous referees as well as participants of the SkiLLS Meeting 2019 (Engelberg, Switzerland), the KOF Research Seminar (Zurich, Switzerland), the 23rd International Conference on Macroeconomic Analysis and International Finance (Rethymno, Greece), the 10th Nordic Econometrics Meeting (Stockholm, Sweden), the 2nd Baltic Economic Conference (Riga, Latvia) as well as the OeNB Summer School 2019 (Vienna, Austria). The views expressed in this paper are those of the authors and do not necessarily reflect the views of Latvijas Banka or KOF Swiss Economic Institute.

Boriss Siliverstovs: Latvijas Banka, K. Valdemāra iela 2A, Riga, LV-1050, Latvia; e-mail: Boriss.Siliverstovs@bank.lv, KOF Swiss Economic Institute at ETH Zurich, 8092 Zurich, Switzerland.
Daniel Wochner: KOF Swiss Economic Institute at ETH Zurich, 8092 Zurich, Switzerland; e-mail: wochner@kof.ethz.ch.

1. INTRODUCTION

Over the past two decades, the forecasting literature proposed numerous econometric approaches to account for time-series peculiarities across business cycles, such as regime shifts, non-linear dynamics, non-Gaussianity and stochastic volatility to name a few (see, e.g. Chan (2017), Chan and Hsiao (2014), Koop and Korobilis (2010), Stock and Watson (2016; 2017) and Teräsvirta (2018) for comprehensive recent discussions). While these advances yielded improvements for in-sample estimations of forecasting models, the forecasting literature has so far devoted less attention to the very same kind of business cycle asymmetries for the out-of-sample performance evaluations of forecasting models (see Siliverstovs (2020)).

Recently, however, a burgeoning literature started to account for possible performance differences between distinct phases of the business cycle. In a notable chapter "Forecasting Output" in the Handbook of Economic Forecasting, Chauvet and Potter (2013) challenge the widespread standard of assessing models' forecasting performance in terms of simple averages over the *full* evaluation period because it implicitly discards possible business cycle-related performance asymmetries and may obscure potential differences in predictive performance (see also Siliverstovs (2017a)). Chauvet and Potter (2013) therefore complement their full sample assessments with evaluations that examine a model's performance for both expansionary and recessionary subsamples *separately* and find systematic performance differences between these phases of the business cycle for a large number of frequently employed macroeconomic forecasting models. Their forecasting exercise of US GDP growth over pre- and post-millennial years (1992–2011) yields two main conclusions: on the one hand, they find recessionary periods to be generally harder to predict than the expansionary ones as the models typically produce larger *absolute* forecasting errors in the former rather than latter periods. On the other hand, they show the *relative* improvement of sophisticated forecasting models over AR benchmarks to be typically much more pronounced in bust than in boom cycles. In fact, they find that during booming periods the performance of sophisticated models is often identical to that of simple AR benchmarks or even worse (see also, e.g. Fossati (2018), Kim and Swanson (2016) for related findings and discussions).

These asymmetries were recently corroborated in Siliverstovs (2020) who extends Chauvet and Potter (2013)'s single-frequency data setting with point forecasts to a mixed-frequency environment with both point and density forecasts. Likewise, Siliverstovs (2017b) extends the subsample evaluation for booms and busts to 10 industrialized countries when assessing the predictability of excess returns of stock markets. Siliverstovs (2017a; 2020) refines Chauvet and Potter (2013)'s subsample analyses (which we will refer to as 'meso' level) even further by opening the time dimension to the full extent and employing evaluation metrics that assess the contribution of each individual time period ('micro' level) to the overall predictive performance.

More generally, our paper also relates to the literature on forecasting under instability. Rossi (2013) provides a comprehensive overview of common findings in the forecasting literature and reviews numerous techniques to diagnose and tackle forecasting in the presence of instabilities. The author discusses forecasting performance asymmetries for a few series on a case-by-case basis (see Rossi (2013, p. 1223 ff.), while the recent macroeconomic forecasting literature has started to

examine model performance not only for a single but dozens or even hundreds of macroeconomic target variables (Korobilis (2017), Pesaran et al. (2011), Stock and Watson (2012b)). However, this multi-target literature has so far devoted only little attention to potential business cycle-related asymmetries in their forecast evaluations (see Siliverstovs (2020)), and the literature lacks a systematic and thorough state-dependent forecast evaluation for a broad number of different macroeconomic indicators.

We close this gap in the literature by bridging these distinct strands of the literature and perform a direct forecasting experiment in (quasi) real-time for quarterly, semi-annual and annual forecasting horizons based on a novel large-scale macroeconomic dataset of the US economy (FRED-QD) with 216 indicators over almost 60 years (1960–2018) (see, e.g. McCracken (2019), McCracken and Ng (2019b)). We will focus on DFMs, which constitute a leading class of forecasting models for large macroeconomic datasets and were repeatedly shown to perform well in a range of settings and against a range of competing approaches (see, e.g. Stock and Watson (2012b; 2017), Korobilis (2017), Chauvet and Potter (2013) and references therein; see also Stock and Watson (2006; 2016)). Hence, this paper advances the current state of research in three main ways. First, we contribute to the nascent boom and bust evaluation literature by extending the analyses of Chauvet and Potter (2013) and Siliverstovs (2017a; 2017b; 2020) with systematic assessments of dynamic factor models for not only a single but over 200 US macroeconomic indicators. This allows us to examine whether their findings can be generalized to alternative key indicators (e.g. unemployment rate, inflation, exports) as well as entire groups of macroeconomic indicators (e.g. interest rates, housing, industrial production). Second, the paper advances the recent factor model literature on multiple target variables (Korobilis (2017), Pesaran et al. (2011), Stock and Watson (2012b)) by evaluating the models' forecasting performance separately for booms and busts rather than the full sample. We may therefore determine the sources of predictive performance in greater granularity than previous studies. Third, multi-period ahead direct forecasting models with AR predictors typically suffer from inconsistent and inefficient estimation because overlaps in datasets cause serial dependencies among the model's residuals (Pesaran et al. (2011)). While most studies above did not explicitly account for these issues, we will examine to what extent the results are affected as we address these serial correlations with the GLS-based SURE estimation procedure of Pesaran et al. (2011).

On the whole, our results are consistent with the recent business cycle-related evaluation literature (e.g. Chauvet and Potter (2013), Siliverstovs (2017a; 2017b; 2020), also Stock and Watson (2011)) and extend their relevance to a large number of macroeconomic indicators in general as well as to a few selected leading indicators in particular. Specifically, we find systematic model performance differences across phases of business cycles (longitudinal) as well as different groups of variables (cross-sectional). For a majority of series, we find dynamic factor models to achieve strong and statistically significant improvements over benchmarks during recessionary but not during expansionary periods (see, e.g. Chauvet and Potter (2013), Siliverstovs (2017b)). Performance metrics and tests for the full sample thus tend to be decisively distorted towards the performance in recessions and may qualify models as overall significantly superior to the benchmark when they are in fact inferior most of the time (see, e.g. Siliverstovs (2020) and Fossati (2018)). In light of the importance of forecasts for policy decision making (Wieland and Wolters (2013)), the disregard of

such asymmetries may lead to suboptimal or even erroneous decisions, which matters to executives and researchers at governments, central banks and businesses alike (see Siliverstovs (2017a; 2020)).

The remainder of this paper is organized as follows: Section 2 describes the dataset and relevant data pre-processing steps. Section 3 lays out the formal forecasting framework and predictive models, the estimation and evaluation methods used and builds mainly upon Stock and Watson (2012b), Chauvet and Potter (2013) and Siliverstovs (2017a; 2017b; 2020). Section 4 systematically assesses the results of the forecasting experiment for the main model specification on three different levels of granularity (macro, meso and micro level). Section 4 also shows the results for the extended specification using the SURE estimator as well as multiple robustness checks to the main specification. Section 5 concludes and provides directions for future research.

2. DATA

2.1 Observations and variables

The analysis is based on the FRED provided by the Federal Reserve Bank of St. Louis. Specifically, we employ the novel quarterly dataset FRED-QD (see McCracken (2019), McCracken and Ng (2019b); for the monthly series, see McCracken and Ng (2016; 2019a)).[1] This dataset comprises over two hundred macroeconomic indicators for the US and aims to provide researchers' access to a regularly updated version of the Stock and Watson (2012a) dataset (see discussions in McCracken (2019) and McCracken and Ng (2016)).

Three common data transformations were applied prior to empirical analysis. First, following Korobilis (2017) (who also employed the FRED-QD dataset) and Stock and Watson (2012b), we apply distinct stationarity transformations for independent and dependent variables (see Table A.1). The specific transformations are directly suggested in McCracken and Ng (2019b). Second, to cope with the unbalanced nature of the dataset, indicators whose data were not available in the first quarter of 1960 or had ragged edges were not considered. Finally, the dataset is cleaned for outliers as proposed by Stock and Watson (2012b) (see their online appendix).[2] There were 106, 106, 90 and 63 outliers corrected in $\mathbf{X}_t$, $\mathbf{Y}_{t+1}^{(1)}$, $\mathbf{Y}_{t+2}^{(2)}$ and $\mathbf{Y}_{t+4}^{(4)}$ respectively – which amounts to less than 0.25% of observations of the whole dataset.

These data preparations yield a balanced dataset that contains $K = 216$ variables over $T = 233$ quarters (from the second quarter of 1960 to the second quarter of 2018). McCracken and Ng (2019b) divide these 216 indicators into 14 distinct and differently-sized groups as displayed in Table 1.

[1] The dataset is novel in that it was released as part of their new service (launched in May 2018) for real-time updates of their quarterly dataset FRED-QD (see McCracken (2019)).

[2] Specifically, they propose to classify an observation of the stationarity transformed series as an outlier if its absolute median deviation is larger than six times the series' interquartile range and were then replaced by the median of the previous five values (see Stock and Watson (2012b), online appendix, Section B).

Table 1
FRED-QD time series groups

Group	Group name	Number of time series
G1	National Income and Product Accounts (NIPAs)	22
G2	Industrial Production	15
G3	Employment and Unemployment	44
G4	Housing	11
G5	Inventories, Orders and Sales	6
G6	Prices	46
G7	Earnings and Productivity	10
G8	Interest Rates	18
G9	Money and Credit	14
G10	Households Balance Sheets	9
G11	Exchange Rates	4
G12	Other	1
G13	Stock Prices	5
G14	Non-household Balance Sheets	11
ALL	Total	216

Notes. This table illustrates how the 216 variables used for the analysis are subsumed into 14 distinct groups according to McCracken and Ng (2019b). The table provides the group abbreviations used, group names and the number of time series belonging to each group. For certain tables and graphical illustrations below, the mnemonics of each series were augmented by their group abbreviations, e.g. the indicator for real Gross Domestic Product GDPC1 becomes G1_GDPC1.

2.2 Recessions and expansions

In line with the state-dependent forecast evaluation literature (e.g. Chauvet and Potter (2013)), we divide the full sample into recessionary and expansionary subsamples by employing the definitions of peaks and troughs provided by the Business Cycle Committee of the NBER.[3]

For one-quarter ahead predictions, our definitions of recessions and expansions correspond exactly to the ones of NBER. However, as we not only applied the stationarity transformations to the independent but also the dependent variables, the recession and expansion dates ought to be suitably adjusted at higher forecasting horizons. To see this, consider Figure 1a, which displays the stationarity transformed series of real GDP over the period of the 2008 financial crisis for all forecasting horizons. As shown in Figure 1a, the trough of the stationarity transformed series shifts to the right as h increases. While the original definition of NBER recessions captures the crisis fairly well for $h = 1$, it does less well for $h = 2$ and $h = 4$.[4]

Based on these considerations, at least two adjustments of the NBER recession definitions appear sensible. Our first alternative extends the NBER recession end-date by $(h - 1)$ quarters to the right, such that the recession windows grow as h increases (hereafter, the "growing" adjustment). The resulting recession window is depicted in Figure 1b for $h = 4$ and shows a greater ability to reflect the prolonged period of crisis than the original version in Figure 1a. Moreover, it has an additional benefit of increasing the limited number of recessionary time periods for a few additional

[3] Source: https://www.nber.org/cycles.html. Observe that we will use the notions of recessions, busts, downturns and contractions as well as expansions, booms, upswings and growth periods synonymously.
[4] For $h = 8$ (not shown in Figure 1a), the trough would even lie beyond the NBER dated recessionary window, which is clearly sub-optimal.

observations. As a second alternative, rather than growing the window, we may shift both the start- and end-periods of the original NBER recession ($h-1$) periods ahead (hereafter, the "shifting" adjustment). This variant is depicted in Figure 1c. While we will examine all three versions (without growing and shifting adjustments) and show robustness of the results for each of them (see Subsection 4.3.1), the main results are based on the first alternative in which all recessionary periods in the forecasting window are grown. The expansionary periods then correspond to the remaining observations in the forecasting window.

3. FORECASTING FRAMEWORK

3.1 Notation and setup

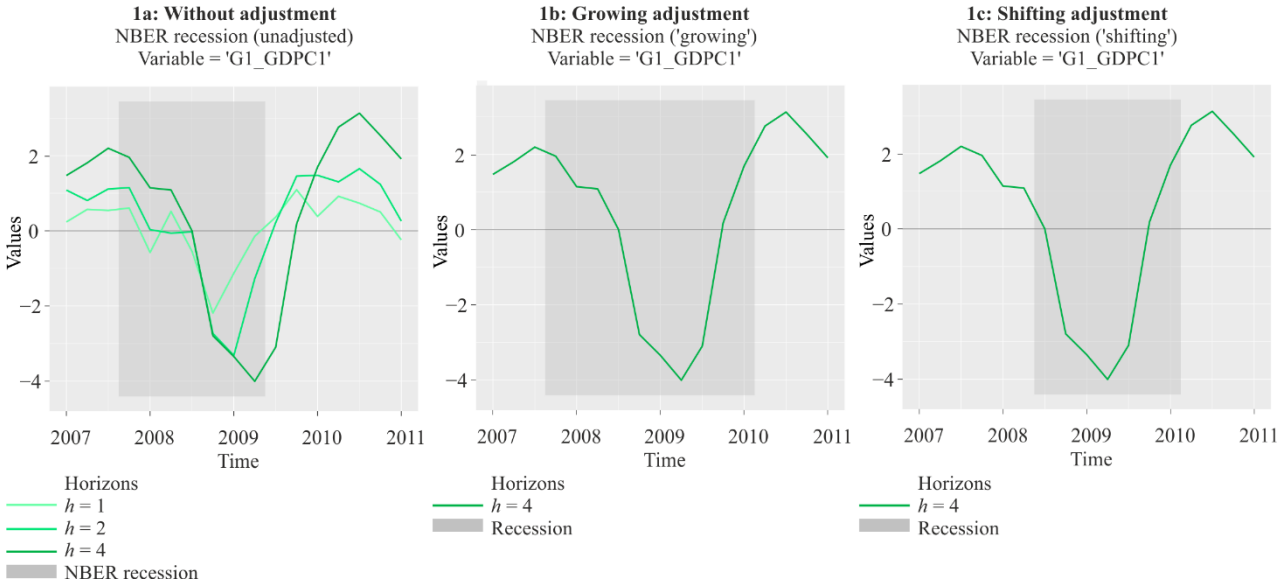
For the formal description of the forecasting exercise, we borrow the notation and methodical conventions applied in the relevant literature (see, e.g. Chauvet and Potter (2013), McCracken (2007), Pesaran et al. (2011), Stock and Watson (2006; 2012b; 2016), Kim and Swanson (2014), Siliverstovs (2017a; 2020) among many others). Let $t \in \{1, \dots, T\}$ be the quarterly time index and $h \in \{1,2,4\}$ the quarterly forecasting horizon. The analyses divide the timeline into an estimation window (from the second quarter of 1960 to the fourth quarter of 1984) and a forecasting window (from the first quarter of 1985 to the second quarter of 2018). Let S denote the last observation of the first estimation window (the fourth quarter of 1984, $S = 99$). Under a recursively expanding estimation scheme, it contains observations $t \in \mathcal{S}_e = \{1 + h, \dots, \tau\}$ with $\tau \in \{S, S+1, \dots, T-h\}$,whereas a rolling estimation scheme keeps the estimation window at a fixed length. We distinguish two different forecasting windows: the forecasting window containing the target date observations is given as $t \in \mathcal{S}_f^t = \{S + h, \dots, T\}$ and the forecasting window containing the time periods at which the forecasts are originated is $t \in \mathcal{S}_f^o = \{S+1, \dots, T-h+1\}$. The forecasting windows are of size $P = |\mathcal{S}_f^t| = |\mathcal{S}_f^o| = T - (S+h) + 1$ where $|\cdot|$ denotes the cardinality of set $\mathcal{S}_f$. The recessionary periods are determined according to our previous discussion (see Subsection 2.2) and denoted as $\mathcal{S}_{f,rec}^t$ and $\mathcal{S}_{f,rec}^o$ respectively. The set of expansionary periods is then given as the remaining set of observations, $\mathcal{S}_{f,exp}^t = \mathcal{S}_f^t \backslash \mathcal{S}_{f,rec}^t$ and $\mathcal{S}_{f,exp}^o = \mathcal{S}_f^o \backslash \mathcal{S}_{f,rec}^o$.

Concerning the variables, let $\mathbf{Z}_t^{(h)} = (Y_t^{(h)}, \mathbf{W}_t) \in \mathbb{R}^K$ designate the vector of stationarity transformed values of the 216 indicators in period t (vectors and matrices are consistently denoted in bold letters). Let the k-th variable in $\mathbf{Z}_t^{(h)}$ be the dependent variable, $Y_t^{(h)} = Z_{k,t}^{(h)}$, and the remaining ones the independent variables, $\mathbf{W}_t = \mathbf{Z}_{-k,t}^{(h)}$, with index $-k$ indicating all but the k-th variable.

In line with Stock and Watson (2012b), we avoid double counting and extract the common factors only from non-aggregated, lower-level variables. Hence, the set of independent variables $\mathbf{W}_t = (\mathbf{X}_t, \widetilde{\mathbf{X}}_t)$ can further be subdivided into 102 series, $\mathbf{X}_t$ is used to derive the common factors and the remaining 113 series in $\widetilde{\mathbf{X}}_t$ (assuming that the current dependent variable k is an aggregated variable; otherwise there are 101 series in $\mathbf{X}_t$ and 114 in $\widetilde{\mathbf{X}}_t$; mccrackenng2019online_QD explicitly indicate which indicators belong to $\mathbf{X}_t$ and $\widetilde{\mathbf{X}}$ respectively). As each of the 216 variables will serve as a dependent variable, the $\widetilde{\mathbf{X}}_t$ will not be used for factor estimation for the current k but will only serve as dependent variables for alternative k. Finally, as a result of the

distinct stationarity transformations for dependent and independent variables (see Table 14), the super-index (h) indicates the base frequency measurement arising from h-quarter stationarity transformation of the dependent variable (e.g. quarterly, semi-annual, annual growth rates for $h = 1, h = 2, h = 4$ respectively; see Korobilis (2017), Stock and Watson (2012b)). As explained above, the observable right-hand-side variables, instead, are all measured on a quarterly basis.

Figure 1
Recession adjustments



Notes. Recessionary periods are highlighted in grey. Subfigure (a) shows the stationarity transformed series of real GDP from the FRED-QD dataset (FRED-QD mnemonic: GDPC1) over the period of the 2008 financial crisis for all forecasting horizons and the original NBER recessions as grey shaded areas (no adjustment). Subfigures (b) and (c), instead, show the NBER recessions for 'growing' and 'shifting' adjustments proposed in Subsection 2.2.

3.2 Forecasting models

While the literature has put forth a comprehensive number of forecasting models, we devote our attention to factor-augmented models, which consistently show good empirical performance, particularly in high-dimensional settings (see Chauvet and Potter (2013), Stock and Watson (2012b; 2017) for comprehensive discussions). For the formulation of our main two forecasting models, their estimation and evaluation, we follow primarily Stock and Watson (2012b) and construct direct h-period ahead forecasts, $\hat{Y}_{t+h}^{(h)}$, from AR factor-augmented models (DFM5) and AR benchmarks (AR4). Popular alternative benchmark specifications are considered as extensions to these two main models: namely, the HMN (Siliverstovs (2017b)), univariate AR process (AR1) (Korobilis (2017)) as well as CADL models (Kim and Swanson (2014)).

3.2.1 Factor models

Autoregressive factor-augmented model (DFM5)

The AR factor-augmented model is given as

$$Y_{t+h}^{(h)} = \varphi + \sum_{p=0}^{P-1} \phi_p Y_{t-p}^{(1)} + \sum_{r=1}^{R} \psi_r F_{r,t}^{(1)} + \varepsilon_{t+h}^{(h)} \quad (1)$$

where $Y_{t+h}^{(h)}$ denotes the h-period ahead value of the dependent variable, $Y_{t-p}^{(1)}$ – the pth AR lag, $F_{r,t}^{(1)}$ – the r-th common factor and $\varepsilon_{t+h}^{(h)}$ – the stochastic error term (see, e.g. Korobilis (2017), Stock and Watson (2002a; 2002b; 2006; 2012b)). The factors can be estimated and derived as the first *R* principal components, $\hat{F}_{r,t}^{(1)}$ standardized (mean zero, unit variance) set of predictors $\mathbf{X}_t^{(1)}$ (see Stock and Watson (2012b), also Stock and Watson (2006; 2016; R Core Team (2019) for implementation).

Our main results are based on a specification with $P = 4$ AR terms and $R = 5$ factors (AR4-DFM5 or simply DFM5); alternative specifications are considered as extensions to this main model.

3.2.2 Benchmark models

Autoregressive benchmark model (AR4, AR1)

To assess the performance of the DFM5 against univariate AR processes, define

$$Y_{t+h}^{(h)} = \varphi + \sum_{p=0}^{P-1} \phi_p Y_{t-p}^{(1)} + \varepsilon_{t+h}^{(h)} \quad (2)$$

with $P = 4$ lags, such that the main AR4 benchmark model is nested in the DFM5 above (see, e.g. Stock and Watson (2012b)). Moreover, we will also consider $P = 1$ as an extension to enhance comparison of the results with Korobilis (2017) who uses AR1 benchmarks.

HMN model

The HMN model is a simple and robust benchmark specification:

$$Y_{t+h}^{(h)} = \varphi + \varepsilon_{t+h}^{(h)} \quad (3)$$

and corresponds to the reduced form of an AR benchmark with $P = 0$ AR terms (see, e.g. Hill et al. (2011), Siliverstovs (2017b)).

CADL model

While the previous benchmarks in equations (2) and (3) are exclusively based on information from the target variable itself, the CADL benchmark additionally incorporates contemporaneous information on the predictors $\mathbf{X}_t$ by combining forecasts of multiple AR distributed lag models, each of which is augmented with a single predictor (see Kim and Swanson (2014)). Formally, the CADL model can be defined as:

$$Y_{t+h}^{(h)} = \sum_{k=1}^{K} \omega_k \hat{Y}_{k,t+h}^{(h)} + \varepsilon_{t+h}^{(h)}, \quad \hat{Y}_{k,t+h}^{(h)} = \varphi + \sum_{p=0}^{P-1} \phi_p Y_{t-p}^{(1)} + \beta_k X_{k,t}^{(1)} \quad (4)$$

where the predictions of each AR distributed lag model, $\hat{Y}_{k,t+h}^{(h)}$, are equally weighted with $\omega_k = \omega = 1/K$ and $P = 4$ (see Kim and Swanson (2014)).

3.3 Model estimation

3.3.1 OLS estimation

In accord with most of the relevant literature above (see, e.g. Kim and Swanson (2014), Korobilis (2017), Siliverstovs (2017b), Stock and Watson (2012b)), our main results will estimate the forecasting models via OLS, the factors are re-estimated and the forecasting models in equations (1) to (4) are updated for every vintage date in the

forecasting window, $t \in \mathcal{S}_f^o$, each forecasting horizon, $h \in \{1,2,4\}$, and every target variable, $k \in \{1, \dots, K\}$.

3.3.2 SURE estimation

While the OLS-based estimation allows us to compare our results with Stock and Watson (2012b) in particular and the existing literature in general, the model parameters may neither be consistently nor efficiently estimated (Pesaran et al. (2011)). This is because the overlaps of datasets for direct multi-step ahead predictions ($h > 1$) induce serially dependent error terms with an MA($h-1$) structure in our main equations (1) and (2) (Pesaran et al. (2011), PPT heretofore). To remove the serial dependence in the errors, we employ PPT's SURE estimator. Their GLS-based estimator offers not only attractive statistical properties of asymptotic consistency and efficiency but also provides a fairly simple implementation, which is mainly based on re-ordering of observations (to re-establish consistency) and pooled estimation that exploits the cross-dependence of errors (to re-establish efficiency) (see Pesaran et al. (2011), Section 3 for details).[5] Notice, for $h = 1$ the errors do not exhibit a serial dependence and the OLS and SURE estimators are equivalent (see Pesaran et al. (2011)).

3.4 Model evaluation

The estimated models $m \in \{1, \dots, M\}$ are used to derive h-period ahead out-of-sample forecasts, $\hat{Y}_{t+h,m}^{(h)}$, for each forecast origin in the forecasting window, $t \in \mathcal{S}_f^o$ (see, e.g. Siliverstovs (2017b), Stock and Watson (2012b)). The accuracies of these predictions will then be assessed on three different levels of temporal granularity (macro, meso, micro) as outlined below.

3.4.1 Macro level: Aggregate accuracy measures

We assess the absolute forecasting performance of model m in terms of its RMSFE:

$$\text{RMSFE}_m^{(h)} = \left[\frac{1}{(T-h)-S+1}\sum_{t=S}^{T-h}\left(Y_{t+h}^{(h)} - \hat{Y}_{t+h,m}^{(h)}\right)^2\right]^{1/2} \tag{5}$$

(Korobilis (2017), Stock and Watson (2012b)). The forecasting performance of the m-th model is assessed against the b-th benchmark with rRMSFE:

$$\text{rRMSFE}_{m,b}^{(h)} = \frac{\text{RMSFE}_m^{(h)}}{\text{RMSFE}_b^{(h)}} \tag{6}$$

[5] Our implementation of the SURE estimator's covariance matrix, $\delta_h(\ell)$, uses Bartlett weights, which are given as $\omega_h(\ell) = 1 - 1_{\{h>1\}}\ell/(h+1)$ (with $\ell \in \{0,1, \dots, h\}$, denoting the lag and $1_{\{h>1\}}$ as binary indicator function equal to one if $h > 1$ and zero else (see, e.g. Newey and West (1994, p. 640)) in order to down-weigh the auto-covariances of the error terms, $\gamma_h(\ell)$, for $h > 1$, as suggested in Pesaran et al. (2011, p. 176). The auto-covariances between $\varepsilon_{t+h}^{(h)}$ and $\varepsilon_{t+h-\ell}^{(h)}$ can be consistently estimated via $\hat{\gamma}_h(\ell) = 1/\tau\sum_{t=1}^{\tau}\hat{\varepsilon}_{t+h}^{(h)}\hat{\varepsilon}_{t+h-\ell}^{(h)}$ (ibid., p. 176; notice: should the variances still be negative despite down-weighting, the SURE estimator was re-estimated by dropping the first observation of the sample until the variances are non-negative). Hence, the auto-covariances used in our implementation are given as $\delta_h(\ell, h) = \omega_h(\ell)\gamma_h(\ell)$.

for $b \in \{1, \dots, M\}, b \neq m$ and $\text{rRMSFE}_{m,b}^{(h)} < 1$ indicates superior average performance of model m compared to model b; and vice versa for $\text{rRMSFE}_{m,b}^{(h)} > 1$ (see, e.g. Siliverstovs (2017a; 2020), Stock and Watson (2006)).

Additionally, we may decompose the MSFE into its bias, variance and covariance components:

$$\text{MSFE}_m^{(h)} = \underbrace{\left(\hat{\mu}_{Y_{t+h}^{(h)}} - \hat{\mu}_{\hat{Y}_{t+h,m}^{(h)}}\right)^2}_{\text{Bias}} + \underbrace{\left(\hat{\sigma}_{Y_{t+h}^{(h)}} - \hat{\sigma}_{\hat{Y}_{t+h,m}^{(h)}}\right)^2}_{\text{Variance}} + \underbrace{2\left(1 - \hat{\rho}_{Y_{t+h}^{(h)}, \hat{Y}_{t+h,m}^{(h)}}\right)\hat{\sigma}_{Y_{t+h}^{(h)}}\hat{\sigma}_{\hat{Y}_{t+h,m}^{(h)}}}_{\text{Covariance}} \quad (7)$$

where $\hat{\mu}$, $\hat{\sigma}^2$ and $\hat{\rho}$ denote the sample mean, variance and correlation coefficients respectively (see Theil (1958, p. 34), Pindyck and Rubinfeld (1998, p. 210 ff.), Chauvet and Potter (2013)). The bias and variance components measure the systematic error of the forecasts and account for the extent to which the averages and variations of predictions differ from those of the realizations, whereas the covariance component measures the remaining unsystematic error arising from imperfect comovement between predictions and realizations (Theil (1958, p. 34 ff.), Pindyck and Rubinfeld (1998, p. 211)).

3.4.2 Meso level: Disaggregate accuracy measure (subset RMSFE)

By averaging across the full sample, aggregate measures as in (6) implicitly discard the fact that the forecasting performance of any two models may differ over time (e.g. Chauvet and Potter (2013), Siliverstovs (2017a; 2017b; 2020)). To obtain a more granular understanding about the forecasting performance in different time periods, we employ two distinct approaches proposed in the recent forecasting evaluation literature. First, Chauvet and Potter (2013) proposed to assess the average forecasting performance for recessionary and expansionary subsamples separately by computing:

$$\text{RMSFE}_{m,rec}^{(h)} = \left[\frac{1}{|\mathcal{S}_{f,rec}^o|}\sum_{t \in \mathcal{S}_{f,rec}^o}(Y_{t+h}^{(h)} - \hat{Y}_{t+h,m}^{(h)})^2\right]^{1/2} \quad (8a)$$

and

$$\text{RMSFE}_{m,exp}^{(h)} = \left[\frac{1}{|\mathcal{S}_{f,exp}^o|}\sum_{t \in \mathcal{S}_{f,exp}^o}(Y_{t+h}^{(h)} - \hat{Y}_{t+h,m}^{(h)})^2\right]^{1/2} \quad (8b)$$

where $|\cdot|$ denotes the number of elements in set $\mathcal{S}_{f,s}^o$ (cardinality) for subsample $s \in \{\text{exp}, \text{rec}\}$. The relative subset RMSFE for a suitable benchmark b may then be defined analogously to equation (6) (see, e.g. Chauvet and Potter (2013)):

$$\text{rRMSFE}_{m,b,s}^{(h)} = \frac{\text{RMSFE}_{m,s}^{(h)}}{\text{RMSFE}_{b,s}^{(h)}} \quad (9).$$

In analogy to equation (6), $\text{rRMSFE}_{m,b,s}^{(h)} < 1$ indicates superior average performance of model m compared to benchmark b over the subsample period s; and conversely for $\text{rRMSFE}_{m,b,s}^{(h)} > 1$ (see Subsection 3.4.1). Furthermore, we may be interested in

the rrRMSFE over subsample s compared to the rRMSFE for the full sample, i.e. the ratio of equation (8) or (9) to equation (6):

$$\text{rrRMSFE}_{m,b,s}^{(h)} = \frac{\text{rRMSFE}_{m,b,s}^{(h)}}{\text{rRMSFE}_{m,b}^{(h)}} \tag{10}.$$

For the recessionary sub-sample, $\text{rrRMSFE}_{m,b,rec}^{(h)} < 1$ indicates that the relative improvement over the benchmark during recessionary periods is more pronounced than for the full sample and vice versa for the case where $\text{rrRMSFE}_{m,b,rec}^{(h)} > 1$.

3.4.3 Micro level: Disaggregate accuracy measure (scaled SFED)

An alternative approach is employed in Siliverstovs (2017a; 2020), which is based on Welch and Goyal (2008) and enables an even more granular view of the forecasting performance by determining the contribution of each individual forecast to the overall RMSFE. Specifically, they consider measures based on the SFEDs:

$$\eta_{t+h,b,m}^{(h)} \equiv \text{SFED}_{t+h,b,m}^{(h)} = (\varepsilon_{t+h,b}^{(h)})^2 - (\varepsilon_{t+h,m}^{(h)})^2 \tag{11},$$

such that $\eta_{t,b,m}^{(h)} > 0$ indicates superior performance of model m in period t compared to benchmark b and vice versa for $\eta_{t,b,m}^{(h)} < 0$ (Siliverstovs (2017a, p. 294 ff.)).[6] The CSSFEDs from time t_0 to t_1 reads:

$$\text{CSSFED}_{t_0,t_1}^{(h)} = \sum_{t=t_0}^{t_1} \eta_{t+h,b,m}^{(h)},\ t_0, t_1 \in \mathcal{S}_f^o,\ t_0 \le t_1 \tag{12},$$

such that $\text{CSSFED}_{t_0=S+1,t_1=T-h+1}^{(h)} > 0$ is equivalent to $\text{rRMSFE}_{b,m}^{(h)} < 1$ and indicates superior overall performance of model m compared with benchmark b and vice versa for $\text{CSSFED}_{t_0=S+1,t_1=T-h+1}^{(h)} < 0$ (see Siliverstovs (2017a; 2020)).

3.4.4 Testing superior predictive ability

We test the models' superior forecasting performance by means of the widely used Diebold–Mariano (1995) tests, which posit under the null equal predictive accuracy between models m and b:

$$\text{H}_0\text{: E}\left[\varepsilon_{t+h,b}^{(h)}\right] = \text{E}\left[\varepsilon_{t+h,m}^{(h)}\right] \tag{13a}$$

and test it against the common alternative

$$\text{H}_\text{A}\text{: E}[\varepsilon_{t+h,b}^{(h)}] > \text{E}[\varepsilon_{t+h,m}^{(h)}] \tag{13b}$$

for all $t \in \mathcal{S}_f^o$, and the used test statistic employs heteroscedasticity and autocorrelation consistent standard errors for $h > 1$ (see Hyndman et al. (2019), Zeileis et al. (2019) and references therein). Likewise, we apply the tests for the recessionary and expansionary subsets, $\mathcal{S}_{f,exp}^o$ and $\mathcal{S}_{f,rec}^o$ respectively. Moreover,

[6] For better comparability of SFED, when multiple dependent variables are used, we normalize the variation among the different series and use scaled SFED, given as $\ddot{\eta}_{t,b,m}^{(h)} = \eta_{t,b,m}^{(h)} / \sqrt{\sigma_\eta^2}$ with $\sigma_\eta^2 = \text{Var}(\eta_{t,b,m}^{(h)})$.

where necessary, the tests account for the fact that models *m* and *b* are nested by using McCracken (2007)'s critical values for both recursive and rolling estimation schemes.

Table 2
Main results: Distribution of rRMSFE

Horizon	Sample	Percentiles of rRMSFE					rRMSFE	rrRMSFE
		5%	25%	50%	75%	95%	<1	<1
h = 1	Overall	0.852	0.958	1.000	1.022	1.107	49.5%	–
	Expansion	0.925	0.985	1.009	1.044	1.165	38.9%	32.9%
	Recession	0.587	0.787	0.953	1.036	1.183	62.0%	67.1%
h = 2	Overall	0.848	0.951	1.006	1.046	1.161	45.8%	–
	Expansion	0.943	1.001	1.038	1.113	1.284	25.0%	26.9%
	Recession	0.637	0.789	0.943	1.045	1.171	63.0%	73.1%
h = 4	Overall	0.840	0.934	1.006	1.047	1.186	47.7%	–
	Expansion	0.898	1.002	1.065	1.173	1.396	24.5%	19.4%
	Recession	0.674	0.800	0.913	1.024	1.194	69.4%	80.6%

Notes. The table entries show the percentiles of distributions of the rRMSFE of the DFM5 model against the AR4 benchmark for all forecasting horizons as in Stock and Watson (2012b) (settings: OLS estimation, recursive scheme, NBER recession adjustment 'grow'; first vintage: the first quarter of 1985). The 'overall' forecast evaluation sample ranges from the fourth quarter of 1984 + h until the second quarter of 2018. The table splits the overall sample into expansion and recession subsamples, as described in Subsection 2.2. The two last columns indicate the fraction of variables for which $\text{rRMSFE}^{(h)}_{\text{DFM5,AR4}} < 1$ and $\text{rrRMSFE}^{(h)}_{\text{DFM5,AR4},s} < 1$, $s \in \{\text{exp}, \text{rec}\}$. For more details, see Subsection 4.1.

4. RESULTS

Our main results in Subsection 4.1 examine the forecasting performance of the DFM5 model against the AR4 benchmark under OLS estimation. The extension to these main results in Subsection 4.2 considers the same models with SURE estimations, and the robustness results in Subsection 4.3 test a variety of alternative specifications.

4.1 Main results

4.1.1 Macro level: Full sample performance

Tables 2 and 3 summarize the main results of our forecasting exercise. To compare our results with those of Stock and Watson (2012b), Table 2 summarizes the distribution of rRMSFEs of all 216 variables similar to their table (ibid., see Table 2, p. 486) as well as the percentage of variables for which the rRMSFE and rrRMSFE are lower than unity (last two columns).

Table 2 indicates that our results for the overall evaluation sample are well in line with those of Stock and Watson (2012b) despite slight differences in forecasting setup.[7] Stock and Watson (2012b) report that the DFM5 model is more accurate than the AR4 benchmark for about half of all dependent variables. Our results for the full sample confirm these findings and show that the median rRMSFE is very close to unity at each horizon, and Table 3 indicates that for about 40% of all variables the improvement is statistically significant (at the level 5%). Likewise, in 5% of all cases

[7] Our setup differs from Stock and Watson (2012b) (see also the online supplement), for example, in terms of the dataset used (Global Insights and Conference Board's Indicators Database vs. FRED-QD), forecasting window (from the first quarter of 1985 to the fourth quarter of 2008 vs. the first quarter of 1985 to the second quarter of 2018), the number of predicted time series (143 vs. 216), the number of time series used for factor extraction (109 vs. 102).

the DFM5 yields an improvement over the AR4 of about 15% in terms of RMSFE, which is again comparable to Stock and Watson (2012b).

Table 3
Main results: Distribution of p-values

Horizon	Sample	DM-tests		
		P-values <1%	*P*-values <5%	*P*-values <10%
h = 1	Overall	25.50%	44.00%	52.30%
	Expansion	12.50%	28.70%	39.40%
	Recession	24.10%	38.00%	44.40%
h = 2	Overall	20.40%	40.30%	48.10%
	Expansion	8.80%	20.80%	25.00%
	Recession	21.80%	43.50%	50.90%
h = 4	Overall	20.40%	42.60%	50.50%
	Expansion	9.30%	20.40%	25.00%
	Recession	17.60%	48.10%	57.90%

Notes. The table shows the results of (one-sided) predictive ability tests when comparing the DFM5 model against the AR4 benchmark for all forecasting horizons (settings: OLS estimation, recursive scheme, NBER recession adjustment 'grow', first vintage: the first quarter of 1985). Specifically, the table entries provide the percentage of all variables for which the one-sided hypothesis of the Diebold and Mariano (1995) test are statistically significant at the 1%, 5% and 10% level respectively. To account for the nested model comparison, we use McCracken (2007)'s critical values and for $h>1$ heteroscedasticity and autocorrelation consistent standard errors are used (see Subsection 3.4.4). The table splits the overall sample into expansion and recession subsamples (see Subsection 2.2). For more details, see Section 4.

4.1.2 Meso level: Sub-sample performance

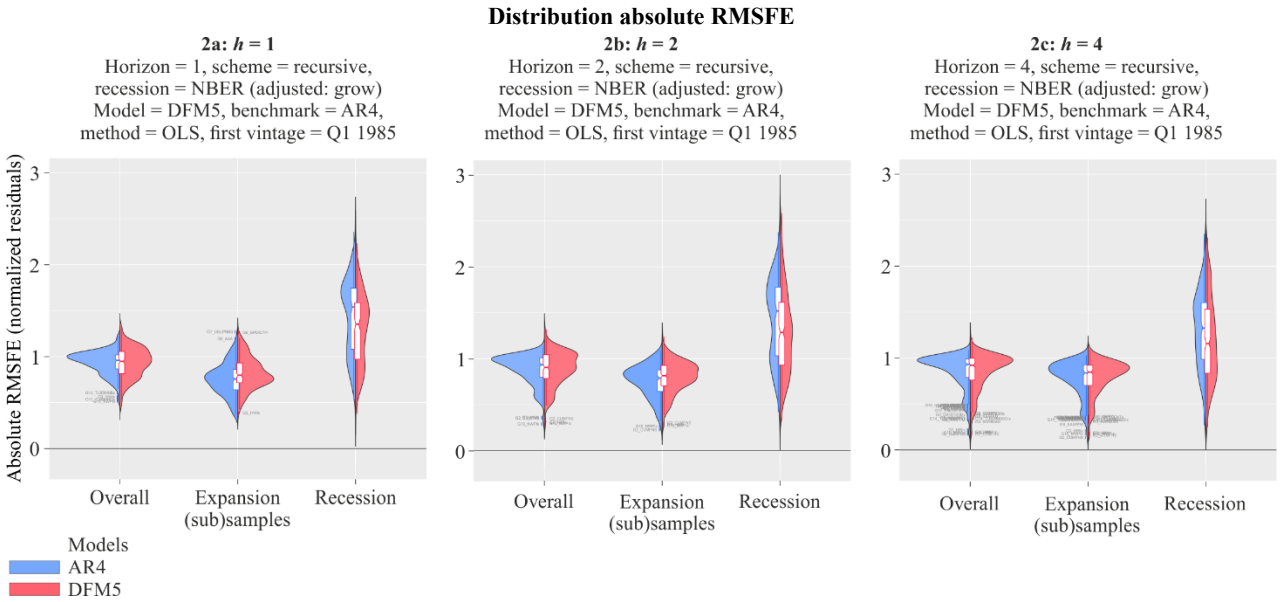
Univariate distributions: Separate subsamples over all variables

However, a more interesting picture emerges as we deviate from Stock and Watson (2012b) and examine the absolute and relative predictive performance for the subsets of expansionary and recessionary periods separately (see Subsections 3.4.2 and 3.4.3). In accord with previous research, we find pronounced performance asymmetries in both absolute and relative terms (e.g. Chauvet and Potter (2013), Siliverstovs (2017a, 2017b, 2020)). First, Figure 2 displays the distribution of absolute RMSFE for the DFM5 and AR4 for the full and both subsamples, which indicates that the absolute size of forecasting errors tends to be much larger in economic downturns than upswings (see Chauvet and Potter (2013)).

Second, in terms of the relative size of RMSFE, the results for the boom and bust subsamples show both systematic and substantial distributional shifts of rRMSFEs, i.e. in virtually all percentiles depicted in Table 2, the rRMSFE is lower for recessions than for expansions (see Chauvet and Potter (2013), Siliverstovs (2017b)).[8] For instance, for one-quarter ahead predictions the DFM5 for the top 5% of series improves the benchmark by at least 41.3% during recessions but only by 7.5% during expansions. Similar asymmetries in forecasting performance are observed at $h = 2$ and $h = 4$ (see Figure B.4). Moreover, the second last column in Table 2 displays the percentage of variables for which the DFM5 model improves upon the AR4 benchmark (inverse percentiles). While this share ranges between 25% and 39% during expansions, it ranges between 62% and 69% during recessions – a distributional shift with about twice as many variables that are more accurately predicted in turbulent times than in calm times.

[8] Figure B.4 in the Appendix visualizes the main results in Table 2 in terms of density plots of rRMSFE (overall, expansion vs. recession).

Figure 2
Distribution of absolute RMSFEs for AR4 and DFM5



Notes. The figures show the distributions of absolute RMSFEs as density plots and boxplots for the DFM (red) and AR4 (blue) at each forecasting horizon (settings: Recursive estimation scheme, NBER adjustment 'grow', estimation ofDFM5 and AR4 via OLS, first vintage: the first quarter of 1985). The absolute RMSFEs were delivered from normalized residuals of the DFM5 and AR4 were both divided by the standard deviation of the corresponding dependent variable. All three figures are measured on the same scale. For more details, see Section 4.1.

Third, the last column in Table 3 provides the fraction of variables for which the relative improvement of the DFM5 over the benchmark during recessionary periods is more pronounced than for the full sample (see the last row at each horizon). The results indicate that this is the case for 67.1%, 73.1% and 80.6% of all variables at $h = 1$, $h = 2$ and $h = 4$ respectively ($\text{rrRMSFE}^{(h)}_{\text{DFM5,AR4,rec}} < 1$). Conversely, these fractions indicate the shares of variables for which the relative accuracy of the DFM5 is overstated during expansionary periods. Thus, at any given horizon, the overall rRMSFE systematically understates performance during recessions and overstates it during expansions for at least two-thirds of all variables (see, e.g. Siliverstovs (2020)).

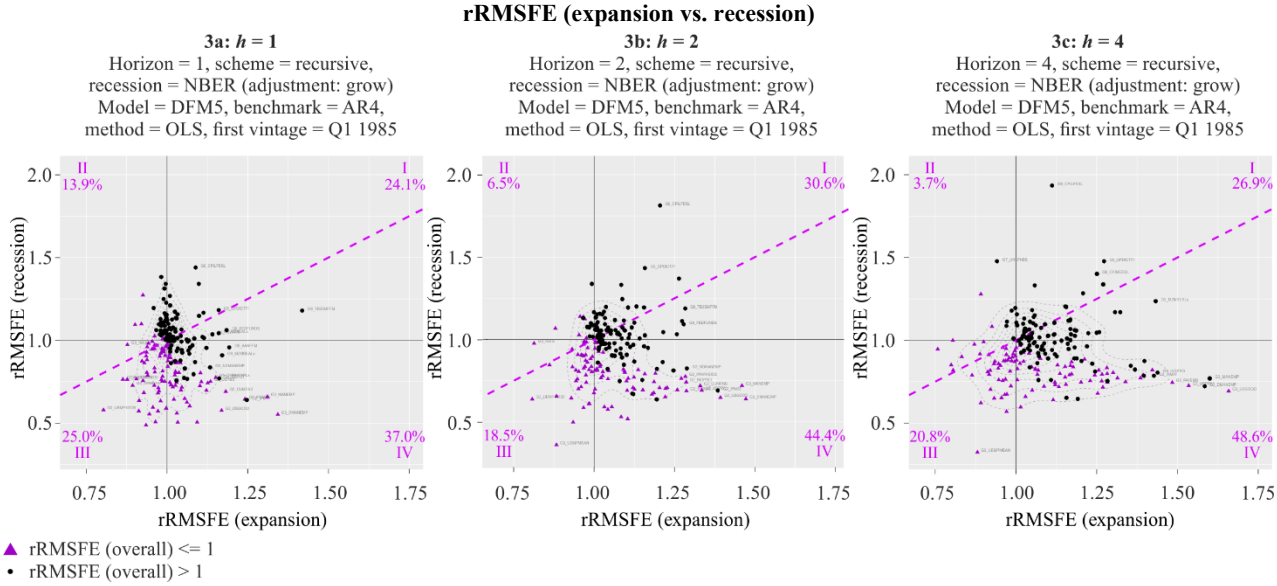
Fourth, the results from the DM-tests in Table 3 further support these findings with statistical evidence. Specifically, the fractions of variables for which the DM test rejects the null hypothesis of equal predictive accuracy is typically substantially higher during recessions than expansions at all forecasting horizons. Hence, the DFM5 and AR4 tend to perform equally well during expansions but statistically significantly differently during recessions (see, e.g. Chauvet and Potter (2013), Fossati (2018), Siliverstovs (2017b) for related findings). In light of the higher variation during recessions, we perceive these results to be a fairly strong indication of systematic differences in predictive performance between the two subsamples.

Bivariate distributions: Joint subsamples over groups

Figure 3 displays the bivariate distributions of rRMSFE for expansions (x-axis) and recessions (y-axis) as scatterplots for all horizons. Each scatterplot is divided into four

quadrants ($rRMSFE_{DFM5,AR4,exp} \lessgtr 1$; $rRMSFE_{DFM5,AR4,rec} \lessgtr 1$).[9] The data points in Figure 3 scatter as follows across the three horizons: the vast majority (about 75%) of the data points lie below the 45°-line, which corresponds to the previously discussed distributional downward shift for the recessionary sub-samples. In each of these cases, recessions help to reduce the overall RMSFE and serve as "breadwinners" to economic forecasters. The DFM5 model is superior [inferior] to the AR4 benchmark for about 20% [30%] of indicators in both subsamples (proportion of points in Quadrant III [I]). For approximately 45% of all series, the DFM5 is benchmark superior to the AR4 only in recessionary but not in expansionary periods (Quadrant IV), and for the remaining small proportion of indicators the DFM5 is inferior during recessions but superior during expansions (Quadrant II).

Figure 3
Scatterplots of rRMSFE in expansions vs. recessions



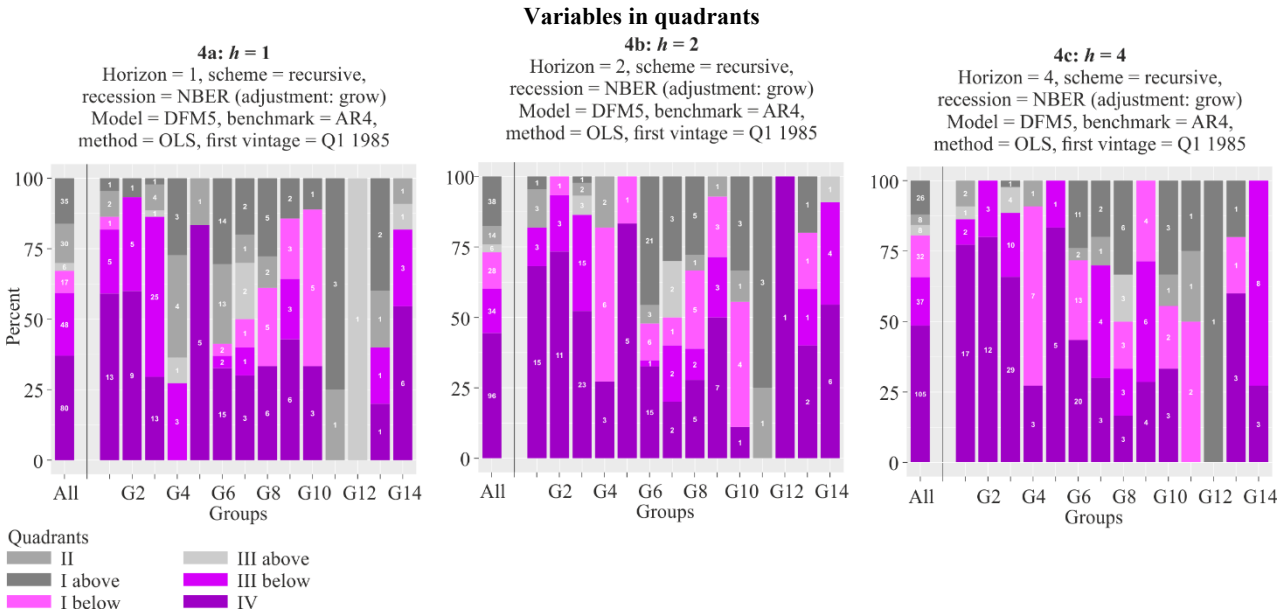
Notes. The figures show the distributions of absolute RMSFEs as density plots and boxplots for the DFM (red) and AR4 (blue) at each forecasting horizon (settings: recursive estimation scheme, NBER adjustment 'grow', estimation of DFM5 and AR4 via OLS, first vintage: the first quarter of 1985). The absolute RMSFEs were derived from normalized residuals, i.e. residuals of the DFM5 and AR4 were both divided by the standard deviation of the corresponding dependent variable. All three figures are measured on the same scale. For more details, see Subsection 4.1.

Figure 4 provides a group-based summary of Figure 3 and summarizes the absolute and relative occurrences of variables in each quadrant by variable groups. The violet bars in Figure 4 correspond to the three distinct quadrant areas below the 45°-line. By classifying the 14 variable groups into four categories (good, good-moderate, moderate-poor, poor performance) based on the DFM5's ability to improve upon the AR4, the figure reveals systematic (cross-sectional) differences between the groups across forecasting horizons: dynamic factor models typically perform better,

[9] The top-right Quadrant I contains the time-series for which the DFM5 fails to improve upon the AR4 in both recessions and expansions. The top-left Quadrant II [bottom-right Quadrant IV] comprises the series for which the DFM5 produces more accurate forecasts during expansions [recessions] but not during recessions [expansions]. The bottom left Quadrant III contains variables for which the DFM5 improves over the AR4 both during expansions and recessions. Furthermore, during recessions the points below the 45°-line have lower rRMSFE than during expansions ($rRMSFE_{DFM5,AR4,rec} < rRMSFE_{DFM5,AR4,exp}$).

especially during recessions, as well as during expansions when predicting indicators of the groups 'National Income and Product Accounts' (G1), 'Industrial Production' (G2), 'Employment and Unemployment' (G3), 'Inventories, Orders and Sales' (G5) as well as 'Non-household Balance Sheets' (G14). For most series, the DFM5 shows good to moderate performance for variables in groups 'Earnings and Productivity' (G7), 'Interest Rates' (G8), 'Money and Credit' (G9) as well as consumer sentiment indices in group 'Others' (G12). Moreover, we find dynamic factor models to perform moderately to poorly for variables from 'Housing' (G4), 'Prices' (G6), 'Households Balance Sheets' (G10) as well as 'Stock Prices' (G13) groups. For a large fraction of indicators in these groups, the DFM5 is often unable to beat AR4 in both recessions and expansions. Finally, the Exchange Rates (G11) dynamics projected by dynamic factor models tend to be worse than those predicted by the AR4 at any time – in recessions or expansions.

Figure 4
Barplot of groups in quadrants



Notes. The figures provide a group-based summary of how the data points (dependent variables) in Figure 6 divide into the four quadrants. Quadrants I and III further distinguish whether the point lies above or below the 45°-line. All violet bars correspond to the areas below the 45°-line and the grey bars – to the areas above the 45°-line. The numbers provided in the bars give the absolute count, and the height of the bar indicates the group-specific fraction of dependent variables in the corresponding quadrant. The height of the leftmost bars for Quadrants III and IV [III and II] together indicate the fraction of variables for which $\text{rRMSFE}_{\text{DFM5,AR4,rec}} < 1$ [$\text{rRMSFE}_{\text{DFM5,AR4,exp}} < 1$] holds. Likewise, the height of the leftmost violet [grey] bars together indicate the fraction of variables for which $\text{rrRMSFE}_{\text{DFM5,AR4,rec}} < 1$ [$\text{rrRMSFE}_{\text{DFM5,AR4,exp}} < 1$] holds. For more details, see Subsection 4.1.

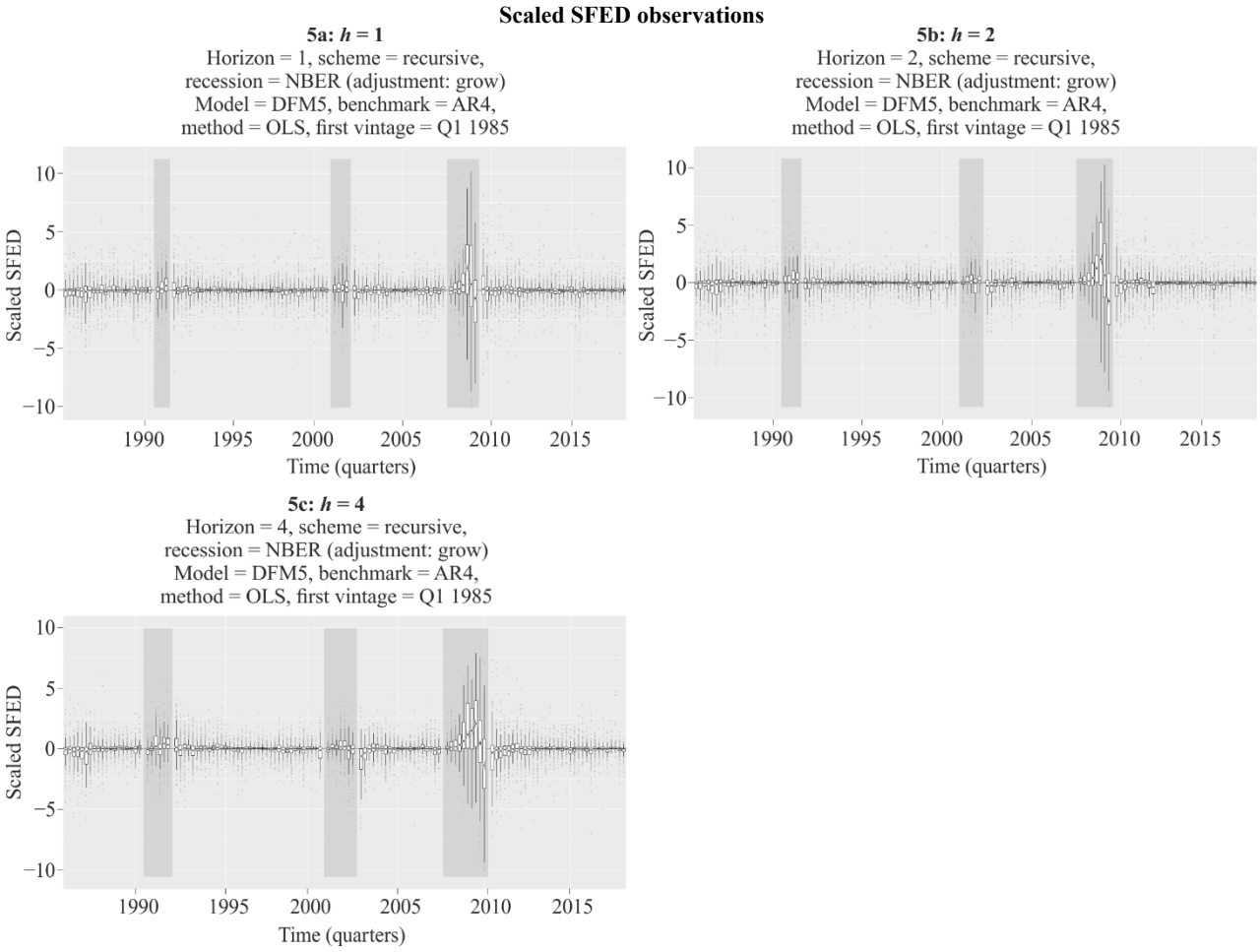
4.1.3 Micro level: Time period specific performance

Individual time periods over all variables

To shift from macro and meso to the micro level, we open the time dimension to the full extent and assess the forecasting performance of variables at each individual time period *t* based on their squared forecast error differences (see Subsection 3.4.3 and, e.g. Siliverstovs (2017a; 2020)). Figure 5 illustrates the histograms of the (unit-variance scaled) SFED of all variables for each individual time period and reveals clearly distinct patterns across the phases of the business cycle. Recall from Subsection 3.4.3 that a positive [negative] SFED indicates superior [inferior]

performance of the DFM5 compared to the AR4. Two features stand out: first, during expansionary periods, the median values of the SFED histograms typically directly hit the zero-value line, whereas the medians are usually positive during recessionary periods. Second, the interquartile range is substantially more concentrated during expansions than during recessions. The figures show surprisingly strong concentrations around the zero line at all horizons, especially after prolonged growth periods, such as the 1990s or 2010s.

Figure 5
Time period specific histograms of scaled SFED



Notes. The figures show for each time period t the histogram of (unit-variance scaled) SFED, $\ddot{\eta}^{(h)}_{t,\text{AR4,DFM5}}$, of all series under the main specifications (settings: recursive estimation scheme, NBER adjustment 'grow', estimation of DFM5 and AR4 via OLS, first vintage: the first quarter of 1985). A median value of the histogram above zero indicates that for more than half of all variables the DFM5 model produces a smaller error than the AR4; and vice versa for values below zero (see Subsection 3.4.3). The grey crosses beyond the histogram's whiskers indicate outliers and the grey shaded rectangles indicate the recessionary periods as defined in Subsection 2.2. For more details, see Subsection 4.1.

This generalizes findings of previous research (e.g. Siliverstovs (2017a; 2020)) and illustrates, on a more granular level, that model performance differs systematically across the business cycle for a large number of target variables.

Individual time periods over individual variables and groups

The most granular assessment opens not only the time-dimension but also the cross-sectional dimension. Figure B.1 (Appendix B) provides the most comprehensive

picture of our analyses and amounts to a three-dimensional illustration depicting the SFED values (z-axis) for each individual time period (x-axis) as well as each individual time-series (y-axis). Red [blue] coloured cells in the heat-plot indicate a positive [negative] SFED value, i.e. superiority [inferiority] of the DFM5 against the AR4 benchmark at the particular time period, whereas white cells indicate equal performance. The last CSSFED value of each series is provided in the rightmost column and corresponds to the sum of all SFED.

Figure B.1 visualizes three key findings. First, consistent with the results above, it displays that relative model performance is state-dependent: in good times most cells are whitish, whereas in bad times the cells typically turn red and for some series blue, which indicates stronger differences in forecast accuracy during recessions than we usually observe during expansions (see, e.g. Siliverstovs (2017a; 2020)). Second, it reveals the two generic drivers of transient and persistent model performance differences: while the former are typically large but temporary, the latter are small but steady (see Figure B.2 and, e.g. Siliverstovs (2017a)). Among these, transient effects appear to be more dominant in Figure B.1.[10] Third, concerning the cross-sectional dimension, the horizontal fine grey lines in Figure B.1 divide the vertically listed variables into the 14 groups and provide a more granular depiction of the group-specific performance.[11] The overall group patterns of red and blue series (CSSFED) also tend to be in line with Stock and Watson (2011)'s assessment that nominal series (such as inflation, exchange rates and prices) are more difficult to forecast than real series (such as industrial production, employment or real manufacturing). While some series naturally diverge from the previously described group trends, the within-group variation is rather limited, which substantiates our previous result of systematic differences in forecasting ability between different variable groups.

4.1.4 10 key macroeconomic indicators

Instead of looking at 216 series, we may narrow our focus on 10 key indicators. In this spirit, we selected 10 leading indicators that are similar or most closely related to those considered in recent studies with multiple dependent variables, such as Kim and Swanson (2014) and Guérin et al. (2018). In particular, Table 4 reports the rRMSFE together with the significance values for (1) GDP: Real Gross Domestic Product, (2) CON: Real Personal Consumption, (3) INV: Real Private Investment, (4) EXP: Real Exports, (5) IMP: Real Imports, (6) UNR: Unemployment Rate, (7) HRS: Business

[10] This argument is qualitatively reflected in the fact that rather few series show persistently superior (or inferior) performance over time, whereas almost all series show pronounced performance differences at a few points in time – typically during recessions. Quantitatively, the proportion of series for which the overall relative performance corresponds to the relative performance in recessions (i.e. for which $\text{sign}(\text{rRMSFE}_{m,b}^{(h)} - 1) = \text{sign}(\text{rRMSFE}_{m,b,rec}^{(h)} - 1)$) amounts to 78.2%, 77.3% and 75.5% for $h = 1$, $h = 2$ and $h = 4$ respectively.

[11] For example, one may wonder why the DFM5 performs systematically worse than the AR4 for particular groups (such as G6 for 'Prices'). A visual analysis of CSSFED patterns shows that a great majority of G6 variables can be attributed to one of the three groups summarized in Figures B.2 (transiency, persistency, combination): first, about 50% of all G6 series show clear transient drivers of performance differences (see G6 CPIAUCSL in Figure B.2d and also, e.g. G6 WPSFD49207, G6 WPSFD49207). For these series, the DFM5 and AR4 perform equally well most of the time except for a few observations in which the DFM5 predicts the actual development (much) worse than the AR4. Second, about 10% of all G6 series show persistent inferiority of the DFM5 (see G6 DDURRG3Q086SBEA in Figure B.2e and also, e.g. G6 DMOTRG3Q086SBEA, G6 CPILFESL) and about 20% show a combination of transient and persistent inferiority (see G6 GDPICTPI in Figure B.2f and also, e.g. G6 CPIMEDSL, G6 CUSR0000SAD). The remaining series are more difficult to qualify clearly. Transient effects therefore also dominate performance differences in group G6.

Hours of All Persons, (8) CPI: Consumer Price Index, (9) RMB: Adjusted Real Monetary Base, (10): PER: Price Earnings Ratio of S&P500.

Table 4
Main results: 10 key macroeconomic indicators

Horizon	Sample	rRMSFE									
		GDP	CON	INV	EXP	IMP	UNR	HRS	CPI	RMB	PER
h = 1	Overall	1.018	1.003*	0.795***	0.980**	0.835***	0.837***	0.899***	1.109	1.015	0.913***
	Expansion	1.166	1.087	0.935**	1.103	1.093	0.964**	1.030	1.000*	1.030	0.923***
	Recession	0.768*	0.794**	0.488***	0.803***	0.505***	0.507***	0.560***	1.338	0.873**	0.875**
h = 2	Overall	0.951**	0.992*	0.820***	1.014	0.767***	0.830***	0.911**	1.212	1.015	0.893***
	Expansion	1.227	1.157	1.055	1.200	0.983*	1.043	1.180	1.088	1.019	0.948**
	Recession	0.707**	0.791**	0.591***	0.818**	0.611**	0.576***	0.668***	1.327	0.937*	0.760**
h = 4	Overall	0.911**	0.954**	0.883**	1.035	0.837***	0.816***	0.891**	1.188	0.983**	0.977**
	Expansion	1.230	1.129	1.121	1.186	0.967**	1.029	1.201	1.158	0.991*	1.121
	Recession	0.756**	0.851**	0.747**	0.887**	0.765**	0.660**	0.747**	1.203	0.918*	0.732**

Notes. The table entries show the rRMSFE of the DFM5 model against the AR4 benchmark for all forecasting horizons for 10 key macroeconomic indicators (settings: OLS estimation, recursive scheme, NBER recession adjustment 'grow'; first vintage: the first quarter of 1985). The entries '*', '**', '***' denote significance of the DM-test (using McCracken (2007)'s critical values) at the 90%, 95% and 99% level respectively. The variable acronyms refer to the following variables (FRED mnemonics are provided in parentheses): GDP: Real Gross Domestic Product (GDPC1), CON: Real Personal Consumption (PCECC96), INV: Real Private Investment (GPDIC1), EXP: Real Exports (EXPGSC1), IMP: Real Imports (IMPGSC1), UNR: Unemployment Rate (UNRATE), HRS: Business Sector Hours of All Persons (HOABS), CPI: Consumer Price Index (CPIAUCSL), RMB: Adjusted Real Monetary Base (AMBSLREALx), PER: S&P 500 Price-Earnings Ratio (S.P.PE.ratio). For more details, see the notes in Tables 2 and 3 as well as Section 4.

Table 4 paints a clear image: for all 10 key indicators (except the CPI[12]), the DFM5 is significantly better than the AR4 in bust phases and typically indistinguishable from the AR4 during boom phases – which is again in accord with the state-dependent forecast evaluation literature (see, e.g. Chauvet and Potter (2013)). Furthermore, similar to Siliverstovs (2020) and Fossati (2018), we see that the average performance for the full evaluation sample is often strongly distorted towards the performance during recessions and can lead to misleading expectations of model performance in terms of size and significance. To see this, consider the 'HRS' variable, for instance: while the overall assessment in Table 4 qualifies the DFM5 model as significantly superior to the benchmark at all three horizons, it is, in fact, worse than the benchmark model most of the time because it is only (substantially and significantly) superior in the few recessionary periods but not in expansionary periods, which make up most of the time periods.

4.2 Extended results

As an extension to the main results above and further deviation from Stock and Watson (2012b), this subsection considers the effects of using Pesaran et al. (2011)'s

[12] The fact that US inflation (CPI) is generally difficult to predict is consistent with previous work (see, e.g. Koop and Potter (2004)). Concerning the significant differences during expansions at $h = 1$, observe that the 90%, 95% and 99% percentiles of McCracken (2007)'s critical values for the settings of the main model are 0.021, 0.308 and 0.979 respectively (see Table 1 in McCracken (2007, p. 729)). Hence, despite the fact that the rRMSFE for CPI during expansions is slightly higher than unity (rRSMFE = 1.0003607) and therefore has a negative DM-statistic (DM = −0.014; determined as weighted loss differential, see McCracken (2007)), the distributional skew arising from nested model comparisons still rejects the null hypothesis of equal predictive accuracy between DFM5 and AR4 at the 10% level.

SURE estimator to account for the MA($h - 1$) process in multi-period ahead predictions. The analyses below extend Pesaran et al. (2011)'s full sample evaluation of the SURE estimator for multiple dependent variables with a state-dependent forecast evaluation (analogously to Subsection 4.1) and over a longer time horizon (1985–2018 instead of 1979–2002).

Table 5
Extended results: rRMSFEs and DM-tests (SURE estimation)

Horizon	Sample	Percentiles of rRMSFE					rRMSFE	rrRMSFE	DM-test
		5%	25%	50%	75%	95%	< 1	< 1	<5%
h = 1	Overall	**0.852**	**0.958**	**1.000**	**1.022**	**1.107**	49.50%	–	44.00%
	Expansion	**0.925**	**0.985**	**1.009**	**1.044**	**1.165**	38.90%	32.90%	28.70%
	Recession	**0.587**	**0.787**	**0.953**	**1.036**	**1.183**	62.00%	67.10%	38.00%
h = 2	Overall	**0.844**	**0.932**	**0.999**	**1.036**	**1.133**	50.90%	–	45.80%
	Expansion	**0.929**	**1.002**	**1.036**	**1.107**	**1.257**	24.50%	24.50%	19.90%
	Recession	**0.596**	**0.775**	**0.913**	**1.034**	**1.162**	65.70%	75.50%	49.10%
h = 4	Overall	**0.809**	**0.888**	**0.963**	**1.023**	**1.171**	64.40%	–	60.60%
	Expansion	**0.832**	**0.969**	**1.019**	**1.092**	**1.258**	38.90%	27.30%	31.50%
	Recession	0.684	0.805	**0.887**	**1.017**	**1.193**	69.40%	72.70%	58.30%

Notes. Analogously to Tables 2 and 3, the table entries show the percentiles of distributions of the rRMSFE of the DFM5 model against the AR4 benchmark as well as the corresponding DM-test results at the 5% level (settings: SURE estimation, recursive scheme, NBER recession adjustment 'grow'; first vintage: the first quarter of 1985). An entry is bold [underlined] if the percentiles [percentages] are equal or smaller [larger] than the corresponding entry in Tables 2 and 3 respectively. For more details, see the notes in Tables 2 and 3 and Subsection 4.2.

Tables 5 and 6 provide the SURE-based counterparts of the OLS-based results in Tables 2–4. To simplify their pairwise comparison, entries are bold [underlined] if the SURE-based percentiles [percentages] are at least as low [high] as the corresponding results for OLS in Tables 2–4. The results in Tables 5 and 6 reveal that our main findings of the previous section are robust to the kind of estimation methodology employed. The SURE-based results paint the very same kind of asymmetries between boom and bust periods in terms of both size and significance. In fact, the DFM5 tends to be even better able to improve upon the AR4 under SURE estimation. These improvements may stem from two sources: they can either be attributed to superior performance of the DFM5(SURE) compared with the DFM5(OLS) and/or the inferior performance of the AR4(SURE) compared to the AR4(OLS). Tables 7 and B.1 shed more light on these sources and show that both effects are relevant in the present case: for multi-period ahead forecasts, the DFM5(SURE) performs better than DFM5(OLS) for about a third of all variables in the full sample, whereas the AR4(SURE) performs worse than AR4(OLS) for the majority of variables (see rRMSFE<1 columns in Tables 7 and B.1).

A close inspection of the DFM5 results highlights two additional aspects. First, Table 7 further corroborates the importance of distinguishing the performance across business cycles: for a majority of variables SURE-based estimations for the DFM5 typically yield more accurate predictions, i.e. in normal times. In turbulent times, however, OLS predictions tend to have smaller errors. Figure 6 provides a more granular description of this fact by decomposing the MSFEs of the top 10 variables into their bias, variance and covariance components based on Theil's inequality (see Subsection 3.4.1). More precisely, the figure illustrates the differences in MSFE components for the DFM(SURE) vs. DFM(OLS) and shows that SURE-based estimations of the DFM5 tend to yield slight improvements over OLS during

expansions but strong deteriorations during recessions, which bends the overall assessment towards the performance in busts. Hence, the conventional overall assessment would misleadingly qualify SURE as inferior when it is in fact better than OLS most of the time. Second, Figure 6 also shows that during expansions, SURE estimation is often able to reduce bias at the expense of an increase in variance and yields typically better co-movements with the actual realizations (see the discussion in Subsection 3.4.1). During recessions, instead, we usually see a deterioration of both bias and variance for the top 10 indicators. Moreover, the results for the AR4 benchmark are provided in the Appendix and show, by and large, comparable patterns as for the DFM5 (see Table B.1 and Figure B.3).[13]

Table 6
Extended results: 10 key macroeconomic indicators (SURE estimation)

Horizon	Sample	rRMSFE									
		GDP	CON	INV	EXP	IMP	UNR	HRS	CPI	RMB	PER
h = 1	Overall	**1.018**	**1.003***	**0.795*****	**0.980****	**0.835*****	**0.837*****	**0.899*****	**1.109**	**1.015**	**0.913*****
	Expansion	**1.166**	**1.087**	**0.935****	**1.103**	**1.093**	**0.964****	**1.03**	**1.000***	**1.03**	**0.923*****
	Recession	**0.768***	**0.794****	**0.488*****	**0.803*****	**0.505*****	**0.507*****	**0.560*****	**1.338**	**0.873****	**0.875****
h = 2	Overall	**0.901****	**0.944****	**0.799*****	**0.999***	**0.733*****	**0.816*****	**0.870****	1.218	1.019	0.959**
	Expansion	**1.208**	**1.122**	1.058	**1.177**	**0.931****	1.062	**1.175**	1.104	1.022	1.037
	Recession	**0.676*****	**0.774*****	**0.583*****	**0.816****	**0.597*****	0.579*****	**0.650*****	**1.314**	0.951	0.771**
h = 4	Overall	**0.848*****	**0.898*****	**0.850*****	**0.992***	**0.815*****	**0.772*****	**0.834*****	1.218	**0.973*****	0.992*
	Expansion	**1.065**	**0.980***	**0.997***	**1.106**	**0.870*****	**0.897*****	**1.025**	1.264	**0.980****	**1.119**
	Recession	0.776*****	0.864****	0.780*****	**0.874****	0.787****	0.693****	0.766*****	**1.195**	0.928*	0.766*****

Notes. Analogously to Table 4, the table entries show the rRMSFE of the DFM5 model against the AR4 benchmark for all forecasting horizons for 10 key macroeconomic indicators (settings: SURE estimation, recursive scheme, NBER recession adjustment 'grow'; first vintage: the first quarter of 1985). A table entry is bold if the rRMSFE under SURE estimation is equal or smaller than the rRMSFE under OLS estimation from Table 4. For more details, see the notes in Tables 2 and 3 as well as Section 4.

Table 7
Distribution of rRMSFE for DFM5(SURE) vs. DFM5(OLS)

Horizon	Sample	Distribution of rRMSFE of DFM5					rRMSFE	rrRMSFE
		5%	25%	50%	75%	95%	<1	<1
h = 1	Overall	1.000	1.000	1.000	1.000	1.000	0.00%	–
	Expansion	1.000	1.000	1.000	1.000	1.000	0.00%	0.00%
	Recession	1.000	1.000	1.000	1.000	1.000	0.00%	0.00%
h = 2	Overall	0.959	0.992	1.008	1.023	1.047	34.70%	–
	Expansion	0.946	0.978	0.998	1.014	1.043	52.30%	71.30%
	Recession	0.951	0.995	1.029	1.056	1.097	28.20%	28.70%
h = 4	Overall	0.921	0.988	1.017	1.056	1.110	36.60%	–
	Expansion	0.867	0.935	0.982	1.037	1.119	58.80%	62.50%
	Recession	0.826	0.981	1.053	1.110	1.202	31.90%	37.50%

Notes. Analogously to Tables 2 and 3, the table entries show the percentiles of distributions of the rRMSFE of the DFM5 model (SURE estimation) against the DFM5 model (OLS estimation) (settings: recursive scheme, NBER recession adjustment 'grow', first vintage: the first quarter of 1985). For more details, see the notes in Table 2 and and Subsection 4.2.

[13] Table B.1 and Figure B.3 also show systematic differences across the business cycles, where the AR4(SURE) tends to perform worse than AR4(OLS) in recessions than in expansions. These differences are particularly pronounced at $h = 2$, but less – at $h = 4$.

In a nutshell, our main results are not critically affected when we account for the MA ($h-1$) process in the errors by means of Pesaran et al. (2011)'s SURE estimator. In particular, we find SURE estimations to be more accurate than OLS in normal times but less accurate in turbulent times.

4.3 Robustness test results

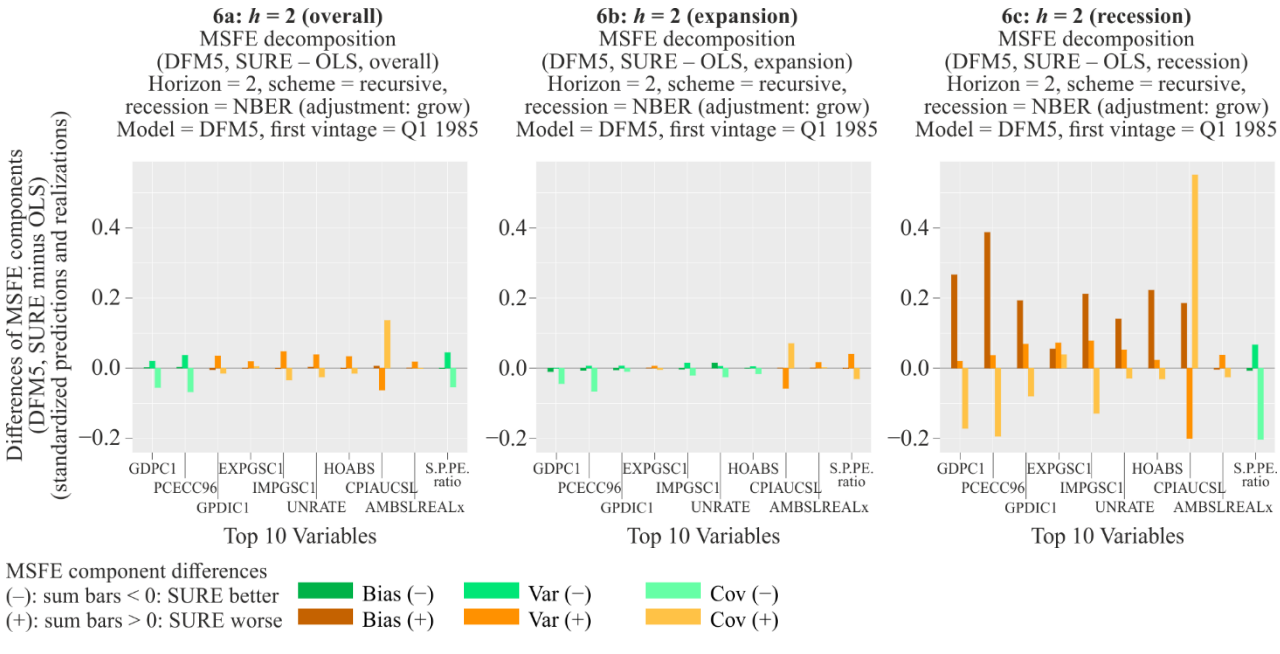
To test the robustness of our main results, this section considers an alternative model and environmental specifications of our forecasting experiment. Each robustness test below considers one particular deviation from the main model specifications in Subsection 4.1. To simplify pairwise comparisons, an entry is again bold [underlined] if the percentile [percentage] entries are equal to or lower [higher] than the corresponding entry for the main results in Tables 2 and 3 respectively.

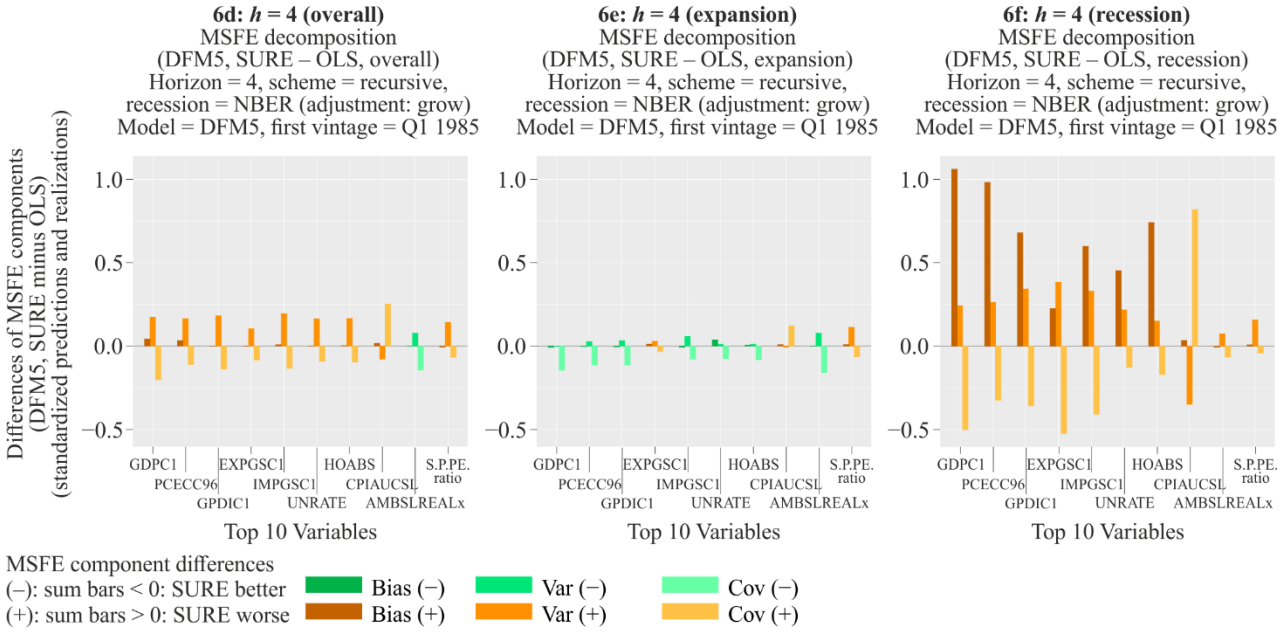
4.3.1 Recession definitions

As explained in Subsection 2.2, we adapted the definition of recessions based on the 'growing' recession adjustment for our main results. This subsection considers two alternatives: the proposed 'shifting' adjustments where both the start and end-date are shifted ($h-1$) periods apart (see Table C.1) as well as the effect of using the original NBER definitions without adjustments (see Table C.2) despite the drawbacks discussed in Subsection 2.2. Also, recall from these discussions that both alternatives do only change the results for multi-period ($h>1$) but not for single-period ($h=1$) ahead predictions.

The results in Tables C.1 and C.2 indicate that the reported main findings are robust to the kind of recessionary definition employed. If anything, we notice a slightly enhanced ability of the DFM5 to outperform the AR4 benchmark in both recessionary and expansionary periods for both alternatives. Yet, this does not materially affect the distinct performance asymmetries across business cycles.

Figure 6
Differences in MSFE components for DFM5(SURE) vs. DFM5(OLS)





Notes. The figures show the differences of MSFE components of the DFM5(SURE) vs. DFM5(OLS) for the top 10 macroeconomic indicators of Subsection 4.1.4 in all (sub-)samples (settings: recursive scheme, NBER recession adjustment 'grow', first vintage: the first quarter of 1985). The MSFEs were decomposed into bias, variance and covariance as outlined in Subsection 3.4.1, and the bars indicate the components' differences (SURE minus OLS). Thus, the bars above [below] the x-axis indicate that SURE has a larger [smaller] error component than OLS. All three component bars of any particular variable are green [orange] if SURE is better [worse] than OLS in terms of MSFE (sum of all components). Moreover, MSFEs were derived from normalized realizations and predictions to have similar scales (using standard deviation of the target variable). For better comparability, the y-axis of all three plots at a particular horizon has the same scale. For more details, see Subsection 4.2.

4.3.2 Rolling estimation scheme

The literature typically considers both recursive and rolling estimation windows (see, e.g. Stock and Watson (2012b), Kim and Swanson (2014)), and Table C.3 provides the results under rolling estimation scheme. The asymmetries between business cycles still persist, but are slightly less pronounced compared to recursively expanding windows in terms of size and significance.

4.3.3 Alternative benchmarks

In addition to the AR4 benchmark, we consider a number of robust alternative specifications, which are frequently employed in the relevant literature (see Subsection 3.2.2). Tables C.4 and C.5 show the estimation results when using the HMN and AR1 benchmarks, and the results indicate even more pronounced differences across the business cycles for these benchmarks. This observation appears to be sensible, because the HMN and AR1 models are (far) more limited in terms of flexibility than our main AR4 benchmark and may therefore not react to strong movements of the underlying series as well (see Siliverstovs (2020); see James et al. (2013) for a general discussion of model flexibility and performance). However, even if the AR4 benchmark is fairly flexible, a natural drawback of AR models is their inability to respond contemporaneously, i.e. they can only react after good or bad news have materialized. This is why we additionally assess the DFM5 against the CADL benchmark, which incorporates, in addition to its AR processes, the contemporaneous

information of all predictors.[14] The results in Table C.6 confirm that the CADL is more competitive against the DFM5. Yet, as can also be seen, the performance differences across business cycles still persist in terms of size and significance at all horizons and therefore leave our main findings unaffected.

4.3.4 Alternative factor models

Tables C.7 and C.8 provide the results for two alternative dynamic factor models with one factor (DFM1) and 10 factors (DFM10) respectively. They show that our main results in Subsection 4.1 are robust to the kind of factor model employed and reveal a similar state-dependent performance. Interestingly, we also see that the DFM1 model tends to perform superior compared to the DFM5 in normal times for almost all variables, whereas the DFM10 shows a mixed performance. This is broadly consistent with our expectations in that the DFM1 is most closely associated with the AR4, which tends to perform reasonably in expansions (see Section 4), whereas the flexible DFM10 may be more prone to overfitting (see James et al. (2013) and also discussions in Kim and Swanson (2016)).

4.3.5 Extension of forecasting window

A major drawback of assessing the forecasting performance for expansions and recessions separately is a general lack of a sufficiently large number of recessionary periods. To address this aspect, we enlarged our forecasting window by shifting its starting period from the first quarter of 1985 to the first quarter of 1980. This allows us to capture two additional recessions in the early 1980s without having to reduce the estimation window too much. Table C.9 provides further evidence in support of our main findings. Specifically, the extension of the forecasting window often results in superior ability of the DFM5 against the AR4 benchmark and have more power to reject the hypothesis of equal predictive accuracy. Hence, the extension of the forecasting window tells the same story.

5. CONCLUSIONS

This study systematically re-examines notable recent contributions on the forecasting performance of dynamic factor models for hundreds of macroeconomic target variables (mainly Stock and Watson (2012b), also Korobilis (2017), Pesaran et al. (2011)) and devotes particular attention to possible performance asymmetries across business cycles in the spirit of Chauvet and Potter (2013) and Siliverstovs (2017a; 2017b; 2020). Our (quasi) real-time forecasting experiment is based on a novel and rich dataset for the US economy (FRED-QD) spanning 216 quarterly indicators for almost 60 years (1960–2018) (see McCracken (2019) and McCracken and Ng (2019b)).

Our results are consistent with the recent business cycle-related evaluation literature (e.g. Chauvet and Potter (2013) and Siliverstovs (2017a; 2017b; 2020)) and systematically broaden their relevance to a large number of macroeconomic indicators in general and 10 key indicators in particular. First, in accord with Chauvet and Potter (2013), we document systematic model performance asymmetries in both absolute

[14] Notice that the DFM5 and CADL are no longer nested in a strict sense. We therefore applied the standard rather than McCracken (2007)'s critical t-values for the DM test, which results in overall fewer significant improvements but still maintains our key finding of pronounced differences between business cycle subsamples.

and relative terms across business cycle phases. On the one hand, for a majority of series, absolute forecasting errors tend to be larger in recessions than in expansions for both dynamic factor models and AR benchmarks (see Chauvet and Potter (2013)). On the other hand, there are clear distributional shifts in relative performance, and dynamic factor models typically perform strongly and statistically significantly better than the benchmark during recessions but only insignificantly differently during expansions (see, e.g. Chauvet and Potter (2013), Siliverstovs (2017b)). We have shown that both performance metrics and tests for the overall sample can be decisively distorted towards the performance of recessionary subsamples and may qualify models as overall significantly superior to the benchmark even if they are inferior most of the time (see, e.g. Fossati (2018), Siliverstovs (2020)). Second, concerning the cross-sectional dimension, we document clear differences in models' predictive ability for different groups of indicators (low within-group heterogeneity). While the DFM performs superior to the AR4 for most of the 14 groups, it tends to perform worse for the series related to exchange rates, stocks, prices and inflation (see also Stock and Watson (2011)). Finally, we find Pesaran et al. (2011)'s GLS-based SURE estimator to yield even more pronounced results in relative terms. In absolute terms, GLS-based model predictions are found to be more accurate than OLS in normal times but less accurate in turbulent times. Moreover, we show that asymmetries persist under a broad range of alternative robustness specifications, which include different recession definitions, estimation schemes, benchmarks, factor models as well as evaluation window sizes.

In a nutshell, our results indicate that the widespread practice of relying on *full sample* forecast evaluation metrics and tests may not be ideal: for at least two thirds of all 216 macroeconomic indicators at any forecast horizon, rRMSFE measures systematically over-estimate DFM model performance in expansions and systematically under-estimate it in contractions (see also Siliverstovs (2017a; 2020)). This is both good and bad news. The good news is that overall evaluations have so far masked the genuine predictive power of factor models: for a majority of variables, they perform best precisely in the periods when public and private sector executives care most about accurate assessments of current and future macroeconomic conditions (see, e.g. Siliverstovs (2017a; 2020), Stock and Watson (2017)). However, in light of the importance of forecasts for policy making (Wieland and Wolters (2013)), the bad news is that overall evaluation measures tend to be critically distorted and may thus give rise to suboptimal or even wrong decisions (see, e.g. Siliverstovs (2017a; 2020)). Our results thus clearly encourage forecasters to explicitly evaluate full *and* subsample performances.

We wish to close with Stock and Watson (2017)'s recent assessment of the forecasting literature: "[D]espite advances in data availability, theory, and computational power, we have not seen dramatic improvements in forecast accuracy over the past decades." (ibid., p. 70). In light of our results, we may only partly agree with this assessment. We disagree on a disaggregated level as we find dramatic improvements in forecast accuracy during recessionary periods for a broad number of macroeconomic indicators. However, we agree on an aggregate level as average forecasting performance measures over the full evaluation period dilute these dramatic improvements and thereby mask substantial improvements in forecast accuracy (see Siliverstovs (2017a; 2020)).

This paper can be advanced along several promising avenues: on the empirical side, we mainly focused on a number of variations of dynamic factor models as well as multiple benchmark specifications, but future research may explore an even richer set of models well-suited to cope with big macroeconomic datasets (see, e.g. Kim and Swanson (2014) for interesting recent work along these lines). Furthermore, the state-dependent evaluation literature has so far predominantly focused on developed countries. Less is known about possible asymmetries in developing countries many of which experienced even more pronounced economic downturns and backlashes over the past decades than developed countries (Chauvet and Potter (2013)). On the conceptual side, alternative metrics and testing procedures are needed that account more directly for business cycle-related asymmetries (see, e.g. Fossati (2018) for notable recent advances in this regard). This may hopefully also shed light on the sources of transiency or persistency in forecasting power. Moreover, the forecast combination literature may be able to put these performance asymmetries to good use (see, e.g. Del Negro et al. (2016) and Kim and Swanson (2016); see Elliott and Timmermann (2016, p. 310 ff.) for a review). Both conceptual and empirical contributions along these lines are likely to enjoy high practical significance over the coming years.

APPENDIX A. Data

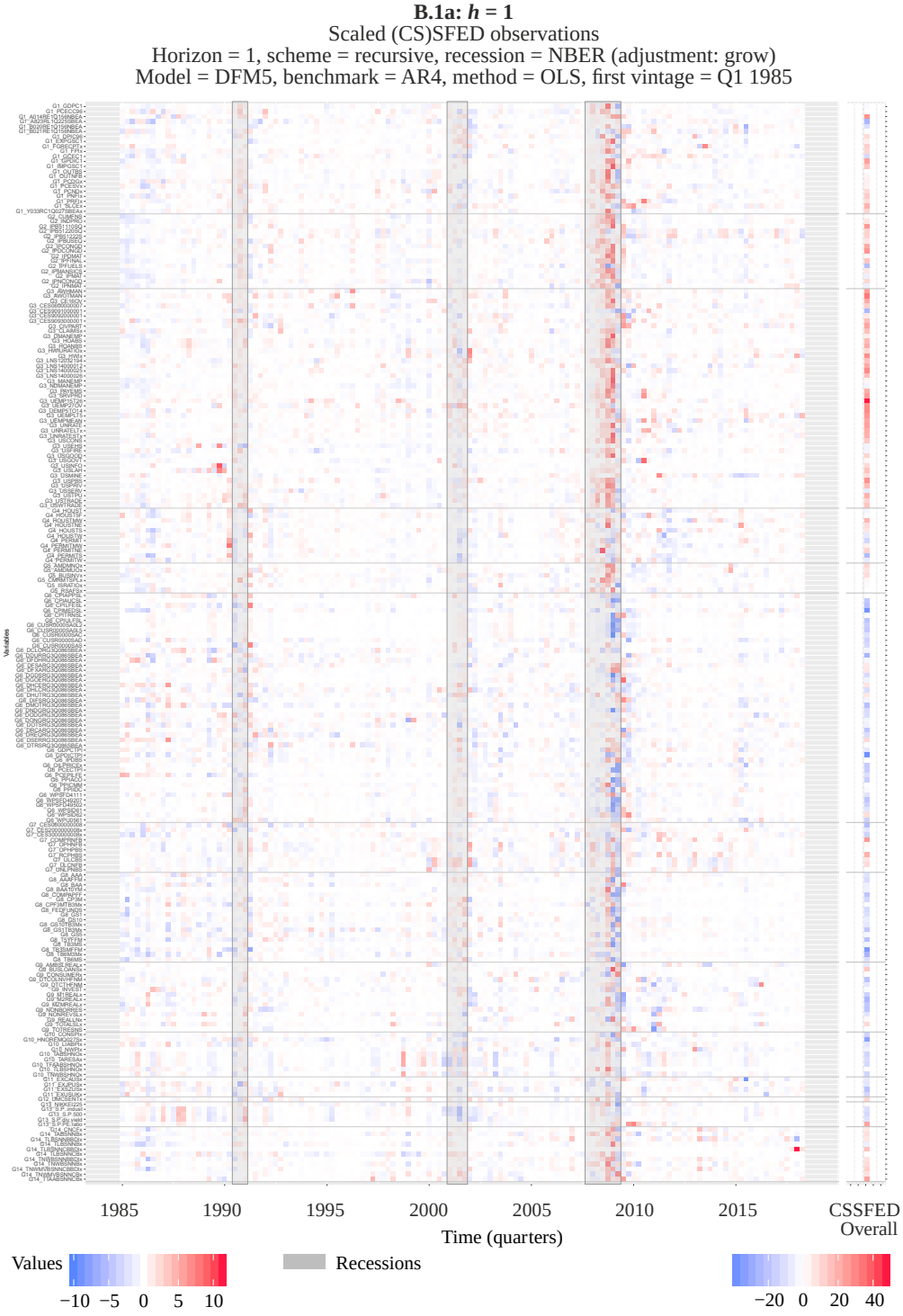
Table A.1
Stationarity transformations

TC	Transformations for $X_{l,t}$
1	$W_{l,t} = Z_{l,t}$
2	$W_{l,t} = Z_{l,t} - Z_{l,t-1}$
3	$W_{l,t} = (Z_{l,t} - Z_{l,t-1}) - (Z_{l,t-1} - Z_{W,t-2})$
4	$W_{l,t} = \ln(Z_{l,t})$
5	$W_{l,t} = \ln(Z_{l,t}) - \ln(Z_{l,t-1})$
6	$W_{l,t} = [\ln(Z_{l,t} - \ln(Z_{l,t-1})] - [\ln(Z_{l,t-1} - \ln(Z_{l,t-2})]$
7	$W_{l,t} = (Z_{l,t}/Z_{l,t-1} - 1) - (Z_{l,t-1})/Z_{l,t-2} - 1)$
TC	Transformations for $Y_t^{(h)}$
1	$Y_t^{(h)} = Z_{k,t}$
2	$Y_t^{(h)} = Z_{k,t} - Z_{k,t-h}$
3	$Y_t^{(h)} = h^{-1}(Z_{k,t} - Z_{k,t-h}) - (Z_{k,t-h} - Z_{k,t-h-1})$
4	$Y_t^{(h)} = \ln(Z_{k,t})$
5	$Y_t^{(h)} = \ln(Z_{k,t}) - \ln(Z_{k,t-h})$
6	$Y_t^{(h)} = h^{-1}[\ln(Z_{k,t} - \ln(Z_{k,t-h})] - [\ln(Z_{k,t-h} - \ln(Z_{k,t-h-1})]$
7	$Y_t^{(h)} = h^{-1}(Z_{k,t}/Z_{k,t-h} - 1) - (Z_{k,t-h})/Z_{k,t-h-1} - 1)$

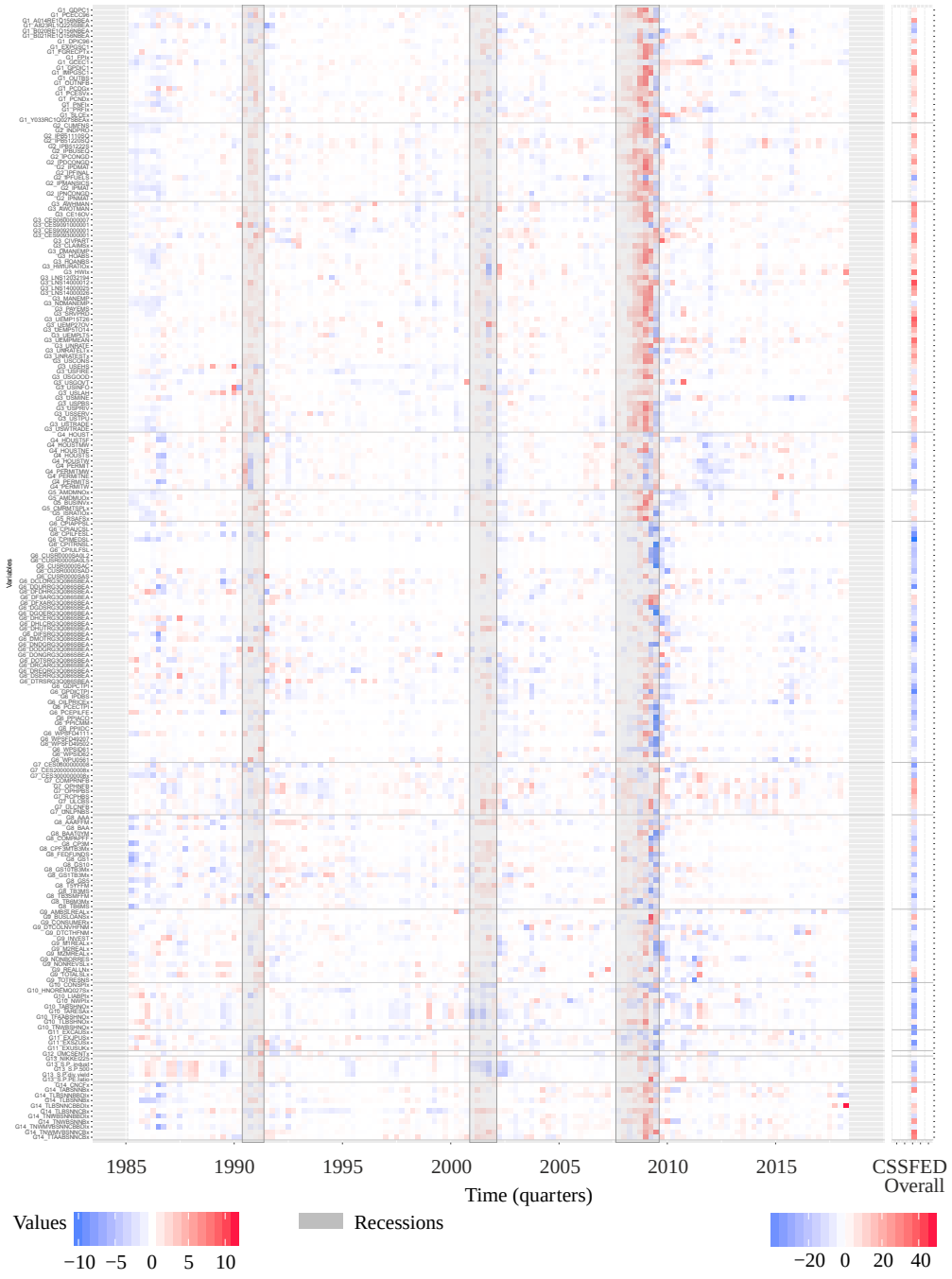
Notes. This table summarizes the stationarity transformations applied to the dependent variables ($Y_t^{(h)}$) and independent variables ($W_{l,t}, l \in \{1, \dots, K\}\backslash\{k\}$) as in Stock and Watson (2012b) (see the online appendix) and Korobilis (2017). TC stands for the corresponding transformation code provided in McCracken and Ng (2019b). The three variables with mnemonics TLBSNNBBDIx, NWPIx and HWIx have very large entries as they were not stationarity transformed (TC = 1). These variables were therefore rescaled by the factors 100'000, 1'000 and 1'000 respectively.

APPENDIX B. Main results

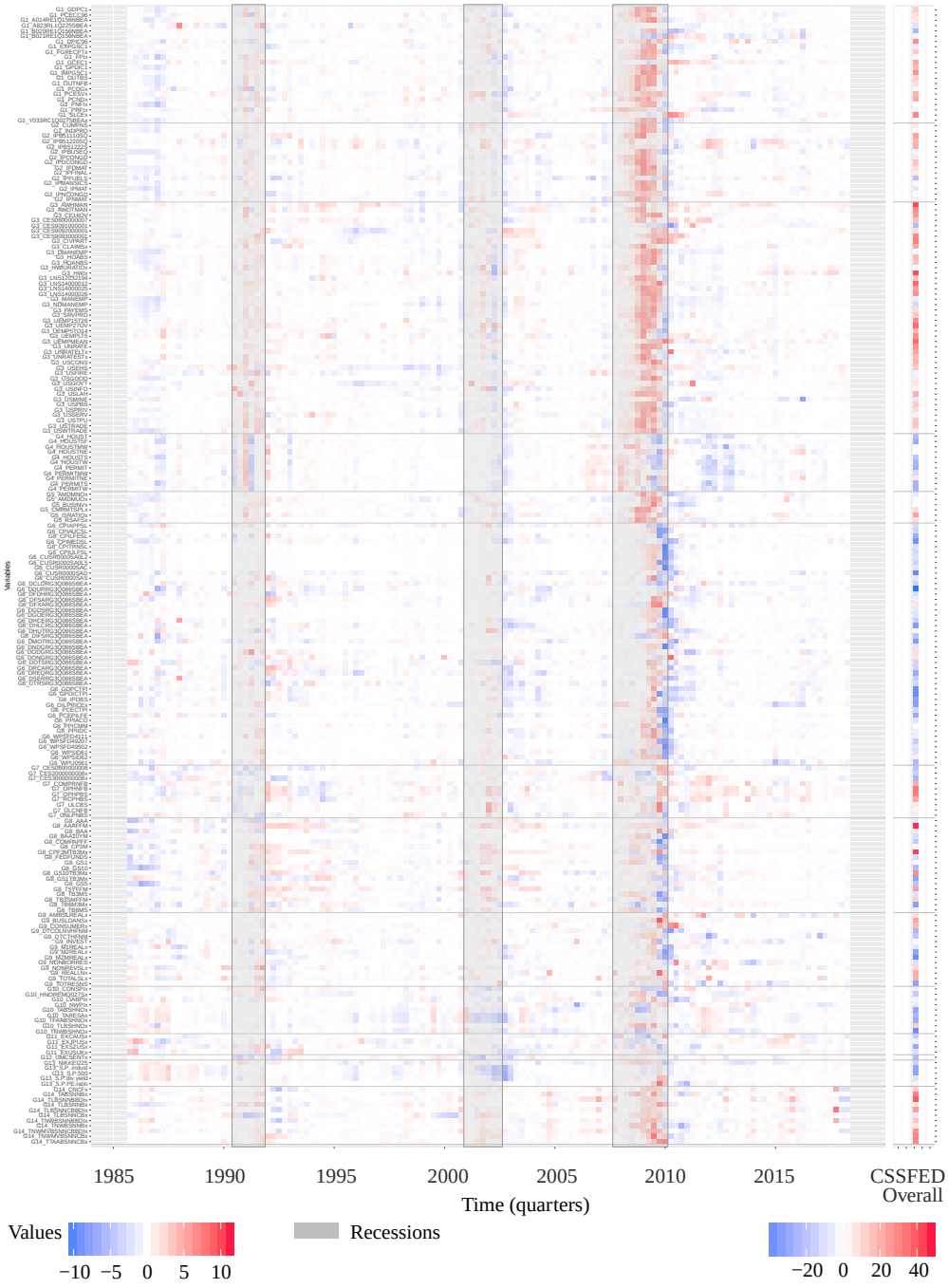
Figure B.1
Heatplots of SFED



B.1b: *h* = 2
Scaled (CS)SFED observations
Horizon = 2, scheme = recursive, recession = NBER (adjustment: grow)
Model = DFM5, benchmark = AR4, method = OLS, first vintage = Q1 1985



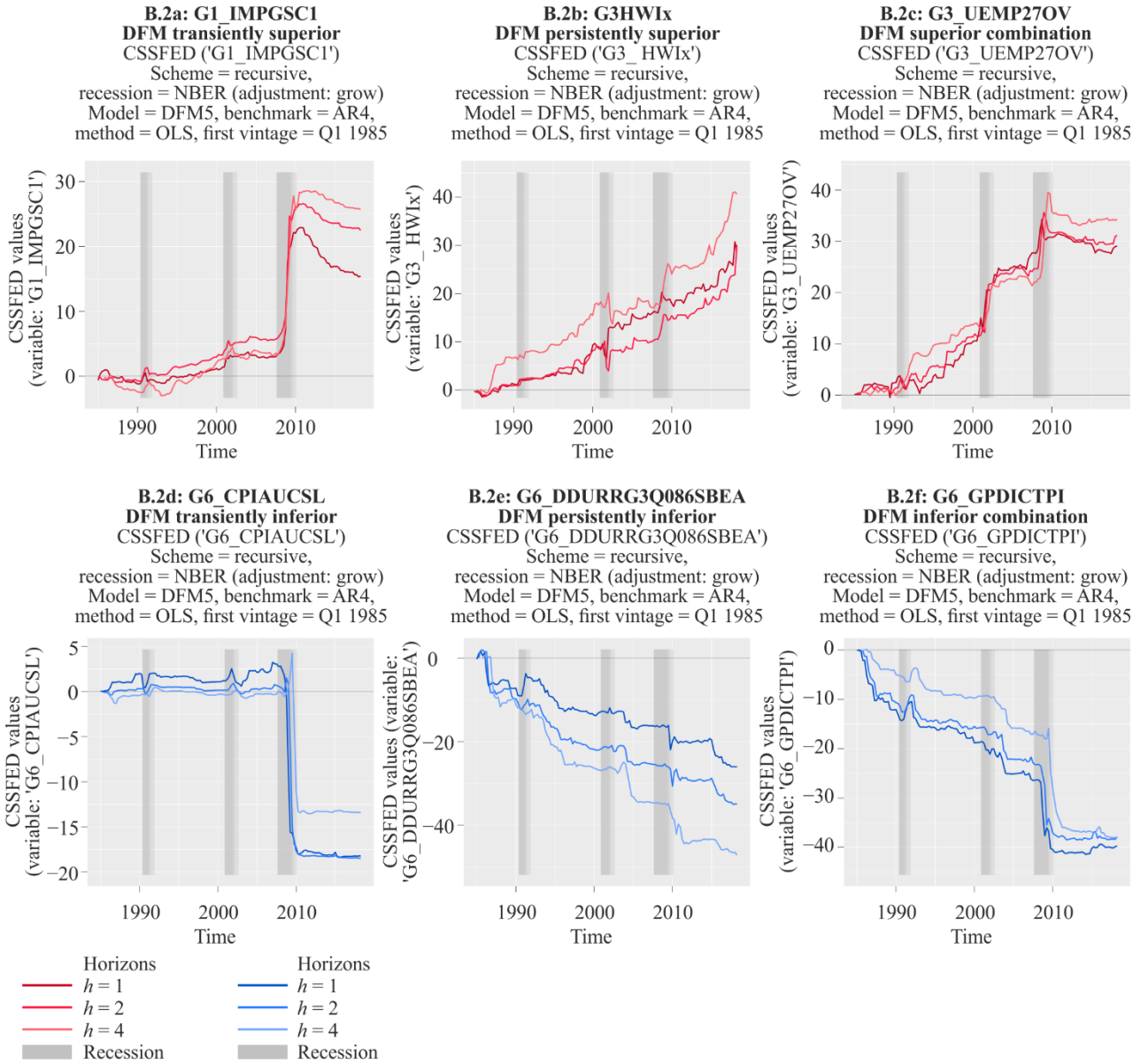
B.1c: $h = 4$
Scaled (CS)SFED observations
Horizon = 4, scheme = recursive, recession = NBER (adjustment: grow)
Model = DFM5, benchmark = AR4, method = OLS, first vintage = Q1 1985



Notes. The figures provide a three-dimensional illustration for the SFED values (z-axis, colour of cells) for each individual time period (x-axis) as well as each individual time-series (y-axis) in the case of the main model specifications (settings: recursive estimation scheme, NBER adjustment 'grow', estimation of DFM5 and AR4 via OLS, first vintage: the first quarter of 1985). The rightmost column illustrates the terminal value of CSSFED of a particular series (see Subsection 3.4.3 for more details). The different variable groups are distinguished by fine grey horizontal lines and the grey shaded areas correspond to recessionary periods as defined in Subsection 2.2. For more details, see Subsection 4.1.

Figure B.2

CSSFED: Transient and persistent model performance differences



Notes. The figures display the CSSFED patterns for six specific variables over the entire forecasting window, $\forall\, t_1 \in \mathcal{S}^{o}_{f,}$ in equation (13) (settings: recursive scheme, NBER recession adjustment 'grow', first vintage: the first quarter of 1985). The left and centre figures depict the four stylized patterns described in Siliverstovs (2017a; 2020) and allows assessing whether the DFM5 performs transiently or persistently superior [inferiorto the AR4 benchmark (see the upper-left and upper-centre figures) [(see the lower-left and lower-centre figures)]. Transient performance improvements [deteriorations] are characterized by a few large upward [downward] movements in CSSFED, whereas persistent performance improvements [deteriorations] show a continuous upward [downward] trend (see, e.g. Siliverstovs (2017a; 2020); notice: Siliverstovs refers to transient improvements as "jumps" and persistent improvements as "trends"). The figures on the right show a combination of both temporary and persistent superiority of the DFM [inferiority] (see the upper-right figure) [(see the lower-right figure)]. For more details, see Subsection 4.1.

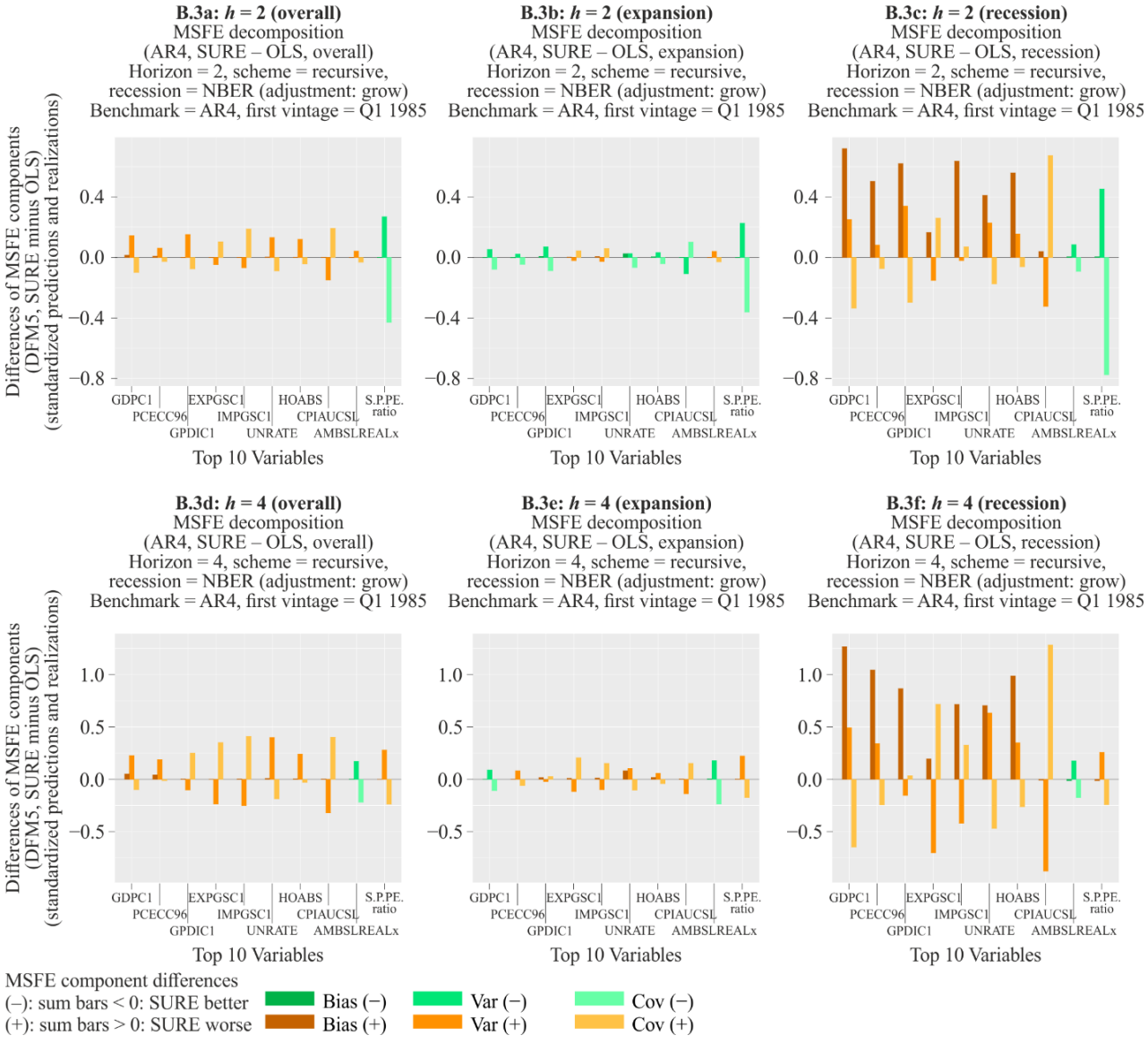
Table B.1

Extended results: Distribution of rRMSFE for AR4(SURE) vs. AR4(OLS)

Horizon	Sample	Distribution of rRMSFE of AR4					rRMSFE	rrRMSFE
		5%	25%	50%	75%	95%	<1	<1
h = 1	Overall	1.000	1.000	1.000	1.000	1.000	0.00%	–
	Expansion	1.000	1.000	1.000	1.000	1.000	0.00%	0.00%
	Recession	1.000	1.000	1.000	1.000	1.000	0.00%	0.00%
h = 2	Overall	0.961	1.000	1.022	1.040	1.071	24.10%	–
	Expansion	0.939	0.976	1.000	1.026	1.059	49.50%	77.80%
	Recession	0.947	1.017	1.048	1.075	1.105	17.60%	22.20%
h = 4	Overall	0.957	1.023	1.065	1.096	1.142	16.70%	–
	Expansion	0.950	1.011	1.048	1.085	1.138	20.80%	53.20%
	Recession	0.890	1.000	1.070	1.114	1.165	25.00%	46.80%

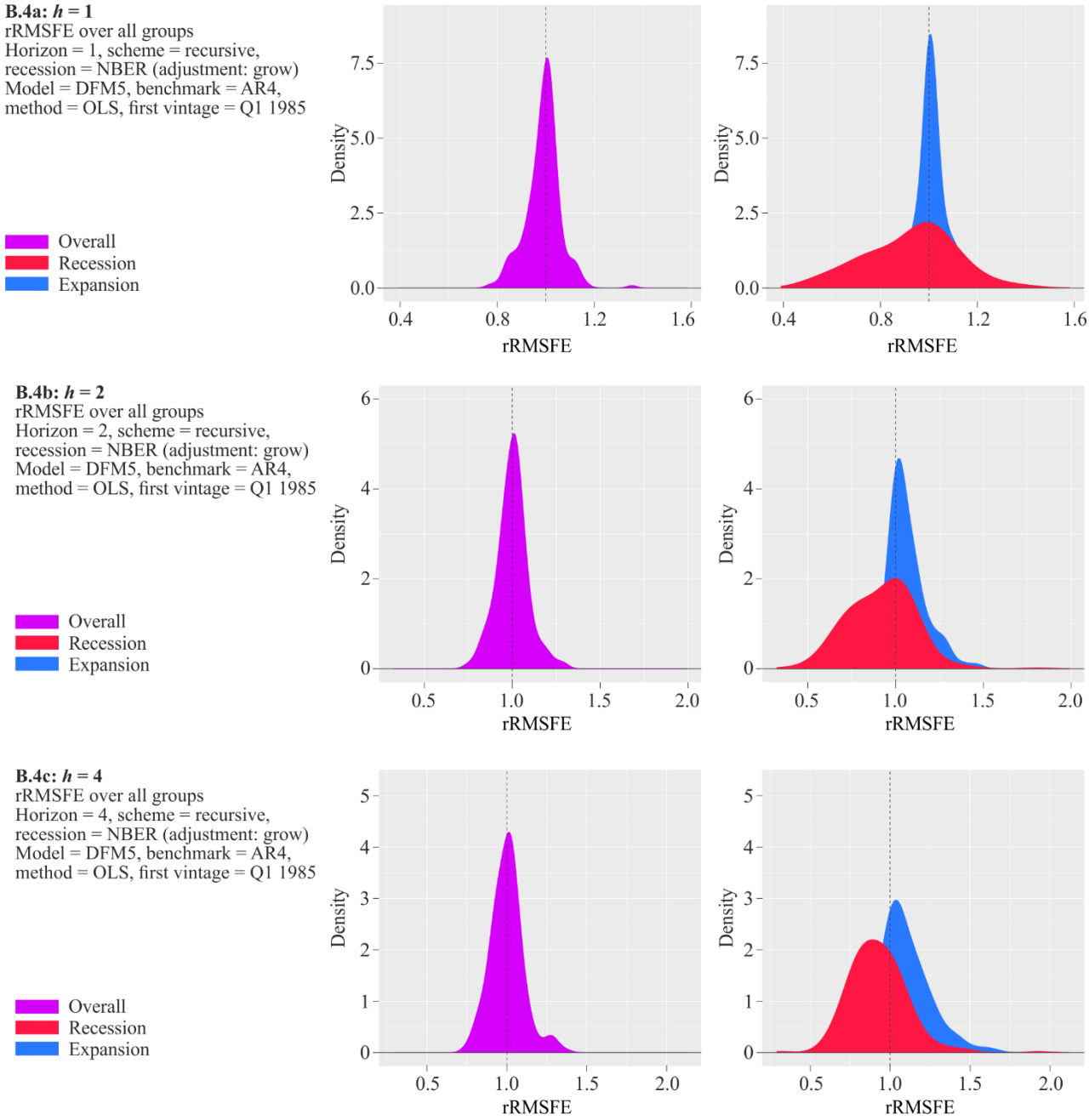
Notes. Analogously to Tables 2 and 3, the table entries show the percentiles of distributions of the rRMSFE of the AR4 model (SURE estimation) against the AR4 model (OLS estimation) (settings: recursive scheme, NBER recession adjustment 'grow', first vintage: the first quarter of 1985). For more details, see notes in Tables 2 and 3 and Subsection 4.2.

Figure B.3
Differences in MSFE components for AR4(SURE) vs. AR4(OLS)



Notes. The figures show the differences of MSFE components of the AR4(SURE) vs. AR4(OLS) for the top 10 macroeconomic indicators of Subsection 4.1.4 in all (sub-)samples (settings: recursive scheme, NBER recession adjustment 'grow'; first vintage: the first quarter of 1985). For more details, see notes under Figure 6 and Subsection 4.2.

Figure B.4
RRMSFE densities (overall, expansions vs. recessions)



Notes. The figure displays (smoothened) density plots of the rRMSFE of the main model specification for the overall evaluation period (left) and for the expansionary and recessionary subsamples (right) respectively (see Subsection 2.2). It corresponds therefore to a visualization of the results contained in Table 2. For more details, see Subsection 4.1.

APPENDIX C. Robustness test results

Table C.1
Robustness tests: Results for shifting recession adjustments

Horizon	Sample	Percentiles of rRMSFE					rRMSFE	rrRMSFE	DM-test
		5%	25%	50%	75%	95%	<1	<1	<5%
h = 1	Overall	**0.852**	**0.958**	**1.000**	**1.022**	**1.107**	49.50%	–	44.00%
	Expansion	**0.925**	**0.985**	**1.009**	**1.044**	**1.165**	38.90%	32.90%	28.70%
	Recession	**0.587**	**0.787**	**0.953**	**1.036**	**1.183**	62.00%	67.10%	35.20%
h = 2	Overall	**0.848**	**0.951**	**1.006**	**1.046**	**1.161**	45.80%	–	40.30%
	Expansion	**0.938**	**0.996**	**1.036**	**1.102**	**1.262**	25.90%	27.30%	20.40%
	Recession	**0.609**	**0.786**	**0.939**	1.050	1.196	64.40%	72.70%	39.40%
h = 4	Overall	**0.840**	**0.934**	**1.006**	**1.047**	**1.186**	47.70%	–	42.60%
	Expansion	**0.906**	**0.994**	**1.048**	**1.127**	**1.268**	28.20%	26.40%	20.80%
	Recession	**0.618**	**0.781**	**0.900**	1.033	1.300	67.10%	73.60%	44.90%

Notes. Analogously to Tables 2 and 3, the table entries show the percentiles of distributions of the rRMSFE of the DFM5 model against the AR4 benchmark as well as the corresponding DM test results at the 5% level (settings: OLS estimation, recursive scheme, NBER recession adjustment 'shift', first vintage: the first quarter of 1985). An entry is bold [underlined] if the percentiles [percentages] are equal or smaller [larger] than the corresponding entry in Tables 2 and 3 respectively. For more details, see notes in Tables 2 and 3 and Subsection 4.3.

Table C.2
Robustness tests: Results without recession adjustments

Horizon	Sample	Percentiles of rRMSFE					rRMSFE	rrRMSFE	DM-test
		5%	25%	50%	75%	95%	<1	<1	<5%
h = 1	Overall	**0.852**	**0.958**	**1.000**	**1.022**	**1.107**	49.50%	–	44.00%
	Expansion	**0.925**	**0.985**	**1.009**	**1.044**	**1.165**	38.90%	32.90%	28.70%
	Recession	**0.587**	**0.787**	**0.953**	**1.036**	**1.183**	62.00%	67.10%	35.20%
h = 2	Overall	**0.848**	**0.951**	**1.006**	**1.046**	**1.161**	45.80%	–	40.30%
	Expansion	**0.937**	**0.995**	1.042	1.118	1.293	26.90%	22.20%	20.40%
	Recession	**0.599**	**0.744**	**0.930**	**1.027**	**1.155**	68.10%	77.80%	44.40%
h = 4	Overall	**0.840**	**0.934**	**1.006**	**1.047**	**1.186**	47.70%	–	42.60%
	Expansion	**0.876**	**0.981**	**1.045**	**1.118**	**1.284**	31.00%	20.40%	23.60%
	Recession	0.698	**0.786**	**0.891**	**0.975**	**1.145**	79.20%	79.60%	58.80%

Notes. Analogously to Tables 2 and 3, the table entries show the percentiles of distributions of the rRMSFE of the DFM5 model against the AR4 benchmark as well as the corresponding DM test results at the 5% level (settings: OLS estimation, recursive scheme, NBER recession adjustment 'none', first vintage: the first quarter of 1985). An entry is bold [underlined] if the percentiles [percentages] are equal or smaller [larger] than the corresponding entry in Tables 2 and 3 respectively. For more details, see notes in Tables 2 and 3 and Subsection 4.3.

Table C.3
Robustness tests: Results for rolling scheme

Horizon	Sample	Percentiles of rRMSFE					rRMSFE	rrRMSFE	DM-test
		5%	25%	50%	75%	95%	<1	<1	<5%
h = 1	Overall	0.868	**0.957**	1.007	1.037	**1.092**	46.80%	–	47.70%
	Expansion	**0.912**	**0.973**	**1.006**	**1.038**	**1.126**	44.90%	40.30%	45.80%
	Recession	0.705	0.871	0.982	1.053	**1.181**	56.00%	59.70%	28.20%
h = 2	Overall	0.874	0.980	1.029	1.064	**1.128**	33.30%	–	35.60%
	Expansion	**0.914**	**0.992**	**1.033**	**1.075**	**1.182**	28.20%	42.10%	29.20%
	Recession	0.759	0.892	0.997	1.105	1.216	50.90%	57.90%	24.50%
h = 4	Overall	0.873	0.966	1.019	1.065	**1.140**	39.80%	–	44.00%
	Expansion	**0.883**	**0.978**	**1.033**	**1.109**	**1.248**	33.80%	38.90%	33.80%
	Recession	0.744	0.901	0.986	1.067	1.224	52.80%	61.10%	28.20%

Notes. Analogously to Tables 2 and 3, the table entries show the percentiles of distributions of the rRMSFE of the DFM5 model against the AR4 benchmark as well as the corresponding DM test results at the 5% level (settings: OLS estimation, rolling scheme, NBER recession adjustment 'grow', first vintage: the first quarter of 1985). An entry is bold [underlined] if the percentiles [percentages] are equal or smaller [larger] than the corresponding entry in Tables 2 and 3 respectively. For more details, see notes in Tables 2 and 3 and Subsection 4.3.

Table C.4
Robustness tests: Results for HMN benchmark

Horizon	Sample	Percentiles of rRMSFE					rRMSFE	rrRMSFE	DM-test
		5%	25%	50%	75%	95%	<1	<1	<5%
h = 1	Overall	**0.392**	**0.734**	**0.907**	**0.984**	**1.054**	79.60%	–	80.10%
	Expansion	**0.446**	**0.815**	**0.943**	**1.009**	**1.104**	73.10%	36.60%	72.70%
	Recession	**0.296**	**0.515**	**0.851**	**1.004**	1.303	74.10%	63.40%	56.90%
h = 2	Overall	**0.486**	**0.735**	**0.888**	**1.021**	**1.101**	69.90%	–	70.40%
	Expansion	**0.517**	**0.800**	**0.965**	**1.056**	**1.223**	61.60%	34.30%	61.60%
	Recession	**0.399**	**0.584**	**0.826**	**1.015**	1.291	73.60%	65.70%	58.30%
h = 4	Overall	**0.578**	**0.767**	**0.894**	**1.022**	**1.123**	71.80%	–	71.80%
	Expansion	**0.565**	**0.841**	**0.995**	**1.103**	**1.274**	51.90%	28.70%	51.90%
	Recession	**0.525**	**0.680**	**0.836**	**1.002**	**1.192**	74.50%	71.30%	60.60%

Notes. Analogously to Tables 2 and 3, the table entries show the percentiles of distributions of the rRMSFE of the DFM5 model against the HMN benchmark as well as the corresponding DM test results at the 5% level (settings: OLS estimation, recursive scheme, NBER recession adjustment 'grow', first vintage: the first quarter of 1985). An entry is bold [underlined] if the percentiles [percentages] are equal or smaller [larger] than the corresponding entry in Tables 2 and 3 respectively. For more details, see notes in Tables 2 and 3 and Subsection 4.3.

Table C.5
Robustness tests: Results for AR1 benchmark

Horizon	Sample	Percentiles of rRMSFE					rRMSFE	rrRMSFE	DM-test
		5%	25%	50%	75%	95%	<1	<1	<5%
h = 1	Overall	**0.846**	**0.938**	**0.987**	1.026	**1.105**	61.60%	–	61.10%
	Expansion	**0.900**	**0.965**	**1.002**	1.048	1.180	47.20%	32.90%	43.50%
	Recession	**0.581**	**0.769**	**0.935**	1.042	1.223	63.00%	67.10%	43.50%
h = 2	Overall	**0.825**	**0.920**	**0.991**	**1.045**	1.169	54.60%	–	54.20%
	Expansion	**0.872**	**0.981**	1.041	1.118	1.290	31.50%	24.50%	28.70%
	Recession	**0.614**	**0.763**	**0.916**	**1.041**	1.210	68.10%	75.50%	49.50%
h = 4	Overall	**0.803**	**0.916**	**0.985**	1.053	**1.167**	54.20%	–	54.20%
	Expansion	**0.844**	**0.986**	1.072	1.183	**1.395**	30.60%	15.30%	27.80%
	Recession	**0.649**	**0.786**	**0.875**	**1.013**	**1.192**	72.20%	84.70%	55.10%

Notes. Analogously to Tables 2 and 3, the table entries show the percentiles of distributions of the rRMSFE of the DFM5 model against the AR1 benchmark as well as the corresponding DM test results at the 5% level (settings: OLS estimation, recursive scheme, NBER recession adjustment 'grow', first vintage: the first quarter of 1985). An entry is bold [underlined] if the percentiles [percentages] are equal or smaller [larger] than the corresponding entry in Tables 2 and 3 respectively. For more details, see notes in Table 2 and Table 3 and Subsection 4.3.

Table C.6
Robustness tests: Results for CADL benchmark

Horizon	Sample	Percentiles of rRMSFE					rRMSFE	rrRMSFE	DM-test
		5%	25%	50%	75%	95%	<1	<1	<5%
h = 1	Overall	0.887	0.970	1.005	1.027	**1.104**	43.50%	–	10.20%
	Expansion	0.936	0.992	1.014	1.049	1.167	33.30%	33.30%	4.20%
	Recession	0.630	0.824	0.971	1.044	**1.174**	58.80%	66.70%	22.20%
h = 2	Overall	0.878	0.970	1.013	1.051	**1.149**	41.20%	–	3.20%
	Expansion	0.964	1.009	1.044	1.120	**1.282**	21.30%	26.90%	1.90%
	Recession	0.671	0.814	0.970	**1.045**	**1.164**	60.20%	73.10%	20.40%
h = 4	Overall	0.864	0.952	1.017	1.057	**1.173**	43.50%	–	5.10%
	Expansion	0.932	1.011	1.067	1.193	**1.391**	21.80%	19.40%	2.30%
	Recession	0.708	0.834	0.932	1.035	**1.181**	65.70%	80.60%	16.70%

Notes. Analogously to Tables 2 and 3, the table entries show the percentiles of distributions of the rRMSFE of the DFM5 model against the CADL benchmark as well as the corresponding DM test results at the 5% level (settings: OLS estimation, recursive scheme, NBER recession adjustment 'grow', first vintage: the first quarter of 1985). An entry is bold [underlined] if the percentiles [percentages] are equal or smaller [larger] than the corresponding entry in Tables 2 and 3 respectively. Apart from Table 3, this table uses the conventional critical *t*-values, because the DFM5 and CADL are not nested in a strict sense. For more details, see notes in Tables 2 and 3 and Subsection 4.3.

Table C.7
Robustness tests: Results for DFM1

Horizon	Sample	Percentiles of rRMSFE					rRMSFE	rrRMSFE	DM-test
		5%	25%	50%	75%	95%	<1	<1	<5%
h = 1	Overall	0.921	0.983	1.003	**1.020**	**1.078**	42.60%	–	24.50%
	Expansion	0.961	0.990	**1.003**	**1.021**	**1.151**	40.70%	41.70%	23.60%
	Recession	0.686	0.865	0.976	1.043	1.251	56.90%	58.30%	36.10%
h = 2	Overall	0.921	0.982	1.010	**1.028**	**1.091**	35.60%	–	24.50%
	Expansion	0.951	**0.994**	**1.009**	**1.041**	**1.215**	30.10%	38.00%	16.20%
	Recession	0.717	0.875	0.990	1.058	1.190	54.60%	62.00%	35.60%
h = 4	Overall	0.903	0.983	1.009	**1.041**	**1.119**	37.00%	–	19.00%
	Expansion	0.935	**0.999**	**1.019**	**1.082**	**1.219**	26.40%	30.10%	16.20%
	Recession	0.730	0.907	0.994	1.044	**1.188**	53.70%	69.90%	38.40%

Notes. Analogously to Tables 2 and 3, the table entries show the percentiles of distributions of the rRMSFE of the DFM1 model against the AR4 benchmark as well as the corresponding DMtest results at the 5% level (settings: OLS estimation, recursive scheme, NBER recession adjustment 'grow', first vintage: the first quarter of 1985). An entry is bold [underlined] if the percentiles [percentages] are equal or smaller [larger] than the corresponding entry in Tables 2 and 3 respectively. For more details, see notes in Tables 2 and 3 and Subsection 4.3.

Table C.8
Robustness tests: Results for DFM10

Horizon	Sample	Percentiles of rRMSFE					rRMSFE	rrRMSFE	DM-test
		5%	25%	50%	75%	95%	<1	<1	<5%
h = 1	Overall	0.860	**0.953**	1.005	1.040	1.151	45.80%	–	51.40%
	Expansion	**0.923**	**0.981**	1.021	1.070	1.212	36.10%	32.40%	38.00%
	Recession	**0.558**	**0.779**	0.958	1.066	1.234	58.30%	67.60%	42.60%
h = 2	Overall	**0.839**	0.952	1.017	1.064	1.170	42.60%	–	45.40%
	Expansion	**0.933**	**0.998**	1.054	1.131	1.294	25.50%	28.70%	26.90%
	Recession	**0.598**	**0.775**	**0.937**	1.068	1.285	59.30%	71.30%	47.20%
h = 4	Overall	0.848	0.946	1.018	1.073	1.222	43.10%	–	46.80%
	Expansion	0.928	1.016	1.087	1.202	1.418	18.50%	19.90%	20.40%
	Recession	**0.661**	0.807	**0.902**	1.050	1.233	62.50%	80.10%	50.50%

Notes. Analogously to Tables 2 and 3, the table entries show the percentiles of distributions of the rRMSFE of the DFM10 model against the AR4 benchmark as well as the corresponding DM test results at the 5% level (settings: OLS estimation, recursive scheme, NBER recession adjustment 'grow', first vintage: the first quarter of 1985). An entry is bold [underlined] if the percentiles [percentages] are equal or smaller [larger] than the corresponding entry in Tables 2 and 3 respectively. For more details, see notes in Tables 2 and 3 and Subsection 4.3.

Table C.9
Robustness tests: Results for forecasting window extension

Horizon	Sample	Percentiles of rRMSFE					rRMSFE	rrRMSFE	DM-test
		5%	25%	50%	75%	95%	<1	<1	<5%
h = 1	Overall	**0.848**	**0.927**	**0.986**	**1.019**	**1.051**	58.30%	–	57.40%
	Expansion	**0.910**	**0.975**	**1.005**	**1.029**	**1.093**	46.30%	27.30%	40.70%
	Recession	0.711	0.812	**0.944**	**1.013**	**1.106**	70.80%	72.70%	45.80%
h = 2	Overall	**0.834**	**0.913**	**0.990**	**1.035**	**1.094**	52.80%	–	51.90%
	Expansion	**0.901**	**0.983**	**1.033**	**1.076**	**1.168**	32.40%	26.90%	31.00%
	Recession	0.699	0.802	**0.934**	**1.031**	**1.130**	64.40%	73.10%	47.70%
h = 4	Overall	**0.829**	**0.908**	**0.992**	**1.046**	**1.109**	53.20%	–	52.30%
	Expansion	**0.866**	**0.997**	**1.055**	**1.121**	**1.308**	26.40%	19.90%	24.10%
	Recession	0.709	0.806	0.927	**1.013**	**1.103**	70.80%	80.10%	50.90%

Notes. Analogously to Tables 2 and 3, the table entries show the percentiles of distributions of the rRMSFE of the DFM5 model against the AR4 benchmark as well as the corresponding DM test results at the 5% level (settings: OLS estimation, recursive scheme, NBER recession adjustment 'grow', first vintage: the first quarter of 1980). An entry is bold [underlined] if the percentiles [percentages] are equal or smaller [larger] than the corresponding entry in Tables 2 and 3 respectively. For more details, see notes in Tables 2 and 3 and Subsection 4.3.

BIBLIOGRAPHY

CHAN, Joshua C. C. (2017). *Notes on Bayesian Macroeconometrics*. Version 1.4, June 2017. 140 p. Retrieved 13.02.2019 from *http://joshuachan.org/papers/BayesMacro.pdf*.

CHAN, Joshua C. C., HSIAO, Cody Yu-Ling. (2014). Estimation of Stochastic Volatility Models with Heavy Tails and Serial Dependence. *In: Bayesian Inference in the Social Sciences*. Ed. by Ivan Jeliazkov and Xin-She Yang. John Wiley & Sons, Inc., September 2014, pp. 155–176.

CHAUVET, Marcelle, POTTER, Simon (2013). Forecasting Output. *In: Handbook of Economic Forecasting.* Elsevier, vol. 2, pp. 141–194.

DEL NEGRO, Marco, HASEGAWA, Raiden B., SCHORFHEIDE, Frank (2016). Dynamic Prediction Pools: An Investigation of Financial Frictions and Forecasting Performance. *Journal of Econometrics*, Elsevier, vol. 192, issue 2, pp. 391–405.

DIEBOLD, Francis X., MARIANO, Robert S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, vol. 20, issue 1, pp. 134–144 (online January 2012).

ELLIOTT, Graham, TIMMERMANN, Allan (2016). *Economic Forecasting*. Princeton University Press, May 2016. 568 p.

FOSSATI, Sebastian (2018). *A Test for State-Dependent Predictive Ability Based on a Markov-Switching Framework.* University of Alberta, May 2018. 39 p.

GUÉRIN, Pierre, LEIVA-LEON, Danilo, MARCELLINO, Massimiliano (2018). Markov-Switching Three-Pass Regression Filter. *Journal of Business and Economic Statistics*, October 2018, pp. 1–18.

HILL, R. Carter, GRIFFITHS, William E., LIM, Guay C. (2011). *Principles of Econometrics*. Fourth Edition. John Wiley & Sons, Inc. 762 p.

HYNDMAN, Rob J., ATHANASOPOULOS, George, BERGMEIR, Christoph, CACERES, Gabriel, CHHAY, Leanne, O'HARA-WILD, Mitchell, PETROPOULOS, Fotios, RAZBASH, Slava, WANG, Earo, YASMEEN, Farah, R Core Team, IHAKA, Ross, REID, Daniel, SHAUB, David, TANG, Yuan, ZHOU, Zhenyu (2019). Package 'forecast'. *CRAN R Project*, December 2019. 143 p. Retrieved from *https://cran.r-project.org/web/packages/ forecast/forecast.pdf*.

JAMES, Gareth, WITTEN, Daniela, HASTIE, Trevor, TIBSHIRANI, Robert (2013). *An Introduction to Statistical Learning: with Applications in R.* Springer, 426 p.

KIM, Hyun Hak, SWANSON, Norman R. (2014). Forecasting Financial and Macroeconomic Variables Using Data Reduction Methods: New Empirical Evidence. *Journal of Econometrics*, vol. 178, issue P2, pp. 352–367.

KIM, Kihwan, SWANSON, Norman R. (2016). *Mixing Mixed Frequency and Diffusion Indices in Good Times and in Bad.* Korea Energy Economics Institute Working Paper, August 2016. 39 p.

KOOP, Gary, KOROBILIS, Dimitris (2010). Bayesian Multivariate Time Series Methods for Empirical Macroeconomics. *Foundations and Trends® in Econometrics*, vol. 3, issue 4, 2009, pp. 267–358.

KOOP, Gary, POTTER, Simon (2004). Forecasting in Dynamic Factor Models Using Bayesian Model Averaging. *The Econometrics Journal*, vol. 7, issue 2, pp. 550–565.

KOROBILIS, Dimitris (2017). *Forecasting with Many Predictors Using Message Passing Algorithms.* Retrieved 15.12.2018 from *https://ssrn.com/abstract=2977838*.

MCCRACKEN, Michael W. (2007). Asymptotics for Out of Sample Tests of Granger Causality. *Journal of Econometrics*, vol. 140, issue 2, pp. 719–752.

MCCRACKEN, Michael W. (2019). *FRED-MD: A Monthly Database for Macroeconomic Research.* Federal Reserve Bank of St. Louis. Retrieved 12.03.2019 from *https://research.stlouisfed.org/econ/mccracken/fred-databases/*.

MCCRACKEN, Michael W., NG, Serena (2016). FRED-MD: A Monthly Database for Macroeconomic Research. *Journal of Business & Economic Statistics*, vol. 34, pp. 574–589.

MCCRACKEN, Michael W., NG, Serena (2019a). *FRED-MD: A Monthly Database for Macroeconomic Research.* Federal Reserve Bank of St. Louis. Retrieved 12.03.2019 from *https://s3.amazonaws.com/files.fred.stlouisfed.org/fred-md/Appendix_Tables_Update.pdf* (Online Appendix for FRED-MD. 4 p.)

MCCRACKEN, Michael W., NG, Serena (2019b). *FRED-QD: A Quarterly Database for Macroeconomic Research.* Federal Reserve Bank of St. Louis. Retrieved 12.03.2019 from *https://s3.amazonaws.com/files.fred.stlouisfed.org/fred-md/FRED-QDappendix.pdf* (Online Appendix for FRED-QD).

NEWEY, Whitney K., WEST, Kenneth D. (1994). Automatic Lag Selection in Covariance Matrix Estimation. *The Review of Economic Studies*, vol. 61, issue 4, October 1994, pp. 631–653.

PESARAN, M. Hashem, PICK, Andreas, TIMMERMANN, Allan (2011). Variable Selection, Estimation and Inference for Multi-Period Forecasting Problems. *Journal of Econometrics*, vol. 164, issue 1, September 2011, pp. 173–187.

PINDYCK, Robert S., RUBINFELD, Daniel L. (1998). *Econometric Models and Economic Forecasts*. 4th edition. McGraw-Hill international editions, Economics series, Boston, Mass.: Irwin/McGraw-Hill. 634 p.

R Core Team (2019). R: A Language and Environment for Statistical Computing [Computer software manual]. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from *https://www.R-project.org/* (R Package 'stats').

ROSSI, Barbara (2013). Advances in Forecasting under Instability. *In: Handbook of Economic Forecasting.* Elsevier, vol. 2, pp. 1203–1324.

SILIVERSTOVS, Boriss (2017a). Dissecting Models' Forecasting Performance. *Economic Modelling*, Elsevier, vol. 67, issue C, pp. 294–299.

SILIVERSTOVS, Boriss (2017b). International Stock Return Predictability: On the Role of the United States in Bad and Good Times. *Applied Economics Letters*, vol. 24, issue 11, pp. 771–773.

SILIVERSTOVS, Boriss (2020). Assessing Nowcast Accuracy of US GDP Growth in Real Time: The Role of Booms and Busts. *Empirical Economics*, Springer, vol. 58, issue 1, January 2020, pp. 7–27.

STOCK, James H., WATSON, Mark W. (2002a). Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association*, vol. 97, No. 460, pp. 1167–1179.

STOCK, James H., WATSON, Mark W. (2002b). Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business & Economic Statistics*, vol. 20, issue 2, April 2002, pp. 147–162.

STOCK, James H., WATSON, Mark W. (2006). Forecasting with Many Predictors. *In: Handbook of Economic Forecasting*, vol. 1, pp. 515–554.

STOCK, James H., WATSON, Mark W. (2011). Dynamic Factor Models. *In: The Oxford Handbook of Economic Forecasting*, pp. 35–59.

STOCK, James H., WATSON, Mark W. (2012a). *Disentangling the Channels of the 2007–2009 Recession.* National Bureau of Economic Research, Inc., NBER Working Paper, No. 18094, May 2012. 53 p. Retrieved from *https://ideas.repec.org/p/nbr/nberwo/18094.html*.

STOCK, James H., WATSON, Mark W. (2012b). Generalized Shrinkage Methods for Forecasting Using Many Predictors. *Journal of Business & Economic Statistics*, vol. 30, issue 4, pp. 481–493.

STOCK, James H., WATSON, Mark W. (2016). Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics. *Handbook of Macroeconomics*, vol. 2, pp. 415–525.

STOCK, James H., WATSON, Mark W. (2017). Twenty Years of Time Series Econometrics in Ten Pictures. *Journal of Economic Perspectives*, vol. 31, issue 2, pp. 59–86.

TERÄSVIRTA, Timo (2018). Nonlinear models in macroeconometrics. *In: Oxford Research Encyclopedias in Economics and Finance.* Oxford University Press. 25 p.

THEIL, Henri (1958). Economic Forecasts and Policy. Amsterdam: North-Holland Pub. Co. 562 p.

WELCH, Ivo, GOYAL, Amit (2008). A Comprehensive Look at the Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies*, vol. 21, issue 4, July 2008, pp. 1455–1508.

WIELAND, Volker, WOLTERS, Maik (2013). Forecasting and Policy Making. *In: Handbook of Economic Forecasting.* Ed. by G. Elliott, C. Granger and A. Timmermann. Elsevier, ed. 1, vol. 2, pp. 239–325.

ZEILEIS, Achim, LUMLEY, Thomas, BERGER, Susanne, GRAHAM, Nathaniel (2019). *Package 'Sandwich'.* CRAN – R Project, April 2019. 42 p. Retrieved from *https://cran.r-project.org/web/packages/sandwich/sandwich.pdf*.